IBM

International Technical Support Organization

Implementing High Availability Cluster Multi-Processing (HACMP) Cookbook

Octavian Lascu
Shawn Bodily
Maria-Katharina Esser
Michael Herrera
Patrick Pothier
Dusan Prelec
Dino Quintero
Ken Raymond
Viktor Sebesteny
Andrei Socoliuc
Antony (Red) Steel

December 2005

Note: Before using this information and the product it supports, read the information in “Notices”.

First Edition (December 2005)

This edition applies to Version 5, Release 3, of IBM High Availability Cluster Multi-Processing (product number 5765-F62).

© Copyright International Business Machines Corporation 2005. All rights reserved.
Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

IBM

Международная организация технической поддержки

Рецепты по реализации High Availability Cluster Multi Processing (HACMP)

Октавиан Ласку
Шон Бодили
Мария-Катарина Эссер
Майкл Херрера
Патрик Потье
Дюсан Прелец
Дино Квинтеро
Кен Реймонд
Виктор Себестени
Андрей Соцолиус
Энтони (Ред) Стил

Москва 2007

Перед использованием данного руководства и продукта ознакомьтесь с информацией в разделе «Примечания».

Переводчик
М. Аницкий

Научный редактор
IBM Certified Advanced Technical Expert – IBM System p5
Д. Миронов

Первое издание на русском языке (2007 г.)

Издание охватывает IBM High Availability Cluster Multi-Processing, версию 5, релиз 3 (номер продукта 5765-F62).

© Copyright International Business Machines Corporation 2005. All rights reserved.
Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

© IBM Corporation (Корпорация International Business Machines Corporation), 2005 г.
Все права защищены.

Ограничение прав пользователей правительством США: использование, копирование и распространение сведений настоящего руководства ограничено условиями контракта GSA ADP Schedule с корпорацией IBM.

Содержание

Предисловие ..... **xvi**

О людях, написавших эту книгу xvi

Приглашаем авторов к сотрудничеству xix

Мы будем рады вашим отзывам.
Ваши отзывы очень важны для нас! xxi

Часть 1. Введение **1**

Глава 1. Введение в HACMP **3**

1.1 Что такое HACMP 4
1.1.1 Высокая доступность *4*
1.1.2 Кластерная мультиобработка *4*

1.2 Решения по обеспечению доступности: обзор 5
1.2.1 Время отключения (простоя) *7*
1.2.2 Единая точка отказа *7*

1.3 История и эволюция 9
1.3.1 HACMP Version 4 ..... *9*
1.3.2 HACMP Version 5 Release 1 ..... *9*
1.3.3 HACMP Version 5 Release 2 ..... *10*
1.3.4 HACMP Version 5 Release 3 ..... *11*

1.4 Терминология и понятия высокой доступности 12
1.4.1 Терминология *12*
1.4.2 Понятия *12*

1.5 Высокая доступность и отказоустойчивость 14
1.5.1 Отказоустойчивые системы *14*
1.5.2 Системы высокой доступности *14*

1.6 Планирование программного обеспечения 15
1.6.1 Уровень AIX и соответствующие требования *15*
1.6.2 Лицензирование *17*

1.7 Установка программного обеспечения HACMP 17
1.7.1 Проверка соответствия требованиям *18*
1.7.2 Новая установка *18*
1.7.3 Установка HACMP *19*
1.7.4 Пути и варианты миграции *20*
1.7.5 Преобразование снимка кластера *21*
1.7.6 Поузловая миграция *24*
1.7.7 Опции обновления *30*

Глава 2. Составляющие высокой доступности **33**

2.1 Данные конфигурации HACMP 34

2.2 Компоненты программного обеспечения 35

2.3 Топология кластера 37

2.3.1 Мониторинг пульса RSCT и HACMP *42*
2.3.2 Мониторинг пульса через IP-синонимы *48*
2.3.3 Сети TCP/IP *51*
2.3.4 Механизмы перехвата IP-адреса *52*
2.3.5 Постоянная IP-метка/адрес *55*
2.3.6 Сети устройств, или Последовательные сети *56*
2.3.7 Сетевые модули *59*
2.3.8 Клиенты *60*
2.3.9 Аспекты сетевой безопасности *61*

2.4 Ресурсы и группы ресурсов 63
2.4.1 Определения *63*
2.4.2 Ресурсы *64*
2.4.3 NFS *75*
2.4.4 Серверы приложений *76*
2.4.5 Мониторы приложений *77*
2.4.6 Коммуникационные адаптеры и каналы *80*
2.4.7 Накопители на магнитной ленте *80*
2.4.8 Ресурсы Fast Connect *80*
2.4.9 Интеграция с WLM *81*
2.4.10 Группы ресурсов *82*

2.5 Подключаемые модули HACMP 98

2.6 Возможности (HACMP 5.1, 5.2 и 5.3) 99
2.6.1 Новые возможности *99*
2.6.2 Неподдерживаемые возможности *102*

2.7 Ограничения 102

2.8 Аспекты хранения 103
2.8.1 Общий LVM *106*
2.8.2 Режим неодновременного доступа *107*
2.8.3 Режим одновременного доступа *109*
2.8.4 Группы томов в режиме расширенного одновременного доступа *110*
2.8.5 Быстрый перехват дисков *112*

2.9 Конфигурация хранилища с общим доступом 113
2.9.1 Требования общего LVM *114*
2.9.2 Неодновременный доступ, одновременный доступ и расширенный одновременный доступ *114*

Часть 2. Планирование, установка и миграция 119

Глава 3. Планирование 121

3.1 Планирование высокой доступности 122

3.2 Планирование HACMP 123
3.2.1 Инструменты планирования *124*

3.3 Приступая к работе 125
3.3.1 Текущая среда *125*
3.3.2 Устранение единых точек отказа *127*
3.3.3 Первоначальная схема кластера *127*
3.3.4 Заполнение обзорной таблицы планирования кластера *129*

3.4 Планирование оборудования кластера 130
3.4.1 Заполнение таблицы планирования оборудования кластера *131*

3.5 Планирование программного обеспечения кластера 132

3.5.1 Уровни AIX и RSCT ... *132*
3.5.2 Поддержка Virtual LAN и Virtual SCSI ... *132*
3.5.3 Обязательные наборы файлов ... *132*
3.5.4 Наборы файлов безопасности AIX ... *133*
3.5.5 Программное обеспечение, необходимое для работы WebSmit ... *133*
3.5.6 Наборы файлов HACMP ... *134*
3.5.7 Файлы AIX, изменяемые HACMP ... *136*
3.5.8 Программное обеспечение приложений ... *140*
3.5.9 Лицензирование ... *140*
3.5.10 Заполнение таблицы планирования программного обеспечения ... *141*

3.6 Аспекты операционной системы ... 141

3.7 Планирование безопасности ... 142
3.7.1 Безопасность кластера ... *142*
3.7.2 Администрирование пользователей ... *145*
3.7.3 Группа HACMP ... *145*
3.7.4 IP-порты HACMP ... *146*
3.7.5 Планирование наборов файлов HACMP ... *146*

3.8 Планирование сетей кластера ... 147
3.8.1 Терминология ... *148*
3.8.2 Общие аспекты сети ... *149*
3.8.3 Планирование перехвата IP-адреса ... *157*
3.8.4 Мониторинг пульса через синонимы ... *161*
3.8.5 Планирование сети, отличной от IP ... *163*
3.8.6 Планирование последовательных сетей RS232 ... *166*
3.8.7 Планирование мониторинга пульса через диски ... *167*
3.8.8 Дополнительные аспекты планирования сети ... *168*
3.8.9 Заполнение таблиц планирования сети ... *171*

3.9 Планирование требований к хранилищу ... 173
3.9.1 Внутренние диски ... *173*
3.9.2 Общие диски ... *173*
3.9.3 Пример конфигурации дисков ... *174*
3.9.4 Группы томов в режиме расширенного одновременного доступа (ECM) ... *175*
3.9.5 Общие логические тома ... *176*
3.9.6 Быстрый перехват диска ... *177*
3.9.7 Заполнение таблиц планирования хранения ... *178*

3.10 Планирование приложений ... 179
3.10.1 Серверы приложений ... *181*
3.10.2 Мониторинг приложения ... *181*
3.10.3 Инструмент анализа доступности ... *182*
3.10.4 Приложения, интегрируемые с HACMP ... *182*
3.10.5 Заполнение таблиц планирования приложений ... *182*

3.11 Планирование групп ресурсов ... 184
3.11.1 Атрибуты группы ресурсов ... *186*
3.11.2 Заполнение таблицы планирования групп ресурсов ... *188*

3.12 Подробная схема кластера ... 189

3.13 Разработка плана тестирования кластера ... 189
3.13.1 Настраиваемый план тестирования ... *189*
3.13.2 Инструмент Cluster Test Tool ... *191*

3.14 Разработка плана установки HACMP ... 194

3.15 Резервное копирование конфигурации кластера ... 195

3.16 Документирование кластера ... 196

3.16.1 Экспорт файла определения кластера с применением SMIT *197*
3.16.2 Создание файла определения кластера из снимка с использованием SMIT *198*
3.16.3 Создание отчета о конфигурации *198*

3.17 Управление изменениями и проблемами 200

3.18 Инструменты планирования 200
3.18.1 Схема кластера *200*
3.18.2 Система автоматизированного планирования *201*
3.18.3 Таблицы планирования на бумаге *206*

Глава 4. Сценарии установки кластера 207

4.1 Основные этапы внедрения кластера HACMP 208

4.2 Установка и конфигурирование WebSMIT 210
4.2.1 Установка веб-сервера Apache и обязательных пакетов *210*
4.2.2 Конфигурирование WebSMIT *212*
4.2.3 Запуск веб-сервера Apache *219*
4.2.4 Доступ к страницам WebSMIT из браузера *220*
4.2.5 Введение в WebSMIT *221*
4.2.6 Меню WebSMIT: конфигурирование и управление кластером *222*

4.3 Конфигурирование HACMP 223
4.3.1 Стандартный путь конфигурирования – Two-Node Configuration Assistant *228*
4.3.2 Использование расширенного пути конфигурирования и C-SPOC *234*

Глава 5. Миграция кластера на HACMP V5.3 241

5.1 Выбор пути миграции 242
5.1.1 Поддерживаемые пути миграции *242*

5.2 Требования 242

5.3 Аспекты 245

5.4 Основные этапы миграции 247

5.5 Протестированные сценарии 249
5.5.1 Сценарий 1: AIX 5.1 и HAES 4.5 ..... *249*
5.5.2 Сценарий 2: AIX 5.2 и HA 5.1 ..... *253*
5.5.3 Сценарий 3: AIX 5.2 и HA 5.2 ..... *256*

5.6 Действия после миграции 260

5.7 Устранение неполадок при отказе миграции 261
5.7.1 Возврат после отказа миграции *261*
5.7.2 Информация о версии кластера в HACMP ODM *263*
5.7.3 Устранение неполадок при использовании снимков *264*
5.7.4 Ошибка DARE при синхронизации *264*
5.7.5 Ошибка «config_too_long» во время миграции *265*

Часть 3. Сценарии и задачи администрирования кластера 267

Глава 6. Сценарий: добавление двух узлов в кластер 269

6.1 Конфигурация из двух узлов 270
6.1.1 Топология кластера *270*
6.1.2 Ресурсы кластера *271*

6.2 Конфигурация из четырех узлов 274

6.2.1 Топология кластера ... *275*
6.2.2 Конфигурация сетей пульса через диски ... *276*
6.2.3 Конфигурация групп ресурсов ... *276*

6.3 Реконфигурирование кластера ... 280
6.3.1 Выполнение предварительных требований ... *280*
6.3.2 Добавление новых узлов в кластер ... *281*
6.3.3 Конфигурирование топологии кластера ... *282*
6.3.4 Конфигурирование ресурсов ... *288*
6.3.5 Запуск HACMP на новых узлах ... *293*

Глава 7. Обслуживание кластера ... 295

7.1 Контроль и тестирование изменений ... 296
7.1.1 Тестовый кластер ... *296*

7.2 Запуск и остановка кластера ... 297
7.2.1 Службы кластера ... *298*
7.2.2 Запуск служб кластера ... *300*
7.2.3 Остановка служб кластера ... *302*

7.3 Управление группами ресурсов и приложениями ... 304
7.3.1 Перевод группы ресурсов в отключенное состояние через SMIT ... *305*
7.3.2 Перевод группы ресурсов в подключенное состояние через SMIT ... *306*
7.3.3 Перемещение группы ресурсов через SMIT ... *307*
7.3.4 Расположение, отменяющее приоритет ... *309*
7.3.5 Приостановка/возобновление мониторинга приложения ... *310*

7.4 Сценарии ... 311
7.4.1 «Горячая замена» сетевой карты PCI ... *311*
7.4.2 Исправления ... *314*
7.4.3 Хранение ... *316*
7.4.4 Приложения ... *317*

Глава 8. Управление кластером ... 319

8.1 CSPOC DP ... 320
8.1.1 Общее описание C-SPOC ... *320*
8.1.2 Меню C-SPOC SMIT ... *321*

8.2 Наборы файлов ... 322
8.2.1 Предопределенные наборы файлов ... *323*
8.2.2 Управление наборами файлов ... *325*

8.3 Администрирование пользователей ... 330
8.3.1 Администрирование пользователей и групп через CSPOC ... *331*
8.3.2 Управление паролями ... *340*

8.4 Управление общим хранением ... 345
8.4.1 Обновление компонентов LVM ... *345*
8.4.2 Диспетчер логических томов C-SPOC ... *349*
8.4.3 Управление логическими томами с одновременным доступом через C-SPOC ... *350*
8.4.4 Управление физическими томами через C-SPOC ... *351*
8.4.5 Примеры ... *352*

8.5 Синхронизация времени ... 363

8.6 Верификация и синхронизация кластера ... 364
8.6.1 Верификация и синхронизация кластера через SMIT ... *364*
8.6.2 Динамическая реконфигурация кластера – DARE ... *368*
8.6.3 Файлы журналов верификации ... *369*

8.6.4 Автоматическое выполнение корректирующих действий во время верификации ... 370
8.6.5 Автоматическая верификация кластера ... 372

8.7 Мониторинг HACMP ... 373
8.7.1 Утилиты проверки состояния кластера ... 374
8.7.2 Утилиты проверки состояния кластера и служб ... 376
8.7.3 Команды информации о топологии ... 377
8.7.4 Команды информации о группах ресурсов ... 379
8.7.5 Файлы журналов ... 380
8.7.6 Уведомление об ошибках ... 382
8.7.7 Мониторинг приложения ... 383
8.7.8 Измерение доступности приложения ... 386

Глава 9. Безопасность кластера ... 389

9.1 Безопасность кластера и демона clcomd ... 390
9.1.1 Файл /usr/es/sbin/cluster/etc/rhosts ... 391
9.1.2 Отключение демона коммуникаций кластера ... 391
9.1.3 Дополнительные средства безопасности кластера ... 391

9.2 Использование зашифрованных межузловых коммуникаций ... 392
9.2.1 Управление ключами шифрования ... 392
9.2.2 Настройка шифрования сообщений ... 393
9.2.3 Устранение неполадок аутентификации и шифрования сообщений ... 399
9.2.4 Проверка текущих параметров аутентификации сообщений ... 399

9.3 Защищенное удаленное выполнение команд в HACMP ... 400
9.3.1 Установка SSH ... 401
9.3.2 Настройка SSH для удаленного выполнения команд без пароля ... 402

9.4 Безопасность WebSmit ... 403
9.4.1 Параметры безопасности WebSMIT ... 403
9.4.2 Разрешение или запрет на доступ к определенным панелям SMIT в WebSMIT ... 405
9.4.3 Журналы WebSMIT ... 406

9.5 HACMP и брандмауэры ... 406

9.6 Безопасность RSCT ... 407
9.6.1 RSCT и HACMP ... 407
9.6.2 Обзор служб безопасности кластера (CtSec) ... 409
9.6.3 Компоненты служб безопасности кластера (CtSec) ... 410
9.6.4 Абстрактный уровень механизма (MAL) ... 410
9.6.5 Подключаемый модуль механизма (MPM) ... 411
9.6.6 Подключаемый модуль UNIX MPM ... 411
9.6.7 Аутентификация на основе хостов с использованием ctcasd ... 411
9.6.8 Служба IDM (Identity Mapping Service) ... 413
9.6.9 Таблица управления доступом RMC ... 414

Часть 4. Расширенные задачи администрирования (с примерами) ... 417

Глава 10. Динамические LPAR (DLPAR) и виртуализация (VIO) ... 419

10.1 Реализация DLPAR в HACMP ... 420
10.1.1 Требования ... 420
10.1.2 Обеспечение приложений ресурсами ... 422

10.1.3 Конфигурирование DLPAR в HACMP ... *429*
10.1.4 Устранение ошибок верификации HMC ... *439*
10.1.5 Конфигурация тестового кластера ... *441*
10.1.6 Результаты тестирования ... *442*

10.2 HACMP и виртуализация ... 448
10.2.1 Требования ... *449*
10.2.2 Обеспечение приложений ресурсами ... *449*

10.3 Сценарии конфигурирования HACMP и виртуализации ... 453
10.3.1 Сценарий 1 ... *453*
10.3.2 Аспекты производительности и архитектуры (сценарий 2) ... *460*

Глава 11. Расширение возможностей группы ресурсов ... 463

11.1 Время установления ... 464

11.2 Политика распределения узлов ... 467
11.2.1 Конфигурирование политики распределения групп ресурсов на основе узлов ... *468*
11.2.2 Сценарий тестирования политики распределения на основе узлов ... *469*

11.3 Динамический приоритет узлов (DNP) ... 470
11.3.1 Конфигурирование политики динамического приоритета узлов ... *470*
11.3.2 Изменение существующей группы ресурсов для использования политики DNP ... *472*
11.3.3 Принцип работы динамического приоритета узлов ... *472*
11.3.4 Сценарий тестирования динамического приоритета узлов ... *474*

11.4 Расположение, отменяющее приоритет (POL) ... 480

11.5 Таймер отсроченного возврата после восстановления ... 484

11.6 Зависимости групп ресурсов ... 490
11.6.1 Зависимость дочернего объекта для группы ресурсов ... *491*
11.6.2 Зависимость расположения для группы ресурсов ... *492*
11.6.3 Ограничения на сочетание зависимостей ... *497*
11.6.4 Вывод зависимостей группы ресурсов ... *497*
11.6.5 Сценарий тестирования зависимостей групп ресурсов ... *498*

Глава 12. Настройка событий ... 503

12.1 Написание скриптов для настраиваемых событий ... 504

12.2 Команды пред- и постобработки событий HACMP ... 504
12.2.1 Установка скриптов пред- и постобработки ... *506*

12.3 Уведомление об ошибках ... 510
12.3.1 Автоматическое уведомление об ошибках ... *510*
12.3.2 Использование уведомления об ошибках ... *512*
12.3.3 Мониторинг общих дисков с использование функции уведомления об ошибках HACMP ... *516*

Глава 13. Аспекты подсистемы хранения ... 523

13.1 Типы групп томов ... 524
13.1.1 Группы томов с расширенным одновременным доступом ... *524*
13.1.2 Группы томов без одновременного доступа ... *525*
13.1.3 Группы томов с одновременным доступом ... *526*
13.1.4 Группы томов с одновременным доступом RAID ... *526*

13.2 Резервирование дисков ... 527

13.3 Принудительная активизация групп томов ... 528

13.4 Быстрый перехват дисков 529
13.4.1 Требования 529
13.4.2 Принцип работы быстрого перехвата дисков 529
13.4.3 Включение быстрого перехвата дисков 531
13.4.4 Преимущества 532
13.4.5 Известные проблемы 532
13.5 Мониторинг пульса через диски 533
13.5.1 Обзор 533
13.5.2 Требования 533
13.5.3 Аспекты производительности 534
13.5.4 Конфигурирование мониторинга пульса через диски 534
13.5.5 Тестирование связи в сети пульса через диски 536
13.5.6 Мониторинг 537

Глава 14. Сетевые функции 539

14.1 Etherchannel 540
14.1.1 Реализация EtherChannel в среде HACMP 541
14.2 Варианты размещения сервисных IP-синонимов 548
14.2.1 Конфигурирование политики размещения сервисных IP-адресов 549
14.3 Общее представление о файле netmon.cf 552
14.4 Общее представление о файле clhosts 554
14.5 Общее представление о файле clinfo.rc 555

Часть 5. Аварийное восстановление 557

Глава 15. Концепции и планирование HACMP Extended Distance 559

15.1 Компоненты HACMP/XD 560
15.1.1 HACMP/XD HAGEO 560
15.1.2 Функция интеграции HACMP/XD PPRC 563
15.2 Аспекты аварийного восстановления 564
15.3 Дополнительные сведения 566

Глава 16. HACMP и Cross-Site LVM 567

16.1 Введение в межсайтовое зеркальное отображение Cross-Site LVM mirroring 568
16.1.1 Требования 569
16.2 Аспекты инфраструктуры 569
16.3 Конфигурирование межсайтового зеркального отображения Cross-Site LVM mirroring 570
16.3.1 Конфигурирование кластера Cross-Site LVM 570
16.3.2 Конфигурирование сайтов кластера 571
16.3.3 Конфигурирование зависимостей сайтов с зеркальным отображением Cross-Site LVM 572
16.3.4 Настройка групп томов зеркального отображения cross-site LVM 574
16.3.5 Конфигурирование группы ресурсов, содержащей группу томов с поддержкой зеркального отображения Cross-Site LVM 575
16.4 Тестирование межсайтового зеркального отображения Cross-Site LVM 576
16.4.1 Протестированные сценарии 578

Глава 17. Сценарий аварийного восстановления в HAGEO 581

17.1 Описание сценария и планирование 582
17.1.1 Планирование конфигурации сети 582
17.1.2 Планирование конфигурации логических томов 583
17.1.3 Определение GMD 584
17.2 Установка и конфигурирование HAGEO 585

Глава 18. Понятия и конфигурирование GLVM 603

18.1 HACMP/XD GLVM 604
18.1.1 Определения и понятия 605
18.2 Миграция – переход от HAGeo к GLVM 615
18.2.1 Установка наборов файлов GLVM и конфигурирование GLVM 617
18.2.2 Аспекты производительности 620
18.2.3 Устранение неполадок 621
18.3 Этапы миграции с HAGEO на GLVM 623

Часть 6. Приложения 631

Приложение A. Таблицы планированияна бумаге 633

Two-Node Cluster Configuration Assistant 634
Таблицы планирования узлов 634

Список сокращений 641

Рекомендуемая литература 644

Книги серии IBM Redbooks 644
Прочие публикации 644
Сетевые ресурсы 644
Как приобрести книги серии IBM Redbooks 645
Поддержка от компании IBM 645

Примечания

Приведенная в данной книге информация была разработана для продуктов и служб, распространяемых на территории США.

Предложения IBM по услугам, товарам и их возможностям, описанным в руководстве, могут не действовать в других странах. За информацией о текущем ассортименте доступных продуктов и услуг обращайтесь в местные представительства IBM. Явные и неявные упоминания услуг, продуктов и их возможностей не означают необходимость их применения. Допускается их замена любыми функционально эквивалентными продуктами и службами сторонних производителей, не нарушающими прав на интеллектуальную собственность IBM, при этом ответственность за проверку совместимости и продуктивности решений сторонних производителей принимает на себя пользователь.

IBM может обладать патентами или патентными заявками на технологии, описанные в настоящем руководстве, предоставление которых не означает наличия лицензии на технологии. Письменные запросы лицензий следует направлять по адресу IBM Director of Licensing, IBM Corporation, North Castle Drive Armonk, NY 10504-1785 USA.

Приведенный ниже абзац не относится к Соединенному Королевству и иным странам, законодательству которых противоречит следующее условие: КОРПОРАЦИЯ IBM ПРЕДОСТАВЛЯЕТ ДАННОЕ РУКОВОДСТВО «КАК ЕСТЬ» И НЕ ДАЕТ НИКАКИХ ЯВНЫХ ИЛИ ПОДРАЗУМЕВАЕМЫХ ГАРАНТИЙ, ВКЛЮЧАЯ (НО НЕ ОГРАНИЧИВАЯСЬ) ПОДРАЗУМЕВАЕМЫЕ ГАРАНТИИ ПАТЕНТНОЙ ЧИСТОТЫ, КОММЕРЧЕСКОЙ ПРИГОДНОСТИ И СООТВЕТСТВИЯ КОНКРЕТНОМУ НАЗНАЧЕНИЮ. В отдельных государствах отказ от явных и подразумеваемых гарантий по ряду сделок запрещен, так что указанное ограничение может вас не коснуться.

В руководстве возможны опечатки и технические неточности. Приводимые в нем сведения регулярно обновляются, соответствующие изменения будут внесены в новую редакцию руководства. Корпорация IBM оставляет за собой право в любое время без уведомления модифицировать описанные в этом руководстве продукты и программные средства.

Любые ссылки на сайты сторонних компаний в этом руководстве носят исключительно информационный характер и не свидетельствуют об их поддержке корпорацией IBM; риск, связанный с применением ресурсов этих сайтов, принимает на себя пользователь.

По своему усмотрению и без каких-либо обязательств IBM может использовать и распространять любые предоставленные сведения.

За информацией о сторонних продуктах обращайтесь к производителям, их публикациям и другим открытым источникам. IBM не тестировала эти продукты и не может подтвердить точность оценок производительности, совместимости и прочих параметров. Соответствующие вопросы следует направлять поставщикам этих продуктов.

Настоящее руководство содержит в качестве примеров данные и отчеты, используемые в повседневной практике предприятий. Для наиболее полной иллюстрации примеров в руководстве встречаются имена лиц, названия компаний, торговых марок, товаров. Все они вымышлены, и любые совпадения с именами и данными реально существующих компаний случайны.

Авторское право

Данное руководство содержит примеры исходного кода прикладных программ, которые иллюстрируют приемы программирования на различных платформах. Вы можете копировать, менять и распространять их в любом виде без отчислений в пользу IBM, руководствуясь целями разработки прикладных программ, включая коммерческие, в соответствии с интерфейсом прикладного программирования платформы, для которой предназначены эти программы. Примеры не подвергались всеобъемлющему тестированию, поэтому IBM не может явно или неявно гарантировать надежность, удобство в обслуживании и работоспособность приведенных примеров. Вам разрешено копировать, изменять и распространять программы-примеры, содержащиеся в данном руководстве, в любом виде без отчислений в пользу IBM с целью разработки, применения, в том числе коммерческого, и дистрибуции прикладных программ, сообразуясь с интерфейсами прикладного программирования корпорации IBM.

Торговые марки

@server®
Redbooks (logo) ™
ibm.com®
pSeries®
AIX 5L™
AIX®
Cross-Site®
DB2 Universal Database™
DB2®
Enterprise Storage Server®
FlashCopy®
HACMP™
IBM®
Magstar®
MVS™
NetView®
POWER4™
POWER5™
Redbooks™
Requisite®
RS/6000®
Seascape®
Tivoli®
TotalStorage®
WebSphere®

Ниже приведены торговые марки корпорации IBM в США и (или) других странах:

Другим компаниям принадлежат следующие торговые марки:

IPC, Java, Solaris, Sun, Ultra и все торговые марки, связанные с Java, являются торговыми марками Sun Microsystems, Inc. в США и (или) других странах.

Microsoft, Windows, логотип Windows – торговые марки Microsoft Corporation в США и (или) других странах.

Intel, логотип Intel Inside, MMX и Pentium являются торговыми марками или зарегистрированными торговыми марками Intel Corporation в США и (или) других странах.

UNIX – зарегистрированная торговая марка Open Group в США и (или) других странах.

Linux – торговая марка, принадлежащая Линусу Торвальдсу (Linus Torvalds) в США и (или) других странах.

Названия других компаний, товаров и услуг могут быть торговыми марками соответствующих владельцев.

Предисловие

Эта книга из серии IBM® Redbooks поможет вам в установке, конфигурировании и настройке новой системы – HACMP™ V5.3, а также описывает новые и усовершенствованные возможности, в частности интеграцию с динамическими разделами (DLPAR), виртуальный ввод-вывод (Virtual I/O) и конфигурации аварийного восстановления (Disaster Recovery).

Эта книга содержит всестороннее описание архитектуры HACMP и HACMP Extended Distance (HACMP/XD). Если вы планируете устанавливать, осуществлять миграцию или просто администрировать кластер высокой доступности, эта книга предназначена для вас. Кроме того, подробно описаны элементы аварийного восстановления и реализация этих требований с использованием HACMP.

Эта книга поможет специалистам по AIX®, нуждающимся во всестороннем и проблемно-ориентированном руководстве для получения знаний и навыков, необходимых для проектирования и внедрения кластеров HACMP, а также для повседневного системного администрирования. Содержание книги представляет сочетание теории и практического опыта.

Эта книга будет особенно полезна для системных администраторов, уже использующих и кластеры HACMP/ES, и кластеры HACMP Extended Distance (XD), которым может потребоваться консолидировать свою среду и перейти на HACMP V5.3. Книга содержит подробное описание поузловой миграции на HACMP/ES 5.3 и всестороннее обсуждение подготовки к обновлению или миграции.

О людях, написавших эту книгу

Эта книга была создана командой специалистов со всего мира, работающих в International Technical Support Organization, Остин (Austin).

Октавиан Ласку (Octavian Lascu) – руководитель проектов в International Technical Support Organization, Поукипси (Poughkeepsie), работающий в IBM с 1992 г. Он много пишет и проводит по всему миру курсы IBM по всем сферам, связанным с кластерами на pSeries® и Linux®. До перехода в ITSO Октавиан работал в IBM Global Services Romania в должности сервисного менеджера по программному и аппаратному обеспечению. Он имеет степень магистра Политехнического института в Бухаресте по электронике, а также имеет сертификацию IBM Certified Advanced Technical Expert по AIX/PSSP/HACMP.

Шон Бодили (Shawn Bodily) – старший специалист по информационным технологиям компании ATS Americas в Далласе, Техас. Он 7 лет работал в IBM и имеет 10-летний опыт работы с AIX и 7 лет специализации в HACMP. Он имеет сертификации по

HACMP и ATE[1] версий V4 и V5. Он разрабатывает и проводит презентации на тему высокой доступности и систем хранения. Также он является соавтором двух книг серии IBM Redbooks™.

Мария-Катарина Эссер (Maria-Katharina Esser) – специалист в области информационных технологий по препродажной технической поддержке; работает в IBM System and Technology Group (STG) в Мюнхене, Германия. Она работает в IBM 17 лет и имеет 6 лет опыта в AIX, RS/6000® и IBM @server® pSeries. Ее область знаний включает системы хранения и HACMP.

Майкл Херрера (Michael Herrera) – инженер по программному обеспечению в IBM AIX Software Support Center в Далласе, Техас. Он имеет 7 лет опыта в поддержке AIX, RS/6000, SAN и pSeries. Он имеет степень бакалавра университета Коннектикута по управляющим информационным системам, а также степень магистра по бизнес-администрированию (MBA) университета Далласа. Он специализируется в средах AIX/HACMP и SAN и имеет сертификацию IBM Advanced Technical Expert.

Патрик Потье (Patrick Pothier) в течении 5 лет работает специалистом по информационным технологиям в компании STG France, специализируясь в FTSS pSeries. Он имеет 13 лет опыта в сфере UNIX® и 11 лет опыта в HACMP. Его область знаний включает операционные системы (MVS™, AIX, Solaris™, Linux), высокую доступность (HACMP), решения резервного копирования (ADSM/6000, Tivoli® Storage Manager и Netbackup), а также администрирование ERP систем (SAP R/3 BC).

Дюсан Прелец (Dusan Prelec) – специалист по поддержке pSeries в IBM Global Services, Словения. Он 3 года работал в службе поддержки AIX и HACMP в Центральной и Восточной Европе. Он участвует в различных проектах внедрения кластеров и систем хранения в Словении и странах Центральной и Восточной Европы. Он является сертифицированным специалистом IBM по AIX и HACMP и преподает курсы по HACMP. Прежде чем перейти в команду IBM Global Services, он работал в отделе препродажной технической поддержки IBM CEMA pSeries.

Дино Квинтеро (Dino Quintero) – старший сертифицированный специалист-консультант в информационных технологиях в ITSO в Поукипси, Нью-Йорк. До перехода в ITSO он работал в должности аналитика производительности в Enterprise Systems Group, а также в качестве разработчика систем аварийного восстановления для IBM Global Services. Его области знаний включают аварийное восстановление и решения по кластеризации pSeries. Он имеет сертификацию по системному администрированию и технологиям кластеризации pSeries. Кроме того, он является старшим сертифицированным специалистом IBM в сфере технологий pSeries. В настоящее время он руководит техническими командами, доставляя решения IBM Redbook по технологиям кластеризации pSeries, а также проводя технические семинары по всему миру.

Кен Рэймонд (Ken Raymond) – старший специалист по информационным технологиям, работающий в IBM Global Services в Оттаве, Канада. Он имеет более 26 лет опыта работы в IBM, в прошлом 10 лет интенсивно работая с заказчиками AIX в области управления системами и высокой доступности. Его области знаний

---

[1] Скорее всего, имеется в виду сертификация IBM Certified Advanced Technical Expert по AIX. *Примеч. науч. ред.*

включают управление системами AIX и их техническую поддержку, проектирование решений pSeries (логические разделы и виртуализация), внедрение AIX и поддержку миграции, внедрение и поддержку кластеризации (HACMP, CSM, PSSP), а также службы управления проектами.

Виктор Себестени (Viktor Sebesteny) – специалист по информационным технологиям в Венгрии. Он работал в IBM Global Services в течение 9 лет. Он имеет степень бакалавра университета KKMF в области вычислительной техники. Его области знаний включают pSeries, AIX, HACMP и CSM.

Андрей Соцолиус (Andrei Socoliuc) – инженер по поддержке программного обеспечения в IBM Global Services в Румынии. Он имеет степень магистра в области вычислительной техники от Политехнического института в Бухаресте, Румыния. Он имеет 6 лет опыта в AIX, pSeries и кластеризации. Его области знаний включают AIX, PSSP, HACMP, TSM и Linux. Он написал множество публикаций о кластерах pSeries и PSSP.

Энтони (Ред) Стил [Antony (Red) Steel] – старший специалист по информационным технологиям в ITS Australia. Он имеет 12 лет опыта в области UNIX, преимущественно в AIX и Linux. Он имеет ученую степень университета Сиднея в области теоретической химии. Его области знаний включают разработку сценариев, настройку систем, производительность, сетевые технологии, высокую доступность и решение проблем. Он разрабатывает и проводит презентации на тему LVM, TCP/IP и высокой доступности как в Австралии, так и в Азиатско-Тихоокеанском регионе. Также он является соавтором трех книг серии IBM Redbooks.

Выражаем благодарность следующим людям за их вклад в этот проект:

Майкл Коффи (Michael K. Coffey)
IBM, Поукипси (Poughkeepsie)

Пол Мойер (Paul Moyer)
IBM, Поукипси (Poughkeepsie)

Томас Уивер (Tomas Weaver)
IBM, Остин (Austin)

Дэвид Труонг (David Truong)
IBM, Даллас (Dallas)

Скип Расселл (Skip Russell)
IBM, Поукипси (Poughkeepsie)

Стив Товсимак (Steve Tovcimak)
IBM, Поукипси (Poughkeepsie)

Элейн Крэкоуер (Elaine Krakower)
IBM, Поукипси (Poughkeepsie)

Джеффри Мэттс (Geoffrey Mattes)
IBM, Австралия

Габриэль Велез (Gabrielle Velez)
International Technical Support Organization

Приглашаем авторов к сотрудничеству

Приглашаем вас к участию в программе размещения (residency program) продолжительностью от двух до шести недель! Вы поможете нам написать книгу, посвященную определенным продуктам или системам, и при этом получите практический опыт работы с новейшими технологиями. У вас будет возможность работать в одной команде с техническими специалистами компании IBM, нашими бизнес-партнерами и/или клиентами.

Ваши усилия помогут повысить принятие продукта и улучшить степень удовлетворенности пользователей. Кроме того, вы сможете развить связи с сотрудниками лабораторий для разработчиков (development labs) компании IBM и повысить свою продуктивность и профессионализм.

Узнайте больше о программе размещения (residency program), просмотрите список вакантных программ размещения (residency index) и подайте заявку на сайте

ibm.com/redbooks/residencies.html

Мы будем рады вашим отзывам. Ваши отзывы очень важны для нас!

Мы стремимся сделать наши книги максимально полезными. Направляйте нам ваши отзывы об этой книге или других книгах из серии IBM Redbooks, используя один из следующих способов:

– используйте веб-форму **Contact us** по адресу

  ibm.com/redbooks

– направляйте ваши отзывы по электронной почте

  redbook@us.ibm.com

– отправляйте ваши отзывы по адресу:

  IBM Corporation, International Technical Support Organization
  Dept. JN9B Building 905
  11501 Burnet Road
  Austin, Texas 78758-3493

Часть 1

Введение

Часть 1 содержит обзор HACMP и описывает компоненты HACMP в контексте эффективного внедрения. Так как HACMP является развитым продуктом, мы считаем важным рассказать о недавней истории HACMP, что поможет в планировании будущих действий, таких, как миграция существующих конфигураций на последнюю версию и использование новых возможностей HACMP V5.2 и V5.3.

Также мы представим основные понятия управления HACMP с рекомендациями, позволяющими облегчить работу системного администратора.

1

Введение в HACMP

Эта глава представляет собой введение в концепцию высокой доступности[1] и, в частности, содержит описание IBM HACMP:

- что такое HACMP (High Availability Cluster Multi-Processing);
- решения по обеспечению доступности;
- история и эволюция;
- терминология и понятия высокой доступности;
- высокая доступность и отказоустойчивость (fault tolerance);
- планирование установки программного обеспечения;
- аспекты установки HACMP.

[1] Термин «high availability» переводят как «высокая надежность» или «высокая доступность». В данной публикации мы используем, как более точный на наш взгляд, перевод «высокая доступность», так как основной целью HA-решения является именно обеспечение доступности приложений (сервисов) конечным пользователям. *Примеч. науч. ред.*

1.1 Что такое HACMP

HACMP является сокращением от «High Availability Cluster Multi-Processing». Главными составляющими здесь являются высокая доступность и мультиобработка в кластерной (многоузловой) среде.

1.1.1 Высокая доступность

В современных сложных средах обеспечение непрерывности работы приложений является основным компонентом эффективной реализации IT-системы. Высокая доступность – одна из составляющих, способствующих обеспечению непрерывности обслуживания клиентов приложения путем маскирования (скрытия) или устранения как запланированных, так и незапланированных простоев (остановок) в работе систем и приложений. Это достигается путем устранения единых точек отказа (Single Point Of Failure, SPOF[1]) в аппаратном и программном обеспечении. Решение высокой доступности гарантирует, что отказ любого компонента решения (аппаратного обеспечения, программного обеспечения или управления системой) не вызовет длительной недоступности приложения и его данных для конечного пользователя.

Решения высокой доступности должны устранять единые точки отказа посредством соответствующего проектирования, планирования, выбора оборудования, конфигурирования программного обеспечения, управления приложениями, тщательного управления средой и дисциплин управления изменениями.

Вкратце высокую доступность можно определить как процесс обеспечения доступности приложения для использования посредством дублирования и/или реализации общего доступа к аппаратным ресурсам, управляемым специализированным программным компонентом.

1.1.2 Кластерная мультиобработка

Помимо высокой доступности, HACMP также содержит компонент, обеспечивающий мультиобработку. В основе функции мультиобработки лежит тот факт, что кластер содержит множество аппаратных и программных ресурсов, управляемых HACMP и обеспечивающих сложную функциональность приложений и более эффективное использование ресурсов.

Краткое определение кластерной мультиобработки может иметь такой вид: множество приложений, выполняющихся на нескольких узлах с общим или одновременным доступом к данным.

Хотя компонент кластерной мультиобработки и является желательным, эффективное использование им всех доступных в многоузловой (кластерной) среде ресурсов зависит от возможностей приложения и реализации системы. Их реализация должна начинаться на этапе планирования и проектирования кластеров.

---

[1] Здесь и далее мы будем упоминать в скобках оригинальные названия терминов (на английском языке), чтобы читателю было проще соотнести термины, используемые в этом руководстве, с теми терминами, которые он встретит в англоязычной документации и при непосредственной работе с HACMP. *Примеч. науч. ред.*

HACMP – всего лишь одна из технологий высокой доступности, построенная на все более надежных операционных системах, более надежном оборудовании с возможностью выполнения «горячей замены», все более устойчивых приложениях и предлагающая функции мониторинга и автоматического реагирования.

Решение высокой доступности на основе HACMP обеспечивает автоматическое обнаружение сбоев, диагностику, восстановление приложений и реинтеграцию узлов. При использовании соответствующего приложения HACMP также может обеспечивать одновременный доступ к данным для приложений параллельной обработки, предлагая таким образом отличную горизонтальную и вертикальную масштабируемость (с добавлением возможностей управления динамическим LPAR).

IBM также разработала расширенную версию HACMP, содержащую функции аварийного восстановления, интегрированные в решение под названием HACMP Extended Distance (HACMP/XD), которое поддерживает функциональность HACMP между двумя географически распределенными сайтами. HACMP/XD поддерживает множество различных методов репликации данных и подробно обсуждается в гл. 15, «Понятия и планирование HACMP Extended Distance».

1.2 Решения по обеспечению доступности: обзор

Существует множество решений, обеспечивающих широкий диапазон опций доступности. В табл. 1.1 описаны различные типы решений обеспечения доступности и их характеристики.

Таблица 1.1. Типы решений обеспечения доступности

Решение	Время простоя (отключения)	Доступность данных	Стоимость
Автономное (Standalone)	Дни	С последней резервной копии	Базовые затраты на аппаратное и программное обеспечение ($)
Улучшенное автономное (Enhanced standalone)	Часы	До последней транзакции	Двойные затраты на аппаратное обеспечение ($$)
Кластеры высокой доступности (High availability clusters)	Минуты	До последней транзакции	Двойные затраты на аппаратное обеспечение и дополнительное обслуживание ($$+)
Отказоустойчивые вычислительные системы (fault-tolerant computing)	Никогда не отключаются	Без потери данных	Специализированное аппаратное и программное обеспечение, очень дорогостоящее ($$$$$$)
HACMP/XD	Минуты	До последней транзакции	Двойная или тройная стоимость аппаратного обеспечения + дополнительные затраты на связь ($$$$)

Решения по обеспечению высокой доступности в целом предлагают следующие преимущества:

- стандартные аппаратные и сетевые компоненты (могут использоваться с существующим оборудованием);
- работают практически со всеми приложениями (зависит только от умения того, кто осуществляет внедрение);

- работают с широким диапазоном типов дисков и сетей;
- отличная доступность при приемлемой стоимости.

Решение высокой доступности от компании IBM для IBM eserver® pSeries предлагает различные преимущества, включая следующие:

- решение является проверенным (результат более чем 15 лет разработки);
- гибкость (практически любое приложение, выполняющееся на автономной системе AIX, может быть защищено с использованием HACMP);
- использование доступных аппаратных компонентов (pSeries);
- проверенное качество поддержки наших клиентов.

При планировании внедрения решения HACMP следует учитывать следующие аспекты:

- тщательное проектирование и подробное планирование;
- устранение единых точек отказа;
- выбор соответствующего оборудования;
- надлежащее внедрение (без использования упрощенных подходов);
- упорядоченные методы системного администрирования и управление изменениями;

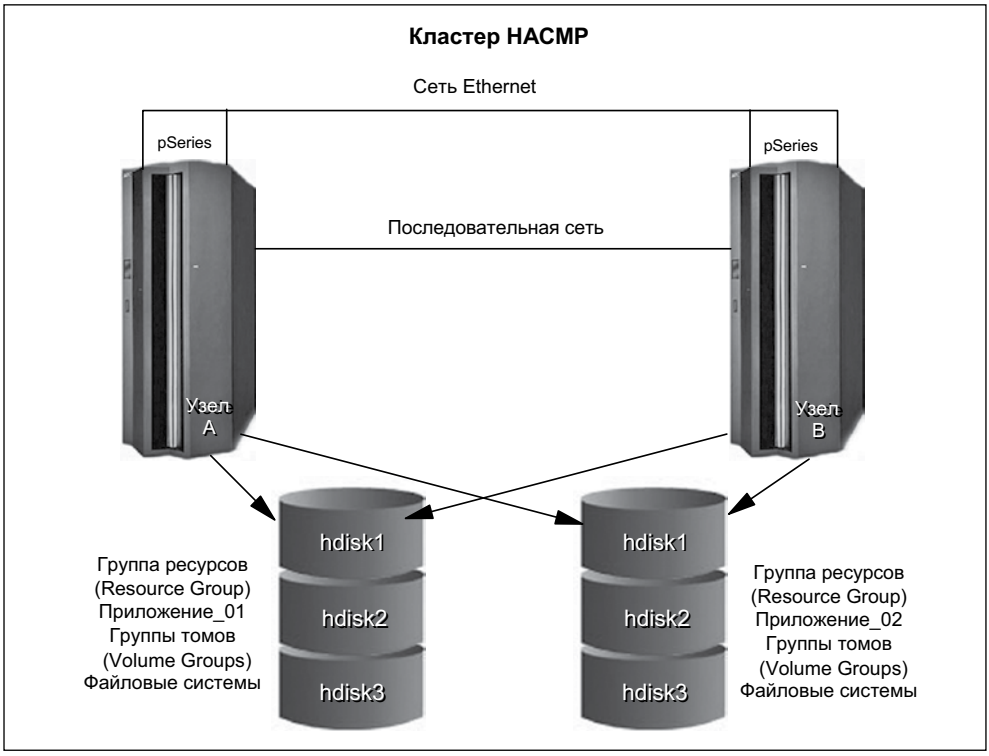

Рис. 1.1. Кластер HACMP

- документированные процедуры эксплуатации;
- всесторонний тестовый план и тщательное тестирование.

Пример среды HACMP представлен на рис. 1.1.

1.2.1 Время отключения (простоя)

Время отключения (простоя) означает период, когда приложение недоступно для обслуживания своих клиентов. Можно разделить отключения на две категории:

- запланированные:
  - замена оборудования;
  - обслуживание;
  - обновление программного обеспечения;
  - резервное копирование (автономное резервное копирование);
  - тестирование (периодическое тестирование необходимо для проверки кластеров);
  - разработка;
- незапланированные:
  - ошибки администратора;
  - отказы приложений;
  - отказы оборудования;
  - ошибки операционной системы;
  - стихийные бедствия.

Таким образом, роль HACMP состоит в том, чтобы обеспечивать доступность приложения как при незапланированных отказах, так и при выполнении обычных повседневных административных задач. HACMP обеспечивает мониторинг и автоматическое восстановление ресурсов, от которых зависит приложение.

1.2.2 Единая точка отказа

Единой точкой отказа (single point of failure, SPOF) является любой отдельно взятый компонент в составе кластера, который в случае сбоя делает приложение недоступным для конечных пользователей.

Хорошее проектирование позволяет устранить единые точки отказа (узлы, устройства хранения, сети) в кластере. HACMP осуществляет управление этими компонентами, а также управление ресурсами, необходимыми для приложения (включая скрипты запуска-перезапуска приложений).

В конечном счете цель любого информационного решения в критической среде состоит в том, чтобы обеспечить непрерывную доступность приложения и защиту данных. Высокая доступность – только один из элементов в обеспечении непрерывной работы. В основе высокой доступности лежит доступность оборудования, программного обеспечения (операционной системы и ее компонентов), приложения и сетевых компонентов.

Во избежание возникновения единых точек отказа необходимы:

- избыточные (резервные) серверы;
- избыточные сетевые пути;
- избыточные пути хранения (данных);
- избыточное хранение (зеркальное отображение/RAID);
- мониторинг;
- обнаружение и диагностика отказов;
- автоматическое перемещение (failover) приложения при сбое;
- автоматическая реинтеграция ресурсов.

Как говорилось выше, хорошее проектирование позволяет устранить единые точки отказа, и HACMP осуществляет управление доступностью приложения во время отключений. В табл. 1.2 перечислены все объекты кластеров, сбой которых может вызвать недоступность приложения. Каждый объект кластера может представлять собой физический или логический компонент.

Таблица 1.2. Единая точка отказа

Объекты кластера	Способ устранения единой точки отказа
Узел (серверы)	Использование нескольких узлов
Электропитание	Использование нескольких цепей или источников питания и/или ИБП
Сетевой адаптер	Избыточность сетевых адаптеров
Сеть	Несколько сетей, подключенных ко всем узлам, избыточные сетевые пути с независимым оборудованием между каждым узлом и клиентами
Подсистема TCP/IP	Использование сетей «точка-точка» для подключения каждого узла к соседнему узлу в кольце
Адаптер ввода-вывода	Избыточность адаптеров ввода-вывода
Контроллеры	Избыточность контроллеров
Хранение	Избыточность оборудования, дисковые стойки, зеркальное отображение/технология RAID, избыточность путей данных
Приложение	Конфигурирование мониторинга приложения и резервных узлов для «подхвата» приложения и его данных
Сайты	Использование более одного сайта для аварийного восстановления
Группы ресурсов	Использование групп ресурсов для управления всеми ресурсами, требуемыми приложению

HACMP также обеспечивает оптимизацию доступности, допуская динамическое реконфигурирование работающих кластеров. Задачи обслуживания, такие, как добавление или удаление узлов, можно выполнять без остановки и перезапуска кластера.

Кроме того, на работающем кластере можно выполнять другие задачи управления, такие, как изменение конфигурации системы хранения, управление пользователями с применением единой точки управления кластером (Cluster Single Point of Control, C-SPOC), не прерывая доступ пользователя к приложению, выполняющемуся на узлах кластера (cluster nodes). C-SPOC также обеспечивает репликацию изменений, выполненных на одном узле, по всему кластеру последовательным образом.

1.3 История и эволюция

История IBM High Availability Cluster Multi-Processing началась в начале 1990-х гг. Разработка HACMP была начата в 1990 г. с целью обеспечить решение высокой доступности для приложений, выполняющихся на серверах RS/6000. Мы не представляем информацию о самых ранних версиях, так как эти версии уже не поддерживались и не использовались на момент написания этой книги; вместо этого мы сосредоточились на наиболее поздних версиях.

Изначально HACMP разрабатывался как автономный продукт (называемый сейчас HACMP Classic), однако, как только появилась инфраструктура высокой доступности Reliable Scalable Clustering Technology (RSCT) от компании IBM, она была применена в HACMP, в результате чего была разработана технология HACMP Enhanced Scalability (HACMP/ES), обеспечивающая более высокую производительность и функциональность по сравнению с «классической» версией.

1.3.1 HACMP Version 4

HACMP V4.2.2 Наряду с HACMP Classic (HAS) эта версия включала версию с улучшенной масштабируемостью (enhanced scalability, ES), основанную на топологии RSCT (Reliable Scalable Clustering Technology), группах и службах управления событиями, унаследованных от PSSP (Parallel Systems Support Program).

HACMP V4.3.X Помимо прочего, эта версия включала поддержку 32 узлов для HACMP/ES, усовершенствования C-SPOC, поддержку сетей ATM, HACMP Task Guides (графический интерфейс для упрощения конфигурирования кластеров), множество скриптов пред- и постобработки событий (pre-event, post-event scripts), функцию переноса (takeover) MAC-адреса FDDI, усовершенствования мониторинга и поддержку администрирования, поузловую миграцию и поддержку AIX Fast Connect.

HACMP V4.4.X К новым элементам в этой версии относятся интеграция с Tivoli, мониторинг приложений, каскадирование без возврата после восстановления (cascading without fallback), усовершенствования C-SPOC, улучшенная поддержка миграции, интеграция функциональных возможностей HA-NFS и электронная документация (HTML и PDF).

HACMP V4.5 В этой версии обязательным является использование AIX 5L; версия содержит функцию автоматического определения конфигурации, допускает применение нескольких сервисных меток (service labels) для каждого сетевого адаптера [посредством использования IP-синонимов (alias)], обеспечивает поддержку постоянных (persistent) IP-адресов, содержит 64-разрядный API и функции мониторинга и восстановления после потери кворума группы томов.

1.3.2 HACMP Version 5 Release 1

В этой версии появилось множество крупных изменений, от упрощения конфигурирования и повышения производительности до изменения терминологии HACMP:

HACMP Classic (HAS) не был включен; пакет содержал только HACMP/ES, основанный на технологии IBM Reliable Scalable Cluster Technology.

Ниже приведены некоторые наиболее важные новые возможности HACMP V5.1.

- Пути конфигурации (процедуры) SMIT «Standard» («Базовая») и «Extended» («Расширенная»).

- Автоматическое определение конфигурации.
- Настраиваемые группы ресурсов (Custom Resourse Groups).
- Сети, отличные от IP, основанные на передаче пакетов мониторинга (пульса, heartbeating) через диски.
- Быстрый перехват диска (Fast Disk Takeover).
- Принудительная активизация (varyon) групп томов.
- Передача пакетов мониторинга (пульса) через IP-синонимы (Heartbeating over IP aliases).
- Повышенная безопасность посредством использования демона коммуникаций (cluster communication daemon) в кластере (что устраняет необходимость использования r-команд AIX, устраняя и необходимость в файле /.rhosts).
- Повышенная производительность процесса конфигурирования и синхронизации кластера.
- Упорядочение терминологии HACMP (ее согласование с другими продуктами HA).
- Упрощение конфигурирования и обслуживания.
- Усовершенствования системы автоматизированного планирования (Online Planning Worksheets).
- Настраиваемые группы ресурсов.
- Мониторинг пульса сервисных IP-адресов/меток на резервных узлах (узлах подхвата).
- Мониторинг пульса через IP-синонимы.
- Мониторинг пульса через диски.
- Различные усовершенствования C-SPOC.
- Интеграция с GPFS.
- Усовершенствования в проверке кластеров.
- Улучшенное управление группами ресурсов.

1.3.3 HACMP Version 5 Release 2

Выпущенная в июле 2004 г. версия HACMP V5.2 содержала дополнительные усовершенствования в управлении, упрощении конфигурирования, автоматизации и производительности: Two-Node Configuration Assistant (упрощенное конфигурирование кластера из двух узлов), содержащий как меню SMIT, так и интерфейс Java™ (в дополнение к путям конфигурации SMIT «Standard» и «Extended»). Ниже приведен обзор усовершенствований, реализованных в HACMP V5.2:

- Наборы файлов (File Collections).
- Управление паролями пользователей.
- Отказ от использования классических групп ресурсов; заменены настраиваемыми группами ресурсов.
- Автоматические процедуры тестирования.

- Автоматическая проверка кластеров.
- Усовершенствованная система автоматизированного планирования (OLPW) теперь может импортировать конфигурацию из существующего HACMP-кластера.
- Управление событиями (event management, EM) было заменено подсистемой мониторинга и управления ресурсами (resource monitoring and control, RMC; стандартна в AIX).
- Повышенная безопасность.
- Зависимости групп ресурсов.
- Возможности самовосстановления кластеров (исправление некоторых ошибок конфигурации кластера).
- HACMP Smart Assist для WebSphere® Application Server.

1.3.4 HACMP Version 5 Release 3

Начиная с июля 2005 г. новая версия, HACMP V5.3, продолжает развитие HACMP, добавляя дальнейшие усовершенствования в областях управления, упрощения конфигурирования, автоматизации и производительности. Ниже приведен обзор усовершенствований в HACMP V5.3.

- Проверка (верификация) кластера при его запуске.
- При проверке выполняются дополнительные корректирующие действия.
- clverify предупреждает о распознаваемых единых точках отказа.
- clverify интегрирует опции HACMP/XD – PPRC; GeoRM; GLVM.
- clverify автоматически заполняет файл clhosts.
- Использование формата XML для файлов OLPW, а также возможность преобразования существующих файлов снимков конфигурации (snapshots) в XML-файлы конфигурации кластера.
- Поддержка томов и файловых систем сторонних производителей (OEM).
- Дальнейшая интеграция HACMP с RSCT.
- Большее количество опций «Smart Assist» – DB2® и Oracle Application Server.
- Удаление некоторых ограничений, относящихся к сайтам, из HACMP.
- Добавление зависимости расположения для групп ресурсов.
- Усовершенствование безопасности WebSMIT:
  - проверка клиентских данных перед выполнением команд HACMP;
  - проверка параметров на стороне сервера;
  - интеграция инструментов аутентификации WebSMIT с механизмами аутентификации AIX.
- Непрерывное выполнение диспетчера кластера Cluster Manager (**clstrmgrES**), вне зависимости от состояния кластера (работает/не работает), для поддержки дальнейшей автоматизации конфигурирования и расширенного администрирования кластера.

- Демон расширений multi-peer (cluster multi-peer extension daemon, **clsmuxpdES**) и демон информации кластера (cluster information daemon, **clinfoES**) больше не используют общие сегменты памяти.

> **Примечание.** На момент публикации существовали версии HACMP 5.1, 5.2 и 5.3, однако в этой книге для тестирования использовалась HACMP V5.3.

1.4 Терминология и понятия высокой доступности

Чтобы понимать функциональные возможности HACMP и эффективно их использовать, необходимо знать некоторые важные термины и понятия.

1.4.1 Терминология

Начиная с HACMP V5.1 терминология, используемая для описания конфигурации и работы HACMP, значительно изменилась. Цель такого изменения заключается в упрощении общего использования и обслуживания HACMP, а также в том, чтобы согласовать терминологию с прочими продуктами IBM.

Например, в предыдущей версии HACMP термин «адаптер» в зависимости от контекста мог иметь различные значения, что затрудняло конфигурирование.

В этой книге используются следующие термины:

Кластер (Cluster)	Слабосвязанный набор независимых систем (узлов) или логических разделов (Logical Partitions, LPAR), объединенный в сеть с целью совместного доступа к ресурсам и обмена данными друг с другом. HACMP определяет связи между совместно работающими системами, где соседние одноранговые узлы кластера обеспечивают работу служб, предоставляемых узлом кластера в том случае, если он не способен это делать. Эти отдельные узлы вместе ответственны за обеспечение функционирования одного или нескольких приложений в случае отказа любого из компонентов кластера.
Узел (Node)	Компьютер IBM eserver pSeries (или LPAR), на котором выполняются AIX и HACMP и который определен как часть кластера. Каждый узел имеет набор ресурсов (диски, файловые системы, IP-адреса и приложения), которые могут быть переданы другому узлу в кластере в случае сбоя узла или компонента.
Клиенты	Клиентом является система, которая может осуществлять доступ к приложению, выполняющемуся на узлах кластера через локальную сеть. На клиентах выполняется клиентское приложение, осуществляющее подключение к серверу (узлу), на котором выполняется приложение.

1.4.2 Понятия

Основные понятия HACMP можно систематизировать следующим образом:

Топология	Включает основные компоненты кластера: узлы, сети, коммуникационные интерфейсы, коммуникационные устройства и коммуникационные адаптеры.
Ресурсы	Логические компоненты или объекты, к которым обеспечивается высокий уровень доступа [например, файловые системы, устройства прямого доступа (raw devices), сервисные IP-метки и приложения] путем перемещения с одного узла на другой. Все ресурсы, которые совместно составляют приложение или службу с высокой доступностью, группируются в группы ресурсов (resource groups, RG).

	HACMP обеспечивает высокую доступность группы ресурсов как единого объекта, который может быть перемещен с одного узла на другой в случае отказа компонента или узла. Группы ресурсов могут быть доступны с одного узла либо, в случае одновременно выполняющихся приложений, могут быть одновременно доступны с нескольких узлов. Кластер может содержать более одной группы ресурсов, что позволяет эффективно использовать узлы кластера (отсюда и понятие «мультиобработка» в HACMP).
Сервисная IP-метка (Service IP label)	Метка, соответствующая сервисному IP-адресу и используемая для связи между клиентами и узлом. Сервисная IP-метка является частью группы ресурсов, что означает, что HACMP осуществляет ее мониторинг и обеспечивает ее высокую доступность.
Перехват[1] IP-адреса (IP address takeover)	Процесс, при котором IP-адрес перемещается с одного адаптера на другой адаптер в той же логической сети. Этот адаптер может быть расположен либо на том же узле, либо на другом узле в кластере. При использовании IP-синонимов в качестве метода назначения адресов адаптерам одному адаптеру может соответствовать более одного адреса.
Перехват ресурса (Resourse takeover)	Представляет собой операцию передачи ресурсов между узлами внутри кластера. При отказе одного компонента или узла в связи с проблемами оборудования или операционной системы его группы ресурсов передаются на другой узел.
Перемещение при сбое (Fallover)	Представляет собой перемещение группы ресурсов с одного активного узла на другой узел (резервный узел) при сбое на этом активном узле.
Возврат после восстановления (Fallback)	Представляет собой перемещение группы ресурсов обратно с резервного узла на прежний узел, когда он станет доступным. Это перемещение обычно выполняется после реинтеграции ранее отказавшего узла.
Пакет «пульсации» (Heartbeat packet)	Пакет, передаваемый между коммуникационными интерфейсами в кластере, используемый различными демонами кластера для мониторинга состояния компонентов кластера – узлов, сетей, адаптеров.
Демоны RSCT	Включают два процесса (службы топологии и службы групп), осуществляющие мониторинг состояния кластера и каждого узла. Диспетчер кластера получает информацию о событиях, генерируемую этими демонами, и предпринимает соответствующие (ответные) действия в случае сбоя.
Лидер группы (Group Leader)	Узел с наивысшим IP-адресом, определенным в одной из сетей HACMP (первой доступной сети), выступающий в качестве центрального репозитория (хранилища) для всех данных топологии и групп, поступающих от демонов RSCT, относящихся к состоянию кластера.
Резервный лидер группы (Group Leader Backup)	Узел со вторым по величине IP-адресом в той же произвольно выбранной сети, выступающий в качестве резервного узла для лидера группы и принимающий на себя роль лидера группы в случае выхода его из кластера.
Распорядитель (Mayor)	Узел, выбранный лидером группы RSCT (узел с третьим по величине IP-адресом, если таковой существует; в противном случае им является резервный лидер группы). В его обязанности входит информирование других узлов о любых изменениях в кластере, определенных лидером группы.

[1] Мы используем термины «перехват» или «подхват», так как takeover происходит тогда, когда компонент (элемент топологии) кластера не может больше по какой-либо причине «держать» ресурс, за который он отвечает, и резервный компонент «подхватывает» его нагрузку. *Примеч. науч. ред.*.

Примечание. В более ранних версиях HACMP перемещение группы ресурсов при сбое компонента на одном узле называлось «failover»[4.3].

Более подробно эти понятия описаны в гл. 2, «Составляющие высокой доступности».

1.5 Высокая доступность и отказоустойчивость

В зависимости от времени реагирования и ответного действия на обнаруженные отказы системы кластеры и системы можно разделить:

- на отказоустойчивые системы;
- на системы высокой доступности.

1.5.1 Отказоустойчивые системы

Отказоустойчивые системы предназначены для обеспечения бесперебойной работы вне зависимости от типов отказов, которые могут возникнуть (исключение, пожалуй, составляет только полный отказ сайта вследствие стихийного бедствия). В таких системах ВСЕ компоненты как программного, так и аппаратного обеспечения по меньшей мере дублируются.

Все эти компоненты, процессоры, память и диски имеют специальный дизайн, обеспечивающий непрерывную работу даже при отказе одного субкомпонента. Только специализированное программное обеспечение работает на отказоустойчивом оборудовании.

Такие системы являются очень дорогостоящими и чрезвычайно специализированными. Внедрение отказоустойчивого решения требует больших затрат и значительных усилий на настройку всех компонентов системы.

В средах, где отключения абсолютно неприемлемы (в жизненно важных системах), использование отказоустойчивого оборудования и решений является обязательным.

1.5.2 Системы высокой доступности

Системы высокой доступности представляют собой сочетание аппаратных и программных компонентов, сконфигурированное для совместной работы, обеспечивающее в случае сбоя автоматическое восстановление с минимальным допустимым временем отключения.

В таких системах используемое специальное программное обеспечение обнаруживает проблемы в среде и управляет живучестью приложения, выполняя его перезапуск на том же или на другом доступном компьютере (сохраняя идентичность первоначального компьютера-узла).

Таким образом, очень важно устранить все единые точки отказа в среде. Например, если компьютер имеет только один сетевой интерфейс (подключение), на этом узле необходимо установить второй сетевой интерфейс (подключение), который выполнял бы требуемые функции в случае отказа основного сетевого интерфейса.

[1] В англоязычной документации по HACMP V5.3 можно встретить оба термина, по сути, являющиеся синонимами. *Примеч. науч. ред.*

Другая важная задача состоит в том, чтобы защитить данные посредством зеркального отображения и их записи в совместно используемые дисковые области, доступные с любого компьютера в кластере.

Программное обеспечение HACMP (High Availability Cluster Multi-Processing) гарантирует инфраструктуру и набор инструментов для интеграции приложений в систему высокой доступности.

Приложения, интегрируемые в кластер HACMP, требуют значительной настройки, причем не на уровне приложения, а скорее на уровне платформ HACMP и AIX.

HACMP представляет собой гибкую платформу, позволяющую осуществлять интеграцию обычных приложений, выполняющихся на платформе AIX, что дает возможность создавать системы высокой доступности с приемлемой стоимостью.

Важно помнить, что HACMP не является отказоустойчивым решением, и при внедрении это необходимо учитывать.

1.6 Планирование программного обеспечения

В процессе планирования кластера HACMP один из наиболее важных этапов состоит в том, чтобы определить параметры программного обеспечения, которое будет выполняться на узлах кластера.

Определяющими факторами при планировании программного обеспечения узла являются:

- требования операционной системы: версия AIX и рекомендуемые параметры;
- совместимость приложения: необходимо убедиться в соответствии параметров приложения всем установленным требованиям и в том, что оно поддерживается в кластерных средах;
- ресурсы: используемые типы ресурсов (IP-адреса, конфигурация системы хранения данных, использование NFS и т. д.).

1.6.1 Уровень AIX и соответствующие требования

Перед установкой HACMP необходимо проверить другие требования уровня программного обеспечения.

Актуальный список рекомендованных параметров для HACMP V5.1, 5.2 и V5.3 см. на веб-сайте IBM по адресу

http://www-912.ibm.com/eserver/support/fixes/fcgui.jsp

Табл. 1.3 содержит рекомендованные уровни обновлений HACMP и другого программного обеспечения на момент написания этой книги.

Таблица 1.3. Требования к уровню ОС для HACMP V5.1 и V5.2

Версия HACMP	Уровень ОС AIX и прочего ПО	AIX APAR	Уровень RSCT
HACMP V5.1	5100-05	IY50579, IY48331	2.2.1.30 или выше
HACMP V5.1	5200-02	IY48180, IY44290	2.3.1.0 или выше
HACMP V5.2	5100-06	IY54018, IY53707, IY54140, IY55017	2.2.1.30 или выше

Окончание табл. 1.3

Версия HACMP	Уровень ОС AIX и прочего ПО	AIX APAR	Уровень RSCT
HACMP V5.2	5200-03	IY56213	2.3.3.0 или выше
HACMP V5.3	5200-04	IY72082, IY72946, IY72928	2.3.6 или выше
HACMP V5.3	5300-02	IY71500, 72852, IY72916, IY72928	2.4.2 или выше
HACMP 5.3, CBU, DLPAR, CUoD	5.2 / 5.3	IY73050, IY73051	
HACMP/XD:HAGeo -		Дополнительные требования отсутствуют	
HACMP/XD:GLVM	5200-04	IY66555	
HACMP/XD:GLVM	5300-02	IY68029, IY68300	

Примечание. Для использования C-SPOC с дисками VPATH требуется Subsystem Device Driver (SDD) 1.3.1.3 или более поздней версии.
Для применения HACMP Online Planning Worksheets требуется AIX 5L Java Runtime Environment 1.3.1 или более поздней версии и графический дисплей (локальный или удаленный).
HACMP V5.1 и V5.2 поддерживают использование драйверов устройств MPIO (Multi-path I/O) AIX 5L V5.2 для доступа к дисковым подсистемам.

Требования HACMP/XD с использованием ESS/PPRC:

- AIX 5L Java 1.3.0.13 или более поздней версии;
- микрокод ESS 2.1.1 или более поздней версии;
- интерфейс командной строки 2105 (ibm2105cli.rte.32.6.200.13 или ibm2105esscli.rte.2.1.0.15);
- драйвер устройств подсистемы IBM 2105 (ibmSdd_510nchacmp.rte 1.3.3.6 или выше);
- ESS eRCMF V2.0 для HACMP/XD для eRCMF.

Требования HACMP/XD с использованием SVC/PPRC:

- openssh 3.6.1 или более поздней версии;
- драйвер устройств подсистемы IBM 2145 -devices.fcp.disk.ibm.rte (1.0.0.0), devices.sdd.5.2.rte, devices.fcp.disk.ibm2145.rte.

Следующие компоненты базовой операционной системы AIX являются обязательными для работы HACMP:

- bos.adt.lib,
- bos.adt.libm,
- bos.adt.syscalls,
- bos.net.tcp.client,
- bos.net.tcp.server,
- bos.rte.SRC,
- bos.rte.libc,
- bos.rte.libcfg,

- bos.rte.libcur,
- bos.rte.libpthreads,
- bos.rte.odm,
- bos.data.

При использовании одновременного доступа к ресурсам [(enhanced) concurrent] также необходимы следующие компоненты:

- bos.rte.lvm.5.1.0.25 или выше (для AIX 5L V5.1);
- bos.clvm.enh (необходим для работы LVM).

Полный список рекомендованных пакетов обновлений (maintenance levels) для AIX 5L V5.1 и V5.2 см. на веб-странице IBM по адресу

http://www-912.ibm.com/eserver/support/fixes/fcgui.jsp

1.6.2 Лицензирование

Большинство производителей программного обеспечения требуют наличия лицензий для каждого приложения, для каждого физического компьютера или процессора многопроцессорного компьютера (SMP). Обычно код активизации лицензии вводится во время установки.

Однако в среде HACMP при перехвате, в случае перезапуска приложения на другом узле нужно убедиться в наличии необходимых кодов активизации (лицензий) для нового компьютера; в противном случае приложение может не запуститься должным образом.

Приложение также может требовать наличия уникальной лицензии, привязанной к узлу (отдельного файла лицензии на каждом узле).

Некоторые приложения также содержат ограничения на количество плавающих лицензий для этого приложения в кластере. Во избежание этой проблемы необходимо убедиться, что вы имеете достаточно лицензий для каждого узла кластера, чтобы приложение могло выполняться одновременно на нескольких узлах (особенно это касается действующих приложений, работающих в параллельном (конкурентном) режиме).

1.7 Установка программного обеспечения HACMP

Программное обеспечение HACMP содержит набор средств, которые можно использовать для обеспечения высокой доступности ваших приложений. Необходимо помнить о том, что защита HACMP распространяется не на все компоненты системы или приложения.

Например, если все данные критического приложения находятся на одном диске, этот диск является единой точкой отказа для всего кластера и защита HACMP на него не распространяется. В этом случае необходимо применять защиту с использованием диспетчера логических томов (LVM) AIX или подсистем хранения. HACMP только лишь обеспечивает перехват диска на резервном узле, обеспечивая доступ к данным.

Планирование HACMP потому является столь важным, что главная цель в процессе планирования состоит в устранении единых точек отказа. Единая точка отказа имеет место в том случае, когда критическая функция кластера обеспечивается одним компонентом. При отказе этого компонента кластер не имеет альтернативного способа обеспечения этой функции и приложение или служба, зависящие от этого компонента, становятся недоступными.

Также необходимо помнить о том, что при хорошем планировании кластер является простым в установке, обеспечивает более высокую доступность приложения, корректно работает и требует меньше обслуживания, чем при плохом планировании.

1.7.1 Проверка соответствия требованиям

По окончании планирования необходимо убедиться в том, что ваша система соответствует требованиям, предъявляемым к HACMP; это позволит избежать множества потенциальных ошибок.

HACMP V5.1 требует наличия одного из следующих компонентов операционной системы:

- AIX 5L V5.1 ML5 с RSCT V2.2.1.30 или выше;
- AIX 5L V5.2 ML2 с RSCT V2.3.1.0 или выше (рекомендуемая версия – 2.3.1.1);
- AIX 5L V5.3 ML2 с RSCT V2 или выше (рекомендуемая версия – 2);

Поддержка C-SPOC vpath требует использования SDD 1.3.1.3 или выше.

Наиболее актуальные сведения о требованиях и APAR см. в файле README, поставляемом вместе с продуктом, а также на следующей веб-странице IBM:

http://techsupport.services.ibm.com/server/cluster/

HACMP V5.2 требует наличия одного из следующих компонентов операционной системы:

- AIX 5L V5.1 ML5 с RSCT V2 или выше;
- AIX 5L V5.2 ML2 с RSCT V2 или выше (рекомендуемая версия – 2.3);
- AIX 5L V5.3 ML2 с RSCT V2 или выше (рекомендуемая версия – 2);

Поддержка C-SPOC vpath требует использования SDD 1 или выше.

HACMP V5.3 требует наличия одного из следующих компонентов операционной системы:

- AIX 5L V5.3 ML2 с RSCT V2 или выше (рекомендуемая версия – 2.3);

Поддержка C-SPOC vpath требует использования SDD 1. или выше.

1.7.2 Новая установка

HACMP может быть установлен с использованием программы AIX Network Installation Management (NIM), включающей опцию Alternate Disk Migration. Вам необходимо установить наборы файлов (filesets) HACMP на каждом узле кластера. Установку наборов файлов HACMP можно выполнить либо с использованием NIM, либо из локального хранилища программного обеспечения.

Установка с использованием сервера NIM

Мы рекомендуем использование NIM просто потому, что это позволяет загрузить программное обеспечение HACMP на другие узлы с сервера быстрее, чем с другого носителя. Кроме того, это является гибким методом распространения, обновления и администрирования узлов. Он позволяет установить несколько узлов параллельно и обеспечить среду для выполнения обновления программного обеспечения. В больших средах это может быть очень полезно и позволяет сэкономить время; в небольших средах может быть достаточно локального хранилища.

Если вы решили использовать NIM, необходимо скопировать все наборы файлов HACMP на сервер NIM и определить ресурс lpp_source, прежде чем продолжать установку.

Установка с компакт-диска или жесткого диска

Если ваша среда содержит лишь несколько узлов или если вы считаете использование NIM излишним, можно воспользоваться установкой с компакт-диска или создать локальное хранилище, скопировав наборы файлов HACMP на локальный диск и воспользовавшись командой **exportfs**; это позволит другим узлам осуществлять доступ к данным с использованием NFS.

Прочие примеры установки, например пример установки на SP-системах, а также инструкции по созданию сервера установки см. в части 3, «Network Installation», руководства **AIX 5L Version 5.2 Installation Guide and Reference,** SC23-4389.

1.7.3 Установка HACMP

Перед установкой HACMP настоятельно советуем просмотреть документацию к релизу (release notes) HACMP V5.1, находящуюся в каталоге /usr/es/lpp/cluster/doc. на наличие последней информации о требованиях или известных проблемах.

Для установки программного обеспечения HACMP на серверном узле необходимо выполнить следующие действия:

1. При инсталляции непосредственно с установочного носителя, например с компакт-диска или из локального хранилища, введите быстрый путь (fast path) **smitty install_all**. SMIT выведет экран Install and Update from ALL Available Software (Установка и обновление программного обеспечения).
2. Введите имя устройства, соответствующего установочному носителю, или каталог установки в поле INPUT device/directory for software (Устройство/каталог с программным обеспечением) и нажмите Enter.
3. Введите соответствующие значения полей.

Чтобы выбрать устанавливаемое программное обеспечение, нажмите F4 для вывода списка программного обеспечения или введите `all` для установки всех образов серверов и клиентов. Выберите пакеты, которые требуется установить, в соответствии с конфигурацией вашего кластера. Некоторые из пакетов могут потребовать использования систем, отсутствующих в вашей среде (например, Tivoli Monitoring).

Образы cluster.es и cluster.cspoc (содержащие исполняемый модуль HACMP) являются обязательными и должны быть установлены на всех серверах.

> **Примечание.** При установке диспетчера одновременного доступа к ресурсам (Concurrent Resource Manager) необходимо установить LPP cluster.es.clvm, и при выборе cluster.es и cluster.cspoc необходимо также выбрать соответствующие пакеты сообщений.

В поле Accept new license agreements (Принять условия лицензионного соглашения) необходимо выбрать Yes. Это нужно для продолжения установки. Если выбрать No, установка может прерваться с выдачей предупреждения о том, что для одного или нескольких наборов файлов требуются лицензии на программное обеспечение. Для каждого узла лицензионное соглашение принимается только один раз.

4. Нажмите Enter для запуска процесса установки.

Действия после установки

После установки программного обеспечения HACMP необходимо выполнить следующие действия:

1. Проверьте установленное программное обеспечение с использованием команды AIX **lppchk**, а также просмотрите установленные каталоги на наличие требуемых файлов.
2. Выполните команды **lppchk -v** и **lppchk -c cluster***. Если установка прошла без ошибок, обе команды выполняются без выдачи сообщений; в противном случае следует использовать соответствующие методы устранения неполадок.
3. Каждый узел кластера необходимо перезагрузить.

1.7.4 Пути и варианты миграции

При обновлении или преобразовании кластера HACMP возможны следующие варианты: поузловая миграция и миграция на основе снимка (snapshot).

Поузловая миграция

Вариант поузловой миграции используется в том случае, если необходимо обеспечить доступность приложения во время процесса миграции. Ниже перечислены действия, выполняемые при поузловой миграции.

1. Остановите службы кластера на одном из узлов кластера.
2. Выполните обновление программного обеспечения HACMP.
3. Выполните реинтеграцию узла в кластер.

Этот процесс также называется *циклической миграцией (rolling migration)*. Этот вариант миграции имеет ряд ограничений; дополнительные сведения см. в разделе 1.7.6, «Поузловая миграция».

Если вы можете отключить приложение для его обслуживания, миграция будет включать следующие действия:

1. Остановите службы кластера на всех узлах кластера.
2. Обновите программное обеспечение HACMP на каждом узле.
3. Поочередно запустите службы кластера на каждом узле.

Миграция на основе снимка

Кроме того, можно перевести весь кластер на HACMP V5.1 с использованием функции снимка кластера. Однако во время этого процесса кластер будет недоступен и все узлы должны быть обновлены до повторной активизации кластера. Дополнительные сведения см. в разделе 1.7.5, «Преобразование снимка кластера».

1.7.5 Преобразование снимка кластера

Этот метод миграции предназначен для случаев, когда для AIX и HACMP необходимо выполнить обновление и миграцию одновременно (например, при переходе с AIX V4.3.3 и HACMP V4.4.1 на AIX 5L™ V5.1 и HACMP V5.1).

> **Важно!** Не оставляйте кластер в режиме одновременного использования различных версий надолго, так как при этом нельзя гарантировать высокую доступность.

При миграции с ранней поддерживаемой версии HACMP (HAS) на HACMP V5.X можно выполнить миграцию кластера без снимка. Необходимо сохранить таблицу планирования (planning worksheet) и конфигурационные файлы из текущей конфигурации для последующего использования в случае, если потребуется сконфигурировать кластер HACMP так же, как он был сконфигурирован в предыдущей установке. Затем требуется удалить компоненты программного обеспечения HACMP, повторно их установить в более поздней версии HACMP и сконфигурировать их в соответствии с сохраненными файлами планирования и конфигурации.

> **Примечание.** Нужно помнить о том, что после миграции или обновления ни одна из функций HACMP V5.X еще не является активной. Для активизации новых функций (усовершенствований) необходимо сконфигурировать опции и синхронизировать кластер.

Для перехода с поддерживаемой версии HAS на HACMP необходимо выполнить следующие действия:

1. Убедитесь, что текущее программное обеспечение является committed (не находится в статусе applied).
2. Сохраните конфигурацию своего кластера HAS в виде снимка, а также сохраните все настроенные скрипты обработки событий, которые требуется оставить.
3. Удалите программное обеспечение HAS на всех узлах кластера.
4. Установите программное обеспечение HACMP V5.1.
5. Проверьте установленное программное обеспечение.
6. Преобразуйте и примените сохраненный снимок.

Инструмент создания снимков кластера позволяет сохранить конфигурацию кластера в файл посредством выполнения следующих действий:

1. Переустановите модифицированные вами скрипты обработки событий.
2. Перезагрузите каждый узел.

3. Выполните синхронизацию и проверьте конфигурацию HACMP V5.1.

   Эти действия подробно описаны в следующем разделе.

Проверка предыдущих версий HACMP

Чтобы просмотреть, существует ли программное обеспечение HACMP Classic (HAS) в вашей системе, введите следующую команду:

 # lslpp -h "cluster*"

Если выходные данные команды `lslpp` покажут, что установлена HACMP версии ниже V4.5, то, прежде чем использовать утилиту преобразования снимков, необходимо сначала выполнить обновление как минимум до версии V4.5. Дополнительные сведения см. в руководстве **HACMP for AIX 5L V5.1 Administration and Troubleshooting Guide, SC23-4862-02**.

Сохранение конфигурации кластера и модифицированных скриптов обработки событий

Для сохранения конфигурации кластера HACMP (HAS) версии V4.5 или выше, необходимо создать снимок в HACMP (HAS). Если у вас есть модифицированные скрипты обработки событий, они также должны быть сохранены.

> **Внимание!** Не сохраняйте конфигурацию своего кластера или модифицированные скрипты обработки событий в каталогах /usr/sbin/cluster, /usr/es/sbin/cluster и /usr/lpp/cluster. При установке новых пакетов HACMP эти каталоги удаляются и создаются заново.

Как удалить программное обеспечение HACMP Classic (HAS)

Для удаления программного обеспечения HACMP и конфигурации кластера на узлах кластера и клиентах, необходимо выполнить следующие действия:

1. Введите быстрый путь **smitty install_remove**. Должен будет появиться экран, представленный в примере 1.1.

Пример 1.1. Удаление установленного программного обеспечения

```
                        Remove Installed Software

 Type or select values in entry fields.
 Press Enter AFTER making all desired changes.

                                                      [Entry Fields]
 * SOFTWARE name                                     [cluster*]              +
   PREVIEW only? (remove operation will NOT occur)    yes                    +
   REMOVE dependent software?                         no                     +
   EXTEND file systems if space needed?               no                     +
   DETAILED output?                                   no                     +

   F1=Help             F2=Refresh          F3=Cancel           F4=List
   F5=Reset            F6=Command          F7=Edit             F8=Image
   F9=Shell            F10=Exit            Enter=Do
```

Установка HACMP V5.1

При установке программного обеспечения следуйте инструкциям, приведенным в разделе 1.7.3, «Установка HACMP».

> **Примечание.** Не выполняйте перезагрузку до преобразования и применения сохраненного снимка.

Проверка установленного программного обеспечения

После установки HACMP проверьте наличие требуемых файлов с использованием команды **`lppchk`**. Дополнительные сведения см. в разделе «Действия после установки».

Преобразование и применение сохраненного снимка

После установки HACMP V5.1 на узлах кластера необходимо преобразовать и применить снимок, сохраненный в прежней конфигурации.

> **Важно!** Преобразование снимка необходимо выполнять перед перезагрузкой узлов кластера.

Для преобразования и применения снимка:

1. Используйте утилиту **clconvert_snapshot**, указав номер версии HACMP (HAS) и имя преобразуемого файла снимка. Флаг -C выполняет преобразование снимка HACMP (HAS) в формат снимка HACMP V5.1:

   clconvert_snapshot -C -v версия -s <имя_файла>

2. Примените снимок.

Переустановка сохраненных модифицированных скриптов обработки событий

Переустановите требуемые модифицированные скрипты обработки событий, сохраненные в предыдущей конфигурации.

> **Примечание.** Некоторые скрипты обработки событий, использовавшиеся в прежних версиях, могут оказаться ненужными в HACMP V5.1, особенно в группах ресурсов, применяющих параллельную обработку.

Перезагрузка узлов кластера

Перезагрузка узлов кластера необходима для активизации коммуникационного демона нового кластера (clcomdES).

Проверка и синхронизация конфигурации кластера

После установки программного обеспечения HACMP и перезагрузки каждого узла необходимо выполнить проверку (верификацию) и синхронизацию топологии кластера. При верификации удостоверяется согласованность определения кластера на всех узлах с выдачей соответствующих ошибок и/или предупреждений. В следующем разделе мы кратко рассмотрим процесс проверки кластера.

Выполните команду **smitty hacmp** и выберите пункт **Extended Configuration → Extended Verification and Synchronization (Расширенная конфигурация -> Расширенная верификация и синхронизация)**, выберите **Verify changes only (Верифицировать только изменения)**, после чего нажмите Enter (пример 1.2).

Пример 1.2. Верификация и синхронизация HACMP

```
   HACMP Verification and Synchronization (Active Cluster on a Local Node)

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                      [Entry Fields]
* Emulate or Actual                                  [Actual]                +
  Force synchronization if verification fails?       [No]                    +
* Verify changes only?                               [No]                    +
* Logging                                            [Standard]              +


F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

> **Важно!** Нельзя выполнить синхронизацию в кластере со смешанными версиями. При обновлении не следует надолго оставлять кластер со смешанными версиями HACMP. Новые функции версии V5.1 доступны только после обновления всех узлов и синхронизации кластера.

1.7.6 Поузловая миграция

При выполнении поузловой (циклической) миграции необходимо учитывать следующее:

- На всех узлах кластера должно быть установлено состояние committed ПО HACMP V4.5.
- Поузловая миграция выполняется только с версии HACMP (HAS) V4.5 на версию HACMP V5.1.
- Все узлы в кластере должны быть включены и на них должно выполняться программное обеспечение HAS V4.5.
- Кластер должен находиться в стабильном состоянии.

Должно быть достаточно дискового пространства, чтобы в процессе миграции хватило места как для программного обеспечения HAS, так и для программного обеспечения HACMP:

  – приблизительно 120 Мб в каталоге /usr;
  – приблизительно 1.2 Мб в каталоге / (root);

- После завершения миграции требования к дисковому пространству снижаются до обычных показателей, необходимых для работы HACMP V5.1.
- Узлы должны иметь достаточно памяти для одновременного выполнения демонов HACMP (HAS) и HACMP. Это составляет как минимум 64 Мб оперативной памяти. Рекомендуется иметь 128 Мб оперативной памяти.
- Убедитесь, что у вас отсутствуют типы сетей, неподдерживаемые в HACMP. После начала миграции уже нельзя будет внести изменения в конфигурацию. Необходимо удалить или изменить неподдерживаемые типы заранее. Дополнительные сведения см. в гл. 3, «Planning Cluster Network Connectivity», руководства **HACMP for AIX 5L V5.1 Planning and Installation Guide**, SC23-4861-02.

> **Важно!** Как и при любой миграции, не пытайтесь внести какие-либо изменения в топологию или ресурсы кластера после начала процесса миграции.

- Если какие-либо узлы в кластере в настоящее время настроены на автоматический запуск служб кластера при перезагрузке, измените этот параметр до начала процесса миграции. Следующие процедуры описывают, как отключить автоматический запуск кластера.
  - Используйте C-SPOC для отключения автоматического запуска служб кластера при перезапуске системы.
  - Используйте быстрый путь SMIT **smitty clstop** и выберите опции, показанные в примере 1.3.

Пример 1.3. Остановка служб кластера

```
                         Stop Cluster Services

 Type or select values in entry fields.
 Press Enter AFTER making all desired changes.

                                                      [Entry Fields]
 * Stop now, on system restart or both                on system restart      +
   Stop Cluster Services on these nodes               [p630n01]              +
   BROADCAST cluster shutdown?                         true                  +
 * Shutdown mode                                       graceful              +


 F1=Help             F2=Refresh          F3=Cancel           F4=List
 F5=Reset            F6=Command          F7=Edit             F8=Image
 F9=Shell            F10=Exit            Enter=Do
```

Если вы не используете C-SPOC, необходимо изменить этот параметр на каждом узле в отдельности.

Как осуществить поузловую миграцию

Для осуществления поузловой миграции с HACMP V4.5 на HACMP V5.1 необходимо выполнить следующие действия:

1. Сохраните текущую конфигурацию в снимке (в качестве меры предосторожности). Поместите его в безопасный каталог (не затрагиваемый процедурами установки). Не используйте каталог /usr/sbin/cluster.
2. Остановите службы кластера на одном из узлов, на котором выполняется HAS V4.5, с использованием метода постепенной остановки с передачей ресурсов на резервные узлы (graceful with takeover). Для остановки служб кластера из командной строки введите

   # /usr/es/sbin/cluster/utilities/clstop -gr
3. Прежде чем продолжить, убедитесь в том, что службы кластера на узле остановлены и что ресурсы кластера с этого узла были переданы на резервные узлы.
4. Установите HACMP V5.1 на узле. Инструкции по установке см. в разделе «Установка программного обеспечения HACMP».
5. Выполните проверку установленного программного обеспечения с использованием команды AIX lppchk. См. раздел «Действия после установки».
6. Перезагрузите узел.
7. Перезапустите программное обеспечение HACMP:
   а) введите быстрый путь smitty hacmp;
   б) перейдите в System Management (C-SPOC);
   в) выберите Manage HACMP Services;
   г) выберите Start Cluster Services.

Перезапуск служб кластера

При перезапуске также используется программное обеспечение HACMP: на узле запускаются службы кластера HACMP, после чего узел заново присоединяется к кластеру.

Узел заново получает ресурсы типа *cascading (каскадные),* для которых он является первичным узлом (это зависит от параметра Inactive Takeover).

Теперь на узле выполняется и старая и новая версия HACMP (т. е., HACMP V4.5 и HACMP/ES V5.1), но только HACMP Classic (HAS) управляет событиями и ресурсами кластера. При выводе списка демонов, управляемых контроллером системных ресурсов (system resource controller, SRC), он будет содержать следующие демоны на этом гибридном узле (табл. 1.4).

Таблица 1.4. Список демонов, используемых HACMP

HACMP	HACMP/ES	RSCT
clstmgr	clstmgrES	grpsvcs
cllockd (необязательный)	cllockdES (необязательный)	topsvcs
clsmuxpd	clsmuxpES	emsvcs
clinfo (необязательный)	clinfoES (необязательный)	grpglsm
	clcomdES	emaixos

8. Повторите действия пп. 2–6 для всех узлов в кластере.

Внимание! Запуск служб кластера на последнем узле можно считать точкой необратимости.
После перезапуска HACMP (при этом выполняется перезапуск обеих версий HACMP) на последнем узле, когда миграция уже была начата, ее нельзя обратить.
Если после этого вам потребуется возвратиться к прежней конфигурации HACMP, придется переустановить программное обеспечение HACMP и применить сохраненный снимок. До этого момента можно отменить установку HACMP и вернуться к прежней конфигурации кластера HACMP. Если вам потребуется это сделать, см. раздел «Процедура возврата».

Во время процесса установки и миграции, при перезапуске каждого узла происходит запуск обоих продуктов, где контроль обработки событий кластера осуществляет HACMP clstrmgr, а clstrmgrES работает в пассивном режиме.

После перезапуска служб кластера на последнем узле миграция на HACMP продолжается автоматически. Полное управление кластером автоматически передается демонам HACMP V5.1.

Сообщения, описывающие процесс миграции, записываются в файл /tmp/hacmp.out, а также в файлы журналов /tmp/cm.log и /tmp/clstrmgr.debug.

После окончания миграции, когда все узлы кластеров включены и на них выполняется HACMP V5.1, происходит удаление программного обеспечения HACMP (HAS).

9. После обновления и перезагрузки всех узлов кластер перейдет в стабильное состояние и нужно будет выполнить синхронизацию и проверку конфигурации.

Также после миграции следует обязательно протестировать корректность выполнения операций перемещения при сбое и восстановления.

Примечание. В процессе поузловой миграции с HAS 4.5 на HACMP V5.1 будут выдаваться следующие предупреждения:
sysck: 3001-036 WARNING: File /etc/cluster/lunreset.lst is also owned by fileset cluster.base.server.events.
sysck: 3001-036 WARNING: File /etc/cluster/disktype.lst is also owned by fileset cluster.base.server.events.
Можете благополучно игнорировать эти предупреждения и продолжать установку.

Сообщение config_too_long

После завершения процесса миграции во время удаления наборов файлов HACMP может появиться сообщение **config_too_long**. Это сообщение возникает, когда диспетчер кластера обнаруживает, что обработка события занимает больше установленного времени. Сообщения config_too_long добавляются в файл hacmp.out, пока не завершится событие. При возникновении таких сообщений следует периодически убеждаться в том, что событие все еще действует и что не произошел отказ.

Можно избежать вывода этих сообщений, увеличив время ожидания до вызова события config_too_long (с использованием SMIT). Чтобы изменить интервал, заданный для обработки события, нужно выполнить следующие действия:

1. Введите быстрый путь **smitty hacmp**.

2. Перейдите в **Extended Configuration**.
3. Выберите **Extended Event Configuration**.
4. Выберите **Change/Show Time Until Warning**.

Это необходимо сделать на каждом узле. Изменения вступают в действие после перезапуска служб кластера.

Как работает процесс поузловой миграции

После установки HACMP на всех узлах кластера (когда все узлы находятся в гибридном состоянии) запуск служб кластера на последнем узле кластера автоматически вызывает передачу управления в HACMP V5.1 следующим образом:

1. При установке HACMP V5.1 выполняется установка файла восстановления firstboot на узле кластера, а также происходит создание файла миграции (.mig), используемого в процессе миграции в качестве флага.
2. Драйвер восстановления HACMP направляет сообщение диспетчеру кластера HACMP, сообщая о необходимости инициировать события waiting и waiting_complete.
   - HACMP использует службы групп RSCT для проверки стабильности и членства в кластере.
   - На каждом узле файл firstboot перемещается в активный каталог (/etc).
   - На всех узлах флаг миграции (файл .mig), созданный при установке, передается из каталога HACMP V5.1 в каталог HACMP V4.5.

После перемещения файла firstboot в активный каталог и передачи файла .mig на всех узлах передача управления в HACMP продолжается инициированием события migrate.

3. Драйвер восстановления HACMP инициирует событие migrate.
   - HACMP V5.1 останавливает демоны HACMP V4.5 с использованием опции принудительной остановки (режим forced).
   - Активизируются демоны HACMP V5.1 clinfoES и clsmuxpdES, используя порты, которые прежде применялись соответствующими демонами HACMP V4.5.
4. Драйвер восстановления HACMP V5.1 инициирует событие migrate_complete.
   - Выполняется удаление HACMP V4.5. Файлы конфигурации, общие для обоих продуктов, не затрагиваются.
   - Выполняется повторная установка ссылок на базовые каталоги.
   - Происходит удаление файлов /etc/firstboot.
   - Выполняется удаление флага миграции (файл .mig) в каталоге /usr/sbin/cluster.
5. Миграция завершена.

Снимки кластера, сохраненные во время миграции

Ранее существовавшие снимки HACMP сохраняются в каталоге /usr/es/sbin/cluster/snapshots.

Обработка отказа узла в процессе миграции

При отказе узла в процессе миграции после перемещения файла firstboot в активный каталог он завершает процесс миграции во время перезагрузки узла. Однако после реинтеграции в кластер отказавший узел может иметь несинхронизированный HACMP ODM. В этом случае следует выполнить синхронизацию топологии и ресурсов кластера до реинтеграции отказавшего узла в кластер.

Процедура возврата

Если по какой-то причине вы решите не продолжать процесс миграции, вы можете удалить программное обеспечение HACMP V5.1 с узлов, на которых оно было установлено, в любой момент процесса до запуска HACMP на последнем узле.

> **Примечание.** Удаление программного обеспечения HACMP следует проводить только на локальном узле. Во время миграции не следует выбирать вариант удаления программного обеспечения с нескольких узлов.

Удаление программного обеспечения HACMP

Для этого проделайте следующее:

1. На каждом узле поочередно остановите службы кластера (остановка служб кластера показана в примере 1.3). Прежде чем продолжить, убедитесь в том, что службы кластера на узле остановлены и что соответствующие ресурсы кластера были переданы на резервные (takeover) узлы.
2. После того как вы убедитесь в том, что ресурсы на узле были должным образом переданы на резервный узел, удалите программное обеспечение HACMP V5.1 (см. раздел «Как удалить программное обеспечение HACMP Classic (HAS)».
3. Запустите HACMP на этом узле. Когда вы будете уверены в том, что ресурсы были должным образом переданы обратно на этот узел (если это необходимо), повторите эти действия на следующем узле.
4. Продолжайте этот процесс, пока HACMP не будет удален на всех узлах в кластере.

Обработка отказов синхронизации при поузловой миграции

Если вы попытаетесь внести изменение в топологию или ресурсы кластера до завершения миграции, произойдет отказ процесса синхронизации. Возникнет следующее сообщение:

cldare: Migration from HACMP V4.5 to HACMP V5.1 Detected. cldare cannot be run until migration has completed.

Для возврата после изменения необходимо восстановить активный ODM. Для этого нужно выполнить следующие действия:

1. Запустите **smitty hacmp.**
2. Перейдите в **Problem Determination Tools**.
3. Выберите **Restore HACMP Configuration Database from Active Configuration**.

1.7.7 Опции обновления

В этом разделе мы обсудим обновления для HACMP.

Поддерживаемые обновления для HACMP V5.1

Утилиты преобразования HACMP предлагают простой путь обновления до версии V5.1 с нижеперечисленных версий:

– с HACMP/ES V4.4.1 до HACMP V5.1;

– с HACMP/ES V4.5 до HACMP V5.1.

Если вам потребуется выполнить преобразование на HACMP V5.1 с более ранних версий, чем указаны выше, необходимо сначала выполнить обновление до одной из поддерживаемых версий. Затем вы сможете выполнить преобразование на HACMP V5.1. Например, для преобразования с HACMP/ES 4.2.2 на HACMP V5.1 сначала нужно будет выполнить обновление установки до HACMP/ES 4.4.1 или выше и затем выполнить обновление до HACMP V5.1.

Чтобы выполнить обновление до HACMP V5.1, нужно выполнить следующие действия:

1. Если необходимо, выполните обновление до AIX 5L V5.1 Maintenance Level 5 или выше.
2. Если необходимо, выполните проверку установки AIX.
3. Переведите текущее программное обеспечение HACMP на всех узлах в состояние commited.
4. Остановите HACMP/ES на одном узле (gracefully with takeover – постепенная остановка с передачей ресурсов на резервные узлы) с использованием команды **clstop**.
5. После успешного перемещения ресурсов с остановленного узла на резервный узел установите новое программное обеспечение HACMP (инструкции по установке программного обеспечения HACMP V5.1 см. в разделе 1.7, «Установка программного обеспечения HACMP»). Проверьте установку программного обеспечения с использованием команды AIX **lppchk**, после чего проверьте установленные каталоги на наличие требуемых файлов:

lppchk -v or lppchk -c “cluster.*”

Если установка прошла без ошибок, обе команды выполняются без выдачи сообщений.

6. Перезагрузите первый узел.
7. Запустите программное обеспечение HACMP на первом узле командой **smitty clstart** и убедитесь в успешном подключении первого узла к кластеру.
8. Повторите предыдущие действия для остальных узлов кластера поочередно.
9. Убедитесь в том, что устройство tty сконфигурировано как последовательная сеть (serial network).

10. Убедитесь в том, что с первого узла доступны все внешние диски (с помощью команды **lspv** проверьте PVID для каждого диска). Если PVID для дисков не выводятся, может потребоваться удалить диск и выполнить повторное конфигурирование.
11. После обновления всех узлов выполните синхронизацию конфигурации узлов и топологии кластера с первого узла на все узлы, как описано в следующем разделе, «cl_convert и clconvert_snapshot». Не пропускайте проверку во время синхронизации.

> **Важно!** Во время обновления не следует выполнять синхронизацию с обновленного узла, если необновленный узел остается в кластере со смешанными версиями. Утилита cl_convert назначает идентификаторы узлов (node ID) последовательно для всех узлов в кластере. Эти новые идентификаторы могут конфликтовать с уже существующими.

12. Восстановите объектный класс ODM событий HACMP, чтобы сохранить все скрипты обработки событий, сконфигурированные для вашего кластера.
13. Если нужно, внесите дополнительные изменения в кластер.
14. Проведите тестирование кластера прежде, чем перевести его в рабочую среду.

cl_convert и clconvert_snapshot

HACMP содержит утилиты преобразования **cl_convert** и **clconvert_snapshot**.

Обновление программного обеспечения HACMP/ES до последней версии HACMP включает преобразование ODM с предыдущей версии до текущей версии. При установке HACMP утилита **cl_convert** запускается автоматически. Однако в случае сбоя при установке необходимо запустить **cl_convert** из командной строки.

В случае сбоя при преобразовании следует выполнить **cl_convert** с флагом -F. Например, чтобы выполнить преобразование с HACMP/ES V4.5 на HACMP V5.1, следует использовать флаги -F и -v (version) следующим образом (обратите внимание на добавление «0» к номеру версии V4.5):

```
# /usr/es/sbin/cluster/conversion/cl_convert -F -v 4.5.0
```

Для запуска утилиты преобразования требуются:

- привилегии пользователя root;
- версия HACMP, с которой выполняется преобразование.

Утилита **cl_convert** записывает информацию о выполнении преобразования в файл /tmp/clconvert.log, что дает возможность оценить успешность выполнения преобразования. Этот файл журнала генерируется (перезаписывается) при каждом выполнении **cl_convert** или **clconvert_snapshot.**

Утилита **clconvert_snapshot** не запускается автоматически во время установки, ее следует запускать из командной строки. Выполните **clconvert_snapshot** для обновления снимков кластера при миграции с HACMP (HAS) на HACMP, как описано в этом разделе.

Обновление диспетчера одновременного доступа к ресурсам

Для установки средства одновременного доступа на узлах кластера необходимо установить диспетчер одновременного доступа к ресурсам (Concurrent Resource Manager, CRM), следуя процедуре, описанной в разделе 1.7, «Установка программного обеспечения HACMP».

AIX 5L V5.1 поддерживает расширенный режим одновременного доступа (enhanced concurrent mode, ECM). Если вы выполняете установку HACMP вместе с диспетчером одновременного доступа к ресурсам (Concurrent Resource Manager), см. гл. 2, «Initial Cluster Planning», руководства **HACMP for AIX 5L V5.1 Planning and Installation Guide**, SC23-4861-02.

Сведения о расширенном режиме одновременного доступа и о поддерживаемых дисковых устройствах общего доступа производства IBM см. в гл. 5, «Planning Shared LVM Components», руководства **HACMP for AIX 5L V5.1 Planning and Installation Guide**, SC23-4861-02. Кроме того, если вы хотите использовать диски других производителей, см. прил. D, «OEM Disk Accommodation», руководства **HACMP for AIX 5L V5.1 Planning and Installation Guide**, SC23-4861-02.

Проблемы при установке

Если при установке возникают проблемы, программа установки автоматически выполняет процесс очистки. Если по какой-либо причине после неудачной установки не была выполнена очистка, нужно выполнить следующие действия:

1. Введите **smitty install.**
2. Выберите **Software Maintenance and Utilities**.
3. Выберите **Clean Up After a Interrupted Installation**.
4. Просмотрите выходные данные SMIT (или файл /smit.log) на наличие указания причины прерывания установки.
5. Исправьте выявленные проблемы с использованием методов определения проблем AIX и повторите процесс установки.

2

Составляющие высокой доступности

В этой главе мы обсудим следующие темы:

- данные конфигурации HACMP;
- компоненты программного обеспечения;
- топология кластера;
- ресурсы и группы ресурсов;
- подключаемые модули HACMP;
- возможности (HACMP 5.1, 5.2 и 5.3);
- ограничения;
- аспекты хранения;
- конфигурация хранилища с общим доступом.

2.1 Данные конфигурации HACMP

Существует две основные составляющие конфигурации кластера:

- Топология кластера – описывает базовую инфраструктуру – узлы, сети и системы хранения. HACMP использует эту инфраструктуру для обеспечения высокой доступности другой основной составляющей – ресурсов.
- Ресурсы кластера – компоненты, которые HACMP может перемещать с одного узла на другой, например сервисные IP-метки, файловые системы и приложения.

После конфигурирования кластера топология кластера и информация о ресурсах вводятся на одном из узлов, выполняется процесс верификации, после чего выполняется синхронизация данных на других узлах кластера. HACMP хранит эти данные в своих классах ODM (Object Data Manager) на каждом узле в кластере.

Хотя конфигурирование и изменение настроек HACMP можно осуществлять с любого узла в кластере, рекомендуется выполнять административные операции с одного узла, чтобы обеспечить последовательность определений HACMP в кластере; это позволяет избежать обновления конфигурации кластера с нескольких узлов, что может привести к несогласованности данных.

Мы рекомендуем выполнить следующие основные действия по конфигурированию кластера:

- определите кластер и узлы;
- изучите (проведите процесс обнаружения - discovery) дополнительную информацию (сети, диски);
- определите топологию;
- выполните верификацию и синхронизацию топологии, после чего запустите службы кластера;
- определите ресурсы и группы ресурсов;
- выполните верификацию и синхронизацию ресурсов.

Конфигурация AIX

Вы должны знать, что HACMP при установке и/или запуске вносит некоторые изменения в систему.

Изменения при установке

- Изменения в файлах:
  - /etc/inittab;
  - /etc/rc.net;
  - /etc/services;
  - /etc/snmpd.conf;
  - /etc/snmpd.peers;
  - /etc/syslog.conf;
  - /etc/trcfmt;
  - /var/spool/cron/crontabs/root.

- Добавление группы **hacmp**.
- Кроме того, при конфигурировании и верификации кластера можно также внести изменения в файл /etc/hosts путем добавления или удаления записей.
- Изменение значений следующих сетевых опций:
  - routerevalidate. Устанавливается значение «1» – маршрут каждого подключения, содержащийся в кеше, следует подтверждать при добавлении каждого нового маршрута в таблицу маршрутизации. Это позволяет обеспечить использование приложениями, поддерживающими одно и то же подключение открытым в течение длительного времени, правильного маршрута после внесения изменений в таблицу маршрутизации.
  - nonlocsrcroute. Устанавливается значение «1» – позволяет осуществлять адресацию пакетов с флагом «source route» на узлы за пределами локальной сети.
  - ipsrcrouterecv. Устанавливается значение «1» – позволяет осуществлять прием системой пакетов с флагом «source route».

Настройка параметров операционной системы

В прошлом поддерживалась идея настройки AIX для работы HACMP, однако в настоящее время мы придерживаемся мнения, что система должна быть настроена на работу приложения, а не HACMP. Например, если система на время зависает, а HACMP реагирует, систему следует настроить таким образом, чтобы приложение не зависало. Хотя можно настроить систему так, чтобы HACMP был менее чувствительным, не существует общих правил настройки AIX для работы HACMP.

2.2 Компоненты программного обеспечения

Компоненты программного обеспечения кластера HACMP описываются следующей многоуровневой моделью.

- **Уровень приложения.** Любое приложение, для которого обеспечивается высокая доступность с использованием служб HACMP.
- **Уровень HACMP.** Программное обеспечение, реагирующее на изменения в кластере и обеспечивающее высокую доступность управляемых приложений.
- **Уровень RSCT.** Демоны, осуществляющие мониторинг членства узлов, коммуникационного интерфейса и работоспособности устройств и соответствующим образом информирующие HACMP.
- **Уровень AIX.** Обеспечивает поддержку HACMP через уровень LVM, осуществляющий управление хранением, и уровень TCP/IP, обеспечивающий связь.
- **Уровень LVM.** Обеспечивает доступ к хранилищу и сообщает информацию состояния в HACMP.
- **Уровень TCP/IP.** Обеспечивает надежную связь как между различными узлами, так и между узлом и клиентом.
- Уровень приложения может содержать:
  - код приложения (программы, демоны, расширения ядра и т. д.);
  - конфигурационные данные приложения (файлы или двоичные данные);
  - данные приложения (файлы или устройства прямого доступа).

- Уровень HACMP содержит:
  - код HACMP (двоичные файлы – демоны и исполняемые команды, библиотеки, скрипты);
  - конфигурацию HACMP (ODM, ASCII-файлы);
  - файлы журналов HACMP;
  - службы:
    - демон коммуникаций кластера (clcomdES);
    - диспетчер кластера (clstrmgrES);
    - демон информации кластера (clinfoES); и т. д.

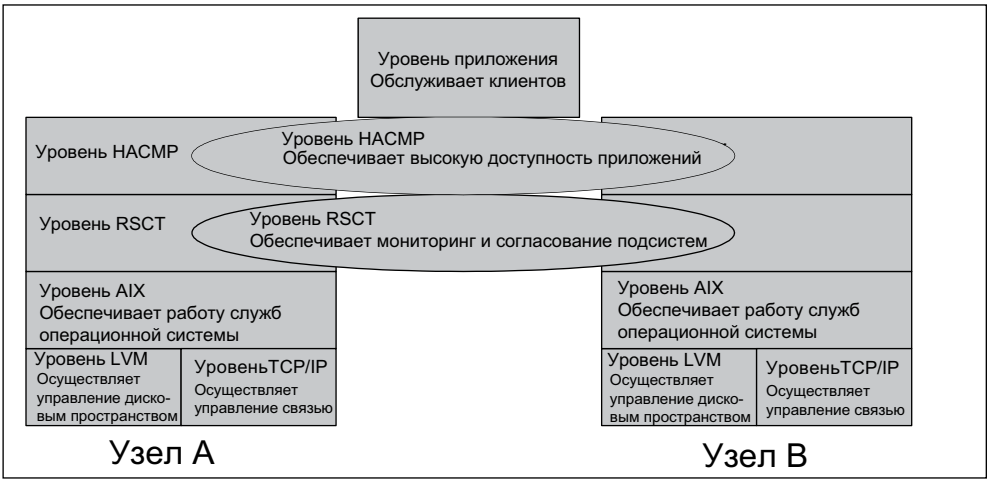

Рис. 2.1. Модель программного обеспечения кластера HACMP

- Уровень RSCT содержит:
  - код RSCT (двоичные файлы – демоны и команды, библиотеки, скрипты);
  - файлы конфигурации (двоичный регистр и ASCII-файлы);
  - службы:
    - топологии и групп (topsvcs и grpsvcs);
    - мониторинга и управления ресурсами (RMC).
- Уровень AIX содержит:
  - ядро, демоны и библиотеки;
  - драйверы устройств;
  - сетевой уровень и уровень TCP/IP;
  - диспетчер логических томов (Logical Volume Manager, LVM);
  - файлы конфигурации (ODM, ASCII).

2.3 Топология кластера

Топология кластера обозначает физическое представление кластера и соединений аппаратных компонентов кластера через сети (IP и отличные от IP). Чтобы понять работу HACMP, необходимо прежде понять базовую топологию кластера – роль каждого компонента и взаимодействие в HACMP. В этом разделе описываются:

- кластер HACMP (HACMP cluster);
- узлы (Nodes);
- сайты (Sites);
- сети (Networks);
- коммуникационные интерфейсы/устройства (Communication interfaces/devices);
- постоянные IP-метки/адреса узла (Persistent node IP labels/addresses);
- сетевые модули (Network [Interface] Modules, NIM);
- службы топологии и групп (Topology and group services);
- клиенты (Clients).

На рис. 2.2 представлена типичная топология кластера, включающего:

- три узла;

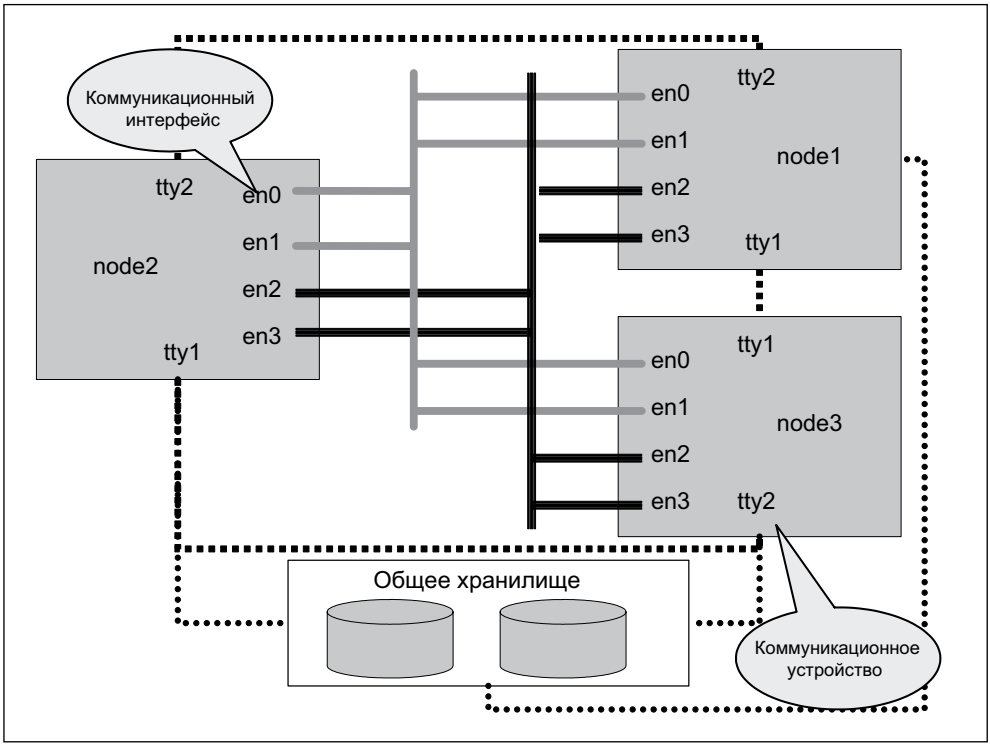

Рис. 2.2. Пример топологии кластера

- две IP-сети (логические сети HACMP) с резервированием интерфейсов на каждом узле;
- общее хранилище;
- соединения «точка-точка», отличные от IP (последовательные), между узлами, сконфигурированные в качестве независимых физических сетей, но соединяющие узлы в виде кольцевой конфигурации.

Кластер HACMP

Кластеру присваивается имя длиной до 32 символов (из символов [a–z], [A–Z], [0–9], «_») и начинающееся с буквы. Также с кластером связывается идентификатор (ID) кластера (число). HACMP 4.5 и более поздние версии генерируют уникальный идентификатор кластера автоматически. Этот идентификатор используется во всех пакетах пульса (heartbeat packets), поэтому два кластера в одной сети не должны иметь одинаковый идентификатор.

Узлы кластера

Узлы составляют ядро кластера HACMP. Узел представляет собой сервер, на котором выполняется образ операционной системы AIX (автономный или раздел), код HACMP и программное обеспечение приложения. Максимальное количество узлов, поддерживаемое кластером HACMP, – 32.

При определении узла кластера необходимо назначить ему уникальное имя и путь для связи (communication path) (IP-адрес или преобразуемая в адрес IP-метка, связанная с одним из интерфейсов на этом узле). Начиная с HACMP 5.1 в качестве имени узла может использоваться короткое имя узла, полное доменное имя узла (hostname. domain.name) или любое имя длиной до 32 символов (из символов [a–z], [A–Z], [0–9], «_») и начинающееся с буквы.

Путь для связи сначала используется HACMP для подтверждения наличия доступа к узлу, затем используется для наполнения ODM на каждом узле в кластере после установления безопасных соединений между узлами. Однако после конфигурирования топологии кластера HACMP может использовать любой интерфейс для связи между узлами в кластере.

HACMP больше не требует, чтобы имя узла представляло преобразуемую в адрес IP-метку, т. е. адрес на одном из IP-интерфейсов. В целях согласованности рекомендуем использовать имя хоста (hostname), подлежащее разрешению в постоянный IP-адрес (persistent IP address), связанный с узлом, однако это не является обязательным условием.

Внимание! На момент публикации ситуация такова, что при конфигурировании HACMP с CUoD или DLPAR имена LPAR (определенные на HMC) должны соответствовать именам узлов HACMP и именам хостов (hostname) в AIX.

Сайты

Использование сайтов не является обязательным. Они предназначены для применения в конфигурациях с межсайтовым зеркальным отображением (cross-site mirroring) и/или HACMP/XD. Сайт состоит из одного или нескольких узлов, сгруппированных в определенном месте. HACMP поддерживает разделение кластера на два сайта. Взаимоотношения сайтов также могут быть частью определения группы ресурсов, однако этот атрибут надо игнорировать (поставить в значение ignore), если сайты не определены/используются.

Можно применять сайты вне конфигураций HACMP/XD и зеркального отображения, однако в этом случае необходимо реализовать соответствующие методы настройки для обеспечения операций сайта. Если сайты определены, события сайтов обрабатываются во время событий node_up и node_down.

Кроме того, существует две характеристики сайтов, которые необходимо определить:

- **Доминирование (dominance).** Какой из сайтов является доминирующим.
- **Резервные связи сайтов (site backup communications).** Могут быть либо не установлены, либо установлены **dbfs** (dial back fail safe) или **sgn** (для сети geo_secondary).

Методы резервной связи используются при отказе основной IP-сети связи между двумя сайтами во избежание разделения сайтов (т. н. «split brain»[1]). Сайт, не являющийся доминирующим, попытается связаться с доминирующим сайтом, используя сеть резервной связи сайтов, и, если доминирующий сайт все еще работает, он остановится.

Сети

В HACMP термин «сеть» используется для определения логического объекта, объединяющего коммуникационные интерфейсы и устройства, используемые для связи между узлами в кластере, а также для доступа клиентов. В HACMP сети могут быть определены как IP-сети (IP networks) или как сети, отличные от IP (non-IP networks).

При описании сетевых функций HACMP используются следующие термины:

- IP-адрес: десятичный IP-адрес с разделяющими точками.
- IP-метка (IP label): метка, связанная с конкретным IP-адресом, определенная методом разрешения имен (DNS или статический, т. е. /etc/hosts).
- Базовая IP-метка/адрес (Base IP label/address): заданная по умолчанию IP-метка/адрес, установленная для интерфейса операционной системой AIX при запуске. Базовый адрес интерфейса.
- Сервисная IP-метка/адрес (Service IP label/address): IP-метка/адрес, через который предоставляется обслуживание; может быть привязан к одному узлу или совместно использоваться несколькими узлами. Хотя эти адреса не являются частью топологии, HACMP обеспечивает их высокую доступность.

---

[1] Ситуация «split brain» возникает, когда каждый сайт (или узел в кластере) считает, что соседний сайт (узел) полностью вышел из строя, в то время как он на самом деле продолжает работать. В результате каждый сайт пытается завладеть ресурсами, что может привести к непредсказуемым (и даже катастрофическим) результатам. *Примеч. науч. ред.*.

- Загрузочный интерфейс (Boot interface). В ранних версиях HACMP использовались термины «загрузочный адаптер» (boot adapter) и «резервный адаптер» (standby adapter), в зависимости от функции. Эти термины были объединены в один термин, описывающий любой интерфейс IP-сети, который может использоваться HACMP для содержания сервисной IP-метки/адреса.
- IP-синонимы (IP aliases). IP-синоним представляет собой IP-адрес, добавляемый к интерфейсу, но не заменяющий его базовый IP-адрес. Является функцией AIX, поддерживаемой HACMP, хотя HACMP все еще требует использования только одной маски подсети для всех адресов, связанных с адаптером.
- Логический сетевой интерфейс (Logical network interface). Имя, в которое AIX осуществляет разрешение порта (например, en0) физического сетевого адаптера.

Рекомендуется, чтобы все вышеперечисленные IP-адреса были определены в одном файле /etc/hosts и чтобы этот файл был одинаковым на всех узлах кластера. При этом, конечно же, необязательно использовать полные доменные имена. Когда HACMP осуществляет обработку изменений в сети, переменная NSORDER установлена в значение local (т. е. для разрешения имен используется /etc/hosts); и все же рекомендуется, чтобы это было указано в файле /etc/netsvc.conf.

Коммуникационные интерфейсы HACMP

Термин «коммуникационный интерфейс» (или просто «интерфейс») обозначает физический адаптер, поддерживающий протокол TCP/IP и представленный IP-адресом. Сетевые интерфейсы, подключенные к общей физической сети, объединяются в логические сети, используемые HACMP.

Каждый интерфейс может иметь несколько TCP/IP-адресов. При конфигурировании кластера определяются IP-адреса, для которых HACMP осуществляет мониторинг с использованием RSCT (базовые или загрузочные IP-адреса), а также IP-адреса, для которых следует обеспечивать высокую доступность (сервисные IP-адреса и постоянные синонимы).

Коммуникационные устройства HACMP

Топология HACMP также включает отличные от IP сети (non-IP networks) типа «точка-точка», такие, как последовательная сеть RS232, target mode SCSI, target mode SSA, и подключения мониторинга пульса через диски (disk heartbeat). На обоих концах сети «точка-точка» находятся устройства AIX (определенные в каталоге /dev), такие, как /dev/tty1, /dev/tmssa1, /dev/tmscsi1 и /dev/hdisk1.

Например, при мониторинге пульса через диски имя дискового устройства (например, /dev/hdisk2) используется в качестве устройства, сконфигурированного в HACMP на каждом конце подключения.

Эти отличные от IP сети представляют собой соединения «точка-точка» между двумя узлами кластера и используются RSCT для управления трафиком сообщений и мониторинга пульса. Эти сети обеспечивают дополнительный уровень защиты кластера HACMP на случай отказа IP-сетей или подсистемы TCP/IP на узлах.

Коммуникационные адаптеры и связи включают адаптер X.25, используемый для обеспечения коммуникационной связи с высокой доступностью. В качестве ресурсов HACMP может быть определено следующее:

- SNA, сконфигурированный через адаптеры локальной сети;
- SNA, сконфигурированный через адаптеры X.25;
- собственные связи X.25.

HACMP осуществляет управление этими связями в составе групп ресурсов, обеспечивая, таким образом, коммуникационные связи высокой доступности. В случае отказа физического сетевого интерфейса, отказа связи X.25 или отказа узла коммуникационная связь высокой доступности переносится на другой доступный адаптер на том же узле или на резервный узел (вместе со всеми ресурсами в той же группе ресурсов).

Физические и логические сети

Физическая сеть соединяет два или больше физических сетевых устройства. Существует множество типов физических сетей, и в HACMP они разделяются на IP-сети и сети, отличные от IP:

- сети TCP/IP, такие, как Ethernet и Token Ring;
- сети устройств, такие, как RS-232, target mode SCSI (tmscsi), target mode SSA (tmssa), или сети мониторинга пульса через диски (disk heartbeat).

HACMP, подобно AIX, поддерживает понятие логических сетей. Два или более сетевых интерфейса в одной физической сети могут быть сгруппированы, представляя логическую сеть. Эти логические сети имеют уникальное имя (например, net_ether_01, если оно присваивалось HACMP) и могут включать одну или несколько подсетей. Логическая сеть может быть представлена как группа интерфейсов, используемых HACMP для содержания одной или нескольких сервисных IP-меток/адресов. RSCT образует собственные сети, соединяющие интерфейсы в одной подсети, и при необходимости может обеспечивать временную маршрутизацию между подсетями.

Определения сетей можно добавить через экраны SMIT HACMP, однако мы рекомендуем выполнить процесс обнаружения (discovery) перед конфигурированием сетей. Выполнение процесса обнаружения заполнит ниспадающие списки, которые могут использоваться в процессе конфигурирования. В процессе обнаружения информация берется из файла **/etc/hosts**, определенных интерфейсов, определенных адаптеров, устройств target mode и существующих дисков с расширенным режимом одновременного доступа (enhanced concurrent) и создаются следующие файлы:

- clip_config. Содержит сведения об обнаруженных интерфейсах, используемых в SMIT-списках <F4>.
- clvg_config. Содержит сведения о каждом физическом томе (PVID, имя VG, состояние, старший номер устройства и т. д.), а также список свободных старших номеров устройств (major numbers).

В процессе обнаружения также могут быть выявлены некоторые несогласованности в сети вашего сайта.

Глобальная сеть

Глобальная сеть представляет собрание нескольких сетей HACMP одного типа, например Ethernet. Как обсуждалось выше, логические сети HACMP могут состоять из любого сочетания физически различающихся сетей и/или различных подсетей.

Для HACMP важно знать, где именно в сети произошел отказ и не произошел ли отказ глобальной сети, так как при отказе глобальной сети перемещение группы ресурсов на другой узел ничего не даст.

2.3.1 Мониторинг пульса RSCT и HACMP

Диспетчер кластера HACMP cluster manager использует несколько источников для получения информации о возможных отказах:

- RSCT осуществляет мониторинг состояния сетевых интерфейсов и устройств;
- AIX LVM осуществляет мониторинг состояния дисков, логических томов и групп томов.
- мониторы приложений (HACMP Application monitors) осуществляют мониторинг состояния приложений.

HACMP, подобно многим другим типам кластеров, использует пакеты пульса (keep alive, KA) для мониторинга доступности сетевых интерфейсов, коммуникационных устройств и IP-меток (сервисных, несервисных и постоянных). HACMP может использовать как IP-сети, так и сети, отличные от IP, для обмена пакетами или сообщениями пульса между узлами. Посредством мониторинга пульса HACMP получает информацию о состоянии интерфейсов, устройств и адаптеров и, таким образом, о доступности узлов кластера.

Начиная с HACMP V5.1 мониторинг пульса основан исключительно на службах топологии RSCT. До этого HACMP Classic (до HACMP V4.5) использовал собственный код для модулей сетевых интерфейсов (Network Interface Modules, NIM). Демоны RSCT для передачи пакетов пульса используют протокол UDP. При запуске HACMP на узле HACMP передает сетевую топологию, сохраненную в конфигурации HACMP ODM в RSCT. RSCT использует эту информацию для построения своих групп связи («колец пульса», heartbeat rings) и, в свою очередь, передает в HACMP уведомления об отказах.

Технология RSCT была разработана в начале 1990-х гг. для систем IBM SP и затем стала инфраструктурой для HACMP/ES (Enhanced Scalability).

RSCT содержит следующие компоненты:

- **Службы мониторинга и управления ресурсами (Resouse monitoring and control, RMC).** HACMP 5.1 и более ранние версии использовали подсистему управления событиями. RMC представляет собой распределенную подсистему, обеспечивающую набор служб высокой доступности. RMC создает события, сопоставляя текущее состояние системных ресурсов с информацией о требуемом клиентами состоянии ресурсов. После этого клиенты могут использовать уведомления о событиях для вызова действий восстановления. Тем не менее диспетчер событий (event manager) все еще используется для поддержки Oracle RAC.
- **Диспетчеры ресурсов (Resourse managers).** Представляют собой демоны, которые в действительности являются частью RMC и представляют административную задачу или системную функцию. HACMP использует RMC для динамической установки приоритетов узлов, мониторинга приложений и пользовательских событий. Например, монитор ресурсов, сообщающий показатель бездействия процессора, используется при перемещении ресурсов на узел с наиболее высоким показателем бездействия процессора.

- **Службы групп (Group services).** Обеспечивают системное средство мониторинга и согласования изменений в состоянии приложения, выполняющегося на наборе узлов, с высокой доступностью.
- **Службы топологии (Topology services).** Обрабатывают мониторинг пульса через несколько сетей в кластере. Имеют сведения о конфигурации сети и сообщают информацию о состоянии сетевых интерфейсов и адаптеров, а также самих узлов.

На рис. 2.3 показаны некоторые из демонов RSCT, а также их взаимодействие с другими демонами HACMP (в HACMP V5.3).

Мониторинг пульса осуществляется путем обмена сообщениями (пакетами пульса) между узлами через каждый интерфейс и коммуникационное устройство, определенные в кластере (топологии). Каждый узел отправляет пакет пульса и ожидает получить пакет через каждую сеть с интервалом, определенным чувствительностью сети. Так как каждый узел в каждой кольцевой сети связан только с двумя соседними узлами, узел будет получать только один пакет с определенного узла каждые два интервала пульса. Это важно при вычислении того, сколько времени занимает определение отказа.

RSCT осуществляет мониторинг только для базовых адресов интерфейсов (если только не выбран мониторинг пульса через IP-синонимы) и не осуществляет мониторинг сервисных IP-меток (при использовании перехвата IP-адреса IPAT посредством синонимов) или постоянных IP-меток.

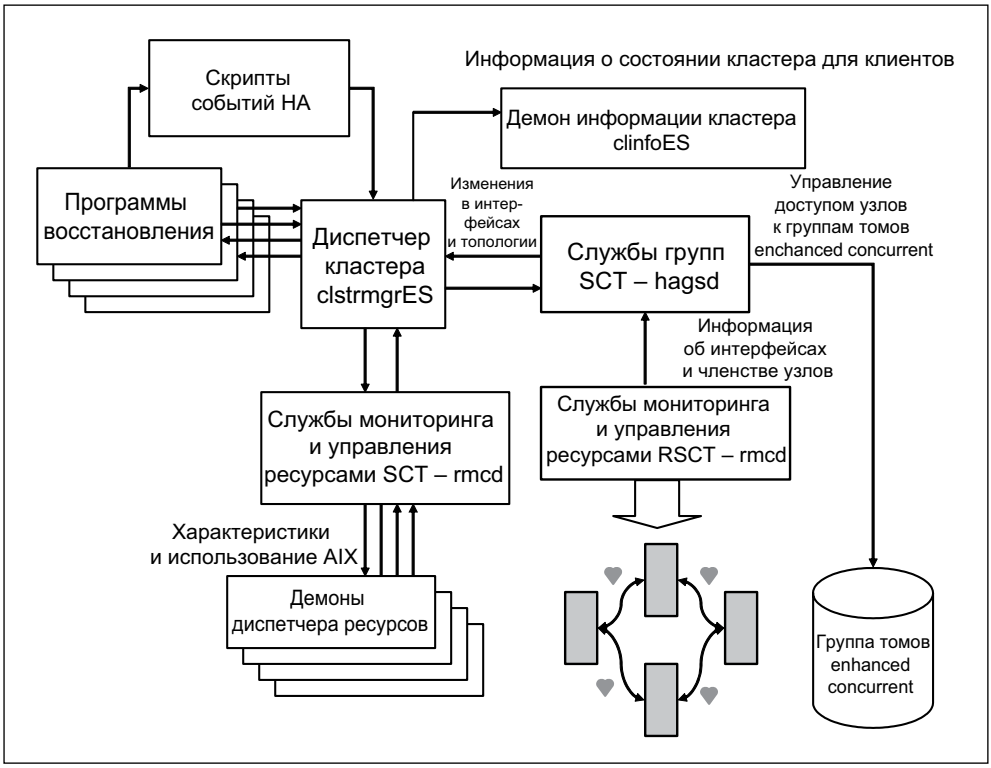

Рис. 2.3. RSCT и важные демоны кластера

HACMP отвечает за отслеживание синонимов меток (сервисных IP-меток при IPAT посредством синонимов и постоянных синонимов меток) как по состоянию базового интерфейса и состоянию связи, так и посредством мониторинга счетчика полученных пакетов (подобно выходным данным команды **netstat**). HACMP V5.2 и более поздние версии пытаются восстановить работоспособность интерфейса, если он находится в состоянии «down» или «detached» (по выходным данным команды **lsattr**), но при этом физическое соединение остается активным.

> **Примечание.** Таким образом, если вы решите командой ifconfig отключить адаптер для тестирования, HACMP вновь его подключит без какой-либо обработки событий HACMP.

RSCT определяет, что на одном из интерфейсов или адаптеров узла произошел отказ, если от него не приходят пакеты пульса, однако продолжает приходить информация пульса через другие интерфейсы и адаптеры узла. В этом случае HACMP сохраняет связь с узлом, переводя сервисные (и постоянные) IP-метки на другой сетевой интерфейс той же сети на том же узле.

Если все интерфейсы этой сети HACMP на этом узле становятся недоступными, HACMP переносит все группы ресурсов, содержащие IP-метки, на другой узел с доступными интерфейсами в этой же сети. Если RSCT не получает пакеты пульса через все интерфейсы или адаптеры узла, то считается, что на этом узле произошел отказ и HACMP попытается привести затронутые этим сбоем группы ресурсов в рабочее состояние на другом узле.

Коммуникации RSCT

HACMP отвечает за запуск RSCT (служб топологии и групп) на узлах, входящих в кластер. RSCT организует свои сети и межузловые связи в зависимости от типа сети следующим образом:

- IP-сети. Кольцо (группа связи RSCT) создается для каждой логической подсети для интерфейсов, определенных в HACMP, в порядке IP-адресов. Каждый узел связан с двумя соседними узлами – узлами с большим и меньшим по очередности IP-адресом. Каждое IP-кольцо (также называемое группой связи RSCT – communication group) изменяется при добавлении каждого нового узла в кластер.
- Последовательные сети. RSCT также создает группу связи для каждой пары коммуникационных устройств или последовательной сети HACMP. Затем RSCT строит логическую сеть между этими группами связи для передачи информации между узлами с использованием коммуникаций, отличных от IP.

Учитывая топологию кластера, представленную на рис. 2.2, RSCT создает три сети пульса – по одной для каждой IP-подсети, и одну для кольца устройств, отличного от IP, как показано на рис. 2.4.

В более ранних версиях HACMP количество последовательных (отличных от IP) коммуникационных устройств одного типа на узел ограничивалось двумя, так что для трех или более узлов была возможна только кольцевая конфигурация. Более новые версии HACMP (5.1 и выше) поддерживают конфигурацию с сетями, отличными от IP, с подключением между всеми узлами (каждого с каждым), если на каждом узле достаточно устройств.

RSCT использует определенные узлы для управления связью между группами связи. Узлы, выполняющие эти задачи, выбираются динамически и могут изменяться при каждом входе или выходе узла из кластера.

- **Лидер группы (Group Leader).** Узел с наивысшим IP-адресом в первой созданной группе связи, называется лидером группы. Этот узел хранит информацию о других узлах в кластере и о конфигурации сети.

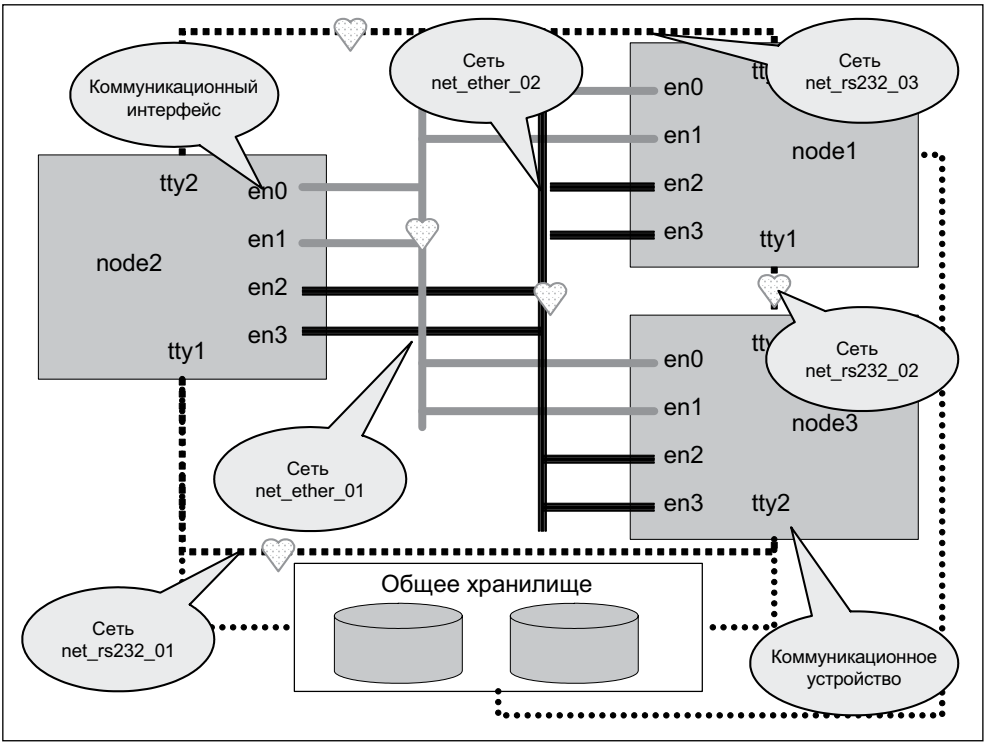

Рис. 2.4. Наложение сетей HACMP на топологию

- **Резервный лидер группы (Group Leader Backup).** Узел со вторым по величине IP-адресом в первой группе связи называется резервным лидером группы. Этот узел содержит резервную копию данных топологии лидера группы и принимает на себя роль лидера группы в случае выхода лидера группы из кластера.
- **Распорядитель (Mayor).** Узел с третьим по величине IP-адресом в первой группе связи (если третий узел недоступен, используется резервный лидер группы). Этот узел отвечает за информирование всех узлов о любых изменениях в топологии кластера.

На рис. 2.5 показан пример двух групп связи, сформированных для двух IP-подсетей, и RSCT-роли узлов.

Как уже обсуждалось, получение информации о состоянии топологии кластера в HACMP в значительной мере зависит от RSCT и от данных, получаемых в пакетах пульса. Тем не менее должна быть полная уверенность в том, что на узле действительно произошел отказ, прежде чем предпринимать какие-либо действия. Если

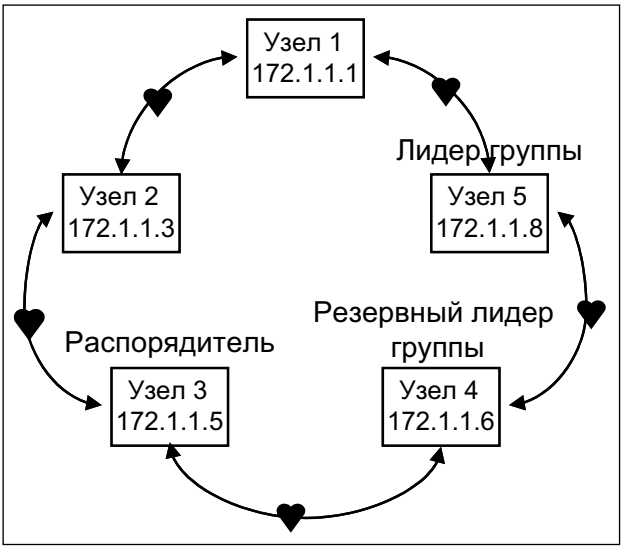

Рис. 2.5. Пример, иллюстрирующий группы связи RSCT и роли узлов

в сети отсутствует избыточность, HACMP может прийти к неправильным выводам о состоянии узлов. Например, если RSCT использует только сеть TCP/IP, отказ сетевого компонента (коммутатора, маршрутизатора, концентратора) или подсистемы TCP/IP будет некорректно интерпретироваться как отказ одного или нескольких узлов. На рис. 2.6 представлен пример кластера, получающего данные пульса исключительно через TCP/IP.

В этом примере узлы *1* и *2* будут считать, что на узлах *3* и *4* произошел отказ, и будут пытаться восстановить работоспособность их ресурсов. Подобным же образом

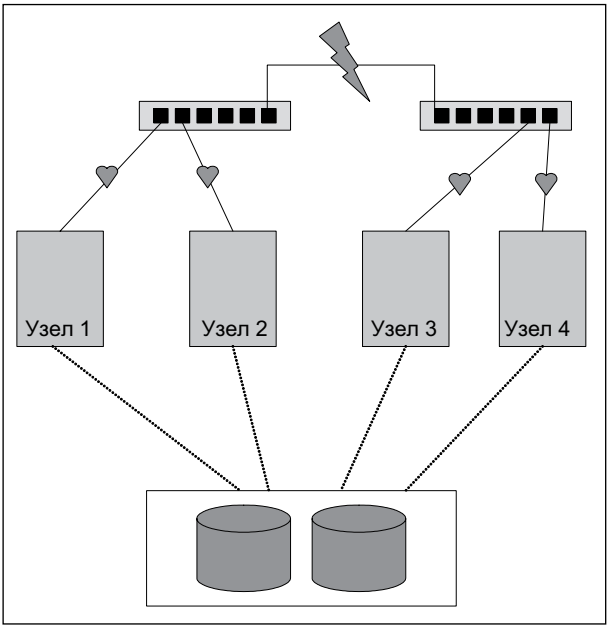

Рис. 2.6. Разделение кластера, вызванное отказом сетевого компонента

узлы *3* и *4* будут считать, что отказ произошел на узлах *1* и *2*. Такая ситуация, называемая *разделенным кластером (partitioned cluster),* может привести к повреждению данных, так как каждая группа узлов будет пытаться одновременно осуществлять доступ к данным и запускать приложения.

Чтобы HACMP мог отличить действительный отказ узла от отказа подсистемы TCP/IP, требуется использовать другой путь связи между узлами – такой, который бы не был основан на TCP/IP. Для этого HACMP использует последовательные сети, отличные от IP (сети «точка-точка» или сети устройств). RSCT осуществляет мониторинг как для сетей устройств, так и для сетей TCP/IP, поэтому HACMP может использовать эту информацию, чтобы отличать отказ узла от отказа IP-сети/подсистемы. Рекомендуется, чтобы каждый кластер имел как минимум одну сеть, отличную от IP, для каждого узла в кластере, чтобы не допустить разделения кластера. Если бы в примере, представленном на рис. 2.6, использовались последовательные сети, HACMP мог бы правильно распознать отказ и не возникла бы опасность потери или повреждения данных (рис. 2.7).

На рис. 2.7 представлена рекомендованная конфигурация кластера из двух узлов.

Другая ситуация, при которой RSCT и HACMP не смогут точно определить состояние интерфейса, заключается в возникновении сети с одним интерфейсом. Если по причине какого-либо отказа интерфейс обнаружит, что он является единствен-

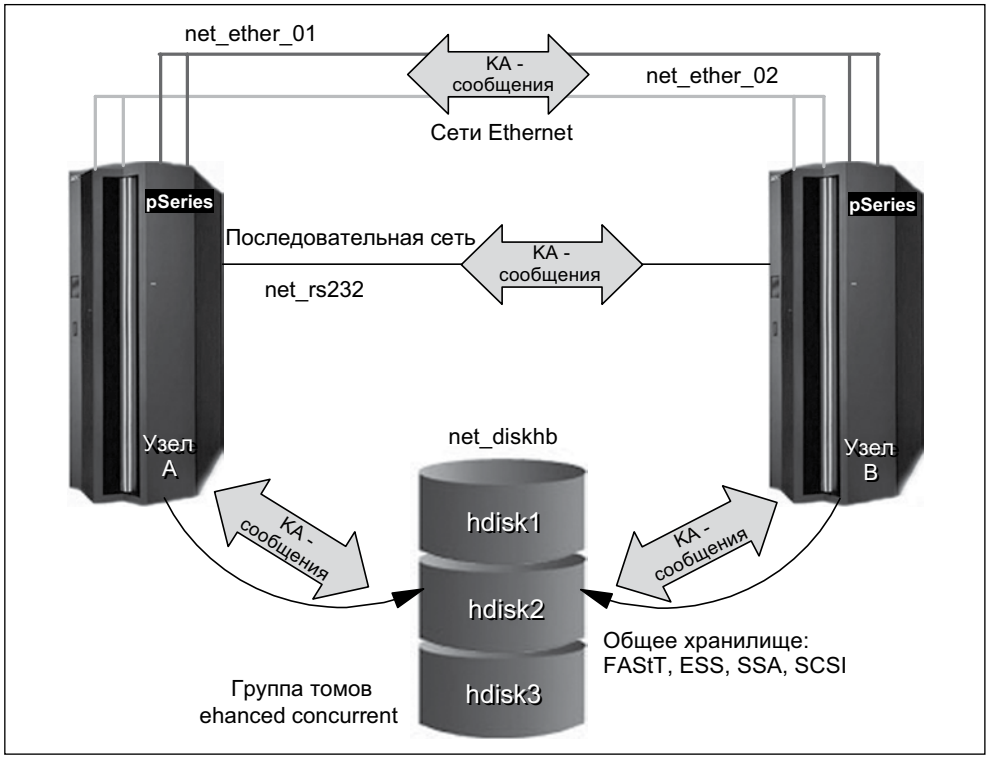

Рис. 2.7. Мониторинг пульса в кластере HACMP

ным интерфейсом в сети, это означает, что не существует другого адреса, с которым интерфейс мог бы обмениваться пакетами пульса. В случае возникновения сети с одним интерфейсом RSCT может определить, работает ли интерфейс, путем мониторинга счетчика полученных пакетов следующим образом:

1. Путем широковещательного ping-опроса.
2. Путем отправки пакетов ICMP ECHO (ping) на все адреса, указанные в файле netmon.cf.
3. Путем создания временного маршрута от одной из групп связи к другой с последующим тестированием связи.

Файл /usr/es/sbin/cluster/etc/netmon.cf

Должен а) содержать список IP-адресов или меток, преобразуемых в адреса, по одной в строке и б) существовать (и быть одинаковым) на всех узлах, входящих в кластер.

Подсети и RSCT

AIX 5L поддерживает использование нескольких маршрутов к одному пункту назначения в таблице маршрутизации ядра. Это означает, что, если несколько совпадающих маршрутов имеют одни и те же критерии, маршрутизация может осуществляться поочередно с использованием каждого из маршрутов подсетей. Это также называется *чередованием маршрутов*.

Таким образом, действие нескольких интерфейсов в одной подсети на одном узле состоит в том, что пакеты будут отправляться через каждый из интерфейсов поочередно. Это означает, что другие узлы, а значит и RSCT, не смогут определить, с какого интерфейса пришел пакет пульса. Чтобы избежать ситуации, при которой RSCT, а значит и HACMP, не смогут точно определить состояние интерфейсов, существуют строгие правила конфигурирования подсетей. Эти правила зависят от конфигурации сети и обсуждаются в разделах, посвященных перехвату IP-адреса.

Примечание. В AIX 5.3 реализована новая опция, mpr_policy, позволяющая настроить TCP/IP таким образом, чтобы пакеты для определенного пункта назначения поступали только с одного адаптера. Чтобы настроить TCP/IP на применение адаптера в зависимости от пункта назначения пакета, следует использовать значение mpr_policy = 5. Мы рекомендуем устанавливать это значение в том случае, если какие-либо из приложений восприимчивы к тому, с какого адаптера приходит пакет, например для NFS.

2.3.2 Мониторинг пульса через IP-синонимы

HACMP теперь поддерживает мониторинг пульса через IP-синонимы. Такая конфигурация устраняет ограничения подсетей на мониторинг базовых интерфейсов, обсуждавшиеся в предыдущем разделе. Теперь существует возможность сконфигурировать базовые IP-адреса без каких-либо ограничений подсетей, что позволяет HACMP и RSCT сконфигурировать и использовать набор отдельных подсетей для мониторинга пульса.

Эти подсети не обязательно должны быть маршрутизируемыми и позволяют сконфигурировать IP-адреса скорее в соответствии с требованиями сайта, чем в со-

ответствии с требованиями HACMP. Например, в случаях, когда сетевой администратор требует, чтобы базовые IP-адреса для каждого адаптера относились к одной подсети. Без мониторинга пульса через IP-синонимы HACMP не поддерживал бы эту конфигурацию, так как подсистема RSCT не смогла бы осуществлять наблюдение за состоянием каждого адаптера.

Тем не менее все же рекомендуется, чтобы сервисные IP-адреса относились к другой подсети по отношению к базовым IP-адресам интерфейсов, чтобы HACMP мог осуществлять точный мониторинг сервисных IP-адресов (если только вы не воспользуетесь опцией mpr_policy в AIX 5.3).

Для конфигурирования мониторинга пульса через IP-синонимы необходимо задать в конфигурации HACMP базовый (начальный) адрес синонима для мониторинга пульса (heartbeat alias address). При запуске HACMP выполняется построение сети пульса через синонимы (alias heartbeat network) начиная с этого адреса путем вычисления IP-адреса для каждого узла на основании номера узла. Эта сеть определяется как отдельная сеть HACMP с количеством подсетей, соответствующим количеству интерфейсов узла. При указании базового адреса синонима мониторинга пульса применяются следующие правила:

- HACMP поддерживает только использование маски подсети базового адаптера для сети пульса через IP-синонимы.
- Маска подсети базового адаптера должна быть больше, чем количество узлов, так как каждый узел будет иметь адрес в этой подсети
- Должно быть достаточно адресного пространства над заданным базовым адресом, чтобы можно было обеспечить по одной подсети для каждого интерфейса узла.
- Сайт не должен содержать адреса в диапазоне синонимов, создаваемых HACMP. Эти адреса не должны также входить в диапазон DNS и т. д.
- HACMP все же требует, чтобы каждый интерфейс мог связываться со всеми другими интерфейсами т. е. чтобы они находились в одной физической сети.

После конфигурирования мониторинга пульса через IP-синонимы HACMP создает требуемые адреса синонимов в соответствии с вышеперечисленными правилами и загружает эту информацию в HACMP ODM. Когда HACMP осуществляет подключение узла в кластер и запускается RSCT, происходит добавление адресов синонимов для каждого адаптера под управлением HACMP. Затем RSCT использует эти адреса для построения своих групп связи. RSCT осуществляет мониторинг именно для этих IP-адресов синонимов, а не для базовых IP-адресов интерфейса.

На рис. 2.8 представлен пример кластера из трех узлов, где каждый узел имеет три интерфейса в одной физической сети и в одной подсети (табл. 2.1). Базовые адаптеры имеют маску подсети 255.255.255.0.

Таблица 2.1. Базовые IP-адреса

	Узел 1	Узел 2	Узел 3
en0	135.2.5.12	135.2.5.22	135.2.5.27
en1	135.2.5.13	135.2.5.23	135.2.5.28
en2	135.2.5.14	135.2.5.24	135.2.5.29

В этом примере HACMP создает три подсети IP-синонимов (по одной для каждого интерфейса) с тремя адресами для каждой (по одному для каждого узла). В табл. 2.2 указаны IP-адреса, используемые HACMP для мониторинга пульса через IP-синонимы, если был сконфигурирован базовый адрес 198.10.1.1.

Таблица 2.2. Конфигурирование мониторинга пульса через IP-адреса синонимов в HACMP (базовый адрес – 198.10.1.1)

	Узел 1	Узел 2	Узел 3
en0 - группа связи 1	198.10.1.2	198.10.1.3	198.10.1.4
en1 - группа связи 2	198.10.2.2	198.10.2.3	198.10.2.4
en2 - группа связи 3	198.10.3.2	198.10.3.3	198.10.3.4

Внимание. Если выбран базовый адрес x.x.x.1, HACMP запустится с использованием x.x.x.2 в качестве первого адреса первого интерфейса. Однако если выбрать адрес x.x.x.0, HACMP будет использовать этот адрес и он не будет работоспособен.

Примечание. Ни одна из трех подсетей – 198.10.1/24, 198.10.2/24, 198.10.3/24) – не должна обязательно быть маршрутизируемой.

На рис. 2.8 представлены три кольца мониторинга пульса (группы связи), используемые RSCT.

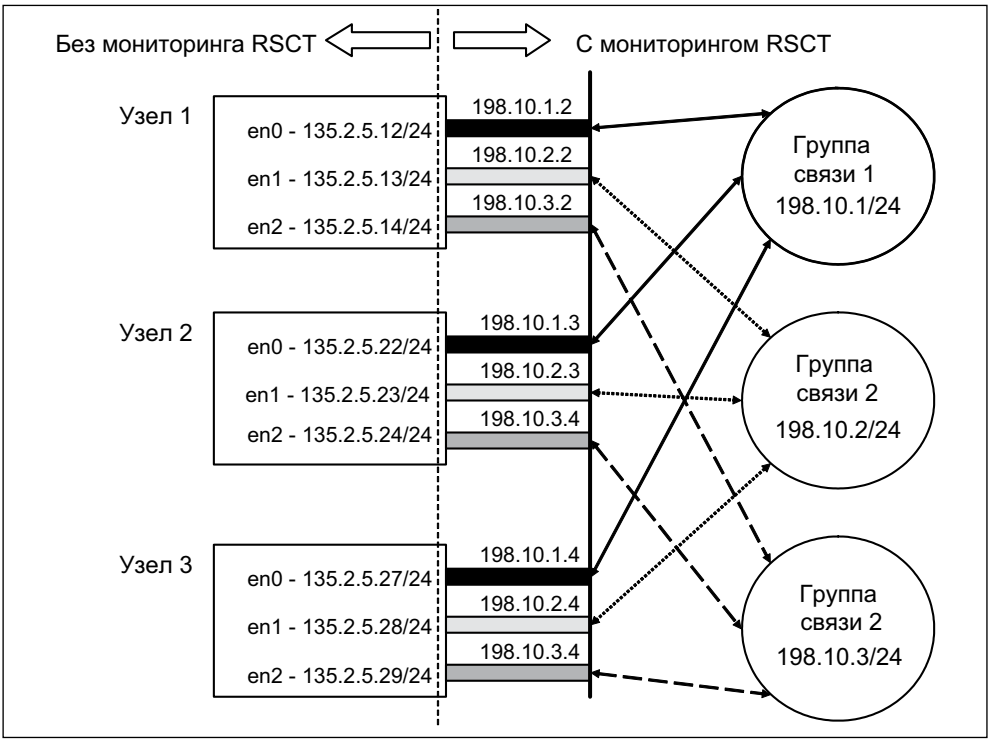

Рис. 2.8. Кластер из трех узлов с пульсом через IP-синонимы

Мониторинг пульса через IP-синонимы поддерживает оба механизма перехвата IP-адресов:

- **Перехват IP-адреса посредством замены (IPAT via replacement).** Сервисная IP-метка заменяет загрузочный IP-адрес интерфейса. IP-адрес синонима мониторинга пульса остается без изменений.
- **Перехват IP-адреса посредством синонимов (IPAT via aliasing).** Сервисная IP-метка добавляется в качестве синонима к интерфейсу с IP-синонимом мониторинга пульса.

2.3.3 Сети TCP/IP

В HACMP 5.1 / 5.2 / 5.3 поддерживаются следующие типы IP-сетей:

- Ethernet (ether);
- Token-ring (token);
- Fiber Distributed Data Interface - FDDI (fddi);
- SP Switch1 и SP Switch2 (hps);
- Asynchronous transfer mode – ATM и эмуляция ATM (atm);
- Etherchannel (ether).

Не поддерживаются следующие типы IP-сетей (начиная с HACMP 5.1):

- Virtual IP Address (VIPA);
- Serial Optical Channel Converter (SOCC);
- Serial Line IP (SLIP);
- Fibre Channel Switch (FCS);
- IEEE 802.3;
- IP Version 6 (IPV6).

> **Примечание.** HACMP поддерживает использование интерфейсов связи IP через агрегированный Ethernet (Etherchannel, IEEE 802.3ad) для перехвата IP-адреса в AIX 5L. Не поддерживается использование Etherchannel:
> - для перехвата аппаратного адреса (hardware address takeover);
> - «горячего подключения» PCI.

HACMP предназначен для работы с любой сетью TCP/IP; эти сети используются для того, чтобы:

- позволять клиентам осуществлять доступ к узлам (т. е. к приложениям);
- позволять узлам обмениваться сообщениями пульса;
- преобразовывать доступ к данным в последовательную форму (в средах с одновременным доступом, например Oracle Real Application Cluster).

Сети TCP/IP можно разделить:

- **На открытые.** Логические сети, предназначенные для связи клиента с узлами. Каждая такая сеть состоит из набора IP-адаптеров, так что каждая сеть может со-

держать несколько подсетей. Так как эти сети предназначены для доступа клиентов, в них поддерживается перехват IP-адреса.

- **Закрытые.** Такие сети предназначены для использования приложениями, не поддерживающими перехват IP-адреса, например Oracle RAC или географическими сетями HAGEO. Все интерфейсы определяются как сервисные, и пакеты пульса пересылаются через эти сети.

2.3.4 Механизмы перехвата IP-адреса

Одна из основных ролей HACMP состоит в том, чтобы обеспечивать высокую доступность сервисных IP-меток/адресов. В HACMP это достигается путем запуска и остановки каждого сервисного IP-адреса при необходимости на подходящем интерфейсе. При активной группе ресурсов на узле HACMP поддерживает два метода активизации сервисных IP-адресов:

- Путем замены базового IP-адреса (установленного во время загрузки ОС) интерфейса на сервисный IP-адрес. Этот метод называется перехватом IP-адреса посредством замены IP (IP address takeover, IPAT via IP replacement). Этот метод также позволяет осуществлять перехват локально администрируемого аппаратного адреса (locally administered hardware address, LAA) – перехват аппаратного адреса.
- Путем добавления сервисного IP-адреса в качестве синонима интерфейса, т. е. в дополнение к базовому IP-адресу. Этот метод называется *перехватом IP-адреса посредством IP-синонимов (IPAT via IP aliasing)*. Он используется по умолчанию в HACMP 5.1 и выше.

Настройки по умолчанию можно изменить путем изменения свойств сети через расширенные меню конфигурирования HACMP.

Следует сказать и о том, что каждый метод предполагает ограничения подсетей для загрузочных интерфейсов и сервисных IP-меток, если только не используется мониторинг пульса через IP-синонимы.

Перехват IP-адреса посредством замены

Сервисная IP-метка/адрес заменяет существующий адрес интерфейса. Таким образом, только одна сервисная IP-метка/адрес может быть сконфигурирована для одного интерфейса. Сервисная IP-метка должна находиться в той же подсети, в которой находится и один из базовых IP-адресов. Этот интерфейс будет в первую очередь использоваться сервисной IP-меткой при активизации группы ресурсов на узле. Другие интерфейсы этого узла, которые не могут находиться в той же подсети, традиционно называются *дежурными* или *резервными (standby) интерфейсами* и используются при перемещении группы ресурсов с другого узла или при отказе загрузочного интерфейса (рис. 2.9). Этот метод может сохранить подсети, однако требует использования дополнительного оборудования.

При перехвате IP-адреса посредством замены (также называемом *классическим перехватом IP-адреса*) также можно выполнить конфигурирование перехвата аппаратного адреса (hardware address takeover, HWAT). При перехвате аппаратного адреса локально администрируемый MAC-адрес становится частью определения сервисной

IP-метки и во время замены MAC-адрес интерфейса также будет изменен. Это позволяет обойтись без обновления ARP-кешей в подсети, и при этом приложения, использующие MAC-адреса, все равно будут указывать на правильный узел.

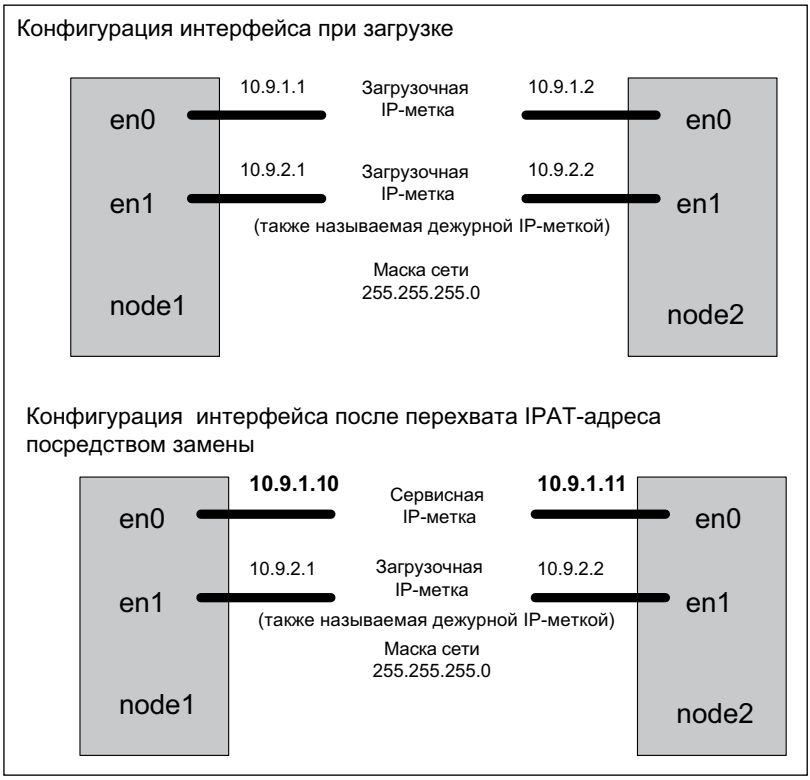

Рис. 2.9. Перехват IP-адреса посредством замены

В случае отказа интерфейса, содержащего сервисный IP-адрес, при перехвате посредством замены HACMP перемещает сервисный IP-адрес на другой доступный интерфейс того же узла и той же сети; в этом случае соответствующая группа ресурсов не затрагивается.

При отсутствии доступного интерфейса на том же узле группа ресурсов вместе с сервисными IP-метками перемещается на другой узел с доступным интерфейсом в той же логической сети. При отсутствии доступных узлов или интерфейсов группа ресурсов переходит в состояние ERROR (Ошибка). Когда HACMP обнаруживает адаптеры, выполняется проверка возможности перевода групп ресурсов, находящихся в состоянии ERROR, обратно в рабочее состояние.

Ограничение. При перехвате IP-адреса посредством замены RSCT и HACMP не смогут корректно осуществлять мониторинг узла, имеющего больше одной группы ресурсов, если обе сервисные IP-метки находятся в одной подсети, так как узел в этом случае будет иметь два интерфейса с базовыми адресами в одной подсети. По этой причине рекомендуется использовать мониторинг пульса через IP-синонимы или установить в AIX 5.3 для опции mpr_policy значение 5.

Перехват IP-адреса посредством синонимов

Для сервисной IP-метки/адреса выполняется создание синонима на интерфейсе без удаления базового загрузочного IP-адреса с использованием команды **ifconfig** (рис. 2.10). Начиная с HACMP 5.1 этот метод применяется по умолчанию. Кроме того, перехват IP-адреса посредством синонимов отменяет понятие дежурных интерфейсов – все сетевые интерфейсы считаются загрузочными.

При добавлении IP-адресов к интерфейсу посредством синонимов на одном интерфейсе может сосуществовать больше одной сервисной IP-метки. Перехват IP-адреса посредством синонимов устраняет необходимость использования одного интерфейса на сервисный IP-адрес, поэтому этот метод является более гибким и в некоторых случаях требует меньше оборудования. Перехват IP-адреса посредством синонимов также сокращает время на перемещение при сбое, так как добавление синонима к интерфейсу требует гораздо меньше времени, чем удаление базового IP-адреса с последующим применением сервисного IP-адреса.

Несмотря на то что перехват IP-адреса посредством синонимов поддерживает несколько сервисных IP-меток/адресов, все равно рекомендуется выполнять конфигурирование нескольких интерфейсов на узел на сеть. Переключение интерфейсов гораздо безопаснее, чем перемещение группы ресурсов на другой узел.

Перехват IP-адреса посредством синонимов осуществляется только в сетях, поддерживающих функцию gratuitous ARP. Эта функция предполагает, что узел отправляет ARP-пакет до использования IP-адреса; при этом ARP-пакет содержит запрос на полу-

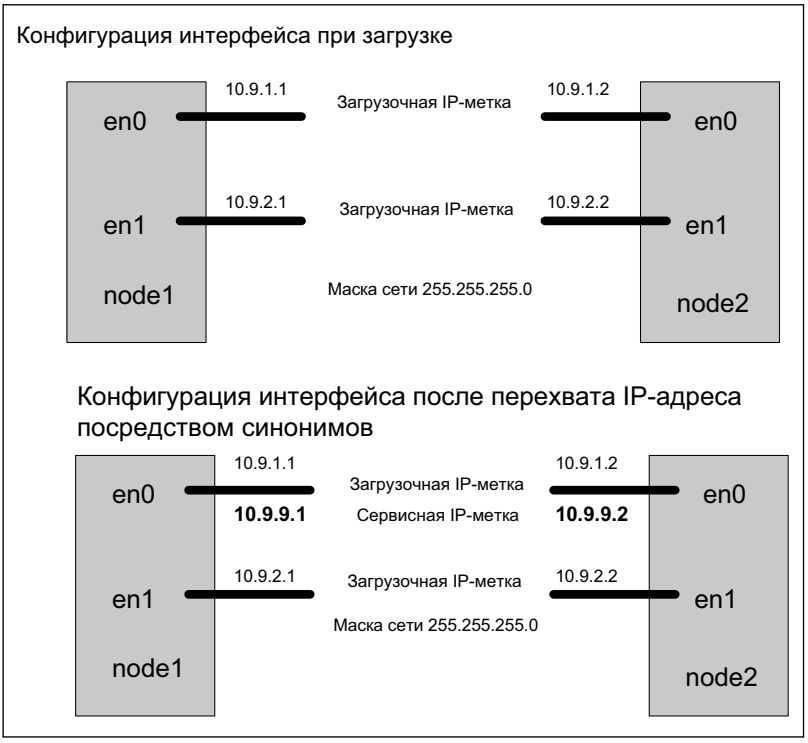

Рис. 2.10. Перехват IP-адреса посредством IP-синонимов

чение IP-адреса. Это позволяет удостовериться в том, что этот адрес не используется ни одним другим узлом, а также обеспечить обновление ARP-кеша с записью нового адреса на каждом компьютере в подсети.

При использовании нескольких активных сервисных IP-меток/адресов на одном узле HACMP по умолчанию равномерно распределяет их среди доступных интерфейсов в логической сети. Однако HACMP 5.3 позволяет управлять этим размещением. HACMP распределяет синонимы на узле, сортируя доступные интерфейсы по количеству уже имеющихся на них синонимов, и размещает новые синонимы соответствующим образом. Это выполняется только при интеграции и перемещении при сбое, поэтому, когда становится доступным новый интерфейс, перераспределение не выполняется.

При перехвате IP-адреса посредством синонимов каждый загрузочный интерфейс на узле должен находиться в другой подсети, хотя интерфейсы с различных узлов могут, конечно же, находиться в одной подсети, если только (как говорилось выше) не используется мониторинг пульса через IP-синонимы. Сервисные IP-метки могут находиться в одной или нескольких подсетях, однако они не могут быть одинаковыми, как и любая из подсетей загрузочных интерфейсов.

Важно! В сетях с перехватом IP-адресов посредством синонимов в HACMP какое-то время будут одновременно активными сервисные IP-адреса отказавшего интерфейса и интерфейса перехвата, что позволяет сохранить маршрутизацию. Это может вызвать запись о дублировании IP-адресов (DUPLICATE IP ADDRESS) в журнале ошибок error log, которую можно игнорировать.

2.3.5 Постоянная IP-метка/адрес

Постоянная IP-метка узла (Persistent node IP label) представляет собой IP-синоним, который может быть назначен определенному узлу сети и, кроме того:

- всегда находится на одном узле (привязана к узлу);
- сосуществует с другими IP-метками, существующими на том же интерфейсе;
- не требует установки дополнительного физического интерфейса на узле;
- не входит в какую-либо группу ресурсов.

Назначение постоянной IP-метки узлу сети обеспечивает привязанный к узлу адрес в сети кластера с высокой доступностью. Этот адрес может использоваться в целях администрирования, так как он всегда указывает на определенный узел, вне зависимости от того, работает ли HACMP.

Примечание. Можно сконфигурировать только одну постоянную IP-метку для узла одной сети. Например, если существует узел, подключенный к двум сетям, определенным в HACMP, этот узел может быть идентифицирован с использованием двух постоянных IP-меток (адресов), по одному для каждой сети.

Постоянные IP-метки определяются в конфигурации HACMP и становятся доступными сразу же после синхронизации определения кластера. Постоянная IP-метка остается доступной на интерфейсе, для которого она была сконфигурирована, даже

при остановке HACMP на узле или при перезагрузке узла. В случае отказа интерфейса, которому была назначена IP-метка во время работы HACMP, постоянная IP-метка будет перемещена на другой интерфейс в той же логической сети и на том же узле.

В случае отказа узла или всех интерфейсов логической сети на узле постоянная IP-метка будет недоступна.

На постоянные IP-метки распространяются следующие ограничения по формированию подсетей:

- **для сетей с перехватом IP-адреса посредством замены.** Постоянный IP-синоним не должен располагаться в одной подсети с дежурными интерфейсами и может располагаться в одной подсети с загрузочными интерфейсами (той же, что сервисные IP-метки);
- **для сетей с перехватом IP-адреса посредством синонимов.** Постоянный IP-синоним не должен располагаться в одной подсети с загрузочными интерфейсами.

Постоянные IP-метки узлов могут создаваться для следующих типов IP-сетей:

- Ethernet;
- Token Ring;
- FDDI;
- ATM LAN Emulator.

Ограничение. Невозможно сконфигурировать постоянные IP-метки узлов в сетях SP Switch, Classical IP Over ATM и сетях, отличных от IP.

2.3.6 Сети устройств, или Последовательные сети

Последовательные сети (serial networks) представляют альтернативный метод обмена информацией с помощью пакетов пульса между узлами кластера. В случае отказа подсистемы IP или физической сети HACMP, при наличии и работоспособности независимого пути, сможет отличить отказ сети от отказа узла.

Последовательные сети представляют собой сети типа «точка-точка», поэтому, если кластер содержит больше двух узлов, последовательные связи должны иметь кольцевую конфигурацию, связывая все узлы в кластере. Несмотря на то что каждый узел будет иметь информацию только о состоянии своих непосредственных соседей, демоны RSCT обеспечивают передачу информации о любых изменениях в состоянии любых узлов лидеру группы.

Несмотря на то что существует возможность конфигурирования кластера HACMP без использования сетей, отличных от IP, мы настоятельно рекомендуем использовать как минимум по одному соединению, отличному от IP, между всеми узлами в кластере.

В настоящее время HACMP поддерживает следующие типы сетей устройств, отличных от TCP/IP, для обмена пакетами пульса между узлами кластера:

- последовательная RS232 (rs232);
- Target mode SCSI (tmscsi);

- Target mode SSA (tmssa);
- мониторинг пульса через диски (diskhb).

Могут использоваться следующие типы последовательных сетей.

RS232

Последовательная сеть, использующая порты RS232, либо встроенные последовательные порты, либо многопортовый последовательный адаптер.

> **Примечание.** Необходимо быть внимательным и выбирать такие порты, которые поддерживают мониторинг пульса.

По умолчанию скорость передачи в сетях RS232 равна 38 400 бит/с; если ваш модем не поддерживает такую скорость, это значение следует изменить (в том случае, если вы хотите использовать эту сеть для связи между удаленными узлами).

Target mode SCSI

Другим вариантом сети, отличной от IP, является SCSI-подключение в target mode. При употреблении общего SCSI-устройства можно использовать шину SCSI для обмена пакетами пульса. Target mode SCSI (tmscsi) поддерживается только устройствами SCSI-2 Differential или SCSI-2 Differential Fast/Wide. SCSI-1 Single-Ended и SCSI-2 Single-Ended не поддерживают последовательные сети в кластере HACMP.

Target mode SSA

При использовании SSA-устройств общего доступа в качестве сети, отличной от IP, в HACMP можно использовать Target mode SSA. Он основан на встроенных возможностях SSA-адаптеров (применяющих протокол связи SCSI). SSA-устройства в кольце SSA (диски и адаптеры) используют связь между «инициатором» и «целевым объектом»; SSA-диски являются «целевыми объектами», тогда как SSA-адаптер может выступать как в роли «инициатора», так и в роли «целевого объекта». Таким образом, tmssa-подключение использует эти возможности для установления последовательного соединения между узлами HACMP. Такая сеть представляет собой коммуникационную сеть типа «точка-точка», в которой связь осуществляется только между двумя узлами.

Чтобы сконфигурировать tmssa-сеть между двумя узлами кластера, SSA-адаптеры на каждом из узлов в одном SSA-кольце составляют конечные точки.

Сеть пульса через диски

В некоторых ситуациях подключения RS232 tmssa и tmscsi могут считаться слишком дорогостоящими или сложными для реализации. В этом случае мониторинг пульса через диски (diskhb) обеспечивает простой в конфигурировании альтернативный вариант, не требующий дополнительного оборудования. Единственное требование состоит в том, чтобы «диски» (физические диски или номера логических устройств во внешнем хранилище) работали в расширенном режиме одновременного доступа (enhanced concurrent). Диски, работающие в расширенном режиме одновременного доступа, используют службы групп RSCT для управления блокировкой и освобождением сектора диска, применяемого для связи. Раньше этот сектор использовался для

дисков режима одновременного доступа SSA (SSA Concurrent mode), теперь он применяется для записи информации пульса.

Любой диск, входящий в группу томов с расширенным одновременным доступом (enhanced concurrent VG), может использоваться в сети diskhb, включая диски, применяемые для хранения данных. Более того, группа томов, содержащая диск, используемый в сети diskhb, не обязательно должна быть активизирована (varied on).

Любой тип диска может быть сконфигурирован как часть группы томов с расширенным одновременным доступом, что делает этот тип сети чрезвычайно гибким. «Адаптеры» на конечных точках этой сети определяются как пара «узел – физический том».

В случае мониторинга пульса через диски рекомендуется использовать одну сеть «точка-точка», содержащую один диск на пару узлов на физический дисковую стойку. Один физический диск не может применяться для двух сетей типа «точка-точка».

Примечание. Группа томов с расширенным одновременным доступом (enhanced concurrent volume group) не то же самое, что группа томов с одновременным доступом (concurrent volume group). Если группа второго типа представляет собой часть группы ресурсов с одновременным доступом (concurrent resourse group), первое понятие скорее указывает на режим блокировки с использованием RSCT.

Мониторинг пульса через диски

Мониторинг пульса через диски (diskhb) представляет собой новую функцию, впервые реализованную в HACMP V5.1 с целью обеспечить дополнительную защиту против разделения кластера и упростить конфигурирование сети, отличной от IP, особенно в средах, где применение соединений RS232, target mode SSA или target mode SCSI является слишком сложным или невозможным.

Этот тип сети может использовать общее дисковое хранилище любого типа (Fibre Channel, SCSI или SSA), если только диск, применяемый для обмена KA-сообщениями, является частью группы томов AIX с расширенным одновременным доступом. Диски, применяемые в сетях пульса, не выделяются исключительно для этой цели; их можно использовать для хранения общих данных приложений (дополнительные сведения см. на рис. 2.7).

Наши клиенты запрашивали поддержку подключений target mode Fibre Channel, однако в связи с разнообразием (нестандартными функциями инициатора и целевого объекта) сред FC (включающими адаптеры, подсистемы хранения, SAN, коммутаторы и концентраторы) такие подключения сложны в реализации и поддержке.

Используя общие диски для обмена сообщениями, внедрение сети, отличной от IP, является более надежным и не зависит от применяемого типа оборудования.

Кроме того, в среде SAN при использовании оптоволоконных соединений между устройствами длина этого соединения, отличного от IP, имеет такие же ограничения, как и в SAN, что позволяет создавать протяженные сети типа «точка-точка».

При определении диска в составе группы томов с расширенным одновременным доступом часть диска не будет использоваться для какой-либо из операций LVM, и эта часть диска (сектор) применяется для обмена сообщениями между двумя узлами.

Ниже перечислены спецификации для использования мониторинга пульса через диски.

- Один диск может применяться для одной сети между двумя узлами. Используемый диск уникальным образом определяется на обоих узлах по идентификатору физического тома (PVID), назначенного LVM.
- Рекомендуемая конфигурация для сетей с мониторингом пульса через диски состоит из одного диска на пару узлов на дисковую стойку.
- Требуется, чтобы используемый диск был частью группы томов с расширенным одновременным доступом, хотя для группы томов не обязательно быть активной или входить в группу ресурсов (с одновременным доступом или без него). Единственное условие состоит в том, что группа томов должна быть определена на обоих узлах.

> **Примечание.** Механизм блокировки кластера для групп томов с расширенным одновременным доступом не использует зарезервированное дисковое пространство для связи (как «классический» clvmd); вместо этого он применяет службы групп RSCT.

2.3.7 Сетевые модули

В HACMP скорость обнаружения отказов определяется для каждого типа сети и может быть либо задана с помощью трех предопределенных значений, либо настроена. Предопределенные значения это «медленное» (slow), «нормальное» (normal) (по умолчанию) и «быстрое» (fast), тогда как при настройке задаются интервал между импульсами и так называемый цикл отказа (failure cycle).

Интервал между импульсами (скорость пульса, heartbeat rate – **hbrate**) определяет скорость отправления службами кластера пакетов «keep alive» между интерфейсами и устройствами в кластере. Цикл отказа (**cycle**) представляет количество идущих подряд импульсов, которое может быть пропущено, прежде чем будет считаться, что произошел отказ интерфейса. HACMP 5.3 поддерживает интервал пульсации в доли секунды.

Время обнаружения отказа (в секундах) определяется по формуле

hbrate * cycle * 2

В HACMP цикл отказа умножается на два и на время обнаружения отказа, что позволяет как узлу, так и соседним узлам сделать корректный вывод о сбое. Кроме того, в целях сокращения сетевого трафика HACMP отправляет только один пакет и ожидает получить один пакет на логическую сеть на интервал пульсации.

При использовании последовательных сетей HACMP объявляет соседний узел неработающим после завершения времени обнаружения отказа (failure detection rate), заданного для этого типа сети. HACMP будет ожидать в течение такого же периода, прежде чем объявить устройство неработающим, если за это время все еще не будет получено ни одного пакета пульса от соседнего узла. HACMP не инициирует событие network_down, пока не откажут как локальное, так и удаленное устройства.

Однако если последовательная сеть является последней сетью, подключенной к определенному узлу, событие node_down будет вызвано, как только будет обнаружен отказ интерфейса.

Различие сетей устройств и IP-сетей состоит в том, что сети устройств не могут отличить отказ интерфейса от отказа сети. Исключение составляют сети с мониторингом пульса через диски. Если узел может получить доступ к диску, интерфейс считается работающим, тогда как сеть считается работающей, если осуществляется обмен сообщениями.

Существует еще одна характеристика, определяемая для каждой сети. Эта характеристика называется *отсрочкой (grace period)* и представляет собой период времени после обнаружения определенного отказа сети, в течение которого отказы такого же типа игнорируются. Это дает кластеру время на внесение изменений в конфигурацию сети без обнаружения ложных отказов.

Внимание! Время на обнаружение отказа должно быть одинаковым во всех сетях, используемых кластером.

2.3.8 Клиенты

Клиентом называется система, которая может получить доступ к узлам кластера через сеть. На клиенте выполняется некоторое клиентское приложение, которое обменивается данными с приложением, выполняющимся в кластере, через сервисные IP-метки. HACMP обеспечивает высокую доступность приложения для клиентов, но сами по себе они не обладают высокой доступностью.

Клиенты AIX 5L могут использовать службы информации кластера (clinfo) для получения сообщений о событиях кластера. Clinfo включает API, отображающий состояние кластера.

Во время перехвата группы ресурсов приложение запускается на другом узле, так что клиенты должны быть уведомлены о предпринимаемом действии. В некоторых случаях клиент приложения использует ARP-кеш на клиентском компьютере для переподключения к серверу. В этом случае возможны два варианта, если клиент находится в одной подсети с узлами кластера:

- если в сети, используемой сервисными IP-метками приложений, настроен перехват IP-адреса посредством замены, тогда одновременно происходит и перехват MAC-адреса, так что не возникает необходимости обновлять ARP-кеш на клиентском компьютере;
- если в сети настроен перехват IP-адреса посредством синонимов, тогда происходит отправка пакета gratuious ARP, так что опять же нет необходимости обновлять кеш ARP на клиентском компьютере.

Но если клиент не поддерживает пакеты gratuious ARP, его кеш можно обновить с использованием /usr/es/sbin/cluster/etc/clinfo.rc. При возникновении изменений в сети clinfo.rc отправляет по одному ping-запросу на каждый адрес или метку в списке переменной PING_CLIENT_LIST. Таким образом, чтобы обеспечить обновление ARP-кеша на клиентах, нужно добавить адреса и метки всех клиентов в переменную PING_CLIENT_LIST в clinfo.rc. Однако если клиент находится в другой подсети, тогда приведенные выше условия относятся к маршрутизатору.

Клиенты, на которых выполняется демон clinfo, смогут быстро переподключиться к кластеру после события в кластере.

2.3.9 Аспекты сетевой безопасности

Безопасность HACMP важна для того, чтобы ограничить как несанкционированный доступ к узлам, так и неавторизованный перехват сообщений между узлами. Более ранние версии HACMP использовали rsh для выполнения команд на других узлах. При этом было сложно обеспечить защиту и существовала опасность подмены IP-адреса. Теперь HACMP употребляет собственный демон коммуникаций кластера (**clcomdES**) для управления связью между узлами.

HACMP обеспечивает безопасность кластера путем:

- управления доступом пользователей к HACMP;
- обеспечения безопасности связи между узлами.

Дополнительные сведения о доступе пользователей и безопасности кластера см. в гл. 15 и 16 руководства **High Availability Cluster Multi-Processing Administration Guide**, SC23-4862-06.

Аутентификация и шифрование подключения

При аутентификации проверяется источник и целостность сообщения, тогда как шифрование гарантирует, что содержание сообщения будет известно только отправителю и получателю.

Существуют следующие способы аутентификации подключения:

- **Стандартная аутентификация.** Используется по умолчанию. Демон коммуникаций (clcomd) осуществляет аутентификацию по IP-адресу и ограничивает команды, которые могут выполняться с привилегиями «root». Существует набор команд HACMP (перечисленных в **/usr/es/sbin/cluster**), которые могут выполняться с привилегиями «root», остальные команды выполняются под учетной записью «nobody».
- **Аутентификация Kerberos.** Аутентификация Kerberos поддерживается в среде SP.
- **Виртуальная частная сеть.** Для связи между узлами может быть сконфигурирована виртуальная частная сеть (VPN); для определения туннелей следует использовать постоянные метки синонимов.

HACMP поддерживает следующие способы шифрования:

- Message Digest 5 (MD5) с Data Encryption Standard (DES);
- MD5 с Triple DES;
- MD5 с Advanced Encryption Standard (AES).

Файлы ключей хранятся в каталоге /usr/es/sbin/cluster/etc.

> **Примечание.** Это шифрование распространяется только на clcomdES, но не на диспетчер кластера.

Демон коммуникаций кластера

С появлением **clcomdES** нет необходимости в конфигурировании файла **/.rhosts.** Однако некоторые приложения все же могут требовать наличия этого файла. Демон комму-

никаций кластера выполняет удаленные команды, основываясь на принципе «наименьших привилегий». Это означает, что нельзя будет выполнить произвольную команду на удаленном узле с привилегиями «root». Только несколько команд HACMP считаются «доверенными» и допускают запуск с привилегиями «root»; эти команды перечислены в /usr/es/sbin/cluster. Остальные команды выполняются под учетной записью «nobody».

Демон коммуникаций кластера запускается средством inittab, где соответствующая запись создается при установке HACMP. Управление демоном осуществляет контроллер системных ресурсов SRC, так что для него доступны команды startsrc, stopsrc и refresh. В частности, refresh используется для повторного чтения /usr/es/sbin/cluster/etc/rhosts и перемещения файлов журнала.

Демон коммуникаций кластера использует порт 6191 и осуществляет аутентификацию входящих подключений, выполняя следующие проверки.

Если файл /usr/es/sbin/cluster/etc/rhosts не существует, все подключения отклоняются.

Если файл существует, то выполняется проверка подключений (по порядку):

- в ODM-классе HACMPnode;
- в ODM-классе HACMPadapter;
- в файле /usr/es/sbin/cluster/etc/rhosts.

Наполнение файла /usr/es/sbin/cluster/etc/rhosts происходит при первой синхронизации с адресом интерфейса с синхронизирующего узла (на синхронизирующем узле он остается пустым). После первой синхронизации происходит наполнение ODM-классов HACMP, так что содержимое файла rhosts можно удалить.

По своему назначению файл используется до первой синхронизации кластера в небезопасной среде. Наполните этот файл на каждом узле только адресами интерфейсов узлов в кластере, и ни одна другая система не сможет связаться через clcomdES.

Замечание. После первоначальной синхронизации кластера происходит наполнение файла /usr/es/sbin/cluster/etc/rhosts адресами интерфейсов синхронизирующего узла.

Запрашивающий узел должен передать свою IP-метку, которая должна соответствовать адресу в приведенном выше расположении, и, если получен правильный ответ, подключение разрешается. Если все вышеописанные записи являются пустыми, демон предполагает, что кластер не был сконфигурирован и примет входящие записи.

Важно! При наличии недопустимой записи в файле /usr/es/sbin/cluster/etc/rhosts демон коммуникаций clcomdES будет отклонять все подключения.

Демон коммуникаций кластера обеспечивает транспортную среду для верификации кластера HACMP, глобальных изменений ODM и удаленного выполнения команд. Демон коммуникаций clcomdES используется следующими командами (применение пользователем не поддерживается):

clrexec выполняет специфические и потенциально опасные команды;

cl_rcp копирует файлы конфигурации AIX;

cl_rsh используется кластером для выполнения команд в удаленной оболочке.

Демон коммуникаций кластера также отличается от традиционных rsh/rcp-подключений более высокой производительностью (r-команды медленны); кроме того, clcomdES оставляет подключения сокетов открытыми, вместо того чтобы закрывать их после каждой операции. Так как многие операции администрирования HACMP требуют доступа к ODM, демон коммуникаций кластера также кеширует копии ODM каждого узла. Когда необходим доступ к ODM, clcomdES сравнивает контрольную сумму кешированных записей ODM с ODM на каждом узле в кластере и осуществляет обновление только в случае необходимости.

Демон коммуникаций кластера также отправляет собственные пакеты пульса на каждый узел и в случае отказа сети пытается повторно установить подключение.

Демон коммуникаций кластера также используется (в HACMP 5.2 и выше):

- для наборов файлов (file collections);
- автоматической синхронизации и верификации кластера;
- администрирования пользователей/паролей;
- C-SPOC.

В отличие от HACMP 5.1, начиная с HACMP 5.2 остановка демона коммуникаций кластера при активном кластере невозможна.

2.4 Ресурсы и группы ресурсов

Этот раздел описывает следующие понятия ресурсов HACMP:

- определения;
- ресурсы;
- группы ресурсов.

2.4.1 Определения

HACMP использует базовую топологию для обеспечения высокой доступности управляемых приложений и требуемых ими ресурсов. К таким ресурсам относятся:

- сервисные IP-метки/адреса;
- физические диски;
- группы томов;
- логические тома;
- файловые системы;
- сетевые файловые системы;
- серверы приложений (приложения);
- адаптеры и каналы связи;
- накопители на магнитной ленте;
- ресурсы Fast Connect;
- интеграция с WLM.

Конфигурация приложений и требуемых ресурсов происходит в группах ресурсов. Группы ресурсов управляются HACMP как единые объекты, работу которых можно отрегулировать в соответствии с требованиями клиентов/пользователей.

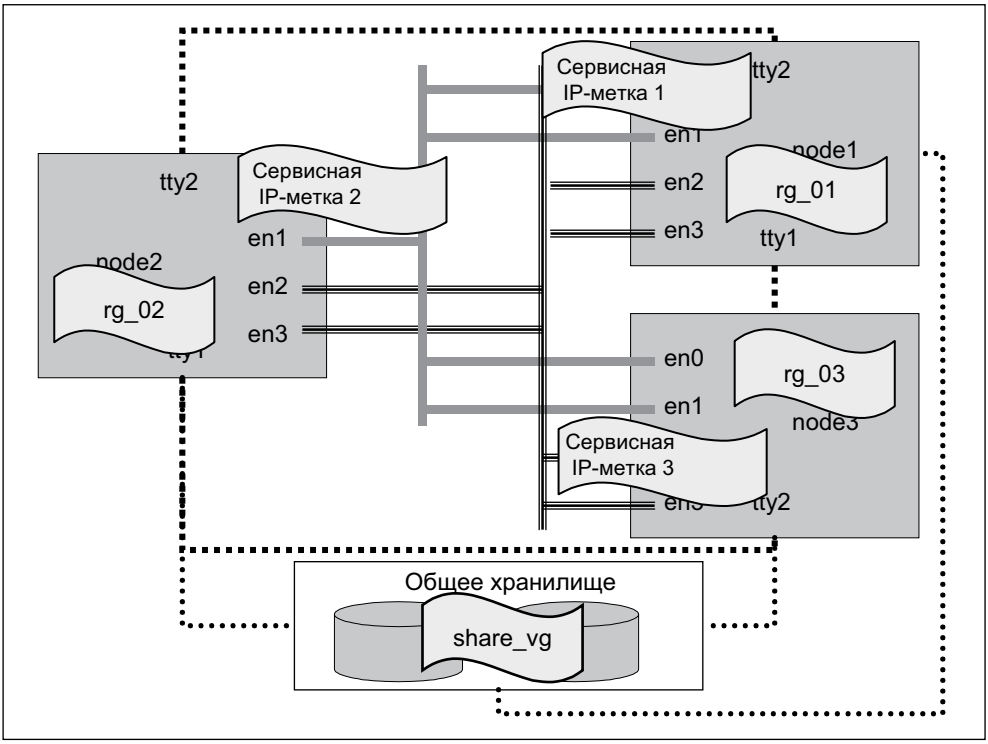

Рис. 2.11. Наложение ресурсов высокой доступности на топологию кластера

На рис. 2.11 представлено наложение ресурсов, для которых HACMP обеспечивает высокую доступность, на базовую топологию кластера:

- сервисные IP-метки;
- приложения, общие для узлов;
- хранилище, общее для узлов.

2.4.2 Ресурсы

Ресурсами считаются следующие компоненты кластера HACMP.

Сервисный IP-адрес/метка

Как обсуждалось выше, сервисным IP-адресом является IP-адрес, используемый клиентами для доступа к приложениям или узлам. Этот сервисный IP-адрес (и соответствующая метка) является частью группы ресурсов, и для него осуществляется мониторинг в HACMP. Существует два типа сервисных IP-адресов (меток):

- Общий (shared) сервисный IP-адрес (метка). IP-который может быть сконфигурирован на нескольких узлах и являющийся частью группы ресурсов, которая в любой момент времени может быть активной только на одном узле.
- Сервисный IP-адрес (метка), привязанный к узлу (node-bound). IP-адрес, который может быть сконфигурирован только на одном узле (не может совместно использоваться несколькими узлами). Обычно сервисные IP-адреса этого типа связаны с группами ресурсов с одновременным доступом (concurrent).

Сервисные IP-адреса становятся активными, когда HACMP переводит соответствующую группу ресурсов в активное (ONLINE) состояние.

В HACMP 5.3 при размещении сервисных IP-меток можно задавать следующие варианты размещения:

- **Без совместного размещения (Anti-Collocation).** Используется по умолчанию; HACMP распределяет сервисные IP-метки по всем загрузочным IP-интерфейсам в одной сети HACMP на узле.
- **С совместным размещением (Collocation).** HACMP размещает все сервисные IP-адреса на одном загрузочном IP-интерфейсе.
- **С совместным размещением и с постоянной меткой (Collocation with persistent label).** HACMP размещает все сервисные IP-адреса на загрузочном IP-интерфейсе, содержащем постоянную IP-метку синонима. Это может быть полезно в средах с VPN и брандмауэром, где только один интерфейс имеет внешний выход.
- **Без совместного размещения и с постоянной меткой (Anti-Collocation with persistent label).** HACMP размещает все сервисные IP-метки по всем загрузочным IP-интерфейсам в одной логической сети, не содержащим постоянную IP-метку синонима. Если другие интерфейсы недоступны, сервисные IP-метки совместно используют адаптер с постоянной IP-меткой синонима.

Следует отметить, что при недостаточном количестве интерфейсов, не соответствующем выбранному варианту размещения, HACMP выполняет распределение IP-меток с использованием доступных интерфейсов, чтобы обеспечить доступность сервисных IP-меток.

Вариант размещения IP-меток также может динамически изменяться, однако изменения вступают в силу только для последующих событий кластера. Это нужно для того, чтобы избежать лишних перебоев в обслуживании. Команда **cltopinfo-w** отображает политики.

Хранилище

Следующие типы хранилищ могут быть сконфигурированы как ресурсы:

- группы томов (AIX и Veritas VM);
- логические тома (все логические тома в определенной группе томов);
- файловые системы (jfs и jfs2) – либо все для определенных групп томов, либо заданные отдельно;
- диски прямого доступа (raw disks) – заданные по идентификатору PVID.

Если хранилище должно совместно использоваться некоторыми или всеми узлами в кластере, то все компоненты должны располагаться во внешнем хранилище и быть сконфигурированы таким образом, чтобы отказ одного узла не влиял на доступ с других узлов (например, при использовании SSA необходимо тщательно выполнять проверку правил циклов).

Существует два способа доступа к хранилищу:

- Конфигурации без одновременного доступа (non-concurrent), где владельцем дисков является один узел, предоставляющий клиентам доступ к ним с использованием других ресурсов, требуемых приложением. При отказе этого узла HACMP определит следующий узел, который примет владение дисками, перезапустит приложения и предоставит клиентам доступ. Диски с расширенным одновременным доступом (enhanced concurrent) часто используются в конфигурациях без одновременного доступа; помните, что понятие режима расширенного одновременного доступа относится к методу блокировки доступа к дискам и не определяет, является ли одновременным сам доступ.
- Конфигурации с одновременным доступом (concurrent), где один или несколько узлов смогут осуществлять одновременный доступ к данным и где управление блокировкой осуществляет приложение. Диски должны входить в группу томов с одновременным доступом (concurrent volume group).

HACMP поддерживает использование следующих дисковых технологий в качестве общих внешних дисков:

- SCSI;
- SSA;
- дисковые системы с подключенным оптоволоконным каналом.

Поддерживаемые устройства:

- традиционные SCSI-диски и стойки;
- SSA-диски и стойки;
- серверы хранения FastT/DS4xxx;
- серверы хранения 2105 Enterprise Storage Server, а также DS8xxx и 6xxx;
- некоторые устройства хранения сторонних производителей.

Важно! IBM может не поддерживать подсистемы и устройства хранения сторонних производителей. Список устройств хранения сторонних производителей можно найти на сайте
http://www.availant.com

Программное обеспечение предоставления множественных путей:

- устройства пути данных (data path devices, vpath);
- MPIO;
- DAC/DAR.

Поддерживаемые параллельные SCSI-устройства:

- дисковые устройства SCSI;

- дисковые стойки SCSI;
- серверы хранения FastT/DS4xxx;
- серверы хранения 2105 Enterprise Storage Server (с SCSI-подключениями).

Как правило, параллельные SCSI-устройства можно сконфигурировать в кластерах размером до четырех узлов, где все узлы подключены к одной шине SCSI. К одной шине SCSI можно подключить до 16 устройств (включая SCSI-адаптеры). Однако не рекомендуется подключать устройства, отличные от дисков (такие, как приводы CD-ROM и накопители на магнитной ленте).

Серверы хранения IBM 2105 Enterprise Storage Server

Сервер хранения IBM 2105 Enterprise Storage Server® обеспечивает подключение с одновременным доступом и общий доступ к дисковому хранилищу для различных серверов открытых систем.

Из-за множества платформ, поддерживаемых в среде с общим хранилищем, во избежание помех очень важно сконфигурировать защищенный доступ к хранилищу посредством соответствующих конфигураций маскировки LUN и разделения на зоны.

ESS использует технологию дисков IBM SSA. ESS обеспечивает встроенную доступность и защиту данных. Для защиты данных используется технология RAID. Кроме того, диски содержат внутренние средства упреждающего анализа отказов, позволяющие прогнозировать ошибки до того, как они повлияют на доступность данных.

В ESS реализовано дублирование практически всех компонентов и защита при отказе любого из внутренних компонентов. ESS осуществляет управление внутрен-

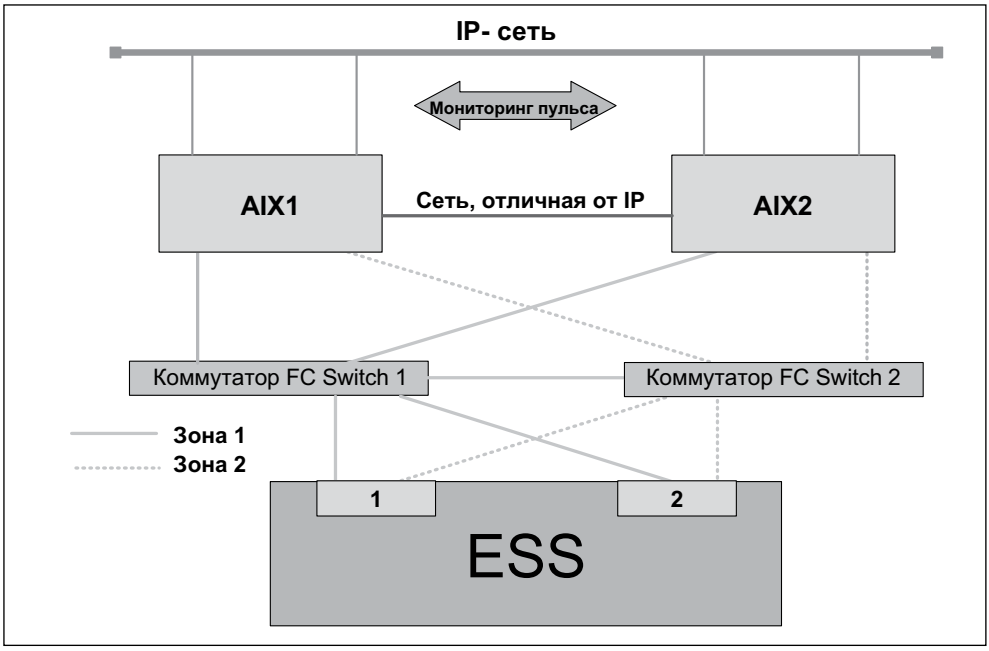

Рис. 2.12. Хранилище ESS

ним хранением (SSA-дисками) встроенным кластером из двух узлов, соединенных через высокоскоростную внутреннюю шину, где каждый узел обеспечивает одни и те же функциональные возможности. Таким образом, при отказе одного из внутренних узлов доступ клиентских систем к хранилищу сохраняется.

Дополнительные сведения о планировании и использовании сервера хранения 2105-800 Enterprise Storage Server (включая диаграммы подключения и т. д.) см. на веб-сайте

http://www.storage.ibm.com/disk/ess/index.html

Пример стандартного кластера HACMP, использующего ESS в качестве общего хранилища, представлен на рис. 2.12.

Серверы хранения IBM FAStT/DS4xxx

Серверы хранения IBM FAStT/DS4xxx обеспечивают гибкость, высокую производительность и надежность хранения для приложений в средах с несколькими узлами.

Хотя архитектура FAStT/DS4xxx и не настолько сложна, как реализованная в ESS, она также основана на максимизации избыточных компонентов – контроллеров хранения, блоков питания и адаптеров подключения хранилищ.

В архитектуре FAStT/DS4xxx реализован протокол Fibre Channel как на стороне узла, так и на стороне хранилища. Он не обеспечивает поддержку SCSI и не предоставляет выделенную высокоскоростную шину для связи между двумя контроллерами, однако он обеспечивает функцию перемещения контроллера при сбое для непрерывных операций, а также кеширование данных на стороне узла.

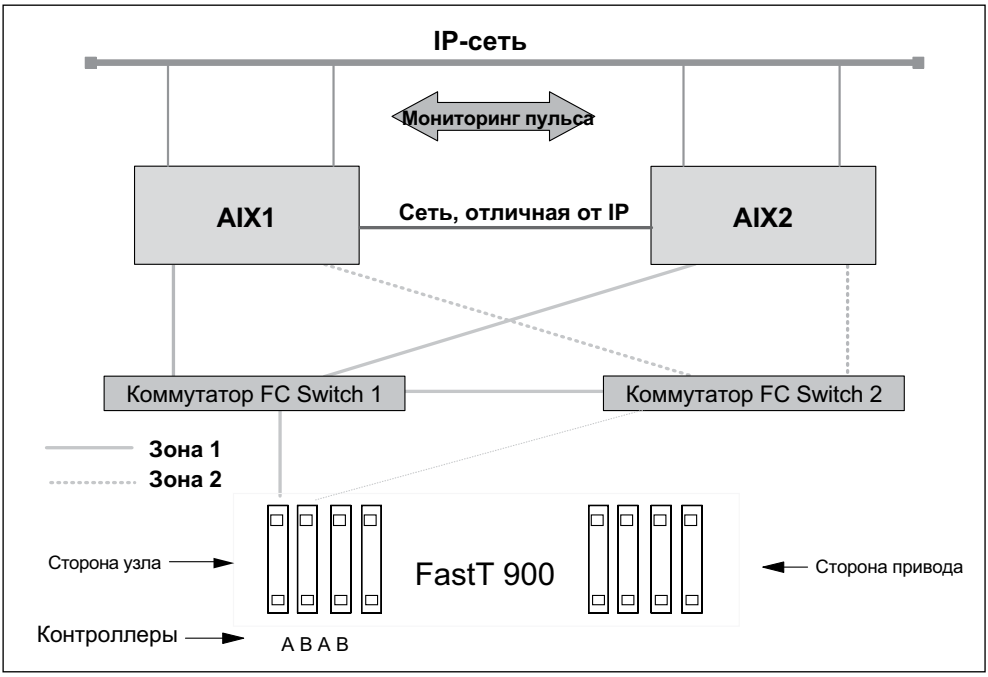

Рис. 2.13. Хранилище DS4xxx

Полные сведения о системах хранения IBM см. на веб-сайте

http://www.storage.ibm.com/disk/fastt/index.html

Стандартное подключение FAStT/DS4xxx к кластеру HACMP представлено на рис. 2.13.

Дисковая подсистема IBM Serial Storage Architecture

Подсистемы хранения SSA (Serial Storage Architecture) обеспечивают решение с большей «дискретностью компонентов» и содержат средства, позволяющие сократить количество единых точек отказа.

Хранилище SSA обеспечивает высокую доступность в среде HACMP посредством использования избыточного оборудования (блоков питания и подключений к хранилищу) и функции «горячей замены» (одновременного обслуживания) для блоков питания и дисков.

Хранилище SSA также позволяет реализовать возможности RAID на уровне адаптера (Host Bus Adapter, HBA).

Примечание. При использовании SSA RAID количество узлов HACMP, допускающих совместное использование одних данных, ограничено двумя.

SSA-дисковые подсистемы IBM 7133 могут использоваться в качестве внешних общих дисковых устройств хранения, предоставляющих одновременный доступ в конфигурации кластера HACMP.

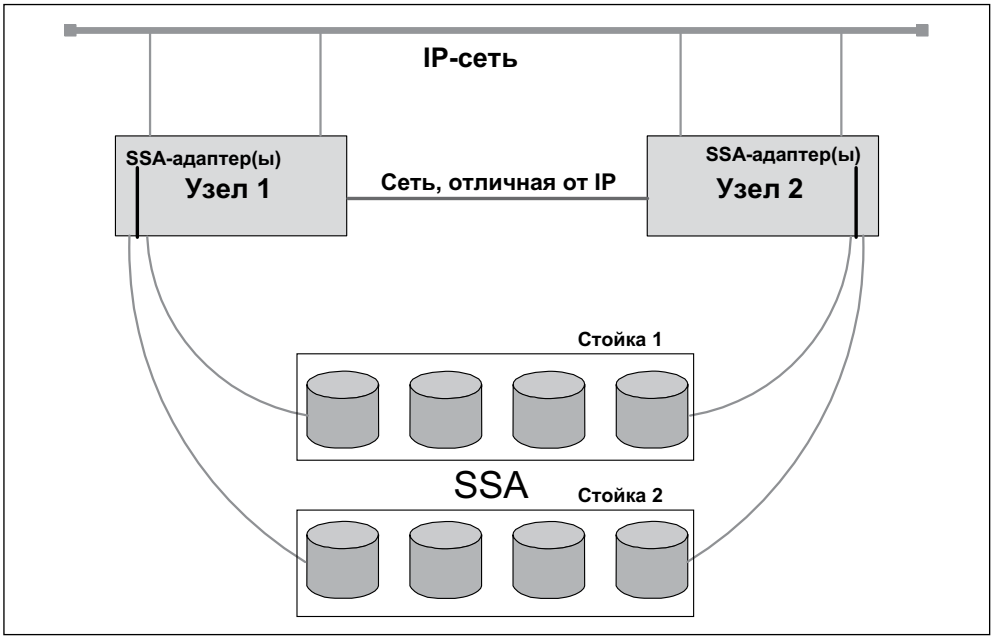

Рис. 2.14. SSA-хранилище

SSA-хранилище обеспечивает гибкий, достаточно простой и более «тонкий» подход к конфигурированию кластеров HACMP с существующими или устаревшими приложениями и ограниченным количеством узлов. Мы рекомендуем осуществлять внедрение всех новых конфигураций с использованием новых технологий (FC-хранилище).

На рис. 2.14 представлен пример кластера HACMP, состоящего из двух узлов.

Выбор способа защиты данных

Защита хранилища (данных или чего-либо другого) осуществляется независимо от HACMP; для обеспечения высокой доступности хранилища необходимо использовать хранилище с надлежащим уровнем избыточности и отказоустойчивости. HACMP не имеет какого-либо контроля над доступностью хранилища. Для защиты данных можно технологию RAID либо применять (на уровне хранилища или адаптера), либо зеркальное отображение AIX LVM (RAID 1).

Дисковые массивы RAID (Redundant Array of Independent Disks) представляют собой группы совместно работающих дисковых приводов, обеспечивающих более высокую скорость передачи, чем одиночные (независимые) приводы. Массивы также могут обеспечивать избыточность данных, чтобы не возникало потерь данных при отказе одного привода (физического диска) в массиве. В зависимости от уровня RAID выполняется либо зеркальное отображение данных, либо чередование данных, либо совмещение обоих методов. Ниже дается описание подуровней RAID.

- *RAID-0*. Также называется «чередованием данных» (striping). Обычно файл последовательно записывается на один диск. При чередовании информация разделяется на фрагменты – chunks (при фиксированном размере такие фрагменты обычно называются блоками – blocks), а фрагменты параллельно записываются на набор дисков (или считываются с него). Это дает следующие преимущества с точки зрения производительности: более высокую скорость передачи данных при последовательных операциях в связи с наложением нескольких потоков ввода-вывода; более высокую пропускную способность при произвольном доступе, так как устраняется сдвиг схемы доступа в связи с распределением данных. Это означает, что при равномерном распределении данных по нескольким дискам требуемая информация, скорее всего, будет расположена на нескольких дисках, вследствие чего повышается пропускная способность нескольких приводов. RAID-0 предназначен только для повышения производительности. Избыточность отсутствует, так что каждый диск является единой точкой сбоя.
- *RAID-1*. RAID-1 также называется «зеркальным отображением дисков». При такой схеме разные диски содержат идентичные копии каждого фрагмента данных; другими словами, у каждого диска есть «двойник», содержащий точную реплику (зеркальный образ) информации. При отказе какого-либо диска в массиве наличие зеркального диска обеспечивает доступность данных. Это также позволяет повысить производительность чтения, так как всегда используется тот диск, у которого привод (головка диска) расположен ближе к требуемым данным, что сокращает время поиска. Время реакции при записи может быть дольше, чем при одном диске, в зависимости от политики записи; операции записи могут выполняться либо параллельно (для более быстрой реакции), либо последовательно (для надежности).

- *RAID-2 и RAID-3.* Представляют собой массивы с механизмом параллельной обработки, где все приводы в массиве работают в согласии. Подобно чередованию данных, записываемая на диск информация разделяется на фрагменты (фиксированные блоки данных) и каждый фрагмент записывается в одну и ту же физическую позицию на различных дисках (параллельно). При чтении на каждый диск могут направляться одновременные запросы данных. Такая архитектура требует записи данных четности для каждой полосы данных; различие между RAID-2 и RAID-3 состоит в том, что RAID-2 может использовать несколько дисковых приводов для записи данных четности, тогда как RAID-3 может использовать только один дисковый привод. При отказе привода система может воссоздать потерянные данные по оставшимся дискам и данным четности. При больших объемах данных производительность очень высокая, однако на небольших запросах производительность низкая, так как во всех операциях всегда применяются все приводы и не может иметь место наложение или независимое выполнение операций.
- *RAID-4.* Эта технология призвана решить некоторые недостатки технологии RAID-3 путем использования фрагментов данных более крупного размера и чередования данных по всем дискам, кроме одного, зарезервированного для данных четности. Использование чередования данных означает, что запросы ввода-вывода должны только указывать привод, на котором требуемые данные находятся в действительности. Это означает, что возможно одновременное и независимое выполнение операций чтения. Однако запросы записи требуют использования цикла «чтение/изменение/обновление», что создает «узкое место» на одном диске четности. Необходимо считать каждую полосу, вставить новые данные, после чего вычислить новые данные четности, прежде чем записать полосу обратно на диск. Затем на диск четности записываются новые данные четности, и этот диск не может использоваться для других операций записи, пока не будет завершена эта операция. Из-за этого «узкого места» технология RAID-4 не используется настолько часто, как RAID-5, в которой тот же процесс реализован без «узкого места».
- *RAID 5.* Эта технология очень похожа на RAID-4. Разница состоит в том, что данные четности также распределяются по тем же дискам, которые используются для данных, устраняя, таким образом, «узкое место». Данные четности никогда не сохраняются на том же приводе, на котором сохраняются и защищаемые им фрагменты. Это позволяет выполнять одновременные операции чтения и записи, увеличивая производительность по причине доступности дополнительного диска (диска, который ранее использовался для данных четности). Существуют другие функции, позволяющие еще больше повысить скорость передачи данных, такие как кеширование одновременных операций чтения с дисков и передача этой информации во время чтения следующих блоков. Это позволяет достичь скорости передачи данных на уровне скорости адаптера.

  Как и в технологии RAID-3, при отказе диска информация может быть воссоздана с других приводов. Массив RAID-5 всегда использует данные четности, однако все равно важно регулярно создавать резервные копии данных в массиве. Массивы RAID-5 распределяют данные по всем дискам массива, по одному сегменту зараз (сегмент может содержать несколько блоков). В массиве из n приводов полоса содержит сегменты данных, записываемые на n-1 дисков, и сегмент четности, запи-

сываемый на n-й диск. Использование этого механизма также означает, что не все дисковое пространство доступно для записи данных. Например, в массиве из пяти дисков по 72 Гб, несмотря на то что суммарный размер составляет 360 Гб, только 288 Гб доступны для записи данных.

- *RAID 0+1 (RAID-10)*, также известный как IBM RAID-1 Enhanced или RAID-10, представляет сочетание RAID-0 (чередование данных) и RAID-1 (зеркальное отображение данных). RAID-10 имеет такие же преимущества производительности, как и RAID-0, обеспечивая при этом доступность данных, характерную для RAID-1. В конфигурации RAID-10 осуществляется чередование как для данных, так и для их зеркального отображения по всем дискам массива. Первая полоса представляет полосу данных, тогда как вторая полоса представляет зеркальное отображение, где зеркальное отображение записывается на другой физический диск. Реализации RAID-10 обеспечивают отличную производительность записи, так как им не приходится вычислять и записывать данные четности. RAID-10 может быть реализован программным способом (AIX LVM), аппаратным способом (уровень подсистемы хранения) либо сочетанием аппаратного и программного способов. Выбор подходящего решения зависит от общих требований. RAID-10 имеет такие же стоимостные характеристики, как и RAID-1.

Таблица 2.3. Характеристики широко используемых уровней RAID

Уровень RAID	Доступное дисковое пространство, %	Производительность операций чтения-записи	Стоимость	Защита данных
RAID-0	100	Высокая и для чтения и для записи	Низкая	Нет
RAID-1	50	Средневысокая для чтения, средняя для записи	Высокая	Есть
RAID-5	80	Высокая для чтения, средняя для записи	Средняя	Есть
RAID 0+1	50	Высокая и для чтения и для записи	Высокая	Есть

Важно! Несмотря на то, что на всех уровнях технологии RAID (кроме RAID-0) реализуется избыточность данных, необходимо регулярно осуществлять резервное копирование данных. Это единственный способ восстановления данных в случае случайного повреждения или удаления файла или каталога.

Наиболее распространенные уровни RAID, используемые в современных реализациях информационных систем, приведены в табл. 2-3.

Аспекты кворума LVM

Для групп с одновременным доступом томов должен быть включен кворум, так как каждый узел может осуществлять доступ к разным дискам, вызывая несоответствие данных.

Если кворум оставлен включенным (по умолчанию), то в случае потери кворума происходит перемещение при сбое для группы ресурсов и для группы томов будет

выполнена принудительная активизация (forced varyon) на другом узле, если принудительная активизация групп томов включена. Если принудительная активизация групп томов включена, HACMP проверяет:

- наличие как минимум одной копии каждого зеркального набора в группе томов;
- читаемость каждого диска;
- наличие как минимум одной доступной копии для каждого логического раздела на каждом логическом томе.

Если эти условия выполняются, HACMP осуществляет принудительную активизацию группы томов.

Использование групп томов с режимом расширенного одновременного доступа

В ранних версиях контроль доступа к группе томов осуществлялся блокировками SCSI, тогда как HACMP включал утилиты для отключения этих блокировок в случае, если узел не освободил группы томов надлежащим образом. В AIX 5.x появились группы томов с расширенным одновременным доступом (enhanced concurrent volume groups) и для блокировки используются RSCT. Еще одно усовершенствование состоит в возможности активизации группы томов в одном из двух режимов:

- **Активное состояние (active).** Группа томов работает так же, как при традиционной активизации, допускается выполнение операций для группы томов и логических томов, файловые системы могут быть смонтированы.
- **Пассивное состояние (passive).** При пассивном состоянии допускается ограниченный доступ к VGDA и LVCB только для чтения.

При интеграции узла в кластер HACMP осуществляет построение списка всех групп томов с расширенным одновременным доступом, являющихся ресурсом в любой группе ресурсов, содержащей узел. Эти группы томов затем активизируются в пассивном режиме.

Когда группа ресурсов переходит на узле в подключенное (online) состояние (или, по-другому, активизируется), тогда группы томов с расширенным одновременным доступом активизируются в активном режиме. Когда группа ресурсов на узле переходит в отключенное (offline)состояние (деактивизируется), группа томов возвращается в пассивный режим.

Для отображения состояния группы томов с режимом расширенного одновременного доступа используются команды **lspv** и **lsvg**; команда **lspv** выводит значения «active» и «passive», тогда как команда **lsvg** выводит значение «active» в поле VG STATE при обоих режимах, а в поле VG PERMISSION выводит значение либо «read / write», либо «passive-only».

> **Важно!** При использовании групп томов с расширенным одновременным доступом важно, чтобы для мониторинга пульса RSCT использовалось несколько сетей. При отсутствии SCSI-блокировки разделенный кластер может очень быстро активизировать группу томов и таким образом повредить данные.

Режим одновременного доступа RAID и SSA

Начиная с HACMP 5.x системы с 64-разрядным ядром должны использовать группы томов с режимом расширенного одновременного доступа (Enhanced Concurrent Mode, ECM). В системах с 32-разрядным ядром осталась поддержка режимов одновременного доступа RAID и SSA, однако начиная с AIX 5.2 невозможно создавать группы томов с одновременным доступом в любом режиме, кроме расширенного.

Режим расширенного одновременного доступа является рекомендуемым, так как:

- он поддерживает мониторинг пульса дисков – обеспечивает большую гибкость, чем при использовании последовательных соединений;
- он обеспечивает быстрый перехват диска (fast disk takeover) – сокращает время на перемещение при сбое;
- в нем легче обеспечить согласованность VGDA между узлами.

Общие физические тома

Для приложений, осуществляющих доступ к диску прямого доступа (raw disk), в качестве ресурса в группу ресурсов может быть добавлен идентификатор физического тома (Physical Volume Identifier, PVID).

Общие логические тома

Хотя общие логические тома и не конфигурируются явным образом как часть группы ресурсов, каждый логический том в общей группе томов будет доступен на узле, когда группа ресурсов будет в подключенном режиме. Эти общие логические тома могут быть сконфигурированы либо для доступа с одного узла, либо для одновременного доступа с нескольких узлов. Если нужно изменить владение логическим томом, не забудьте переустанавливать его при каждом импортировании вышестоящей группы томов.

Хотя это и не относится непосредственно к HACMP, помните о том, что некоторые приложения, использующие логические тома прямого доступа (raw logical volumes), начинают запись с начала устройства, перезаписывая таким образом контрольный блок логического тома (logical volume control block, LVCB).

Настраиваемые методы работы с дисками

Расширенные меню ресурсов (extended resource) SMIT позволяют создавать настраиваемые методы работы с дисками, томами и файловыми системами. Для создания настраиваемого метода необходимо определить в HACMP соответствующие скрипты управления устройствами в среде высокой доступности, например:

- **для настраиваемых дисков:** HACMP предоставляет скрипты выявления фантомных дисков (ghost disks), определяет, есть ли резервирование (блокировка) диска узлом, сбрасывает бит резервирования и делает диск доступным;
- **для групп томов:** HACMP предоставляет скрипты для вывода имен групп томов, вывода дисков в группе томов, перевода группы томов в подключенное и отключенное состояние;

- **для файловых систем:** HACMP предоставляет скрипты подключения (монтирования), отключения, вывода списка и проверки состояния.

HACMP 5.3 содержит настраиваемые методы для Veritas Volume Manager (VxVM), использующего Veritas Foundation Suite v4.0.

Файловые системы (jfs и jfs2) - fsck и logredo

Собственные файловые системы AIX используют технологии ведения журналов баз данных для поддержки своей структурной целостности. Таким образом, после отказа AIX использует журнал файловой системы (JFSlog) для восстановления последнего согласованного состояния файловой системы. Этот метод работает быстрее, чем при использовании утилиты **fsck**. В случае отказа процесса восстановления с использованием журнала JFSlog возникнет ошибка и файловая система подключена не будет.

Утилита **fsck** выполняет проверку согласованности файловой системы, проверяя индексные дескрипторы (inodes), структуру каталогов и файлы. Хотя при ее использовании больше шансов восстановить поврежденные файловые системы, ее выполнение занимает больше времени.

> **Важно!** Восстановление согласованного состояния файловой системы не гарантирует согласованности данных, за это отвечает приложение.

2.4.3 NFS

HACMP работает с сетевой файловой системой AIX (NFS), обеспечивая высокую доступность NFS-сервера, позволяя резервному NFS-серверу восстановить рабочее состояние NFS в случае отказа основного NFS-сервера. Такая возможность доступна только для кластеров из двух узлов, так как HACMP сохраняет блокировки файловых систем NFS и корректно обрабатывает кеш дублирования запросов. При «подхвате» группы ресурсов NFS другим узлом подключенные клиенты испытают такое же отключение, как при перезагрузке NFS-сервера.

При конфигурировании NFS через HACMP можно осуществлять управление:

- Сетью, которую HACMP использует для подключения NFS.
- Экспортом и подключениями NFS на уровне каталогов.
- Способами экспорта для экспортируемых каталогов и файловых систем NFS. Эта информация хранится в файле **/usr/es/sbin/cluster/etc/exports**, имеющем такой же формат, как и файл exports операционной системы AIX – **/etc/exports**.

Ограничения NFS и HACMP

- Кластер может содержать только два узла
- Общие группы томов, содержащие файловые системы, подлежащие экспорту с использованием NFS, должны иметь один и тот же старший номер устройства (major number) на всех узлах; в противном случае восстановление клиентских приложений во время перемещения при сбое будет невозможно.

- Если операции экспорта NFS определены на узле посредством HACMP, все операции экспорта NFS должны контролироваться HACMP. Нельзя смешивать операции экспорта AIX и HACMP NFS.
- Если в группе ресурсов определены операции экспорта NFS, в поле file systems mounted before IP configured нужно установить значение true.
- По умолчанию для группы ресурсов, содержащей экспортированные файловые системы NFS, автоматически выполняется перекрестное подключение (cross-mount). Это подразумевает также, что каждый узел в группе ресурсов будет действовать как NFS-клиент, так что он должен иметь IP-метку в той же подсети, в которой ее имеет и сервисная IP-метка NFS-сервера.

Перекрестные подключения NFS

Перекрестные подключения (cross-mounts) NFS работают следующим образом:

- узел, содержащий группу ресурсов NFS, осуществляет экспорт всех файловых систем NFS;
- каждый узел в группе ресурсов монтирует все файловые системы NFS, определенные в группе ресурсов;
- при «подхвате» группы ресурсов другим узлом этот узел выполнит монтирование файловых систем NFS и их повторный экспорт.

  Например:
- узел1 с сервисной IP-меткой svc1 монтирует и экспортирует /fs1;
- узел2 монтирует svc1:/fs1 к /mntfs1;
- узел1 также монтирует svc1:/fs1 к /mntfs1.

2.4.4 Серверы приложений

Практически любое приложение, которое может выполняться на автономном сервере AIX, может выполняться и в кластерной среде, защищенной HACMP. Приложение должно иметь возможность запуска и остановки из скриптов, а также возможность восстановления из скрипта после неожиданной остановки.

Приложения определяются в HACMP как сервер приложений (application server) со следующими атрибутами:

- **Скрипт запуска (start script).** Этот скрипт должен быть способен запустить приложение как после обычной, так и после неожиданной остановки. Выходные данные скрипта записываются в файл журнала **`hacmp.out`**. Код завершения скрипта отслеживается HACMP.
- **Скрипт остановки (stop script).** Этот скрипт должен быть способен успешно остановить приложение. Выходные данные скрипта также записываются в журнал, и код завершения также отслеживается.
- **Мониторы приложений (application monitors).** Для обеспечения высокой доступности приложений HACMP может осуществлять мониторинг самого приложения, а не только требуемых ресурсов.

Начиная с HACMP 5.2 эти скрипты должны быть одинаковыми на всех узлах, так что, если необходима специальная настройка для определенных узлов, процедура должна проверять имя узла.

В ходе мониторинга кодов завершения скриптов приложений HACMP предполагает, что ненулевой код завершения скрипта означает сбой скрипта и что запуск или остановка приложения были неуспешными. В этом случае группа ресурсов переходит в ошибочное состояние и записывается сообщение (событие) **config_too_long**.

При конфигурировании приложения для работы в HACMP необходимо учитывать следующее:

- Совместимо ли приложение с версией AIX.
- Совместима ли среда хранения с кластером высокой доступности.
- Нужно хорошо понимать взаимозависимости приложения и платформы. Расположение кода приложения, данных, временных файлов, сокетов, каналов и прочих компонентов системы, например принтеров, должно реплицироваться по всем узлам, на которых будет располагаться приложение.
- Как уже обсуждалось, должна быть реализована возможность запуска и остановки приложения без вмешательства оператора, в частности после неожиданной остановки узла. Скрипты запуска и остановки приложения должны быть тщательным образом протестированы перед внедрением и при каждом изменении в среде.
- Группа ресурсов, содержащая приложение, должна содержать все ресурсы, требуемые для работы приложения, или быть дочерней группой ресурсов для такой группы ресурсов.
- Необходимо учитывать лицензирование приложения. Многие приложения имеют лицензии, основанные на коде процессора; необходимо провести тщательное планирование, чтобы убедиться в том, что приложение может быть запущено на любом узле из списка узлов группы ресурсов. Также следует быть внимательным с количеством процессоров на каждом узле и т. п., так как некоторые лицензии восприимчивы к таким показателям.

2.4.5 Мониторы приложений

HACMP использует мониторы приложений (application monitors) для обеспечения высокой доступности приложений. Начиная с HACMP 5.2 каждое приложение может иметь несколько мониторов.

HACMP использует два типа мониторов приложений:

- **Мониторы процессов (process monitors).** Обнаруживают прерывание одного или нескольких процессов с использованием RSCT Resource Monitoring and Control (RMC). Следует быть внимательным, чтобы выбрать правильное имя процесса, – используйте выходные данные команды **ps -el**, чтобы быть уверенным в том, что выбран требуемый процесс. Можно осуществлять мониторинг нескольких процессов. Один монитор приложения может осуществлять мониторинг определенного количества экземпляров одного процесса или единичных экземпляров множества процессов.

- **Настраиваемые мониторы (custom monitors).** Тестируют состояние приложения с использованием настраиваемого пользовательского скрипта. В качестве метода должна применяться исполняемая программа (может использоваться shell-скрипт), тестирующая приложение и завершающаяся после этого. При возврате нулевого значения приложение считается нормально работающим, ненулевое значение указывает на неудачный запуск или отказ приложения.

Например, монитор базы данных может выполнять запросы к базе данных и возвращать код завершения, соответствующий ответу базы данных.

Начиная с HACMP 5.2 каждый тип монитора может иметь различные режимы работы:

- **Монитор запуска (startup monitor).** Монитор приложения проверяет, был ли сервер приложения успешно запущен в заданный интервал стабилизации. Данный тип монитора специально предназначен для родительских групп ресурсов в отношениях зависимости между родительскими и дочерними группами ресурсов, так как HACMP запускает скрипт запуска приложения в фоновом режиме и не ожидает его завершения. При использовании монитора запуска дочерние группы ресурсов не переводятся в подключенный режим, пока монитор не подтвердит, что приложение было успешно запущено. Если для приложения сконфигурировано несколько мониторов запуска, используется самый длинный стабилизационный интервал, соответствующий времени события.
- **Монитор выполнения (long running monitor).** Монитор приложения периодически проверяет, работает ли сервер приложения после завершения интервала стабилизации. Этот режим используется по умолчанию.
- **Двойной монитор (both).** Монитор приложения проверяет, было ли приложение запущено за период стабилизации, после чего осуществляет периодический мониторинг после завершения периода стабилизации.

При конфигурировании мониторов процессов необходимо определить следующее:

- **Имя монитора.** Уникальное имя.
- **Режим монитора.** Монитор запуска, монитор выполнения или двойной монитор.
- **Сервер приложения.** Сервер приложения, подлежащий мониторингу.
- **Интервал стабилизации.** В зависимости от режима:
  - В режиме запуска осуществляется регулярный мониторинг приложения в течение заданного периода, чтобы определить, было ли оно успешно запущено. Если монитор сообщает, что приложение было успешно запущено, то HACMP закрывает монитор и продолжает обработку. Однако если приложение не было успешно запущено к концу периода, то считается, что произошел отказ при «подхвате» группы ресурсов и HACMP предпринимает действия по восстановлению.
  - В режиме выполнения HACMP в течение заданного периода ожидает стабилизации приложения перед началом мониторинга приложения.
  - В двойном режиме выполняется периодическая проверка приложения в течение заданного периода, чтобы убедиться в том, что приложение было запущено успешно; после завершения этого периода осуществляется мониторинг приложения, чтобы убедиться, что оно продолжает выполняться.

- **Счетчик перезапусков.** Указывает количество попыток перезапуска приложения, прежде чем HACMP предпримет действия, описанные ниже. При мониторинге запуска этот параметр имеет значение 0.
- **Интервал перезапуска.** Указывает время, в течение которого приложение должно быть стабильным, прежде чем сбросить счетчик перезапусков. Рекомендуется, чтобы этот показатель содержал значение, равное или больше чем значение **счетчик перезапуска x интервал стабилизации**. По умолчанию задается 110 % этого значения.
- **Действие, выполняемое при отказе приложения*.** Можно задать действие, предпринимаемое HACMP, если приложение не удалось перезапустить по истечении количества перезапусков, заданных счетчиком перезапусков. По умолчанию задано уведомление, которое инициирует событие уведомления. Другой вариант – перемещение при сбое, при котором HACMP выполняет «подхват» группы ресурсов на другом узле из списка узлов групп ресурсов либо на следующий узел, соответствующий критериям динамического приоритета узла.
- **Метод уведомления.** Используется при отказе приложения.
- **Метод удаления*.** Необязательный параметр, определяющий метод, используемый HACMP для удаления приложения после его отказа перед попыткой перезапуска приложения. Если используется только один сервер приложения, по умолчанию этот параметр указывает на его скрипт остановки; однако необходимо быть внимательным, так как, если приложение уже остановлено, этот скрипт может дать сбой.
- **Метод перезапуска*.** Необязательный параметр, определяющий метод, используемый HACMP при попытке перезапуска приложения, если счетчик перезапусков не равен нулю. И опять же, если используется только один сервер приложения, по умолчанию этот параметр указывает на его скрипт запуска.

Для мониторов процессов следует определить следующее:

- **Наблюдаемый процесс.** Имя процесса, для которого осуществляется мониторинг (чтобы определить корректное имя процесса, воспользуйтесь командой **ps-el**).
- **Владелец процесса.** Имя пользователя, владеющего всеми процессами, для которых осуществляется мониторинг (например, **db2adm**).
- **Счетчик экземпляров.** Должно быть задано количество экземпляров. Если мониторинг осуществляется только для одного процесса, то значение этого параметра должно быть равным количеству экземпляров, и в случае, если оно будет большим или меньшим, выдается ошибка. Если осуществляется мониторинг нескольких процессов, то допускается только один экземпляр для каждого процесса.

  Для настраиваемых мониторов следует определить:
- **Метод мониторинга.** Имя программы, используемое HACMP для мониторинга сервера приложений. При возвращении ненулевого кода завершения HACMP предполагает, что возникла проблема с приложением.

---

* Не относится к мониторам запуска.

- **Интервал мониторинга.** Количество секунд, в течение которых HACMP ожидает возврата кода завершения от метода мониторинга. Если за это время монитор не отреагирует, HACMP предположит, что он завис.
- **Сигнал при зависании монитора.** Сигнал, отправляемый HACMP монитору приложения в том случае, если ответ не был получен за время, заданное интервалом мониторинга. По умолчанию отправляется сигнал **SIGKILL**.

Если используется несколько мониторов для определенного приложения, HACMP обрабатывает их следующим образом:

- если сначала вызывается монитор с политикой перемещения при сбое, тогда запускается процесс перемещения при сбое и другие мониторы для этого приложения отключаются до тех пор, пока приложение не будет перезапущено;
- если сначала вызывается монитор с политикой уведомления, то запускается определенный метод уведомления, тогда как другие мониторы будут продолжать работать.

Доступность приложения

HACMP также содержит инструмент анализа доступности приложения (application availability analysis tool), который может использоваться для аудита общей доступности приложения, а также для оценки кластерной среды.

2.4.6 Коммуникационные адаптеры и каналы

HACMP поддерживает три типа коммуникационных каналов (communication links):

- SNA, сконфигурированная через интерфейс локальной сети;
- SNA через X.25;
- X.25.
- Из-за способа использования X.25 эти интерфейсы воспринимаются как отдельный класс интерфейсов или как устройства, не включенные в топологию HACMP, и поэтому управление ими осуществляется с использованием нестандартных методов. В частности, для мониторинга интерфейсов X.25 не применяется мониторинг пульса. Новый демон **clcommlinkd** для мониторинга состояния ссылки X.25 использует **x25status.**

2.4.7 Накопители на магнитной ленте

Некоторые накопители на магнитной ленте с подключением SCSI или Fibre Channel могут быть сконфигурированы как ресурсы высокой доступности в составе любой группы ресурсов (кроме concurrent).

2.4.8 Ресурсы Fast Connect

Сервер приложений Fast Connect не требует конфигурирования скриптов запуска и остановки, так как они уже интегрированы в HACMP. После конфигурирования Fast Connect в качестве ресурса HACMP система HACMP начинает поддерживать запуск, остановку, перемещение при сбое, возврат после восстановления и восстановление

служб Fast Connect. Службы Fast Connect не могут выполняться в кластере в момент его старта, так как HACMP требуется осуществлять управление Fast Connect.

Если сконфигурирован перехват IP-адреса и аппаратного адреса, клиентам не требуется повторно устанавливать свое подключение.

2.4.9 Интеграция с WLM

Диспетчер рабочей нагрузки (workload manager, WLM) представляет собой инструмент администрирования ресурсов AIX, позволяющий устанавливать целевые значения и ограничения по использованию процессора, физической памяти и пропускной способности дискового ввода-вывода для приложений и пользователей. Могут быть сконфигурированы WLM-классы, каждому из которых соответствует определенный набор системных ресурсов. Затем определяются правила назначения приложений или групп пользователей классу и, таким образом, набор используемых ими ресурсов.

Применяя конфигурацию HACMP, WLM запускается либо при подключении узла к кластеру, либо в результате использования WLM системой DARE, и только на узлах, входящих в группы ресурсов, содержащие классы WLM. HACMP работает с WLM двумя способами:

- если WLM уже выполняется, HACMP сохраняет работающую конфигурацию, останавливает WLM и выполняет перезапуск с файлами конфигурации HACMP. Когда HACMP останавливается на узле, активизируется предыдущая конфигурация WLM;
- если WLM не выполняется, он запускается с конфигурацией HACMP и останавливается, когда HACMP останавливается на узле.

Внимание! HACMP может выполнять только ограниченную проверку конфигурации WLM. Необходимо заранее выполнить надлежащее планирование.

Конфигурация, используемая WLM на узле, настраивается отдельно для каждого узла и групп ресурсов, которые могут быть переведены в подключенный режим на этом узле. Классы диспетчера нагрузки могут быть назначены группам ресурсов как:

- основные классы (Primary classs);
- дополнительные классы (Secondary class).

При интеграции узла в кластер HACMP выполняет проверку каждой группы ресурсов, соответствующей данному узлу в списке узлов. Используемые классы WLM зависят от политики запуска для каждой группы ресурсов и приоритета узлов в списке узлов.

Основной класс WLM

Если политика группы ресурсов предусматривает ее запуск либо только на домашнем узле (online on home node only), либо на первом доступном узле (online on first available node), то:

- если узел имеет наивысший приоритет в списке узлов, используется основной класс WLM;

- если узел не является узлом с наивысшим приоритетом и не определен дополнительный класс WLM, узел использует основной класс WLM;
- если узел не является узлом с наивысшим приоритетом и определен дополнительный класс WLM, узел использует дополнительный класс WLM.

Если группа ресурсов имеет политику запуска либо для подключенного режима на всех доступных узлах (online on all available nodes, одновременный доступ), либо для подключенного режима с использованием политики распределения узлов (online using a node distribution policy), то узел применяет основной класс WLM.

Дополнительный класс WLM

Является необязательным и используется только на узлах, не являющихся основными узлами для групп ресурсов с политикой запуска либо только на домашнем узле, либо на первом доступном узле.

2.4.10 Группы ресурсов

Для обеспечения высокой доступности в HACMP каждый ресурс должен быть включен в группу ресурсов (resource group, RG). Группы ресурсов позволяют HACMP осуществлять управление соответствующим набором ресурсов как единым объектом. Например, приложение может содержать скрипты запуска и остановки, базу данных и IP-адрес. Эти ресурсы включаются в группу ресурсов, которой HACMP управляет как единым объектом.

HACMP обеспечивает высокую доступность групп ресурсов путем их перемещения с узла на узел при изменении состояния кластера. Существуют следующие основные состояния кластера и соответствующие им действия групп ресурсов:

- **Запуск кластера.** Узлы кластера стартуют, и после этого группы ресурсов распределяются по ним в соответствии с политикой запуска.
- **Отказ/восстановление ресурса.** Когда определенный ресурс, являющийся частью группы ресурсов, становится недоступным, группу ресурсов можно переместить на другой узел. Подобным образом можно выполнить ее обратное перемещение, когда ресурс становится доступным.
- **Завершение работы HACMP на узле.** Существует множество способов остановить HACMP на узле. Один метод вызывает перемещение групп ресурсов узла на другие узлы. Другой метод переводит группы ресурсов в отключенный режим. При определенных обстоятельствах можно остановить службы кластера на узле, оставив ресурсы активными.
- **Отказ/восстановление узла.** При отказе узла группы ресурсов, которые были активными на узле, распределяются по другим узлам в кластере, в зависимости от их политик распределения перемещения при сбое. После восстановления узла и его реинтеграции в кластер группы ресурсов могут быть снова перемещены на него, в зависимости от политик перемещения при восстановлении.
- **Завершение работы кластера.** После завершения работы кластера все группы ресурсов переводятся в отключенный режим. Однако существует несколько кон-

фигураций, при которых ресурсы могут остаться активными, но ресурсы кластера останавливаются.

Прежде чем разбирать режимы работы и атрибуты, конфигурируемые для групп ресурсов, необходимо рассмотреть следующие термины:

- **Список узлов (node list).** Представляет перечень узлов, способных содержать определенную группу ресурсов. Каждый узел должен быть способен получить доступ к ресурсам, составляющим группу ресурсов.
- **Приоритет узла по умолчанию (default node priority).** Представляет порядок, в котором узлы определяются в группе ресурсов. При отказе узлов группа ресурсов с атрибутами по умолчанию будет перемещаться с узла на узел в этом порядке.
- **Домашний узел (home node).** Узел с наивысшим приоритетом в списке узлов по умолчанию. По умолчанию указывает на узел, на котором группа ресурсов будет изначально активизироваться. Под этим термином не подразумевается узел, на котором группа ресурсов активна в данный момент.
- **Запуск (startup).** Процесс перевода группы ресурсов в подключенное состояние (online).
- **Перемещение при сбое (fallover).** Процесс перемещения группы ресурсов, находящейся в подключенном состоянии, с одного узла на другой узел в кластере в ответ на событие.
- **Возврат после восстановления (fallback).** Процесс перемещения группы ресурсов, находящейся в подключенном состоянии, с узла, не являющегося ее домашним узлом, на узел, для которого выполняется реинтеграция.

Режимы работы группы ресурсов-политики и атрибуты

Работа группы ресурсов определяется путем конфигурирования политик и режимов работы группы ресурсов. Ранние версии HACMP (до 5.1) поддерживали три предопределенные группы ресурсов:

- **Каскадные группы ресурсов (cascading).** Изначально предназначались для того, чтобы группы ресурсов имели сродство (affinity) с определенным узлом – имели преимущество запуска на этом узле и осуществляли возврат после восстановления, когда он снова становится доступным. Режим работы может контролироваться сочетанием трех атрибутов:
  - **Переход при неактивности (Inactive takeover, ITO).** Управляет переводом группы ресурсов в подключенное состояние узлами, отличными от узла с наибольшим приоритетом в списке узлов группы ресурсов.
  - **Каскадирование без возврата после восстановления (Cascading without fallback, CWOF).** Определяет, возможен ли переход группы ресурсов на узел с более высоким приоритетом при его интеграции в кластер.
  - **Динамический приоритет узла (Dynamic Node Priority, DNP).** Соответствует такому же атрибуту в настраиваемых группах ресурсов (custom resource groups).

- **Ротационные группы ресурсов (rotating).** Позволяют осуществлять распределение групп ресурсов по узлам в кластере. При запуске узла он пытается запустить все неактивные ротационные группы ресурсов, для которых он является узлом с наивысшим приоритетом[1]. При интеграции другого узла в кластер активные ротационные группы ресурсов не перемещаются.

  При наличии нескольких ротационных групп ресурсов в кластере приоритет узла определяется порядком узлов в списке узлов. При подключении каждого узла к кластеру на него перемещается ротационная группа ресурсов, для которой он является узлом с наивысшим приоритетом. Если количество групп ресурсов превышает количество узлов, в подключенный режим переводятся дополнительные группы ресурсов. Если не определено несколько сетей HACMP, то каждый узел получает по одной группе ресурсов для каждой сети.
- **Группы ресурсов с одновременным доступом (concurrent).** Предназначены для приложений с одновременным доступом (работающих в параллельном режиме) – группа ресурсов активна на каждом узле из списка узлов; она переходит в отключенный режим на узле, на котором возникают проблемы, и переходит в подключенный режим на узле, интегрируемом в кластер.

Настраиваемые группы ресурсов

Что в действительности важно для проектировщиков и администраторов HACMP, так это работа групп ресурсов при запуске, перемещении при сбое и возврате после восстановления. HACMP 5.1 поддерживает и «настраиваемые» (custom), и «классические» группы ресурсов, однако начиная с версии 5.2 доступны только «настраиваемые» группы ресурсов. Существуют следующие опции работы настраиваемых групп ресурсов.

Опции запуска (startup options)

Эти опции контролируют работу группы ресурсов при первом запуске.

- **Подключение только на домашнем узле (online on home node only).** Группа ресурсов переводится в подключенный режим, когда ее домашний узел подключается к кластеру. Если домашний узел недоступен, группа ресурсов будет находиться в отключенном состоянии, пока он не будет доступен (рис. 2.15).
- **Подключение на первом доступном узле (online on first available node).** Группа ресурсов переводится в подключенный режим при подключении первого узла из списка узлов к кластеру (рис. 2.16).
- **Подключение на всех доступных узлах (online on all available nodes).** Группа ресурсов будет подключена на всех узлах из списка узлов при их подключении к кластеру (рис. 2.17).
- **Подключение с использованием политики распределения (online using distribution policy).** Группа ресурсов подключается, только если на узле не подключена другая группа ресурсов такого же типа.

---

[1] Другими словами, ротационная группа ресурсов запускается на первом доступном узле в кластере, а каскадная (без атрибута ITO) дожидается запуска своего домашнего узла. *Примеч. науч. ред.*

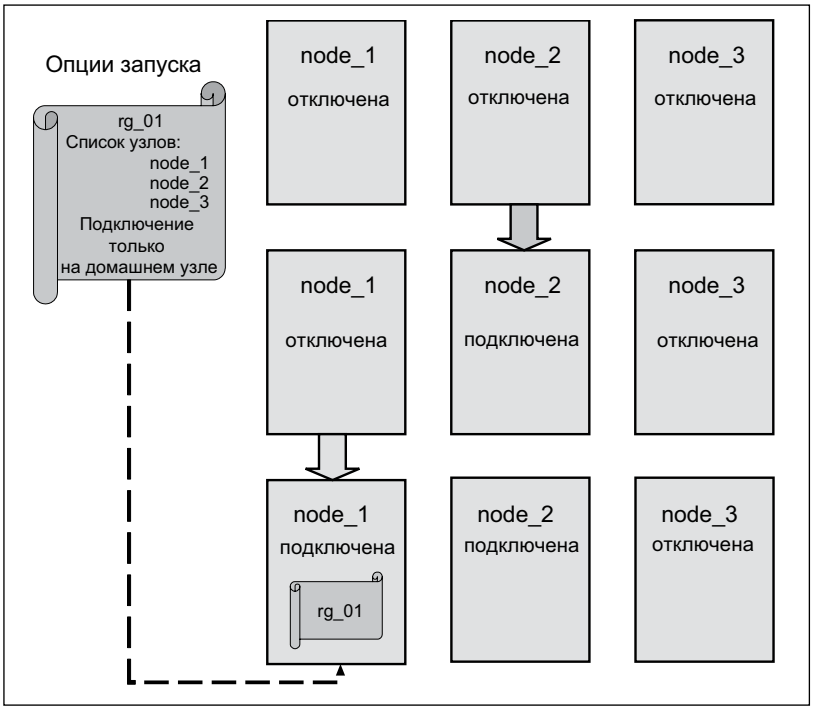

Рис. 2.15. Подключение только на домашнем узле

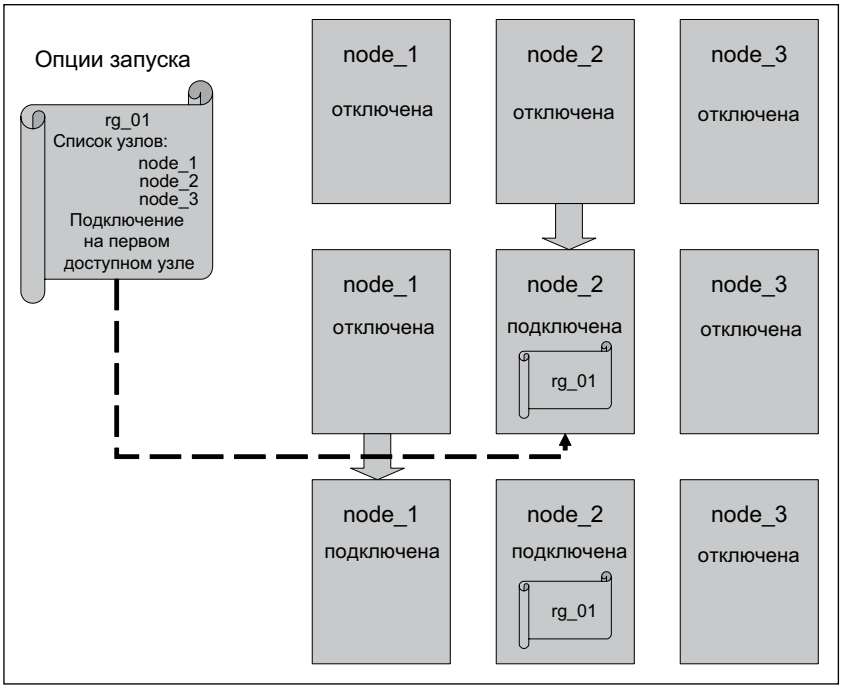

Рис. 2.16. Подключение на первом доступном узле

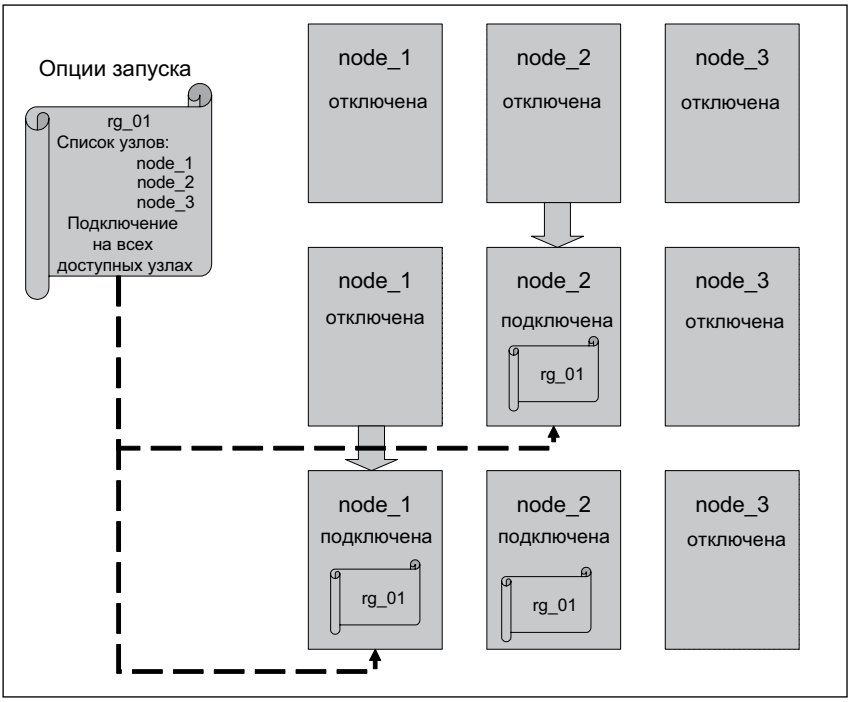

Рис. 2.17. Подключение на всех доступных узлах

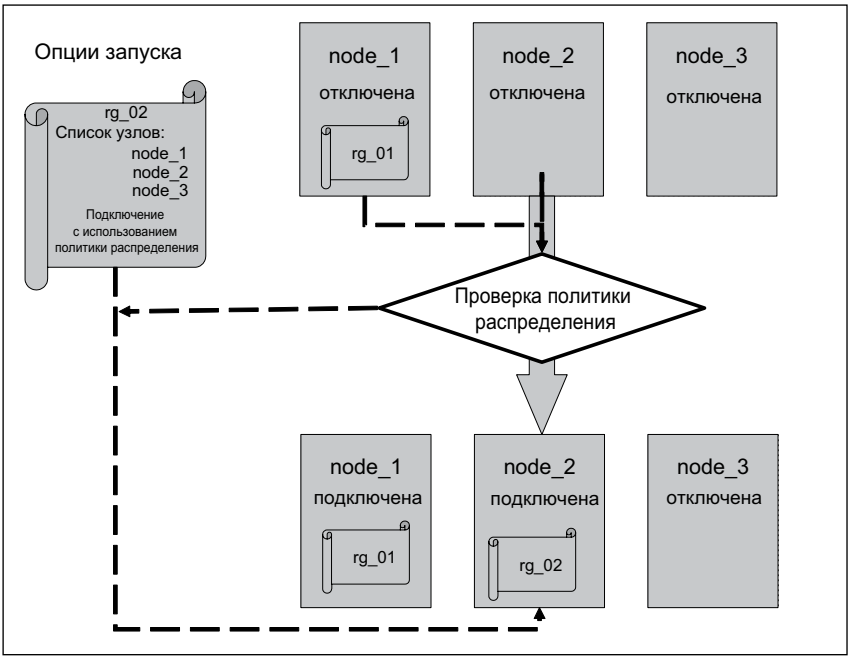

Рис. 2.18. Подключение с использованием политики распределения

Если при подключении узла к кластеру существует несколько групп ресурсов такого типа, HACMP выбирает группу ресурсов с меньшим количеством узлов в списке узлов. Если этот показатель одинаков для всех групп ресурсов, HACMP выбирает первый узел в алфавитном порядке. Однако если один из узлов имеет зависимую группу ресурсов (т. е. является родительским объектом в отношениях зависимости), он будет иметь приоритет (рис. 2.18).

Опции перемещения при сбое (fallover options)

Эти опции контролируют работу группы ресурсов, если HACMP придется переместить ее на другой узел в ответ на событие.

- **Перемещение при сбое на узел из списка со следующим приоритетом (fallover to next priority node in list).** Группа ресурсов выполняет перемещение при сбое на следующий узел в списке узлов группы ресурсов (рис. 2.19).

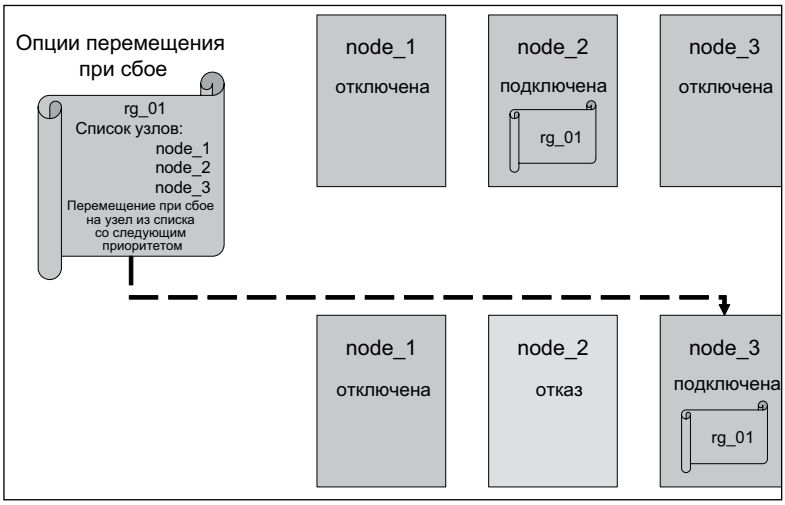

Рис. 2.19. Перемещение при сбое на узел из списка со следующим приоритетом

- **Перемещение при сбое с использованием динамического приоритета узла (fallover using dynamic node priority).** Узел, на который выполняется перемещение, может быть выбран на основе доступности процессора, доступности памяти или наименьшего использования дисков. HACMP применяет RSCT для сбора данных по выбранному показателю со всех узлов в списке узлов, после чего осуществляется перемещение группы ресурсов на узел, который лучше всего соответствует критериям (рис. 2.20).
- **Перевод в отключенное состояние (только на узле с ошибкой) [Bring offline (on error only)].** В случае ошибки группа ресурсов переводится в отключенное состояние. Эта опция предназначена для групп ресурсов с подключением на всех доступных узлах (рис. 2.21).

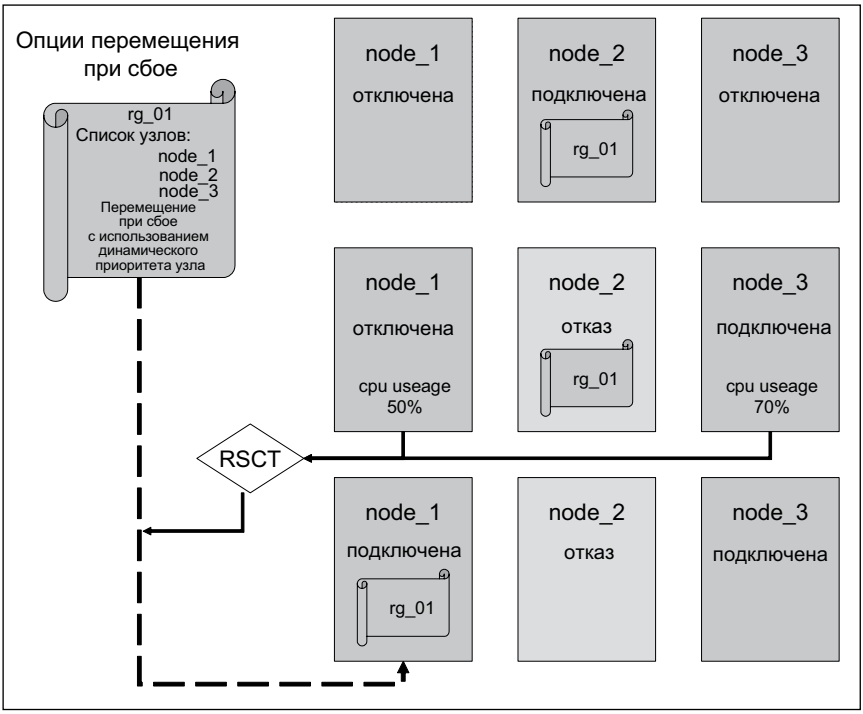

Рис. 2.20. Перемещение при сбое с использованием динамического приоритета узла

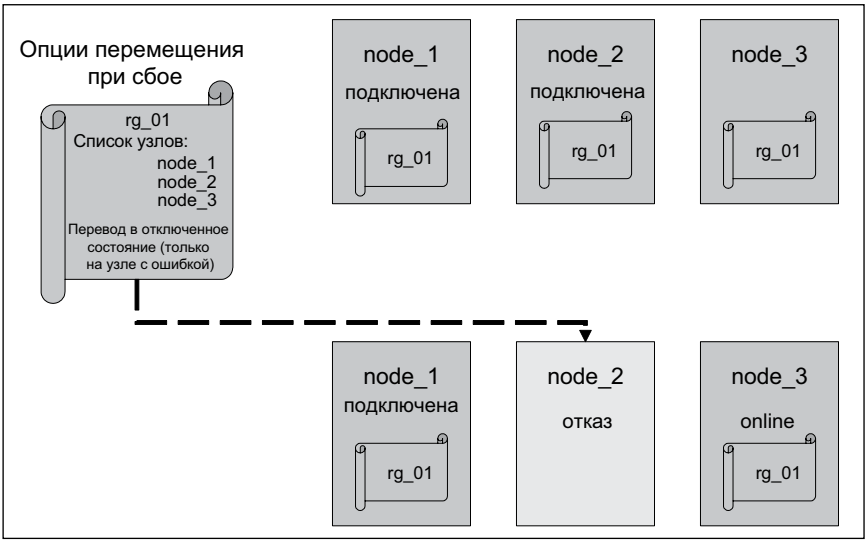

Рис. 2.21. Перевод в отключенное состояние (только на узле с ошибкой)

Опции возврата после восстановления (Fallback options)

Эти опции контролируют работу подключенной группы ресурсов при подключении узла к кластеру.

- **Возврат после восстановления на узел с более высоким приоритетом в списке (fallback to higher priority node in list).** Группа ресурсов осуществляет возврат после восстановления на узел с более высоким приоритетом при его подключении к кластеру (рис. 2.22).
- **Без выполнения возврата после восстановления (never fallback).** Группа ресурсов не переносится на узел с более высоким приоритетом при его подключении к кластеру. Эта опция должна использоваться в группах ресурсов с подключением на всех доступных узлах. См. Рисунок 2.23 на странице 101.

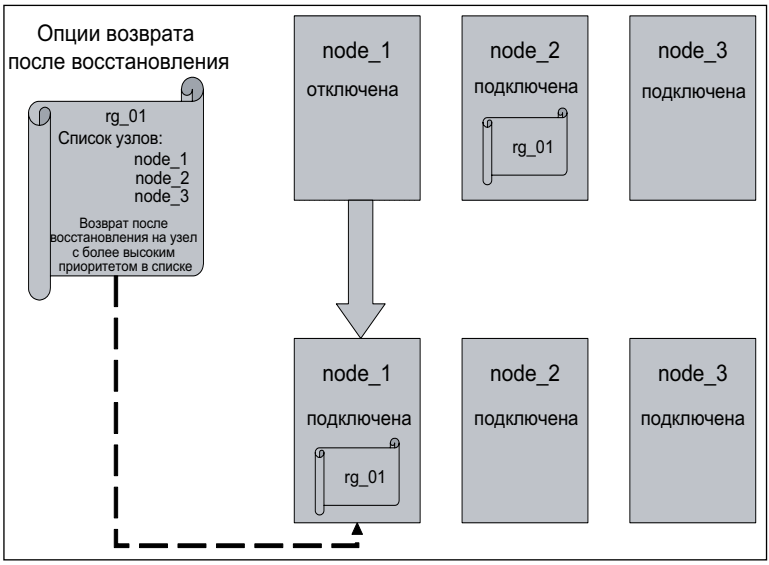

Рис. 2.22. Возврат после восстановления на узел с более высоким приоритетом в списке

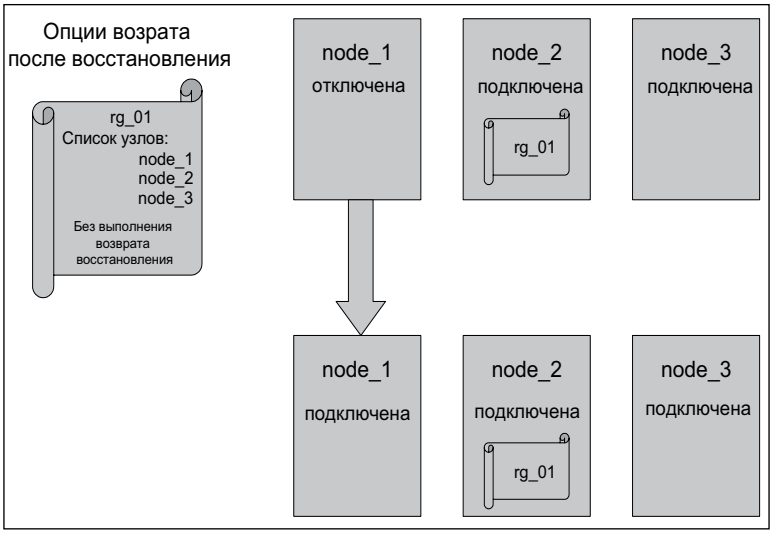

Рис. 2.23. Без выполнения возврата после восстановления

Таблица 2.4. Сравнение работы групп ресурсов

Старая конфигурация	Запуск	Перемещение при сбое	Возврат после восстановления
Каскадная ITO = false CWOF = false DNP = false	Подключение только на домашнем узле	Перемещение при сбое на узел со следующим приоритетом в списке	Возврат после восстановления на узел с более высоким приоритетом в списке
Каскадная ITO = true CWOF = false DNP = false	Подключение на первом доступном узле	Перемещение при сбое на узел со следующим приоритетом в списке	Возврат после восстановления на узел с более высоким приоритетом в списке
Каскадная ITO = false CWOF = true DNP = false	Подключение только на домашнем узле	Перемещение при сбое на узел со следующим приоритетом в списке	Без выполнения возврата после восстановления
Каскадная ITO = true CWOF = true DNP = false	Подключение на первом доступном узле	Перемещение при сбое на узел со следующим приоритетом в списке	Без выполнения возврата после восстановления
Каскадная ITO = false CWOF = false DNP = true	Подключение только на домашнем узле	Перемещение при сбое с использованием динамического приоритета узла	Возврат после восстановления на узел с более высоким приоритетом в списке
Каскадная ITO = true CWOF = false DNP = true	Подключение на первом доступном узле	Перемещение при сбое с использованием динамического приоритета узла	Возврат после восстановления на узел с более высоким приоритетом в списке
Каскадная ITO = false CWOF = true DNP = true	Подключение только на домашнем узле	Перемещение при сбое с использованием динамического приоритета узла	Без выполнения возврата после восстановления
Каскадная ITO = true CWOF = true DNP = true	Подключение на первом доступном узле	Перемещение при сбое с использованием динамического приоритета узла	Без выполнения возврата после восстановления
Ротационная	Подключение с использованием политики распределения	Перемещение при сбое на узел со следующим приоритетом в списке	Без выполнения возврата после восстановления
С одновременным доступом	Подключение на всех доступных узлах	Отключение только на узле с ошибкой	Без выполнения возврата после восстановления

Примечание. ITO – Inactive takeover (Переход при неактивности); CWOF – Cascading without fallback (Каскадирование без возврата после восстановления); DNP – Dynamic Node Priority (Динамический приоритет узла).

Табл. 2.4 показывает соответствие опций запуска, перемещения при сбое и возврата после восстановления оригинальным каскадным группам ресурсов, ротационным группам ресурсов и группам ресурсов с одновременным доступом.

Атрибуты группы ресурсов

Работу группы ресурсов можно настроить с использованием таких атрибутов группы ресурсов, как:

- время установления (settling time);
- таймеры отсроченного возврата после восстановления (delayed fallback timers);
- политика распределения (distribution policy);
- динамические приоритеты узлов (dynamic node priorities);
- порядок обработки групп ресурсов (resource group processing order);
- расположение, отменяющее приоритет (Priority override location);
- зависимости групп ресурсов – отношения «родительский объект/дочерний объект» (resource group dependencies – parent/child);
- зависимости групп ресурсов – расположение (resource group dependencies – location).

Таблица 2.5. Атрибуты группы ресурсов и их влияние на работу группы ресурсов

Атрибут	Запуск	Перемещение при сбое	Возврат после восстановления
Время установления	P		
Таймер отсроченного возврата после восстановления			P
Политика распределения	P		
Динамический приоритет узла		P	
Порядок обработки групп ресурсов	P	P	P
Расположение, отменяющее приоритет	P		P
Зависимость групп ресурсов «родительский объект/дочерний объект»	P	P	P
Зависимость групп ресурсов – расположение	P	P	P

Время установления (settling time)

Представляет собой атрибут кластера, влияющий на работу групп ресурсов с политикой запуска, настроенной на подключение на первом доступном узле. Если атрибут не установлен, эти группы ресурсов запускаются на первом узле в группе ресурсов, интегрируемом в кластер. Если время установления для группы ресурсов определено, и узел, интегрируемый в кластер, является узлом с наивысшим приоритетом, группа ресурсов подключается немедленно, в противном случае выполняется ожидание в течение времени установления на случай, если подключится другой узел с более высоким приоритетом.

Это позволяет предотвратить запуск группы ресурсов на первом подключенном узле с низким приоритетом с последующим перемещением на узлы с более высоким приоритетом по мере их подключения.

Таймеры отсроченного возврата после восстановления (delayed fallback timers)

Используются для настройки времени, в которое следует выполнить возврат группы ресурсов. Может быть задано либо выполнение в заданный день и время, либо выполнение в определенное время ежедневно, еженедельно, ежемесячно или ежегодно.

Таймер отсроченного возврата после восстановления обеспечивает возврат группы ресурсов на узел с наивысшим приоритетом в определенное время. Это обеспечивает возникновение небольшого перерыва в обслуживании в удобное для пользователей время. При этом ресурс не должен находиться на узле с наивысшим приоритетом.

Политика распределения (distribution policy)

Политика распределения на основе узлов обеспечивает «подхват» каждым узлом при запуске только одной группы ресурсов с заданным набором политик.

В HACMP 5.2 была реализована также политика распределения на основе сетей, которая обеспечивала подключение только одной группы ресурсов в одной сети и на одном узле, так что узлы с несколькими сетями могли содержать несколько групп ресурсов одного типа.

Динамический приоритет узла (dynamic node prioritiy)

При наличии трех и более узлов в кластере можно осуществлять настройку политики динамического приоритета узла. Для определения того, на какой узел следует переместить группу ресурсов, можно выбрать один из трех показателей:

- максимальный показатель свободной памяти;
- максимальный показатель бездействия процессора;
- минимальная нагрузка на диск.

Диспетчер кластера ведет таблицу значений этих показателей для каждого узла в кластере. Так что на момент перемещения при сбое диспетчер кластера может быстро определить, какой узел лучше всего соответствует требуемым критериям. Эти значения обновляются каждые 2 мин., если только не отсутствует доступ к узлу; в этом случае предыдущие значения остаются без изменений.

> **Важно!** При определении максимального показателя свободной памяти HACMP вычисляет показатель используемой виртуальной памяти, а не показатель применяемой физической памяти.

Для вывода текущих значений используется команда **lssrc -ls clstrmgrES**.

Пример 2.1. Проверка значений динамического приоритета узлов с точки зрения clstrmgrES

```
odin:/# lssrc -ls clstrmgrES
Current state: ST_STABLE
sccsid = "@(#)36   1.135.1.37   src/43haes/usr/sbin/cluster/hacmprd/main.C,
hacmp.pe, 51haes_r530, r5300525a 6/20/05 14:13:01"
i_local_nodeid 1, i_local_siteid 1, my_handle 2
ml_idx[1]=0     ml_idx[2]=1     ml_idx[3]=2
```

```
There are 0 events on the Ibcast queue
There are 0 events on the RM Ibcast queue
CLversion: 8
Example 2-1cluster fix level is "0"
The following timer(s) are currently active:
Current DNP values
DNP Values for NodeId - 1  NodeName - frigg
    PgSpFree = 0  PvPctBusy = 0  PctTotalTimeIdle = 0.000000
DNP Values for NodeId - 2  NodeName - odin
    PgSpFree = 130258  PvPctBusy = 0  PctTotalTimeIdle = 99.325169
DNP Values for NodeId - 3  NodeName - thor
    PgSpFree = 0  PvPctBusy = 0  PctTotalTimeIdle = 0.000000
```

Порядок обработки групп ресурсов (resource group processing order)

При попытке подключения узлом нескольких групп ресурсов по умолчанию происходит объединение всех ресурсов в одну большую группу ресурсов с последующей их обработкой как одной «группы ресурсов». Это называется *параллельной обработкой*, хотя и не является действительной параллельной обработкой, так как обработка выполняется в едином потоке.

Такой режим работы по умолчанию может быть изменен, и для определенных групп ресурсов может быть задана последовательная обработка путем определения списка последовательной активизации. Этот список определяет порядок обработки на определенном узле, а не по всем узлам. При последовательной обработке происходит следующее:

- заданные группы ресурсов будут обрабатываться по порядку;
- группы ресурсов, содержащие только NFS-подключения, будут обрабатываться параллельно;
- остальные группы ресурсов будут обрабатываться по порядку;
- освобождение ресурсов будет осуществляться в обратном порядке.

Расположение, отменяющее приоритет (Priority override location, POL)

При перемещении группы ресурсов на другой узел или сайт, ее отключении или подключении администратором устанавливается расположение, отменяющее приоритет (POL) для узла и группы ресурсов. Так как такое действие не согласуется с установленным режимом работы групп ресурсов, задается атрибут POL, чтобы остановить немедленный возврат группы ресурсов на соответствующий узел. Этот атрибут замещает атрибут «sticky» из предыдущих версий. Группы ресурсов с конфигурацией подключения на всех доступных узлах могут подключаться и отключаться на отдельных узлах без использования атрибута POL.

Атрибут POL может быть задан как:

- **Постоянный.** Продолжает действовать после перезапуска служб кластера на всех узлах в кластере.
- **Непостоянный**: Действует только до перезапуска служб кластера на всех узлах в кластере. После перезагрузки кластера группа ресурсов возвращается к стандартному режиму работы.

Примечание. При перемещении группы ресурсов с политикой «без выполнения возврата после восстановления» устанавливается атрибут POL и для группы ресурсов выполняется возврат на этот узел, пока атрибут POL не будет удален.

При перемещении подключенной группы ресурсов на другой узел предлагаются следующие варианты:

- **Список узлов (Node list).** Узел, на который следует переместить группу ресурсов. Этот узел будет задан в качестве атрибута POL.
- **Восстановление порядка приоритетов узлов (Restore node priority order).** Группа ресурсов перемещается на узел с наивысшим приоритетом, и атрибут POL не устанавливается.
- Эта информация хранится в файле **/usr/es/sbin/cluster/etc/clpol** на каждом узле в кластере.

Зависимости групп ресурсов (resource group dependencies)

Может быть задано сочетание двух типов зависимостей групп ресурсов:

- зависимость «родительский объект/дочерний объект»;
- зависимости расположения.

Отношения «родительский объект / дочерний объект» между группами ресурсов предназначены для многоуровневых приложений, где одна или несколько групп ресурсов не могут быть успешно запущены, пока определенная группа ресурсов не станет активной. При определении отношения «родительский объект/дочерний объект» родительская группа ресурсов должна быть подключена до подключения любой дочерней группы ресурсов на каком-либо узле. При необходимости отключения родительской группы ресурсов сначала нужно отключить дочерние группы.

Может быть задано до трех уровней зависимости, т. е. родительская группа ресурсов может иметь дочерние группы ресурсов, которые, в свою очередь, являются родительскими объектами по отношению к другим группам ресурсов. Однако циклические зависимости не допускаются.

Рис. 2.24 иллюстрирует пример, где группа ресурсов *2* имеет две дочерние группы ресурсов, одна из которых также имеет свою дочернюю группу ресурсов. Таким образом, группа ресурсов *2* должна быть подключена прежде, чем можно будет подключить группы ресурсов *3* и *4*. Подобным образом группа ресурсов *4* должна быть подключена прежде, чем можно будет подключить группу ресурсов *5*. Группа ресурсов *3* имеет две родительские группы ресурсов (*1* и *2*), которые должны быть подключены прежде, чем можно будет ее подключить.

Так как HACMP запускает приложения в фоновом режиме (при этом зависание скрипта не останавливает работу HACMP), важно, чтобы работали мониторы запуска приложений для родительских групп ресурсов в каждой зависимости «родительский объект/дочерний объект». Как только монитор (или мониторы) запуска приложения подтвердят успешный запуск приложения, можно будет начинать обработку дочерних групп ресурсов.

В группах ресурсов HACMP 5.3 также могут быть определены зависимости расположения. Возможны следующие варианты:

- **Подключение на том же узле (online on same node)**: Для заданных групп ресурсов запуск, перемещение при сбое и возврат после восстановления всегда осуществляются на одном и том же узле, т. е. узлы осуществляют перемещение набора групп. Группа ресурсов с зависимостью может быть подключена только на том узле, на котором уже подключены другие группы ресурсов из того же набора, если только она не является первой подключаемой группой ресурсов в наборе.

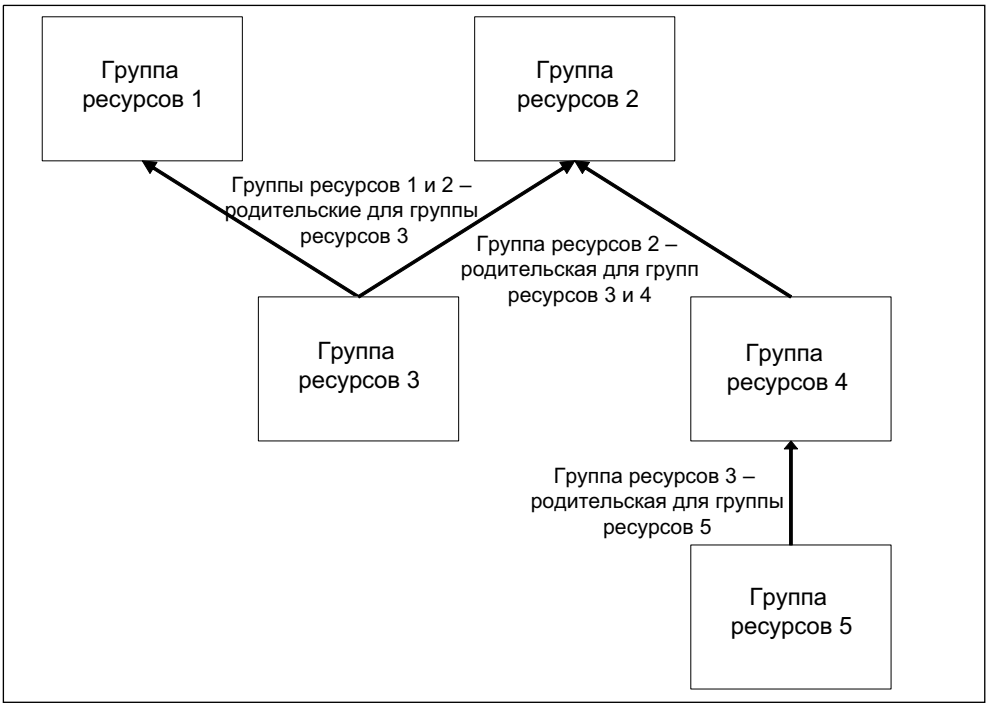

Рис. 2.24. Отношения «родительский объект / дочерний объект»

- **Подключение на разных узлах (online on different nodes).** Для заданных групп ресурсов запуск, перемещение при сбое и возврат после восстановления осуществляются на разных узлах. Группам ресурсов назначается приоритет, так что группы ресурсов с более высоким приоритетом обслуживаются первыми и сохраняются в подключенном состоянии при ограниченном количестве узлов. Группы ресурсов с низким приоритетом отключаются, если группа ресурсов с более высоким приоритетом не имеет узла. Группы ресурсов со средним приоритетом не отключаются. Группа ресурсов с такой зависимостью может быть подключена только на узле, на котором не подключены другие группы ресурсов, являющиеся частью этой зависимости.
- **Подключение на том же сайте (online on same site).** Заданные группы ресурсов всегда подключаются на одном и том же сайте.

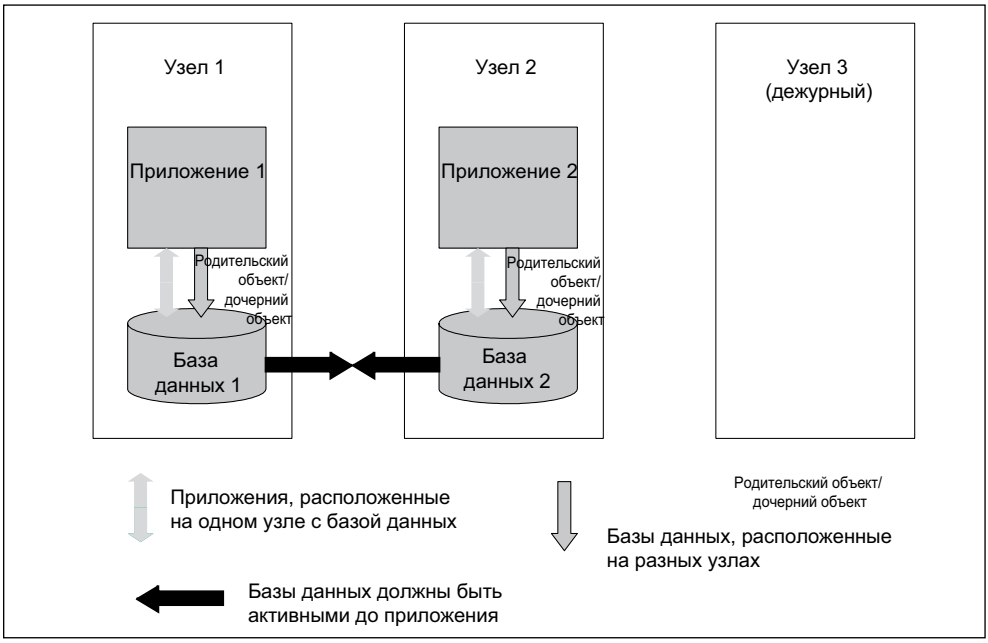

Рисунок 2.25. Зависимости групп ресурсов

Группа ресурсов с этой зависимостью может быть подключена только на том же сайте, на котором уже подключены другие группы ресурсов с этой зависимостью, если только она не является первой подключаемой группой ресурсов с этой зависимостью.

На рис. 2.25 представлен пример кластера из трех узлов с двумя базами данных и двумя приложениями. Приложения не могут быть запущены до подключения баз данных, поэтому устанавливается зависимость «родительский объект/дочерний объект». В целях производительности базы данных должны располагаться на разных узлах, а приложения должны располагаться на одних узлах с базами данных, поэтому устанавливаются зависимости расположения.

Для установки или вывода зависимостей групп ресурсов можно воспользоваться командой **`clrgdependency`**, как показано в примере 2.2.

Пример 2.2. Изменение и проверка зависимостей групп ресурсов

```
odin:># clrgdepdency -t [PARENT_CHILD | NODECOLLOCATION | ANTICOLLOCATION |
SITECOLLOCATION ] -sl
odin:># clrgdepdency -t PARENT_CHILD -sl
#Parent                Child
rg1                    rg2
rg1                    rg3

odin:># clrgdepdency -t NODECOLLOCATION -sl

odin:># clrgdepdency -t ANTILOCATION -sl
```

```
#HIGH:INTERMEDIATE:LOW
rg01::rg03frigg

odin:># clrgdepdency -t SITECOLLOCATION -sl
rg01 rg03 frigg
```

Другой способ проверки состоит в использовании команды **odmget HACMPrg_loc_dependency**.

Управление группой ресурсов

Для групп ресурсов могут выполняться следующие операции:

- **Подключение (online).** Группа ресурсов может быть подключена на узле из списка узлов группы ресурсов. Группа ресурсов должна быть до этого отключена, если она не является группой ресурсов, подключаемой на всех доступных узлах.
- **Отключение (offline).** Группа ресурсов может быть отключена на определенном узле.
- **Перемещение на другой узел с подключением.** Группа ресурсов, являющаяся подключенной на одном узле, может быть отключена и перемещена на другой узел из списка узлов группы ресурсов и подключена. Это может включать и перемещение группы ресурсов на другой сайт.

Атрибут (priority override location) устанавливается в соответствии с приведенным выше описанием.

Некоторые изменения являются недопустимыми:

- родительская группа ресурсов не может быть отключена или перемещена, если существует дочерняя группа ресурсов в подключенном состоянии;
- дочерняя группа ресурсов не может быть запущена, пока не будет подключена родительская группа ресурсов.

Состояния групп ресурсов

В HACMP 5x способ обработки отказов групп ресурсов был изменен и, по сути, вмешательство оператора не всегда является обязательным.

Если узел при подключении к кластеру не может перевести группу ресурсов в подключенное состояние, группа ресурсов остается в ошибочном состоянии (ERROR). При возникновении отказа, если группа ресурсов не настроена на подключение на всех доступных узлах, HACMP попытается перевести группу ресурсов в подключенное состояние на другом активном узле из списка узлов группы ресурсов.

Начиная с HACMP 5.2 каждый узел, подключаемый к кластеру, автоматически пытается подключить все группы ресурсов, находящиеся в ошибочном состоянии.

Если узлу не удалось выполнить «подхват» группы ресурсов во время перемещения при сбое, группа ресурсов помечается как «восстанавливаемая» (recoverable), и HACMP попытается подключить группу ресурсов на других узлах из списка узлов группы ресурсов. Если это не удается сделать на всех узлах, группа ресурсов остается в ошибочном состоянии.

При отказе сети на определенном узле HACMP определяет, какие группы ресурсов были затронуты (которые имели сервисные IP-метки в этой сети), после чего пытается выполнить подключение на другом узле. В случае отсутствия узлов с требуемыми сетевыми ресурсами группы ресурсов остаются в ошибочном состоянии. Если какие-либо интерфейсы становятся доступными, HACMP решает, какие группы ресурсов в ошибочном состоянии могут быть подключены, после чего пытается их подключить.

> **Совет.** Если требуется отменить автоматический режим подключения группы ресурсов в ошибочном состоянии, нужно указать, что она должна оставаться отключенной на узле, установив для нее POL.

Выборочные перемещения при сбое (selective fallovers)

- Отказ интерфейса. Если возможно, HACMP заменяет интерфейсы; в противном случае выполняется перемещение группы ресурсов на узел с наивысшим приоритетом с доступным интерфейсом, и в случае неуспешного перемещения группа ресурсов переводится в ошибочное состояние.
- Отказ сети:
  - локальный – перемещение затронутых групп ресурсов на другой узел;
  - глобальный – вызов события node_down для всех узлов.
- Отказ приложения. Если монитор приложения указывает на отказ приложения, то в зависимости от конфигурации HACMP попытается сначала перезапустить приложение на том же узле (обычно выполняется три попытки), затем, если попытки были неудачными, HACMP перемещает группу ресурсов на другой узел, и, если это тоже не удается, группа ресурсов переводится в ошибочное состояние.
- Отказ коммуникационного канала:
  - HACMP пытается переместить группу ресурсов на другой узел;
  - если настроено выборочное перемещение при сбое для группы томов при возникновении ошибки «LVM_SA_QUORCLOSE», то HACMP пытается переместить затрагиваемые группы ресурсов на другой узел.

2.5 Подключаемые модули HACMP

Подключаемые программные модули для HACMP содержат примеры скриптов, позволяющие выполнить конфигурирование следующих служб в составе кластера высокой доступности:

- сервер имен;
- сервер печати;
- сервер DHCP.

Каждый модуль содержит скрипты запуска и остановки приложений, скрипты мониторов приложений и скрипты удаления. В состав также входит скрипт, подтверждающий наличие корректных файлов конфигурации в общей файловой системе. Каждый подключаемый модуль содержит файл README с подробной информацией. Они находятся в каталоге /usr/es/sbin/cluster/plugin/<имя_модуля>.

2.6 Возможности (HACMP 5.1, 5.2 и 5.3)

В этом разделе перечисляются некоторые новые возможности и усовершенствования, а также то, что больше не поддерживается.

2.6.1 Новые возможности

Усовершенствования внутрикластерной связи в демоне clinfo. Демон clinfo теперь содержит информацию о версии и имеет новый файл журнала **/tmp/clinfo.debug**.

В демоне диспетчера кластера были добавлены функциональные возможности SMUX peer daemon (clsmuxpd), так что выполнение SNMP-запросов возможно даже при неактивном кластере. Были созданы два новых состояния: not_configured и not_synced.

Диспетчер кластера теперь имеет два файла журнала:

/tmp/clstrmgr.debug	**файл журнала** с настраиваемым расположением, содержащий стандартную регистрируемую информацию диспетчера кластера;
/tmp/clsmuxtrmgr.debug	**новый файл** журнала, предназначенный для трассировки новой функции SNMP диспетчера кластера.

Усовершенствования верификации кластера

Автоматическая верификация и синхронизация. HACMP верифицирует конфигурацию узлов при запуске (либо на первом узле в кластере, либо при подключении к активному кластеру). Выполняется верификация (и, при необходимости, коррекция) следующих условий:

- согласованность количества экземпляров RSCT;
- соответствие конфигурации IP-интерфейсов заданной в RSCT;
- отключение автоматической активизации для общих групп томов;
- отключение функции автоматического монтирования файловых систем.

Если конфигурация подключаемых узлов не соответствует конфигурации работающего кластера, она будет синхронизирована с одним из работающих узлов.

Верификация также выявляет потенциальные единые точки отказа, которые ранее выявлялись только при автоматическом уведомлении об ошибках.

Дополнительная верификация кластера

HACMP также выполняет следующие дополнительные проверки:

- одинакова ли версия RSCT на всех узлах;
- одинаковы ли настройки MTU на каждом IP-интерфейсе, а также согласованы ли параметры AIX и HACMP для IP-интерфейсов;
- согласованы ли сетевые опции AIX, используемые в HACMP и RSCT, на всех узлах;
- согласованы ли группы томов и PVID на узлах, являющихся членами владеющей группы ресурсов;

- если установлен HACMP/XD, политика управления сайтом не может быть настроена на игнорирование (ignore);
- если установлен HACMP/XD, то выполняется верификация конфигурации GeoRM, PPRC или GLVM.

Автоматическое наполнение файла clhosts

Файл clhosts, используемый многими программами мониторинга, имеет две версии:

- **Серверная версия.** Этот файл находится на всех узлах в каталоге /usr/es/sbin/cluster/etc и определяет добавление элемента 127.0.0.1 при установке HACMP.
- **Клиентская версия.** Файл clhosts.client находится в каталоге /usr/es/sbin/cluster/etc и наполняется при верификации кластера всеми адресами и метками каждого интерфейса и определенным сервисным IP-адресом. При этом версии с отметками времени сохраняются.

Файл определения кластера в формате XML

Формат XML является наиболее распространенным форматом для файлов определения кластера, создаваемых пользователем, и файлов системы автоматизированного планирования (Online Planning Worksheets). Для преобразования существующих файлов снимков кластера в XML-файл определения кластера можно использовать SMIT.

Тома OEM и Veritas и интеграция файловой системы

Теперь HACMP может без сложностей осуществлять управление группами томов OEM и соответствующими файловыми системами. Эта функция означает, что диски, тома и файловые системы OEM можно включить в группу ресурсов HACMP. Для этого могут использоваться либо имеющиеся методы, либо специально разработанные методы.

В частности, HACMP автоматически определяет группы томов, созданные диспетчером томов Veritas с использованием Veritas Foundation Suite (v4.0).

Функция SMS

Был добавлен новый метод удаленного уведомления. Теперь можно отправлять сообщения удаленного уведомления на любой адрес, например на мобильный телефон или на адрес электронной почты.

Зависимости расположения группы ресурсов

Помимо политик, определяющих зависимости типа «родительский объект/дочерний объект» для групп ресурсов, и политики распределения при запуске, HACMP теперь предлагает зависимости в масштабе кластера для групп ресурсов:

- с подключением на одном узле;
- с подключением на разных узлах;
- с подключением на одном сайте.

> **Примечание.** Политика распределения при запуске основана на узлах; в HACMP 5.2 был выбор между политикой на основе узлов и политикой на основе сетей.

Параметр распределения IP-меток/адресов

По умолчанию HACMP распределяет сервисные IP-метки/адреса по доступным интерфейсам. Перед активизацией сервисной IP-метки HACMP определяет количество адресов синонимов, уже существующих на каждом интерфейсе, после чего использует интерфейс с наименьшим количеством синонимов.

- **Без совместного размещения (Anti-Collocation).** Используется по умолчанию; HACMP распределяет сервисные IP-метки по всем загрузочным IP-интерфейсам в одной сети HACMP на узле.
- **С совместным размещением (Collocation).** HACMP размещает все сервисные IP-адреса на одном загрузочном IP-интерфейсе.
- **С совместным размещением и с постоянной меткой (Collocation with persistent label).** HACMP размещает все сервисные IP-адреса на загрузочном IP-интерфейсе, содержащем постоянную IP-метку синонима. Это может быть полезно в средах с VPN и брандмауэром, где только один интерфейс имеет внешний выход.
- **Без совместного размещения и с постоянной меткой (Anti-Collocation with persistent label).** HACMP размещает все сервисные IP-метки по всем загрузочным IP-интерфейсам в одной логической сети, не содержащим постоянную IP-метку синонима. Если другие интерфейсы недоступны, сервисные IP-метки совместно используют адаптер с постоянной IP-меткой синонима.

HACMP/XD

Параллельная обработка основного и дополнительного экземпляров реплицируемых групп ресурсов HACMP/XD выполняется по умолчанию, однако может быть задана и последовательная обработка. Процессы DARE и rg_move поддерживают параллельную обработку между сайтами.

Могут быть заданы политики управления сайтом при запуске, перемещении при сбое и возврате после восстановления, как для основного, так и для дополнительного экземпляра группы ресурсов.

- Межсайтовые операции группы ресурсов с одновременным доступом (подключение на всех доступных узлах) могут быть совмещены с политикой сайта без одновременного доступа.
- Могут быть заданы отношения зависимости типа «родительский объект/дочерний объект».
- Может использоваться политика распределения при запуске.
- Поддерживаются группы ресурсов как с совместным размещением, так и без совместного размещения.
- При верификации кластера также происходит верификация HACMP/XD-конфигураций, однако конфигурацию нужно распространять вручную на другие узлы, так как часто приходится выполнять большие объемы операций настройки.

Усовершенствования безопасности в WebSMIT

Осуществляется подтверждение параметров, передаваемых в WebSMIT, перед выполнением.

Средства аутентификации WebSMIT в большей мере интегрированы с механизмами аутентификации AIX.

Программы Smart assist для HACMP

HACMP теперь поддерживает:

- Smart assist для WebSphere – хотя этот инструмент поддерживался и в версии 5.2, поддержка была обновлена;
- Smart assist для DB2 – включает поддержку мониторинга и восстановления для DB2 Universal Database™ Enterprise Server Edition;
- Smart assist для Oracle – обеспечивает помощь при установке сервера приложений Oracle 10g.

2.6.2 Неподдерживаемые возможности

- cllockd и cllockdES теперь не поддерживаются.
- clinfo теперь не использует общую память, вместо этого применяются очереди сообщений.
- clsmuxpd теперь не поддерживается, и его функции включены в диспетчер кластера.
- Больше не поддерживаются каскадные группы ресурсов, ротационные группы ресурсов и группы ресурсов с одновременным доступом.
- Подсистема управления событиями (event management) была заменена подсистемой RSCT Resource Monitoring and Control (RMC).
- cldiag больше не поддерживается в командной строке.
- clverify больше не поддерживается в командной строке.

2.7 Ограничения

Этот раздел описывает некоторые наиболее распространенные ограничения, свойственные HACMP. Эти ограничения представлены в табл. 2.6.

Таблица 2.6. Ограничения HACMP

Компоненты	Максимальное количество, поддерживаемое кластером
Узлы	32
Группы ресурсов	64
Сети	48
Сетевые интерфейсы, устройства и метки	256
Ресурсы кластера	Несмотря на то что clinfo может обслуживать не более 128, кластер может содержать больше
Зависимости «родительский объект/дочерний объект»	Не больше трех уровней

Окончание табл. 2.6

Компоненты	Максимальное количество, поддерживаемое кластером
Сайты	2
Интерфейсы	7 интерфейсов на узел на сеть
Мониторы приложений в сайте	128
netmon.cf	30 имен или IP-адресов
Постоянные IP-синонимы	один на узел на сеть
XD-сети	1 на кластер
GLVM-режимы	Синхронные, без одновременного доступа
GLVM-устройства	Все PV, поддерживаемые AIX, должны быть одинаковыми (локальными или удаленными); улучшенный режим одновременного доступа не поддерживается
Динамическая реконфигурация	Невозможна в HACMP/XD:HAGEO и HACMP/XD:GLVM

Требования подсетей

Таблица маршрутизации ядра AIX 5L поддерживает наличие нескольких маршрутов к одному пункту назначения. Если несколько совпадающих маршрутов имеют одинаковый коэффициент, каждый из маршрутов подсетей используется поочередно. При этом в HACMP возникает следующая проблема: если один узел имеет несколько интерфейсов, совместно использующих один маршрут, то HACMP не имеет средства определения его состояния.

Поэтому мы рекомендуем, чтобы каждый интерфейс на узле относился к отдельной подсети, чтобы можно было осуществлять мониторинг каждого интерфейса. Альтернативный вариант состоит в использовании мониторинга пульса через синонимы.

2.8 Аспекты хранения

Этот раздел содержит список некоторых наиболее часто используемых подсистем хранения и соответствующих программ управления, их характеристики, а также описание возможностей управления хранением в HACMP.

Подсистемы хранения серии IBM DS4xxx

Эти устройства прежде назывались FAStT-серверами (Fiber Attach Storage Server, сервер хранения с оптоволоконным подключением). HACMP поддерживает несколько различных моделей подсистем хранения DS4xxx. Описание всех моделей выходит за рамки данной книги.

Мы покажем принципы настройки хранилища DS4xxx на примере сервера хранения DS4500.

Сервер хранения DS4500

Сервер хранения DS4500 поддерживает прямое подключение до четырех узлов, каждый из которых содержит по два адаптера и которые предназначены для обеспечения максимальной избыточности как на стороне узла, так и на стороне хранилища. При исполь-

зовании внешних FC-коммутаторов в сочетании с сервером хранения DS4500 можно подключить к серверу хранения DS4500 до 64 узлов (с двумя адаптерами каждый).

Перед конфигурированием хранилища DS4500 необходимо удостовериться в наличии всех аппаратных и кабельных подключений, обязательных для конфигурации. Дополнительные сведения о кабельных подключениях DS4500 см. в руководстве **IBM TotalStorage DS4500 Fibre Channel Storage Server Installation Guide**, GC26-7530.

Программное обеспечение DS4xxx Storage Manager

Единственный способ конфигурирования хранилища DS4500 состоит в использовании программного обеспечения DS4xxx Storage Manager, которое доступно в большинстве популярных операционных систем, в частности в AIX, Linux и Windows® XP/2000. Используя DS4xxx Storage Manager, можно выполнить конфигурирование поддерживаемых уровней RAID, логических устройств и разделов. К поддерживаемым уровням RAID относятся RAID-0, RAID-1, RAID-5 и RAID 0+1.

В DS4xxx Storage Manager отсутствует опция конфигурирования RAID-10. При выборе RAID-1 с несколькими дисками DS4xxx Manager обеспечивает чередование и зеркальное отображение данных.

Это средство дает возможность пользователям форматировать логические диски в соответствии с требованиями операционных систем. Существует несколько версий Storage Manager.

К новым возможностям, поддерживаемым DS4xxx Storage Manager, относятся:

- FlashCopy®. Логический диск FlashCopy представляет логический мгновенный образ другого логического диска, называемого *базовым логическим диском*, находящегося в подсистеме хранения. FlashCopy представляет собой логический аналог полной физической копии, однако его создание происходит быстрее и требует меньше дискового пространства (20 % от первоначального логического диска).
- Удаленное зеркальное отображение (remote mirror). Опция удаленного зеркального отображения используется для репликации данных между подсистемами хранения на удаленные расстояния в подключенном состоянии и в реальном времени.
- Volumecopy (копирование тома). Опция **volumecopy** представляет собой механизм репликации данных с логических дисков в пределах массива хранения на основе микропрограммного обеспечения. Пользователи отправляют запросы **volumecopy**, указывая два совместимых диска. Один диск является источником, а другой – целевым диском. Запрос **volumecopy** является постоянным, так что результаты процесса копирования могут быть сообщены пользователю.
- Разделение хранилища (storage partitioning). Дает возможность пользователю представить все тома хранилища в SAN через различные разделы путем сопоставления томов хранилища номерам LUN, где каждый раздел соответствует LUN от 0 до 255. Такое сопоставление томов или LUN относится только к порту или портам, настроенным на доступ к этому LUN. Эта функция также позволяет осуществлять поддержку одновременного подключения нескольких узлов, использующих различные операционные системы, и их подсистем дискового хранения к одному серверу хранения DS4xxx.

Сервер хранения Enterprise Storage Server (ESS/Shark)

Серверы хранения IBM Enterprise Storage Server (ESS) представляют собой второе поколение дисковой системы хранения Seascape®, обеспечивающей наилучшие в отрасли показатели доступности, производительности, управляемости и масштабируемости. Уровни RAID в ESS предопределены в виде нескольких конфигураций и имеют ограниченные возможности изменения. Доступны следующие уровни RAID: RAID-1, RAID-5 и RAID 0+1.

Сервер IBM Enterprise Storage Server (ESS) не просто осуществляет общее хранение между различными промышленными платформами; он может повысить производительность, доступность, масштабируемость и управляемость ресурсов хранения предприятия с использованием различных мощных средств. Некоторые из средств по названию сходны со средствами FAStT/DS4xxx Storage, однако их технические концепции значительно различаются. Ниже приведены некоторые из возможностей.

- FlashCopy. Обеспечивает возможность быстрого дублирования данных. Эта опция позволяет устранить необходимость в остановке приложений на длительные периоды времени для выполнения операций резервного копирования и восстановления.
- Одноранговое удаленное копирование (peer-to-peer remote copy). Эта функция поддерживает наличие синхронной копии данных (всегда соответствующей основной копии) в удаленном расположении. Эту резервную копия данных можно легко использовать для восстановления после сбоя в основной системе без потери каких-либо транзакций; эта опция позволяет обеспечить бесперебойную работу приложений электронного бизнеса.
- Расширенная удаленная копия (Extended remote copy, XRC). Эта функция обеспечивает наличие копии данных в удаленном расположении (которое может быть подключено с использованием телекоммуникационных линий на неограниченном расстоянии), используемой в случае отказа основной системы хранения. В функции XRC в ESS реализована полная поддержка незапланированных отключений. В случае отказа телекоммуникационной связи эта опция позволяет быстро выполнить синхронизацию удаленной резервной копии без дублирования всех данных с основного расположения в целях защиты полного аварийного восстановления.
- Настраиваемые тома. Позволяют выполнять определение томов различных размеров для высокопроизводительных серверов, что дает администраторам возможность настраивать оптимальную производительность систем.
- Разделение хранилища. Позволяет более эффективно использовать устройства хранения, предоставляя каждому серверу доступ к собственному пулу хранения. Пулы хранения могут совместно использоваться несколькими серверами.

Дополнительные сведения о конфигурировании Enterprise Storage Server см. в руководстве **IBM TotalStorage Enterprise Storage Server Service Guide 2105 Model 750/800 and Expansion Enclosure, Volume 1**, SY27-7635.

Серии IBM TotalStorage DS6000 и DS8000

Эти новые подсистемы хранения также поддерживаются в HACMP, однако на момент написания данной книги у нас не было ни достаточной информации, ни оборудования для тестирования.

Подробные сведения о сериях DS6000 и DS8000 и об их поддержке см. на веб-сайте *http://www-1.ibm.com/servers/eserver/pseries/ha/*

Архитектура SSA (Serial Storage Architecture)

Архитектура SSA является отраслевым стандартом интерфейса, обеспечивающего высокопроизводительное отказоустойчивое подключение устройств хранения. В подсистемах SSA осуществляется мультиплексирование передач на несколько узлов назначения; эффективная пропускная способность также повышается путем пространственного повторного использования отдельных связей. Команды автоматически перенаправляются с одного устройства на другое в кольце, пока не будет достигнуто целевое устройство. Одновременно по кольцу может передаваться несколько команд.

SSA поддерживает RAID-0, RAID-1, RAID-5 и RAID 0+1. Чтобы использовать любую из установок RAID, необходимо следовать правилам колец для дисковых стоек SSA. Для создания RAID-массива необходимо использование специальных соединений на дисках. Дополнительные сведения о конфигурации RAID в IBM SSA см. в руководстве **IBM Advanced SerialRAID Adapters Installation Guide,** SA33-3287.

2.8.1 Общий LVM

В кластере HACMP основным элементом являются данные, используемые приложениями высокой доступности. Эти данные хранятся в объектах диспетчера логических томов (Logical Volume Manager, LVM). Кластеры HACMP используют возможности LVM для обеспечения доступа к этим данным с нескольких узлов. Диспетчер логических томов AIX обеспечивает общий доступ к данным с нескольких узлов. К компонентам общего диспетчера логических томов относятся:

- **общая группа томов** (shared volume group) – группа томов, расположенная полностью на внешних дисках, совместно используемых узлами кластера;
- **общий физический том** (shared physical volume) – диск, расположенный в общей группе томов;
- **общий логический том** (shared logical volume) – логический том, расположенный полностью в общей группе томов;
- **общая файловая система** (shared file system) – файловая система, полностью расположенная на общем логическом томе.

Если вы системный администратор кластера HACMP, вам, возможно, придется выполнять следующие задачи, связанные с LVM:

- создание новой общей группы томов;
- расширение, сокращение, изменение или удаление существующей группы томов;

- создание нового логического тома;
- расширение, сокращение, изменение или удаление существующего логического тома;
- создание новой общей файловой системы;
- расширение, изменение или удаление существующей файловой системы;
- добавление и удаление физических томов.

При выполнении любой из этих задач обслуживания общих компонентов LVM необходимо убедиться в том, что после экспорта и последующего реимпорта группы томов права владения и разрешения были переустановлены.

После экспорта и реимпорта владельцем группы томов является пользователь «root», и эта группа является доступной для группы system.

> **Примечание.** Такое изменение может оказать неблагоприятное воздействие на некоторые приложения, в частности на некоторые серверы баз данных, использующие логические тома прямого доступа, в случае изменения прав владения для логического тома прямого доступа. После выполнения этой последовательности действия необходимо восстановить требуемые значения прав владения и разрешений.

Доступ к общему логическому тому может обеспечиваться в любом из следующих режимов доступа к данным:

- режим неодновременного доступа;
- режим одновременного доступа;
- режим расширенного одновременного доступа.

2.8.2 Режим неодновременного доступа

В среде неодновременного (non-concurrent) доступа HACMP обычно использует файловые системы Journaled File System для управления данными, хотя некоторые приложения баз данных могут обходить файловую систему JFS и осуществлять прямой доступ к логическому тому.

В режиме неодновременного доступа LVM поддерживаются как конфигурации с зеркальным отображением, так и конфигурации без зеркального отображения. Дополнительные сведения по созданию логических томов с зеркальным отображением и без него см. в руководстве **HACMP for AIX 5L V5.3 Planning and Installation Guide**, SC23-4861-06.

Для создания на узле общей группы томов с неодновременном доступом необходимо выполнить следующие действия:

1. Используйте быстрый путь **smitty mkvg.**

2. Используйте заданные по умолчанию значения полей, если только ваш сайт не имеет других особых требований:

- VOLUME GROUP name (Имя группы томов). Имя группы томов в кластере должно быть уникальным.

- Activate volume group AUTOMATICALLY at system restart? (Установить автоматическую активизацию группы томов при перезапуске системы?) Установите значение No, чтобы группа томов активизировалась надлежащим образом скриптами событий кластера.
- ACTIVATE volume group after it is created? (Активизировать ли группу томов после ее создания?). Установите значение Yes.
- Volume Group MAJOR NUMBER [Старший номер (устройства) группы томов]. Необходимо использовать одинаковый старший номер на всех узлах. Используйте команду `lvlstmajor` на всех узлах для определения свободного старшего номера, общего для всех узлов.

Для создания на узле общей файловой системы с неодновременным доступом, необходимо выполнить следующие действия:

1. Используйте быстрый путь **smitty crjfs.**

2. Переименуйте логический том и логический том журналов для файловой системы и группы томов. AIX присваивает имя логического тома каждому создаваемому логическому тому. Примеры имен логических томов: /dev/lv00 и /dev/lv01. В кластере HACMP имя любого общего логического тома должно быть уникальным. Кроме того, журнал файловой системы JFS (jfslog) также представляет собой логический том, требующий уникального имени в кластере.

3. Просмотрите значения следующих полей:

- Mount automatically at system restart? (Осуществлять ли подключение при перезапуске системы?). Убедитесь в том, что для этого поля установлено значение No.
- Start Disk Accounting (Запустить учет использования дисков). Установите для этого поля значение No, если только вы действительно не хотите запустить учет использования дисков.

4. Выполните тестирование только что созданной файловой системы путем ее подключения и отключения.

Импорт группы томов на узле перемещения при сбое

Перед импортом группы томов убедитесь в том, что группа томов деактивизирована на основном узле. После этого можно запустить процесс сбора информации discovery HACMP, при котором собираются сведения обо всех группах томов, доступных на всех узлах.

При импорте группы томов на узле перемещения при сбое выполняется синхронизация определения ODM для группы томов на каждом узле, на котором осуществляется ее импорт.

При добавлении группы томов в группу ресурсов можно выбрать импорт группы томов вручную на узел перемещения при сбое либо автоматический импорт на все узлы перемещения при сбое в группе ресурсов.

Дополнительные сведения об импорте групп томов см. в руководстве **HACMP for AIX 5L V5.3 Planning and Installation Guide,** SC23-4861-06.

Примечание. После импорта группы томов на узле перемещения при сбое необходимо изменить состояние запуска группы томов. Для изменения состояния группы томов в соответствии с требованиями HACMP нужно выполнить следующую команду:
chvg -an -Qn <vgname>
Эта команда отключает автоматическую активизацию при перезапуске системы, а также отключает кворум группы томов.

2.8.3 Режим одновременного доступа

Использование одновременного (concurrent) доступа в HACMP требует установки дополнительного набора файлов. Режим одновременного доступа не поддерживается в файловых системах; вместо этого необходимо использовать логические тома прямого доступа или физические диски.

Создание группы томов с одновременным доступом

Должны быть выполнены установка, конфигурирование и обеспечена доступность физических томов (hdisk*). Проверка состояния дисков выполняется с помощью следующей команды:

```
# lsdev -Cc disk
```

Для использования группы томов с одновременным доступом необходимо выполнить ее создание как группы томов с возможностью одновременного доступа (concurrent capable). Группа томов с возможностью одновременного доступа может быть активизирована как в режиме неодновременного доступа, так и в режиме одновременного доступа.

Для создания группы томов с одновременным доступом, нужно выполнить следующие действия:

1. Введите **smit cl_convg**.
2. Выберите **Create a Concurrent Volume Group**.
3. Введите требуемые значения полей.
4. Нажмите Enter.

Импорт группы томов с возможностью одновременного доступа

Импорт группы томов с возможностью одновременного доступа выполняется с использованием следующей команды:

```
# importvg -C -y vg_name physical_volume_name
```

Имя диска в группе томов указывается в качестве аргумента команды **importvg**. По умолчанию AIX автоматически активизирует группы томов без возможности одновременного доступа при их импорте. AIX не выполняет автоматическую активизацию группы томов с возможностью одновременного доступа при их импорте.

Активизация групп томов с возможностью одновременного доступа в режиме неодновременного доступа

Для создания логического тома необходимо выполнить активизацию группы томов с возможностью одновременного доступа в режиме неодновременного доступа. Для активизации группы томов в режиме неодновременного доступа следует использовать команду **varyonvg**:

```
# varyonvg <vgname>
```

Создание логических томов в группе томов с возможностью одновременного доступа

Существует возможность создания логических томов в группе томов с указанием зеркальных отображений логических томов для обеспечения избыточности данных.

Для создания логических томов в группе томов с возможностью одновременного доступа на исходном узле необходимо выполнить следующие действия:

1. Используйте быстрый путь **smit cl_conlv**.
2. Укажите размер логического тома в виде количества логических разделов.
3. Укажите требуемые значения других доступных опций.
4. Нажмите Enter.

Деактивизация группы томов

После создания логического тома необходимо выполнить деактивизацию (varyoff) группы томов с использованием команды **varyoffvg** так, чтобы ее активизацию можно было выполнить скриптами HACMP. Введите

```
# varyoffvg <vgname>
```

Определение группы томов с одновременным доступом в группе ресурсов HACMP

Для одновременного запуска группы томов с одновременным доступом на всех узлах нужно указать имя группы томов в скрипте запуска HACMP.

При запуске кластера вы можете убедиться в том, что группа томов с одновременным доступом активизирована на всех сконфигурированных узлах.

2.8.4 Группы томов в режиме расширенного одновременного доступа

В HACMP V5.1 была реализована возможность создания и использования групп томов с расширенным одновременным доступом (enhanced concurrent mode, ECM). Их можно использовать как для одновременного, так и для неодновременного доступа. Также можно выполнить преобразование существующих групп томов с одновременным доступом (классических) в группы томов в режиме расширенного одновременного доступа с использованием C-SPOC.

Для групп томов с расширенным одновременным доступом, используемых в среде неодновременного доступа, вместо применения механизма резервирования SCSI HACMP V5.1 использует механизм быстрого перехвата дисков, что обеспечивает быстрый перехват и целостность данных.

Примечание. В HACMP V5.1 быстрый перехват дисков доступен только в AIX 5L V5.2.

Группа томов с расширенным одновременным доступом активизируется на всех узлах в кластере, входящих в группу ресурсов. Однако доступ для изменения данных разрешен только для узла с активной (подключенной) группой ресурсов.

Активная и пассивная активизация в режиме расширенного одновременного доступа

Группа томов с расширенным одновременным доступом может быть активизирована на узле в двух режимах: активном или пассивном.

Активная активизация

В активном состоянии разрешены все высокоуровневые операции. При активизации группы томов с расширенным одновременным доступом на узле в активном состоянии возможны следующие операции:

- операции над файловыми системами, в частности монтирование файловых систем;
- операции над приложениями;
- операции над логическими томами, в частности создание логических томов;
- синхронизация групп томов.

Пассивная активизация

При активизации группы томов с расширенным одновременным доступом в пассивном состоянии LVM обеспечивает своего рода ограждение группы томов на уровне LVM. Узел, содержащий группу томов с пассивной активизацией, допускает выполнение ограниченного количества операций чтения для группы томов:

- доступ только для чтения к специальному файлу группы томов[1];
- доступ только для чтения к первым 4 Кб на всех логических томах, входящих в группу томов[2].

Когда группа томов активизирована в пассивном состоянии, не допускается выполнение следующих операций:

- операций над файловыми системами, в частности монтирования;
- любых операций открытия или записи для логических томов;
- синхронизации групп томов.

---

[1] Имеется в виду VGDA. *Примеч. науч. ред.*

[2] Для доступа к LVCB. *Примеч. науч. ред.*

Создание группы томов с расширенным одновременным доступом

При создании групп томов с одновременным доступом в AIX 5L 5.1 или более поздней версии они автоматически создаются в режиме расширенного одновременного доступа.

Для создания группы томов с возможностью одновременного доступа из командной строки AIX следует использовать команду **mkvg**. Например:

```
# mkvg -n -s 32 -C -y myvg hdisk11 hdisk12
```

При этом произойдет создание группы томов с расширенным одновременным доступом для hdisk11 и hdisk12. Флаги выполняют следующие действия:

-n	не активизировать группу томов при запуске;
-s 32	определяет разделы размером 32 Мб;
-C	создает группу томов с расширенным одновременным доступом;
-y	задает имя группы томов.

2.8.5 Быстрый перехват дисков

Представляет собой новую возможность HACMP V5.1, основное назначение которой заключается в следующем:

- сокращение времени отключения приложений с обеспечением более быстрого перемещения при сбое для группы ресурсов;
- одновременный доступ к группе томов (с сохранением целостности данных);
- использование групп томов с расширенным одновременным доступом (ECM);
- использование RSCT для связи.

Группы томов с расширенным одновременным доступом поддерживают активизацию в активном и пассивном режимах и могут быть включены в группу ресурсов с неодновременным доступом.

Быстрый перехват дисков (fast disk takeover) выполняется программным обеспечением HACMP автоматически. Для всех общих групп томов, созданных в режиме расширенного одновременного доступа и содержащих файловые системы, HACMP активизирует функцию быстрого перехвата дисков. При запуске HACMP на всех узлах в группе ресурсов, совместно использующих одну группу томов с расширенным доступом, эта группа томов активизируется в пассивном режиме. При подключении группы ресурсов на узле, осуществляющем «подхват» ресурсов, группа томов активизируется в активном режиме.

Другие узлы обеспечивают активизацию группы томов в пассивном режиме. В этом случае все изменения в группе томов распространяются автоматически на все узлы в группе томов. Переход из активного в пассивный режим и обратно координируется HACMP при запуске кластера, активизации и перемещении при сбое группы ресурсов, а также при переподключении отказавшего узла к кластеру.

Эти функции имеют следующие требования:

- HACMP V5.1,
- AIX 5L 5.2 или выше,

- bos.clvm.5.2.0.11 или выше,
- APAR IY44237.

Дополнительные сведения о быстром перехвате дисков см. в руководстве **HACMP for AIX 5L V5.3 Planning and Installation Guide,** SC23-4861-06.

2.9 Конфигурация хранилища с общим доступом

Большинство конфигураций HACMP требует использования общего хранилища. К дисковым подсистемам IBM с поддержкой доступа с нескольких узлов относятся SCSI, SSA, ESS и FAStT.

Кроме того, можно использовать устройства и подсистемы хранения сторонних производителей (OEM), хотя большинство из них не сертифицированы IBM для использования в HACMP. Прежде чем использовать такие устройства, следует просмотреть соответствующую информацию на веб-сайтах производителей.

Табл. 2.7 содержит список наиболее часто используемых устройств хранения IBM, которые могут употребляться для общего доступа в кластере HACMP.

HACMP также поддерживает накопители на магнитной ленте с общим доступом (SCSI или FC). Накопители на магнитной ленте с общим доступом могут быть подключены через SCSI или FC. Режим одновременного доступа к ленте **не** поддерживается. Некоторые из поддерживаемых подсистем работы с магнитными лентами перечислены в табл. 2.8.

Таблица 2.7. Внешние подсистемы хранения

IBM 7133 SSA, дисковая подсистема, модели D40 и T40 (дисковые модули до 72.8 Гб и до восьми узлов в кольце SSA).
IBM Enterprise Storage Server (ESS), модели E10, E20, F10 и F20 (поддержка до восьми узлов, использующих интерфейсы SCSI и Fibre Channel через IBM FC/FICON, Feature code: 3021, 3022 и 3023)
IBM 2105-800 (ESS) Total Storage Enterprise Storage Server (FS и SCSI)
IBM Total Storage, модели FAStT 200, 500, 600, 700 и 900.
IBM 2106 Total Storage, серии DS6000 и DS8000

Таблица 2.8. Поддержка накопителя на магнитной ленте

IBM 3583 Ultrium Scalable Tape Library, модели L18, L32 и L72
IBM 3584 Ultra™ Scalable Tape Library, модели L32 и D32
IBM Total Storage Enterprise Tape Drive 3590, модель H11
IBM Magstar® 3590 Tape Drive, модели E11 и B11
IBM 3581 Ultrium Tape Autoloader, модели H17 и L17
IBM 3580 Ultrium Tape Drive, модели H11 и L11

Актуальный список поддерживаемых устройств хранения и накопителей на магнитной ленте см. на веб-сайте IBM

http://www-1.ibm.com/servers/eserver/pseries/ha/

HACMP также можно настроить на работу с подсистемами хранения общего доступа (дисковыми подсистемами и накопителями на магнитной ленте) сторонних

производителей. Список устройств хранения сторонних производителей см. на веб-сайтах соответствующих производителей, а также на веб-сайте Availant

http://www.availant.com/

Конфигурирование хранилища – одна из наиболее важных задач, которые необходимо выполнить, прежде чем выполнять конфигурирование кластера HACMP. Конфигурирование хранилища можно считать частью конфигурирования HACMP.

В зависимости от потребностей приложения и от типа хранилища необходимо решить, сколько узлов в кластере будут иметь доступ к общему хранилищу, а также какие группы ресурсов какие диски будут использовать.

Большая часть подсистем хранения производства IBM поддерживается в HACMP. Дополнительные сведения о поддержке серверов хранения см. в руководстве **HACMP for AIX 5L V5.3 Planning and Installation Guide**, SC23-4861-06.

2.9.1 Требования общего LVM

Планирование общего LVM для кластера HACMP зависит от метода общего доступа к диску и от типа общего дискового устройства. При планировании общего LVM следует учитывать следующие аспекты:

- метод защиты данных;
- метод доступа к хранилищу;
- избыточность оборудования хранилища.

Примечание. HACMP сам по себе не обеспечивает защиту хранилища. Защита хранилища обеспечивается двумя последними системами:

- AIX (зеркальное отображение LVM);
- аппаратный RAID.

В этом разделе содержится информация о методах защиты данных на уровне хранилища, а также о режимах доступа к общим дискам LVM:

- режим неодновременного доступа;
- «классический» режим одновременного доступа (диспетчер одновременного доступа к логическим томам HACMP – clvm);
- режим расширенного одновременного доступа (ECM), новая опция в AIX 5L V5.1 и выше.

2.9.2 Неодновременный доступ, одновременный доступ и расширенный одновременный доступ

В конфигурации с неодновременным доступом только один узел кластера может одновременно осуществлять доступ к общим данным. При перемещении группы ресурсов, содержащей общее дисковое пространство, на другой узел новый узел активизирует диски и проверяет текущее состояние групп томов, логических томов и файловых систем.

В конфигурациях с неодновременным доступом диски могут совместно использоваться как:

- физические тома прямого доступа;
- логические тома прямого доступа;
- файловые системы.

В конфигурации с одновременным доступом данные на дисках одновременно доступны для всех дисков. Этот режим не поддерживает использование файловых систем (JFS или JFS2).

Быстрый перехват диска

В HACMP V5.1 используется новый диспетчер, AIX LVM, с расширенным одновременным доступом. В AIX 5L V5.2 каждая новая группа томов с одновременным доступом должна создаваться в режиме расширенного одновременного доступа.

Только в AIX 5L V5.2 группы томов с расширенным одновременным доступом могут также использоваться для размещения на них файловых систем (общих или необщих). Это позволяет ускорить процесс перехвата общих файловых систем в случае перемещения при сбое.

Группы томов с расширенным одновременным доступом активизируются на всех узлах в группе ресурсов, и доступ к данным координируется HACMP. Только узел с активной группой ресурсов может активизировать группу томов в «активном» режиме; остальные узлы активизируют группу томов в «пассивном» режиме. В «пассивном» режиме не разрешено выполнение высокоуровневых операций над группой томов.

Внимание! При использовании групп ресурсов с опцией быстрого перехвата дисков очень важно иметь избыточные сети и сети, отличные от IP. Это позволяет предотвратить повреждение данных (все-таки группы томов находятся в режиме одновременного доступа) в случае разделения кластера («split-brain»).

Требования LVM

Диспетчер логических томов (Logical Volume Manager, LVM) осуществляет управление хранилищем, координируя постановку в соответствие данных в физических и логических хранилищах. Логическое хранилище может быть расширено и реплицировано, а также может охватывать несколько физических дисков и стоек.

Основные компоненты LVM:

- Физический том (physical volume, PV). Представляет единый физический диск, видимый из AIX (hdisk*). Разделяется на физические разделы, которые представляют физические единицы размещения, используемые LVM.
- Группа томов (volume group, VG). Представляет собой набор физических томов, воспринимаемых AIX как непрерывная адресуемая дисковая область. В HACMP группа томов и ее логические тома могут входить в общую группу ресурсов. Группа томов не может входить в несколько групп ресурсов.

- Физический раздел (physical partition, PP). Представляет собой единицу размещения в группе томов. Физические тома разделяются на физические разделы (при добавлении физического тома в группу томов), и физические разделы используются в логических томах (один, два или три физических раздела на логический раздел[1]).
- Область дескриптора группы томов (Volume group descriptor area, VGDA). Представляет зону на диске, содержащую информацию о размещении хранилища в группе томов. Группе томов из одного диска соответствует две копии VGDA. Двухдисковая группа томов содержит три копии VGDA: две на одном диске и одна на другом. Группа томов, состоящая из трех и более физических томов, содержит по одной копии VGDA на каждом диске в группе томов.
- Кворум. Чтобы активная группа томов оставалась активной, должен быть доступен «кворум» (quorum) областей VGDA (50 % + 1). Кроме того, если в группе томов отключена опция кворума, такая группа томов не может быть активизирована (без использования опции принудительной активизации), если отсутствует одна копия VGDA. При включении опции кворума системный администратор должен знать схему группы томов, чтобы обеспечить целостность данных.
- Логический том (logical volume, LV). Представляет набор логических разделов, для которых AIX обеспечивает доступность в качестве единого хранилища. Логические тома могут использоваться как пространство хранения прямого доступа или как хранилища файловой системы. В HACMP логический том, входящий в группу томов, одновременно становится частью группы ресурсов и не может стать частью другой группы ресурсов.
- Логический раздел (logical partition, LP). Представляет собой единицу размещения пространства для логических томов, а также является логическим представлением физического раздела. AIX LVM позволяет осуществлять сопоставление логических разделов одному, двум или трем физическим разделам для реализации зеркального отображения логических томов.

Примечание. Хотя зеркальное отображение LVM может применяться с любым типом диска, при использовании серверов хранения IBM 2105 ESS или FAStT можете пропустить эту опцию. Эти подсистемы хранения (а также некоторые подсистемы сторонних производителей) обеспечивают собственную избыточность данных с использованием различных уровней RAID.

- Файловые системы. В действительности представляет собой простую базу данных для хранения файлов и каталогов. В AIX файловая система хранится на едином логическом томе. Основными компонентами файловой системы (JFS или JFS2) являются логический том, содержащий данные, журнал файловой системы и драйвер устройств файловой системы. HACMP поддерживает использование JFS и JFS2 в качестве общих файловых систем, с тем условием, что журнал должен находиться на отдельном логическом томе [JFS2 также может содержать встроенные журналы (inline log), но они не поддерживаются в HACMP][2].

---

[1] Два или три – в случае зеркалирования. *Примеч. науч. ред.*

[2] В документе HACMP v.5.3 for AIX Concepts and Facilities Guide SC23-4864-06. Chapter 3, раздел «Filesystems», то, что JFS2 поддерживает встроенные журналы, оценивается как преимущество JFS2, но при этом не указано, что они не поддерживаются в HACMP. См. также примечание в разделе 6.1.2 (к этому изданию). *Примеч. науч. ред.*

► Принудительная активизация групп томов

HACMP V5.1 содержит новую функцию принудительной активизации группы томов на узле. Если в процессе перехвата простое выполнение команды **varyon** для этой группы томов будет неуспешным (из-за отсутствия кворума), HACMP проверит наличие хотя бы одной рабочей копии каждого логического раздела для каждого логического тома в этой группе томов, прежде чем выполнять активизацию этой группы томов на узле перехвата.

Принудительная активизация группы томов позволяет подключить группу томов (в составе группы ресурсов) и поддерживать подключенное состояние, пока доступна хотя бы одна рабочая копия данных. Использовать опцию принудительной активизации следует только для групп томов, имеющих логические тома с зеркальным отображением; при этом нужно быть очень внимательным, чтобы не допустить создание разделенного кластера.

> **Примечание.** Следует задать политику размещения «super strict» для логических томов в группах томов, используемых с опцией принудительной активизации. В этом случае LVM проверяет наличие копий логического тома на разных дисках, увеличивая вероятность успешности принудительной активизации после отказа одного или нескольких дисков.

Эта опция полезна при перехвате в том случае, если в группе томов, входящей в группу ресурсов, отказывает один или несколько дисков (VGDA). Если эта опция отключена, группа ресурсов не будет активизирована на узле перехвата, что делает приложение недоступным.

HACMP при использовании принудительной активизации групп томов при перехвате сначала пытается просто выполнить команду **varyonvg**. В случае неуспешного выполнения команды из-за отсутствия кворума HACMP проверяет целостность данных, чтобы убедиться, что существует хотя бы одна доступная копия всех данных в группе томов, прежде чем пытаться выполнить принудительное подключение тома. Если такая копия существует, выполняется команда **varyonvg -f**; в противном случае группа томов остается отключенной и группа ресурсов переходит в ошибочное состояние (ERROR).

> **Примечание.** Пользователи все еще могут применять диски для обеспечения кворума (quorum buster disks) или дополнительные скрипты для принудительной активизации группы томов, однако новый атрибут принудительной активизации в HACMP автоматизирует эту операцию, и пользовательские принудительные процедуры теперь можно отключить.

Дополнительные сведения см. в гл. 5, «Planning Shared LVM Components», руководства **HACMP for AIX 5L V5.3 Planning and Installation Guide**, SC23-4861-06.

Часть 2

Планирование, установка и миграция

Часть 2 содержит сведения о планировании кластера и среды HACMP, а также пример установки кластера.

Кроме того, мы представили примеры миграции кластера с более ранней версии HACMP на версию HACMP 5.3. Наши сценарии содержат пошаговые инструкции и комментарии, а также описание проблем, возникающих при миграции.

3

Планирование

В этой главе обсуждаются аспекты планирования кластера HACMP. Адекватное планирование и подготовка обязательны для успешной установки и обслуживания кластера HACMP. Время, выделенное для надлежащего планирования конфигурации кластера и подготовки среды, окупится созданием кластера, который более прост в установке, обеспечивает более высокую доступность приложений и более прост в обслуживании.

Прежде чем начать планирование кластера, необходимо иметь отчетливое представление об имеющейся среде, приложении и о своих ожиданиях от HACMP. На основе этой информации можно разработать план реализации, который позволит легко интегрировать HACMP в вашу среду и, что более важно, заставить HACMP осуществлять управление доступностью вашего приложения в соответствии с вашими ожиданиями.

В дополнение к информации, представленной в гл. 2, «Составляющие высокой доступности», которая рассматривает понятия и основные аспекты структуры HACMP, эта глава сосредоточена на действиях, необходимых для планирования реализации HACMP 5.3. Для простоты описания мы представим планирование и подготовку кластера на примере кластера из двух узлов со взаимным перехватом (mutual takeover). Эта конфигурация является отправной точкой для более сложных установок.

3.1 Планирование высокой доступности

Основная цель планирования кластера высокой доступности состоит в том, чтобы исключить и свести к минимуму простои в работе требуемого приложения. С этой целью необходимо устранить единые точки отказа как в аппаратном, так и в программном обеспечении. Обычно это достигается путем дублирования оборудования по схеме «N+1», в частности путем дублирования источников питания, сетевых интерфейсов, адаптеров SAN и конфигураций зеркального отображения или RAID-дисков. Все эти компоненты увеличивают стоимость сервера, но могут не защитить приложение в случае отказа сервера или операционной системы.

> **Примечание.** Подробные сведения о планировании приведены в руководстве High Availability Cluster Multi-Processing for AIX 5L Planning and Installation Guide, SC23-4861-06.

Здесь наступает очередь HACMP. В HACMP можно настроить мониторинг серверного оборудования, операционной системы и компонентов приложения. В случае отказа HACMP может предпринять корректирующие действия, в частности выполнить перемещение требуемых ресурсов (сервисных IP-адресов, хранилищ и приложений) на рабочие компоненты кластера, чтобы как можно быстрее восстановить доступность приложения.

HACMP является чрезвычайно гибким продуктом, поэтому разработка кластера под нужды организации требует тщательного планирования. Требования и режим функционирования приложения представляют важную входную информацию плана HACMP и являются основными факторами при определении схемы кластера. При разработке схемы кластера необходимо поставить перед собой следующие вопросы:

- Для каких служб приложений необходимо обеспечить высокую доступность?
- Каковы требования к уровню обслуживания для этих служб приложений (24x7, 8x5) и насколько быстро должно выполняться восстановление обслуживания в случае отказа?
- Каковы потенциальные точки отказа в среде и как их можно устранить?
- Какие точки отказа могут быть автоматически обнаружены HACMP и для каких потребуется написать специальный код для вызова события?
- Каким должен быть уровень квалификации в группе, осуществляющей внедрение и обслуживание кластера?

Хотя за внедрение HACMP обычно отвечают системные администраторы AIX, как правило, они не могут заниматься этим самостоятельно. Для помощи в планировании HACMP следует создать команду, состоящую из следующих представителей, каждый из которых играет важную роль в успешной работе кластера:

- сетевой администратор;
- системный администратор AIX;
- администратор баз данных;
- разработчик приложений;
- персонал поддержки;
- пользователи приложения.

3.2 Планирование HACMP

Основные этапы успешного внедрения HACMP указаны на рис. 3.1. Обратите внимание на то, что внедрение кластера не завершается с успешным конфигурированием кластера. Процедуры тестирования кластера, резервного копирования, документирования и текущего управления одинаково важны для обеспечения текущей целостности кластера.

Используя понятия, описанные в гл. 1, «Введение в HACMP», внедрение HACMP начинается с разработки подробного плана конфигурирования и внедрения кластера HACMP. Для направления этого процесса и записи информации о кластере можно использовать такие инструменты планирования, как таблицы планирования на бумаге (Paper Planning Worksheets), описываемые в руководстве **High Availability Cluster Multi-Processing for AIX 5L Planning and Installation Guide**, SC23-4861-06, и система автоматизированного планирования (Online Planning Worksheets).

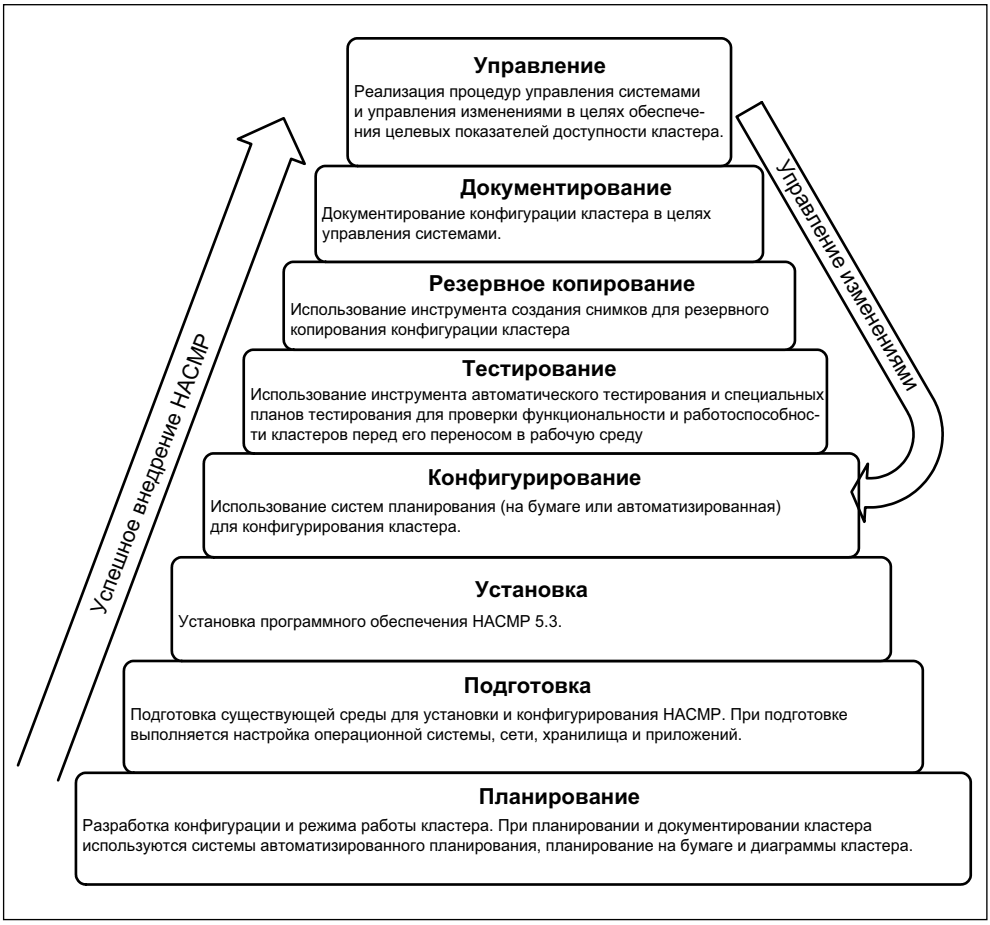

Рис. 3.1. Этапы внедрения HACMP

Важно! Помните, что время, проведенное за планированием кластера, обернется более простым внедрением, так что не торопитесь.

Как показано на рис. 3.1, планирование является основой, на которой строится внедрение. Надлежащее планирование должно затрагивать все аспекты внедрения кластера. Оно должно включать:

- схему и режим работы кластера;
- подробную конфигурацию кластера;
- аспекты и план установки;
- план тестирования целостности кластера;
- стратегию резервного копирования для кластера;
- процедуру документирования кластера;
- план управления проблемами и изменениями в кластере.

Важно! Для успешного внедрения важно, чтобы планирование и подготовка кластера HACMP осуществлялись до установки и конфигурирования HACMP.

Для простоты описания мы представим планирование кластера на примере кластера из двух узлов со взаимным перехватом. Эта конфигурация является отправной точкой для более сложных установок. Глава также содержит образцы таблиц планирования, так как цель главы заключается в том, чтобы показать, как выполняется планирование кластера.

3.2.1 Инструменты планирования

При планировании кластера HACMP можно использовать три инструмента:

- диаграмму кластера;
- таблицы планирования на бумаге (Paper Planning Worksheets);
- система автоматизированного планирования (Online Planning Worksheets).

И диаграмма кластера и таблицы планирования на бумаге представляют собой метод записи информации о кластере вручную. Система автоматизированного планирования представляет простой в использовании интерфейс на основе Java, который может применяться для записи и конфигурирования кластера.

Все три инструмента обсуждаются более подробно в конце этой главы.

Примечание. Если вы решили использовать систему автоматизированного планирования (Online Planning Worksheets, OLPW) или средство упрощенного конфигурирования кластера из двух узлов (Two-Node Cluster Configuration Assistant), все равно важно выполнить планирование и подготовку HACMP. Система автоматизированного планирования и средство упрощенного конфигурирования кластера из двух узлов предназначены только для того, чтобы упростить документирование и конфигурирование кластера; вам все равно нужно иметь отчетливое представление о планировании и подготовке.

3.3 Приступая к работе

Планирование кластера начинается с анализа текущей среды и своих ожиданий от HACMP:

- Для каких приложений необходимо обеспечить высокую доступность?
- Сколько узлов необходимо для поддержки приложений?
- Существуют ли узлы достаточного размера (процессор/память) для выполнения нескольких приложений или устанавливаются новые узлы?
- Как клиенты подключаются к приложению, какова конфигурация сети?
- Какой тип общего диска будет использоваться?
- Каковы ожидания от HACMP?

3.3.1 Текущая среда

На рис. 3.2 показана простая стартовая конфигурация, которую мы будем использовать в качестве примера. Она сосредоточена на обеспечении высокой доступности двух приложений. Это может быть либо пара существующих серверов, либо два новых сервера.

Стартовая конфигурация показывает следующее:

- каждое приложение находится на отдельном узле (сервере);
- клиенты осуществляют доступ к каждому приложению через выделенное Ethernet-подключение на каждом сервере;
- каждый узел имеет приблизительно один размер с точки зрения процессора и памяти, каждый имеет дополнительную мощность;
- каждый узел имеет дополнительные источники питания и внутренние диски с зеркальным отображением;
- приложения расположены на внешнем диске SAN;
- каждое приложение имеет собственные надежные скрипты запуска и остановки;
- существует инструмент мониторинга для проверки состояния каждого приложения;
- AIX 5.3 уже установлен.

Важно! Каждое приложение, подлежащее интеграции в кластер, должно работать в автономном режиме (standalone mode). Кроме того, вы должны иметь возможность полностью контролировать приложение (запускать, останавливать и осуществлять тестирование).

Цель состоит в том, чтобы использовать два узла в конфигурации со взаимным перехватом, где приложение app1 обычно находится на узле node01, а приложение app2 обычно находится на узле node02. В случае отказа нужно, чтобы оба приложения выполнялись на оставшемся сервере. Из диаграммы мы видим, что нужно подготовить среду таким образом, чтобы каждый узел мог выполнять оба приложения.

Внимание! Каждое приложение, подлежащее интеграции в кластер, должно быть способно выполняться в автономном режиме на любом узле, на котором ему, возможно, придется выполняться (как в обычных условиях, так и в случае перемещения при сбое).

Анализируя требования к кластеру HACMP, получаем три основные области, требующие внимания, как показано на рис. 3.2: сеть, приложение и хранилище. Все действия по планированию в какой-то мере предназначены для поддержки одного из трех этих элементов.

1. **Сеть** – каким образом клиенты подключаются к приложению (сервисный адрес). Сервисный адрес может перемещаться между всеми выделенными узлами кластера.

2. **Приложение** – какие ресурсы необходимы для приложения. Приложение должно располагать всем, что ему необходимо для работы на узле перемещения при сбое, включая ресурсы процессора и памяти, лицензии, исполняемые модули и данные конфигурации. Оно должно иметь надежные скрипты запуска и остановки, а также инструмент для мониторинга своего состояния.

3. **Хранилище** – какой тип общего диска будет использоваться. Данные приложения должны располагаться на общем диске, доступном для всех требуемых узлов кластера.

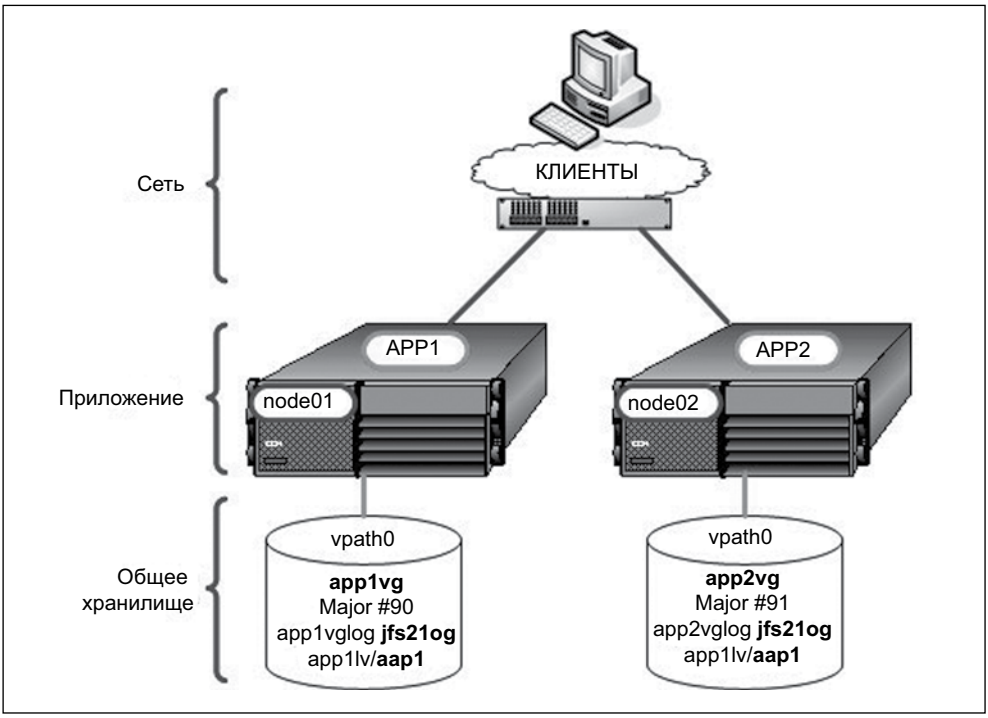

Рис. 3.2. Первоначальная среда

3.3.2 Устранение единых точек отказа

В табл. 3.1 приведен обзор возможных единых точек отказа в инфраструктуре кластера и способов защиты от них. Эти аспекты следует учитывать во время разработки подробной схемы кластера.

Таблица 3.1. Единые точки отказа

Объекты кластера	Способ устранения единой точки отказа	Поддержка в HACMP/AIX
Узлы	Использование нескольких узлов	До 32
Источники питания	Использование нескольких электросетей или источников бесперебойного питания (ИБП)	Столько, сколько нужно
Сети	Использование нескольких сетей для соединения узлов	До 48
Сетевые интерфейсы, устройства и IP-адреса	Использование резервных сетевых адаптеров	До 256
Подсистема TCP/IP	Использование сетей «точка-точка» для соединения соседних узлов и клиентов	Столько, сколько нужно
Дисковые адаптеры	Использование резервных дисковых адаптеров	Столько, сколько нужно
Дисковые контроллеры	Использование резервных дисковых контроллеров	Столько, сколько нужно (ограничение на аппаратном уровне)
Диски	Использование резервного оборудования, а также технологий зеркального отображения, чередования или их сочетания	Столько, сколько нужно
Приложения	Назначение узла для перехвата приложения, конфигурирование мониторов приложений, конфигурирование кластеров с узлами на нескольких сайтах	Столько, сколько нужно
Сайты	Использование нескольких сайтов для аварийного восстановления (disaster recovery)	2
Группы ресурсов	Использование групп ресурсов с указанием, каким образом должен работать набор ресурсов	До 64 в кластере
Ресурсы кластера	Использование нескольких ресурсов кластера	До 128 в Clinfo (кластер может включать больше)

3.3.3 Первоначальная схема кластера

После того как было сформировано представление о текущей среде, уяснены понятия HACMP и выявлены свои ожидания от кластера, можно начать разработку схемы кластера.

На этом этапе следует создать схему кластера HACMP. Рекомендуется начать с простого и постепенно увеличивать степень детализации по мере продвижения процесса планирования. Схема позволяет выявить единые точки отказа, требования приложений и поможет направлять процесс планирования.

Для записи информации о конфигурации и кластере следует также использовать таблицы планирования на бумаге или систему автоматизированного планирования.

Первоначальная схема кластера, используемого в нашем примере, представлена на рис. 3.3. На данный момент основное внимание сосредотачивается на высоко-

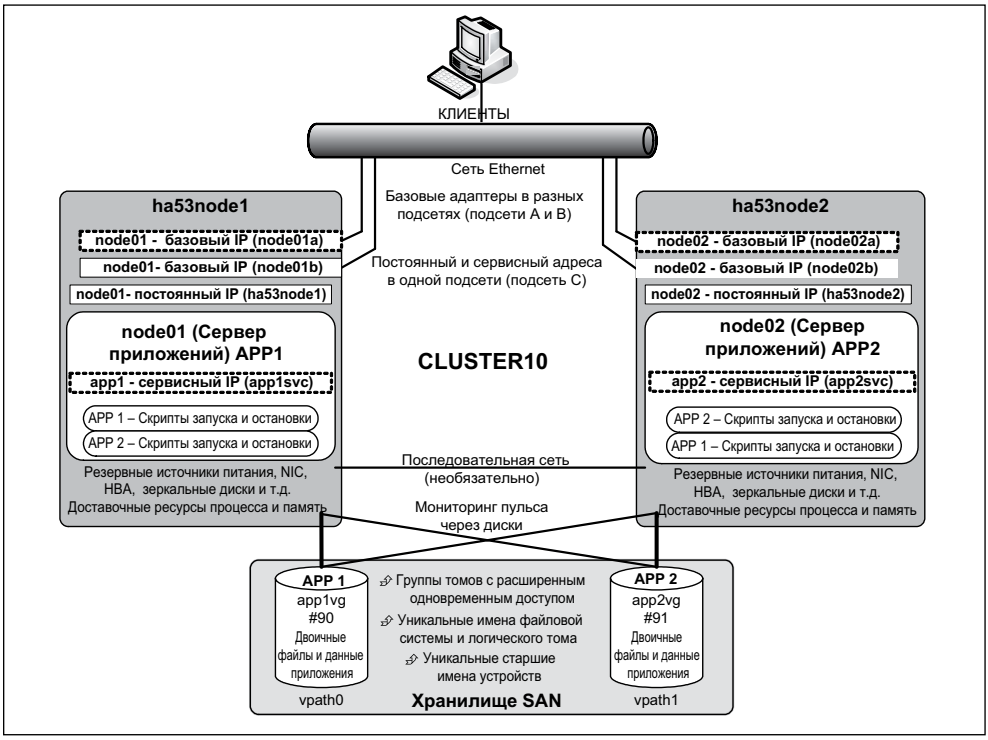

Рисунок 3.3. Первоначальная схема кластера

уровневом функционировании кластера, подробности кластера рассматриваются по мере прохождения этапа планирования.

Мы начинаем принимать решения относительно схемы топологии и режима работы кластера на основании своих требований. Например, в соответствии с требованиями первоначальная схема кластера включает следующее:

- Кластер представляет собой кластер из двух узлов со взаимным перехватом.
- В качестве имен узлов кластера можно использовать имена хостов, однако мы решили задавать имена узлов кластера.

> **Примечание.** Если вы планируете использовать динамические разделы (DLPAR), имя хоста AIX, имя узла кластера и имя HMC LPAR (отображаемое в интерфейсе HMC) должны совпадать.

- Каждый узел содержит одно приложение, но способен выполнять оба (нужно рассмотреть сеть, хранилище, память, процессор, программное обеспечение).
- Каждый узел имеет два Ethernet-адаптера, подключенные к одной физической сети Ethernet, каждый к отдельному коммутатору (или какой-либо тип резервного коммутатора).
- Используется перехват IP-адреса посредством синонимом (а не посредством замены).

- Каждый узел имеет постоянный IP-адрес (IP-синоним, всегда доступный при работе узла) и один сервисный IP-адрес (IP-синоним на одном из адаптеров, находящихся под управлением HACMP). Базовые адреса адаптера Ethernet относятся к разным подсетям. Так как используется мониторинг пульса через синонимы, это условие не является обязательным, при необходимости оба адаптера могут находится в одной подсети.
- Общие диски находятся в SAN и доступны на обоих узлах.
- Все группы томов на общих дисках создаются в режиме расширенного одновременного доступа (Enhanced Concurrent Mode, ECM), что позволяет обеспечить мониторинг пульса через диски и быстрый перехват дисков.
- Последовательное подключение RS232 показано, однако в связи с использованием мониторинга пульса через диски можно опустить этот элемент схемы. Он показан как необязательный.
- Используется мониторинг пульса через IP-синонимы (не показано).
- Каждый узел имеет достаточно ресурсов процессора и памяти для выполнения обоих приложений.
- Каждый узел имеет резервное оборудование и внутренние диски с зеркальным отображением.
- Установлен AIX 5.3 ML02.
- Используется HACMP 5.3.

Этот список кратко описывает основные компоненты схемы кластера. Каждый элемент будет более подробно рассмотрен в процессе планирования.

3.3.4 Заполнение обзорной таблицы планирования кластера

Рекомендуем заполнить первоначальную таблицу, как мы это сделали в нашем примере. В этой главе приведено 11 таблиц планирования, каждая их которых охватывает различные аспекты его планирования. Табл. 3.2 представляет первую таблицу со списком основных элементов кластера.

Таблица 3.2. Обзор кластера

ТАБЛИЦА КЛАСТЕРА HACMP – ЧАСТЬ 1 из 11	**ДАТА: июль 2005**
ОБЗОР КЛАСТЕРА	
ИМЯ КЛАСТЕРА	cluster10
ОРГАНИЗАЦИЯ	IBM ITSO
ИМЯ ХОСТА УЗЛА 1	ha53node1
ИМЯ ХОСТА УЗЛА 2	ha53node2
ИМЯ HACMP УЗЛА 1	node01
ИМЯ HACMP УЗЛА 2	node02
КОММЕНТАРИИ	Это набор таблиц планирования для простого 2-узлового кластера HACMP 5.3 со взаимным перехватом с использованием перехвата IP-адреса посредством синонимов.

3.4 Планирование оборудования кластера

Схема кластера начинается с определения требуемого количества и типа узлов. Эти аспекты в значительной степени зависят от двух факторов:

- количества ресурсов, требуемых каждым приложением;
- режима перемещения при отказе в кластере.

Примечание. Количество узлов в кластере может варьироваться от 2 до 32.

При выборе узлов основное условие заключается в том, чтобы в случае перемещения при сбое оставшийся узел или узлы были способны выполнять приложения с отказавшего узла. Другими словами, если у вас есть кластер из двух узлов и при этом отказывает один узел, оставшийся узел должен иметь ресурсы, требуемые для выполнения приложений с отказавшего узла (помимо собственных приложений). Если это невозможно, можно рассмотреть вариант внедрения дополнительного узла в качестве дежурного или вариант использования динамических разделов (DLPAR, в системах POWER4™ или POWER5™). Как вы сможете заметить, HACMP допускает широкий выбор конфигураций кластера в зависимости от ваших требований.

HACMP работает практически с любым узлом, поддерживаемым AIX, от настольных систем до высокопроизводительных серверов. При выборе типа узла следует учитывать следующее:

- Убедитесь, что все узлы имеют достаточно ресурсов процессора и памяти, чтобы система могла вести себя требуемым образом в случае перемещения при сбое. Ресурсы процессора и памяти должны быть способны поддерживать требуемые приложения в случае перемещения при сбое, иначе у клиентов могут возникнуть проблемы с производительностью. Если вы используете разделы LPAR, вам может быть полезно использовать возможности DLPAR для увеличения ресурсов в случае перемещения при сбое. При использовании автономных серверов такая возможность отсутствует, так что вам, возможно, придется рассмотреть вариант применения дежурного (резервного, standby) сервера.
- На каждом сервере используйте оборудование высокой доступности и резервные компоненты там, где это возможно. Например, применяйте резервные источники питания и подключите их в отдельным электросетям.
- На каждом узле обеспечьте защиту rootvg (копии локальной операционной системы) с использованием зеркального отображения или технологии RAID.
- Установите как минимум два Ethernet-адаптера на каждом узле и подключите их к отдельным коммутаторам, чтобы защитить их от отказа одного из адаптеров или коммутаторов.
- Установите два SAN-адаптера на каждом узле, чтобы защитить их от отказа одного из SAN-адаптеров.

Хотя это не обязательно, мы рекомендуем использовать узлы кластера с похожими конфигурациями оборудования, чтобы упростить распространение ресурсов и выполнение административных операций. Другими словами, не пытайтесь выпол-

нить перемещение при сбое с высокопроизводительного сервера на настольную систему, ожидая что все будет работать надлежащим образом; будьте благоразумны при выборе узлов.

3.4.1 Заполнение таблицы планирования оборудования кластера

Табл. 3.3 содержит сведения об оборудовании для нашего примера. Там, где это возможно, мы использовали избыточное оборудование, дополнительные коммутаторы Ethernet и SAN, а также убедились в достаточности ресурсов для поддержки одновременной работы приложений на каждом узле.

Таблица 3.3. Оборудование кластера

ТАБЛИЦА КЛАСТЕРА НАСМР – ЧАСТЬ 2 из 11		**ДАТА: июль 2005**
ОБОРУДОВАНИЕ КЛАСТЕРА		
КОМПОНЕНТ ОБОРУДОВАНИЯ	**СПЕЦИФИКАЦИИ**	**КОММЕНТАРИИ**
p520	Сервер pSeries	Количество – 2
	2 процессора и 4 Гб памяти	*Последняя версия микрокода
	4 внутренних SCSI-диска	Резервные источники питания.
		Достаточно ресурсов для выполнения обоих приложений на одном сервере
Ethernet-адаптеры	10/100/1000 Ethernet	2 NIC на узел (минимум)
Сетевые коммутаторы	Название производителя - Модель	2 коммутатора.
		Каждый NIC на узле подключен к отдельному коммутатору.
		Все порты сконфигурированы в одной VLAN.
		Коммутаторы поддерживают gratuitous ARP-запросы и отключенный алгоритм Spanning Tree.
		Установлено оптимальное значение скорости порта коммутатора
Адаптеры SAN	6239 2GB Fibre Channel	2 HBA на узел
Коммутаторы SAN	IBM 2109	2 коммутатора.
		Каждый HBA на узле подключен к отдельному коммутатору.
		Общий диск разделен на зоны для всех узлов, требующих доступа
Хранилище SAN	IBM ESS	Модель 800
КОММЕНТАРИИ	Совместимость оборудования проверена.	

3.5 Планирование программного обеспечения кластера

Для обеспечения совместимости требуется провести анализ всех программных компонентов, которые планируется использовать в кластере. Необходимо рассмотреть программное обеспечение AIX, RSCT, HACMP, приложения и хранилища. В этом разделе обсуждаются различные уровни программного обеспечения, а также вопросы совместимости.

3.5.1 Уровни AIX и RSCT

HACMP 5.3 поддерживается в AIX версий 5.2 и 5.3. Использовавшиеся в нашей среде уровни сочетания AIX и RSCT приведены в табл. 3.4.

Таблица 3.4. Уровни AIX и RSCT

Версия AIX	Версия RSCT	Минимальные наборы файлов (filesets) RSCT
AIX 5.3 ML02 (5300-02) или выше, с APARS	2.4.2	rsct.compat.basic.hacmp.2.4.2.0, rsct.compat.clients.hacmp.2.4.2.0, rsct.core.sec.2.4.2.1, rsct.core.rmc.2.4.2.1
AIX 5.2 ML06 (5200-06) или выше, с APARS IYXXXXX	2.3.6	rsct.compat.basic.hacmp.2.3.6.0, rsct.compat.clients.hacmp.2.3.6.0, rsct.core.sec.2.3.6.1, rsct.core.rmc.2.3.6.1

3.5.2 Поддержка Virtual LAN и Virtual SCSI

Для поддержки функций Virtual LAN (виртуальная локальная сеть) и Virtual SCSI, реализованных в IBM P5 Virtual I/O Server, необходимы следующие уровни программного обеспечения:

- AIX 5.3 ML02 (5300-02) с APARs IY70082 и iFIX IY72974;
- VIO Server V1.1 с Fixpack 6.2 и iFIX IY71303.062905.epkg.Z;
- минимальные уровни RSCT:
  - rsct.basic.hacmp 2.4.2.1,
  - rsct.basic.rte 2.4.2.2,
  - rsct.compat.basic.hacmp 2.4.2.0.

3.5.3 Обязательные наборы файлов

Следующие наборы файлов являются обязательными для работы HACMP. Они должны быть установлены вместе с последней версией исправлений к соответствующему уровню AIX перед установкой HACMP.

- bos.adt.lib;
- bos.adt.libm;
- bos.adt.syscalls;
- bos.net.tcp.client;
- bos.net.tcp.server;
- bos.rte.SRC;
- bos.rte.libc;
- bos.rte.libcfg;
- bos.rte.libcur;
- bos.rte.libpthreads;
- bos.rte.odm;
- bos.rte.lvm.rte (необходим только при использовании существующего или унаследованного 32-разрядного диспетчера Concurrent Logical Volume Manager для одновременного доступа);
- bos.clvm.enh (необходим для логических томов с расширенным одновременным доступом; используется при мониторинге пульса через диски).

3.5.4 Наборы файлов безопасности AIX

Следующие наборы файлов являются обязательными, если вы планируете использовать аутентификацию или шифрование сообщений в HACMP для связи между узлами кластера. Их можно установить с диска AIX 5L Expansion Pack CD-ROM.

- rsct.crypt.des – для шифрования данных с использованием алгоритма аутентификации сообщений DES.
- rsct.crypt.3des – для шифрования данных с использованием алгоритма аутентификации сообщений Triple DES.
- rsct.crypt.aes256 – для шифрования данных с использованием алгоритма аутентификации сообщений AES (Advanced Encryption Standard).

Примечание. Эти наборы файлов не поддерживаются в AIX 5.1.

3.5.5 Программное обеспечение, необходимое для работы WebSmit

Следующее программное обеспечение является необходимым, если вы планируете установить и сконфигурировать WebSmit. Представленные версии являются наиболее актуальными на момент написания этой книги.

- Apache-совместимый веб-сервер (Apache или IBM Http Server, IHS).
  – Apache 1.3.31 можно скопировать по адресу

  http://www-1.ibm.com/servers/aix/products/aixos/linux/download.html.

– IBM IHS v2.0.47.1 можно скопировать по адресу

  http://www-306.ibm.com/software/webservers/httpservers/.

- RPM Package Manager (если еще не установлен в системе); expat-195.7-1.aix5.1.ppc.rpm можно скопировать по адресу

  http://www-1.ibm.com/servers/aix/products/aixos/linux/download.html

- openssl может быть скопирован по адресу

  http://www-1.ibm.com/servers/aix/products/aixos/linux/download.html

  Выберите «AIX Toolbox Cryptographic Content» и скопируйте следующие файлы:

  – apache-1.3.31-1ssl.aix5.1.ppc.rpm;
  – mod_ssl-2.8.19-1ssl.aix5.1.ppc.rpm;
  – openssl-0.9.7d-1.ssl.aix5.1.ppc.rpm.

> **Примечание.** Вы должны зарегистрироваться на ibm.com®, при этом необходимо учитывать, что регистрация может занять 24 ч. Если вы НЕ зарегистрированы, вы не сможете скопировать файлы криптографического содержимого.

3.5.6 Наборы файлов HACMP

С установочного носителя можно получить следующие наборы файлов HACMP (за исключением дополнительных языковых наборов файлов):

- cluster.adt.es.client.include;
- cluster.adt.es.client.samples.clinfo;
- cluster.adt.es.client.samples.clstat;
- cluster.adt.es.client.samples.libcl;
- cluster.adt.es.java.demo.monitor;
- cluster.assist.license – HACMP Smart Assist Feature;
- cluster.doc.en_US.assist.db2.html;
- cluster.doc.en_US.assist.db2.pdf;
- cluster.doc.en_US.assist.oracle.html;
- cluster.doc.en_US.assist.oracle.pdf;
- cluster.doc.en_US.assist.websphere.html;
- cluster.doc.en_US.assist.websphere.pdf;
- cluster.doc.en_US.es.html – HTML-документация по HAES;
- cluster.doc.en_US.es.pdf – PDF-документация по HAES (англ.);
- cluster.doc.en_US.glvm.html;
- cluster.doc.en_US.glvm.pdf;
- cluster.doc.en_US.pprc.pdf;
- cluster.doc.en_US.pprc.html;

- cluster.es.assist.common – HACMP Smart Assist Common;
- cluster.es.assist.db2 – HACMP Smart Assist for DB2;
- cluster.es.assist.oracle – HACMP Smart Assist for Oracle;
- cluster.es.assist.websphere;
- cluster.es.cfs.rte;
- cluster.es.client.lib – библиотеки ES-клиента;
- cluster.es.client.rte – среда выполнения ES-клиента;
- cluster.es.client.utils – утилиты ES-клиента;
- cluster.es.client.wsm – WebSmit;
- cluster.es.clvm.rte – ES для одновременного доступа AIX;
- cluster.es.cspoc.cmds – команды ES CSPOC;
- cluster.es.cspoc.dsh – ES CSPOC dsh;
- cluster.es.cspoc.rte – команды среды выполнения ES CSPOC;
- cluster.es.ercmf.cmds;
- cluster.es.ercmf.rte;
- cluster.es.plugins.dns;
- cluster.es.plugins.printserver;
- cluster.es.plugins.dhcp;
- cluster.es.pprc.cmds;
- cluster.es.pprc.rte;
- cluster.es.server.cfgast – двухузловая конфигурация ES;
- cluster.es.server.diag – средства диагностики ES-сервера;
- cluster.es.server.events – события ES-сервера;
- cluster.es.server.rte – базовая среда выполнения ES-сервера;
- cluster.es.server.testtool;
- cluster.es.server.utils – утилиты ES-сервера;
- cluster.es.svcpprc.cmds;
- cluster.es.svcpprc.rte;
- cluster.es.worksheets – система автоматизированного планирования (OLPW);
- cluster.hativoli.client;
- cluster.hativoli.server;
- cluster.license – электронная лицензия HACMP;
- cluster.man.en_US.assist.data;
- cluster.man.en_US.es.data;
- cluster.msg.en_US.assist;
- cluster.msg.en_US.cspoc;
- cluster.msg.en_US.ercmf;

- cluster.msg.en_US.es.client;
- cluster.msg.en_US.es.server;
- cluster.msg.en_US.hativoli;
- cluster.msg.en_US.pprc;
- cluster.msg.en_US.svcpprc;
- cluster.xd.glvm;
- cluster.xd.license;
- glvm.rpv.util – Geographic LVM;
- glvm.rpv.client;
- glvm.rpv.server;
- glvm.rpv.msg.en_US;
- hageo.doc.en_US.data;
- hageo.gmdsizing;
- hageo.man.en_US.message.data;
- hageo.man.en_US.mirror.data;
- hageo.manage.utils;
- hageo.message.ext;
- hageo.message.utils;
- hageo.mirror.ext;
- hageo.mirror.utils;
- hageo.msg.en_US.message;
- hageo.msg.en_US.mirror.

3.5.7 Файлы AIX, изменяемые HACMP

Помните о том, что следующие системные файлы могут быть изменены HACMP в процессе установки, верификации и синхронизации кластера.

/etc/hosts

Скрипты событий кластера используют файл /etc/hosts для разрешения имен. Все IP-интерфейсы узлов кластера должны быть указаны в этом файле на всех узлах. HACMP может изменить этот файл, чтобы обеспечить наличие всей необходимой информации в файле /etc/hosts на всех узлах для корректной работы HACMP.

При удалении сервисных IP-меток из конфигурации кластера с использованием SMIT рекомендуется также удалить их из файла /etc/hosts.

/etc/inittab

Файл /etc/inittab изменяется в следующих случаях:

- При установке HACMP.

При первоначальной установке HACMP добавляется следующая строка (она запускает подсистемы clcomdES и clstrmgrES, если они еще не запущены):hacmp:2:once:/usr/es/sbin/cluster/etc/rc.init >/dev/console 2>&1.

> **Важно!** Эта запись HACMP осуществляет запуск следующих демонов с использованием команды startsrc, если они еще не запущены:
> - startsrc -s syslogd,
> - startsrc -s snmpd,
> - startsrc -s clcomdES,
> - startsrc -s clstrmgrES.

- При конфигурировании перехвата IP-адреса в HACMP:
  - **harc:2:wait:/usr/es/sbin/cluster/etc/harc.net # HACMP network startup**
  - если включен перехват IP-адреса, система редактирует файл /etc/inittab, изменяя rc.tcpip- и inet-зависимые записи с уровня выполнения «**2**» (многопользовательский уровень по умолчанию) на уровень выполнения «**a**»;
  - записи с уровнем выполнения «a» обрабатываются только при выполнении команды telinit с указанием этого уровня выполнения.
- При выборе опции Start at System Restart на панели SMIT System Management (C-SPOC) > Manage HACMP Services > Start Cluster Services:
  - **hacmp6000:2:wait:/usr/es/sbin/cluster/etc/rc.cluster -boot -i # Bring up Cluster;**
  - при загрузке системы файл /etc/inittab вызывает скрипт /usr/es/sbin/cluster/etc/rc.cluster для запуска HACMP;
  - так как inet-демоны не должны запускаться до тех пор, пока интерфейсы, управляемые HACMP, не перейдут на сервисный IP-адрес, HACMP также добавляет следующую запись в конец файла /etc/inittab, указывая таким образом что обработка файла /etc/inittab была завершена:
    - **clinit:a:wait:/bin/touch /usr/es/sbin/cluster/.telinit #HACMP for AIX These must be the last entry in run level “a” in inittab!**
    - **pst_clinit:a:wait:/bin/echo Created /usr/es/sbin/cluster/.telinit > /dev/console #HACMP for AIX These must be the last entry in run level “a” in inittab!**

> **Внимание!** Хотя можно запускать службы кластера из inittab, настоятельно рекомендуем не использовать эту опцию. Лучше контролировать запуск HACMP вручную. Например, в случае отказа лучше определять причину отказа до перезапуска HACMP на узле.

- При установке Concurrent Logical Volume Manager (cluster.es.clvm) в HACMP. В файл /etc/inittab автоматически добавляется следующая запись:

  haclvm_cfg:2:wait:/usr/es/sbin/cluster/clvm/config_mode3.

> **Примечание:** В файл inittab также добавляется запись ha_star. Эта запись вносится вместе с набором файлов bos.rte.control, а не с HACMP.

/etc/rc.net

Файл /etc/rc.net вызывается утилитой cfgmgr (cfgmgr – утилита AIX 5L, осуществляющая конфигурирование устройств, а также, опционально, установку программного обеспечения устройств в системе) для настройки и запуска TCP/IP в процессе загрузки. Он задает имя хоста, шлюз по умолчанию и статические маршруты.

/etc/services

HACMP использует следующие сетевые порты для связи между узлами кластера (они все перечислены в файле /etc/services):

- clinfo_deadman 6176/tcp;
- clsmuxpd 6270/tcp;
- clm_lkm 6150/tcp;
- clm_smux 6175/tcp;
- godm 6177/tcp;
- topsvcs 6178/udp;
- grpsvcs 6179/udp;
- emsvcs 6180/udp;
- #clver 6190/tcp (закомментировано, так как clverify теперь использует clcomd);
- clcomd 6191/tcp;
- clinfo_client 6174/tcp;
- #cllockd 6100/udp (закомментировано – cllockd не поддерживается начиная с HACMP 5.2);
- #clm_mig_1k 6151/tcp (закомментировано – cllockd не поддерживается начиная с HACMP 5.2).

Примечание. При установке HACMP/XD для GLVM в файл /etc/services на всех узлах локальных и удаленных сайтов, на которых установлено программное обеспечение, автоматически добавляется следующая запись с номером порта и протоколом подключения:
rpv 6192/tcp. Требования приложения

Помимо HACMP, следующие порты используются RMC:

- #rmc 657/tcp;
- #rmc 657/udp.

WebSmit обычно использует порт (порт WebSmit является настраиваемым)

#http 42267 (порт WebSmit)

/etc/snmpd.conf

Версия snmpd.conf зависит от того, какая версия системы используется: AIX 5L V5.1 или более поздняя. В более поздних версиях системы AIX 5L, чем V5.1, по умолчанию используется файл snmpdv3.conf.

Демон SNMP считывает файл конфигурации /etc/snmpd.conf при запуске, а также при обновлении или выдаче сигнала kill -1. Этот файл задает имена сообществ (community names) и соответствующие привилегии доступа и представления, хосты уведомления о ловушках (trap), атрибуты журналов, конфигурации параметров snmpd, а также конфигурации SMUX для snmpd. Процесс установки HACMP добавляет clsmuxpd-пароль для этого файла.

Чтобы включить HACMP MIB с управлением из диспетчера кластера (Cluster Manager), в конец файла добавляется запись

smux 1.3.6.1.4.1.2.3.1.2.1.5 “clsmuxpd_password” # HACMP clsmuxpd

/etc/snmpd.peers

Файл /etc/snmpd.peers осуществляет настройку сторон подключения SMUX (snmpd SMUX peers). При установке HACMP добавляет следующую запись, чтобы включить clsmuxpd-пароль для этого файла:

clsmuxpd 1.3.6.1.4.1.2.3.1.2.1.5 "clsmuxpd_password" # HACMP clsmuxpd

/etc/syslog.conf

Файл конфигурации /etc/syslog.conf используется для управления выходными данными демона syslogd, который регистрирует системные сообщения. В процессе установки HACMP добавляет в этот файл записи, которые направляют вывод сообщений, связанных с HACMP в определенные файлы.

Пример:

- # HACMP Critical Messages from HACMP
- local0.crit /dev/console
- # HACMP Informational Messages from HACMP
- local0.info /usr/es/adm/cluster.log
- # HACMP Messages from Cluster Scripts
- user.notice /usr/es/adm/cluster.log
- # HACMP/ES for AIX Messages from Cluster Daemons
- daemon.notice /usr/es/adm/cluster.log

Файл /etc/syslog.conf должен быть одинаковым на всех узлах.

/etc/trcfmt

Файл /etc/trcfmt представляет собой файл-шаблон для утилиты регистрации и создания отчетов трассировки системы, trcrpt. В процессе установки добавляется трассировка HACMP в файл формата трассировки. Трассировка HACMP выполняется для демонов clstrmgr и clinfo.

/var/spool/cron/crontab/root

В процессе установки HACMP в файл /var/spool/cron/crontab/root добавляется чередование файлов журналов HACMP.

0 0 * * * /usr/es/sbin/cluster/utilities/clcycle 1>/dev/null 2>/dev/null # > HACMP for AIX Logfile rotation

3.5.8 Программное обеспечение приложений

Обычно приложения не зависят от версии HACMP; они даже не знают о функционировании HACMP. Другими словами, HACMP просто запускает и останавливает приложения (HACMP также может осуществлять мониторинг приложений, однако обычно для этого подбирается метод, зависящий от приложения).

Тем не менее существует несколько приложений, в частности Oracle RAC 9i, которые в значительной степени зависят от используемой версии HACMP.

Свяжитесь с производителем приложения, чтобы убедиться в отсутствии проблем (например, с лицензированием) при использовании HACMP 5.3.

3.5.9 Лицензирование

Необходимо принимать во внимание два аспекта лицензирования: лицензирование HACMP (функций) и лицензирование приложения.

HACMP

Начиная с HACMP V5.2 лицензирование HACMP основано на количестве процессоров. Под количеством процессоров понимается суммарное количество процессоров, на которых установлен или выполняется HACMP V5. Наличие лицензий HACMP является обязательным на каждом компьютере, на котором устанавливается и выполняется HACMP.

Лицензирование HACMP/XD V5 основано на количестве процессоров. Под количеством процессоров понимается суммарное количество процессоров, на которых установлен или выполняется HACMP/XD V5. Наличие лицензий HACMP V5 и HACMP/XD V5 является обязательным на каждом компьютере, на котором устанавливается и выполняется HACMP/XD.

Лицензирование HACMP V5 Smart Assist основано на количестве процессоров. Под количеством процессоров понимается суммарное количество процессоров, на которых установлен или выполняется HACMP V5 Smart Assist. Наличие лицензий HACMP V5 и HACMP V5 Smart Assist является обязательным на каждом компьютере, на котором устанавливается и выполняется HACMP V5 Smart Assist.

Итак, вот что это значит:

- если у вас есть сервер pSeries с четырьмя процессорами, работающими в режиме использования всех ресурсов в одном разделе (full system partition mode), вам необходима лицензия на 4 процессора;
- если у вас есть сервер pSeries с четырьмя процессорами, обрабатывающими логические разделы, и HACMP выполняется только в разделе с двумя процессорами, вам необходима лицензия на 2 процессора;
- конечно же, вам необходима лицензия для каждого сервера, на котором вы планируете запускать HACMP.

Примечание. Лицензирование HACMP по микроразделам не осуществляется. Необходимо выполнять лицензирование для процессоров целиком.

Приложение

Некоторые приложения имеют особенные требования по лицензированию, например использование отдельной лицензии для каждого процессора, выполняющего приложение; это означает, что необходимо лицензировать приложение путем включения информации, относящейся к процессору, в приложение при его установке. В результате, несмотря на корректную обработку отказов узла программным обеспечением HACMP, оно может быть неспособно перезапустить приложение на узле перемещения при сбое из-за ограничения на количество лицензий для этого приложения в кластере.

Во избежание этой проблемы, убедитесь в наличии лицензии для каждой системы в кластере, которая потенциально может выполнять приложение.

> **Важно!** Свяжитесь с производителем приложения, чтобы убедиться в отсутствии проблем с лицензированием при использовании HACMP 5.3.

3.5.10 Заполнение таблицы планирования программного обеспечения

Табл. 3.5 содержит полный список программного обеспечения, установленного в нашем примере.

Таблица 3.5. Программное обеспечение кластера

ТАБЛИЦА КЛАСТЕРА HACMP – ЧАСТЬ 3 из 11		**ДАТА: июль 2005**
ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ КЛАСТЕРА		
КОМПОНЕНТ ПО	ВЕРСИЯ	КОММЕНТАРИИ
AIX	5.3 ML02	Последняя версия AIX
RSCT	2.4.2.1	Последняя версия RSCT
HACMP	5.3 BASE	Версия GA
IBM SDD	1.6.0.2	ПО обеспечения множественных путей к хранилищу
ПРИЛОЖЕНИЕ	Тестовое приложение, Версия 1	Укажите версии своих приложений
КОММЕНТАРИИ	Для всего ПО выполнена проверка совместимости. При выполнении приложений с HACMP проблем не возникает. Лицензирование HACMP выполняется для четырех процессоров на каждом узле. Лицензирование приложений проверено, и лицензии приобретены для обоих серверов.	

3.6 Аспекты операционной системы

Помимо уровней и наборов файлов операционной системы AIX, существует еще несколько аспектов операционной системы, которые необходимо рассмотреть на этапе планирования.

Требования к дисковому пространству

HACMP требует наличия следующих объемов дискового пространства для установки в группе томов rootvg:

- **/usr** требует **82 Мб** свободного пространства для полной установки HACMP;
- / (root) требует **710 Кб** свободного пространства.

Кроме того, рекомендуется выделить приблизительно 100 Мб свободного пространства в **/var** и **/tmp** для журналов HACMP. (Требуемое пространство зависит от количества узлов в кластере, которое влияет на размер сообщений, записываемых в различные журналы HACMP.)

Синхронизация времени

Синхронизация времени между узлами кластера важна как для приложения, так и для журналов HACMP. Она относится к стандартным задачам системного администратора, и мы рекомендуем использовать NTP-сервер или другую процедуру синхронизации времени между узлами кластера.

> **Примечание.** Синхронизация времени между узлами особенно полезна для аудита и отладки проблем в кластере.

Настройки операционной системы

Для HACMP не требуется дополнительных настроек операционной системы. Используйте обычную настройку AIX в соответствии с рабочей нагрузкой приложений.

3.7 Планирование безопасности

Защита узлов кластера (и приложения) от несанкционированного доступа является важным фактором общей доступности системы. Существуют некоторые общие аспекты безопасности, а также аспекты, связанные с HACMP, которые мы рассмотрим в этом разделе.

3.7.1 Безопасность кластера

Кластеру HACMP нужен способ аутентификации на всех своих узлах для выполнения удаленных команд, связанных с верификацией кластера, синхронизацией и некоторыми административными операциями (C-SPOC).

Защита кластера необходима для предотвращения неавторизованного доступа к узлам кластера. Начиная с HACMP 5.1 дополнительная защита кластера обеспечивается новым механизмом безопасности, использующим возможности демона коммуникаций кластера (clcomdES).

Была устранена зависимость от AIX rsh (и таким образом, от файла /.rhosts). Так как некоторые внешние для HACMP команды, например пользовательские скрипты, могут все еще требовать удаленного выполнения команд с использованием r-команд, вам нужно проанализировать, есть ли необходимость оставить файл ~/.rhosts.

Межузловая связь в HACMP осуществляется с использованием демона кластера (clcomdES), что устраняет необходимость в «классических» удаленных командах AIX. Более подробные объяснения механизма работы clcomdES см. далее в этой главе.

Режимы аутентификации подключения в HACMP

- Стандартный режим безопасности:
  - стандартный режим безопасности (standard security mode) используется по умолчанию;
  - реализуется непосредственно демоном коммуникаций кластера (clcomdES);
  - использует информацию об узлах и адаптерах, хранящуюся в ODM-классах HACMP и файле /usr/es/sbin/cluster/etc/rhosts, для определения допустимых партнеров.
- Расширенный режим безопасности (kerberos):
  - режим безопасности Kerberos (enhanced security mode) доступен только для кластеров HACMP, реализованных в кластере SP;
  - используется метод аутентификации Kerberos.

Для повышения безопасности можно также использовать VPN-туннели между узлами кластера. В этом случае трафик clcomdES для IP-интерфейсов/адресов, сконфигурированных в HACMP, направляется через VPN-туннели, предоставляемые AIX. При использовании VPN следует применять постоянные адреса для VPN-туннелей. Конфигурирование VPN сначала выполняется в AIX, а потом в HACMP. Для упрощения конфигурирования HACMP предоставляет меню SMIT.

В стандартном режиме безопасности при удаленном выполнении команд HACMP в /usr/es/sbin/cluster используется принцип наименьших привилегий. При этом ни одна команда не может быть выполнена на удаленном узле с привилегиями «root», кроме команд, перечисленных в /usr/es/sbin/cluster. Эти команды HACMP считаются доверенными, и для них допускается запуск с привилегиями «root»; все остальные команды выполняются под учетной записью «nobody».

Для управления межузловыми коммуникациями демону коммуникаций кластера необходим список допустимых IP-меток или адресов кластера. Существует два способа предоставить эту информацию:

- автоматическое конфигурирование узлов (используется по умолчанию);
- индивидуальное конфигурирование узлов (вручную).

Автоматическое конфигурирование узлов

Если вы впервые осуществляете конфигурирование HACMP, файл /usr/es/sbin/cluster/etc/rhosts на узле является пустым. Демон clcomdES должен выполнить аутентификацию IP-адреса входящего подключения, чтобы убедиться, что оно исходит от узла в кластере; правила проверки адресов основаны на следующем процессе:

- Если файл /usr/es/sbin/cluster/etc/rhosts пуст и на этом узле не определен кластер HACMP, то первое подключение с другого узла будет аутентифицировано и принято. Содержимое файла /usr/es/sbin/cluster/etc/rhosts будет изменено; оно будет

включать все базовые адреса сетевых адаптеров, доступные для ping-опроса с запрашивающего узла.

- Если на узле уже определен кластер (ODM-класс HACMPcluster не пуст), то clcomdES ищет путь для связи (IP-адрес) в ODM-классе HACMPnode, а затем в HACMPadapter. Если он находит допустимый путь для связи, он берет первое вхождение (сначала из HACMPnode, затем из HACMPadapter); в противном случае выполняется поиск допустимого IP-адреса в файле /usr/es/sbin/cluster/etc/rhosts.
- Если clcomdES не может выполнить аутентификацию входящих подключений, он выдаст ошибку и вам нужно будет вручную исправить файл /usr/es/sbin/cluster/etc/rhosts, после чего повторно запустить демон clcomdES (stopsrc -s clcomdES, startsrc -s clcomdES)

Обычно пользователю не приходится вручную заполнять файл rhosts; это делает clcomdES. Так как после установки этот файл пуст, он будет заполнен при первом подключении с другого узла. Первое подключение обычно выполняется с целью верификации и синхронизации, после чего происходит заполнение ODM-классов HACMPnode и HACMPadapter. После синхронизации кластера файл rhosts можно очистить, но не удалить. Информация из классов HACMPnode и HACMPadapter затем используется для аутентификации clcomd.

Внимание!
- Чтобы гарантировать, что неавторизованный хост не подключится к узлу между установкой программного обеспечения HACMP и инициированием подключения от одного узла кластера к другому, можно вручную заполнить (отредактировать) файл /usr/es/sbin/cluster/etc/rhosts, добавив в него одну или несколько IP-меток/адресов (являющихся частью кластера).
- Если впоследствии вы решите переделать конфигурацию кластера (начать с самого начала или изменить базовые IP-адреса узлов), рекомендуется также очистить содержимое файла rhosts (НЕ УДАЛЯЯ ЕГО) на ВСЕХ узлах, на которых планируется реализовать кластер.

Индивидуальное конфигурирование узлов

В качестве альтернативного решения, если вас в особенности интересует сетевая безопасность (построение кластера может выполняться на основе незащищенной сети), можно поместить все IP-адреса/метки в файл /usr/es/sbin/cluster/etc/rhosts до конфигурирования кластера.

При установке HACMP этот файл создается пустым и имеет разрешения чтения и записи только для пользователя «root».

Примечание. Убедитесь, что каждые IP-адрес/метка допустимы для кластера, иначе в файл /var/hacmp/clcomd/clcomd.log будет записана ошибка.

Настройка файла /usr/es/sbin/cluster/etc/rhosts

1. Откройте файл /usr/es/sbin/cluster/etc/rhosts на узле под учетной записью «root».

2. Отредактируйте файл, добавив в него все возможные IP-адреса сетевых интерфейсов каждого узла.

3. Вводите по одной IP-метке или адресу в каждой строке.

> **Примечание.** Если вы отключите демон коммуникаций кластера или полностью УДАЛИТЕ файл /usr/es/sbin/cluster/etc/rhosts, программы, требующие межузловой связи, такие, как C-SPOC, средства верификации и синхронизации кластера, наборы файлов и средства аутентификации и шифрования сообщений, работать не будут. По этой причине необходимо обеспечить постоянное выполнение clcomdES.

3.7.2 Администрирование пользователей

Большинство приложений требует, чтобы информация о пользователях (идентификатор пользователя, членство в группе, идентификатор группы) была согласованной на всех узлах кластера, чтобы пользователи могли подключиться к работающим узлам без возникновения проблем.

Это особенно важно в ситуациях перемещения при сбое (перехвата). Очень важно, чтобы пользователь приложения был способен осуществлять доступ к общим файлам с любого узла в кластере. Обычно это означает, что идентификатор пользователя (UID) и идентификатор группы (GID), связанные с приложением, должны быть одинаковы на всех узлах.

При подготовке к конфигурированию кластера важно это учитывать и в случае необходимости исправить; в противном случае могут возникнуть проблемы с обслуживанием во время перемещения при сбое.

После установки HACMP содержит средства управления учетными записями пользователей и групп AIX 5L в кластере HACMP. Также имеется утилита для авторизации заданных пользователей и изменения их пароля на всех узлах в кластере HACMP.

> **Внимание!** При управлении учетными записями пользователей с применением таких утилит, как Network Information Service (NIS), средств управления пользователями PSSP или Distributed Computing Environment (DCE) Manager, НЕ употребляйте средства управления пользователями HACMP. Применение средств управления пользователями HACMP в такой среде может вызвать серьезную несогласованность в базах данных аутентификации пользователей.

3.7.3 Группа HACMP

При установке HACMP, если группа hacmp не существует, она будет создана. При создании группы HACMP просто берет следующий доступный идентификатор GID для группы hacmp.

> **Примечание.** Если вы предпочитаете контролировать идентификатор GID группы hacmp, рекомендуем создать группу hacmp перед установкой наборов файлов HACMP.

3.7.4 IP-порты HACMP

Помимо портов, перечисленных в файле /etc/services, следующие службы также требуют использования портов, однако эти порты выбираются произвольным образом при запуске процесса. В настоящее время нет способа выделения определенных портов, просто помните об их наличии. Для примера представлены типичные номера портов (однако при необходимости они могут быть изменены):

- #clstrmgr 870/udp,
- #clstrmgr 871/udp,
- #hatsd 32789/udp,
- #clinfo 32790/udp.

3.7.5 Планирование наборов файлов HACMP

HACMP требует, чтобы определенные файлы были одинаковыми на всех узлах кластера. К таким файлам относятся скрипты обработки событий, скрипты приложений, некоторые файлы конфигурации AIX 5L и файлы конфигурации HACMP. Средство HACMP File Collections позволяет выполнять автоматическую синхронизацию этих файлов между узлами кластеров, выдавая предупреждения в случае неожиданных результатов (например, если один или несколько файлов из набора были удалены или имеют нулевую длину на одном или нескольких узлах кластера).

Управление этими наборами файлов можно осуществлять через меню SMIT. С помощью SMIT можно добавлять, удалять и изменять наборы файлов для соответствия вашим потребностям.

Стандартные наборы файлов HACMP

При установке HACMP происходит установка следующих наборов файлов (file collections):

- Configuration_Files,
- HACMP_Files.

Configuration_Files

Набор Configuration_Files представляет контейнер для следующих важных системных файлов:

- /etc/hosts,
- /etc/services,
- /etc/snmpd.conf,
- /etc/snmpdv3.conf,
- /etc/rc.net,
- /etc/inetd.conf,

- /usr/es/sbin/cluster/netmon.cf,
- /usr/es/sbin/cluster/etc/clhosts,
- /usr/es/sbin/cluster/etc/rhosts.

Можно изменять опции распространения для этого набора файлов, а также добавлять и удалять файлы из этого набора файлов.

HACMP_Files

Набор HACMP_Files представляет контейнер, в котором обычно находятся настраиваемые пользователем файлы конфигурации HACMP, в частности скрипты запуска/остановки, настраиваемые события и т. д. Этот набор файлов не может быть удален или изменен, и файлы из этого набора нельзя удалить, изменить или добавить.

> **Примечание.** Например, при определении сервера приложений в HACMP (создании скриптов запуска, остановки и необязательных скриптов мониторинга) HACMP автоматически включает эти файлы в набор HACMP_Files.

3.8 Планирование сетей кластера

Конфигурация сети представляет основной компонент схемы кластера.

В стандартной кластерной среде клиенты осуществляют доступ к приложениям через сеть TCP/IP (обычно через Ethernet) с использованием сервисного адреса. HACMP обеспечивает высокую доступность этого сервисного адреса, а также при необходимости перемещение этого адреса между коммуникационными интерфейсами в одной сети. HACMP пересылает пакеты пульса между всеми коммуникационными интерфейсами (адаптерами) в сети, чтобы определить состояние адаптеров и узлов и предпринять необходимые меры.

Чтобы не допустить возникновения единой точки отказа в сетевом протоколе TCP/IP, а также разделения кластера, HACMP также использует отличные от IP сети типа «точка-точка» для мониторинга пульса. Это позволяет HACMP определить причину отказа, например отказ TCP/IP или отказ узла.

В этом разделе мы рассмотрим каждый тип сети и выберем соответствующие сетевые подключения и адреса.

На рис. 3.4 представлен обзор сетей, используемых в кластере.

Сеть Ethernet применяется для открытого доступа; к ней с каждого узла подключается несколько адаптеров. Эта сеть содержит базовые IP-адреса, постоянные IP-адреса и сервисные IP-адреса. Можно использовать несколько сетей, однако для простоты описания мы будем применять только одну.

Также показана последовательная сеть RS232. Она представляет собой сеть типа «точка-точка» с прямым подключением, с кабельным соединением между последовательными портами на каждом узле.

Также представлена сеть пульса через диски (а также сеть типа «точка-точка»). При использовании SAN-дисков можно легко реализовать мониторинг пульса через

диски, так как это не требует дополнительного оборудования. Кроме того, в multipath-конфигурации устройств применение **устройства vpath**, в отличие от hdisk, позволяет использовать возможности программного обеспечения SDD. Другими словами, в hdisk применяется только один путь к устройству, тогда как в vpath обычно используется несколько путей. Multipath-устройства могут быть сконфигурированы при употреблении нескольких дисковых адаптеров на узле, нескольких адаптеров хранилища или при одновременном использовании и того и другого.

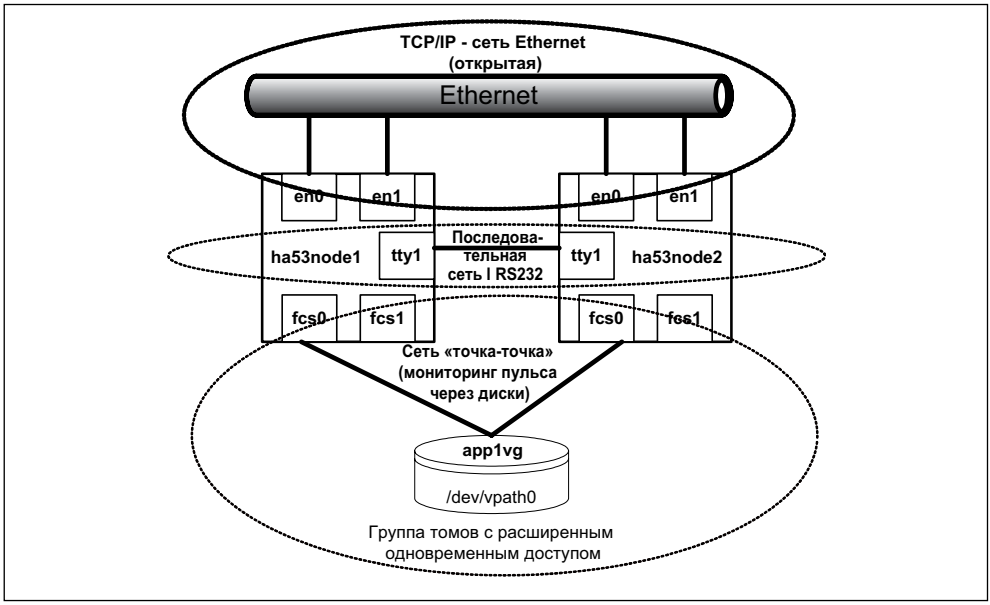

Рис. 3.4. Сети кластера HACMP

Все сетевые подключения применяются HACMP для мониторинга состояния сети, адаптеров и узлов кластера.

В нашем примере мы планируем использовать сеть Ethernet и сеть пульса через диски, а не сеть RS232.

3.8.1 Терминология

Этот раздел содержит краткий обзор терминологии, применяемой в обсуждениях сетевых возможностей HACMP.

IP-метки

IP-метки (IP labels) представляют собой имена, связанные с IP-адресами, подлежащими разрешению системой (/etc/hosts, BIND и т. д.).

Сервисная IP-метка/адрес

Сервисная IP-метка/адрес (service IP label/address) представляет собой IP-адрес или метку, через которую осуществляется обслуживание. Обычно она представляет собой

адрес, используемый клиентами для доступа к приложению. Она может быть привязана к узлу либо совместно использоваться узлами, и HACMP обеспечивает его высокую доступность.

Постоянная IP-метка/адрес

Постоянная IP-метка/адрес (persistent IP label/address) – это IP-синоним, привязанный к узлу, управляемый HACMP. Другими словами, постоянный синоним никогда не перемещается на другой узел; иногда он называется *привязанным к узлу (node-bound)*.

Коммуникационный интерфейс

Коммуникационный интерфейс (communication interface) – это физический интерфейс, поддерживающий протокол TCP/IP. В качестве примера можно привести Ethernet-адаптер. Представлен IP-меткой времени загрузки или базовой IP-меткой.

Коммуникационное устройство

Коммуникационное устройство (communication device) – это физическое устройство, представляющее сторону отличной от IP сети типа «точка-точка». Например: /dev/tty1 и /dev/vpath0.

Коммуникационный адаптер

Коммуникационный адаптер (communication adapter) – это физический адаптер X25, высокая доступность которого поддерживается HACMP.

Сетевая интерфейсная карта (NIC)

Сетевая интерфейсная карта (Network Interface Card, NIC) – это обычный физический адаптер, используемый для предоставления доступа к сети, например Ethernet-адаптер считается сетевой интерфейсной картой.

3.8.2 Общие аспекты сети

Данный раздел содержит множество аспектов, которые следует помнить при разработке конфигурации сети.

Поддерживаемые типы сетей

HACMP позволяет осуществлять межузловую связь в TCP/IP-сетях следующих типов. (следует заметить, что наиболее часто используемой сетью является Ethernet):

- Ethernet;
- Token-Ring;
- Fiber Distributed Data Interchange (FDDI);
- ATM и эмуляция ATM LAN;
- SP Switch1 и SP Switch2;
- Etherchannel (или 802.3ad Link Aggregation).

Следующие TCP/IP-сети **не** поддерживаются:

- Средство VIPA (Virtual IP Address) в системе AIX 5L;
- Serial Optical Channel Converter (SOCC);
- SLIP;
- FC Switch (FCS);
- IBM HPS (High Performance Switch);
- 802_ether;
- IP V6.

Можно сконфигурировать мониторинг пульса посредством сетей «точка-точка» следующих типов:

- последовательная сеть RS232;
- мониторинг пульса через диски (через диски в режиме расширенного одновременного доступа);
- Target mode SSA (почти устарел);
- Target mode SCSI (устарел).

Сетевые подключения

HACMP требует, чтобы каждый узел в кластере имел как минимум одно прямое немаршрутизируемое сетевое подключение ко всем остальным узлам. Эти сетевые подключения передают сообщения пульса между узлами кластера для определения состояния всех узлов кластера, сетей и сетевых интерфейсов.

HACMP также требует, чтобы все коммуникационные интерфейсы для заданной сети кластера были определены в одной физической сети, осуществляли маршрутизацию пакетов и получали ответы друг от друга без помех от другого сетевого оборудования.

Не используйте интеллектуальные коммутаторы, маршрутизаторы или другое сетевое оборудование, которые не осуществляют прозрачную передачу широковещательной рассылки UDP и других пакетов между всеми узлами кластера.

Мосты, концентраторы и прочие пассивные устройства, не изменяющие поток пакетов, можно благополучно размещать между узлами кластера, а также между узлами и клиентами.

На рис. 3.5 показана физическая конфигурация Ethernet с двумя Ethernet-адаптерами на каждом узле, подключенными через два коммутатора; все узлы сконфигурированы в одной физической сети (VLAN). Это иногда называется расположением в одном «домене коллизий» MAC.

Etherchannel

HACMP поддерживает применение Etherchannel (или Link Aggregation) для подключения к сети Ethernet. Etherchannel может быть полезен, если вы посчитаете нужным использовать несколько Ethernet-адаптеров как для увеличения пропускной способности сети, так и для перемещения при сбое, но при этом также требуется сохранить простоту конфигурации HACMP. При употреблении Etherchannel можно просто задать интерфейс

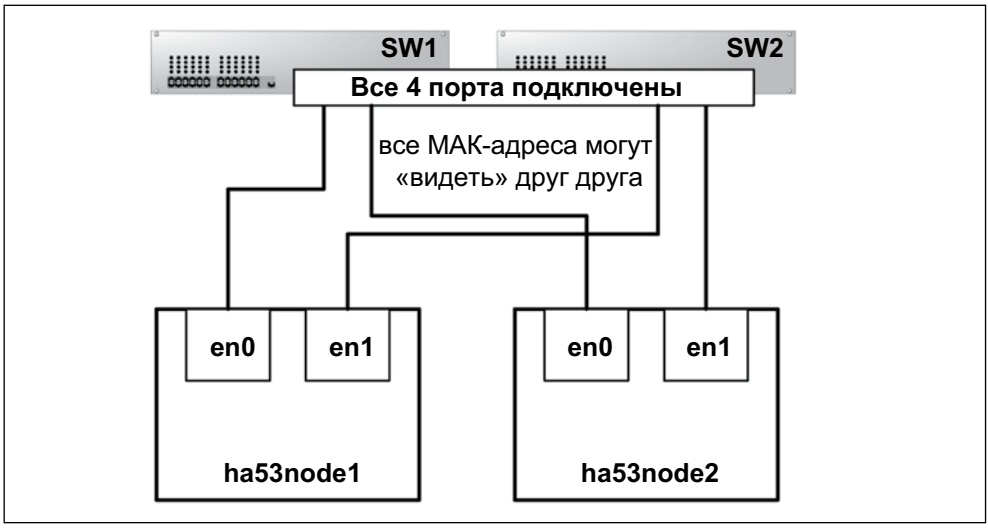

Рис. 3.5. Подключения к коммутаторам в сети Ethernet

Etherchannel в качестве коммуникационного интерфейса; любые отказы в сети Ethernet, за исключением отказа самой сети Ethernet, можно обрабатывать без участия HACMP.

EtherChannel представляет собой технологию агрегации сетевых портов, позволяющую объединять несколько Ethernet-адаптеров для создания одного псевдо-Ethernet-устройства. Например, ent0 и ent1 можно объединить в адаптер EtherChannel под названием ent3; затем можно настроить IP-адрес для интерфейса en3. Система считает эти объединенные адаптеры одним адаптером. Таким образом, IP-адрес настраивается для них как для любого Ethernet-адаптера. Кроме того, всем адаптерам в EtherChannel назначается один аппаратный (MAC) адрес, так что для удаленных систем они представляются как один адаптер.

Основное преимущество технологии EtherChannel состоит в том, что она использует сетевую пропускную способность всех адаптеров. При отказе адаптера сетевой трафик автоматически передается на следующий доступный адаптер без разрыва существующих пользовательских подключений. После восстановления адаптер автоматически возвращается к работе.

В дополнение к функции агрегации также можно выполнить назначение резервного адаптера. Этот адаптер конфигурируется как часть Etherchannel, но остается неактивным, пока не откажут все основные адаптеры. Это называется резервированием сетевого интерфейса (Network Interface Backup, NIB).

При планировании сети Etherchannel необходимо учитывать следующее:

- основные (агрегированные) линии должны быть подключены к одному коммутатору;
- необходимо настроить сетевой коммутатор, указав, какие порты использует Etherchannel;
- резервный сетевой интерфейс должен быть подключен к отдельному коммутатору;

- использование адаптеров с разной скоростью передачи не поддерживается.

На рис. 3.6 представлена простая конфигурация Etherchannel с двумя адаптерами, подключенными к одному коммутатору, и с резервным адаптером, подключенным к отдельному коммутатору. В этой конфигурации HACMP настраивается на использование только одного базового адаптера на узел (en3). В нашем примере базовый, постоянный и сервисный IP-адреса будут находиться на одном адаптере, пока не произойдет перемещение при сбое на другой узел.

На рис. 3.7 показана несколько более сложная конфигурация, в которой каждый узел имеет два адаптера Etherchannel, по одному на каждый коммутатор. При такой конфигурации HACMP использует адаптеры Etherchannel в качестве базовых.

Примечание. Etherchannel может содержать от 1 до 8 основных адаптеров при одном резервном адаптере на Etherchannel.
Это позволяет использовать только один адаптер в качестве основного и один адаптер в качестве резервного, если требуется обрабатывать отказы адаптера или коммутатора без использования HACMP.

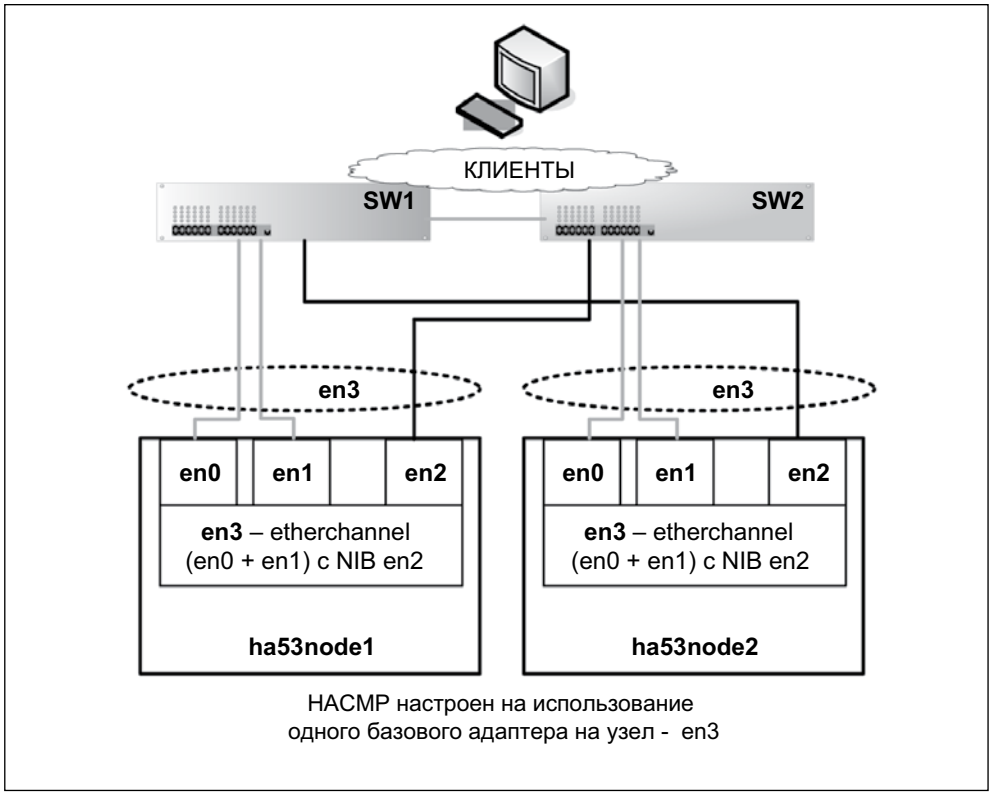

Рис. 3.6. Etherchannel с резервированием сетевого интерфейса

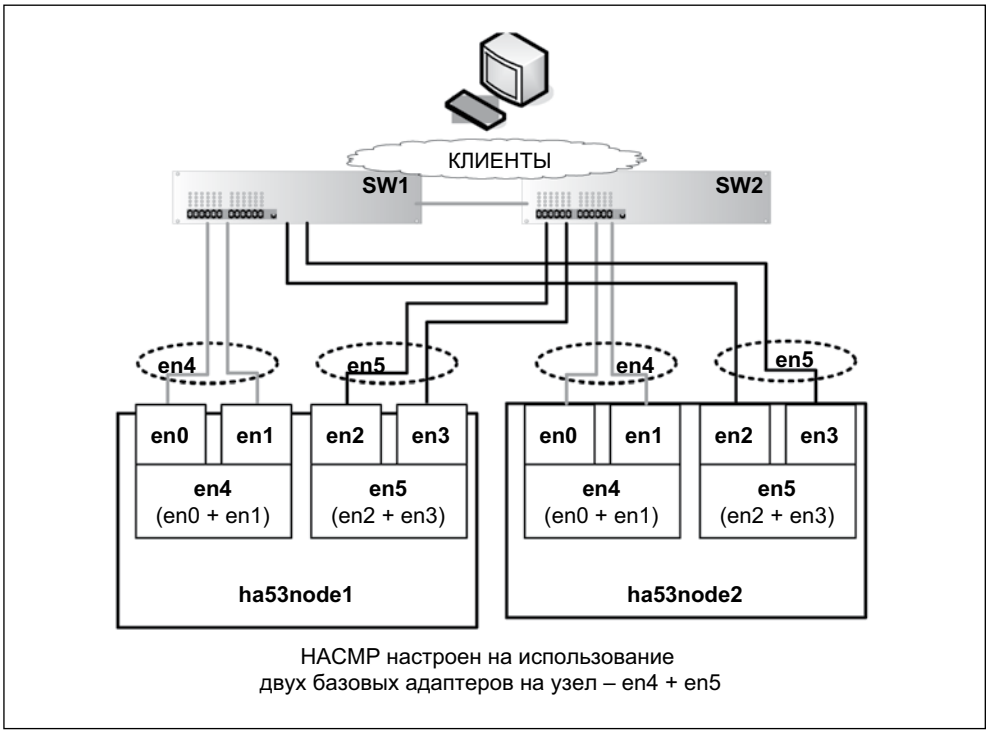

Рис. 3.7. Конфигурация с несколькими адаптерами Etherchannel

Имена хостов и имена узлов

Обычно имя хоста совпадает с именем узла HACMP. При использовании пути конфигурации «Standard» («Базовая») HACMP получает с узла имя хоста и использует его в качестве имени узла. При использовании расширенного пути конфигурирования (Extended Configuration path) можно задать имя узла самостоятельно.

В том случае, если приложение требует переноса имени хоста в AIX 5L в случае перемещения при сбое на другой узел вместе с приложением, следует использовать скрипты пред- и постобработки для изменения имени хоста таким образом, чтобы оно соответствовало сервисной IP-метке при перемещении группы ресурсов, содержащей это приложение, на другой узел.

> **Внимание!** Если вы планируете использовать DLPAR, имя хоста в AIX, имя узла кластера и имя HMC LPAR должны совпадать.

Мы рекомендуем вам избегать таких ситуаций, так как они ограничивают варианты перемещения при сбое. Например, вы не можете изменить имя хоста на резервном узле, если на нем уже выполняется приложение. В этом случае потребуется сконфигурировать дежурный узел.

/etc/hosts

IP-адрес и связанная с ним метка (имя) должны быть заданы в файле /etc/hosts. Мы рекомендуем выбрать один из узлов кластера для выполнения всех изменений в этом файле, а затем использовать FTP или наборы файлов HACMP для распространения файла /etc/hosts на другие узлы.

> **Примечание.** Мы настоятельно рекомендуем провести тестирование прямого и обратного разрешения имен на всех узлах в кластере и соответствующих пунктах управления оборудованием (Hardware Control Points, HMC[1]). Все эти компоненты должны выполнять разрешение имен идентичным образом, в противном случае могут возникнуть проблемы безопасности и другие проблемы, связанные с разрешением имен.

IP-метки

IP-метка представляет собой IP-адрес, сконфигурированный на NIC в дополнение к базовому IP-адресу NIC. Использование IP-синонимов является функцией AIX 5L, поддерживаемой HACMP. AIX 5L поддерживает использование нескольких IP-синонимов на NIC, как в одной, так и в разных подсетях.

> **Примечание.** AIX 5L допускает конфигурирование IP-синонимов с разными масками подсети для интерфейса, однако в HACMP эта функция еще не поддерживается.

Постоянные IP-адреса (синонимы)

Основная цель использования постоянных синонимов состоит в том, чтобы обеспечить доступ к узлу с отключенными службами HACMP. Постоянный синоним представляет собой маршрутизируемый адрес, доступ к которому сохраняется, пока работает узел. Этот синоним требуется сконфигурировать через HACMP. При запуске HACMP выполняется проверка доступности синонима, и если он недоступен, HACMP конфигурирует его на доступном адаптере в выделенной сети. Если синоним уже доступен, HACMP сохраняет его.

> **Важно!** Если на узле существует постоянный IP-адрес, он ДОЛЖЕН быть синонимом, а НЕ базовым адресом адаптера.

Постоянный синоним

- Всегда находится на одном узле (привязан к узлу).
- Сосуществует с другими IP-метками на интерфейсе.
- Не требует установки дополнительного физического интерфейса на узле.
- Не входит в какую-либо группу ресурсов.

---

[1] Имеется в виду Hardware Management Console. *Примеч. науч. ред.*

Мы рекомендуем выполнить конфигурирование постоянного синонима через AIX («smitty inetalias»), прежде чем выполнять конфигурирование HACMP. После этого в качестве пути для связи с узлом рекомендуется использовать постоянный синоним, а не базовые адаптеры. Это дает вам свободу при необходимости впоследствии изменять базовые IP-адреса (при изменении адреса базового адаптера следует обязательно проверить файл /usr/es/sbin/cluster/etc/rhosts). После конфигурирования HACMP следует добавить постоянный синоним в конфигурацию HACMP.

> **Примечание.** Постоянный IP-адрес назначается HACMP на одном коммуникационном интерфейсе, входящем в сеть, определенную в HACMP.

Рис. 3.8 иллюстрирует понятие постоянного адреса. Заметьте, что он представляет собой просто еще один IP-адрес, сконфигурированный на одном из базовых интерфейсов. Команда **`netstat`** отображает его как дополнительный IP-адрес адаптера.

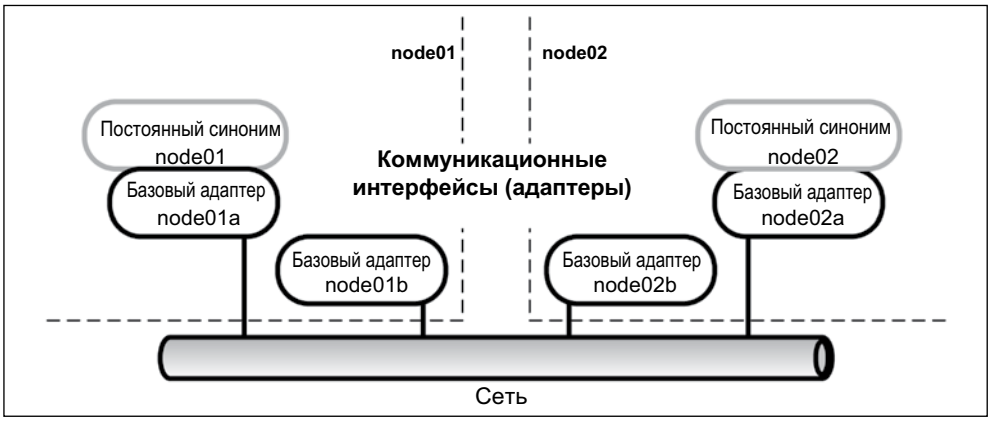

Рис. 3.8. Постоянные синонимы

Подсети

Требования к подсетям колеблются в зависимости от выбранной конфигурации (перехват IP-адреса посредством замены или посредством синонимов), однако в HACMP маски подсетей для всех коммуникационных интерфейсов, определенных в HACMP, должны быть одинаковыми.

При перехвате IP-адреса посредством замены:

- базовый и сервисный IP-адреса на основном адаптере должны относиться к одной подсети;
- все базовые IP-адреса на дополнительных адаптерах должны относиться к разным подсетям (должны быть отличными друг от друга и от основного адаптера).

> **Примечание.** Адреса базового адаптера также называются «загрузочными» IP-адресами, если не используется мониторинг пульса через IP-синонимы (описано далее в этом разделе).

При перехвате IP-адреса посредством синонимов:

- все базовые IP-адреса на узле должны относиться к разным подсетям (если не используется мониторинг пульса через IP-синонимы);
- все сервисные IP-адреса должны относиться к подсети, отличной от любой из базовых подсетей;
- сервисные IP-адреса могут относиться как к одной, так и к различным подсетям;
- постоянный IP-адрес может относиться как к той же подсети, к которой относится и сервисный IP-адрес, так и к другой подсети;
- если вы решите использовать мониторинг пульса через IP-синонимы, то базовые IP-адреса могут относиться как к одной подсети, так и к различным подсетям, так как при этом HACMP не осуществляет их мониторинг, а только мониторинг синонимов, заданных в HACMP.

Аспекты шлюза (маршрута) по умолчанию

В зависимости от конфигурации вашей IP-сети при управлении интерфейсами из HACMP может произойти потеря маршрута по умолчанию.

Если маршрут по умолчанию привязан к одной из подсетей базового адреса, то при отказе этого адаптера произойдет потеря маршрута по умолчанию.

Чтобы не допустить возникновения такой ситуации, мы рекомендуем использовать постоянный адрес и привязать маршрут по умолчанию к этой подсети. Постоянный адрес будет активным, пока будет активным узел; таким образом, маршрут по умолчанию также будет активным.

Если вы решите этого не делать, то придется создать скрипт постобработки события, чтобы восстановить маршрут по умолчанию при возникновении проблемы.

Обновление ARP-кеша

При изготовлении каждой сетевой интерфейсной карте (Network Interface Card, NIC) назначается уникальный аппаратный адрес – MAC-адрес (Media Access Control). MAC-адрес представляет адрес, используемый сетевыми драйверами для отправки пакетов между сетевыми картами в локальной сети. Большинство систем ведут список, содержащий недавно использовавшиеся IP-адреса и соответствующие им MAC-адреса и называемый ARP-кешем. Так как HACMP может перемещать IP-адреса между сетевыми картами, некоторые записи клиентского ARP-кеша могут быть неточными.

После возникновения события в кластере, узлы и сетевые устройства HACMP, поддерживающие смешанный режим (promiscuous mode), автоматически обновляют свои ARP-кеши. Клиенты и сетевые устройства, не поддерживающие смешанный режим, продолжают содержать неправильные записи. Это можно исправить одним из двух способов:

- Посредством использования альтернативных аппаратных адресов. Нужно настроить HACMP на перемещение как IP-адреса, так и MAC-адреса (работает только при перехвате IP-адреса посредством замены).
- Обновлением ARP-кеша посредством использования записей ping_client_list в clinfo.rc.

HACMP в коммутируемой сети

При использовании VLAN все интерфейсы, определенные в HACMP для заданной сети, должны относиться к одной VLAN. Таким образом, все адаптеры в одной сети должны быть подключены к одной физической сети и вести обмен данными друг с другом («видеть» MAC-адреса друг друга).

> **Примечание.** НЕ все адаптеры должны содержать адреса, маршрутизируемые вне VLAN. Только сервисные и постоянные адреса должны быть маршрутизируемыми. Адреса базового адаптера и любые синонимы, используемые для мониторинга пульса, не должны обязательно быть маршрутизируемыми вне VLAN, так как они неизвестны клиентской стороне.

Нужно убедиться, что коммутатор обеспечивает своевременную реакцию на ARP-запросы. Для многих моделей коммутаторов это означает отключение следующих функций:

- алгоритм spanning tree,
- portfast,
- uplinkfast,
- backbonefast.

Если требуется, чтобы алгоритм spanning tree был включен, то функция portfast также должна быть включена.

Настройки скорости передачи в Ethernet

Так как согласование скорости передачи может вызвать проблемы в некоторых комбинациях адаптеров и коммутаторов, рекомендуется не использовать автоматическое согласование, а устанавливать требуемые значения скорости и дуплексного режима.

3.8.3 Планирование перехвата IP-адреса

Перехват IP-адреса (IP Address Takeover, IPAT) представляет собой механизм, используемый HACMP для перемещения сервисных адресов между коммуникационными интерфейсами.

Существует два метода: перехват IP-адреса посредством замены (IPAT via replacement) и перехват IP-адреса посредством синонимов (IPAT via aliases). Конфигурация вашей сети зависит от применяемого метода управления интерфейсами.

Для любой новой инсталляции мы рекомендуем использовать перехват IP-адреса посредством синонимов, так как этот метод прост в реализации и более гибок, чем перехват IP-адреса посредством замены. Можно применять несколько сервисных адресов для одного адаптера в любое время; кроме того, в случае перемещения при сбое имеет место некоторая экономия времени, так как HACMP просто нужно добавить синоним, а не выполнять повторное конфигурирование базового IP-адреса адаптера, что гораздо быстрее.

Некоторые конфигурации потребуют использования мониторинга пульса через синонимы. Например, если оба локальных базовых адаптера относятся к одной подсети или же если все базовые адаптеры на всех узлах относятся к разным подсетям.

Все варианты подробно рассматриваются в следующем разделе. В нашем примере применяется перехват IP-адреса посредством синонимов и мониторинг пульса через синонимы.

Перехват IP-адреса посредством замены

Этот способ конфигурирования сетей HACMP является более традиционным. HACMP при запуске осуществляет замену загрузочного адреса на сервисный адрес.

Для кластера из двух узлов требуется использование по меньшей мере одной подсети на коммуникационный интерфейс на узел (при использовании одинаковой маски подсети во всех подсетях). Для кластера с несколькими коммуникационными интерфейсами на узел требуется следующее:

- базовый и сервисный адреса основного коммуникационного интерфейса должны относиться к одной подсети;
- базовые IP-адреса всех дополнительных коммуникационных интерфейсов должны относиться к различным подсетям (относительно друг друга и относительно основного интерфейса).

Преимущество перехвата IP-адреса посредством замены состоит в том, что оно разрешает выполнять перехват аппаратного адреса (Hardware Address Takeover, HWAT) вместе с перехватом IP-адреса посредством замены. Эта возможность позволяет осуществлять перемещение (локально администрируемого) MAC-адреса адаптера, соответствующего сервисному IP-адресу, вместе с IP-адресом на дежурный адаптер. Это устраняет необходимость обновления ARP-кеша на стороне клиента в случае переноса IP-адресов.

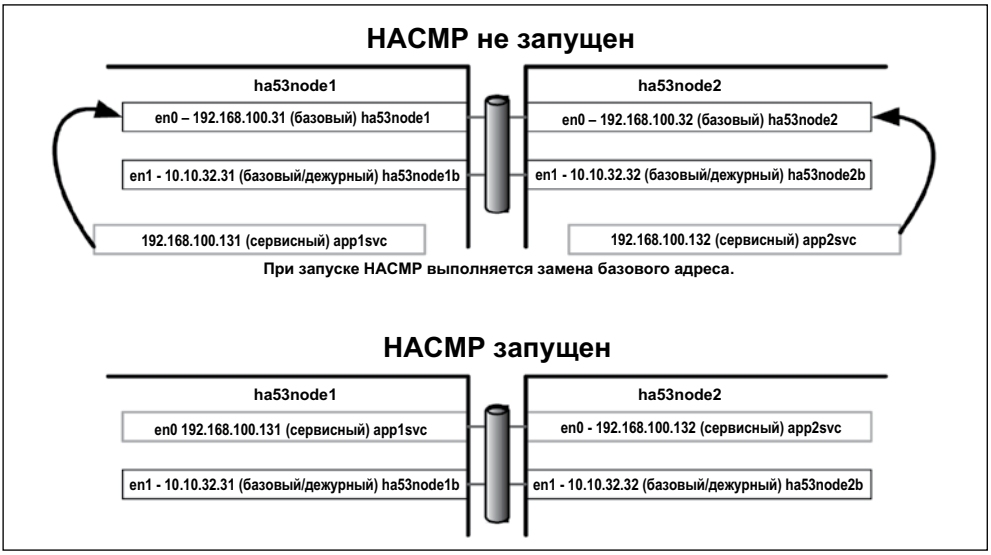

Рис. 3.9. Перехват IP-адреса посредством замены

На рис. 3.9 показано состояние сетевых адаптеров до и после запуска HACMP на узлах. Обратите внимание на то, что HACMP при запуске заменяет загрузочный/базовый адрес сервисным адресом. Перемещение при сбое выполняется на дополнительный (дежурный) адаптер(ы), где, опять же, базовый (дежурный) адрес заменяется сервисным адресом. Количество сервисных адресов ограничено количеством резервных адаптеров, определенных в той же сети HACMP.

Перехват IP-адреса посредством синонимов

Этот метод назначения сервисных адресов является более новым и более гибким, чем перехват IP-адреса посредством замены. Используя перехват IP-адреса посредством синонимов, можно осуществлять назначение нескольких IP-адресов одному коммуникационному интерфейсу.

HACMP позволяет использовать перехват IP-адреса посредством IP-синонимов для следующих типов сетей, поддерживающих gratuious ARP-запросы (в AIX):

- Ethernet;
- Token Ring;
- FDDI;
- SP Switch1 и SP Switch2.

> **Примечание.** Перехват IP-адреса посредством IP-синонимов не поддерживается в сетях ATM.

При запуске HACMP выполняется конфигурирование сервисного синонима поверх существующего базового IP-адреса доступного адаптера.

При использовании перехвата IP-адреса посредством синонимов следует учитывать следующие требования:

- Требования подсетей:
  - Каждый базовый адаптер должен относиться к отдельной подсети, чтобы можно было осуществлять мониторинг пульса. Базовые адреса не обязательно должны быть маршрутизируемыми вне кластера.

> **Примечание.** Это ограничение снимается при использовании мониторинга пульса через синонимы.

  - Сервисные адреса должны относиться к отдельной подсети относительно любой из базовых подсетей. Можно использовать несколько сервисных адресов, и все они могут относиться как к одной подсети, так и к различным подсетям.
  - Постоянный синоним может относиться либо к той же подсети, либо к другой подсети относительно сервисного адреса.
  - Все маски подсети должны быть одинаковыми.
- Несколько сервисных меток могут совместно существовать как синонимы для заданного интерфейса.
- Нельзя сконфигурировать перехват аппаратного адреса (Hardware Address Takeover, HWAT).

Мы рекомендуем использовать постоянный синоним и включить его в одну подсеть с маршрутом по умолчанию. Это обычно означает, что постоянный адрес должен быть включен в одну подсеть с сервисными адресами. Постоянный синоним можно использовать для доступа к узлу при отключенном HACMP, а также для преодоления проблем с маршрутом по умолчанию.

Можно выполнить настройку параметров размещения сервисных IP-меток, сконфигурированных в HACMP V5.3. Можно настроить размещение синонимов через меню SMIT с использованием следующих вариантов:

- **Без совместного размещения (Anti-Collocation).** Используется по умолчанию. HACMP распределяет сервисные IP-метки по всем доступным коммуникационным интерфейсам с использованием выбора по принципу наименьшей загруженности.
- **С совместным размещением (Collocation).** HACMP размещает все сервисные IP-метки на одном коммуникационном интерфейсе (NIC).
- **Без совместного размещения и с постоянной меткой (Anti-Collocation with persistent label).** HACMP размещает все сервисные IP-метки по всем активным коммуникационным интерфейсам, **не** содержащим постоянную IP-метку синонима. HACMP размещает сервисную IP-метку на интерфейсе, содержащем постоянную метку только в том случае, если другие сетевые интерфейсы недоступны. Если постоянные IP-метки не были сконфигурированы, HACMP позволяет выбрать метод размещения Anti-Collocation with Persistent (Без совместного размещения и с постоянной меткой), однако при этом выдается предупреждение и по умолчанию используется обычный метод без совместного размещения.

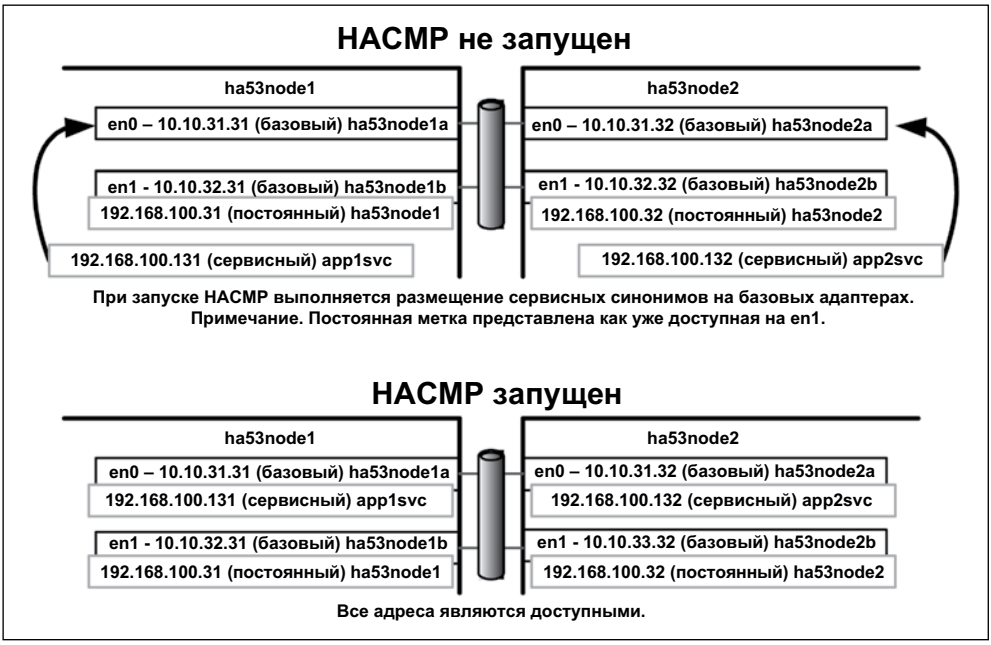

Рис. 3.10. Перехват IP-адреса посредством синонимов

- **С совместным размещением и с постоянной меткой (Collocation with persistent label).** Все сервисные IP-метки располагаются на одной сетевой карте, содержащей постоянную IP-метку. Этот вариант может быть полезен при конфигурациях виртуальной частной сети (VPN) с брандмауэром, где только один интерфейс имеет внешний выход и все IP-адреса (постоянные и сервисные) должны располагаться на одном коммуникационном интерфейсе. Если постоянные IP-метки не были сконфигурированы, HACMP позволяет выбрать метод размещения Collocation with Persistent (С совместным размещением и с постоянной меткой), однако при этом выдается предупреждение и по умолчанию используется обычный метод с совместным размещением.

На рис. 3.10 показано состояние сетевых адаптеров до и после запуска HACMP на узлах. Обратите внимание на то, базовые адреса не изменяются. HACMP добавляет на базовые адаптеры сервисные и постоянные синонимы. Постоянные адреса всегда доступны, тогда как сервисные метки добавляются и удаляются при запуске и остановке HACMP. Перемещение при сбое выполняется путем переноса сервисной метки на другой доступный коммуникационный интерфейс. В нашем примере только сеть 192.168.100/24 является маршрутизируемой вне кластера.

3.8.4 Мониторинг пульса через синонимы

HACMP требует использования отдельной подсети для мониторинга каждого базового адаптера. При конфигурации с применением двух Ethernet-адаптеров на узле требуется две подсети. При употреблении трех адаптеров требуется три подсети. Эти подсети не обязательно должны быть маршрутизируемыми вне сетей кластера.

Чтобы обеспечить средство мониторинга этих адаптеров (без изменения адреса базового адаптера), а также чтобы избежать возникновения проблем с подсетями, HACMP предоставляет функцию мониторинга пульса через синонимы. Этот метод не требует внесения каких-либо изменений в существующие базовые адреса. HACMP просто игнорирует базовые адреса и добавляет собственный набор синонимов для осуществления мониторинга пульса.

При использовании мониторинга пульса через IP-синонимы, IP-адреса, используемые при загрузке, могут располагаться либо в той же подсети, либо в других подсетях; однако IP-адрес, применяемый во время загрузки, должен располагаться в подсети, не включающей сервисные IP-метки. Как оказалось, если все адреса (базовые и сервисные) попадают в одну подсеть, возникают проблемы с маршрутизацией в связи с функцией чередования маршрутов (route striping) операционной системы AIX.

Чтобы установить мониторинг пульса через IP-синонимы, следует настроить параметр «IP Address Offset for Heartbeating over IP Aliases» («Смещение IP-адреса для мониторинга пульса через IP-синонимы») как часть конфигурирования сетей HACMP. IP-адреса, используемые для мониторинга пульса, определяются и назначаются системой HACMP с применением этого значения смещения. Маска подсети совпадает с той, которая используется для сервисных и несервисных адресов.

Например, можно в качестве параметра смещения IP-адреса употреблять значение 1.1.1.1. При использовании сети с двумя NIC на каждом узле следует добавить маску подсети 255.255.255.0; в результате получим следующие IP-синонимы мониторинга пульса:

- node01:
  - en0 - 1.1.1.1;
  - en1 - 1.1.2.1.
- node02:
  - en0 - 1.1.1.2;
  - en1 - 1.1.2.2.

IP-синонимы мониторинга пульса добавляются при запуске HACMP на узле и удаляются при остановке HACMP. Эти IP-синонимы используются только для сообщений пульса. Для них не требуется осуществлять маршрутизацию, и их не следует применять для какого-либо другого трафика. Маска подсети совпадает с используемой для сервисных и несервисных синонимов.

На рис. 3.11 показано состояние сетевых адаптеров до и после запуска HACMP на узлах. Обратите внимание на то, базовые адреса не изменяются. Помимо сервисных и постоянных синонимов, добавляемых HACMP на базовые адаптеры, также добавляются синонимы пульса. Последние удаляются при остановке HACMP вместе с сервисными синонимами. В нашем примере только сеть 192.168.100/24 является маршрутизируемой вне кластера.

Команда **netstat -i** выводит три IP-адреса для каждого адаптера при запущенном HACMP.

Рис. 3.11. Мониторинг пульса через синонимы

3.8.5 Планирование сети, отличной от IP

Сети типа «точка-точка» играют важную роль в обеспечении высокой доступности кластера. Достаточно небезопасно применять одну лишь TCP/IP-сеть для обеспечения доступности кластера и недопущения разделения кластера. Поэтому важно создавать сети типа «точка-точка». При использовании больших кластеров необходимо создавать соответствующие пути между всеми узлами в кластере.

Назначение топологии последовательной сети или сети «точка-точка» состоит в том, чтобы обеспечить достаточное количество путей между узлами кластера, чтобы RSCT могла правильно оценить серьезность сбоя в кластере.

Разделение кластера

Разделение кластера (также называемое *изоляцией узла* или *split brain*) возникает в том случае, когда узел HACMP перестает получать весь трафик пульса с другого узла (по всем доступным сетям) и предполагает, что на этом узле произошел отказ.

Проблема разделения кластера состоит в том, что узлы с одной стороны раздела воспринимают отсутствие пакетов пульса от узлов с другой стороны раздела как признак отказа узлов, генерируя события отказа для этих узлов. После этого узлы с каждой стороны кластера пытаются перехватить ресурсы (если настроен перехват) с узла, который на самом деле все еще активен и потому является законным владельцем этих ресурсов. Такие попытки перехвата могут привести к непредсказуемым результатам в кластере, например к повреждению данных в связи с реинициализацией дисков.

Наилучшая защита от возникновения подобных ситуаций состоит в том, чтобы использовать несколько сетей, как TCP/IP-сетей, так и сетей типа «точка-точка», чтобы обеспечить правильность оценки серьезности проблемы. Помните о том, что HACMP (в частности, RSCT) отправляет и получает пакеты пульса по всем доступным сетям, поэтому чем больше сетей, тем точнее HACMP сможет определить, имеет ли место отказ узла или отказ сети.

Следующий набор из четырех рисунков показывает, каким образом следует осуществлять добавление сетей в кластер, чтобы обеспечить наилучшую защиту от разделения кластера.

На рис. 3.12 показан кластер из четырех узлов с одним Ethernet-подключением к каждому серверу. Так как используется только одна сеть, то при потере любой связи часть кластера будет разделена. В примере показан разрыв связи между двумя Ethernet-коммутаторами, вызывающий отделение двух узлов слева от двух узлов справа. В данном случае могут возникнуть проблемы в результате попыток выполнить перехват ресурсов с активного узла.

Рис. 3.13 несколько более реалистичен; на нем представлены двойные Ethernet-подключения с каждого узла. Каждый Ethernet-адаптер подключен к отдельному коммутатору. В этом случае для разделения кластера необходимо, чтобы произошел отказ двух коммутаторов или же обоих Ethernet-подключений на узле. Однако сама по себе сеть TCP/IP остается единой точкой отказа.

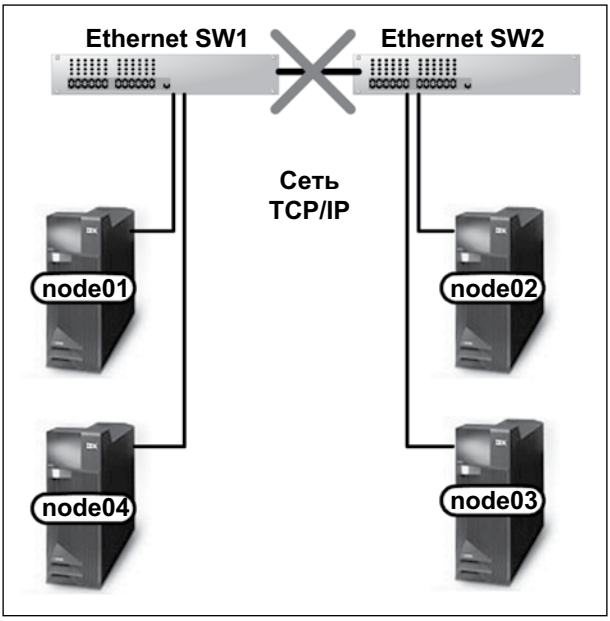

Рис. 3.12. Разделение кластера

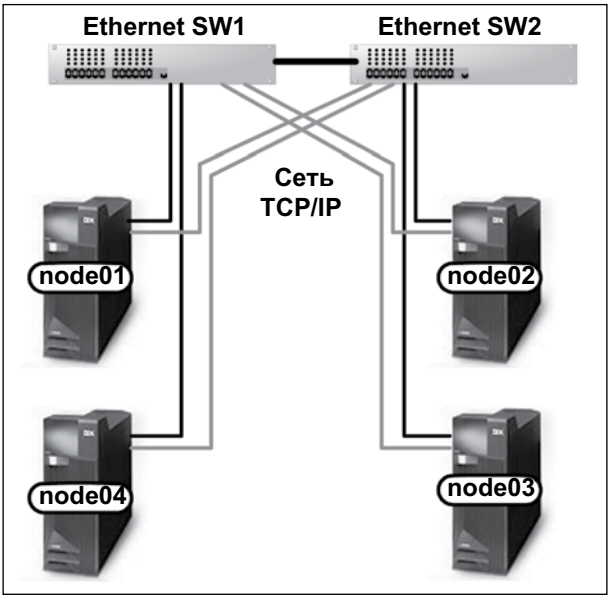

Рис. 3.13. Кластер без разделения с двойными Ethernet-подключениями

На рис. 3.14 представлена рекомендованная конфигурация. Имеются двойные Ethernet-подключения к нескольким Ethernet-коммутаторам, а также добавлена кольцевая сеть типа «точка-точка». В кольцевой сети каждый узел подключен к своим непосредственным соседям. При потере одного подключения RSCT все еще сможет подключиться ко всем оставшимся узлам. Для разделения кластера потребуется возникновение двойного отказа на узле и отказа сети TCP/IP.

Для построения надежной конфигурации рассмотрим реализацию звездной топологии. При такой конфигурации, помимо сети TCP/IP, каждый узел подключен ко всем другим узлам кластера сетями типа «точка-точка». Это позволяет обеспечить связь RSCT с работающими узлами при отказе нескольких узлов. Эта конфигурация представлена на рис. 3.15.

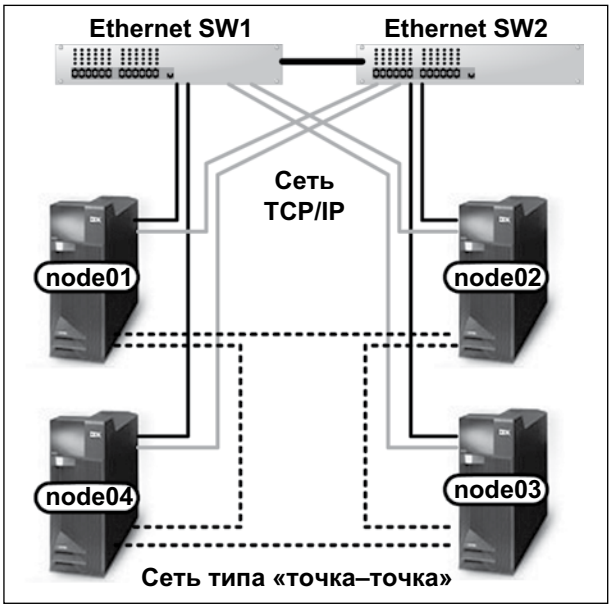

Рис. 3.14. Конфигурация с использованием Ethernet-сети и кольцевой сети типа «точка-точка»

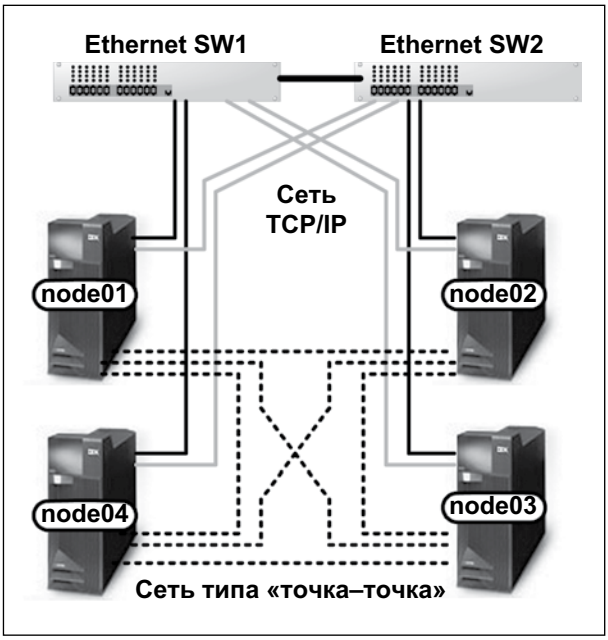

Рис. 3.15. Конфигурация с использованием сети Ethernet и сети типа «точка-точка» звездной топологии

3.8.6 Планирование последовательных сетей RS232

Сеть RS232 содержит по одному последовательному порту на каждом узле, подключенном через последовательный кабель. При наличии нескольких узлов потребуется несколько последовательных кабелей.

Если вы решите использовать последовательную сеть RS232, необходимо учитывать следующее:

- Некоторые серверы pSeries имеют ограничение на использование встроенных последовательных портов, некоторые порты недоступны, а некоторые порты должны назначаться группами (особенно в среде LPAR).
- Если доступные последовательные порты отсутствуют, а ваша запланированная конфигурация HACMP для этого узла использует сеть RS232, то требуется использовать специальный PCI-адаптер для узла кластера (LPAR).
- Все сети RS232, определенные в HACMP, автоматически настраиваются на использование последовательных портов на скорости 38 400 бод. В зависимости от длины последовательного кабеля RSCT поддерживает скорости передачи 38 400, 19 200, 9 600 бод.
- Для мониторинга пульса можно использовать любой последовательный порт, соответствующий следующим требованиям:
- оборудование должно поддерживать применение этого последовательного порта для подключения модема;
- последовательный порт должен быть свободен для исключительного использования в HACMP.

Кабель для соединения двух последовательных портов должен быть подключен как полный нуль-модемный кабель; он не поставляется по умолчанию вместе с оборудованием. рис. 3.16 показывает нуль-модемное подключение. Фактически используемые разъемы подключения зависят от оборудования; скорее всего, будут использоваться разъемы DB9, DB25 или RJ50.

Чтобы определить, соответствуют ли ваши последовательные порты установленным требованиям, см. документацию к оборудованию и сообщения о поддержке HACMP.

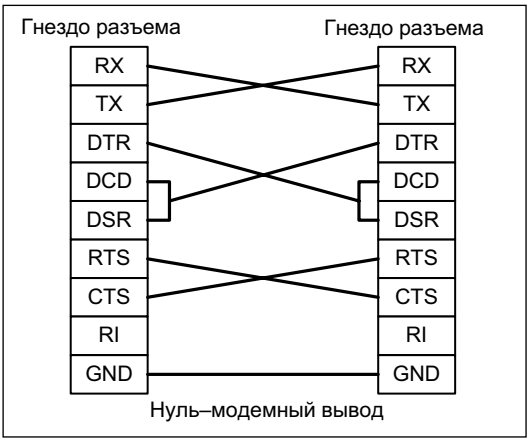

Рис. 3.16. Нуль-модемное кабельное подключение

3.8.7 Планирование мониторинга пульса через диски

Мониторинг пульса через диски представляет еще один тип сетей типа «точка-точка» для обнаружения отказов. В предыдущих версиях HACMP можно было сконфигурировать отличные от IP сети мониторинга пульса через диски SCSI или SSA путем настройки сетей Target Mode SCSI (TMSCSI) и Target Mode SSA (TMSSA) типа «точка-точка».

Начиная с HACMP 5.1, можно также сконфигурировать отличное от IP подключение мониторинга пульса через диски типа «точка-точка» с использованием любого общего диска, входящего в группу томов в режиме расширенного одновременного режима (enhanced concurrent mode, ECM).

В сети мониторинга пульса через диски два узла, подключенные к диску, периодически записывают сообщения пульса и считывают сообщения пульса (записанные другим узлом) с небольшого участка диска, незанятого данными. Хотя сеть пульса через диски соединяет только два узла, кластеры, содержащие больше двух узлов, допускают использование нескольких дисков для мониторинга пульса.

При употреблении SAN-диска в качестве общего диска следует рассмотреть вариант использования мониторинга пульса через диски по следующим причинам:

- можно использовать любой имеющийся общий диск (включая диски, подключенные к SAN).
- не требуется дополнительное оборудование или кабели.

Для того чтобы применять мониторинг пульса через диски, необходимы SAN-диски, доступные для обоих узлов и входящие в группу томов с расширенным одновременным доступом.

Любой общий диск из группы томов в режиме расширенного одновременного доступа может поддерживать подключение пульса типа «точка-точка». Каждый диск может поддерживать одно подключение между двумя узлами. Подключение использует общее дисковое оборудование в качестве пути для связи.

Сеть пульса через диски в кластере содержит:

- Два узла, каждый из которых имеет адаптер SAN. Узел может входить в любое количество сетей пульса с использованием одного диска.
- Диск в режиме расширенного одновременного доступа. Один диск может применяться только в одной сети пульса.

При выборе диска для использования при мониторинге пульса необходимо учитывать следующее:

- Диск, используемый для мониторинга пульса, должен быть членом группы томов в расширенном одновременном режиме. Однако группы томов, связанные с дисками, используемыми для мониторинга пульса, не обязательно должны быть определены как ресурсы в группе ресурсов HACMP.
- Диск, используемый для мониторинга пульса, не должен быть чрезмерно загружен, так как HACMP ожидает, что операции записи будут происходить в пределах заданных интервалов. Если вы решите применять диск со значительной нагрузкой ввода-вывода, необходимо увеличить значение параметра тайм-аута для сети пульса через диски. Вообще рекомендуется использовать диск, не осуществляющий более 60 операций поиска в секунду.

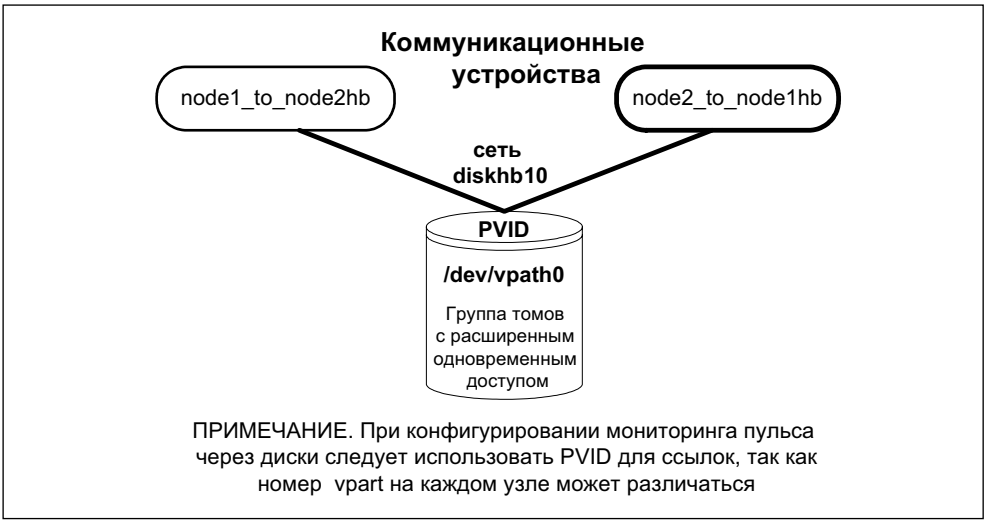

Рис. 3.17. Сеть пульса через диски

- При установке драйвера Subsystem Device Driver (драйвер устройства для серии DS8XXX) и связывании группы томов с расширенным одновременным доступом с активным устройством vpath необходимо убедиться в том, что коммуникационное устройство мониторинга пульса через диски определено для использования устройства /dev/vpath (а не связанного устройства /dev/hdisk); это позволит применять многопутевое программное обеспечение.
- Если в общей группе томов осуществляется зеркальное отображение, как минимум один диск в каждом зеркальном отображении должен использоваться для мониторинга пульса через диски.
- В сети пульса через диски рекомендуется применять один LUN (диск) на пару узлов на дисковую стойку.

На рис. 3.17 показаны основные компоненты сети пульса через диски.

Обратите внимание на то, что номер vpath может отображаться различным образом с разных узлов, что связано с нумерацией дисков в AIX. Поэтому рекомендуется проверять PVID, чтобы убедиться в том, что на всех узлах был выбран один и тот же диск.

Кроме того, рекомендуется запускать процесс обнаружения в HACMP и выбирать требуемые диски из выводимого списка.

3.8.8 Дополнительные аспекты планирования сети

Помимо конфигурирования топологии сети, есть еще два вопроса, которые следует рассмотреть при планировании кластера:

- взаимодействие HACMP с Domain Name Service (DNS) и Network Information Services (NIS);
- сетевые модули HACMP.

Все эти вопросы рассматриваются в данном разделе.

Взаимодействие HACMP с DNS и NIS

Чтобы обеспечить успешное и быстрое выполнение событий в кластере, HACMP отключает разрешение имен хоста в NIS или DNS во время обмена сервисных IP-меток путем установки следующей переменной окружения AIX 5L: NSORDER = local. Поэтому файл /etc/hosts на каждом узле кластера должен содержать все IP-метки, определенные в HACMP для всех узлов кластера.

После завершения операции обмена доступ к DNS восстанавливается.

Мы предлагаем поместить запись

hosts = local, bind4

в файл /etc/netsvc.conf, чтобы обеспечить чтение файла /etc/hosts перед попыткой поиска в DNS.

Сетевые модули

Каждая поддерживаемая сеть кластера имеет соответствующий сетевой модуль RSCT (также называемый *сетевым интерфейсным модулем – network interface module, NIM*), осуществляющий мониторинг трафика пульса через сеть кластера. Сетевые модули обеспечивают взаимное подключение в кластере, через которое диспетчеры кластера на всех узлах отправляют друг другу сообщения «keep-alive».

В новых версиях HACMP передает соответствующие параметры настройки в сетевые модули RSCT для поддержки связи через следующие типы сетей:

- Ethernet;
- последовательная сеть (RS232);
- сеть пульса через диски (с использованием дисков с расширенным одновременным доступом);
- Target-mode SCSI;
- Target-mode SSA;
- Token-Ring;
- FDDI;
- SP Switch;
- ATM.

Скорость обнаружения отказов

Скорость обнаружения отказов (failure detection rate) определяет, через какой промежуток времени подключение считается отказавшим. Скорость обнаружения отказов включает два компонента:

- Цикл отказа (cycle). Количество импульсов, которое может быть пропущено, прежде чем произойдет обнаружение отказа.
- Скорость пульса (hbrate). Количество секунд между импульсами.

Время, требуемое для обнаружения отказа, может быть определено с использованием следующей формулы:

(скорость пульса) X (цикл отказа) X 2.

Скорость обнаружения отказов для сетевого модуля можно изменить двумя способами:

- Выбрать одно из предопределенных значений скорости (медленная, нормальная или быстрая). Для сетей типа Ether применимо следующее:
  - быстрая – 10 с (5 x 1 x 2);
  - нормальная – 20 с (10 x 1 x 2);
  - медленная – 48 с (12 x 2 x 2).
- Изменить составляющие параметров cycle или hbrate. Можно использовать меню SMIT «Change a Cluster Network Module using Custom Values» («Изменение сетевого модуля кластера с использованием настраиваемых значений»).

Предопределенные значения для каждого типа сети определяются таким образом, чтобы получались обоснованные результаты. Вам может потребоваться рассмотреть вариант изменения скорости обнаружения отказов с целью:

- сократить время перемещения при сбое;
- не допустить перегрузки процессора узла и последующих ложных перехватов.

Сведения о чувствительности сети (также называемой *скоростью обнаружения отказов*) можно получить от служб топологии, как показано в примере 3.1.

Пример 3.1. Чувствительность сети типа Ether

```
p630n01-# lssrc -ls topsvcs
................. Omitted lines......................
NIM's PID: 19978
net_ether_01_1 [1]    3    1 S 10.10.31.31     10.10.31.31
net_ether_01_1 [1] en0          0x42d56868      0x42d56872
HB Interval = 1.000 secs. Sensitivity = 10 missed beats
................. Omitted lines......................
```

usr/sbin/cluster/netmon.cf

В конфигурациях кластера, содержащих сети, которые при определенных условиях могут стать сетями с одним адаптером, может быть сложно точно определить отказ того или иного адаптера. В таких ситуациях RSCT использует файл netmon.cf.

Службы топологии RSCT сканируют конфигурацию netmon.cf во время запуска кластера. Когда сетевому монитору требуется проверить сеть, чтобы убедиться в функционировании адаптера, он направляет запросы ICMP ECHO на каждый IP-адрес. После отправления запроса на все адреса сетевой монитор проверяет счетчик входящих пакетов, прежде чем определить, произошел ли отказ адаптера.

Этот файл может содержать до 30 адресов или меток; для него применимы следующие указания:

- файл netmon.cf содержит по одному IP-адресу или IP-метке на кабель;
- каждый IP-адрес и соответствующую метку из файла netmon.cf необходимо включить в файл /etc/hosts.

3.8.9 Заполнение таблиц планирования сети

Следующие таблицы содержат требуемую информацию о сети. Табл. 3.6 содержит спецификации сети Ethernet из нашего примера.

Таблица 3.6. Ethernet-сети кластера

ТАБЛИЦА КЛАСТЕРА HACMP – ЧАСТЬ 4 из 11					**ДАТА: июль 2005**
ETHERNET-СЕТИ КЛАСТЕРА					
ИМЯ СЕТИ	ТИП СЕТИ	МАСКА СЕТИ	ИМЕНА УЗЛОВ	IPAT ЧЕРЕЗ СИНОНИМЫ	Смещение IP-адреса для мониторинга пульса через IP-синонимы
ether10	ethernet (общего доступа)	255.255.255.0	node01, node02	Включен	1.1.1.1
КОММЕНТАРИИ	ПРИМЕЧАНИЯ. Смещение IP-адреса добавляет IP-синонимы для каждого интерфейса Ethernet при запуске HACMP. Эти синонимы затем используются для мониторинга пульса, тогда как для адресов базового адаптера мониторинг не осуществляется. Выберите заданную по умолчанию скорость обнаружения отказов (ether = normal = 20 с)				

Табл. 3.7 описывает сети типа «точка-точка», используемые в кластере. В нашем примере применяется только сеть пульса через диски, однако мы для образца дополнительно включили описание сети RS232.

Таблица 3.7. Сети типа «точка-точка»

ТАБЛИЦА КЛАСТЕРА HACMP – ЧАСТЬ 5 из 11					**ДАТА: июль 2005**
ПОСЛЕДОВАТЕЛЬНЫЕ СЕТИ КЛАСТЕРА И СЕТИ ТИПА «ТОЧКА-ТОЧКА»					
ИМЯ СЕТИ	ТИП СЕТИ	ИМЕНА УЗЛОВ	Устройство	ИМЯ ИНТЕРФЕЙСА	МЕТКА АДАПТЕРА
serial10	serial (закрытого доступа)	node01, node02	Неприменимо	/dev/tty0 /dev/tty0	node01_tty1 node02_tty1
diskhb10	diskhb	node01, node02	vpath0 vpath0	Неприменимо	node1_to_node2hb node2_to_node1hb
КОММЕНТАРИИ	ПРИМЕЧАНИЯ. Связи RS232, Target mode SCSI, Target mode SSA и мониторинга пульса через диски не используют протокол TCP/IP и не требуют использования маски сети или IP-адреса. Для сети serial10 конфигурация не выполняется, она выполняется только для сети diskhb10.				

После описания сетей необходимо выполнить описание интерфейсов и IP-адресов, используемых HACMP, как показано в табл. 3.8.

Таблица 3.8. Коммуникационные интерфейсы и IP-адреса кластера

ТАБЛИЦА КЛАСТЕРА HACMP – ЧАСТЬ 6 из 11					**ДАТА: июль 2005**
ИНТЕРФЕЙСЫ И IP-АДРЕСА					
node01					
IP-метка	Размещение IP-синонима	СЕТЕВОЙ ИНТЕРФЕЙС	ИМЯ СЕТИ	ФУНКЦИЯ ИНТЕРФЕЙСА	IP-АДРЕС/МАСКА
node01a	Неприменимо	en0	ether10	Базовый (несервисный)	10.10.31.31 255.255.255.0
node01b	Неприменимо	en1	ether10	Базовый (несервисный)	10.10.32.31 255.255.255.0
ha53node1	Без совместного размещения (по умолчанию)	Неприменимо	ether10	Постоянный	192.168.100.31 255.255.255.0
app1svc	Без совместного размещения (по умолчанию)	Неприменимо	ether10	Сервисный	192.168.100.131 255.255.255.0
node02					
IP-метка	Размещение IP-синонима	СЕТЕВОЙ ИНТЕРФЕЙС	ИМЯ СЕТИ	ФУНКЦИЯ ИНТЕРФЕЙСА	IP-АДРЕС/МАСКА
node02a	Неприменимо	en0	ether10	Базовый (несервисный)	10.10.31.32 255.255.255.0
node02b	Неприменимо	en1	ether10	Базовый (несервисный)	10.10.32.32 255.255.255.0
ha53node2	Без совместного размещения (по умолчанию)	Неприменимо	ether10	Постоянный	192.168.100.32 255.255.255.0
app2svc	Без совместного размещения (по умолчанию)	Неприменимо	ether10	Сервисный	192.168.100.132 255.255.255.0
КОММЕНТАРИИ	Каждый узел содержит два базовых адаптера, каждый из которых находится в отдельной подсети. Каждый узел также содержит постоянный (привязанный к узлу) адрес и сервисный адрес. IPAT посредством синонимов используется так же, как и мониторинг пульса через синонимы (начальный диапазон = 1.1.1.1)				

3.9 Планирование требований к хранилищу

При планировании хранилища кластера необходимо учитывать следующие аспекты:

- Физические диски:
  - Убедитесь в том, что для всех дисковых систем обеспечивается высокая доступность. Это может быть реализовано посредством зеркального отображения, технологии RAID и резервного оборудования.
  - Внутренние диски. Обычно являются местом расположения rootvg.
  - Внешние диски. Должны быть местом расположения данных приложения.
- Компоненты LVM:
  - все общие хранилища имеют уникальные имена логических томов и файловых систем;
  - старшие номера устройств должны быть уникальными;
  - требуется ли обеспечить зеркальное отображение данных?

3.9.1 Внутренние диски

Внутренний диск узла обычно содержит rootvg, а также может содержать исполняемые файлы приложений. Мы рекомендуем организовать зеркальное отображение внутреннего диска для обеспечения высокой доступности; другими словами, это позволяет устранить необходимость перемещения при сбое для всего узла по причине простого отказа внутреннего диска.

3.9.2 Общие диски

Данные приложений располагаются на внешнем диске, что позволяет обеспечить доступ к ним со всех нужных узлов. Такие диски называются *общими дисками (shared disks)*.

> **Важно!** Все общие диски должны быть разделены на зоны, соответствующие узлам кластера, требующим доступа к определенным томам. Другими словами, общий диск должен иметь возможность активизации и доступа с любого узла, которому требуется выполнять определенное приложение.

Мы советуем убедиться в том, что общая группа томов может быть активизирована вручную на каждом узле, прежде чем требовать управления активизацией от HACMP.

В кластере HACMP общие диски подключаются к нескольким узлам кластера. В конфигурации без одновременного доступа в любой момент времени только один узел является владельцем дисков. При отказе узла-владельца, чтобы восстановить обслуживание клиентов, другой узел кластера из списка узлов группы ресурсов перехватывает владение общими дисками и перезапускает приложения.

Как правило, в зависимости от количества дисков в группе ресурсов и от метода перехвата дисков перехват может занимать от 30 до 300 с.

HACMP поддерживает использование следующих дисковых технологий производства IBM в качестве общих внешних дисков в кластере высокой доступности:

- SCSI-приводы, включая подсистемы RAID;
- SSA-адаптеры и дисковые подсистемы производства IBM;
- адаптеры и дисковые подсистемы Fibre Channe;
- устройства Data path (VPATH) – SDD 1.3.1.3 или выше.

Дисковые подсистемы сторонних производителей также могут поддерживаться, однако нужно предварительно получить сведения об их поддержке у изготовителя оборудования.

При работе с общей группой томов:

- Не включайте внутренний диск в общую группу томов, так как он не будет доступен для других узлов.
- Не активизируйте (vary on) общие группы томов в кластере HACMP при загрузке системы. Для этого следует использовать скрипты обработки событий кластера. Убедитесь в том, что атрибут автоматической активизации в AIX 5L ODM для общих групп томов, перечисленных в группе ресурсов, установлен в значение No. Для исправления этого атрибута можно использовать утилиту верификации кластера.

Важно! При определении группы томов в HACMP не следует осуществлять для нее управление вручную с любого узла вне HACMP во время работы HACMP. Это может привести к непредсказуемым результатам. Если вам требуется выполнить некоторые действия над группой томов отдельно от HACMP, следует остановить службы кластера, выполнить задачу управления группой томов, оставить группу томов в деактивизированном состоянии (varied off), после чего перезапустить HACMP.

3.9.3 Пример конфигурации дисков

На рис. 3.18 показаны различные дисковые конфигурации, которые могут использоваться в кластере. Обратите внимание на то, что для внутренних дисков осуществляется зеркальное отображение (rootvg) и что на каждом узле существует несколько адаптеров Fiber Channel (HBA), каждый из которых подключен к отдельному SAN-коммутатору (для обеспечения избыточности).

Для проверки дисков следует использовать идентификаторы PVID, так как вполне возможно, что номер vpath на разных узлах может различаться в связи с нумерацией устройств в AIX.

Общие диски представлены разделенными на зоны для обоих узлов. Разделение на зоны выполняется через программное обеспечение управления хранилищем (SAN Storage Manager).

Важно! Нужно обязательно следовать правилам конфигурирования хранилища (разделение на зоны, маскировка LUN), соответствующим вашей среде. Это особенно важно в случаях, когда узлы HACMP совместно используют подсистемы хранения и SAN с другими серверами.

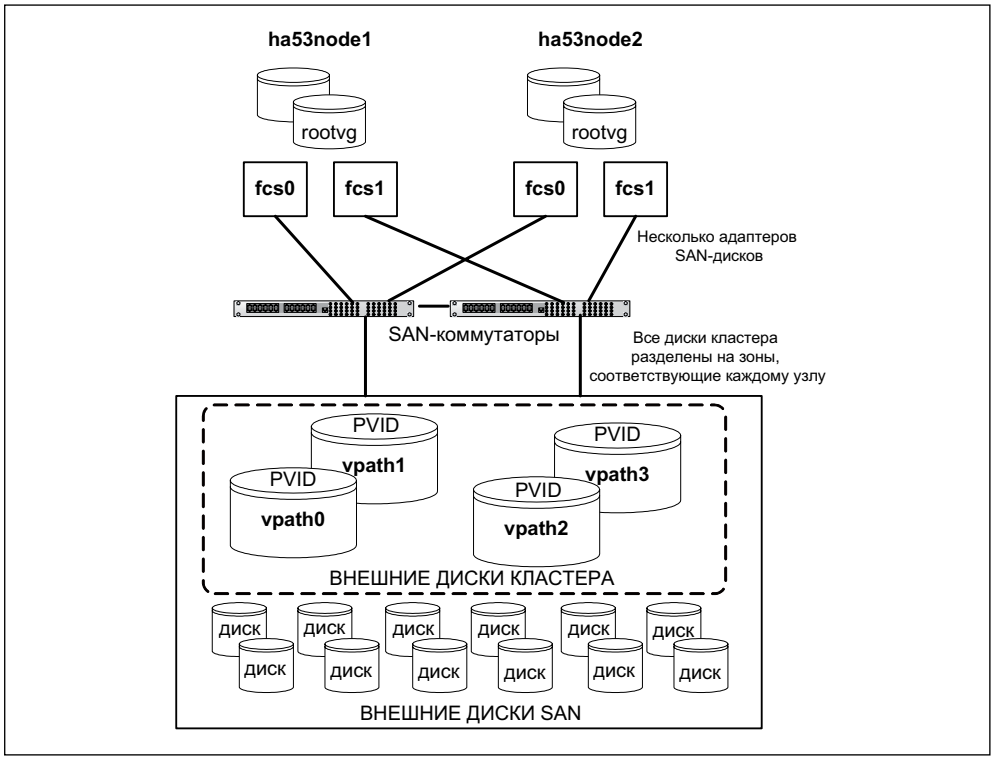

Рис. 3.18 Конфигурация физических дисков

3.9.4 Группы томов в режиме расширенного одновременного доступа (ECM)

Любой диск, для которого HACMP обеспечивает поддержку подключения к нескольким узлам, может использоваться для создания группы томов в режиме расширенного одновременного доступа (enhanced concurrent mode) как в средах с одновременным доступом, так и в средах с без одновременного доступа.

- **Среды с одновременным доступом.** Приложение одновременно выполняется на всех активных узлах кластера. Чтобы такие приложения могли осуществлять доступ к своим данным, выполняется активизация групп томов с одновременным доступом на всех активных узлах кластера. После этого приложение отвечает за обеспечение согласованности доступа к данным.
- **Среды без одновременного доступа.** Приложение в любой заданный момент времени выполняется только на одном узле. К группам томов не осуществляется одновременный доступ; в любой момент времени доступ осуществляется с одного узла.

При активизации группы томов в режиме расширенного одновременного доступа LVM разрешает доступ к группе томов для всех узлов. Однако высокоуровневые под-

ключения например, подключения (mounts) JFS или NFS, запрещены для всех узлов, кроме узла, который на данный момент является владельцем группы томов в HACMP.

В AIX 5.x группы томов в режиме ECM служат заменой классической опции одновременного доступа в HACMP. В новых версиях AIX (5.2 и выше) можно создавать только группы томов в режиме расширенного одновременного доступа. Хотя все еще можно использовать «старые» 32-разрядные группы томов с одновременным доступом, следует проанализировать ограничения, вызываемые обслуживанием таких групп томов в своем кластере. Дополнительные сведения см. в руководстве **High Availability Cluster Multi-Processing for AIX 5L Planning and Installation Guide**, SC23-4861-06.

> **Примечание.** Несмотря на то что можно задавать группы томов с расширенным одновременным доступом, это НЕ обязательно означает, что вам нужно использовать их для одновременного доступа, т.е., вы можете определить и использовать эти группы томов для обычного доступа к общей файловой системе. Тем не менее нельзя определять файловые системы в группах томов, предназначенных для одновременного доступа.

3.9.5 Общие логические тома

Планирование общих логических томов связано с вопросом доступности данных. Обеспечение высокой доступности данных с использованием зеркального отображения или технологии RAID является ключевым требованием. Помните, что защита от отказов дисков в HACMP обеспечивается использованием механизмов LVM и хранилища (RAID), поэтому очень важно обеспечить высокую доступность дисковой инфраструктуры.

При планировании общих компонентов LVM необходимо учитывать следующие принципы:

- Копии логических томов и RAID-массивы обеспечивают защиту от потери данных в результате отказа физического диска.
- Все файлы операционной системы должны располагаться в корневой группе томов (root volume group, rootvg), а все пользовательские данные должны находиться вне этой группы.
- Группы томов, содержащие по меньшей мере три физических тома, обеспечивают максимальную доступность при реализации зеркального отображения.
- Если вы планируете задать для групп томов атрибут «Use Forced Varyon of Volume Groups if Necessary» («При необходимости использовать принудительную активизацию групп томов»), то требуется применять максимально строгую политику выделения дисков для зеркальных физических томов – super strict.
- При зеркальном отображении LVM каждый физический том, содержащий копию, должен использовать отдельный источник питания. При отказе одного источника питания другие источники питания обеспечат отсутствие единой точки отказа.
- При планировании группы томов следует учитывать аспекты кворума. При включенном кворуме в группе томов из двух дисков возникает опасность потери кворума и доступа к данным. Следует либо построить трехдисковые группы томов

(например, с использованием дисков/LUN для обеспечения кворума), либо отключить кворум.

- Необходимо помнить о разработанной конфигурации кластера. Узел, для которого не сконфигурирован перехват ресурсов, не должен владеть критическими группами томов.
- Убедитесь в том, что запланировано регулярное выполнение резервного копирования.

После создания дисковой инфраструктуры высокой доступности при планировании общих групп томов следует учитывать следующие элементы:

- Все общие группы томов должны иметь уникальные имена логических томов и файловых систем. Это относится и к файлам журналов jfs/jfs2.
- **Не следует** использовать встроенные журналы с общими файловыми системами JFS2.
- Старшие номера устройств для каждой группы томов должны быть уникальными (особенно если вы планируете использовать NFS).

На рис. 3.19 представлены основные компоненты внешнего хранилища. Обратите внимание на то, что имена всех логических томов и файловых систем являются уникальными, как и старший номер устройства для каждой группы томов. Обеспечение высокой доступности данных достигается путем использования диска SAN и дублирования путей к устройствам.

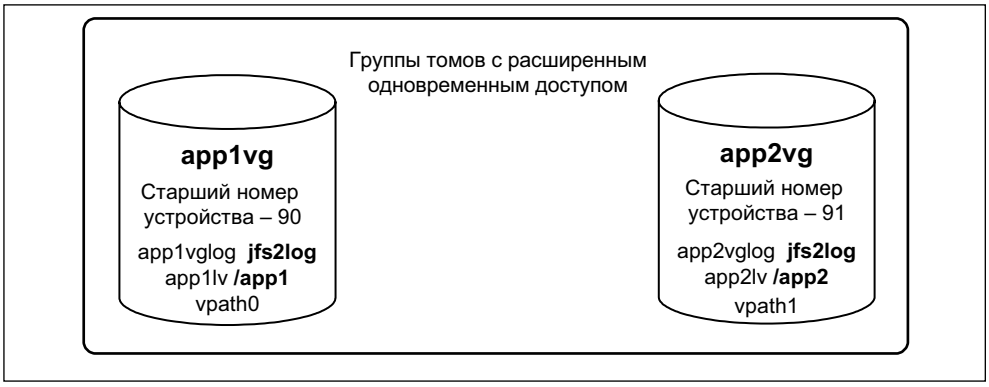

Рис. 3.19. Внешний диск SAN

3.9.6 Быстрый перехват диска

HACMP автоматически обнаруживает отказы узлов и инициирует перехват дисков в процессе перехвата группы ресурсов. Традиционный процесс перехвата диска предполагает сброс бита резервирования диска SCSI (или SSA) перед активизацией (varying on) группы томов на резервном узле. Этот процесс может быть длительным, особенно при большом количестве дисков в группе томов.

Начиная с HACMP 5.1 в связи с поддержкой групп томов с расширенным одновременным доступом в AIX стало возможным сократить время перехвата диска (а

значит, и группы ресурсов) с использованием опции быстрого перехвата диска. Если группа томов, используемая для общего доступа к файловой системе, была определена в режиме расширенного одновременного доступа, HACMP автоматически это обнаруживает и активизирует группу томов на всех узлах, входящих в эту группу ресурсов. Это устраняет необходимость сброса резервирования диска в случае перехвата. Для работы функции быстрого перехвата диска необходимо выполнение следующих требований:

- AIX 5L v.5.2 или выше;
- HACMP 5.1 или выше с набором файлов (fileset) bos.clvm.enh, установленным на всех узлах в кластере;
- использование групп томов в режиме расширенного одновременного доступа в группах ресурсов без одновременного доступа.

Для существующих групп томов, входящих в группы ресурсов без одновременного доступа, можно выполнить преобразование этих групп томов в группы томов с расширенным одновременным доступом после обновления программного обеспечения HACMP.

Фактическое время быстрого перехвата диска в любой конфигурации зависит от множества факторов, выходящих за рамки контроля HACMP, таких, как вычислительная мощность узлов и активность узлов во время перемещения при сбое.

3.9.7 Заполнение таблиц планирования хранения

Следующие таблицы содержат необходимую информацию об общих группах томов. Совместно они дают достаточно полное представление об общей дисковой конфигурации.

Описание общих групп томов и физических дисков приведено в табл. 3.9:

Таблица 3.9. Общие диски

ТАБЛИЦА КЛАСТЕРА HACMP – ЧАСТЬ 7 из 11					**ДАТА: июль 2005**
ОБЩИЕ ДИСКИ					
node01			node02		
VGNAME	VPATH	HDISK	HDISK	VPATH	VGNAME
app1vg	vpath0	hdisk0, hdisk1, hdisk2, hdisk3	hdisk0, hdisk1, hdisk2, hdisk3	vpath0	
	vpath1	hdisk4, hdisk5, hdisk6, hdisk7	hdisk4, hdisk5, hdisk6, hdisk7	vpath1	app2vg
КОММЕНТАРИИ. Все диски видимы с обоих узлов; app1vg обычно располагается на узле node01, app2vg обычно располагается на узле node02.					

Запишите сведения об общих группах томов, как показано в табл. 3.10.

Таблица 3.10. Общие группы томов

ТАБЛИЦА КЛАСТЕРА HACMP – ЧАСТЬ 8 из 11		**ДАТА: июль 2005**
ОБЩИЕ ГРУППЫ ТОМОВ (БЕЗ ОДНОВРЕМЕННОГО ДОСТУПА)		
ГРУППА РЕСУРСОВ	ГРУППА ТОМОВ 1	ГРУППА ТОМОВ 1
C10RG1	app1vg Основной номер = 90 журнал = app1vglog Логический том 1 = app1lv1 Файловая система 1 = /app1 (20 Гб)	Неприменимо
C10RG2	app2vg Основной номер = 91 журнал = app2vglog Логический том 1 = app2lv1 Файловая система 1 = /app2 (20 Гб)	Неприменимо
КОММЕНТАРИИ	Создается общая группа томов на первом узле, после чего выполняется импорт на второй узел; #importvg -y app1vg -V 90 vpath0 (может быть необходимо сделать pv доступным с использованием chdev -l vpath0 -a pv=yes); #chvg -an app1vg (отключает автоматическую активизацию для группы томов); #mount /app1 (для того, чтобы убедиться в возможности подключения файловой системы); #umount /app1; #varyoffvg app1vg (оставляет группу томов в отключенном режиме, оставляя HACMP управление группой)	

3.10 Планирование приложений

Практически любое приложение, работающее на автономном сервере AIX, может быть интегрировано в кластер HACMP, так как приложения не знают о функционировании HACMP. Другими словами, HACMP запускает и останавливает приложения.

При планировании обеспечения высокой доступности приложения нужно понимать, какие ресурсы необходимы приложению, и знать расположение этих ресурсов в кластере. Это даст возможность реализовать решение, которое позволит обеспечить корректную обработку ресурсов в HACMP при отказе узлов.

Вы должны полностью понимать, как приложение выполняется в среде с одним узлом и в среде с несколькими узлами. Мы рекомендуем в процессе подготовки приложения к работе в HACMP протестировать выполнение приложения вручную на обоих узлах, прежде чем передавать управление приложением в HACMP. Не следует ограничиваться предположениями о работе приложения в условиях перемещения при сбое.

> **Примечание.** Основное условие обеспечения высокой доступности приложения состоит в том, чтобы сначала проверить корректность выполнения приложения в автономном режиме на каждом узле, на котором оно может находиться.

Мы рекомендуем убедиться в корректности выполнения приложения на всех требуемых узлах прежде, чем конфигурировать управление приложением из HACMP.

Необходимо проанализировать и учитывать следующие аспекты:

- код приложения – двоичные файлы, скрипты, ссылки, файлы конфигурации и т. д.;
- переменные окружения – все переменные окружения, которые требуется передать в приложение для его корректного выполнения;
- данные приложения;
- настройка сети – IP-адреса, имя хоста;
- лицензирование приложения;
- пользователи системы, определенные приложением.

При планировании защиты приложения в кластере HACMP необходимо учитывать следующее:

- Приложение должно быть совместимо с используемой версией AIX.
- Приложение должно быть совместимо с системой общего хранилища, так как она будет содержать данные приложения.
- Убедитесь в том, что приложения успешно работают в среде с одним узлом. Выполнять отладку приложения в кластере гораздо сложнее, чем на одном сервере.
- Следует распределить приложение и его данные таким образом, чтобы на общих внешних дисках находились только данные. Такое распределение не только предотвращает нарушение лицензий программного обеспечения, но и упрощает восстановление после сбоя.
- Если вы планируете включить в группы ресурсов кластера с зависимостями «родительский объект/дочерний объект» многоуровневые приложения, например базы данных или сервер приложений, HACMP предоставляет простые в использовании меню SMIT, чтобы задать такое отношение.
- Необходимо написать надежные скрипты как для запуска, так и для остановки приложения на узлах кластера. Скрипт запуска должен быть способен восстановить приложение после аварийного завершения. Убедитесь, что скрипты корректно выполняются в среде с одним узлом, прежде чем использовать их в HACMP.
- Проверьте требования к лицензированию. Некоторые производители требуют использования отдельной лицензии для каждого процессора, на котором выполняется приложение; это означает, что вы должны лицензировать приложение, включив информацию о процессоре в приложение при его установке. В результате, несмотря на правильную обработку отказов узла программным обеспечением HACMP, оно может быть неспособно перезапустить приложение на резервном узле из-за недостаточного количества лицензий приложения, доступных в кластере. Во избежание этой проблемы убедитесь в наличии лицензии для каждого компонента системы в кластере, на котором приложение потенциально может выполняться.
- Если требуется обеспечить одновременный доступ, проверьте, использует ли приложение собственный механизм блокировки.

3.10.1 Серверы приложений

В HACMP под сервером приложения понимается просто набор скриптов, используемых для запуска и остановки приложения.

Сконфигурируйте свой сервер приложения, задав имя для использования в HACMP и выполнив привязку к скриптам запуска и остановки.

После создания сервера приложения следует связать его с группой ресурсов. Затем HACMP применяет эту информацию для управления приложением.

3.10.2 Мониторинг приложения

HACMP может осуществлять мониторинг приложения с использованием одного из двух методов:

- мониторинга процесса – обнаруживает завершение процесса с использованием средства RSCT Resource Monitoring and Control (RMC).
- настраиваемого мониторинга – осуществляет мониторинг состояния приложения с использованием заданного метода мониторинга, например скрипта.

Начиная с HACMP 5.2 можно применять несколько мониторов для одного приложения.

При определении собственного настраиваемого метода мониторинга необходимо помнить следующее:

- Можно сконфигурировать несколько мониторов приложения с уникальными именами и связать их с одним или несколькими серверами приложения.
- Метод мониторинга должен представлять собой исполняемую программу, например скрипт оболочки (shell script), тестирующий приложение и на выходе возвращающий целочисленное значение, указывающее состояние приложения. Код завершения должен быть нулевым при нормальном состоянии приложения и ненулевым при отказе приложения.
- HACMP не осуществляет передачу аргументов в метод мониторинга.
- По умолчанию метод мониторинга записывает сообщения в файл /tmp/clappmond.application_monitor_name.monitor.log. Кроме того, по умолчанию при каждом запуске приложения происходит перезапись файла журнала мониторинга.
- Метод не должен быть слишком сложным. Если метод мониторинга не возвращает значение в течение заданного интервала опроса, он удаляется.

> **Важно!** Так как процесс мониторинга зависит от времени выполнения, ВСЕГДА тестируйте свой метод мониторинга при различных нагрузках, чтобы подобрать правильное значение интервала опроса.

- Убедитесь, что контроллер системных ресурсов (System Resource Controller, SRC) настроен на перезапуск приложения.

3.10.3 Инструмент анализа доступности

Для измерения точного количества времени, в течение которого любое из приложений, определенных в HACMP, является доступным, можно использовать инструмент анализа доступности (application availability analysis tool). Программное обеспечение HACMP осуществляет сбор, простановку отметок времени и регистрацию информации по следующим событиям:

- определение, изменение или удаление монитора приложения;
- запуск, остановка или отказ приложения;
- отказ или прекращение работы узла либо возобновление работы узла;
- отключение или перемещение группы ресурсов;
- приостановка или возобновление мониторинга приложения с использованием нескольких мониторов.

3.10.4 Приложения, интегрируемые с HACMP

Некоторые приложения, включая Fast Connect Services и Workload Manager, могут быть сконфигурированы непосредственно как ресурсы высокой доступности без серверов приложений или дополнительных скриптов. Кроме того, верификация HACMP обеспечивает корректность и согласованность некоторых аспектов конфигурации Fast Connect Services или Workload Manager.

3.10.5 Заполнение таблиц планирования приложений

Следующие таблицы описывают требуемую информацию для каждого приложения.

Заполните таблицу приложений, включив в нее всю требуемую информацию, как показано в табл. 3.11.

Таблица 3.11. Таблица приложений

ТАБЛИЦА КЛАСТЕРА HACMP – ЧАСТЬ 9 из 11				**ДАТА: июль 2005**
ТАБЛИЦА ПРИЛОЖЕНИЙ				
APP1				
ЭЛЕМЕНТ	КАТАЛОГ	ФАЙЛОВАЯ СИСТЕМА	РАСПОЛОЖЕНИЕ	ДОСТУП
ИСПОЛНЯЕМЫЕ ФАЙЛЫ	/app1/bin	/app1	Хранилище SAN	Общий
ФАЙЛЫ КОНФИГУРАЦИИ	/app1/conf	/app1	Хранилище SAN	Общий
ФАЙЛЫ ДАННЫХ	/app1/data	/app1	Хранилище SAN	Общий
ФАЙЛЫ ЖУРНАЛОВ	/app1/logs	/app1	Хранилище SAN	Общий
СКРИПТ ЗАПУСКА	/cluster/local/app1/start.sh	/	rootvg	Необщий (должен располагаться на обоих узлах)

Окончание табл. 3.11.

ТАБЛИЦА КЛАСТЕРА НАСМР – ЧАСТЬ 9 из 11				ДАТА: июль 2005
ТАБЛИЦА ПРИЛОЖЕНИЙ				
СКРИПТ ОСТАНОВКИ	/cluster/local/app1/stop.sh	/	rootvg	Необщий (должен располагаться на обоих узлах)
СТРАТЕГИЯ ПЕРЕМЕЩЕНИЯ ПРИ СБОЕ	Перемещение при сбое на узел node02			
КОМАНДЫ И ПРОЦЕДУРЫ ОБЫЧНОГО ЗАПУСКА	Убедиться в том, что сервер APP1 работает			
КОМАНДЫ И ПРОЦЕДУРЫ ВЕРИФИКАЦИИ	Выполнить следующую команду и убедиться в том, что APP1 активно. В противном случае отправить уведомление			
КОМАНДЫ И ПРОЦЕДУРЫ ОБЫЧНОЙ ОСТАНОВКИ	Убедиться в том, что сервер APP1 останавливается корректно			
РЕИНТЕГРАЦИЯ УЗЛА	Должна осуществляться во время запланированного окна обслуживания, чтобы свести к минимуму нарушение работы клиентов			
APP2				
ЭЛЕМЕНТ	КАТАЛОГ	ФАЙЛОВАЯ СИСТЕМА	РАСПОЛОЖЕНИЕ	ДОСТУП
ИСПОЛНЯЕМЫЕ ФАЙЛЫ	/app2/bin	/app2	Хранилище SAN	Общий
ФАЙЛЫ КОНФИГУРАЦИИ	/app2/conf	/app2	Хранилище SAN	Общий
ФАЙЛЫ ДАННЫХ	/app2/data	/app2	Хранилище SAN	Общий
ФАЙЛЫ ЖУРНАЛОВ	/app2/logs	/app2	Хранилище SAN	Общий
СКРИПТ ЗАПУСКА	/cluster/local/app2/start.sh	/	rootvg	Необщий (должен располагаться на обоих узлах)
СКРИПТ ОСТАНОВКИ	/cluster/local/app2/stop.sh	/	rootvg	Необщий (должен располагаться на обоих узлах)
СТРАТЕГИЯ ПЕРЕМЕЩЕНИЯ ПРИ СБОЕ	Перемещение при сбое на узел node01			
КОМАНДЫ И ПРОЦЕДУРЫ ОБЫЧНОГО ЗАПУСКА	Убедиться в том, что сервер APP2 работает			
КОМАНДЫ И ПРОЦЕДУРЫ ВЕРИФИКАЦИИ	Выполнить следующую команду и убедиться в том, что APP2 активно. В противном случае отправить уведомление			
КОМАНДЫ И ПРОЦЕДУРЫ ОБЫЧНОЙ ОСТАНОВКИ	Убедиться в том, что сервер APP2 останавливается корректно			
РЕИНТЕГРАЦИЯ УЗЛА	Должна осуществляться во время запланированного окна обслуживания, чтобы свести к минимуму нарушение работы клиентов			
КОММЕНТАРИИ	Обзор приложений			

Следует также заполнить таблицу мониторинга приложений, включив в нее всю информацию, необходимую для инструментов мониторинга приложений (табл. 3.12).

Таблица 3.12. Таблица мониторинга приложений

ТАБЛИЦА КЛАСТЕРА HACMP – ЧАСТЬ 10 из 11	**ДАТА: июль 2005**
МОНИТОРИНГ ПРИЛОЖЕНИЙ	
APP1	
Можно ли осуществлять мониторинг приложения с использованием монитора процессов?	Да
Наблюдаемые процессы	app1
Владелец процесса	root
Счетчик экземпляров	1
Стабилизационный интервал	30
Счетчик перезапуска	3
Интервал перезапуска	95
Действие при отказе приложения	Перемещение при сбое
Метод уведомления	/usr/es/sbin/cluster/events/notify_app1
Метод очистки	/usr/es/sbin/cluster/events/stop_app1
Метод перезапуска	/usr/es/sbin/cluster/events/start_app1
APP2	
Можно ли осуществлять мониторинг приложения с использованием монитора процессов?	Да
Наблюдаемые процессы	app2
Владелец процесса	root
Счетчик экземпляров	1
Стабилизационный интервал	30
Счетчик перезапуска	3
Интервал перезапуска	95
Действие при отказе приложения	Перемещение при сбое
Метод уведомления	/usr/es/sbin/cluster/events/notify_app2
Метод очистки	/usr/es/sbin/cluster/events/stop_app2
Метод перезапуска	/usr/es/sbin/cluster/events/start_app2

3.11 Планирование групп ресурсов

В HACMP управление ресурсами осуществляется с использованием групп ресурсов.

Каждая группа ресурсов обрабатывается как контейнер, который может содержать следующие типы ресурсов: IP-метки, приложения, файловые системы и группы томов. Каждая группа ресурсов имеет настройки, которые определяют, когда и каким образом осуществляется ее перехват или освобождение. Можно выполнить точную настройку работы группы ресурсов без одновременного доступа при запуске узла, перемещении группы ресурсов на другой узел при отказе узла или при возврате группы ресурсов при реинтеграции узла.

К ресурсам и группам ресурсов применимы следующие правила и ограничения:

- Для обеспечения высокой доступности ресурса кластера в HACMP он должен быть частью группы ресурсов. Если требуется, чтобы ресурс содержался отдельно, можно определить группу для одного этого ресурса. В группе ресурсов может быть определен один или несколько ресурсов.
- Ресурс не может входить в несколько групп ресурсов.
- Мы рекомендуем поместить сервер приложения вместе с требуемыми им ресурсами в одну группу ресурсов (если нет причин поступить иначе).
- При включении одного узла в списки узлов для нескольких групп ресурсов следует убедиться, что этот узел способен обслуживать все группы ресурсов одновременно.

На рис. 3.20 упрощенно показана взаимосвязь между приложениями, группами томов и сервисными адресами, а также их совмещение в группах ресурсов. Группировка позволяет объединить приложение вместе с его общим дисковым хранилищем и сервисной IP-меткой в одну группу ресурсов. При этом все, что необходимо приложению, будет доступно, когда HACMP активизирует группу ресурсов.

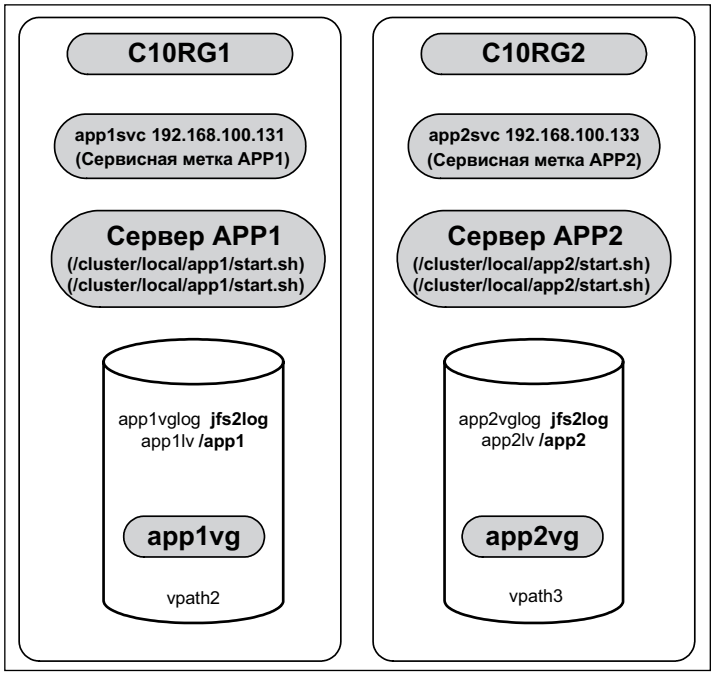

Рис. 3.20. Группы ресурсов

После того как было решено, какие компоненты должны быть включены в группу ресурсов, необходимо выполнить планирование режимов работы группы ресурсов.

В таблице 3.13 приведены основные режимы работы (запуск, перемещение при сбое, возврат после восстановления), которые можно сконфигурировать для групп ресурсов в HACMP 5.3.

Таблица 3.13. Режимы работы группы ресурсов

Запуск	Перемещение при сбое	Возврат после восстановления
Подключение только на домашнем узле (Online on home node only, OHNO) для группы ресурсов	Перемещение на следующий по приоритету узел в списке	Без возврата после восстановления
	Перемещение с использованием динамического приоритета узла	Возврат после восстановления на узел с более высоким приоритетом в списке
Подключение с использованием политики распределения узлов	Перемещение на следующий по приоритету узел в списке	Без возврата после восстановления
	Перемещение с использованием динамического приоритета узла	
Подключение на первом доступном узле (Online on first available node, OFAN)	Перемещение на следующий по приоритету узел в списке	Без возврата после восстановления
	Перемещение с использованием динамического приоритета узла	Возврат после восстановления на узел с более высоким приоритетом в списке
	Перевод в отключенный режим (только на ошибочном узле)	
Подключение на всех доступных узлах	Перевод в отключенный режим (только на отказавшем узле)	Без возврата после восстановления

3.11.1 Атрибуты группы ресурсов

Время установления при запуске

Атрибут времени установления (settling time) относится только к группам ресурсов с подключением на первом доступном узле (Online on First Available Node, OFAN) и устанавливает ожидание в течение заданного количества времени, прежде чем активизировать группу ресурсов. По истечении времени установления HACMP активизирует группу ресурсов на доступном узле с наивысшим приоритетом. Этот атрибут используется для того, чтобы избежать многократного перемещения групп ресурсов с одного узла на другой по мере подключения узлов с более высоким приоритетом.

Если узел запуска является домашним узлом для заданной группы ресурсов, время установления не выдерживается, вместо этого HACMP сразу же пытается подхватить группу ресурсов на этом узле.

> **Внимание!** Этот атрибут действует в масштабе кластера и устанавливается для всех групп томов с подключением на первом доступном узле.

Политика динамического приоритета узла

Настройка политики динамического приоритета узла (dynamic node priority, DNP) позволяет выбирать резервный узел на основе определенных критериев производительности. При этом для выбора резервного узла используется переменная ресурсов RMC, например «наименьшая нагрузка на процессор». При включенной политике динамического приоритета порядок резервных узлов определяется состоянием кластера на момент события, измеряемый выбранной переменной ресурсов RMC.

Если вы решите определить политики динамического приоритета узлов с использованием переменных ресурсов RMC для определения резервного узла для группы ресурсов, следует учитывать следующее:

- политика динамического приоритета узла наиболее полезна в кластере, где все узлы имеют одинаковую вычислительную мощность и объем памяти;
- политика динамического приоритета узла неприменима в кластерах, содержащих меньше трех узлов;
- политика динамического приоритета узла неприменима для групп ресурсов с одновременным доступом.

Необходимо помнить о том, что выбор резервного узла также зависит от таких факторов, как доступность сетевого интерфейса на узле.

Таймер отсроченного возврата после восстановления

Таймер отсроченного возврата (delayed fallback timer) после восстановления позволяет выполнить для группы ресурсов возврат после восстановления на узел с более высоким приоритетом в заданное время. Группа ресурсов, для которой сконфигурирован таймер отсроченного возврата после восстановления и которая в заданный момент времени находится не на домашнем узле, выполняет возврат на узел с более высоким приоритетом.

Зависимости групп ресурсов

HACMP 5.3 предлагает множество конфигураций, где можно задавать отношения между группами ресурсов, для которых следует обеспечивать управление запуском, перемещением при сбое и возвратом после восстановления.

Можно сконфигурировать:

- зависимости «родительский объект/дочерний объект», чтобы все связанные приложения в различных группах ресурсов обрабатывались в правильном порядке;
- зависимости расположения, чтобы определенные приложения в различных группах ресурсов подключались вместе на одном узле или сайте либо подключались на разных узлах.

Несмотря на то что по умолчанию все группы ресурсов обрабатываются параллельно, обработка зависимых групп ресурсов в HACMP осуществляется в соответствии с порядком, определенным зависимостью, и не обязательно параллельно. Зависимости групп ресурсов имеют действие в масштабе кластера и замещают любые настройки последовательного порядка обработки для любых групп ресурсов, входящих в зависимость.

Зависимости между группами ресурсов представляют прогнозируемый и надежный способ построения кластеров с многоуровневыми приложениями.

Метод перехвата IP-адреса и группы ресурсов

Нельзя смешивать метки IPAT посредством синонимов и IPAT посредством замены в одной группе ресурсов. Это ограничение применяется во время верификации ресурсов кластера.

Перехват IP-адреса не применяется к группам ресурсов с одновременным доступом.

Группа ресурсов может включать несколько сервисных IP-меток. При перемещении группы ресурсов с перехватом IP-адреса посредством синонимов все сервисные метки в группе ресурсов перемещаются в виде синонимов на доступный сетевой интерфейс.

Планирование диспетчера рабочей нагрузки

Диспетчер рабочей нагрузки (Workload Manager, WLM) дает возможность пользователям задавать целевые значения и ограничения применения процессора, физической памяти и пропускной способности дисковой подсистемы для различных процессов и приложений. Это обеспечивает более эффективный контроль использования критических системных ресурсов при пиковых нагрузках. HACMP позволяет выполнять конфигурирование классов WLM в группах ресурсов HACMP таким образом, чтобы запуск и остановка WLM и активная конфигурация WLM были под контролем кластера.

HACMP не проверяет каждый аспект конфигурации WLM, так что ответственность за целостность файлов конфигурации WLM возлагается на вас. После добавления классов WLM в группу ресурсов HACMP утилита верификации проверяет только лишь существование требуемых классов WLM. Таким образом, вы должны в полной мере понимать принципы работы WLM и внимательно выполнять конфигурирование классов WLM.

3.11.2 Заполнение таблицы планирования групп ресурсов

Таблица планирования групп ресурсов содержит всю требуемую информацию о планировании для групп ресурсов (табл. 3.14).

Таблица 3.14. Таблицы групп ресурсов

ТАБЛИЦА КЛАСТЕРА HACMP – ЧАСТЬ 11 из 11		**ДАТА: июль 2005**
ГРУППЫ РЕСУРСОВ		
ИМЯ РЕСУРСА	C10RG1	C10RG2
Политика межсайтового управления	Игнорируется	Игнорируется
Имена узлов-участников	node01 node02	node02 node01
Политика запуска	Подключение только на домашнем узле (OHNO)	Подключение только на домашнем узле (OHNO)
Политика перемещения при сбое	Перемещение на следующий по приоритету узел в списке (FONP)	Перемещение на следующий по приоритету узел в списке (FONP)
Политика возврата после восстановления	Возврат на узел с более высоким приоритетом (FBHP)	Возврат на узел с более высоким приоритетом (FBHP)
Таймер отсроченного возврата		
Время установления		
Политики времени выполнения		
Политика динамического приоритета узла		
Порядок обработки (параллельная, последовательная или настраиваемая)		
Сервисная IP-метка	app1svc	app2svc
Серверы приложений	app1	app2
Группы томов	app1vg	app2vg

Окончание табл. 3.14.

ТАБЛИЦА КЛАСТЕРА HACMP – ЧАСТЬ 11 из 11		ДАТА: июль 2005
ГРУППЫ РЕСУРСОВ		
Файловые системы	/app1	/app2
Проверка согласованности файловой системы	fsck	fsck
Метод восстановления файловой системы	Последовательный	Последовательный
Файловые системы или каталоги для экспорта		
Файловые системы или каталоги для подключения NFS		
Сеть для подключения NFS	ether10	ether10
Основной класс WLM		
Автоматический импорт групп томов	Нет	Нет
Подключение файловых систем перед конфигурированием IP	Нет	Нет
КОММЕНТАРИИ	Обзор двух групп ресурсов	

3.12 Подробная схема кластера

Собрав все вместе и используя информацию, собранную при планировании кластера и записанную в таблицах планирования, теперь можно перейти к построению простой для восприятия подробной схемы кластера. На рис. 3.21 представлена подробная схема кластера из нашего примера. Эту схему можно использовать в качестве вспомогательного средства при конфигурировании кластера и диагностике проблем.

3.13 Разработка плана тестирования кластера

Разработка соответствующего плана тестирования для проверки работы кластера при отказах так же важна, как и планирование и конфигурирование кластера HACMP. Другими словами, нужно проверить, соответствует ли обработка отказов кластером ожиданиям. Необходимо протестировать (или подтвердить) возможности восстановления кластера, прежде чем кластер станет частью рабочей среды.

3.13.1 Настраиваемый план тестирования

Как и в прежних версиях HACMP, вам следует разработать локальный набор тестов для проверки целостности кластера. Это обычно включает отключение сетевых кабелей, отключение интерфейсов и завершение работы узлов кластера в целях проверки возможностей восстановления кластера. Это упражнение все еще остается полезным, так как вы имеете возможность имитировать отказы и наблюдать за поведением кластера. Если что-то идет неправильно или не так, как ожидалось, следует остановить тестирование и изучить проблему. После успешного завершения всех тестов кластер можно переносить в рабочую среду.

В табл. 3.15 представлен пример плана тестирования, который можно использовать для тестирования кластера.

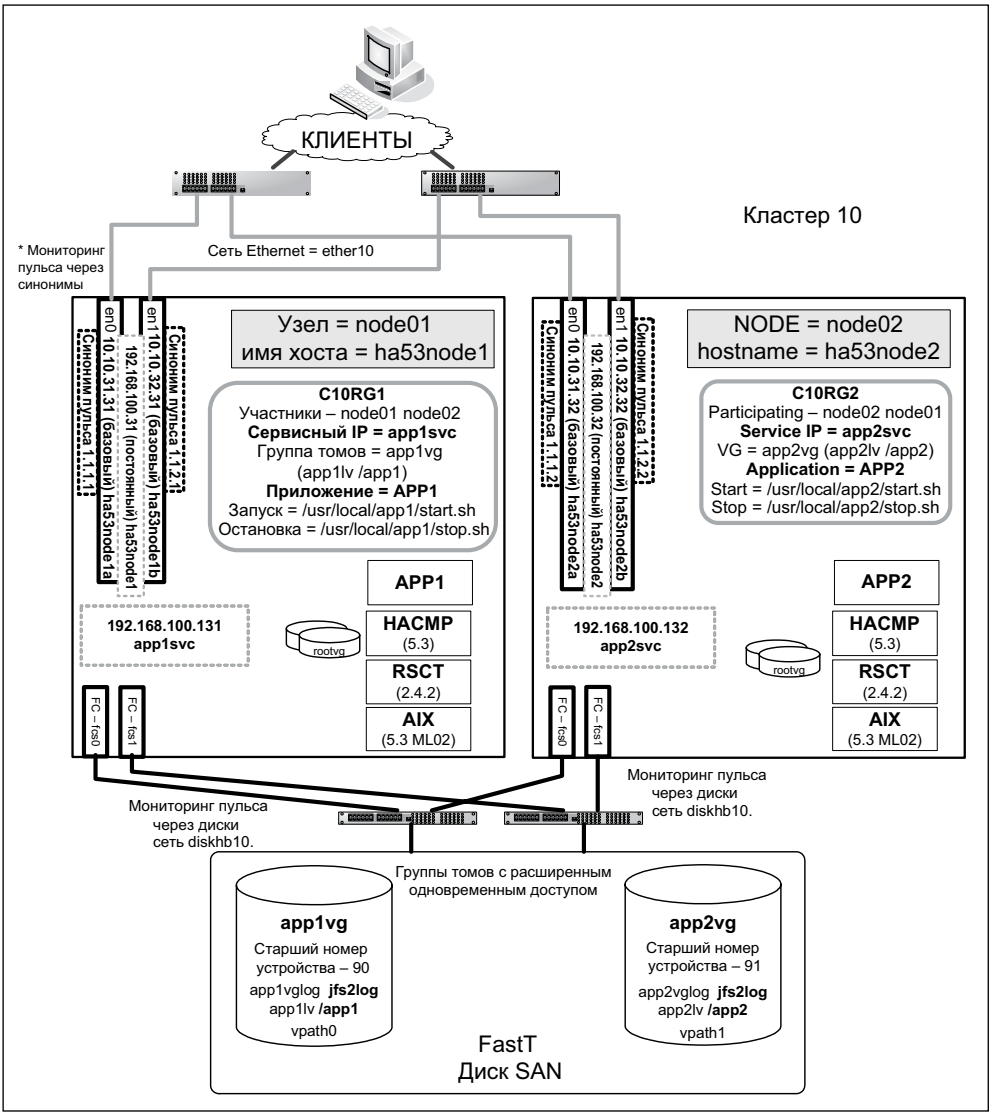

Рис. 3.21. Подробная схема кластера

Таблица 3.15. Пример плана тестирования

План тестирования кластера			
Тест №	Описание теста	Комментарии	Результаты
1	Запуск HACMP на node01	node01 запускается и перехватывает группу ресурсов C10RG1	
2	Запуск HACMP на node02	node02 запускается и перехватывает группу ресурсов C10RG2	
3	Постепенная остановка без перехвата (graceful stop without takeover) на node01	Группа ресурсов C10RG1 отключается	

Окончание табл. 3.15

План тестирования кластера			
Тест №	Описание теста	Комментарии	Результаты
4	Запуск HACMP на node01	node01 запускается и перехватывает группу ресурсов C10RG1	
5	Постепенная остановка с перехватом (graceful stop with takeover) на node01	Группа ресурсов C10RG1 перемещается на node02	
6	Запуск HACMP на node01	node01 запускается и требует группу ресурсов C10RG1	
7	Отказ (отключение) сервисного интерфейса на node01	Перемещение сервисного IP-адреса на второй базовый адаптер	
8	Переподключение сервисного интерфейса на node01	Сервисный IP-адрес остается на втором базовом адаптере	
9	Отказ (отключение) сервисного интерфейса на node01 (теперь на втором адаптере)	Сервисный (и постоянный) IP-адрес перемещается на первый базовый адаптер	
10	Выполнение на node01 команды halt -q для принудительного отключения операционной системы	node01 останавливается – группа ресурсов C10RG1 перемещается на node02	
11	Перезагрузка node01 и перезапуск HACMP	node01 перезагружается. После запуска HACMP, node01 требует C10RG1	
12	Постепенная остановка без перехвата (graceful stop without takeover) на node02	Группа ресурсов C10RG1 отключается	
13	Запуск HACMP на node02	node02 запускается и перехватывает группу ресурсов C10RG2	
14	Постепенная остановка с перехватом (graceful stop with takeover) на node01	Группа ресурсов C10RG2 перемещается на node01	
15	Запуск HACMP на node02	node02 запускается и требует группу ресурсов C10RG2	
16	Отказ (отключение) сервисного интерфейса на node02	Перемещение сервисного IP-адреса на второй базовый адаптер	
17	Переподключение сервисного интерфейса на node02	Сервисный IP-адрес остается на втором базовом адаптере	
18	Отказ (отключение) сервисного интерфейса на node02 (теперь на втором адаптере)	Сервисный (и постоянный) IP-адрес перемещается на первый базовый адаптер	
19	Выполнение на node02 команды halt -q для принудительного отключения операционной системы	node02 останавливается – группа ресурсов C10RG2 перемещается на node01	
20	Перезагрузка node02 и перезапуск HACMP	node02 перезагружается. После запуска HACMP, node02 требует C10RG2	

3.13.2 Инструмент Cluster Test Tool

Чтобы упростить тестирование кластера, HACMP 5.2 и 5.3 включают инструмент Cluster Test Tool, позволяющий протестировать функции кластера, прежде чем он станет частью рабочей среды.

Инструмент Cluster Test Tool работает только в кластере с HACMP 5.2 или более поздней версии с верифицированной и синхронизированной конфигурацией. Инструмент может работать в двух режимах:

- Автоматическое тестирование. Процедура автоматического тестирования (предопределенный набор тестов), поставляемая вместе с инструментом, используется для выполнения базового тестирования кластера. Выполнение установки не требуется. Нужно просто запустить тест из SMIT и просмотреть результаты тестирования в файле журнала Cluster Test Tool.
- Настраиваемое тестирование. Если вы являетесь опытным администратором HACMP и хотите выполнить точную настройку тестирования кластера для своей среды, можно создать настраиваемые тесты, которые можно запускать из SMIT. После установки настраиваемой среды тестирования выполняется запуск процедуры тестирования из SMIT, после чего результаты тестирования просматриваются в файле журнала Cluster Test Tool.

Cluster Test Tool использует демон коммуникаций кластера HACMP для организации связи между узлами кластера с целью обеспечения безопасности кластера HACMP.

Автоматическое тестирование

Инструмент тестирования содержит автоматический метод, предназначенный для быстрого тестирования функционирования кластера. Его выполнение обычно занимает от 30 до 60 мин., в зависимости от сложности кластера; при этом выполняются тесты, перечисленные ниже. Для выполнения этих тестов вы должны иметь доступ под записью root.

Тесты общей топологии кластера

Cluster Test Tool выполняет тесты общей топологии кластера в следующем порядке:

- запуск служб кластера на всех доступных узлах;
- постепенная (graceful) остановка служб кластера на узле;
- перезапуск служб кластера на узле, который был остановлен;
- остановка служб кластера с перехватом на другом узле;
- перезапуск служб кластера на узле, который был остановлен;
- принудительная остановка служб кластера на другом узле;
- перезапуск служб кластера на узле, который был остановлен.

Тесты групп ресурсов с неодновременным доступом

Если кластер включает одну или несколько групп ресурсов с неодновременным доступом, инструмент тестирования выполняет каждый из нижеперечисленных тестов в заданном порядке для каждой группы ресурсов:

- отключение локальной сети на узле, чтобы инициировать перемещение группы ресурсов;
- восстановление ранее отключенной сети;
- отключение сервера приложений и восстановление после отказа приложения.

Тест групп ресурсов с одновременным доступом

Если кластер содержит одну или несколько групп ресурсов с политикой управления запуском, настроенной на подключение на всех доступных узлах (online on all available nodes, OAAN), инструмент тестирования выполняет один тест, состоящий в отключении сервера приложений и восстановлении после отказа приложения.

Тест на фатальный отказ

Инструмент выполняет один тест на фатальный отказ, который останавливает диспетчер кластера на произвольно выбранном узле, на котором в данный момент находится как минимум одна активная группа ресурсов.

> **Примечание.** Если инструмент останавливает диспетчер кластера на контрольном узле, вам может потребоваться перезагрузить этот узел.

Выполнение автоматических тестов

Общие рекомендации советуют периодически подтверждать конфигурацию кластера. Существует два инструмента автоматизации выполнения этой задачи:

- Инструмент автоматического тестирования кластера. Используется непосредственно для тестирования кластера.
- Автоматическая верификация конфигурации кластера. Этот инструмент периодически проверяет и информирует о любых изменениях конфигурации, чтобы администратор кластера мог предпринять корректирующие действия (выполнить синхронизацию и повторное тестирование кластера).

Эти инструменты можно использовать для реализации стандартной процедуры проверки. После выполнения первоначального теста тестирование вручную выполнять необязательно. Однако так как автоматический инструмент тестирования кластера предпринимает действия, которые могут вызвать перерыв в обслуживании, необходимо назначить использование этого инструмента на время, соответствующее окну обслуживания.

Инструмент Cluster Test Tool выполняет заданный набор тестов, произвольным образом выбирая узлы, сети, группы ресурсов и т. д. для тестирования. В процессе тестирования инструмент тестирует различные компоненты кластера.

> **Важно!** Прежде чем начать выполнение автоматического тестирования, убедитесь, что кластер не работает в режиме обслуживания в рабочей среде.

Для запуска автоматической процедуры тестирования:

- Введите **smit hacmp.**
- В SMIT выберите Initialization and Standard Configuration > HACMP Cluster Test Tool и нажмите Enter.
- Появится сообщение «Are you sure». Если вы еще раз нажмете Enter, запустится автоматическое тестирование.

3.14 Разработка плана установки HACMP

После планирования конфигурации и составления схемы кластера, нужно выполнить подготовку к установке.

При внедрении HACMP на существующих серверах следует выделить достаточное окно обслуживания для установки, конфигурирования и тестирования кластера.

Если выполняется новая инсталляция, нужно выделить время на конфигурирование и тестирование базового кластера. После конфигурирования и тестирования кластера можно выполнять интеграцию необходимых приложений во время запланированного окна обслуживания.

Возвратившись к рис. 3.1, можно увидеть, что установке HACMP предшествует этап подготовки. Этот этап необходим для того, чтобы обеспечить готовность инфраструктуры к установке HACMP. Обычно это предполагает использование таблиц планирования и схемы кластера для подготовки узлов для установки HACMP 5.3.

- Убедитесь в том, что установлены необходимые пакеты программного обеспечения узлов и операционной системы.
- Убедитесь в правильности конфигурирования сети.
- Убедитесь в правильности конфигурирования общих дисков.
- Убедитесь в том, что требуемые приложения способны выполняться на всех узлах.

Этап подготовки может занять некоторое время, в зависимости от сложности среды и количества используемых групп ресурсов и узлов. Нужно выделить достаточно времени на подготовку среды, так как нет смысла пытаться устанавливать HACMP в неподготовленной среде. Это обернется напрасной тратой времени на устранение неполадок в плохой инсталляции. Помните о том, что построение хорошо сконфигурированного кластера происходит на основе надежной инфраструктуры.

После завершения планирования кластера и подготовки среды узлы готовы к установке HACMP.

Установка кода достаточно проста. При установке с компакт-диска следует просто использовать SMIT для установки требуемых наборов файлов. При установке из хранилища программного обеспечения можно выполнить NFS-подключение каталога, после чего использовать SMIT для установки из этого каталога.

Убедитесь в том, что у вас есть лицензии на все устанавливаемые функции, такие, как Smart Assist и HACMP/XD.

После того как были установлены необходимые наборы файлов на всех узлах кластера, используйте таблицы планирования для конфигурирования кластера. У вас есть несколько инструментов, которые можно использовать для конфигурирования кластера:

- Можно настроить WebSmit и использовать его для конфигурирования кластера.
- В кластере из двух узлов можно использовать Two-Node Configuration Assistant, реализованный на основе Java.
- Можно использовать ASCII-экран и SMIT для выполнения конфигурирования.

- Существует множество способов облегчить конфигурирование кластера. В следующей главе каждый из них будет рассмотрен подробно, но, если вкратце, вы можете:
  - Использовать Two-Node Configuration Assistant для конфигурирования кластера. Этот инструмент позволяет сконфигурировать базовый двухузловой кластер с одной группой ресурсов.
  - Использовать панели SMIT «HACMP Standard Configuration» для конфигурирования кластера в стандартном формате.
  - Использовать панели SMIT «HACMP Extended Configuration» для конфигурирования кластера вручную.
  - Применить к кластеру файл *.haw, сгенерированный системой автоматизированного планирования (Online Planning Worksheets).

  Применить снимок кластера для конфигурирования кластера.

> **Примечание.** При конфигурировании кластера мы рекомендуем начать с конфигурирования (сетевой) топологии кластера. После конфигурирования топологии кластера следует выполнить конфигурирование, верификацию и синхронизацию кластера, прежде чем переходить к ресурсам (общим группам томов, сервисным IP-адресам и приложениям).
> После успешной верификации и синхронизации топологии следует запустить службы кластера и проверить, все ли работает в соответствии с ожиданиями.
> Это позволяет выявить любые проблемы в сети до конфигурирования ресурсов кластера.

После конфигурирования, верификации и синхронизации кластера следует запустить инструмент автоматического тестирования кластера, чтобы проверить функционирование кластера. Просмотрите результаты работы инструмента тестирования; если они являются успешными, выполните необходимые собственные тесты для дальнейшей верификации кластера.

Проверьте все включенные вами уведомления об ошибках.

После успешного выполнения тестирования создайте резервные копии системы (mksysb) для каждого узла, а также снимок кластера с одного из узлов кластера.

На этом этапе кластер должен быть готов к переносу в рабочую среду.

Теперь к управлению доступностью приложения применимы стандартные процессы управления изменениями и проблемами.

3.15 Резервное копирование конфигурации кластера

Основным средством резервного копирования кластера HACMP является снимок кластера. Хотя файл описания кластера системы автоматизированного планирования (Online Planning Worksheets) тоже описывает конфигурацию кластера, он менее полный, так как он не включает записи ODM.

Основной информацией, сохраненной в снимке кластера, являются данные, находящиеся в классах базы данных конфигурации HACMP (таких, как HACMPcluster, HACMPnode, HACMPnetwork, HACMPdaemons). Эта информация используется для воссоздания конфигурации кластера при применении снимка кластера.

Снимок кластера не сохраняет какие-либо пользовательские скрипты, приложения и прочие параметры конфигурации, не связанные с HACMP. Например, имена серверов приложений и расположение их скриптов запуска и остановки сохраняются в объектном классе HACMPserver базы данных конфигурации. Однако сами скрипты, как и вызываемые ими приложения, не сохраняются.

Утилита создания снимков кластера сохраняет данные в двух разных файлах:

- Файле данных ODM (.odm). Этот файл содержит все данные, сохраненные в объектных классах базы данных конфигурации HACMP для кластера. Этому файлу назначается определяемое пользователем имя с расширением .odm. Так как информация в базе данных конфигурации практически одинакова на всех узлах кластера, снимок кластера сохраняет значения только с одного узла.
- Файле информации о состоянии кластера (.info). Этот файл содержит стандартные выходные данные AIX 5L и HACMP. Этому файлу назначается то же определяемое пользователем имя с расширением .info. По умолчанию этот файл больше не содержит информацию журнала кластера. Обратите внимание на то, что через SMIT можно задать сбор журналов кластера в этом файле при создании снимка кластера.

Для полного резервного копирования следует создать резервную копию (mksysb) каждого узла кластера с использованием стандартных методов. Выберите один узел для создания снимка кластера и сохраните снимок в безопасном месте в целях аварийного восстановления.

При возможности создайте снимок до создания резервной копии (mksysb) узла, чтобы он был включен в резервную копию системы.

Важно! Можно создать снимок с любого узла в кластере, даже при отключенном HACMP. Однако применить снимок к кластеру можно только в том случае, если все узлы доступны и выполняют одну версию HACMP (HACMP может осуществлять связь между узлами с использованием clcomdES).

3.16 Документирование кластера

Для эффективного управления кластером важно выполнять документирование конфигурации кластера. Хорошее документирование кластера позволяет обеспечить более эффективный контроль изменений и быстрое разрешение проблем на всех этапах, от управления изменениями в кластере до устранения неполадок.

Мы рекомендуем вам вести точную схему кластера, которую можно использовать для управления изменениями и проблемами.

Кроме того, HACMP обеспечивает инструменты для упрощения сбора данных конфигурации кластера посредством использования системы автоматизированного планирования (OLPW).

В этом разделе описывается создание файла определения кластера через SMIT с последующим его использованием для создания отчета о конфигурации кластера через OLPW. Итоговый отчет имеет формат HTML и может быть просмотрен через веб-браузер.

Основные этапы создания отчета кластера следующие:

- Экспорт файла определения кластера с одного из узлов кластера с использованием SMIT:
  - файл обычно сохраняется в формате *.haw;
  - при использовании OLPW на рабочей станции следует передать файл определения через ftp на рабочую станцию.
- Использование OLPW для открытия существующего файла определения.
- Использование OLPW для создания отчета о конфигурации. При этом создается файл *.html.
- Использование веб-браузера для просмотра файла. Рекомендуем сохранить файл на другом сервере или рабочей станции в целях аварийного восстановления.

3.16.1 Экспорт файла определения кластера с применением SMIT

Можно создать файл определения кластера из активного кластера HACMP, после чего открыть этот файл с использованием приложения Online Planning Worksheets.

Для создания файла определения кластера из SMIT проделайте следующее:

- Введите **smit hacmp.**
- Выберите Extended Configuration (Расширенное конфигурирование). Выберите Export Definition File for Online Planning Worksheets (Экспорт файла определения для OLPW) и нажмите Enter (пример 3.2).

Пример 3.2. Экспорт файла определения из SMIT

```
                    Extended Configuration

Move cursor to desired item and press Enter.
Discover HACMP-related Information from Configured Nodes
Extended Topology Configuration
Extended Resource Configuration
Extended Cluster Service Settings
Extended Event Configuration
Extended Performance Tuning Parameters Configuration
Security and Users Configuration
Snapshot Configuration
Export Definition File for Online Planning Worksheets

Extended Verification and Synchronization
HACMP Cluster Test Tool
```

- Введите значения следующих полей и нажмите Enter:
  - File Name (Имя файла). Полный путь к файлу определения кластера. По умолчанию используется путь **/var/hacmp/log/cluster.haw**.

- Cluster Notes (Примечания кластера). Любые дополнительные комментарии, относящиеся к вашему кластеру. Введенная здесь информация будет выводиться в панели Cluster Notes (Примечания кластера) в Online Planning Worksheets.

- Откройте файл определения кластера в Online Planning Worksheets.

3.16.2 Создание файла определения кластера из снимка с использованием SMIT

Кроме того, файл определения кластера также можно создать из снимка кластера HACMP, после чего его можно открыть с использованием приложения Online Planning Worksheets.

Для создания файла определения кластера из снимка с применением SMIT проделайте следующее:

- введите **smit hacmp**;
- выберите Extended Configuration (Расширенное конфигурирование):
  - Snapshot Configuration (Конфигурация снимка) ->;
  - Convert Existing Snapshot For Online Planning Worksheets (Преобразовать существующий снимок для OLPW);
  - выберите предварительно созданный снимок;
- после создания файла определения кластера откройте его в Online Planning Worksheets.

3.16.3 Создание отчета о конфигурации

Отчет о конфигурации позволяет записать информацию о состоянии конфигурации кластера в формате HTML.

Отчет содержит обзорную информацию, включающую следующее:

- имя каталога, содержащего изображения, используемые в отчете;
- версия приложения Online Planning Worksheets;
- автор и компания, указываемые на панели Cluster Configuration (Конфигурация кластера);
- примечания кластера, добавленные с панели Cluster Notes (Примечания кластера);
- последние дата и время, в которые система Online Planning Worksheets сохранила файл определения кластера.

Отчет также содержит разделы по следующим вопросам:

- узлы и пути для связи;
- приложения;
- сети;
- экспорт NFS;
- IP-метки;
- серверы приложений;

- глобальная сеть;
- мониторы приложений;
- сайты;
- пейджеры и мобильные телефоны;
- диски;
- удаленные уведомления;
- группы ресурсов;
- ресурсы накопителей на магнитной ленте;
- группы томов;
- политики времени выполнения для групп ресурсов;
- логические тома;
- обзор узлов;
- наборы файлов;
- верификация кластера;
- межсайтовое зеркальное отображение LVM.

Для создания отчета о конфигурации:

- выберите File (Файл) > Create Report (Создать отчет);
- в диалоговом окне Save (Сохранение) введите имя и расположение файла отчета.

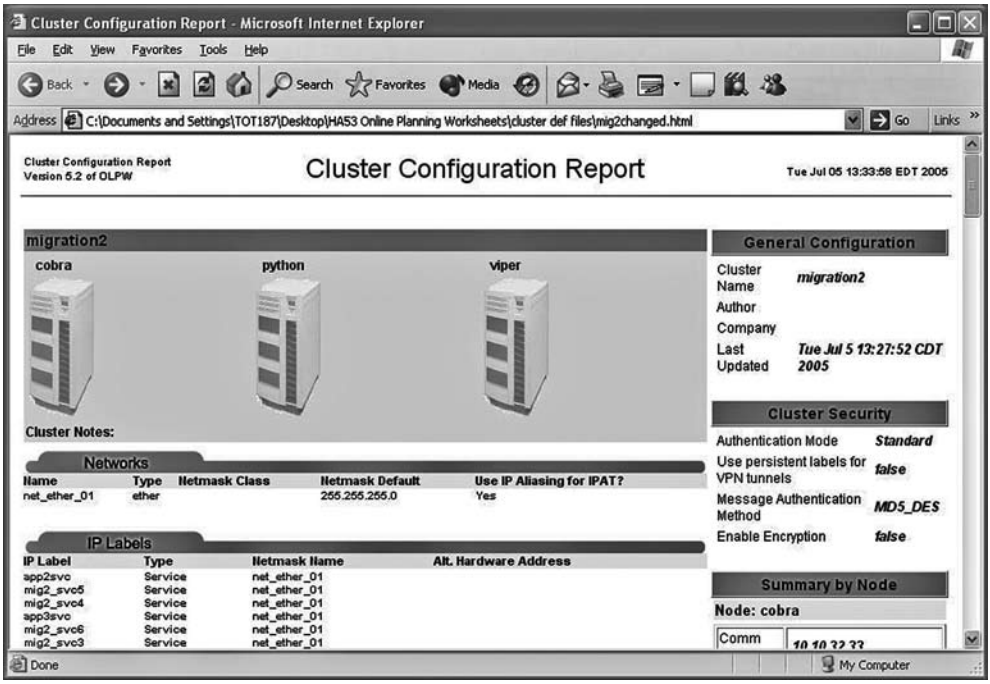

Рис. 3.22. Образец отчета о конфигурации

При создании отчета в каталоге, содержащем отчет, создается каталог olpwimages. Например, при сохранении файла отчета в каталоге /home/pat/reports в качестве каталога изображений используется /home/pat/reports/olpwimages. Каталог olpwimages содержит графические файлы, связанные с отчетом. При каждом создании отчета происходит замена файлов отчета и файлов в каталоге изображений.

На рис. 3.22 показан скриншот созданного отчета. Можно выполнять прокрутку страницы для просмотра информации.

3.17 Управление изменениями и проблемами

После запуска кластера начинается работа по управлению изменениями и проблемами.

Эффективные процессы управления изменениями и проблемами необходимы для обеспечения доступности кластера. В целях эффективности текущая конфигурация кластера всегда должна быть «под рукой». Можно использовать OLPW для создания html-версии конфигурации, а также (рекомендуется) схемы текущего кластера.

Любые изменения в кластере должны быть всесторонне исследованы с точки зрения их воздействия на функционирование кластера. Даже те изменения, которые непосредственно не влияют на HACMP (например, добавление дополнительной нагрузки, не связанной с HACMP), могут отразиться на работе кластера. Необходимо выполнять планирование, назначение и документирование изменений, а после их внесения – тестирование кластера.

Чтобы упростить внедрение изменений в кластере, HACMP обеспечивает набор SMIT-меню C-SPOC (Cluster Single Point of Control). Всегда, когда это возможно, следует использовать меню C-SPOC для внесения изменений. Используя C-SPOC, можно вносить изменения на одном узле, после чего они распространяются на другие узлы кластера.

Все возникающие проблемы в кластере следует немедленно изучать и исправлять. Так как основная задача HACMP состоит в том, чтобы скрывать любые возникающие ошибки от приложений, даже при наличии инструментов мониторинга вы можете не узнать о перемещении при сбое. Убедитесь в том, что включены уведомления об ошибках, сообщающие об отказах соответствующему персоналу.

3.18 Инструменты планирования

В этом разделе подробно рассматриваются три основных инструмента планирования. Также включен образец схемы кластера и таблицы планирования.

3.18.1 Схема кластера

Создание схемы для кластера HACMP позволяет четко представить работу кластера и помогает выявить единые точки отказа. Образец схемы кластера из двух узлов представлен на рис. 3.23.

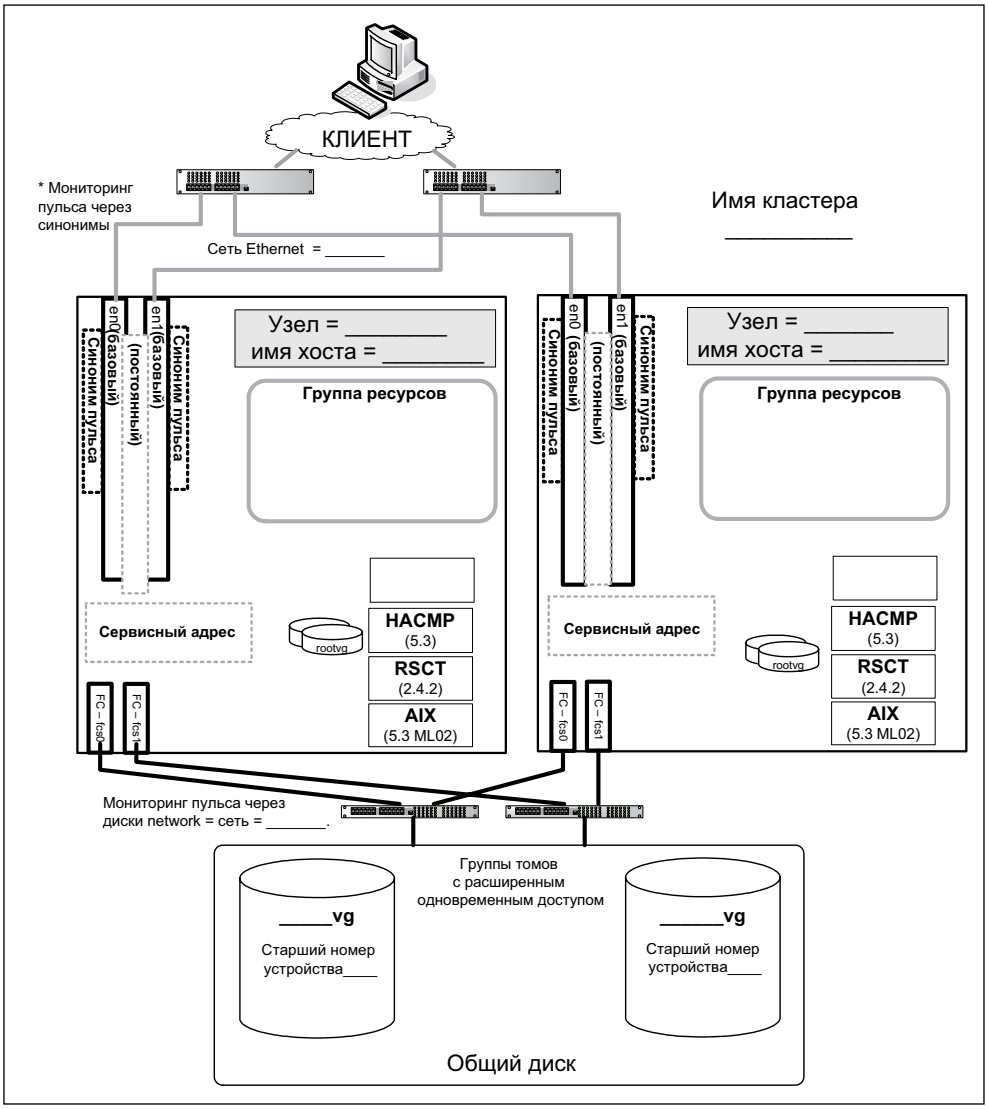

Рис. 3.23. Образец схемы кластера

3.18.2 Система автоматизированного планирования

Система автоматизированного планирования (Online Planning Worksheets, OLPW) представляет Java-версию таблиц планирования на бумаге. Используя это приложение, можно либо импортировать информацию о конфигурации HACMP из существующего кластера и отредактировать ее, либо ввести всю информацию о конфигурации вручную.

Приложение сохраняет информацию в виде файла определения кластера, применимого для конфигурирования кластера. Приложение также проверяет данные на наличие всей необходимой информации.

Следующий порядок действий представляет эффективное использование инструмента OLPW при планировании, внедрении и документировании кластера.

- Подготовка к планированию кластера путем ознакомления с понятиями HACMP и со своей средой.
- Запуск программы Online Planning Worksheet на своей рабочей станции с компакт-диска HACMP.
- Заполнение таблиц OLPW и создание файла определения кластера.
- Установка HACMP на узлах кластера.
- Копирование файла определения кластера на один из узлов кластера.
- Применение файла определения кластера в кластере с использованием команды cl_opsconfig.
- Создание снимка кластера.
- Документирование кластера путем создания отчета, при котором создается html-файл, который можно использовать для управления системами.

Внимание! Online Planning Worksheets представляет собой инструмент для конфигурирования и записи состояния кластера. Вы должны иметь четкое представление о планировании кластера, прежде чем использовать этот инструмент.

Запуск OLPW с компакт-диска в Windows

- Вставьте компакт-диск HACMP в соответствующий привод.
- Найдите файл olpw/worksheets.bat и запустите его.

Примечание. Не закрывайте командное окно, использовавшееся для запуска приложения. При закрытии этого окна произойдет закрытие приложения.

Запуск приложения с компакт-диска в AIX 5L

В системе AIX 5L для запуска приложения Online Planning Worksheets с компакт-диска HACMP проделайте следующее:

- Убедитесь в том, что в переменной окружения PATH установлен путь к JRE следующим образом:
  - AIX 5.3 - /usr/java141/bin;
  - AIX 5.2 - /usr/java131/bin;
  - AIX 5.1 - /usr/java130/bin.
- Подключите установочный носитель с использованием следующей команды:

mount -v cdrfs -p -r cd_location mount_directory, где cd_location – расположение компакт-диска; mount_directory – имя подключаемого каталога (точки монтирования). Например: mount -v cdrfs -p -r /dev/cd0 /mnt.

- Запустите приложение с использованием следующей команды:

  java -jar mount_directory/olpw/worksheets.jar, где mount_directory – каталог, описанный выше.

Установка OLPW

- Установка приложения в системе AIX 5L. Установка приложения Online Planning Worksheets выполняется с установочного носителя программного обеспечения HACMP. Устанавливаемый образ приложения:

  cluster.es.worksheets

  Приложение Online Planning Worksheets устанавливается в каталоге /usr/es/sbin/cluster/worksheets.

- Запуск приложения OLPW из графического интерфейса AIX 5L. Выполните следующую команду:

  /usr/es/sbin/cluster/worksheets/worksheets

  Приложение проверяет наличие установленной требуемой версии JRE перед запуском приложения в фоновом режиме.

Установка приложения OLPW в системе Windows

- Установите Online Planning Worksheets с установочного носителя HACMP в системе AIX 5L.
- Скопируйте файлы worksheets.bat и worksheets.jar в требуемый каталог в системе Windows.

> **Примечание.** При копировании файлов через FTP обязательно задайте режим ASCII для .bat-файла и режим binary для .jar-файла.

Запуск приложения из системы Windows

Выполните команду worksheets.bat из командной строки либо сделайте двойной щелчок мышью по значку worksheet.jar в проводнике.

Описание главного окна

При открытии приложения Online Planning Worksheets выводится его главное окно, представленное на рис. 3.24:

Главное окно содержит две панели:

- левая панель позволяет осуществлять навигацию по компонентам кластера;
- правая панель выводит панели, связанные со значками, выбранными на левой панели.

  Информация о конфигурации вводится в правой панели.

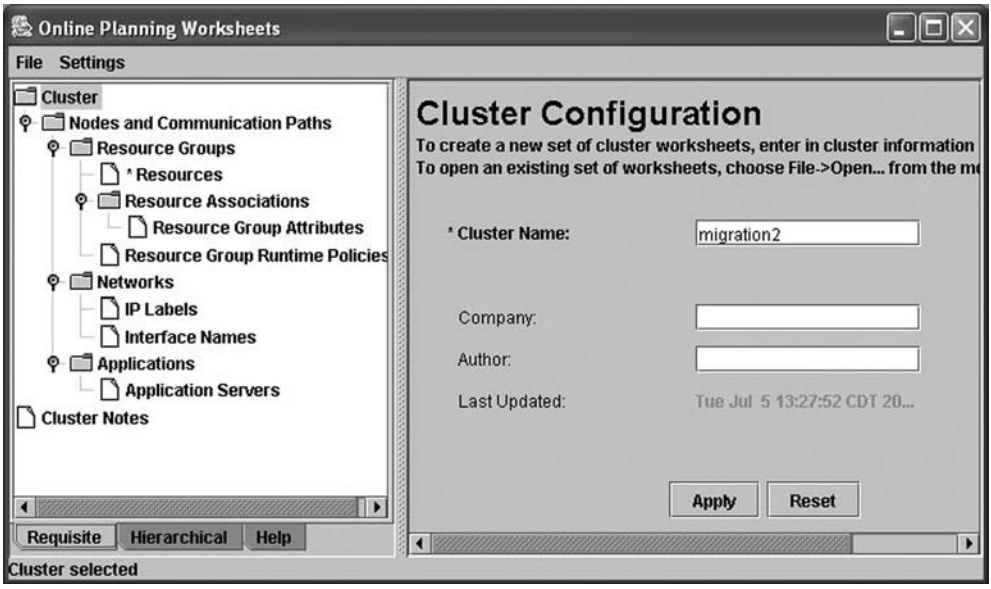

Рис. 3.24. Основное меню Online Planning Worksheets

Создание нового файла описания кластера

В начале планирования кластера можно либо ввести всю информацию вручную, либо считать информацию о конфигурации кластера в OLPW, после чего ввести остальную информацию вручную.

Для создания файла определения кластера проделайте следующее:

- Введите все данные вручную с использованием информации планирования

  или считайте информацию о конфигурации непосредственно из кластера HACMP следующим образом:

  – Используя меню SMIT, создайте файл определения.

  – В OLPW выберите File (Файл) > Import HACMP Definition (Импорт определения HACMP), после чего выберите файл определения кластера.

  – Выводится диалоговое окно Import Validation (Проверка импорта). Здесь можно просмотреть информацию об ошибках проверки либо получить подтверждение об успешном импорте файла определения HACMP.

  – Введите всю дополнительную информацию вручную

- Сохраните созданный файл определения.

Открытие существующего файла определения кластера

Примечание. Файл определения кластера должен находиться на том же узле, на котором выполняется приложение Online Planning Worksheets. Приложение Online Planning Worksheets поддерживает открытие файлов определения кластера со следующими расширениями:

- .haw. Это расширение является предпочтительным. Оно поддерживается в HACMP 5.2 и 5.3.
- .xml. Это расширение также может использоваться. Оно поддерживается в HACMP 5.3.
- .ws. Этот файловый формат поддерживается в HACMP 5.1.0.1. В целях обратной совместимости в HACMP 5.3 существует возможность открытия .ws-файлов; однако их следует сохранять с расширением .xml или .haw.

Чтобы открыть файл определения кластера, выберите File (Файл) > Open (Открыть). Одновременно можно открыть только один файл определения кластера.

Программа запросит вас о необходимости сохранения текущего файла, независимо от того, вносились ли в него какие-либо изменения, после чего главное окно выводится без информации о конфигурации.

Добавление примечаний о конфигурации кластера

При планировании конфигурации в файл определения кластера можно добавлять примечания. Для этого:

- на левой панели, либо в представлении Requisite®, либо в иерархическом представлении выберите Cluster Notes (Примечания кластера);
- на панели Cluster Notes (Примечания кластера) введите информацию, которую требуется сохранить;
- нажмите кнопку Apply (Применить).

Сохранение файла определения кластера

Это можно сделать двумя способами:

1. Выберите File (Файл) > Save (Сохранить), чтобы использовать имя кластера в качестве имени файла.

2. Выберите File (Файл) > Save As (Сохранить как), чтобы ввести другое имя файла. В диалоговом окне Save (Сохранение) введите имя и расположение файла определения кластера; убедитесь, что имя файла имеет расширение .haw (или .xml), и нажмите Save (Сохранить).

При сохранении файла OLPW автоматически проверяет определение кластера, если только автоматическая проверка не была выключена.

Применение данных таблиц в кластере HACMP

После заполнения панелей конфигурации в приложении OLPW можно сохранить файл, после чего применить его на узле кластера. При использовании приложения Online Planning Worksheets в системе Windows необходимо сначала скопировать файл определения кластера на узел кластера, прежде чем его применять.

Предварительные условия

Прежде чем применять файл определения кластера в кластере, убедитесь в том, что выполнены следующие условия:

- Программное обеспечение HACMP установлено на всех узлах кластера.
- Все аппаратные устройства, заданные в конфигурации кластера, подключены.
- В случае замены существующей конфигурации вся существующая информация кластера в базе данных конфигурации HACMP была сохранена в снимке.
- Службы кластера остановлены на всех узлах.
- Рабочий файл /usr/es/sbin/cluster/etc/rhosts присутствует на всех узлах кластера. Это необходимо для выполнения утилиты cl_opsconfig.

Применение файла конфигурации кластера

- Из приложения Online Planning Worksheets проверьте файл определения кластера.
- Создайте отчет для документирования конфигурации кластера.
- Сохраните файл и выйдите из приложения. Если файл конфигурации кластера находится в системе Windows, скопируйте файл на узел HACMP.
- С узла кластера выполните команду cl_opsconfig:

  /usr/es/sbin/cluster/utilities/cl_opsconfig your_config_file

  где your_config_file – имя файла конфигурации на узле.

Утилита cl_opsconfig выполняет проверку файла (если приложение Online Planning Worksheets установлено локально), применяет информацию в кластере, выполняет синхронизацию и верификацию. Во время верификации на экран выводятся сообщения о возникающих событиях, а также предупреждения и ошибки. Сообщения об ошибках cl_opsconfig можно просматривать на экране либо перенаправлять в файл журнала.

Перенаправление стандартного потока вывода сообщений об ошибках осуществляется следующим образом (в оболочке korn shell; в других оболочках могут быть отличия):

```
/usr/sbin/cluster/utilities/cl_opsconfig your_config_file 2> output_file
```

3.18.3 Таблицы планирования на бумаге

Подробно таблицы планирования на бумаге представлены в прил. «А» в руководстве **HACMP 5.3 Planning and Installation Guide**.

Мы считаем, что полезно привести эти таблицы в формат, подходящий к вашей среде. Для этого мы включили ряд примеров специализированных таблиц, чтобы помочь вам в планировании простого кластера. Эти таблицы см. в прил. «А», «Таблицы планирования на бумаге».

4

Сценарии установки кластера

В этой главе рассматриваются следующие темы:

- подготовка аппаратного и программного обеспечения кластера;
- конфигурирование WebSMIT;
- общие аспекты конфигурирования кластера;
- стандартный путь конфигурирования – Two-Node Configuration Assistant;
- использование расширенного пути конфигурирования и C-SPOC.

4.1 Основные этапы внедрения кластера HACMP

В этом разделе мы представим основные этапы, которым нужно следовать при внедрении кластера HACMP. Хотя целевая конфигурация в разных реализациях может слегка различаться, основные этапы одинаковы; могут иметь место лишь небольшие изменения в последовательности.

Основные этапы внедрения кластера высокой доступности

1. Планирование. Этот этап наиболее важен, так как он требует глубоких знаний и четкого представления о своей среде. Тщательное планирование является основой успешного внедрения кластера. Дополнительные сведения о методах планирования см. в гл. 3, «Планирование».

> **Примечание.** Помимо конфигурации кластера, этап планирования должен также обеспечивать план тестирования кластера. Этот план тестирования следует использовать на последнем этапе внедрения, а также при периодических проверках кластера.

2. Установка и подключение оборудования. На этом этапе аппаратная среда уже должна быть подготовлена в соответствии с конфигурацией, определенной на этапе планирования. Должны быть выполнены следующие задачи:

- установка оборудования pSeries (стойки, источники питания, консоль управления оборудованием HMC и т. д.);
- конфигурирование разделов (где применимо);
- подключение компьютеров к локальной сетевой среде;
- подключение компьютеров к хранилищу (SAN).

3. Установка и конфигурирование базовой операционной системы (AIX) и выполнение обязательных условий для работы HACMP. На данном этапе нужно выполнить следующие задачи:

- установка базовой операционной системы, приложения и обязательных пакетов для работы HACMP (CD-ROM, Network Install Manager) в соответствии с локальными правилами;
- конфигурирование локальной сетевой среды (конфигурирование TCP/IP – интерфейсы, разрешение имен и т. д.);
- конфигурирование пользователей, групп, аутентификации и т. д.

4. Конфигурирование общего хранилища. Может включать (в зависимости от используемой подсистемы хранения):

- конфигурирование драйверов устройства хранения и многопутевых расширений (если применимо);
- конфигурирование соответствия логического хранилища физическому (RAID-массивов, LUN и т. д.) и защита хранилища;
- конфигурирование безопасности хранения (маскирование LUN, разделение SAN на зоны) – там, где применимо;

- конфигурирование метода хранения для приложения (файловые системы, логические тома прямого доступа или диски прямого доступа).

5. Установка и конфигурирование приложений. На данном этапе должно быть выполнено конфигурирование приложений и тестирование их выполнения на автономном узле. Также следует вручную выполнить перемещение и тестирование приложения на всех узлах, выделенных для выполнения приложения в кластере высокой доступности:

- создайте и протестируйте скрипты запуска и остановки приложения; убедитесь в том, что приложение способно восстанавливаться после неожиданных отказов, и что скрипты запуска/остановки работают должным образом на всех узлах, выделенных для выполнения приложения;
- создайте и протестируйте скрипты мониторинга приложения (если нужно) на всех узлах, выделенных для выполнения приложения.

6. Установка программного обеспечения HACMP и выполнение перезагрузки всех узлов. Это также обязательно после применения исправлений HACMP.

7. Определение кластера и обнаружение или определение вручную топологии кластера. Так как HACMP содержит несколько инструментов конфигурирования, можно выбрать между «стандартным» (простым) и «расширенным» (более сложным) путем конфигурирования. Кроме того, можно выбрать между введением всех данных топологии вручную и использованием механизма обнаружения HACMP, которое упрощает конфигурирование кластера.

8. Синхронизация топологии кластера и запуск служб HACMP (на всех узлах). На этом этапе мы рекомендуем выполнить верификацию и синхронизацию топологии кластера и запустить службы кластера. Верификация и синхронизация на данном этапе упрощают выполнение последующих этапов внедрения, так как ошибки конфигурации гораздо проще обнаружить и исправить их на данном этапе, обеспечивая надежную топологию кластера для дальнейшего конфигурирования ресурсов.

9. Конфигурирование ресурсов кластера. На данном этапе следует выполнить конфигурирование следующих ресурсов:

- сервисные IP-адреса (метки);
- серверы приложений (скрипты запуска/остановки приложений);
- мониторы приложений (скрипты и действия мониторинга приложений).

10. Конфигурирование групп ресурсов кластера и общего хранилища. Группы ресурсов кластера являются «контейнерами», используемыми для группирования ресурсов, управляемых совместно в HACMP. Изначально группы ресурсов определяются как пустые контейнеры:

- определите группы ресурсов; выполните синхронизацию кластера;
- определите общее хранилище (группы томов, файловые системы, дисковые системы сторонних производителей и т. д.);
- наполните группы ресурсов сервисными IP-метками, серверами приложений, группами томов, мониторами приложений.

11. Выполните синхронизацию кластера. Так как топология HACMP уже сконфигурирована и службы HACMP запущены, то после синхронизации кластера группы ресурсов будут подключены.

Проведите оценку кластера путем проверки сообщений (консоль, /tmp/hacmp.out и т. д.).

12. Проведите тестирование кластера. Как только кластер достигнет *стабильного (stable)* состояния, следует провести тестирование кластера.

> **Примечание.** Несмотря на возможность использования инструмента автоматического тестирования кластера, мы настоятельно рекомендуем также выполнять тщательное тестирование кластера вручную. Инструмент автоматического тестирования кластера особенно полезен: а) для документирования тестов и результатов, б) для обновления документации кластера.

4.2 Установка и конфигурирование WebSMIT

Помимо «классического» конфигурирования с использованием инструмента System Management Interface Tool (SMIT), HACMP V5.2 и более поздние версии также включают веб-интерфейс для конфигурирования кластера. Хотя необходимо выполнить некоторую подготовительную работу, использование веб-интерфейса для доступа к SMIT-панелям для управления является эффективным методом конфигурирования и обслуживания кластера.

WebSMIT также содержит графический интерфейс для мониторинга работы кластера HACMP. Следует помнить о том, что WebSMIT – это всего лишь интерфейс к меню SMIT и информации о состоянии кластера с некоторыми полезными дополнениями (например, выводом структуры меню SMIT), а также с простым для восприятия графическим интерфейсом.

Основные этапы установки и конфигурирования WebSMIT следующие:

- установка пакета HACMP WebSMIT;
- подготовка платформы – установка Apache и обязательных пакетов;
- конфигурирование защищенного доступа на веб-сервере Apache;
- конфигурирование WebSMIT и документирование;
- верификация и запуск страниц WebSMIT;
- конфигурирование и обслуживание кластера HACMP.

4.2.1 Установка веб-сервера Apache и обязательных пакетов

В этом разделе описывается установка и конфигурирование защищенного HTTP-сервера (Apache с использованием SSL) на узлах кластера. Вы можете установить Apache на всех выделенных узлах кластера или только на некоторых из них. Мы рекомендуем выполнить установку на всех узлах в кластере, так как при этом можно осуществлять администрирование кластера через WebSMIT с любого узла в кластере.

Важно. Несмотря на распространенное мнение, что установка веб-сервера на сервере в рабочей среде может вызвать проблемы безопасности, конфигурирование через WebSMIT обеспечивает ЗАЩИЩЕННЫЙ способ администрирования кластера. WebSMIT лишь предоставляет доступ к меню HACMP SMIT (а не к SMIT в целом), а также обеспечивает аутентификацию и шифрование трафика.

Прежде чем начать, следует просмотреть последний файл README в каталоге /usr/es/sbin/cluster/wsm на узлах кластера.

Примечание. Так как Apache и SSL не являются продуктами компании IBM и содержат программы шифрования, на которые распространяются экспортные ограничения в США и других странах, необходимо получить пакеты в соответствии с правилами вашей страны. IBM обеспечивает возможность копирования криптографических пакетов со своего сервера, однако вы ДОЛЖНЫ зарегистрироваться на веб-сайте IBM, прежде чем вам будет предоставлен доступ для копирования.
Процесс регистрации может занять до 24 ч, в зависимости от вашей географической зоны, к этому следует готовиться заранее.

Необходимо наличие следующих наборов файлов:

- rpm.rte,
- expat-XXXX.ppc.rpm,
- apache-XXXX.ppc.rpm,
- mod_ssl-XXXX.ppc.rpm,
- openssl-XXXX.ppc.rpm.

Где «XXXX» должен отражать текущую версию набора файлов, указанную в файле /usr/es/sbin/cluster/wsm/README на узлах кластера. На момент написания этой книги мы использовали следующие версии:

expat-1.95.7-1.aix5.1.ppc.rpm

apache-1.3.31-1ssl.aix5.1.ppc.rpm

mod_ssl-2.8.19-1ssl.aix5.1.ppc.rpm

openssl-0.9.7d-1.aix5.1.ppc.rpm

Проверьте, установлены ли они в вашей системе, с помощью команды

rpm -qa

А также проверьте наличие предыдущего списка пакетов RPM.

В текущих инсталляциях AIX 5L, RPM (Red Hat Package Manager) устанавливается по умолчанию. Если RPM не установлен в вашей системе, скопируйте его по адресу

ftp://ftp.software.ibm.com/aix/freeSoftware/aixtoolbox/INSTALLP/ppc/rpm.rte

и установите

installp -qacXgd rpm.rte

Копирование кода Apache и необходимых пакетов

В браузере введите следующий URL:

http://www-1.ibm.com/servers/aix/products/aixos/linux/download.html

Для сохранения скопированных файлов используйте каталог /tmp. Скопируйте пакет RPM для пакета expat, после чего щелкните по ссылке «AIX Toolbox Cryptographic Content» («Криптографическое содержимое пакета инструментов AIX»), войдите в систему, получите лицензию и скопируйте RPMS для пакетов openssl, apache и mod_ssl.

Установка Apache и обязательных пакетов

После копирования пакетов в каталог /tmp используйте rpm для установки четырех rpm-файлов (порядок установки этих пакетов является важным), как показано в примере 4.1.

Пример 4.1. Установка Apache и обязательных пакетов с использованием rpm

```
# cd /tmp
# rpm -ivh openssl-*.rpm
# rpm -ivh expat-*.rpm
# rpm -ivh apache-*.rpm
# rpm -ivh mod_ssl-*.rpm
```

Альтернативный метод установки rpm-пакетов заключается в использовании SMIT (или команды installp, так как команда installp способна обрабатывать rpm-пакеты) с применением быстрого пути smitty install_latest.

4.2.2 Конфигурирование WebSMIT

Программное обеспечение HACMP, включая набор файлов WebSMIT, к этому моменту уже должно быть установлено. См. раздел «Установка веб-сервера Apache и обязательных пакетов».

Для конфигурирования WebSMIT мы собираемся использовать веб-сервер Apache для предоставления меню HACMP SMIT с «виртуального хоста» («Virtual Host»).

Кроме того, в целях безопасности мы используем:

- аутентификацию пользователей (пользователей операционной системы);
- протокол Secure HTTP (HTTPS), чтобы избежать передачи открытым текстом, например при передаче паролей пользователей.
- отдельный порт для страниц WebSMIT – 42267/TCP (чтобы избежать конфликтов с другими веб-страницами, которые могут быть доступны через стандартные порты 80 и 443).

Примечание. Всегда внимательно просматривайте файл README в каталоге /usr/es/sbin/cluster/wsm. Также проверяйте наличие обновленной версии этого файла на веб-сайте IBM.

Следуйте файлу README и скопируйте файл конфигурации Apache. Эта конфигурация предполагает, что на вашем компьютере не установлен другой веб-сервер. Если же другой веб-сервер установлен, нужно следовать локальным правилам конфигурирования веб-сервера и проконсультироваться с локальным веб-администратором, прежде чем приступать к конфигурированию WebSMIT и Apache.

Начните с сохранения первоначального файла конфигурации Apache (httpd.conf). Затем скопируйте файл /usr/es/sbin/cluster/wsm/httpd.conf.sample, поставляемый с пакетом WebSMIT, в /etc/opt/freeware/apache/httpd.conf.

Измените файл конфигурации в соответствии с файлом README и параметрами своей среды (разрешение имен, путь, конфигурация IP).

Во втором разделе файла httpd.conf следует изменить директиву ServerName в соответствии с примером 4.2.

Пример 4.2. Изменение директивы ServerName в файле httpd.conf

```
..................................... Строки пропущены ...........................
#
#ServerName localhost.your.domain

>>>>>>>>>>>>>>>>> Заменить на: <<<<<<<<<<<<<<<<<<
#
#ServerName localhost.your.domain
ServerName ha53node1 <------- Добавление строки
```

Далее в третьем разделе файла httpd.conf следует изменить фрагмент VirtualHost таким образом, чтобы он обрабатывал страницы WebSMIT в соответствии с локальной конфигурацией. Следует добавить целый раздел, как показано в примере 4.3.

Пример 4.3. Изменение фрагмента VirtualHost

```
............................ Строки пропущены .....................
### Section 3: Virtual Hosts
............................ Строки пропущены ....................
##
## SSL Virtual Host Context
##
<VirtualHost _default_:443>
..................................... Строки пропущены ...............................
<VirtualHost>
##### ------------------ Добавлять отсюда -----------------------
</VirtualHost>
#!!!
```

```
#!!!!!                                                                    !!!!!
#!!!!!  Строки начинаются со следующих записей:                           !!!!!
#!!!!!  NameVirtualHost                                                   !!!!!
#!!!!!  <VirtualHost>                                                     !!!!!
#!!!!!                                                                    !!!!!
#!!!!! ServerName                                                         !!!!!
#!!!!! ServerAdmin                                                        !!!!!
#!!!!!                                                                    !!!!!
#!!!!!
```

```
#   both in parallel (to also allow the use of DSA ciphers, etc.)
SSLCertificateKeyFile /etc/opt/freeware/apache/ssl.key/server.key
#SSLCertificateKeyFile /etc/opt/freeware/apache/ssl.key/server-dsa.key

#   Server Certificate Chain:
#   Point SSLCertificateChainFile at a file containing the
#   concatenation of PEM encoded CA certificates which form the
#   certificate chain for the server certificate. Alternatively
#   the referenced file can be the same as SSLCertificateFile
#   when the CA certificates are directly appended to the server
#   certificate for convinience.
#SSLCertificateChainFile /etc/opt/freeware/apache/ssl.crt/ca.crt

#   Certificate Authority (CA):
#   Set the CA certificate verification path where to find CA
#   certificates for client authentication or alternatively one
#   huge file containing all of them (file must be PEM encoded)
#   Note: Inside SSLCACertificatePath you need hash symlinks
#         to point to the certificate files. Use the provided
#         Makefile to update the hash symlinks after changes.
#SSLCACertificatePath /etc/opt/freeware/apache/ssl.crt
#SSLCACertificateFile /etc/opt/freeware/apache/ssl.crt/ca-bundle.crt

#   Certificate Revocation Lists (CRL):
#   Set the CA revocation path where to find CA CRLs for client
#   authentication or alternatively one huge file containing all
#   of them (file must be PEM encoded)
#   Note: Inside SSLCARevocationPath you need hash symlinks
#         to point to the certificate files. Use the provided
#         Makefile to update the hash symlinks after changes.
#SSLCARevocationPath /etc/opt/freeware/apache/ssl.crl
#SSLCARevocationFile /etc/opt/freeware/apache/ssl.crl/ca-bundle.crl

#   Client Authentication (Type):
#   Client certificate verification type and depth. Types are
#   none, optional, require and optional_no_ca. Depth is a
#   number which specifies how deeply to verify the certificate
#   issuer chain before deciding the certificate is not valid.
#SSLVerifyClient require
#SSLVerifyDepth 10

#   Access Control:
#   With SSLRequire you can do per-directory access control based
#   on arbitrary complex boolean expressions containing server
```

```
#   variable checks and other lookup directives. The syntax is a
#   mixture between C and Perl. See the mod_ssl documentation
#   for more details.
#<Location />
#SSLRequire ( %{SSL_CIPHER} !~ m/^(EXP|NULL)/ \
#               and %{SSL_CLIENT_S_DN_O} eq "Snake Oil, Ltd." \
#               and %{SSL_CLIENT_S_DN_OU} in {"Staff", "CA", "Dev"} \
#               and %{TIME_WDAY} >= 1 and %{TIME_WDAY} <= 5 \
#               and %{TIME_HOUR} >= 8 and %{TIME_HOUR} <= 20 ) \
#               or %{REMOTE_ADDR} =~ m/^192\.76\.162\.[0-9]+$/
#</Location>

#
#   Aliases: Add here as many aliases as you need (with no limit). The format is
#   Alias fakename realname
#
<IfModule mod_alias.c>

#
#   Note that if you include a trailing / on fakename then the server will
#   require it to be present in the URL. So "/icons" isn't aliased in this
#   example, only "/icons/". If the fakename is slash-terminated, then the
#   realname must also be slash terminated, and if the fakename omits the
#   trailing slash, the realname must also omit it.

#
#   ScriptAlias: This controls which directories contain server scripts.
#   ScriptAliases are essentially the same as Aliases, except that
#   documents in the realname directory are treated as applications and
#   run by the server when requested rather than as documents sent to the client.
#   The same rules about trailing "/" apply to ScriptAlias directives as to
#   Alias.
#
        ScriptAlias /cgi-bin/ "/usr/es/sbin/cluster/wsm/cgi-bin/"

#
#   "/opt/freeware/apache/cgi-bin" should be changed to whatever your ScriptAlias ed
#   CGI directory exists, if you have that configured.
#
    <Directory "/usr/es/sbin/cluster/wsm/cgi-bin">
        AllowOverride AuthConfig
        Options None
        Order allow,deny
```

```
        <FilesMatch ".*\.cgi$|wsm_tree">
                Allow from all
        </FilesMatch>
    </Directory>
</IfModule>

#   SSL Engine Options:
#   Set various options for the SSL engine.
#   o   FakeBasicAuth:
#       Translate the client X.509 into a Basic Authorisation. This means that
#       the standard Auth/DBMAuth methods can be used for access control. The
#       user name is the `one line' version of the client's X.509 certificate.
#       Note that no password is obtained from the user. Every entry in the user
#       file needs this password: `xxj31ZMTZzkVA'.
#   o   ExportCertData:
#       This exports two additional environment variables: SSL_CLIENT_CERT and
#       SSL_SERVER_CERT. These contain the PEM-encoded certificates of the
#       server (always existing) and the client (only existing when client
#       authentication is used). This can be used to import the certificates
#       into CGI scripts.
#   o   StdEnvVars:
#       This exports the standard SSL/TLS related `SSL_*' environment variables.
#       Per default this exportation is switched off for performance reasons,
#       because the extraction step is an expensive operation and is usually
#       useless for serving static content. So one usually enables the
#       exportation for CGI and SSI requests only.
#   o   CompatEnvVars:
#       This exports obsolete environment variables for backward compatibility
#       to Apache-SSL 1.x, mod_ssl 2.0.x, Sioux 1.0 and Stronghold 2.x. Use this
#       to provide compatibility to existing CGI scripts.
#   o   StrictRequire:
#       This denies access when "SSLRequireSSL" or "SSLRequire" applied even
#       under a "Satisfy any" situation, i.e. when it applies access is denied
#       and no other module can change it.
#   o   OptRenegotiate:
#       This enables optimized SSL connection renegotiation handling when SSL
#   directives are used in per-directory context.
#SSLOptions +FakeBasicAuth +ExportCertData +CompatEnvVars +StrictRequire
<Files ~ "\.(cgi|shtml|phtml|php3?)$">
SSLOptions +StdEnvVars
</Files>
<Directory "/usr/es/sbin/cluster/wsm/cgi-bin">
```

```
SSLOptions +StdEnvVars
</Directory>

#   SSL Protocol Adjustments:
#   The safe and default but still SSL/TLS standard compliant shutdown
#   approach is that mod_ssl sends the close notify alert but doesn't wait for
#   the close notify alert from client. When you need a different shutdown
#   approach you can use one of the following variables:
#   o   ssl-unclean-shutdown:
#       This forces an unclean shutdown when the connection is closed, i.e. no
#       SSL close notify alert is send or allowed to received. This violates
#       the SSL/TLS standard but is needed for some brain-dead browsers. Use
#       this when you receive I/O errors because of the standard approach where
#       mod_ssl sends the close notify alert.
#   o   ssl-accurate-shutdown:
#       This forces an accurate shutdown when the connection is closed, i.e. a
#       SSL close notify alert is send and mod_ssl waits for the close notify
#       alert of the client. This is 100% SSL/TLS standard compliant, but in
#       practice often causes hanging connections with brain-dead browsers. Use
#       this only for browsers where you know that their SSL implementation
#       works correctly.
#   Notice: Most problems of broken clients are also related to the HTTP
#   keep-alive facility, so you usually additionally want to disable
#   keep-alive for those clients, too. Use variable "nokeepalive" for this.
#   Similarly, one has to force some clients to use HTTP/1.0 to workaround
#   their broken HTTP/1.1 implementation. Use variables "downgrade-1.0" and
#   "force-response-1.0" for this.
SetEnvIf User-Agent ".*MSIE.*" \
nokeepalive ssl-unclean-shutdown \
downgrade-1.0 force-response-1.0

#   Per-Server Logging:
#   The home of a custom SSL log file. Use this when you want a
#   compact non-error SSL logfile on a virtual host basis.
CustomLog /usr/es/sbin/cluster/wsm/logs/ssl_request_log \
"%t %h %{SSL_PROTOCOL}x %{SSL_CIPHER}x \"%r\" %b"
</VirtualHost>
```

Примечание. В нашей среде 192.168.100.31 представляет постоянный IP-адрес, соответствующий «ha53node1» в /etc/hosts.

Затем следует подготовить файлы и каталоги WebSMIT для локальной платформы, как показано в примере 4.4.

Пример 4.4. Подготовка файлов WebSMIT

```
# ha53node1_> cd /usr/es/sbin/cluster/wsm
# ha53node1_> chown nobody:sytem logs
# ha53node1_> chmod 775 logs
# ha53node1_> chmod 4511 cgi-bin/wsm_cmd_exec
# ha53node1_> ln -sf /usr/share/man/info/en_US/cluster/HAES \
> /usr/es/sbin/cluster/wsm/htdocs/en_US/HAES
```

> **Примечание.** При обновлении пакетов WebSMIT вам нужно будет перезапустить команды, представленные в Примере 4.4.

Просмотрите файл /ust/es/sbin/cluster/wsm/wsm.conf на наличие следующих параметров (представленных в примере 4.5).

Пример 4.5. Параметры конфигурации WebSMIT

```
AUTHORIZED_PORT=42267
REDIRECT_TO_HTTPS=1
AUTHORIZED_USERS=root
REQUIRE_AUTHENTICATION=1
```

Эти переменные конфигурации (здесь представлены значения по умолчанию) указывают, что трафик WebSMIT перенаправляется в порт 42267/TCP (как сконфигурировано в разделе VirtualHost для Apache), используется протокол связи HTTPS, доступ к меню HACMP SMIT через WebSMIT разрешен для пользователя «root» и что аутентификация пользователя обязательна.

4.2.3 Запуск веб-сервера Apache

После внесения изменений в конфигурацию можно запустить веб-сервер Apache. Рекомендуется тестировать конфигурацию перед каждым запуском сервера с использованием следующей команды:

apachectl configtest

SYNTAX OK

Проверьте наличие сообщений об ошибках. Если на этом этапе будет выдано предупреждение, это указывает на то, что с конфигурацией все нормально, так как еще не был настроен основной веб-сервер (а только виртуальный).

Запустите веб-сервер Apache с опцией SSL с использованием следующей команды:

#apachectl startssl

> **Примечание.** Если не удается запустить сервер с опцией SSL (с использованием команды apachectl start), вы не сможете получить доступ к страницам WebSMIT!

Чтобы остановить веб-сервер Apache, необходимо выполнить следующую команду:

apachectl stop

При каждом обновлении какого-либо из файлов конфигурации следует выполнять тестирование конфигурации, а также остановить и перезапустить Apache, чтобы изменения вступили в силу.

4.2.4 Доступ к страницам WebSMIT из браузера

В результате тестирования WebSMIT была подтверждена полная поддержка его использования с Internet Explorer v5 и выше. Также поддерживаются Mozilla, Firefox и прочие браузеры на основе Gecko. Работают все функции на основе Javascript, за исключением, пожалуй, функциональных клавиш. Функции на основе Javascript не работают в Netscape Navigator v4 и ниже. (Не проводилось тестирование использования с Opera и браузерами на основе KHTML.)

Сначала нужно запустить на своем браузере WebSmit, после чего задать в нем следующий URL:

https://<yourIPaddress>:42267

Появится страница, подобная изображенной на рис. 4.1:

По умолчанию доступ к WebSMIT осуществляется под учетной записью пользователя «root» (в соответствии с параметрами, заданными в файле /usr/es/sbin/cluster/wsm/wsm_smit.conf). Этот параметр можно заменить таким образом, чтобы он указывал на любого пользователя AIX (пользователь должен существовать и иметь допустимый пароль).

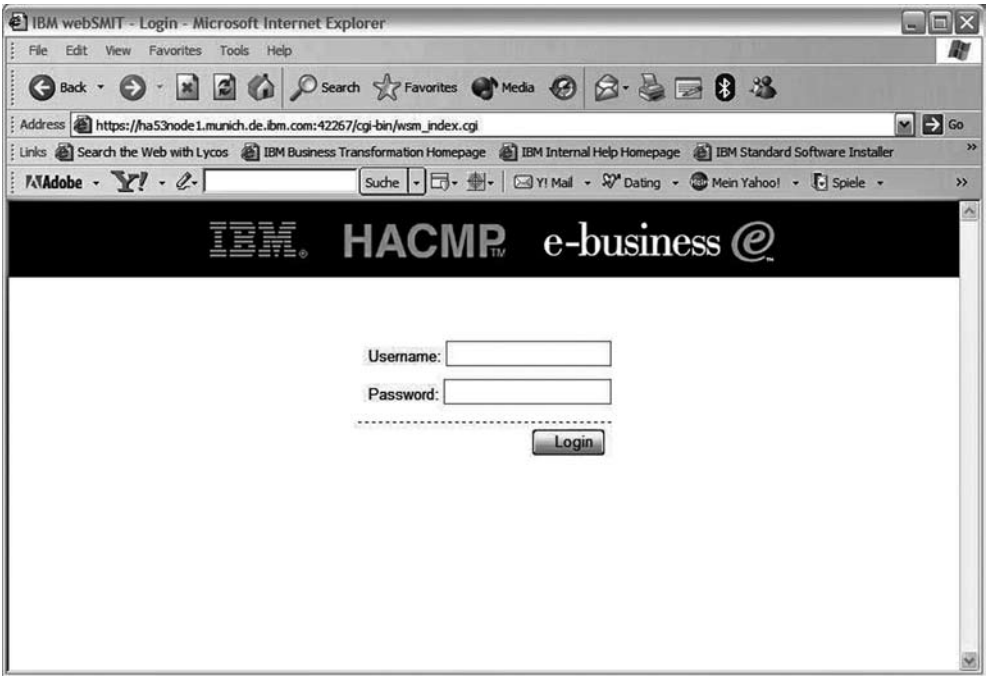

Рис. 4.1. Страница входа в WebSMIT

Внимание! Этот пользователь сможет запускать WebSMIT с правами пользователя «root», даже если у него нет таких прав в AIX.

После определения пользователей в операционной системе AIX вы можете добавлять или изменять пользователей WebSMIT в файле /usr/es/sbin/cluster/wsm/wsm_smit.conf.

4.2.5 Введение в WebSMIT

После успешного входа в WebSMIT появится экран, подобный изображенному на рис. 4.2:

Основная часть окна содержит три элемента:

- **Cluster Status (Состояние кластера)** – представляет обзор кластера и его состояния. Пример сконфигурированного и работающего кластера приведен на рис. 4.3.
- **Cluster Configuration and Management (Конфигурирование и управление кластером)** – предоставляет доступ к инструментам конфигурирования кластера. Описывается в разделе «Меню WebSMIT: Конфигурирование и управление кластером».
- **Online Documentation (Электронная документация)** – предоставляет доступ к системе документации HACMP (HACMP Documentation Bookshelf). Она содержит все руководства по HACMP 5.3.

Рис. 4.2. Главное меню WebSMIT

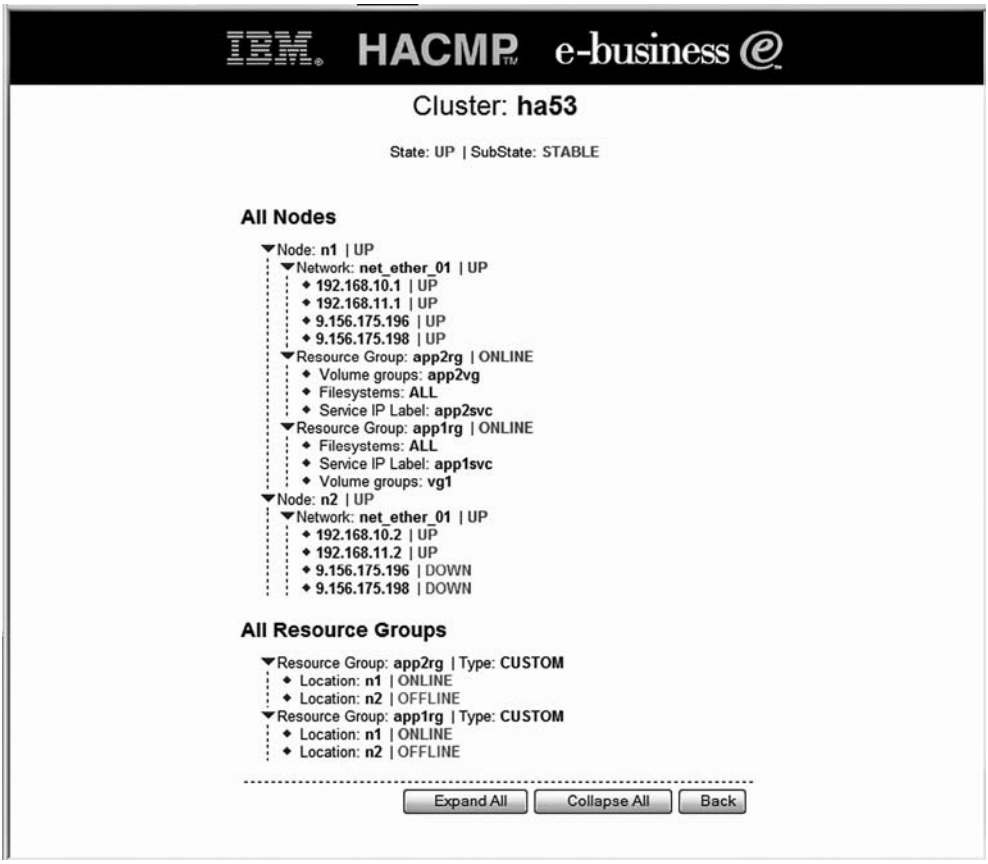

Рис. 4.3. Состояние кластера WebSMIT

> **Важно!** На странице состояния кластера также можно выполнять операции над узлами кластера, ресурсами и группами ресурсов. Обязательно прочитывайте информационные и справочные сообщения, прежде чем выполнять какие-либо действия.

4.2.6 Меню WebSMIT: конфигурирование и управление кластером

Меню Cluster Configuration and Management (Конфигурирование и управление кластером) предоставляет доступ к той же структуре и функциям меню, к которым предоставляет доступ и SMIT с меню HACMP. Меню WebSMIT имеют следующие преимущества в сравнении с меню SMIT:

- Структурное представление (**Treeview**) дает полный структурный обзор всех меню и подменю. Кроме того, оно может использоваться в качестве навигационной панели. Можно щелкнуть мышью по пункту в **Treeview**, и в главном окне WebSMIT произойдет переход в соответствующее меню.

- При перемещении курсора мыши над меню в главном окне выводятся всплывающие окна. Они содержат текст контекстно зависимой справки.
- В нижней части окна всегда выводится текущий быстрый путь. Этот быстрый путь соответствует используемому в меню SMIT.

Меню WebSMIT просты в применении. Помимо обычных функций SMIT, можно осуществлять обратный переход по страницам, используя следующие элементы управления:

- кнопка «Назад» («Back») в окне браузера;
- клавиша возврата (Backspace);
- клавиша F3;
- кнопка F3 в нижней части страницы.

Кнопка F1, показанная в нижней части страницы, выводит контекстно зависимую справку для текущей страницы. Кроме того, при перемещении курсора мыши по экрану выводится контекстно зависимая справка для каждой операции.

4.3 Конфигурирование HACMP

В этом разделе представлено базовое конфигурирование кластера с использованием различных инструментов и меню. Существует два способа базовой установки HACMP:

- Можно начать с использования Two-Node Configuration Assistant для создания базовой конфигурации и затем на основании этой базовой конфигурации настроить требуемую конфигурацию. Этот сценарий обсуждается в разделе 4.3.1, «Стандартный путь конфигурирования – Two-Node Configuration Assistant».
- С другой стороны, можно сначала сконфигурировать только топологию, а затем выполнить конфигурирование всех ресурсов, групп ресурсов и т. д. через расширенные меню. Этот сценарий представлен в разделе 4.3.2, «Использование расширенного пути конфигурирования и C-SPOC».

Прежде чем решить, какой способ следует использовать, нужно обязательно выполнить планирование и убедиться в том, что документация кластера готова к использованию. См. гл. 3, «Планирование».

В этой главе мы выполним конфигурирование с применением обоих сценариев в соответствии с планом, представленным на рис. 3.3.

Общие аспекты методов конфигурирования

Когда следует использовать стандартный путь конфигурирования. Использование стандартного пути конфигурирования позволяет добавлять базовые компоненты в базу данных конфигурации HACMP (ODM) за несколько простых этапов. Такой путь конфигурирования в значительной степени автоматизирует обнаружение и выбор сведений о конфигурации, а также определяет стандартные режимы работы для сетей и групп ресурсов.

Предварительные условия, предположения и параметры по умолчанию в стандартном пути:

- Программное обеспечение HACMP должно быть установлено на всех узлах кластера.
- Все сетевые интерфейсы должны быть как физически, так и логически настроены в AIX. Вы должны иметь возможность подключения с одного узла ко всем другим узлам и наоборот.
- Процесс обнаружения HACMP выполняется на всех серверных узлах, а не только на локальном узле.
- Хотя при использовании стандартного пути конфигурирования требуемая информация находится на удаленных узлах, HACMP автоматически собирает необходимую информацию о кластере. При использовании стандартного пути конфигурирования обнаружение кластера выполняется автоматически.

> **Ограничение.** При использовании стандартного пути конфигурирования нельзя выбрать перехват IP-адреса посредством замены. Эти параметры необходимо задавать в расширенном пути конфигурирования (Extended Configuration path).

- HACMP предполагает, что все сетевые интерфейсы в физической сети относятся к одной сети HACMP.
- В качестве имен узлов используются имена хостов.
- HACMP использует IP-синонимы по умолчанию.
- Осуществляется конфигурирование групп ресурсов с политиками запуска, перемещения при сбое и возврата после восстановления (без настройки политик таймеров возврата после восстановления).
- Стандартный путь конфигурирования позволяет задавать скрипты запуска и остановки приложений, однако скрипты мониторинга приложений можно внедрить только при использовании расширенного пути конфигурирования (Extended Configuration path).

Когда следует использовать расширенный путь конфигурирования. Если конфигурируются менее применяемые элементы кластера либо если связь со всеми узлами кластера недоступна на момент конфигурирования, можно вручную вводить информацию с использованием расширенного пути конфигурирования.

Использование опций в меню расширенного конфигурирования позволяет добавлять в базу данных конфигурации HACMP основные компоненты кластера, а также дополнительные типы режимов работы и ресурсов. Расширенный путь конфигурирования (Extended Configuration path) следует использовать для настройки тех компонентов, политик и опций кластера, которые не реализованы в меню Initialization (Инициализация) и Standard Configuration (Стандартное конфигурирование).

Расширенный путь конфигурирования (Extended Configuration path) следует использовать в следующих ситуациях:

- если нужно, чтобы имя узла отличалось от имени хоста;

- если требуется использовать IPAT посредством замены;
- если нужно добавить или изменить IP-сеть;
- если нужно добавить или изменить сети, отличные от IP, и устройства;
- если требуется сконфигурировать вариант размещения для сервисных IP-адресов;
- если требуется задать политики таймеров возврата после восстановления для групп ресурсов;
- если нужно добавить или изменить конфигурацию кластера, когда один из узлов недоступен, и т. д.

Постоянные IP-адреса

При конфигурировании кластера рекомендуется назначить постоянные IP-адреса для всех узлов. Постоянные IP-адреса назначаются в качестве синонимов поверх существующих IP-адресов для одного интерфейса на каждом узле, после чего они уникально идентифицируют определенный узел.

Роль постоянных IP-адресов состоит в обеспечении доступа к узлу независимо от состояния кластера HACMP. После синхронизации топологии кластера, пока узел включен (даже если HACMP не работает), доступен постоянный IP-адрес узла, позволяя осуществлять доступ для выполнения административных задач.

Возможный сценарий назначения постоянных IP-адресов см. на рис. 4.4.

Чтобы изменить базовые IP-адреса адаптера, нужно подключиться к узлам либо через системную консоль, либо через telnet с использованием интерфейса, который не подвергнется этому изменению. В данном случае следует на узле ha53node1 ис-

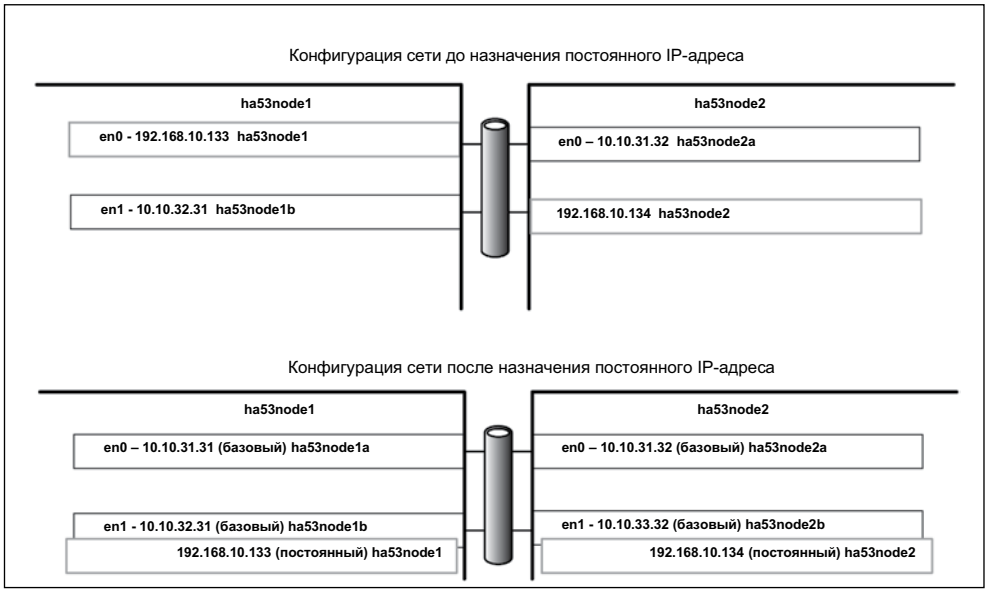

Рис. 4.4. Назначение постоянных IP-адресов

пользовать интерфейс en1 (IP-метка = ha53node1b) либо на узле ha53node2 использовать интерфейс en0 (IP-метка = ha53node2a).

После подключения к узлу следует проверить, какие интерфейсы включены.

netstat -i

Нужно найти сетевой интерфейс, которому на данный момент назначена постоянная IP-метка. После этого следует удалить эту метку:

#smitty

->Communications Applications and Services

->TCP/IP

->Further Configuration

->Network Interfaces

->Network Interface Selection

->Change / Show Characteristics of a Network Interface

Выберите интерфейс, который требуется изменить. Возникнет экран, подобный представленному на рис. 4.5.

На экране SMIT, показанном на рис. 4.5, удалите значения полей INTERNET ADDRESS (dotted decimal) и NETWORK MASK (hexadecimal or dotted decimal). Установите в поле Current STATE значение **down** или **detached.**

Чтобы убедиться в том, что состояние интерфейса было изменено на **down** или на **detached**, нужно выполнить следующую команду:

#netstat -i

Вполне возможно, что это изменение приведет к потере маршрута по умолчанию. Мы выполним конфигурирование маршрутизации позже.

Войдите в тот же экран SMIT, который показан на рис. 4.5. Можно использовать быстрый путь **smitty chinet**. Смените интерфейс и назначьте для него базовый адрес,

```
                    Change / Show a Standard Ethernet Interface

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
  Network Interface Name                                en2
  INTERNET ADDRESS (dotted decimal)                    [192.168.11.2]
  Network MASK (hexadecimal or dotted decimal)         [255.255.255.0]
  Current STATE                                         up                    +
  Use Address Resolution Protocol (ARP)?                yes                   +
  BROADCAST ADDRESS (dotted decimal)                   []
  Interface Specific Network Options
     ('NULL' will unset the option)
    rfc1323                                            []
    tcp_mssdflt                                        []
    tcp_nodelay                                        []
    tcp_recvspace                                      []
    tcp_sendspace                                      []
  Apply change to DATABASE only                         no                    +

F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Рис. 4.5. Изменение IP-адреса интерфейса

связанный с HACMP, через поля INTERNET ADDRESS и NETWORK MASK. Установите в поле Current STATE значение **up**.

Добавьте постоянную IP-метку:

#smitty
->Communications Applications and Services
->TCP/IP
->Further Configuration
->Network Interfaces
->Network Interface Selection
->Configure Aliases (select your IP version - we use IPV4)
->Add an IPV4 Network Alias

Выберите сетевой интерфейс, для которого следует назначить синоним. Скорее всего, этим интерфейсом будет интерфейс, на котором находится второй базовый адрес HACMP. Введите значения полей INTERNET ADDRESS и NETWORK MASK..

Проверьте наличие маршрута по умолчанию:

#netstat -rn (или lsattr -El inet0)

Если путь по умолчанию отсутствует, добавьте его с использованием следующей команды:

mkdev -l inet0

SMIT-панели, связанные с HACMP, и их структура

Использование WebSMIT позволяет отображать структуру (дерево) SMIT-меню, связанных с HACMP. Каждый треугольник указывает на наличие как минимум одного подменю. Это представлено на рис. 4.6.

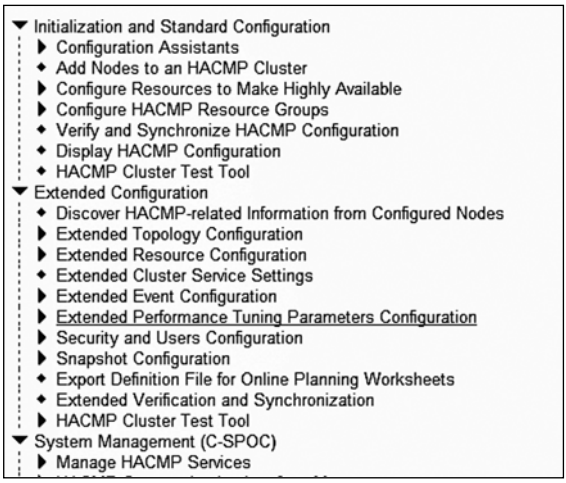

Рис. 4.6. Структура меню, связанных с HACMP, в WebSMIT (фрагмент)

4.3.1 Стандартный путь конфигурирования – Two-Node Configuration Assistant

Сначала мы выполним настройку очень простой топологии кластера с использованием WebSMIT. Мы применим инструмент Two-Node Cluster Configuration Assistant, в котором для базового конфигурирования кластера нужно будет ответить на пять вопросов. На этом этапе уже должна быть настроена сетевая конфигурация, конфигурация LVM и скрипты запуска и остановки приложений.

Эти пять элементов перечислены ниже.

- путь для связи с резервным узлом;
- имя сервера приложений;
- скрипт запуска сервера приложений;
- скрипт остановки сервера приложений;
- сервисная IP-метка.

Прежде чем приступить к конфигурированию, необходимо убедиться в том, что все общие предположения стандартного пути конфигурации соответствуют плану кластера. Более подробно это обсуждается в разделе «Общие аспекты методов конфигурирования».

Чтобы обеспечить требуемые результаты при использовании стандартного пути установки с применением Two-Node Configuration Assistant, необходимо выполнить следующие приготовления, прежде чем приступить к конфигурированию.

Конфигурирование сети

При подготовке постоянных IP-меток в разделе «Постоянные IP-адреса» уже было выполнено конфигурирование базовых сетей. Для запуска Two Node Configuration Assistant этого достаточно. Все остальное будет добавлено позднее.

Конфигурирование хранилища

Чтобы использовать Two Node Configuration Assistant, все, что касается групп томов, логических томов и файловых систем, должно быть сконфигурировано предварительно.

> **Примечание.** Если у вас уже есть группы томов (содержащие логические тома и файловые системы), сконфигурированные на внешних дисках, то, даже если они не будут связаны с приложением, интеграцию которого вы собираетесь выполнить, Two Node Configuration Assistant обнаружит эти группы томов и будет их использовать в составе кластера.

Ниже приведена процедура, которой мы рекомендуем придерживаться (см. далее соответствующие разделы настоящей работы).

1. Проверка конфигурации.
2. Создание группы томов с расширенным одновременным доступом.
3. Создание логического тома журнала для этой группы томов.
4. Создание требуемого количества логических томов.

5. Создание файловых систем для каждого определенного логического тома.
6. Подключение файловых систем.
7. Проверка отсутствия другого логического тома журнала.
8. Отключение всех файловых систем в группах томов, связанных с HACMP.
9. Деактивизация группы томов.
10. Импорт группы томов на другом узле с последующей верификацией.
11. Подключение всех файловых систем.
12. Документирование неподдерживаемых команд (varyonvg -c -P xxxx).

> **Примечание.** Этапы от создания группы томов с расширенным одновременным доступом до деактивизации группы томов должны выполняться только на одном узле.
> Этапы импорта группы томов на другом узле с последующей верификацией и подключения всех файловых систем должны выполняться на всех остальных узлах.

Проверка конфигурации

Для того чтобы сконфигурировать общий компонент LVM, необходимо убедиться в том, что все узлы видят все общие диски.

Для этого следует выполнить следующую команду на обоих узлах:

```
# lspv
```

Сравните выходные данные на обоих узлах, чтобы убедиться в том, что обе стороны видят одни и те же не назначенные в группы томов физические тома с одинаковым идентификатором физического тома (physical volume identifier, PVID).

Если вы не видите PVID для hdisk, следует запустить следующую команду на всех узлах, прежде чем выполнять дальнейшие операции:

```
# chdev -l hdiskX -a pv=yes
```

> **Примечание.** Все общие диски должны иметь одинаковый назначенный PVID. В противном случае вы не сможете создавать компоненты LVM.

Создание группы томов с расширенным одновременным доступом

Мы рекомендуем использовать в качестве общих групп томов только группы томов с расширенным одновременным доступом.

> **Примечание.** Режим расширенного одновременного доступа представляет собой метод определения группы томов с возможностью одновременного доступа (что позволяет использовать ее для приложения с одновременным доступом), независимо от того, будет эта группа томов употребляться в режиме одновременного доступа или нет.

Группа томов с расширенным одновременным доступом используется как для общих групп ресурсов (без совместного доступа), так и для групп ресурсов с сов-

местным доступом. Таким образом, группа томов может работать либо в режиме одновременного доступа (Concurrent mode), либо в общем режиме (Shared mode). Группа томов с возможностью одновременного доступа применяться в группах ресурсов с одновременным доступом под управлением HACMP и RSCT. Если группа томов с возможностью одновременного доступа используется в группе ресурсов без одновременного доступа, она предоставляет дополнительные возможности:

- Можно выполнять мониторинг пульса через диски с применением любой группы томов с расширенным одновременным доступом, не выделяя весь диск для мониторинга пульса через диски.
- Так как группа томов имеет расширенный одновременный доступ, она активизируется на всех узлах, но является активной только на одном узле. В HACMP это используется для выполнения быстрого перехвата дисков. Пассивная активизация группы томов предоставляет узлу информацию о действительном состоянии группы томов. Это позволяет выполнить быстрый переход к активной активизации при возникновении отказа. При этом обеспечивается безопасность и целостность данных в этих группах томов. Только один узел может иметь группу томов в активном состоянии.

> **Примечание.** Не рекомендуется активизировать группы томов с расширенным одновременным доступом вручную. Эту задачу должен координировать HACMP.

Для конфигурирования группы томов следует использовать SMIT:

#smitty mkvg

После этого появится экран, представленный на рис. 4.7:

Убедитесь в том, что в поле **Activate volume group AUTOMATICALLY at system restart? (Установить автоматическую активацию группы томов при перезапуске системы?)** установлено значение **No**.

```
                         Add an Original Volume Group

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                      [Entry Fields]
  VOLUME GROUP name                                  [app2vg]
  Physical partition SIZE in megabytes                                  +
* PHYSICAL VOLUME names                              [hdisk3]           +
  Force the creation of a volume group?               no                +
  Activate volume group AUTOMATICALLY                 no                +
    at system restart?
  Volume Group MAJOR NUMBER                          [110]              +#
  Create VG Concurrent Capable?                       enhanced concurrent +

F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Рис. 4.7. Определение группы томов с использованием SMIT

Если вы планируете использовать NFS для экспорта каталогов, расположенных в файловых системах, определенных в этой группе томов, следует также убедиться в том, что на всех узлах кластера установлено одинаковое уникальное значение параметра **Volume Group MAJOR NUMBER (Старший номер группы томов)**.

При этом не происходит немедленной активизации группы томов. Для активизации следует выполнить команду

```
# varyonvg app2vg
```

Создание логического тома журнала для этой группы томов

Мы рекомендуем вручную создать выделенный логический том для журналов JFS или JFS2. В этом случае вы сможете выбрать имя и расположение логического тома.

> **Важно!** Встроенные журналы для общих файловых систем JFS2 НЕ поддерживаются в HACMP.

Для определения логического тома используется команда

```
# smitty mklv
```

```
                          Add a Logical Volume

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

[TOP]                                                   [Entry Fields]
  Logical volume NAME                                  [app2loglv]
* VOLUME GROUP name                                     app2vg
* Number of LOGICAL PARTITIONS                         [1]                    #
  PHYSICAL VOLUME names                                [hdisk3]              +
  Logical volume TYPE                                  [jfs2log]             +
  POSITION on physical volume                           middle               +
  RANGE of physical volumes                             minimum              +
  MAXIMUM NUMBER of PHYSICAL VOLUMES                   []                     #
    to use for allocation
  Number of COPIES of each logical                      1                    +
    partition
  Mirror Write Consistency?                             active               +
  Allocate each logical partition copy                  yes                  +
[MORE...12]

F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Рис. 4.8. Создание логического тома журнала

Выберите только что созданную и активизированную группу томов. В нашем примере мы будем использовать группу томов `app2vg`, как показано на рис. 4.8.

Для параметра **Number of Logical Partitions (Количество логических разделов)** обычно достаточно задать значение **1**[1]. Убедитесь в том, что для параметра **Logical volume TYPE (Тип логического тома)** задано значение **jfslog** или **jfs2log**, в зависимости от ситуации.

Так как **jfslog** представляет специальный тип логического тома, этот логический том необходимо отформатировать. Для этого используется следующая команда:

[1] Подробнее про журналы файловых систем вы можете прочитать в документации по ОС AIX «Operating system and device management». *Примеч. науч. ред.*

logform /dev/app2loglv

где значение параметра **app2loglv** должно соответствовать имени, заданному в параметре Logical Volume Name (Имя логического тома) на предыдущем этапе.

Система спросит, следует ли ликвидировать (destroy) соответствующее устройство[1], нужно ответить Yes (Да).

Создание требуемого количества логических томов

Выберите повторно ту же группу томов. В нашем примере мы используем группу томов `app2vg`. Заполните значения на следующем экране в соответствии с рис. 4.9.

Пожалуйста, убедитесь в том, что значение типа логического тома соответствует значению, установленному для логического тома журнала. Повторите это действие для всех файловых систем, добавляемых к общей группе томов.

```
                          Add a Logical Volume

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

[TOP]                                                   [Entry Fields]
  Logical volume NAME                                  [app2lv]
* VOLUME GROUP name                                     app2vg
* Number of LOGICAL PARTITIONS                         [30]                    #
  PHYSICAL VOLUME names                                [hdisk3]               +
  Logical volume TYPE                                  [jfs2]                 +
  POSITION on physical volume                           middle                +
  RANGE of physical volumes                             minimum               +
  MAXIMUM NUMBER of PHYSICAL VOLUMES                   []                      #
    to use for allocation
  Number of COPIES of each logical                      1                     +
    partition
  Mirror Write Consistency?                             active                +
  Allocate each logical partition copy                  yes                   +
[MORE...12]

F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Рис. 4.9. Создание логического тома

Создание файловых систем для каждого определенного логического тома

Используйте SMIT:

#smitty

->System Storage Management (Physical &Logical Storage)

->File Systems

->Add /Change / Show Delete File Systems

->Enhanced Journaled File Systems

->Add a Enhanced Journaled File Systems on a Previously Defined Logical volume

Создайте все файловые системы на соответствующих логических томах, как показано на рис. 4.10.

[1] То есть отформатировать журнал. *Примеч. науч. ред.*.

```
                    Add an Enhanced Journaled File System

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                      [Entry Fields]
* LOGICAL VOLUME name                                  app2lv                +
* MOUNT POINT                                         [/app2]
  Mount AUTOMATICALLY at system restart?               no                    +
  PERMISSIONS                                          read/write            +
  Mount OPTIONS                                        []                    +
  Block Size (bytes)                                   4096                  +
  Logical Volume for Log                                                     +
  Inline Log size (MBytes)                             []                     #
  Extended Attribute Format                            Version 1             +
  ENABLE Quota Management?                             no                    +

F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Рис. 4.10. Создание файловых систем на предварительно определенных логических томах

Убедитесь в том, что в поле **Mount AUTOMATICALLY at system restart (Осуществлять ли подключение при перезапуске системы?)** установлено значение **No.**

Подключение файловых систем

Нужно выполнить подключение всех файловых систем командой

```
# mount /app2
```

Проверка отсутствия другого логического тома журнала

Теперь нужно убедиться в том, что используется созданный вами логический том журнала (см. пример 4.6).

Пример 4.6. Проверка логического тома журнала

```
root@ha53node1:/>
root@ha53node1:/> mount /app2
root@ha53node1:/> lsvg -l app2vg
app2vg:
LV NAME             TYPE       LPs   PPs   PVs  LV STATE      MOUNT POINT
app2loglv           jfs2log    1     1     1    open/syncd    N/A
app2lv              jfs2       30    30    1    open/syncd    /app2
root@ha53node1:/>
```

Отключение всех файловых систем в группах томов, связанных с HACMP

Нужно отключить все файловые системы, связанные с HACMP, командой

```
# umount /app2
```

Деактивизация группы томов

Деактивизация группы томов выполняется командой

```
#varryoffvg app2vg
```

Импорт группы томов на другом узле с последующей верификацией

Вы собираетесь импортировать группу томов на другой узел. Поэтому вы должны указать старший номер группы томов, установленный на первом узле. Кроме того, нужно убедиться в том, что задан правильный и соответствующий hdisk.

В нашем случае используется следующая команда:

```
# importvg -V 102 -y app2vg hdisk2
```

Следующий шаг состоит в том, чтобы активизировать группу томов:

```
# varyonvg app2vg
```

Подключение всех файловых систем

Чтобы убедиться в том, что все работает нормально, следует подключить все файловые системы и еще раз проверить, используется ли правильный логический том журнала:

```
# mount /app2
# lsvg -l app2vg
```

Затем нужно выполнить деактивизацию всех групп томов:

```
# varyoffvg app2vg
```

Подготовка приложения

Для того чтобы HACMP обеспечивал обслуживание ваших приложений, нужно подготовить набор из двух или трех скриптов для каждой группы ресурсов.

- Скрипт запуска запускает приложение.
- Скрипт остановки останавливает приложение.[1]

4.3.2 Использование расширенного пути конфигурирования и C-SPOC

В данном сценарии мы начинаем конфигурирование кластера с обнаружения топологии кластера, после чего используем C-SPOC для последующего конфигурирования. Это вносит некоторые изменения в список действий, приведенный в разделе «Основные этапы внедрения кластера HACMP», в частности этап 4, «Конфигурирование общего хранилища», и этап 5.1, «Создайте и протестируйте скрипты запуска и остановки приложения; убедитесь в том, что приложение способно восстанавливаться после неожиданных отказов и что скрипты запуска/остановки работают должным

[1] Очевидно, третий скрипт предназначен для мониторинга приложения (опционально). *Примеч. науч. ред.*

образом на всех узлах, выделенных для выполнения приложения», выполняются после этапа 7, «Определение кластера и обнаружение или определение вручную топологии кластера».

Все последующее конфигурирование кластера выполняется через C-SPOC во время работы кластера.

Конфигурирование топологии с использованием расширенного пути конфигурирования

При использовании расширенного пути конфигурирования вы имеете возможность более тонкой настройки опций и некоторых параметров конфигурации, недоступных через стандартный путь конфигурирования (Standard Configuration path).

Определение кластера

```
#smitty
->Communications Applications and Services
  ->HACMP for AIX
    ->Extended Configuration
      ->Extended Topology Configuration
        ->Configure an HACMP Cluster
          ->Add/Change/Show an HACMP Cluster
```

Выберите имя своего кластера.

Добавление узлов в кластер

После определения имени кластера следует добавить узлы и путь для связи (IP-адрес, назначаемый доступному сетевому интерфейсу) для каждого узла:

```
#smitty
->Communications Applications and Services
  ->HACMP for AIX
    ->Extended Configuration
      ->Extended Topology Configuration
        ->Configure HACMP Nodes
```

Введите имя узла и путь для связи с этим узлом. В расширенном пути конфигурирования можно задать имя узла, отличное от имени хоста. Повторите это действие для всех узлов, которые будут в кластере.

Добавление сетей (IP и отличных от IP)

Используя расширенный путь конфигурирования, нужно создать сеть и сетевые интерфейсы.

```
#smitty
->Communications Applications and Services
  ->HACMP for AIX
```

->Extended Configuration

->Extended Topology Configuration

->Configure HACMP Networks

Выберите **Add a Network** (Добавить сеть). Система запросит у вас тип сети. Проверьте, соответствует ли ваш тип перечисленным в разделе Discovered IP-based Network Types (Обнаруженные типы IP-сетей), и если это не так, выберите требуемый тип в разделе Pre-defined IP-based Network Types (Предопределенные типы IP-сетей). После этого появится экран, представленный на рис. 4.11.

```
                  Add an IP-Based Network to the HACMP Cluster

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                      [Entry Fields]
* Network Name                                       [net_ether_02]
* Network Type                                        ether
* Netmask                                            [255.255.255.0]          +
* Enable IP Address Takeover via IP Aliases          [Yes]                    +
  IP Address Offset for Heartbeating over IP Aliases []

F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Рис. 4.11. Добавление IP-сети

В этом меню можно указать, что не требуется применять метод перехвата IP-адреса посредством синонимов. Если требуется использовать IPAT посредством замены, выберите **No** (**Нет**).

Повторите эти действия для всех сетей (IP и отличных от IP), которые требуется добавить.

Коммуникационные интерфейсы и устройства

Следующий этап состоит в том, чтобы выполнить конфигурирование коммуникационных интерфейсов/устройств (Configure Communication Interfaces/Devices). Выберите соответствующий пункт меню на экране **Extended Topology Configuration (Расширенное конфигурирование топологии)**, на следующем экране выберите пункт **Add Communication Interface/Devices (Добавить коммуникационный интерфейс/устройства)**, после чего выберите пункт **Add Pre-defined Communication Interfaces and Devices (Добавить предопределенные коммуникационные интерфейсы и устройства)** и затем выберите пункт **Communication Interfaces (Коммуникационные интерфейсы)**. Будет выведена ранее созданная сеть. После ее выбора появится экран, представленный на рис. 4.12.

Повторите этот этап для всех коммуникационных интерфейсов и устройств, которые требуется сконфигурировать.

```
                        Add a Communication Interface

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                      [Entry Fields]
* IP Label/Address                                   [ha53node1b]          +
* Network Type                                        ether
* Network Name                                        net_ether_01
* Node Name                                          [node1]               +
  Network Interface                                  []


F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Рис. 4.12. Добавление коммуникационных интерфейсов

Обнаружение информации HACMP со всех узлов

Вместо того чтобы вручную вводить все данные, после определения кластера и всех узлов (с указанием пути для связи с каждым узлом), можно выполнить процедуру обнаружения HACMP:

#smitty

->Communications Applications and Services

->HACMP for AIX

->Extended Configuration

->Discover HACMP related Information from Configured Nodes

При этом выполнится автоматическое обнаружение кластера (IP и не IP-сетей коммуникационных интерфейсов и устройств, общих групп томов).

Обнаруженные данные будут представлены в виде выделения для расширенной топологии и меню ресурсов.

Верификация и синхронизация кластера

На экране SMIT **Extended Configuration (Расширенное конфигурирование)** также представлен следующий этап.

->Extended Verification and Synchronization

После выбора этого пункта появится экран, представленный на рис. 4.13.

Выберите требуемые значения; здесь нужно выбрать значения, представленные выше.

Верификацию и синхронизацию кластера можно выполнить в любое время. Однако существуют этапы, на которых выполнение этих операций является необходимым, например перед запуском служб кластера, после изменения конфигурации с использованием расширенного пути конфигурирования.

```
                    HACMP Verification and Synchronization

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                      [Entry Fields]
* Verify, Synchronize or Both                        [Both]                 +
* Automatically correct errors found during          [Interactively]        +
  verification?

* Force synchronization if verification fails?       [No]                   +
* Verify changes only?                               [No]                   +
* Logging                                            [Standard]             +

F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Рис. 4.13. Расширенная верификация и синхронизация HACMP

Конфигурирование остальных параметров через C-SPOC

Сокращение C-SPOC обозначает Cluster Single Point Of Control (Единая точка управления кластером) и является названием мощного инструмента администрирования, позволяющего выполнять изменения в конфигурации кластера с одного узла.

> **Примечание.** Автоматическое исправление ошибок может выполняться только при полной остановке HACMP на всех узлах в кластере.

Зачем использовать C-SPOC

Как системному администратору кластера HACMP, вам может потребоваться выполнять следующие задачи, связанные с LVM:

- создание новой общей группы томов;
- расширение, сокращение, изменение или удаление существующей группы томов;
- создание нового общего логического тома;
- увеличение, уменьшение[1], изменение или удаление существующего логического тома;
- создание новой общей файловой системы;
- увеличение[2], изменение или удаление существующей файловой системы;
- добавление, удаление физических томов.

---

[1] Уменьшение размера логического тома возможно только как часть уменьшения размера файловой системы. *Примеч. науч. ред.*

[2] В AIX 5.3 возможно уменьшение размера файловой системы JFS2 без ее отключения (с автоматическим уменьшением размера соответствующего логического тома). *Примеч. науч. ред.*

C-SPOC позволяет избежать большинства распространенных отказов, связанных со стабильностью информации кластера на всех узлах.

Команды C-SPOC работают только на тех компонентах LVM с общим доступом или с одновременным доступом, которые определены как часть группы ресурсов HACMP. При использовании SMIT HACMP C-SPOC выполнение команд происходит на узле, который является владельцем компонента LVM (узле, на котором он активизирован[1]).

Команды C-SPOC, выполняющие изменения компонентов LVM, требуют использования имени группы ресурсов в качестве аргумента. Компонент LVM, представляющий целевой объект команды, должен быть сконфигурирован в заданной группе ресурсов. C-SPOC применяет информацию группы ресурсов, чтобы определить, на каких узлах следует выполнять ту или иную операцию.

Единственной задачей, которую можно выполнить в C-SPOC без использования соответствующей группы ресурсов, является создание группы томов.

Удаление файловой системы или логического тома

При удалении файловой системы или логического тома с использованием C-SPOC, целевая файловая система или логический том не должны быть сконфигурированы как ресурс в заданной группе ресурсов. Вы должны удалить их конфигурацию из группы ресурсов перед удалением файловой системы или логического тома.

Обновление компонентов LVM в кластере HACMP

При изменении определения общего компонента LVM в кластере данная операция обновляет данные LVM, описывающие компонент, как на локальном узле, так и в области дескриптора группы томов (Volume Group Descriptor Area, VGDA) на дисках в группе томов. В AIX 5L LVM позволяет всем узлам в кластере получать информацию об изменениях группы томов, логического тома и файловой системы сразу же при внесении изменений, не ожидая получения этой информации во время медленных обновлений.

Дополнительные сведения о C-SPOC см. в гл. 9 руководства **High Availability Cluster Multi-Processing Administration Guide**, SC23-4862-06.

Как пользоваться C-SPOC

Нужно войти в меню Smitty C-SPOC следующим образом:

```
#smitty
->Communications Applications and Services
  ->HACMP for AIX
    ->System Management (C-SPOC)
```

Появится экран, представленный на рис. 4.14.

Дополнительные сведения о C-SPOC см. в разделе 8.1, «C-SPOC DP».

---

[1] Имеется в виду активизация (vary on) группы томов как части группы ресурсов. *Примеч. науч. ред.*

```
                          System Management (C-SPOC)

Move cursor to desired item and press Enter.

  Manage HACMP Services
  HACMP Communication Interface Management
  HACMP Resource Group and Application Management
  HACMP Log Viewing and Management
  HACMP File Collection Management
  HACMP Security and Users Management
  HACMP Logical Volume Management
  HACMP Concurrent Logical Volume Management
  HACMP Physical Volume Management
  Configure GPFS

  Open a SMIT Session on a Node



F1=Help             F2=Refresh          F3=Cancel           F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Рис. 4.14. Главное меню C-SPOC

5

Миграция кластера на HACMP V5.3

Эта глава описывает различные способы миграции кластера на HACMP 5.3. Перед миграцией следует провести планирование и рассмотреть аспекты миграции. Прежде чем начать, вы должны быть знакомы с использованием утилиты создания снимков кластера (cluster snapshot utility). Кроме того, изучите текущую конфигурацию, ее работу и изменения в каждом релизе, до которого выполняется обновление, так как в конечном результате все может работать не так, как ожидается.

Внимание! Всегда просматривайте список действий в руководстве по планированию и установке, если вы незнакомы с процедурой.

5.1 Выбор пути миграции

Существует три пути миграции, с использованием которых можно обновить кластер до более поздней версии. (Выбор пути миграции осуществляется путем ответа на вопрос, допустим ли перерыв в работе целого кластера, или же требуется обеспечивать максимальную доступность).

- **Циклическая миграция (Rolling Migration)** (с ES на ES). Этот метод обеспечивает доступность ресурсов как минимум на одном узле, пока происходит поочередное обновление остальных узлов кластера. Это означает, что во время миграции кластер будет находиться в смешанном режиме.
- **Метод снимков (Snapshot Method)**. Этот метод требует недоступности всех узлов кластера на некоторый период, а также того, чтобы на всех узлах была установлена новая версия HACMP, прежде чем преобразовывать и применять снимок до миграции.
- **Поузловая миграция (Node by Node)** (с HAS на ES). Этот метод подобен методу циклической миграции. Однако во время обновления каждого узла обе версии кода HACMP будут отображаться как установленные одновременно, а непосредственное управление кластером осуществляют старые демоны. Только после обновления последнего узла и его интеграции в кластер выполняется итоговое преобразование.

> **Замечание.** Обратите внимание на то, что в прошлом понятия циклической миграции и поузловой миграции использовались как синонимы.

5.1.1 Поддерживаемые пути миграции

Таблица 5.1. Поддерживаемые обновления версий до HACMP 5.3

Текущая версия	Циклическая миграция	Метод снимков	Поузловая миграция
HACMP/ES 5.2	да	да	неприменимо
HACMP/ES 5.1	да	да	неприменимо
HACMP/ES 4.5	да	да	неприменимо
HACMP 4.5	неприменимо	да	да

5.2 Требования

Перед обновлением HACMP вы должны быть знакомы с высокоуровневыми концепциями низкоуровневых задач, такими, как планирование, обслуживание и устранение неполадок, так как многие разделы в этой главе построены на этом знании.

Требования к дисковому пространству (для миграции с HAS на HAES):

Должно быть достаточно дискового пространства, чтобы в процессе миграции помещалось и программное обеспечение HAS, и программное обеспечение HAES:

- приблизительно 120 Мб в каталоге /usr;
- приблизительно 1.2 Мб в каталоге / (корневом каталоге).

Узлы должны иметь достаточно памяти, чтобы одновременно выполнять оба набора демонов HACMP. Минимальный объем оперативной памяти составляет 64 Мб, однако рекомендуется иметь 128 Мб оперативной памяти (не считая требований приложений).

Требования программного обеспечения кластера и RSCT

Требуется обеспечить одинаковый уровень наборов файлов программного обеспечения кластера и RSCT (включая PTF) на всех узлах до начала миграции. Кроме того, следует обеспечить перевод программного обеспечения в состояние committed (а не только в applied).

Чтобы проверить, установлено ли программное обеспечение в состояние committed:

1. Выполните команду **lslpp -h cluster.***

2. Если под заголовком действия выводится APPLY, введите smit install_commit перед установкой программного обеспечения HACMP. SMIT отобразит панель Commit Applied Software Updates (Remove Saved Files).

3. Введите следующие значения полей:

- SOFTWARE name (Название программного обеспечения). Из списка выберите все требуемые наборы файлов.
- COMMIT old version if above version used it? (COMMIT старую версию поверх используемой версии?). Установите для этого поля значение Yes[1].
- EXTEND file system if space needed? (Расширить файловую систему, если требуется пространство?). Установите для этого поля значение Yes.

Изучение конфигурации

Нужно знать конфигурацию своего кластера. Нужно изучить и понять общую конфигурацию до выполнения обновления. Если среда вам незнакома, выполните команду **cltopinfo** или **cldump**, чтобы получить информацию о среде. Выходные данные команд **cllsif и clshowres** также позволяют получить представление о конфигурации кластера и ожидаемом выполнении перемещения при сбое. Эта информация также полезна при необходимости обращения в службу поддержки AIX.

- /usr/es/sbin/cluster/utilities/cllsif
- /usr/es/sbin/cluster/utilities/clshowres

Первая команда отображает текущую конфигурацию топологии. Это поможет определить доступные сети, соответствующие метки и IP-адреса и их текущие функции. Вторая команда выводит все группы ресурсов и связанные с ними атрибуты. Это может помочь в определении текущих выделенных ресурсов и их функционирования во время перемещения при сбое.

---

[1] Такого поля в smit install_commit нет. (Проверялось в AIX 5.2 и 5.3). *Примеч. науч. ред.*

Рекомендации и проверки перед миграцией

Ниже перечислены действия, которые следует предпринимать перед любой миграцией, чтобы обеспечить ее успешное выполнение и наличие надежного плана возврата.

- Всегда создавайте снимок перед выполнением миграции. Не забудьте сохранить его в нескольких безопасных местах.

Всегда выполняйте резервное копирование системы (**mksysb**) с текущей конфигурацией. Следует проверить содержимое резервной копии и убедиться в том, что с ее помощью можно выполнить восстановление данных.

- При наличии ресурсов рассмотрите вариант alternate disk installation[1] для резервного диска на каждом узле кластера перед выполнением миграции. Это полезно в том случае, если вам придется быстро выполнить возврат к старой конфигурации. Для выполнения возврата нужно изменить список загрузочных устройств (bootlist), указав в нем резервный диск, после чего выполнить перезагрузку узлов, на которых была выполнена миграция.
- Прежде чем начать миграцию, убедитесь, что на всех узлах установлен одинаковый набор файлов программного обеспечения кластера и RSCT (включая PTF).
- Файл /.rhosts нужен только при миграции с более ранних версий, чем HACMP 5.1. После выполнения миграции рекомендуется удалить файл .rhosts в корневом каталоге (/), если другие приложения не требуют rsh для связи между узлами.

> **Примечание.** Если ваше приложение и/или скрипты пред- и постобработки требуют удаленного выполнения команд (на основе AIX rsh), необходимо сохранить файл ~/.rhosts даже после миграции.

- Убедитесь в том, что ваш кластер успешно выполняет перемещение при сбое, до миграции; в противном случае оно может не работать и после обновления.
- Проверьте состояние системы и установленных наборов файлов:
  - выполните команду lppchk -v, -l, -c;
  - выполните команду instfix -i | grep ML;
  - просмотрите выходные данные команды errpt -a на наличие последних соответствующих ошибок;
  - выполните команду df -k и проверьте заполненность файловых систем;
  - выполните команду lsps -s и убедитесь в том, что пространство страничной подкачки (paging space) не заполнено;
  - выполните команду emgr -l и проверьте наличие каких-либо исправлений efixes, загруженных в системе, прежде чем начинать обновление.

---

[1] Имеется в виду клонирование rootvg на резервный диск (Clone the rootvg to an Alternate Disk). *Примеч. науч. ред.*

5.3 Аспекты

Обратите внимание на множество изменений в HACMP 5.3 по сравнению с более ранними версиями. В некоторых случаях они вызывают изменения в работе кластера. Ниже перечислены аспекты, которые следует учитывать при выполнении обновления.

- Что касается HACMP 5.2, типы групп ресурсов cascading (каскадная), rotating (ротационная) и concurrent (с одновременным доступом) были преобразованы в настраиваемые группы ресурсов с соответствующими политиками запуска, перемещения при сбое и возврата после восстановления. При миграции с версий до HACMP 5.2 выполняется преобразование групп ресурсов таким образом, чтобы каждая группа ресурсов использовала соответствующие политики. Подробное описание политик в настраиваемых группах ресурсов приведено в гл. 1.
- Настраиваемые значения в HACMP после обновления кластера будут сброшены. Это включает следующие параметры:
  - Параметры настройки сетевых модулей, такие, как скорость обнаружения отказов (failure detection rate), отсрочка (grace period) и скорость пульса (heartbeat rate). Для них устанавливаются значения по умолчанию.
  - Выполняется сброс изменений в событиях кластера, в частности скриптов пред- и постобработки событий. При сбросе этих изменений файлы и скрипты, используемые при настройке, не удаляются, однако HACMP перестает их видеть.
- Все изменения в стандартном наборе команд HACMP будут сброшены в значения по умолчанию.
- После начала циклической миграции кластер будет работать в смешанном режиме до тех пор, пока не будет выполнено обновление последнего узла и его реинтеграция в кластер. В этот период не следует пытаться применять какие-либо изменения в топологии кластера или ресурсах:
  - не выполняйте верификацию или синхронизацию кластера;
  - не выполняйте принудительное отключение узла;
  - не пытайтесь выполнять какие-либо операции DARE или C-SPOC, кроме функций управления службами HACMP (Manage HACMP Services);
  - не используйте функцию Problem Determination Tools (Инструменты определения проблем) > View Current State (Просмотр текущего состояния);
  - не используйте опцию Extended Configuration (Расширенное конфигурирование) > Snapshot Configuration (Конфигурирование снимков) > Add a Cluster Snapshot (Добавить снимок кластера) и не выполняйте команду clsnapshot;
  - не используйте опцию Problem Determination Tools (Инструменты определения проблем) > Recover from HACMP Script Failure (Восстановление после отказа скрипта HACMP) и не выполняйте команду clruncmd, кроме случаев выполнения команды или опции SMIT с целевого узла, заданного командой.

Важно! Не следует оставлять кластер в гибридном состоянии длительное время, чтобы избежать случайного вызова любой из этих операций.

- При обновлении с версий до HACMP 5.2 инсталляция HACMP создает группу hacmp на всех узлах. Классы (ODM) базы данных конфигурации HACMP (HACMP Configuration Database) были обновлены, и теперь их владельцами являются пользователь root и группа hacmp:
  - Для большинства объектных классов HACMP установлены права доступа к файлам 640. Исключение представляет класс HACMPdisksubsystem, имеющий права доступа 600
  - Все двоичные файлы HACMP, предназначенные для применения пользователями без привилегий «root», устанавливаются с правами доступа 2555. Включается бит setgid, так что программа будет выполняться в контексте группы hacmp.
  - При использовании программ, осуществляющих прямой доступ к ODM, может возникнуть необходимость их перезаписи.

Внимание! Использование информации, получаемой непосредственно из ODM, предназначено только для информационных целей, так как формат разделов (stanzas) может быть различным в разных обновлениях и/или в новых версиях. Таким образом, жесткое кодирование ODM-запросов в пользовательских приложениях не поддерживается и его следует избегать.

  - При использовании средства PSSP File Collections для обеспечения согласованности /etc/group новая группа hacmp может быть потеряна во время следующей синхронизации файлов. Чтобы этого избежать, нужно выполнить следующие действия:
    - отключить синхронизацию PSSP File Collection для /etc/group;
    - включить группу hacmp в мастер-файл (master file) /etc/group и распространить изменение на всех узлах.

Примечание. В целях безопасности не следует расширять полномочия группы hacmp.

Политики динамического приоритета узлов в версиях до HACMP 5.2 были основаны на переменных ресурсов RSCT Event Management. Если ваша конфигурация включает группу ресурсов HACMP 5.1, в которой для политики динамического приоритета узла установлено значение Fallover using Dynamic Node Priority (Перемещение при сбое с использованием динамического приоритета узла), при миграции это значение будет сброшено. Утилита **cl_convert** изменит политику перемещения при сбое для этой группы ресурсов, установив для нее значение Fallover to Next Priority Node in the list (Перемещение при сбое на узел из списка со следующим приоритетом):

- во время миграции используется политика перемещения при сбое по умолчанию;
- обратите внимание на то, что HACMP 5.3 поддерживает только три политики:
  - cl_highest_free_mem;
  - cl_highest_idle_cpu;
  - cl_lowest_disk_busy.

- В HACMP 5.3 поддерживается только одна политика распределения групп ресурсов – политика распределения на основе узлов (node-based distribution policy). Ранее доступная политика распределения на основе сетей (network-based distribution policy) была удалена. Во время миграции она преобразуется в политику на основе узлов.
- События кластера, определенные пользователем, могут не работать, так как `emsvcs` больше не применяются в HACMP. Для исправления проблем следует перезаписать их с использованием `rmcd`.
- Диспетчер блокировки кластера (Cluster Lock Manager, cllockd или cllockdES) более недоступен в HACMP 5.3. Во время установки происходит удаление файлов и определений Lock Manager на узле. Проконсультируйтесь с производителем приложения относительно поддержки одновременного доступа.
- Обратите внимание на то, что режим расширенного одновременного доступа поддерживается только в AIX 5L V5.1 и более поздних версиях. Режим одновременного доступа SSA (SSA concurrent mode) не поддерживается на 64-разрядном ядре. При использовании SSA-дисков в режиме одновременного доступа нельзя запустить 64-разрядное ядро до преобразования всех групп томов в режим расширенного одновременного доступа.

5.4 Основные этапы миграции

Циклическая миграция

Метод циклической миграции используется в тех случаях, когда требуется обеспечить доступность приложения в процессе миграции. Этот тип миграции состоит из следующих этапов (рис. 5.1):

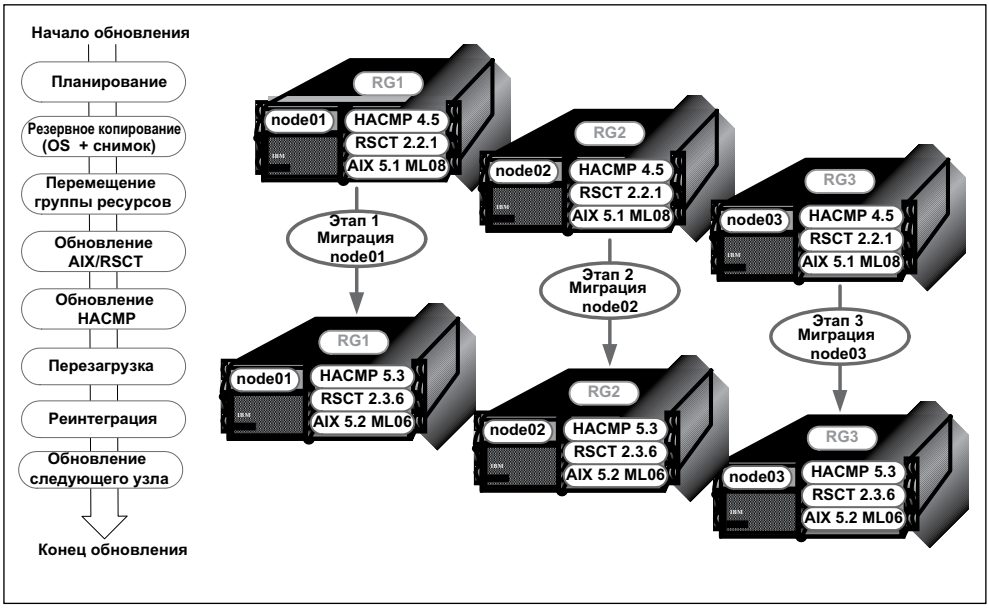

Рис. 5.1. Основные этапы циклической миграции с HA 4.5 до HA 5.3

1. Остановка служб кластера на первом узле с передачей ресурсов на резервный узел.

2. Обновление программного обеспечения AIX и RSCT (если необходимо).

3. Обновление программного обеспечения HACMP (включая последние PTF).

4. Перезагрузка.

5. Реинтеграция узла в кластер и повторное выполнение операций на следующем узле.

На рис. 5.1 представлен пример обновления с кластера HACMP 4.5 под управлением AIX 5.1 до кластера HACMP 5.3 под управлением AIX 5.2.

Миграция с использованием снимков

Также существует возможность миграции кластера с использованием метода снимков. Использование этого метода предполагает применение снимка, созданного до выполнения обновления, для восстановления конфигурации. После удаления и повторной установки всех узлов выполняется преобразование и применение снимка. Однако в процессе миграции целый кластер будет недоступен, и все узлы должны быть обновлены до повторной активизации кластера.

Этот тип миграции состоит из следующих этапов:

1. Остановка служб кластера на всех узлах.

2. Обновление программного обеспечения AIX и RSCT на всех узлах (если необходимо).

3. Деинсталляция текущей версии HACMP и установка HACMP 5.3 на всех узлах (включая последние PTF, если они доступны).

4. Перезагрузка.

5. Преобразование снимка.

Замечание. Если вы планируете использовать метод снимков, помните о том, что при попытке повторного применения снимка на узлах в первую очередь HACMP попытается использовать путь для связи, заданный в объектном классе HACMPnode.

```
#odmget HACMPnode
HACMPnode:
name = “cobra”
object = “COMMUNICATION_PATH”
value = “10.10.32.33”
node_id = 2
node_handle = 2
version = 8
```

Это значение изначально устанавливается при первом определении узлов в кластере с указанием COMMUNICATION PATH. Рекомендуется всегда устанавливать этот путь в качестве постоянного IP-адреса. Если по какой-либо причине этот IP-адрес недоступен на момент применения снимка, следует вручную установить этот синоним на одном из интерфейсов.

Последним файлом, в котором выполняется поиск пути для связи, является файл /usr/es/sbin/cluster/etc/rhosts, в котором при необходимости можно вручную задать все IP-адреса кластера. IP-адреса будут включать: базовые, сервисные и постоянные IP-адреса.

6. Применение снимка (при этом конфигурация будет передана на все узлы).

7. Запуск служб кластера поочередно на каждом узле.

8. Верификация и синхронизация кластера.

5.5 Протестированные сценарии

В процессе написание этой книги мы протестировали выполнение миграции на HACMP V5.3 с трех разных версий:

- HAES 4.5;
- HAES 5.1;
- HAES 5.2.

Для каждой из этих версий мы выполняли циклическую миграцию и миграцию с использованием снимков, выполняя документирование действий и результатов. Обратите внимание на то, что при переходе с HAS на HAES мы решили не рассматривать способ поузловой миграции. При подготовке к обновлению с HAS 4.5 мы рекомендуем по возможности использовать метод преобразования снимка.

Мы выполнили все инсталляции и обновления с применением NIM-сервера с наиболее актуальными PTF для HACMP, AIX, RSCT и SDD.

5.5.1 Сценарий 1: AIX 5.1 и HAES 4.5

В качестве первого теста миграции мы решили выполнить обновление трехузлового кластера AIX 5.1 / HAES 4.5 до AIX 5.2 / HA 5.3. Мы использовали серверы p630 (7028-6C4) в своей конфигурации кластера. На рис. 5.2 представлена диаграмма конфигурации, используемой в нашей тестовой среде.

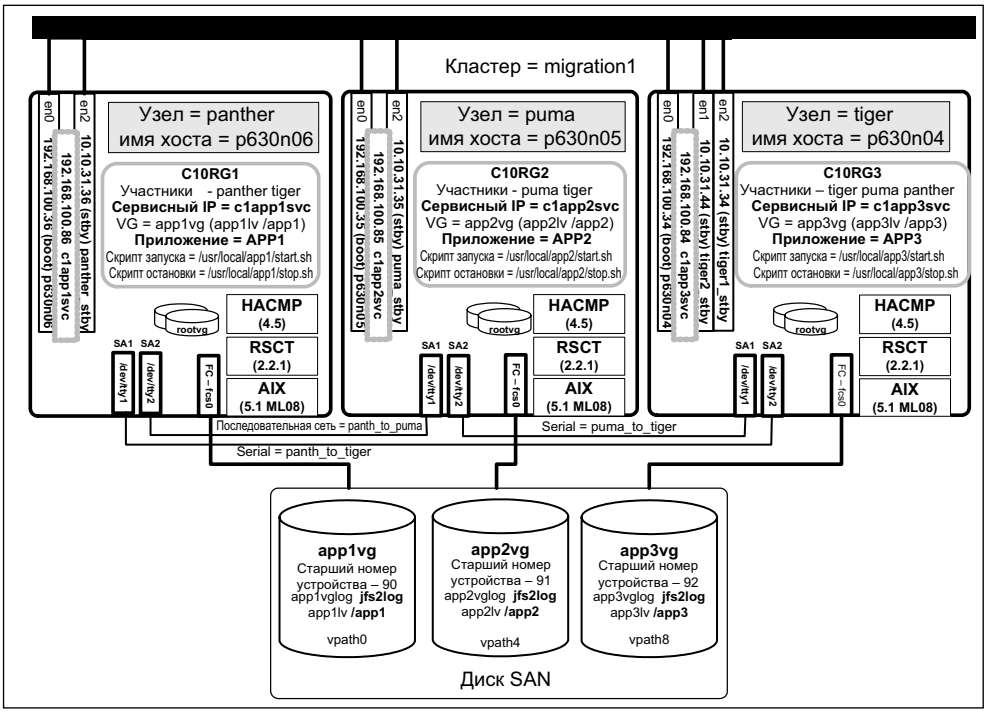

Рис. 5.2. Сценарий 1: среда миграции с HA 4.5 на HA 5.3

В нашем кластере каждый узел осуществлял управление своей собственной группой ресурсов. Каждая группа ресурсов содержала группу томов приложения, сервисный IP-адрес и сервер приложения. Топология была настроена на использование IPAT посредством замены, с целью протестировать переход на HACMP 5.3. Мы выполнили конфигурирование сетей RS232, использующих интегрированные порты на узлах для трафика пульса в сетях, отличных от IP. Хранилище содержало ESS LUN (2105-800) с подключением к коммутатору, а также скрипт подключения хоста (ibm2105.rte) и драйвер SDD.

Этапы циклической миграции: сценарий 1

Мы начали воплощать свой сценарий после того, как выполнили тестирование и убедились в том, что наш кластер из трех узлов находится в стабильном состоянии и может успешно выполнять перемещение при сбое. Мы решили начать обновление с узла **panther**.

Мы выполнили следующие предварительные действия:

1. Сохранение снимка с рабочего кластера (со всеми активными узлами).

> **Внимание!** Не сохраняйте снимок в каталог /tmp, так как при установке AIX содержимое каталога /tmp удаляется. Сохраните снимок в другой каталог или на другой сервер.
> Не забудьте также сохранить скрипты запуска/остановки в надежный каталог.

2. Создание резервной копии (mksysb).

3. Создание alt_disk_install[1] как средства возврата к прежней конфигурации. Это было сделано с целью ускорения повторного тестирования.

Сценарий миграции

Ниже перечислены действия, выполняемые нами в нашем сценарии миграции.

1. Остановка HACMP на узле **panther** (постепенная остановка с передачей ресурсов на резервный узел – graceful with takeover):

– перемещение C10RG1 на узел tiger;

– проверка с использованием /usr/es/sbin/cluster/utilities/clfindres.

2. Инициация миграции AIX5.2/RSCT:

– применение последних исправлений AIX 5.2 (ML06);

– обновление и верификация уровней RSCT (rsct.basic.rte 2.3.6.2);

– удаление и замена SDD текущим драйвером AIX 5.2;

  - **stopsrc -s sddsrv;**
  - **rmdev -dl dpo -R;**

- деинсталляция SDD 5.1 с использованием smitty remove (devices.sdd.51.rte 1.6.0.2);
- установка SDD 5.2 (devices.sdd.52.rte 1.6.0.0 и 1.6.0.2 PTF).

3. Выполнение smit update_all для обновления наборов файлов HACMP до версии 5.3.

---

[1] Имеется в виду так называемое «клонирование» системы. *Примеч. науч. ред.*

4. Перезагрузка узла **panther**:

– верификация инсталляции AIX (lppchk -l / -c / -v, instfix, oslevel, errpt);
– lslpp -l | grep cluster => Нет следов версии 4.5 (проверка отсутствия наборов файлов);
– odmget HACMPcluster => все еще отображалась версия 4.5 [5].

5. Запуск hacmp на узле **panther**.

– в нашей конфигурации группа ресурсов C10RG1 была настроена на перемещение при сбое на узел с более высоким приоритетом при реинтеграции, вследствие чего группа ресурсов возвратилась на узел panther.
– команда cldump отобразила корректное размещение и состояние всех ресурсов.

Внимание! Выполнение синхронизации для кластера в гибридном состоянии нарушит миграцию и оставит кластер в несогласованном состоянии.

6. Повторное выполнение действий для узла **puma.**

7. Повторное выполнение действий для узла **tiger.**

После входа последнего узла в кластер выполняется преобразование ODM, после чего в объектных классах HACMPcluster и HACMPnode отображается версия cluster_version = 8.

Результаты циклической миграции: сценарий 1

Миграция AIX представляла самый медленный этап обновления. Реинтеграция последнего узла кластера была успешной, и мы смогли убедиться в том, что различные разделы HACMP ODM были преобразованы корректно и что все узлы теперь выводили версию cluster_version = 8. Способы проверки номеров версий HACMP см. в табл. 5.2.

После завершения мы проверили состояние кластера с использованием утилиты clstat, а также состояние узлов по выходным данным команды **lssrc -ls clstrmgrES.**

#lssrc -ls clstrmgrES

Current state: ST_STABLE

После завершения мы использовали инструмент тестирования кластера (Cluster Test Tool) для запуска различных тестов и не обнаружили каких-либо проблем. Мы снова выполнили тестирование миграции и не обнаружили проблем с функционированием кластера.

Миграция с использованием снимков: сценарий 1

В том же кластере из трех узлов мы проверили работу метода преобразования снимков. Мы использовали образы установки на резервных дисках для возврата к прежней среде, после чего перезапустили службы кластера на всех узлах, на которых выполнялся HA 4.5. Тестирование содержало следующие действия:

1. Остановка HACMP на всех узлах: panther, puma, tiger.

2. Выполнение команды smit remove и деинсталляция всех наборов файлов cluster.* со всех узлов.

3. Миграция кода AIX/RSCT с использованием NIM.

4. Установка пакетов HACMP с текущими уровнями PTF на всех узлах.

5. Перезагрузка всех узлов кластера.

6. Копирование файла snapshot.odm, предварительно сохраненного в процессе циклической миграции и имеющего расположение /usr/es/sbin/cluster/utilities/snapshot.odm.

7. Выполнение следующей команды для преобразования снимка: #/usr/es/sbin/cluster/conversion/clconvert_snapshot -v 4.5 -s snapshot.odm

8. Применение снимка: smit hacmp > Extended Configuration (Расширенное конфигурирование) > Snapshot Configuration (Конфигурирование снимка) > Apply a Cluster Snapshot (Применение снимка кластера) > выбор снимка и нажатие Enter.

9. Запуск служб кластера поочередно на каждом узле.

Результаты миграции с использованием снимков: сценарий 1

В целом, выполнение миграции с использованием снимков произошло очень быстро. Несмотря на то, что потребовалось полное отключение кластера, преобразование файла снимка заняло не больше минуты, и преобразованный снимок был очень быстро применен на всех трех узлах.

> **Замечание.** Если вы планируете использовать метод снимков, помните о том, что при попытке повторного применения снимка на узлах в первую очередь HACMP попытается использовать путь для связи, заданный в объектном классе HACMPnode.
> #odmget HACMPnode
> HACMPnode:
> name = “cobra”
> object = “COMMUNICATION_PATH”
> value = “10.10.32.33”
> node_id = 2
> node_handle = 2
> version = 8
> Это значение изначально устанавливается при первом определении узлов в кластере с указанием COMMUNICATION PATH. Рекомендуется всегда устанавливать этот путь в качестве постоянного IP-адреса. Если по какой-либо причине этот IP-адрес недоступен на момент применения снимка, следует вручную установить этот синоним на одном из интерфейсов.
> Последним файлом, в котором выполняется поиск пути для связи, является файл /usr/es/sbin/cluster/etc/rhosts, в котором при необходимости можно вручную задать все IP-адреса кластера. IP-адреса будут включать: базовые, сервисные и постоянные IP-адреса.

После завершения миграции мы убедились в корректности преобразования разделов HACMP ODM путем проверки версии кластера в объектных классах `HACMPcluster` и `HACMPnode`. Также мы использовали инструмент тестирования кластера (Cluster Test Tool) после завершения миграции и не обнаружили каких-либо проблем.

5.5.2 Сценарий 2: AIX 5.2 и HA 5.1

Второй тестовый сценарий представлял собой миграцию двухузлового кластера AIX 5.2 / HA 5.1, настроенного на взаимный перехват ресурсов, на AIX 5.3 / HA 5.3. Мы использовали серверы p630 (7028-6C4) в конфигурации кластера. На рис. 5.3 представлена схема используемой конфигурации.

Топология была настроена на использование стандартных IP-синонимов. На каждом узле были сконфигурированы постоянные IP-адреса в той же подсети, в которой находятся и сервисные адреса. Хранилище содержало ESS LUN (2105-800) с подключением к коммутатору, а также скрипт подключения хоста (ibm2105.rte) и драйвер SDD. Каждый узел был настроен на использование собственных групп ресурсов, каждая из которых включает свою группу томов, сервисный IP-адрес и сервер приложения. Группы томов были сконфигурированы как группы томов с расширенным одновременным доступом, чтобы установить сеть пульса через диски, отличную от IP.

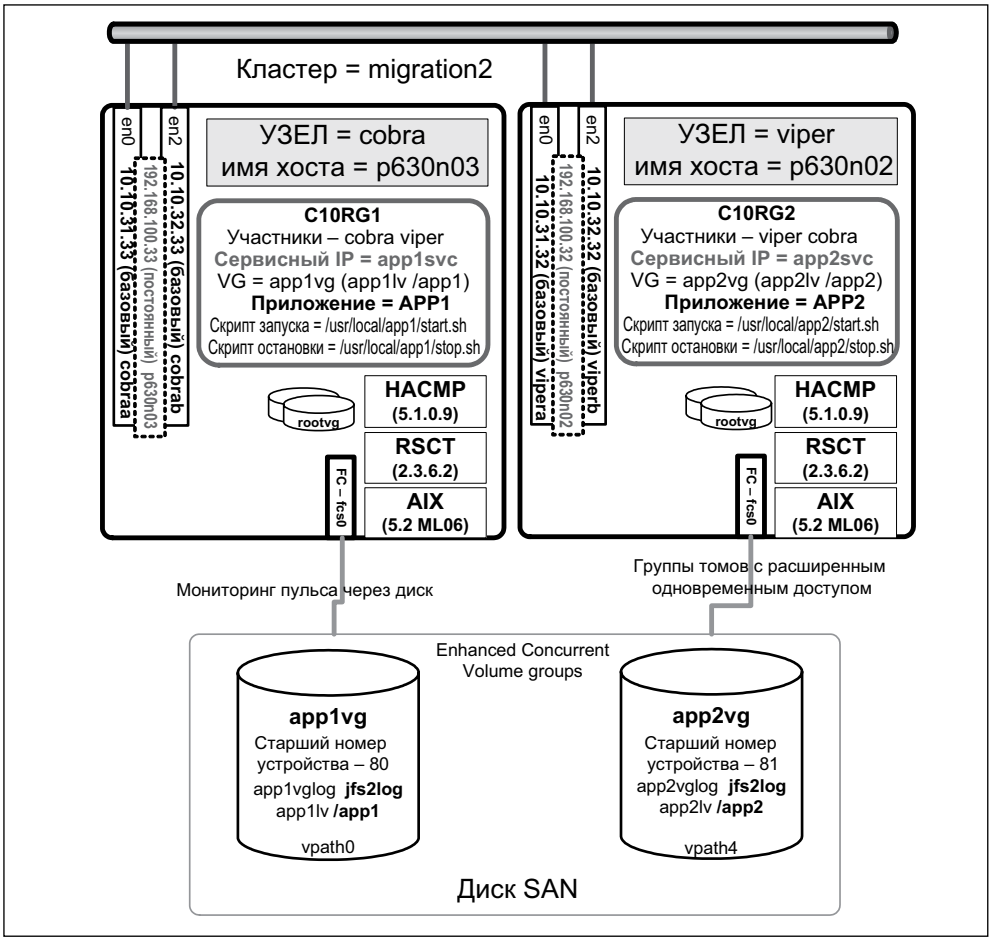

Рис. 5.3. Сценарий 2: среда миграции с HA 5.1 на HA 5.3

Циклическая миграция: сценарий 2

Мы протестировали стабильность кластера и выполнение перемещения при сбое, прежде чем перейти к тестам миграции. Первым протестированным нами методом миграции для этого двухузлового кластера была циклическая миграция.

Мы выполнили следующие предварительные действия:

1. Сохранение снимка с рабочего кластера (со всеми активными узлами).

> **Внимание!** Не сохраняйте снимок в каталог /tmp, так как при установке AIX содержимое каталога /tmp удаляется. Сохраните снимок в другой каталог или на другой сервер. Не забудьте также сохранить скрипты приложения в надежный каталог.

2. Создание резервной копии (mksysb).

3. Создание alt_disk_install как средства возврата к прежней конфигурации. Это было сделано с целью ускорения повторного тестирования.

Сценарий миграции

Сначала мы решили выполнить обновление узла **cobra**. Для этого мы выполнили следующие действия:

1. Остановка HACMP на узле **cobra** (постепенная остановка с передачей ресурсов на резервный узел – graceful with takeover):

– перемещение C10RG1 на узел viper;

– проверка с использованием /usr/es/sbin/cluster/utilities/clRGinfo.

2. Инициация миграции AIX 5.3/RSCT с использованием кода NIM вместе с последними исправлениями: smit remove sdd – должно быть выполнено удаление и замена:

– stopsrc -s sddsrv;

– rmdev -dl dpo -R;

– деинсталляция драйвера SDD 5.2 с использованием smitty remove (devices.sdd.52.rte 1.6.0.2);

– установка драйвера SDD 5.3 (devices.sdd.53.rte 1.6.0.0 и 1.6.0.2).

3. Выполнение smit update_all для загрузки наборов файлов HA 5.3.

4. Перезагрузка узла **cobra**.

5. Реинтеграция узла **cobra** в кластер путем запуска служб кластера. Политика ресурсов была настроена на перемещение при сбое на узел с более высоким приоритетом, вследствие чего группа ресурсов `C10RG1` возвратилась на узел `cobra`.

> **Внимание.** На этом этапе ODM-классы HACMP на узле `cobra` не были преобразованы и не соответствуют новой версии, тогда как наборы файлов соответствуют версии HA 5.3. На данном этапе кластер находится в смешанном режиме или, так называемом гибридном состоянии (hybrid state).

6. Остановка HACMP на узле **viper** (постепенная остановка с передачей ресурсов на резервный узел – graceful with takeover). Группа ресурсов C10RG2 перемещается на узел cobra.

7. Повтор операций установки AIX/RSCT и HACMP на узле viper.

8. Перезагрузка узла viper.

9. Реинтеграция узла viper в кластер путем запуска служб кластера.

Результаты циклической миграции: сценарий 2

Общее тестирование циклической миграции было успешным. Как и в сценарии 1, миграция AIX представляла самый медленный этап обновления. После ее выполнения мы протестировали перемещение при сбое в кластере и не обнаружили проблем. Реинтеграция последнего узла была успешной, и мы убедились в том, что изменения в ODM соответствовали новой версии HACMP.

В целях тестирования мы решили прервать миграцию, когда узлы были в смешанном режиме. Мы остановили узел **cobra**, когда на нем выполнялся HACMP 5.3 и он содержал обе группы ресурсов. В это время на узле **viper** выполнялся процесс установки HACMP 5.3. После перезагрузки узла **cobra** и перезапуска служб кластера, он подхватил свою группу ресурсов (C10RG1). После этого мы вручную подключили группу ресурсов узла viper:

#smit hacmp > System Management (C-SPOC) > HACMP Resource Group and Application Management (Управление группами ресурсов и приложениями HACMP) > Bring a Resource Group online (Подключение группы ресурсов) > выбор C10RG2 и нажатие Enter.

Группа ресурсов была подключена без каких-либо проблем. После этого мы выполнили обновление AIX и HACMP на втором узле, после чего смогли выполнить успешную интеграцию в кластер без каких-либо проблем.

После завершения мы использовали инструмент тестирования кластера (Cluster Test Tool) для запуска различных тестов и не обнаружили каких-либо проблем. Мы снова выполнили тестирование миграции и не обнаружили проблем с функционированием кластера.

Миграция с использованием снимков: сценарий 2

В том же кластере из двух узлов мы проверили работу метода преобразования снимков. Мы использовали образы установки на резервных дисках для возврата к прежней среде, после чего перезапустили службы кластера на обоих узлах, на которых выполнялся HA 5.1. Тестирование содержало следующие действия:

1. Остановка HACMP на всех узлах: cobra viper

2. Выполнение команды smit remove и деинсталляция всех наборов файлов cluster.* со всех узлов.

3. Миграция кода AIX/RSCT с использованием NIM.

4. Установка пакетов HACMP с текущими уровнями PTF на всех узлах.

5. Перезагрузка всех узлов кластера.

6. Копирование файла snapshot.odm, предварительно сохраненного в процессе циклической миграции и имеющего расположение /usr/es/sbin/cluster/utilities/snapshot.odm.

7. Выполнение следующей команды для преобразования снимка: #/usr/es/sbin/cluster/conversion/clconvert_snapshot -v 5.1 -s snapshot.odm.

8. Применение снимка: smit hacmp > Extended Configuration (Расширенное конфигурирование) > Snapshot Configuration (Конфигурирование снимка) > Apply a Cluster Snapshot (Применение снимка кластера) > выбор снимка и нажатие Enter.

9. Запуск служб кластера поочередно на каждом узле.

Результаты миграции с использованием снимков: сценарий 2

В целом миграция с использованием снимка также прошла успешно. Преобразование файла снимка заняло несколько секунд, и его применение на обоих узлах также произошло очень быстро. После завершения миграции мы убедились в корректности преобразования разделов HACMP ODM путем проверки версии кластера в объектных классах HACMPcluster и HACMPnode.

Также после завершения миграции мы использовали инструмент тестирования кластера (Cluster Test Tool) для выполнения нескольких тестов. После обновления не было обнаружено каких-либо проблем функционирования кластера.

Мы считаем, что, если ваша среда допускает перерыв в обслуживании в масштабе кластера, можно использовать этот быстрый и надежный метод, позволяющий полностью избежать функционирования узлов в смешанном режиме. Использование снимка для возврата конфигурации HACMP после обновления всех узлов позволяет избежать потенциальных проблем, которые могут возникнуть в процессе циклической миграции. Дополнительные сведения о потенциальных проблемах см. в разделе «Аспекты», в этой главе.

5.5.3 Сценарий 3: AIX 5.2 и HA 5.2

Третий тестовый сценарий представлял собой миграцию кластера AIX 5.2 / HA 5.2 с двумя активными узлами и одним дежурным резервным узлом на AIX 5.3 / HA 5.3. В своей конфигурации мы использовали три одинаковых LPAR p690 (7040-681). На рис. 5.4 представлена схема используемой конфигурации.

Топология была настроена на использование IP-синонимов вместе с мониторингом пульса через синонимы. Это было сделано, чтобы убедиться в том, что обновление сохранит смещение сети 10.10.1.1 и соответствующие синонимы и что впоследствии проблем в работе кластера не возникнет.

> **Примечание.** Помните, что при использовании мониторинга пульса через синонимы в вашей сетевой топологии мониторинг базовых адресов в HACMP не выполняется.

Мы сконфигурировали базовые адреса для каждого узла в соответствующей подсети и поместили все сервисные и постоянные IP-адреса в другой диапазон подсети. Такая конфигурация IP используется для имитации университетской среды, где компьютеры могут быть разделены определенным расстоянием и могут использовать различные подсети, а также для того, чтобы убедиться в том, что для базовых адресов не выполняется мониторинг с использованием трафика пульса HACMP.

Хранилище содержало ESS LUN (2105-800) с подключением к коммутатору, а также скрипт подключения хоста и драйвер SDD для балансировки нагрузки и многопутевой конфигурации (multipathing). Мы использовали группы томов с расширенным одновременным доступом для создания трех сетей пульса через диски, отличных от IP.

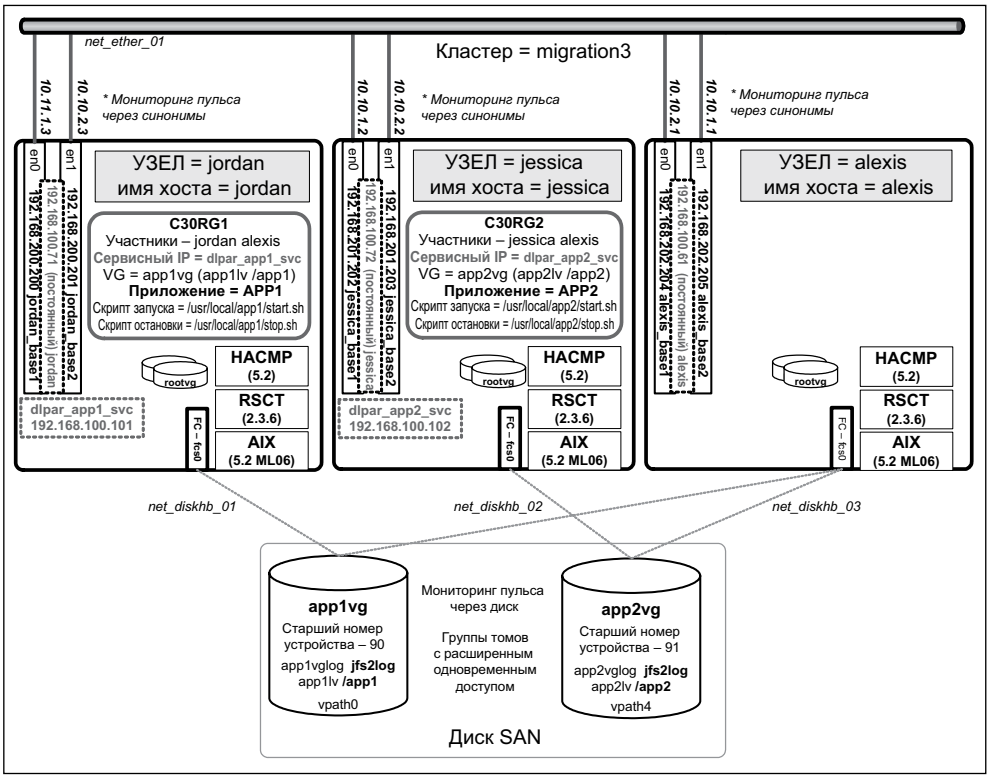

Рис. 5.4. Сценарий 3: среда миграции с HA 5.2 на HA 5.3

Циклическая миграция: сценарий 3

Для этого сценария мы использовали кластер в среде DLPAR, чтобы убедиться в том, что все работает после обновления. Мы выполнили следующие предварительные действия:

1. Сохранение снимка с рабочего кластера (со всеми активными узлами).

2. Создание резервной копии (mksysb).

3. Создание alt_disk_install как средства возврата к прежней конфигурации. Это было сделано с целью ускорения повторного тестирования.

Сценарий миграции

Сначала мы решили выполнить обновление дежурного узла **`alexis`**, чтобы максимально сократить перерыв в работе узлов, содержащих группы ресурсов. Для этого мы выполнили следующие действия:

1. Остановка HACMP на узле **`alexis`** (постепенная остановка с передачей ресурсов на резервный узел – graceful with takeover).

Внимание. Не сохраняйте снимок в каталог /tmp, так как при миграции AIX содержимое каталога /tmp удаляется. Сохраните снимок в другой каталог или на другой сервер. Не забудьте также сохранить скрипты приложения в надежный каталог.

2. Инициация установки базового кода AIX 5.3/ RSCT с использованием NIM вместе с последними исправлениями. Удаление и замена SDD текущим драйвером AIX 5.3:

– **stopsrc -s sddsrv**;

– **rmdev -dl dpo -R**;

– деинсталляция драйвера SDD 5.2 с использованием smitty remove (devices.sdd.52.rte 1.6.0.2);

– установка драйвера SDD 5.3 (devices.sdd.53.rte 1.6.0.0 и 1.6.0.2 PTF).

3. Выполнение smit update_all для загрузки наборов файлов HACMP 5.3.

4. Перезагрузка узла **alexis**:

– Верификация инсталляции AIX (lppchk -l / -c / -v, instfix, oslevel, errpt);

– lslpp -l | grep cluster => Нет следов версии 5.2 (проверка отсутствия наборов файлов);

– odmget HACMPcluster => все еще отображалась версия 7.

5. Реинтеграция узла **alexis** в кластер путем запуска служб кластера.

Внимание. Выполнение синхронизации для кластера в гибридном состоянии нарушит миграцию.

6. Остановка HACMP на узле **jordan** (постепенная остановка с передачей ресурсов на резервный узел – graceful with takeover).

7. Повтор действий пп. 2–5 на узле **jordan**.

8. Остановка HACMP на узле **jessica** (постепенная остановка с передачей ресурсов на резервный узел – graceful with takeover).

9. Повтор действий пп. 2–5 на узле **jessica**.

Результаты циклической миграции: сценарий 3

Как и в двух предыдущих сценариях, мы не столкнулись с какими-либо проблемами при циклической миграции HACMP. После интеграции последнего узла в кластер было выполнено преобразование разделов ODM-классов, и теперь для всех узлов выводилось cluster_version = 8, как и ожидалось. Номера версий HACMP см. в табл. 5.2.

При первоначальной настройке топологии кластера мы выполнили конфигурирование базового, постоянного и сервисного IP-адресов в одной подсети. Хотя такая топология и допустима, она создает некоторые проблемы в среде NIM (Network Install Manager), так как NIM неспособен обрабатывать постоянные синонимы (что вызывает отказы при подключении NFS).

Поэтому мы советуем быть осторожными при внедрении HACMP в среде NIM; это также относится к среде CSM (Cluster Systems Management), которая использу-

ет NIM для установки и обновления программного обеспечения на управляемых узлах. (пример 5.1 представляет конфигурацию в файле hosts (которая создает проблемы с NIM):

Пример 5.1. Разрешение имен в нашей среде

```
_:>># more /etc/hosts
192.168.100.200  jordan_base1
192.168.100.201  jordan_base2
192.168.100.202  jessica_base1
192.168.100.203  jessica_base2
192.168.100.204  alexis_base1
192.168.100.205  alexis_base2
192.168.100.71   p690_1_lpar1   #persistent IP jordan
192.168.100.72   p690_1_lpar2   #persistent IP jessica
192.168.100.61   p690_2_lpar1   #persistent IP alexis
192.168.100.101  dlpar_app1_svc
192.168.100.102  dlpar_app1_svc
```

Примечание. Конфигурация, представленная в примере 5.1, основана на маске сети класса C (255.255.255.0). В качестве альтернативы мы изменили нашу топологию на топологию, представленную на рис. 5.4, и сконфигурировали наши базовые адреса в отдельной подсети от сервисных и постоянных IP-адресов.

Миграция с использованием снимков: сценарий 3

В том же трехузловом кластере мы проверили работу метода преобразования снимков. Мы использовали образы установки на резервных дисках для возврата к прежней среде, после чего перезапустили службы кластера на обоих узлах, на которых выполнялся HACMP 5.2. Тестирование содержало следующие действия:

1. Остановка HACMP на всех узлах: **alexis jordan jessica .**

2. Выполнение команды smit remove и деинсталляция всех наборов файлов cluster.* со всех узлов.

3. Миграция кода AIX/RSCT с использованием NIM.

4. Установка пакетов HACMP с текущими уровнями PTF на всех узлах.

5. Перезагрузка всех узлов кластера.

6. Копирование файла snapshot.odm, предварительно сохраненного в процессе циклической миграции и имеющего расположение /usr/es/sbin/cluster/utilities/snapshot.odm

7. Выполнение следующей команды для преобразования снимка: #/usr/es/sbin/cluster/conversion/clconvert_snapshot -v 5.2 -s snapshot.odm.

8. Применение снимка: smit hacmp > Extended Configuration (Расширенное конфигурирование) > Snapshot Configuration (Конфигурирование снимка) > Apply a Cluster Snapshot (Применение снимка кластера) > выбор снимка и нажатие Enter.

9. Запуск служб кластера поочередно на каждом узле.

Результаты миграции с использованием снимков: сценарий 3

В целом миграция с использованием снимка также прошла успешно. Преобразование файла снимка заняло около 1 мин, и его применение на всех трех узлах также произошло очень быстро. Затем мы убедились в корректности преобразования разделов ODM.

Мы считаем, что, если ваша среда допускает перерыв в обслуживании в масштабе кластера, можно использовать этот быстрый и надежный метод, позволяющий полностью избежать функционирования узлов в смешанном режиме. Использование снимка для возврата конфигурации HACMP после обновления всех узлов позволяет избежать потенциальных проблем, которые могут возникнуть в процессе циклической миграции. Дополнительные сведения о потенциальных проблемах см. в разделе «Аспекты», в этой главе.

5.6 Действия после миграции

Ниже перечислены действия, которые рекомендуется выполнить после осуществления миграции.

- Проверьте наличие наборов файлов HACMP, оставшихся от прежней версии. Могли остаться старые наборы файлов документации, которые, хотя и не создают каких-либо проблем, также должны быть обновлены.
- После выполнения любого типа миграции следует осуществить верификацию и синхронизацию конфигурации кластера.
- После любой миграции следует протестировать выполнение перемещения при сбое и восстановления. Инструмент Cluster Test Tool позволяет выполнять несколько тестов и может быть настроен на использование дополнительных тестов.
- Обратите внимание на то, что после версии 4.5 файл ~/.rhosts больше не нужен (и не используется) в HACMP. Если он не нужен в вашей среде, не забудьте его удалить.

Совет после миграции

При выполнении миграций AIX/HACMP мы использовали NIM-сервер. Мы применяли постоянные IP-адреса в определениях компьютеров NIM-клиентов. В результате мы обнаружили, что при установке в качестве базового адреса для одного из интерфейсов была установлена постоянная IP-метка, тогда как базовый адрес был удален. Кроме того, в качестве имени хоста было установлено имя компьютера, заданное в определении NIM-клиента.

Мы исправили изменения, выполнив следующие действия:

1. smitty chinet => жесткая установка прежнего базового адреса интерфейса.

2. Мы закомментировали строки, добавленные в файл /etc/rc.net:

/bin/hostname <hostname>

/usr/sbin/ifconfig <en#> inet <IP> netmask 255.255.255.0

3. hostname <name> => установка прежнего имени хоста.

При использовании NIM следует проверить, были ли выполнены эти изменения; в противном случае вы получите неоднозначные результаты в HACMP после запуска служб кластера.

Этой проблемы можно избежать путем настройки NIM (это предполагает создание собственного скрипта настройки).

5.7 Устранение неполадок при отказе миграции

Миграция кластера является неотъемлемой частью администрирования кластера. Для успешного выполнения обновления следует тщательно спланировать и пересмотреть способ миграции. Хотя мы не обнаружили каких-либо проблем при тестировании миграции, мы понимаем, что в случае возникновения проблем было бы полезно обратиться к документированным действиям, описывающим способ возврата в стабильное состояние. Данный раздел был написан с этой целью.

5.7.1 Возврат после отказа миграции

Выполнение действий, перечисленных в контрольном списке действий перед миграцией, является наилучшим способом избежать проблем при миграции. Однако в случае возникновения проблем, при которых не выполняется интегрирование узлов в кластер или при которых узлы приостанавливают интеграцию, нужно выполнить отмену этапов миграции и следовать приведенным ниже рекомендациям для возврата к старой конфигурации.

> **Внимание.** В зависимости от этапа миграции, на котором возникли проблемы, может не потребоваться выполнять восстановление кластера целиком.

Если выполняемая миграция кластера HACMP включает обновление AIX и RSCT, после реинтеграции первого обновленного узла ваш кластер будет работать в смешанном или гибридном режиме. При неудачной попытке интеграции этого узла или оставшихся узлов в кластер коррекция миграции кластера может быть очень сложной и вам придется возвратиться к старой конфигурации.

На этом этапе у вас есть следующие варианты:

Вариант 1 – включить узел и попытаться найти и устранить проблему.

Вариант 2 – возвратиться к старой конфигурации.

Вариант 3 – деинсталлировать старое программное обеспечение HACMP, установить новый код и применить способ миграции с использованием снимков (при условии, что у вас есть снимок).

Вариант 1: устранение неполадок при отказе миграции

В случае остановки работы узла включите узел. При включенном компьютере и неактивном HACMP можно попытаться определить причину остановки путем просмотра журналов кластера и системы. Рекомендуется просмотреть следующие журналы и отчеты:

- Просмотрите отчет об ошибках (errpt -a| more). Запишите точное время отказа на основании записей в журнале ошибок. Проанализируйте все недавние релевантные ошибки и попытайтесь найти факты аварийного завершения демонов или создания файлов CORE.
- Просмотрите файл /usr/es/adm/cluster.log на наличие любых релевантных событий кластера. Проанализируйте журнал на наличие событий кластера, имевших место во время отказа. Не забудьте просмотреть журналы на других узлах кластера на предмет возможных различий.
- Просмотрите файл /tmp/clstrmgr.debug на наличие сообщений об аварийных остановках clstrmgrES. При аварийной остановке диспетчера кластера обычно происходит остановка работы компьютера. Большинство сообщений, содержащих время остановки, записываются в конец этого файла. Сообщения могут дать вам представление о причине отказа.
- В зависимости от содержания errpt-сообщений можно прибегнуть к анализу файлов журналов служб групп в /var/ha/log, чтобы попытаться определить проблему. Анализ записей журналов, совпадающих по времени с возникновением проблемы, может помочь в определении причины аварийного завершения демона или других зарегистрированных проблем.

Обычно глубокий анализ этих файлов журналов осуществляется опытным администратором HACMP или представителем службы поддержки программного обеспечения IBM. Помните о том, что анализ этих журналов в процессе миграции займет некоторое время и задержит процесс миграции.

> **Замечание.** При обращении в службу поддержки программного обеспечения IBM требуется выполнить сбор выходных данных команды snap -e на узле, на котором произошел отказ, для того чтобы ускорить анализ и разрешение проблемы.

Вариант 2: возврат к старой конфигурации

Если ваша миграция включает только обновление программного обеспечения HACMP и в середине процесса циклической миграции возникает отказ, самый быстрый метод возврата к прежней конфигурации состоит в деинсталляции наборов файлов HACMP на всех узлах, на которых они уже были обновлены. Затем следует переустановить на этих узлах старую версию HACMP и выполнить синхронизацию конфигурации с оставшихся активных узлов.

В том случае если вместе с обновлением программного обеспечения AIX и RSCT выполняется обновление HACMP, возврат к старой конфигурации затрудняется. Вам потребуется остановить HACMP на всех обновленных узлах и выполнить на них возврат к предыдущей версии AIX либо с использованием резервной копии системы (mksysb), либо с использованием образов установки на резервных дисках, созданных перед началом миграции. После восстановления узлов с резервного источника следует выполнить верификацию и синхронизацию с текущих активных узлов. При успешном завершении можно выполнять реинтеграцию восстановленных узлов в кластер.

После успешного восстановления первоначальной версии кода кластера следует изучить этапы миграции и убедиться в том, что были выполнены действия, описанные в разделе «Требования».

Вариант 3: деинсталляция HACMP и выполнение миграции с использованием снимков

Для использования этой опции необходимо иметь снимок конфигурации кластера на момент выполнения версии HACMP, с которой выполняется обновление. При этом действия в целом совпадают с действиями при миграции с использованием снимков, описанными в разделе «Основные этапы миграции».

1. Остановка служб кластера на всех узлах.

2. Обновление программного обеспечения AIX и RSCT на всех узлах (если необходимо).

3. Деинсталляция текущей версии HACMP и установка HACMP 5.3 на всех узлах (включая последние PTF, если они доступны).

4. Перезагрузка узлов.

5. Преобразование снимка.

6. Применение снимка.

7. Запуск служб кластера поочередно на каждом узле.

8. Верификация и синхронизация кластера.

> **Примечание.** Во всех наших тестовых сценариях миграции переход на HACMP 5.3 выполнялся автоматически без каких-либо проблем. Если у вас возникли проблемы, отличные от описанных в этом разделе, свяжитесь со службой поддержки программного обеспечения IBM для получения дальнейшей информации.

5.7.2 Информация о версии кластера в HACMP ODM

Чтобы получить информацию о версии кластера, выполните команды odmget HACMPcluster или odmget HACMPnode. Важно отметить, что после завершения миграции на HACMP 5.3 уровень версии должен быть равен восьми.

Таблица 5.2. Версия HACMP в разделах ODM

Версия HACMP	В HACMPcluster	В HACMPnode
HACMP 5.3	cluster_version = 8	version = 8
HACMP 5.2	cluster_version = 7	version = 7
HACMP 5.1	cluster_version = 6	version = 6
HACMP 4.5	cluster_version = 5	version = 5
HACMP 4.4.1	cluster_version = 4	version = 4

Если версия не была обновлена при циклической миграции после интеграции последнего узла в кластер, следует просмотреть файл `clconvert.log` на наличие записей о возникновении проблем при миграции. Изменение значений в ODM администратором или персоналом службы поддержки IBM применяется в последнюю очередь.

Внимание! Хотя это и не одобряется, но если вы все же решили вручную отредактировать значения в ODM для исправления несоответствия, следует прежде обратиться в службу поддержки программного обеспечения IBM.

5.7.3 Устранение неполадок при использовании снимков

В некоторых случаях при миграции с использованием снимков вы можете столкнуться с возникновением ошибок при верификации и отказов при попытке применения снимков. В том случае, если при применении снимка возникает ошибка, просмотрите файлы журналов на возможность ее коррекции утилитой верификации HACMP 5.3. Помните, что даже в случае отказа при попытке применения снимка, некоторые обновления конфигурации можно выполнить путем изменения ODM-классов HACMP.

Если ошибка соответствует критериям несогласованности, подлежащей исправлению функцией автокоррекции процесса верификации, можно продолжать процесс обновления с использованием опции принудительного применения снимка. После завершения следует выполнить процесс синхронизации и верификации, установив опцию Automatically Correct Errors during the Cluster Verification (Автоматическое исправление ошибок во время верификации кластера) в значение Interactively (Интерактивное).

Внимание! Использование опции принудительного применения снимка допускается только в том случае, если вы уверены в том, что обнаруженная ошибка может быть автоматически исправлена.

Могут выводиться предупреждения и ошибки следующего вида:

WARNING: «The NFS mount/Filesystem specified for resource group rg1 is using incorrect syntax for specifying an NFS cross mount: /mnt/fs1» (ПРЕДУПРЕЖДЕНИЕ: «NFS-подключение/файловая система, заданные для группы ресурсов rg1, используют некорректный синтаксис при указании перекрестного подключения NFS: /mnt/fs1»).

ERROR: «Disk Heartbeat Networks have been defined, but no Disk Heartbeat Devices. You must configure one device for each node in order for a Disk Heartbeat network to function» (ОШИБКА: «Были определены сети пульса через диски, но не были определены устройства пульса через диски. Для работы сети пульса через диски необходимо сконфигурировать по одному устройству на каждом узле»).

В этих случаях можно выполнять принудительное применение снимка для продолжения процесса миграции на HACMP 5.3. Несмотря на отказ при применении, кластер остается в согласованном состоянии. В этом случае принудительное применение снимка является безопасным.

5.7.4 Ошибка DARE при синхронизации

После выполнения миграции при попытке синхронизации может возникнуть следующее сообщение:

cldare: Migration from HACMPversion to HACMP 5.3 Detected. cldare cannot be run until migration has completed (cldare: Обнаружена миграция с HACMPversion на HACMP 5.3. cldare не может быть запущен до завершения миграции)

Необходимо сначала просмотреть файл **clconvert.log** на наличие отказов, после чего выполнить следующие действия:

1. Войти в **smit hacmp.**

2. Перейти в Problem Determination Tools (Инструменты определения проблем).

3. Выбрать Restore HACMP Configuration Database from Active Configuration (Восстановление базы данных конфигурации HACMP из активной конфигурации).

Если после этого проблема не разрешилась, проверьте наличие блокировочных файлов нулевой длины **/usr/es/sbin/cluster/.esmig** на всех узлах кластера. Эти файлы обычно автоматически удаляются при интеграции последнего узла в кластер.

> **Внимание!** При удалении этих файлов происходит удаление блокировки. Если эти файлы не были удалены в ходе стандартной процедуры, это может указывать на возникновение других ошибок; поэтому, прежде чем продолжить, советуем связаться со службой поддержки программного обеспечения IBM.

5.7.5 Ошибка «config_too_long» во время миграции

Если кластер был в рабочем состоянии до начала процесса миграции, маловероятно, чтобы он перешел в режим восстановления и выдал сообщение **config_too_long.** В том случае, если это произойдет, следует предпринять следующие действия по восстановлению HACMP.

В процессе обновления в каталоге /usr/lpp/save.config сохраняется множество файлов, включая следующие:

/usr/lpp/save.config/usr/es/sbin/cluster/events/node_up.rp

/usr/lpp/save.config/usr/es/sbin/cluster/events/node_down.rp

При обновлении с HACMP/ES 4.5 также сохраняется следующее событие:

/usr/lpp/save.config/usr/es/sbin/cluster/events/rg_move.rp

Если после интеграции последнего узла в кластер в конце миграции не происходит автоматического обновления разделов ODM, это может привести к возникновению сообщения config_too_long, так как система обработки событий кластера не сможет найти первоначальный путь к этим событиям: /usr/es/sbin/cluster/events.

После просмотра файла clconvert.log на наличие отказов миграции можно выполнить удаление фрагмента /usr/lpp/save.config из разделов. Эта операция применяется в последнюю очередь под контролем персонала службы поддержки IBM.

> **Важно!** Если эти пути не были автоматически исправлены во время миграции, это может указывать на проблемы при преобразовании других элементов. В этом случае следует обратиться в службу поддержки программного обеспечения IBM.

Часть 3

Сценарии и задачи администрирования кластера

Часть 3 описывает задачи администрирования кластера и сценарии изменения и обслуживания кластера НАСМР.

Рассматриваются следующие темы:

- сценарий: добавление двух узлов в кластер
- обслуживание кластера
- управление кластером
- безопасность кластера

6

Сценарий: добавление двух узлов в кластер

Эта глава содержит пример изменения конфигурации кластера с двухузловой на четырехузловую. В данной главе представлена пошаговая процедура добавления новых узлов и групп ресурсов с последующим тестированием кластера:

- описание первоначальной двухузловой конфигурации кластера с использованием таблиц планирования HACMP;
- новая четырехузловая конфигурация кластера с указанием дополнительных устройств и ресурсов;
- подробное описание конфигурации мониторинга пульса через диски;
- подробное пошаговое руководство по изменению конфигурации кластера в активном состоянии.

Предполагается, что вы уже прочли главы, посвященные планированию, установке и конфигурированию кластера HACMP, в частности гл. 3, «Планирование» и гл. 4, «Сценарии установки кластера». При необходимости рекомендуем обращаться к руководству по администрированию HACMP.

В этой главе нет описания всех функций HACMP; вместо этого мы сосредоточились на описании изменения конфигурации кластера в активном состоянии. Мы хотим показать вам, как можно безопасно изменять конфигурацию HACMP.

6.1 Конфигурация из двух узлов

В этом разделе описывается конфигурация первоначального кластера из двух узлов с использованием таблиц, подобных применяемым при планировании кластера HACMP.

Конфигурация кластера изображена на рис. 6.1. Основные особенности двухузлового кластера следующие:

- два узла;
- одна открытая (public) Ethernet-сеть (net_ether_01) с перехватом IP-адреса посредством синонимов;
- два базовых IP-адреса и один сервисный адрес на узел;
- одна сеть пульса через диски (net_diskhb_01) для соединения без IP;
- на каждом узле выполняется одно приложение в конфигурации взаимного перехвата;
- одна общая группа томов на узел;
- резервные подключения Fibre Channel.

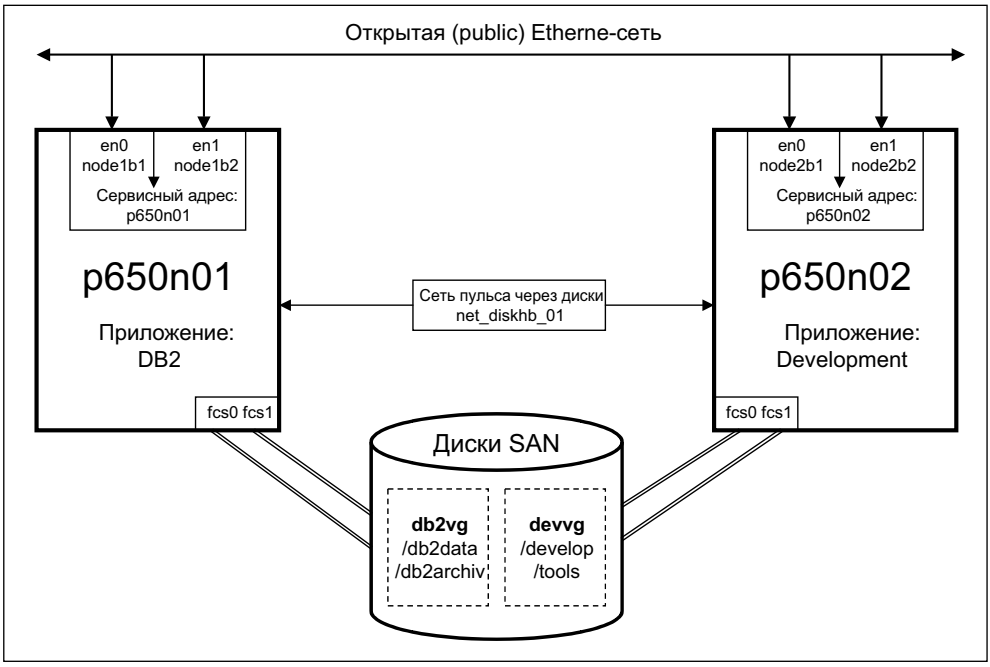

Рис. 6.1. Конфигурация из двух узлов

6.1.1 Топология кластера

Сведения о конфигурации топологии кластера:

Имя кластера: clto24.

Имя узла 1: p650n01.

Имя узла 2: p650n02.

В кластере используется одна IP-сеть и одна сеть пульса через диски, отличная от IP. Также применяется перехват IP-адреса посредством синонимов. Конфигурация сети представлена в табл. 6.1.

Таблица 6.1. Конфигурация сетей

Имя сети	Тип сети	Маска сети	IPAT посредством синонимов
net_ether_01	ethernet (открытая)	255.255.255.0	Включен
net_diskhb_01	Сеть пульса через диски	Неприменимо	Неприменимо

В табл. 6.2 представлена конфигурация сетевых интерфейсов. Каждый узел содержит два сетевых адаптера, и каждый интерфейс содержит один базовый IP-адрес. Кроме того, имеется два сервисных адреса, по одному для каждой группы ресурсов.

Таблица 6.2. Сетевые интерфейсы

IP-метка	Сетевой интерфейс	Функция интерфейса	IP-адрес
node1b1	en0	Базовый на p650n01	192.168.100.51
node1b2	en1	Базовый на p650n01	192.168.145.51
p650n01[1]		Сервисный	192.168.147.51
node2b1	en0	Базовый на p650n02	192.168.100.52
node2b2	en1	Базовый на p650n02	192.168.145.52
p650n02		Сервисный	192.168.147.52

Кластер имеет одну сеть пульса через диски, которая включает одну группу томов с возможностью расширенного одновременного доступа (enhanced concurrent capable volume group). Этот диск используется только для мониторинга пульса. (см. табл. 6.3).

Таблица 6.3. Конфигурация сети пульса через диски

Имя сети	Первый узел	Второй узел	Имя устройства	Имя группы томов
net_diskhb_01	p650n01	p650n02	vpath4	c12vg

> **Примечание.** Для согласованности мы используем одинаковую нумерацию устройств vpath на всех узлах, поэтому на обоих узлах устройства vpath имеют одинаковое имя.

6.1.2 Ресурсы кластера

Далее описывается, какие ресурсы необходимы для выполнения наших приложений в HACMP.

[1] Как видно из этого примера, желательно использовать «значимые» сервисные IP-метки, так как из названия, указанного здесь, понять, что она относится к группе ресурсов rg1 (в свою очередь, отвечающей за приложение DB2), невозможно. Подобные названия зачастую увеличивают время поиска и устранения неисправности. Имя p650n01 в данном примере больше подходит для постоянной IP-метки. *Примеч. науч. ред.*

Наш кластер имеет два приложения. В обычных условиях приложение DB2 выполняется на узле p650n01, тогда как приложение Development выполняется на узле p650n02. Узлы работают в традиционной конфигурации со взаимным перехватом. Каждое приложение имеет собственную группу томов: db2vg и devvg соответственно.

Конфигурация общих групп томов и файловых систем

В табл. 6.4 представлена конфигурация общих групп томов.

Таблица 6.4. Конфигурация общих групп томов

Имя группы томов	Vpath на p650n01	Vpath на p650n02	Старший номер устройства
db2vg	vpath0	vpath0	45
devvg	vpath1	vpath1	46

Табл. 6.5 содержит подробную конфигурацию файловых систем в группе томов db2vg.

Таблица 6.5. Конфигурация файловых систем первой группы томов

Имя группы томов	db2vg
Файловая система 1	/db2data
Файловая система 1 – логический том	/dev/db2datalv
Файловая система 1 – том журнала jfs	/dev/db2loglv
Файловая система 2	/db2archiv
Файловая система 2 – логический том	/dev/db2arclv
Файловая система 2 – том журнала jfs	/dev/db2loglv

Табл. 6.6 содержит конфигурацию файловых систем в группе томов devvg.

Таблица 6.6. Конфигурация файловых систем второй группы томов

Имя группы томов	devvg
Файловая система 1	/develop
Файловая система 1 – логический том	/dev/devlv
Файловая система 1 – том журнала jfs	Встроенный журнал[1]
Файловая система 2	/tools
Файловая система 2 – логический том	/dev/toolslv
Файловая система 2 – том журнала jfs	Встроенный журнал[2]

Серверы приложений

Наш кластер содержит два сервера приложений: DB2 и Development. В табл. 6.7 представлены скрипты запуска и остановки для серверов приложений.

[1] Здесь и далее в примерах в этой главе встроенный журнал (inline log) допустим только при использовании файловой системы JFS2. *Примеч. науч. ред.*

[2] См. также примечание в разделе 2.9.2. *Примеч. науч. ред.*

Таблица 6.7. Конфигурация серверов приложений

Имя сервера приложений	Скрипт запуска	Скрипт остановки
DB2	/usr/ha/start.db2	/usr/ha/stop.db2
Development	/usr/ha/start.development	/usr/ha/stop.development

Группы ресурсов

Мы создали две группы ресурсов. Каждая группа ресурсов содержит один сервер приложений, один сервисный IP-адрес и требуемую группу томов. Табл. 6.8 содержит подробную конфигурацию группы ресурсов.

Таблица 6.8. Конфигурация групп ресурсов

Имя ресурса	rg1[1]	rg2
Имена узлов-участников	p650n01, p650n02	p650n02, p650n01
Политика запуска	Подключение только на домашнем узле (Online On Home Node Only)	Подключение только на домашнем узле (Online On Home Node Only)
Политика перемещения при сбое	Перемещение при сбое на узел из списка со следующим приоритетом (Fallover To Next Priority Node In The List)	Перемещение при сбое на узел из списка со следующим приоритетом (Fallover To Next Priority Node In The List)
Политика возврата после восстановления	Возврат после восстановления на узел с более высоким приоритетом в списке (Fallback To Higher Priority Node In The List)	Возврат после восстановления на узел с более высоким приоритетом в списке (Fallback To Higher Priority Node In The List)
Политика динамического приоритета узлов		
Порядок обработки	Параллельная	Параллельная
Сервисные IP-метки	p650n01	p650n02
Серверы приложений	db2	development
Группы томов	db2vg	devvg
Использование принудительной активизации групп томов при необходимости	Нет	Нет
Автоматический импорт групп томов	Нет	Нет
Файловые системы	ВСЕ	ВСЕ
Проверка согласованности файловой системы	fsck	fsck
Метод восстановления файловой системы	Последовательный	Последовательный

[1] Безусловно, напрямую это не влияет на функционирование кластера, но ресурсные группы гораздо нагляднее называть по имени приложений, например здесь – db2rg и devrg (по аналогии с упомянутыми в этой главе группами томов и другими ресурсами). Такие названия помогают обслуживающему персоналу быстрее разобраться в ситуации в случае возникновения проблем. *Примеч. науч. ред.*

Окончание табл. 6.8

Имя ресурса	rg1[1]	rg2
Подключение файловых систем перед конфигурированием IP	Нет	Нет
Файловые системы/каталоги для экспорта		
Файловые системы/каталоги для NFS-подключения		
Сеть для NFS-подключения		
Основной класс WLM		

6.2 Конфигурация из четырех узлов

В этом разделе обсуждается конфигурация нового кластера из четырех узлов. Мы хотим добавить в кластер два новых узла и два новых приложения вместе с необходимыми ресурсами. Ниже представлена информация, используемая на этапе реконфигурации.

На рис. 6.2 представлена конфигурация нового кластера. Основные особенности кластера:

- четыре узла;
- одна открытая Ethernet-сеть (net_ether_01) с перехватом IP-адреса посредством синонимов;

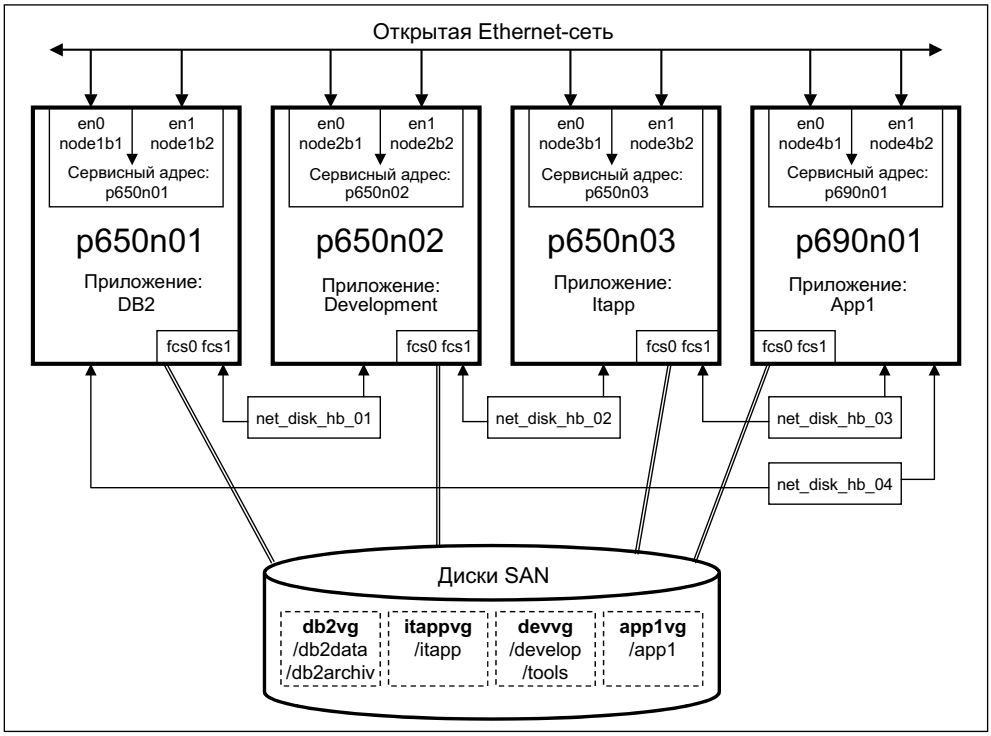

Рис. 6.2. Четырехузловая конфигурация

- два базовых IP-адреса и один сервисный адрес на узел;
- четыре сети пульса через диски для соединения без IP;
- по умолчанию на каждом узле выполняется одно приложение;
- одна общая группа томов на узел;
- резервные подключения Fibre Channel.

6.2.1 Топология кластера

Сведения о конфигурации топологии кластера:

Имя кластера: clto24.

Имя узла 1: p650n01.

Имя узла 2: p650n02.

Имя узла 3: p650n03.

Имя узла 4: p690n01.

Конфигурация сети представлена в табл. 6.9. В новом кластере все еще используется только одна сеть Ethernet с перехватом IP-адреса посредством синонимов. Изменение состоит в том, что теперь кластер содержит четыре сети пульса через диски. Подробно конфигурация сетей пульса через диски описана в разделе 6.2.2, «Конфигурация сетей пульса через диски».

Таблица 6.9. Конфигурация сетей

Имя сети	Тип сети	Маска сети	IPAT посредством синонимов
net_ether_01	ethernet (открытая)	255.255.255.0	Включен
net_diskhb_01	Сеть пульса через диски	Неприменимо	Неприменимо
net_diskhb_02	Сеть пульса через диски	Неприменимо	Неприменимо
net_diskhb_03	Сеть пульса через диски	Неприменимо	Неприменимо
net_diskhb_04	Сеть пульса через диски	Неприменимо	Неприменимо

Конфигурация IP-сети на новых узлах подобна конфигурации на старых узлах: узлы содержат две интерфейсные карты с базовыми адресами. Также мы имеем два новых сервисных адреса для новых групп ресурсов (табл. 6.10).

Таблица 6.10. Сетевые интерфейсы

IP-метка	Сетевой интерфейс	Функция интерфейса	IP-адрес
node1b1	en0	Базовый на p650n01	192.168.100.51
node1b2	en1	Базовый на p650n01	192.168.145.51
p650n01		Сервисный	192.168.147.51
node2b1	en0	Базовый на p650n02	192.168.100.52
node2b2	en1	Базовый на p650n02	192.168.145.52
p650n02		Сервисный	192.168.147.52
node3b1	en0	Базовый на p650n03	192.168.100.53
node3b2	en1	Базовый на p650n03	192.168.145.53

IP-метка	Сетевой интерфейс	Функция интерфейса	IP-адрес
p650n03		Сервисный	192.168.147.53
node4b1	en0	Базовый на p690n01	192.168.100.63
node4b2	en1	Базовый на p690n01	192.168.145.63
p690n01		Сервисный	192.168.147.63

6.2.2 Конфигурация сетей пульса через диски

Четырехузловой кластер требует использования четырех сетей пульса через диски. Каждая сеть соединяет два узла через устройство vpath. Четыре сети составляют кольцо, соединяющее узлы. В случае отказа сети Ethernet узлы могут передавать пакеты пульса друг другу по этому кольцу, например узел 1 может отправлять данные на узел 3. На рис. 6.3 показана топология сети пульса через диски.

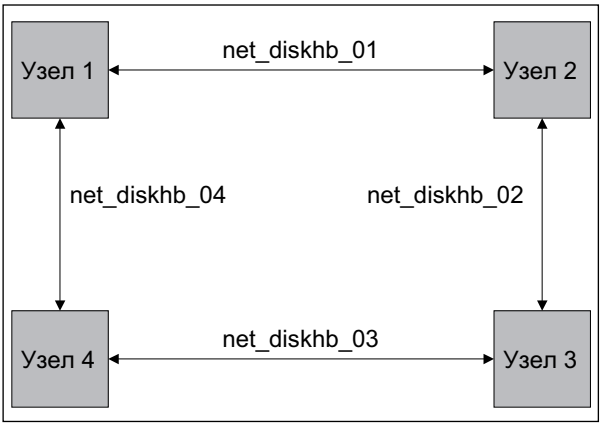

Рис. 6.3. Конфигурация сетей пульса через диски

Подробно конфигурация сети пульса через диски представлена в табл. 6.11.

> **Примечание.** Для согласованности мы используем одинаковую нумерацию устройств vpath на всех узлах, поэтому каждая пара узлов имеет одинаковое имя устройств vpath.

Таблица 6.11. Конфигурация сетей пульса через диски

Имя сети	Узел 1/устройство	Узел 2/устройство	Имя группы томов
net_diskhb_01	p650n01/vpath4	p650n02/vpath4	c12vg
net_diskhb_02	p650n02/vpath5	p650n03/vpath5	c23vg
net_diskhb_03	p650n03/vpath6	p690n01/vpath6	c34vg
net_diskhb_04	p690n01/vpath7	p650n01/vpath7	c41vg

6.2.3 Конфигурация групп ресурсов

Новый кластер содержит два новых приложения: Itapp выполняется на узле p650n03, его резервный узел – p690n01. Домашний узел приложения App1 – p690n01; он применяет динамический приоритет для перемещения при сбое на наименее используемый узел.

Мы не меняем серверы приложений DB2 и Development. В обычных условиях DB2 выполняется на узле p650n01, тогда как Development выполняется на узле p650n02. Узлы работают в традиционной конфигурации со взаимным перехватом.

Ниже описана конфигурация групп ресурсов HACMP, включая информацию об общем хранилище и группе ресурсов.

Конфигурация общих групп томов и файловых систем

В табл. 6.12 представлена конфигурация общей группы томов.

Таблица 6.12. Конфигурация общих групп томов

Имя группы томов	p650n01	p650n02	p650n03	p690n01	Старший номер устройства
db2vg	vpath0	vpath0	vpath0	vpath0	45
devvg	vpath1	vpath1	vpath1	vpath1	46
itappvg	vpath2	vpath2	vpath2	vpath2	47
app1vg	vpath3	vpath3	vpath3	vpath3	48

Табл. 6.13 содержит подробную конфигурацию файловых систем в группе томов db2vg.

Таблица 6.13. Конфигурации файловых систем первой группы томов

Имя группы томов	db2vg
Файловая система 1	/db2data
Файловая система 1 – логический том	/dev/db2datalv
Файловая система 1 – том журнала jfs	/dev/db2loglv
Файловая система 2	/db2archiv
Файловая система 2 – логический том	/dev/db2arclv
Файловая система 2 – том журнала jfs	/dev/db2loglv

Табл. 6.14 содержит конфигурацию файловых систем в группе томов devvg.

Таблица 6.14. Конфигурация файловых систем второй группы томов

Имя группы томов	devvg
Файловая система 1	/develop
Файловая система 1 – логический том	/dev/devlv
Файловая система 1 – том журнала jfs	Встроенный журнал
Файловая система 2	/tools
Файловая система 2 – логический том	/dev/toolslv
Файловая система 2 – том журнала jfs	Встроенный журнал

Табл. 6.15 содержит конфигурацию файловой системы в группе томов itappvg.

Таблица 6.15. Конфигурация файловой системы третей группы томов

Имя группы томов	itappvg
Файловая система 1	/itapp
Файловая система 1 – логический том	/dev/ itapplv
Файловая система 1 – том журнала jfs	Встроенный журнал

Табл. 6.16 содержит конфигурацию файловой системы в группе томов app1vg.

Таблица 6.16. Конфигурация файловой системы четвертой группы томов

Имя группы томов	app1vg
Файловая система 1	/app1
Файловая система 1 – логический том	/dev/ app1lv
Файловая система 1 – том журнала jfs	Встроенный журнал

Серверы приложений

Наш кластер содержит два новых сервера приложений: Itapp и App1. В табл. 6.17 представлены скрипты запуска и остановки для серверов приложений.

Таблица 6.17. Конфигурация серверов приложений

Имя сервера приложений	Скрипт запуска	Скрипт остановки
DB2	/usr/ha/start.db2	/usr/ha/stop.db2
Development	/usr/ha/start.development	/usr/ha/stop.development
Itapp	/usr/ha/start.itapp	/usr/ha/stop.itapp
App1	/usr/ha/start.app1	/usr/ha/stop.app1

Группы ресурсов

Табл. 6.18 содержит подробную конфигурацию для первоначальных групп ресурсов rg1 и rg2.

Таблица 6.18. Конфигурация групп ресурсов-1

Имя ресурса	rg1	rg2
Имена узлов-участников	p650n01, p650n02	p650n02, p650n01
Политика запуска	Подключение только на домашнем узле (Online On Home Node Only)	Подключение только на домашнем узле (Online On Home Node Only)
Политика перемещения при сбое	Перемещение при сбое на узел из списка со следующим приоритетом (Fallover To Next Priority Node In The List)	Перемещение при сбое на узел из списка со следующим приоритетом (Fallover To Next Priority Node In The List)
Политика возврата после восстановления	Возврат после восстановления на узел с более высоким приоритетом в списке (Fallback To Higher Priority Node In The List)	Возврат после восстановления на узел с более высоким приоритетом в списке (Fallback To Higher Priority Node In The List)
Политика динамического приоритета узлов		
Порядок обработки	Параллельная	Параллельная
Сервисные IP-метки	p650n01	p650n02
Серверы приложений	db2	development
Группы томов	db2vg	devvg

Окончание табл. 6.18

Имя ресурса	rg1	rg2
Использование принудительной активизации групп томов при необходимости	Нет	Нет
Автоматический импорт групп томов	Нет	Нет
Файловые системы	ВСЕ	ВСЕ
Проверка согласованности файловой системы	fsck	fsck
Метод восстановления файловой системы	Последовательный	Последовательный
Подключение файловых систем перед конфигурированием IP	Нет	Нет
Файловые системы/каталоги для экспорта		
Файловые системы/каталоги для NFS-подключения		
Сеть для NFS-подключения		
Основной класс WLM		

Табл. 6.19 содержит подробную конфигурацию новых групп ресурсов. Группа ресурсов rg4 применяет политику динамического перемещения при сбое: в случае отказа узла p690n01 выполняется ее перемещение на любой из остальных трех узлов с наименьшей нагрузкой на процессор.

Таблица 6.19. Конфигурация групп ресурсов-2

Имя ресурса	rg3	rg4
Имена узлов-участников	p650n03, p690n01, p650n02	p690n01, p650n03, p650n02, p650n01
Политика запуска	Подключение только на домашнем узле (Online On Home Node Only)	Подключение только на домашнем узле (Online On Home Node Only)
Политика перемещения при сбое	Перемещение при сбое на узел из списка со следующим приоритетом (Fallover To Next Priority Node In The List)	Перемещение при сбое на узел из списка со следующим приоритетом (Fallover To Next Priority Node In The List)
Политика возврата после восстановления	Возврат после восстановления на узел с более высоким приоритетом в списке (Fallback To Higher Priority Node In The List)	Возврат после восстановления на узел с более высоким приоритетом в списке (Fallback To Higher Priority Node In The List)
Политика динамического приоритета узлов		Перемещение при сбое на узел с наименьшей нагрузкой на процессор (Fallover to the node with the highest idle CPU (cl_highest_idle_cpu))
Порядок обработки	Параллельная	Параллельная

Окончание табл. 6.19

Имя ресурса	rg3	rg4
Сервисные IP-метки	p650n03	p690n01
Серверы приложений	itapp	app1
Группы томов	itappvg	app1vg
Использование принудительной активизации групп томов при необходимости	Нет	Нет
Автоматический импорт групп томов	Нет	Нет
Файловые системы	ВСЕ	ВСЕ
Проверка согласованности файловой системы	fsck	fsck
Метод восстановления файловой системы	Последовательный	Последовательный
Подключение файловых систем перед конфигурированием IP	Нет	Нет
Файловые системы/каталоги для экспорта		
Файловые системы/каталоги для NFS-подключения		
Сеть для NFS-подключения		
Основной класс WLM		

6.3 Реконфигурирование кластера

Расширение кластера с двух до четырех узлов состоит из следующих этапов:

1. Выполнение предварительных требований.
2. Добавление новых узлов.
3. Конфигурирование топологии IP-сети.
4. Конфигурирование сетей пульса через диски.
5. Конфигурирование групп ресурсов и серверов приложений.
6. Запуск новых узлов.
7. Тестирование кластера.

Замечание. Весь процесс конфигурирования можно выполнить при запущенном и работающем кластере.

6.3.1 Выполнение предварительных требований

На новых узлах требуется выполнить следующие действия:

- Установить AIX на новых узлах.
- Если возможно, установить на всех узлах последние обновления AIX и критические исправления. Это может потребовать перерыва в работе старых узлов.

- Установить на новых узлах ту же версию кода HACMP и те же исправления.
- Если возможно, установить последние исправления HACMP на всех узлах. Это может потребовать перерыва в работе первых двух узлов.
- Установить и протестировать базовые сетевые интерфейсы на новых узлах.
- Отредактировать файл /etc/hosts, чтобы он включал все IP-адреса всех узлов кластера. Убедитесь в том, что на всех узлах этот файл имеет одинаковое содержимое. Этот файл используется в процессе обнаружения HACMP, а также в процессе конфигурирования IP-сети.
- В целях обеспечения повышенной безопасности мы советуем записать все IP-интерфейсы всех узлов в файл /usr/es/sbin/cluster/etc/rhosts.
- Если вы не используете наборы файлов HACMP, то следует скопировать все необходимые скрипты пред- и постобработки событий, программы уведомления об ошибках, скрипты запуска/остановки и прочие пользовательские файлы на новые узлы.
- Выполнить синхронизацию идентификаторов пользователей и групп на новых узлах.
- Создать требуемые общие группы томов и файловые системы.
- Перезагрузить узлы, чтобы перезапустить clcomdES.

Внимание! Убедитесь в том, что на всех узлах уровни обновлений AIX и уровни обновлений HACMP одинаковы.

6.3.2 Добавление новых узлов в кластер

1. Запустите **smit hacmp** на одном из первоначальных узлов. Мы будем использовать этот узел для всех операций, кроме тех, о которых сказано другое.

2. Выберите **Extended Configuration (Расширенное конфигурирование)**.

3. Выберите **Extended Topology Configuration (Расширенное конфигурирование топологии).**

4. Выберите **Configure HACMP Nodes (Конфигурирование узлов HACMP).**

5. Выберите **Add a Node to the HACMP Cluster (Добавление узла в кластер HACMP).**

6. Введите имя нового узла.

7. Введите путь для связи к новому узлу или нажмите F4 для вывода списка IP-меток из файла /etc/hosts. См. скриншот SMIT на рис. 6.4. Также см. табл. 6.10. Добавьте узлы p630n3 и p690n01, используя один из базовых адресов в качестве путей для связи с узлом.

Повторите эти действия на каждом узле.

Обнаружение информации, относящейся к HACMP, на узлах

1. Запустите **smit hacmp**.

2. Выберите **Extended Configuration (Расширенное конфигурирование).**

```
                         Add a Node to the HACMP Cluster

  Type or select values in entry fields.
  Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
  * Node Name                                          [p650n03]
    Communication Path to Node                         [node4b1] +



  F1=Help             F2=Refresh          F3=Cancel           F4=List
  F5=Reset            F6=Command          F7=Edit             F8=Image
  F9=Shell            F10=Exit            Enter=Do
```

Рис. 6.4. Добавление узла в кластер HACMP

3. Выберите **Discover HACMP-related Information from Configured Nodes (Обнаружение информации, относящейся к HACMP, на сконфигурированных узлах).**

После этого HACMP выполняет обнаружение всей релевантной информации на всех узлах, включая новые. Впоследствии эта информация будет использоваться для конфигурирования кластера.

6.3.3 Конфигурирование топологии кластера

Следуйте нижеперечисленным указаниям, чтобы сконфигурировать параметры IP-сети кластера на новых узлах.

Добавление базовых интерфейсов в топологию кластера

См. параметры IP-сети кластера в табл. 6.9 и в табл. 6.10.

1. Запустите **smit hacmp.**

2. Выберите **Extended Configuration (Расширенное конфигурирование).**

3. Выберите **Extended Topology Configuration (Расширенное конфигурирование топологии).**

4. Выберите **Configure HACMP Communication Interfaces/Devices (Конфигурирование коммуникационных интерфейсов/устройств HACMP).**

5. Выберите **Add Communication Interfaces/Devices (Добавление коммуникационных интерфейсов/устройств).**

6. Выберите **Add Discovered Communication Interface and Devices (Добавление обнаруженных коммуникационных интерфейсов/устройств)**.

7. Выберите **Communication Interfaces (Коммуникационные интерфейсы).**

8. Выберите требуемую сеть (например, net_ether_01).

```
                 Configure HACMP Communication Interfaces/Devices

  Move cursor to desired item and press Enter.

    +-------------------------------------------------------------------+
    |      Select one or more Discovered Communication Interfaces to Add |
    |                                                                   |
    | Move cursor to desired item and press F7. Use arrow keys to scroll. |
    |     ONE OR MORE items can be selected.                            |
    | Press Enter AFTER making all selections.                          |
    |                                                                   |
    |   # Node / Network                                                |
    |   #       Interface         IP Label                     IP Address |
    |                                                                   |
    |   # net_ether_01 / p650n03                                        |
    |           en1               node3b2                      192.168.14 |
    |           en0               node3b1                      192.168.10 |
    |                                                                   |
    |   # net_ether_01 / p690n01                                        |
    |           en0               node4b1                      192.168.10 |
    |                                                                   |
    | F1=Help                 F2=Refresh              F3=Cancel         |
    | F7=Select               F8=Image                F10=Exit          |
  F1| Enter=Do                /=Find                  n=Find Next       |
  F9+-------------------------------------------------------------------+
```

Рис. 6.5. Конфигурирование базового сетевого интерфейса

9. Выберите добавляемый коммуникационный интерфейс. Базовые адреса новых узлов будут выведены в выпадающем списке (см. рис. 6.5).

10. Нажмите Enter для добавления выбранных интерфейсов.

Повторите эти действия для каждого базового интерфейса на всех новых узлах (node3b1, node3b2, node4b1 и node4b2). Если в списке не отображаются требуемые сетевые интерфейсы, нужно тщательно проверить параметры сети для новых узлов и файла /etc/hosts, после чего повторно выполнить процесс обнаружения HACMP.

Проверка новой сетевой топологии

1. Запустить **smit hacmp**.

2. Выполнить **Extended Configuration (Расширенное конфигурирование).**

3. Выполнить **Extended Topology Configuration (Расширенное конфигурирование топологии).**

4. Выполнить **Show HACMP Topology (Отображение топологии HACMP).**

5. Выполнить **Show Cluster Topology (Отображение топологии кластера).** Нажмите Enter, чтобы просмотреть новую информацию о топологии кластера.

> **Замечание.** Можно вывести топологию кластера с использованием утилиты /usr/es/sbin/cluster/utilities/cllsif.

Топология кластера должна соответствовать табл. 6.10, если вы все сделали правильно.

Конфигурирование сетей пульса через диски

Для удобства мы советуем вам создать отдельные группы томов для мониторинга пульса через диски, чтобы эти диски не использовались для других целей. Кроме того, следует импортировать группы томов на каждом узле, чтобы вы знали, какие диски для чего применяются.

Подробное описание конфигурации сетей пульса через диски см. в разделе 6.2.2, «Конфигурация сетей пульса через диски».

1. Создайте требуемое количество групп томов с возможностью расширенного одновременного доступа (enhanced concurrent capable volume groups) на одном узле (в нашем случае необходимо создать еще три группы томов: c23vg, c34vg и c41vg):

а) Запустите **smit lvm**.

б) Выберите **Volume Groups (Группы томов)**.

в) Выберите **Add a Volume Group (Добавить группу томов)**.

г) Введите следующую информацию (см. скриншот SMIT на рис. 6.6).

- Имя группы томов.
- Имена физических томов.

Volume group MAJOR NUMBER (Старший номер группы томов). Задайте номер, доступный на всех узлах. Доступные старшие номера можно просмотреть с использованием команды **lvlstmajor**.

- Установите для параметра **Create VG Concurrent Capable (Создать группу томов с возможностью одновременного доступа)** значение **enhanced concurrent (расширенный одновременный доступ)**. Создайте все три группы томов с одновременным доступом: c23vg, c34vg и c41vg.

> **Замечание.** Задайте группе томов пульса через диски осмысленное имя. Например, мы используем схему наименования cXYvg, где X и Y означают номера двух узлов, применяющих этот диск для мониторинга пульса, например группа томов, используемая для мониторинга пульса между узлами 3 и 4, называется c34vg.

2. Импортируйте новые группы томов на всех узлах. Необходимо везде применять одинаковый старший номер группы томов, например **importvg -V 53 -y c23vg vpath5**, либо использовать **smit importvg**.

3. Перезапустите процесс обнаружения кластера HACMP (См. раздел «Обнаружение информации, относящейся к HACMP, на узлах»).

4. Теперь мы можем добавить пары дисков для мониторинга пульса через диски в конфигурацию кластера:

```
                                Add a Volume Group

  Type or select values in entry fields.
  Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
    VOLUME GROUP name                                  [c23vg]
    Physical partition SIZE in megabytes +
  * PHYSICAL VOLUME names                              [vpath5] +
    Force the creation of a volume group?               no +
    Activate volume group AUTOMATICALLY                 no +
      at system restart?
    Volume Group MAJOR NUMBER                          [53] +#
    Create VG Concurrent Capable?                       enhanced concurrent +
    Create a big VG format Volume Group?                no +
    LTG Size in kbytes                                  128 +


  F1=Help             F2=Refresh          F3=Cancel           F4=List
  F5=Reset            F6=Command          F7=Edit             F8=Image
  F9=Shell            F10=Exit            Enter=Do
```

Рис. 6.6. Создание группы томов с возможностью расширенного одновременного доступа

а) Запустите **smit hacmp**.

б) Выберите **Extended Configuration (Расширенное конфигурирование)**.

в) Выберите **Extended Topology Configuration (Расширенное конфигурирование топологии)**.

г) Выберите **Configure HACMP Communication Interfaces/Devices (Конфигурирование коммуникационных интерфейсов/устройств HACMP)**.

д) Выберите **Add Communication Interfaces/Devices (Добавление коммуникационных интерфейсов/устройств)**.

е) Выберите **Add Discovered Communication Interface and Devices (Добавление обнаруженных коммуникационных интерфейсов/устройств)**.

ж) Выберите **Communication Devices (Коммуникационные устройства)**.

з) Выберите пары дисков для мониторинга пульса через диски, которые следует добавить (рис. 6.7). SMIT выводит список обнаруженных дисков, содержащих группы томов с возможностью расширенного одновременного доступа. Выберите два устройства из списка. Устройства должны иметь одинаковый идентификатор тома (volume ID), они должны представлять один и тот же физический диск. При выборе пар дисков всегда следует обращать внимание на идентификатор PVID (в правой части экрана), так как имена hdisk на разных узлах могут отличаться. Один диск может быть использован только в одной сети мониторинга пульса через диски. Также следует проверить выходные данные команды **lspv**, чтобы определить отношения между группами томов и дисками.

```
                Configure HACMP Communication Interfaces/Devices

  Mo+-------------------------------------------------------------------+
    | Select Point-to-Point Pair of Discovered Communication Devices to Add |
    |                                                                     |
    | Move cursor to desired item and press F7. Use arrow keys to scroll. |
    |     ONE OR MORE items can be selected.                              |
    | Press Enter AFTER making all selections.                            |
    |                                                                     |
    | [MORE...12]                                                         |
    |     p690n01                        vpath3   /dev/vpath3   000154decb5 |
    |     p650n01                        vpath6   /dev/vpath6   000197caca4 |
    |     p650n02                        vpath6   /dev/vpath6   000197caca4 |
    | >   p650n03                        vpath6   /dev/vpath6   000197caca4 |
    | >   p690n01                        vpath6   /dev/vpath6   000197caca4 |
    |     p650n01                        vpath7   /dev/vpath7   000215cad90 |
    |     p650n02                        vpath7   /dev/vpath7   000215cad90 |
    |     p650n03                        vpath7   /dev/vpath7   000215cad90 |
    |     p690n01                        vpath7   /dev/vpath7   000215cad90 |
    | [BOTTOM]                                                            |
    |                                                                     |
    | F1=Help                 F2=Refresh              F3=Cancel           |
    | F7=Select               F8=Image                F10=Exit            |
  F1| Enter=Do                /=Find                  n=Find Next         |
  F9+-------------------------------------------------------------------+
```

Рис. 6.7. Добавление устройств пульса через диски

Например, сеть пульса через диски между узлами p650n03 и p690n01 проходит через c34vg. Группа томов c34vg содержит один диск; этот диск на обоих узлах называется vpath6. Этот диск имеет идентификатор PVID 000197caca4d816f. Так что при выборе пары устройств в меню SMIT сначала нужно выбрать диск с требуемым PVID, после чего выбрать соответствующие узлы.

Повторите этот этап с каждой парой устройств пульса через диски для новых узлов. Описание конфигурации пары устройств пульса через диски см. в разделе 6.2.2, «Конфигурация сетей пульса через диски».

5. Проверьте новую топологию сети:

а) Запустите **smit hacmp.**

б) Выберите **Extended Configuration (Расширенное конфигурирование)**.

в) Выберите **Extended Topology Configuration (Расширенное конфигурирование топологии)**.

г) Выберите **Show HACMP Topology (Отображение топологии HACMP)**.

д) Выберите **Show Topology Information by Network Name (Отображение информации о топологии по имени сети)**.

е) Выберите **Show All Networks (Показать все сети)**. Пример выводимых данных см. на рис. 6.8.

```
                 Configure HACMP Communication Interfaces/Devices

  Mo+-------------------------------------------------------------------+
    | Select Point-to-Point Pair of Discovered Communication Devices to Add |
    |                                                                     |
    | Move cursor to desired item and press F7. Use arrow keys to scroll. |
    |     ONE OR MORE items can be selected.                              |
    | Press Enter AFTER making all selections.                            |
    |                                                                     |
    | [MORE...12]                                                         |
    |     p690n01                        vpath3   /dev/vpath3   000154decb5 |
    |     p650n01                        vpath6   /dev/vpath6   000197caca4 |
    |     p650n02                        vpath6   /dev/vpath6   000197caca4 |
    | >   p650n03                        vpath6   /dev/vpath6   000197caca4 |
    | >   p690n01                        vpath6   /dev/vpath6   000197caca4 |
    |     p650n01                        vpath7   /dev/vpath7   000215cad90 |
    |     p650n02                        vpath7   /dev/vpath7   000215cad90 |
    |     p650n03                        vpath7   /dev/vpath7   000215cad90 |
    |     p690n01                        vpath7   /dev/vpath7   000215cad90 |
    | [BOTTOM]                                                            |
    |                                                                     |
    | F1=Help                 F2=Refresh              F3=Cancel           |
    | F7=Select               F8=Image                F10=Exit            |
  F1| Enter=Do                /=Find                  n=Find Next         |
  F9+-------------------------------------------------------------------+
```

Рис. 6.8. Проверка конфигурации сети пульса через диски

К этому моменту мы успешно сконфигурировали новую топологию кластера. Пожалуйста, еще раз проверьте новую топологию и выполнение предварительных требований, прежде чем начинать процесс верификации и синхронизации.

Выполнение верификации и синхронизации кластера

1. Запустите **smit hacmp**.

2. Выберите **Extended Configuration (Расширенное конфигурирование)**.

3. Выберите **Extended Verification and Synchronization (Расширенная верификация и синхронизация)**.

4. Используйте следующие значения при запуске процесса верификации кластера (рис. 6.9):

- **Emulate or Actual (Эмуляция или действительное выполнение)**: Actual (Действительное выполнение).
- **Verify changes only? (Выполнить только верификацию изменений?)**: No (Нет).
- **Logging (Ведение журналов)**: Standard (Стандартное).

При верификации кластера выполняется проверка и наполнение ODM-классов HACMP информацией о новой конфигурации топологии. Если у вас включены набо-

```
        HACMP Verification and Synchronization (Active Cluster Nodes Exist)

   Type or select values in entry fields.
   Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
   * Emulate or Actual                                 [Actual] +
   * Verify changes only?                              [No] +
   * Logging                                           [Standard] +



   F1=Help             F2=Refresh          F3=Cancel           F4=List
   F5=Reset            F6=Command          F7=Edit             F8=Image
   F9=Shell            F10=Exit            Enter=Do
```

Рис. 6.9. Верификация и синхронизация кластера

ры файлов, HACMP копирует файлы на новые узлы. Вам нужно исправить возможные ошибки и предупреждения, прежде чем продолжать конфигурирование группы ресурсов. После успешной верификации новые узлы и коммуникационные интерфейсы должны выводиться в утилите **/usr/es/sbin/cluster/clstat**.

> **Внимание!** Когда кластер находится в активном состоянии, нельзя выполнять верификацию и синхронизацию HACMP с включенной опцией «Automatically correct errors found during verification» («Автоматически исправлять ошибки, обнаруженные при верификации»)[1]. Кроме того, нельзя выполнить принудительную синхронизацию в случае неудачной верификации.

6.3.4 Конфигурирование ресурсов

На данном этапе предполагается, что уже были выполнены следующие предварительные задачи:

- создано общее хранилище, группы томов и файловые системы;
- установлено приложение;
- созданы и протестированы скрипты запуска и остановки приложения.

Конфигурирование сервисного IP-адреса

Для выполнения этого этапа используйте данные конфигурации сети из табл. 6.10.

1. Запустите **smit hacmp**.
2. Выберите **Extended Configuration (Расширенное конфигурирование)**.

[1] В оригинале ошибка – опция называется «Automatically correct errors found during verification», т. е. автоматически исправлять ошибки, обнаруженные при верификации. *Примеч. науч. ред.*

3. Выберите **Extended Resource Configuration (Расширенное конфигурирование ресурсов)**.

4. Выберите **HACMP Extended Resources Configuration (Расширенное конфигурирование ресурсов HACMP)**.

5. Выберите **Configure HACMP Service IP Labels/Addresses (Конфигурирование сервисных IP-меток/адресов HACMP)**.

6. Выберите **Add a Service IP Label/Address (Добавить сервисную IP-метку/адрес)**.

7. Выберите **Configurable on Multiple Nodes (Настраиваемая на нескольких узлах)**.

8. Выберите свою сеть, например, net_ether_01.

9. Нажмите F4 для вывода списка обнаруженных адресов. Нажмите Enter для подтверждения (рис. 6.10).

```
   Add a Service IP Label/Address configurable on Multiple Nodes (extended)

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
* IP Label/Address                                    p650n03 +
* Network Name                                        net_ether_01
  Alternate Hardware Address to accompany IP Label/A []
  ddress




F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Рис. 6.10. Конфигурирование сервисной IP-метки

Повторите эти действия для новых сервисных IP-меток: p650n03 и p690n01.

Определение новых серверов приложений

1. Запустите **smit hacmp.**

2. Выберите **Extended Configuration (Расширенное конфигурирование)**.

3. Выберите **Extended Resource Configuration (Расширенное конфигурирование ресурсов)**.

4. Выберите **HACMP Extended Resources Configuration (Расширенное конфигурирование ресурсов HACMP)**.

5. Выберите **Configure HACMP Applications (Конфигурирование приложений HACMP)**.

6. Выберите **Configure HACMP Application Servers (Конфигурирование серверов приложений HACMP)**.

7. Выберите **Add an Application Server (Добавить сервер приложения)**.

8. Введите имя сервера приложения, а также полный путь и имя скриптов запуска и остановки. Здесь можно выполнить конфигурирование метода монитора приложения, если он у вас есть (см. скриншот SMIT на рис. 6.11).

```
                              Add Application Server

  Type or select values in entry fields.
  Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
  * Server Name                                        [app1]
  * Start Script                                       [/usr/ha/start.app1]
  * Stop Script                                        [/usr/ha/stop.app1]
    Application Monitor Name(s) +



  F1=Help             F2=Refresh          F3=Cancel           F4=List
  F5=Reset            F6=Command          F7=Edit             F8=Image
  F9=Shell            F10=Exit            Enter=Do
```

Рис. 6.11. Добавление сервера приложения

Создайте серверы приложений itapp и app1. Сведения о конфигурации серверов приложений см. в табл. 6.17.

Определение групп ресурсов

Подробные сведения о параметрах конфигурации групп ресурсов см. в табл. 6.18 и в табл. 6.19.

1. Запустите **smit hacmp**.

2. Выберите **Extended Configuration (Расширенное конфигурирование)**.

3. Выберите **Extended Resource Configuration (Расширенное конфигурирование ресурсов)**.

4. Выберите **HACMP Extended Resource Group Configuration (Расширенное конфигурирование групп ресурсов HACMP)**.

5. Выберите **Add a Resource Group (Добавить группу ресурсов)**.

6. Введите имя группы ресурсов. Выберите имена узлов-участников (нажмите F4 для вывода всплывающего списка имен узлов). Установите политики запуска, перемещения при сбое и возврата после восстановления (нажмите F4 для вывода имеющихся опций) (рис. 6.12).

```
                              Add Application Server

  Type or select values in entry fields.
  Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
  * Server Name                                        [app1]
  * Start Script                                       [/usr/ha/start.app1]
  * Stop Script                                        [/usr/ha/stop.app1]
    Application Monitor Name(s) +


  F1=Help             F2=Refresh          F3=Cancel           F4=List
  F5=Reset            F6=Command          F7=Edit             F8=Image
  F9=Shell            F10=Exit            Enter=Do
```

Рис. 6.12. Добавление группы ресурсов

Повторите эти действия для всех новых групп ресурсов (rg3 и rg4).

Определение атрибутов группы ресурсов

1. Запустите **smit hacmp.**

2. Выберите **Extended Configuration (Расширенное конфигурирование)**.

3. Выберите **Extended Resource Configuration (Расширенное конфигурирование ресурсов)**.

4. Выберите **HACMP Extended Resource Group Configuration (Расширенное конфигурирование групп ресурсов HACMP)**.

5. Выберите **Change/Show Resources and Attributes for a Resource Group (Изменить/показать ресурсы и атрибуты группы ресурсов)**.

6. Выберите из списка группу ресурсов, которую требуется изменить.

7. Выполните конфигурирование группы ресурсов на основе параметров конфигурации группы ресурсов, представленных в табл. 6.19 (см. скриншот SMIT на рис. 6.13).

Повторите эти действия для всех новых групп ресурсов (rg3 и rg4). Если какой-либо из предопределенных ресурсов (например, общая группа томов или файловая система) не отображается, следует еще раз запустить процесс обнаружения HACMP (см. раздел «Обнаружение информации, относящейся к HACMP, на узлах».

Выполнение верификации и синхронизации кластера

К этому моменту было успешно выполнено конфигурирование новых ресурсов кластера. Можно также выполнить любые другие необходимые задачи конфигурирования, например добавить мониторы приложений или скрипты пред- и постобработки событий.

```
          Change/Show All Resources and Attributes for a Resource Group

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

[TOP]                                                   [Entry Fields]
  Resource Group Name                                  rg4
  Participating Nodes (Default Node Priority)          p690n01 p650n02 p650n>
* Dynamic Node Priority Policy                         [cl_highest_idle_cpu] +

  Startup Policy                                       Online On Home Node O>
  Fallover Policy                                      Fallover Using Dynami>
  Fallback Policy                                      Fallback To Higher Pr>
  Fallback Timer Policy (empty is immediate)           [] +
  Service IP Labels/Addresses                          [p690n01] +
  Application Servers                                  [app1] +
  Volume Groups                                        [app1vg] +
  Use forced varyon of volume groups, if necessary     false +
  Automatically Import Volume Groups                   false +
  Filesystems (empty is ALL for VGs specified)         [] +
  Filesystems Consistency Check                        fsck +
  Filesystems Recovery Method                          sequential +
  Filesystems mounted before IP configured             false +
[MORE...15]

F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Рис. 6.13. Конфигурирование атрибутов группы ресурсов

Пожалуйста, ее раз проверьте новые ресурсы, прежде чем начинать процесс верификации и синхронизации, так как некоторые группы ресурсов могут быть активизированы на первоначальных узлах. В нашем примере группы ресурсов rg3 и rg4 будут подключены на узлах p650n01 или p650n02 после успешной синхронизации кластера.

1. Запустите **smit hacmp.**

2. Выберите **Extended Configuration (Расширенное конфигурирование)**.

3. Выберите **Extended Verification and Synchronization (Расширенная верификация и синхронизация)**.

4. Нажмите Enter, чтобы запустить процесс верификации и синхронизации.

Так как два узла уже работают и новая конфигурация ресурсов может на них повлиять, пожалуйста, проверьте файлы /tmp/hacmp.out на двух первоначальных узлах на наличие ошибок.

6.3.5 Запуск HACMP на новых узлах

Если верификация кластера прошла успешно, можно запустить HACMP на новых узлах.

1. Запустите **smit clstart**.

2. Выберите новые узлы и нажмите Enter для запуска HACMP.

Следует проверить файлы /tmp/hacmp.out на всех узлах на наличие ошибок при запуске. Если проектирование и конфигурирование были выполнены должным образом, то на этом этапе все должно быть нормально.

Тестирование кластера

Так как выполнение этой операции предполагает, что кластер должен продолжать работать на первых двух узлах, мы рекомендуем выполнить тестирование новой конфигурации во время окна обслуживания, когда допускается перерыв в работе существующих приложений. Мы предлагаем как можно быстрее выполнить тестирование перехвата и проверить максимально возможное количество сценариев. См. раздел 3.13, «Разработка плана тестирования кластера».

В табл. 6.20 представлено размещение группы ресурсов при отказе одного или двух узлов. При отключении узла p690n01 нельзя предсказать расположение группы ресурсов rg4, так как она будет перемещена на узел с наименьшим показателем использования процессора. В таблице это обозначено путем указания группы ресурсов в скобках.

Таблица 6.20. Размещение групп ресурсов при отказе узла

	p650n01	p650n02	p650n03	p690n01
Все узлы работают	rg1	rg2	rg3	rg4
Отказ p650n01	отказ	rg2, rg1	rg3	rg4
Отказ p650n02	rg1, rg2	отказ	rg3	rg4
Отказ p650n03	rg1	rg2	отказ	rg4, rg3
Отказ p690n01	rg1 (rg4)	rg2 (rg4)	rg3 (rg4)	отказ
Отказ p650n03 и p690n01	rg1 (rg4)	rg2, rg3, (rg4)	отказ	отказ
Отказ p650n01, p650n03 и p690n01	отказ	rg2, rg1, rg3, rg4	отказ	отказ

7

Обслуживание кластера

В этой главе описываются основные практические методы осуществления планирования и выполнения операций обслуживания кластера HACMP. Основная цель состоит в том, чтобы максимально долго обеспечивать активность приложений кластера. Для выполнения этих операций мы используем функциональные возможности HACMP и AIX 5L. Конечно же, приведенные сценарии не являются исчерпывающими.

В этой главе предполагается знание общих процедур устранения неполадок в AIX, а также мониторинга журнала ошибок. Мы не описываем, как определить существующие проблемы как во время устранения проблем, так и при выполнении профилактического обслуживания.

7.1 Контроль и тестирование изменений

Осуществление контроля за изменениями является необходимым условием эффективного обеспечения высокой доступности в кластерных системах. Управление изменениями представляет собой набор всесторонне документированных процедур. Оно включает несколько компонентов и является обязательным.

Управление изменениями среди прочего включает:

- ограниченный административный доступ;
- тщательно документированные и **протестированные** процедуры;
- надлежащее планирование и утверждение изменений.

7.1.1 Тестовый кластер

Тестовый кластер является важным элементом в обеспечении надлежащего управления изменениями и в общей эффективности рабочего кластера. Тестовые кластеры позволяют осуществлять тщательное тестирование административных процедур и/или процедур обслуживания с целью обнаружения проблем до того, как они достигнут рабочего кластера. Тестовые кластеры – это не роскошь, а **обязательный элемент**.

Многие существующие клиенты HACMP имеют тестовый кластер или по меньшей мере начинают с тестового кластера. Однако со временем узлы тестового кластера часто начинают применять в компании для других целей. Использование этих систем требует запланированного окна обслуживания, подобно рабочему кластеру. В такой ситуации не следует считать, что эти узлы больше не представляют тестовый кластер.

В идеале тестовый кластер должен использовать те же версии и обновления AIX, HACMP и приложений, которые использует и рабочий кластер. Предпочтительно также использовать максимально похожее оборудование. В большинстве случаев нецелесообразно полностью копировать рабочую среду, особенно при применении нескольких рабочих кластеров. Существует несколько методов, которые можно применять для обеспечения максимальной эффективности тестового кластера при употреблении нескольких рабочих кластеров с различным программным обеспечением.

Например, часто используются логические разделы (LPAR) и несколько различных образов rootvg с применением **alt_disk_install**. Логические разделы позволяют без труда создать тестовый кластер с минимальным использованием физических ресурсов; например, тестовый кластер даже может разместиться на одном физическом компьютере. Использование нескольких загрузочных образов позволяет клиентам легко сменить кластерную среду путем перезагрузки раздела с другого образа. Это также позволяет осуществлять тестирование многих программных процедур:

- обслуживание AIX;
- применение исправлений HACMP;
- обслуживание приложений.

Этот тип тестового кластера требует употребления как минимум одного диска на образ на LPAR. Например, если тестовый кластер содержит два узла и три разных

образа rootvg, требуется использовать по меньшей мере шесть жестких дисков. Тем не менее это гораздо проще, чем применять шесть отдельных узлов в трех тестовых кластерах.

Общую стоимость тестового кластера можно еще больше сократить с использованием расширенной виртуализации Power5. Расширенная виртуализация (advanced virtualization) позволяет располагать несколько образов rootvg на одном физическом диске. Она также позволяет логическим разделам использовать общую физическую сеть и DASD-адаптеры через VLAN и VSCSI. Дополнительные сведения об этих возможностях см. в книге **Advanced POWER Virtualization on IBM Eserver p5 Servers: Introduction and Basic Configuration,** SG24-7940[1].

Тестовый кластер также позволяет выполнять тестирование процедур обслуживания оборудования. Эти процедуры в числе прочего включают:

- обновление микрокода компьютера;
- обновление микрокода адаптера;
- замена адаптера;
- замена диска.

Другие операции тестирования можно выполнить с использованием инструмента тестирования кластера (Cluster Test Tool) и/или эмуляции событий.

7.2 Запуск и остановка кластера

Под **запуском служб кластера** подразумевается процесс запуска подсистем RSCT, необходимых для работы HACMP, с последующим запуском демонов HACMP, обеспечивающих необходимое согласование между узлами кластера. Во время запуска диспетчер кластера (Cluster Manager) выполняет событие **node_up**, после чего осуществляется подхват групп ресурсов. Под остановкой служб кластера понимается остановка этих демонов на узле, которая в зависимости от типа остановки может вызывать выполнение дополнительных скриптов HACMP.

> **Важно!** Начиная с HACMP V5.3 процесс диспетчера кластера (clstrmgrES) всегда выполняется. Он может находиться в одном из двух состояний, выводимых следующей командой:
> lssrc -ls clstrmgrES
> ST_INIT (событие запуска выполняется)
> ST_NOTCONFIGURED (событие запуска не выполняется)

Изменения в состоянии кластера называются *событиями кластера (cluster events)*. Диспетчер кластера (Cluster Manager) на каждом узле осуществляет мониторинг локальной аппаратной и программной подсистем на такие события, как, например, событие **отказа приложения (application failure)**. При возникновении таких событий диспетчер кластера выполняет один или несколько скриптов, например скрипт **перезапуска приложения (restart application)**. Диспетчеры кластера, выполняющиеся на всех узлах, обмениваются сообщениями в целях согласования необходимых действий в при возникновении событий.

[1] Эта книга также готовится к изданию в переводе на русский язык. *Примеч. науч. ред.*

Во время периодов обслуживания часто бывает необходимо остановить и запустить службы кластера. Однако перед этим необходимо понять, какие при этом возникают взаимодействия между узлами, а также каково воздействие на доступность системы. Кластер должен быть синхронизирован, и верификация не должна обнаруживать ошибок. Следующий раздел кратко описывает эти процессы, а также обработку, выполняемую при запуске или остановке этих служб. Далее в этом разделе приведено описание процедур, необходимых для запуска или остановки служб кластера на узле.

7.2.1 Службы кластера

Ниже перечислены основные демоны HACMP и RSCT.

- **Демон диспетчера кластера (Cluster Manager daemon, clstrmgrES)**. Основной демон HACMP. Он обеспечивает глобальное представление топологии кластера и ресурсов, а также запускает скрипты обработки событий в ответ на изменения состояния узлов, интерфейсов или ресурсов (или по запросу пользователя).

  Диспетчер кластера получает информацию о состоянии интерфейсов от служб топологии (Topology Services). Диспетчер кластера обеспечивает актуальную информацию о расположении и состоянии всех групп ресурсов. Диспетчер кластера является клиентом служб групп (Group Services) и использует последние для обеспечения надежной связи между демонами.

  В версиях 5.1 и 5.2 этот демон запускается после служб кластера. В версии 5.3 демон clstrmgr запускается через процесс init и должен выполняться постоянно. Кроме того, так как в версии 5.3 демон clstrmgr является постоянно выполняющимся процессом, нельзя использовать команду **lssrc -s clstrmgrES** для определения состояния кластера. Вместо этого следует использовать команду **/usr/es/sbin/cluster/utilities/clcheck_server grpsvcs**.

- **Демон диспетчера блокировки кластера (Cluster Lock Manager daemon, cllockd).** Этот демон представляет информационные службы блокировки. Использование демона cllockd на узлах кластера необходимо только в том случае, если эти узлы входят в конфигурацию с одновременным доступом. Обратите внимание на то, что диспетчер блокировки не поддерживается на 64-разрядном ядре и полностью удален начиная с HACMP V5.2.

- **Демон коммуникаций кластера (Cluster Communication Daemon, clcomdES).** Этот демон, впервые появившийся в версии 5.1, обеспечивает безопасную связь между узлами кластера для всех утилит кластера, в частности утилит верификации и синхронизации и управления системой (C-SPOC). Демон **clcomd** запускается автоматически при загрузке через процесс **init**. Начиная с версии 5.2, clcomdES должен быть запущен, прежде чем можно будет запустить какие-либо службы кластера.

- **Демон информации кластера (Cluster Information Program, clinfoES).** Этот демон предоставляет информацию о состоянии кластера узлам кластера и клиентам, а также вызывает скрипт **/usr/es/sbin/cluster/etc/clinfo.rc** при возникновении событий кластера. Демон clinfo является необязательным на узлах кластера и клиентах.

- **Демон Cluster SMUX Peer daemon (clsmuxpd).** Этот демон предоставляет информацию о состоянии объектов кластера. Он публикует информацию о состоянии топологии и ресурсов кластера для MIB. Он является клиентом диспетчера кластера (Cluster Manager). Этот демон работает совместно с демоном протокола SNMP (snmpd). Демон clsmuxpd должен выполняться на всех узлах кластера.

> Примечание. Демон clsmuxpd нельзя запустить при отключенном демоне snmpd. В HACMP 5.3 clsmuxpd больше не используется.

- **Подсистема служб топологии кластера (Cluster Topology Services Subsystem).** Подсистема служб топологии кластера RSCT обеспечивает мониторинг состояния сетевых интерфейсов и публикует состояние для клиентов, осуществляющих доступ к информации посредством членства в службах групп. Основным демоном является демон **hatsd.** Службы топологии также включают сетевые интерфейсные модули **hats_nim***, отправляющие и получающие пакеты пульса. Подсистема служб топологии кластера должна выполняться на всех узлах кластера.
- **Подсистема управления событиями кластера (Cluster Event Management Subsystem).** Эта подсистема RSCT сопоставляет информацию о состоянии системных ресурсов с информацией о состоянии ресурсов, требуемых для клиентских программ (приложений, подсистем и прочих программ). Демон haemd выполняется на всех узлах домена.

> Примечание. Подсистема управления событиями кластера используется только в Oracle 9i; в HACMP 5.2 и более поздних версиях он заменен подсистемой RMC (Resource Monitoring and Control).

- **Монитор ресурсов подсистемы управления событиями операционной системы AIX (Event Management AIX Operating System Resource Monitor).** Этот демон RSCT выступает в качестве монитора ресурсов для подсистемы управления событиями и обеспечивает информацию о параметрах и использовании операционной системы. Демон **emaixos** автоматически запускается подсистемой управления событиями.
- **Подсистема служб групп кластера (Cluster Group Services Subsystem).** Эта подсистема RSCT обеспечивает надежную связь и протоколы, необходимые для работы кластера. Клиенты – это распределенные демоны, такие, как диспетчер кластера HACMP (HACMP Cluster Manager) и диспетчер логических томов с расширенным одновременным доступом (Enhanced Concurrent Logical Volume Manager). Демон hagsd должен выполняться на всех узлах кластера.
- **Демон сервера глобализации кластера (Cluster Globalized Server daemon, grpglsmd).** Этот демон RSCT выполняется в качестве клиента служб групп; его функция состоит в том, чтобы обеспечить глобальное членство адаптера коммутатора по всем узлам кластера. Демон grpglsmd должен выполняться на всех узлах кластера.
- **Подсистема мониторинга и управления ресурсами (Resource Monitoring and Control Subsystem).** Эта подсистема RSCT выполняется в качестве монитора ресурсов для подсистемы управления событиями и обеспечивает информацию

о параметрах и использовании операционной системы. Подсистема RMC должна выполняться на всех узлах кластера. По умолчанию демон **rmcd** после установки настроен на запуск из **inittab**. Скрипт **rc.cluster** обеспечивает работу подсистемы RMC.

7.2.2 Запуск служб кластера

В этом разделе описываются опции запуска служб кластера на одном узле, на нескольких узлах или даже на всех узлах. Запуск служб кластера следует всегда выполнять с использованием SMIT. Экран SMIT представлен на рис. 7.1.

```
                          Start Cluster Services

  Type or select values in entry fields.
  Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
  * Start now, on system restart or both                now +
    Start Cluster Services on these nodes              [Maddi,Melany] +
    BROADCAST message at startup?                       true +
    Startup Cluster Information Daemon?                 false +
    Reacquire resources after forced down ?             false +
    Ignore verification errors?                         false +
    Automatically correct errors found during           Interactively +
    cluster start?


  F1=Help             F2=Refresh          F3=Cancel           F4=List
  F5=Reset            F6=Command          F7=Edit             F8=Image
  F9=Shell            F10=Exit            Enter=Do
```

Рис. 7.1. Меню запуска служб кластера

Из учетной записи пользователя «root» следует выполнить следующие действия для запуска служб кластера на узле:

1. Введите быстрый путь **smit cl_admin.**

2. В SMIT выберите **Manage HACMP Services (Управление службами HACMP)** -> **Start Cluster Services (Запустить службы кластера)** и нажмите **Enter.**

3. Введите следующие значения полей:

Start now, on system restart or both (Запустить сейчас, при перезапуске системы или в обоих случаях). Определяет, когда следует запускать службы кластера и **clinfoES**: при вводе значений в этой панели нажатием Enter (**now**) при перезагрузке операционной системы (**on system restart**) или в обоих случаях (**both**).

Примечание. В рабочей среде обычно не рекомендуется задавать автоматический запуск служб кластера при перезапуске системы.
Причина этого непосредственно связана с последствиями отказа системы. Если происходит отказ системы, владеющей группой ресурсов и AIX настроен на перезагрузку после отказа, может произойти перезапуск службы кластера посреди перехвата ресурсов. В зависимости от конфигурации кластера это может вызвать конфликт группы ресурсов, ошибки обработки группы ресурсов или даже возврат. Все это может увеличить перерыв в обслуживании.
Однако при тестировании и обслуживании, а также на выделенных дежурных узлах использование этой опции может быть удобным.

- **Start Cluster Services on these nodes (Запустить службы кластера на этих узлах).** Следует ввести имена одного или нескольких узлов, на которых требуется запустить службы кластера. Узлы также можно выбрать из списка. При указании нескольких узлов вручную следует разделять имена запятыми, как показано на рис. 7.1.
- **BROADCAST message at startup? (Отправлять широковещательное сообщение при запуске?).** Указывает, требуется ли отправлять широковещательное сообщение на все узлы при запуске служб кластера.
- **Startup Cluster Lock Services (Запустить службу блокировки кластера).** Выберите **true** или **false**, чтобы определить, нужно ли запускать демон cllockd/cllockdES. Используется только в среде с одновременным доступом.

Примечание. HACMP 5.2 и более поздние версии больше не поддерживают cllockd и cllockdES (Cluster Lock Manager, диспетчер блокировки кластера).

- **Startup Cluster Information Daemon? (Запустить демон информации кластера?).** Указывает, требуется ли запустить демон **clinfo**. Если ваше приложение применяет Clinfo, если вы используете монитор **clstat** или если вы хотите запустить эмуляцию событий, установите в этом поле значение **true**. В противном случае установите значение **false**.
- **Reacquire Resources after Forced Down? (Выполнить повторный перехват ресурсов после принудительного отключения?).** По умолчанию установлено значение **false**, так как HACMP ожидает, что ресурсы будут в подключенном состоянии, так как они не были переведены в отключенное состояние. Если службы были предварительно остановлены с использованием опции **forced** и ресурсы были переведены в отключенное состояние после остановки служб и при этом требуется, чтобы ресурсы автоматически переводились в подключенное состояние, надо выбрать **true**.

  В HACMP V5.3 были добавлены две новые опции:
- **Ignore Verification Errors? (Игнорировать ошибки верификации?).** Если требуется запускать службы кластера при обнаружении в процессе верификации ошибок на заданных узлах или в кластере в общем, следует установить значение

true для заданных узлов. Если не требуется запускать службы кластера на заданных узлах при обнаружении ошибок на каком-либо узле в процессе верификации, следует установить значение **false**.

- **Automatically correct errors found during cluster start? (Автоматическое исправление ошибок, обнаруженных при запуске кластера?).** Возможные значения: **Yes (Да)**, **No (Нет)** и **Interactively (Интерактивное)**.

При значении Yes (Да) выполняется автоматическое исправление ошибок без предупреждения. При значении No (Нет) исправление не осуществляется и кластер не запускается при обнаружении ошибок. При значении Interactively (Интерактивное) выполняются запросы к пользователю при запуске с указанием обнаруженных ошибок и выбором действия (исправлять или не исправлять).

Нажмите **Enter**. Система запускает службы кластера на заданном узле, активизируя определенную вами конфигурацию кластера. Продолжительность выполнения команд и скриптов зависит от вашей конфигурации (другими словами, от количества дисков, количества конфигурируемых интерфейсов, количества подключаемых файловых систем и количества запускаемых приложений).

- Во время события **node_up** выполняется подхват группы ресурсов. Продолжительность выполнения каждого события **node_up** зависит от обработки ресурсов во время события. События **node_up** для объединения узлов обрабатываются последовательно.

При выполнении команды, когда службы кластера HACMP запущены на всех заданных узлах, SMIT выводит окно состояния команды. Обратите внимание на то, что, когда панель SMIT указывает завершение запуска кластера, обработка событий в большинстве случаев еще не завершена. Чтобы проверить, работают ли узлы, можно использовать **clstat**, **WebSmit** или даже просмотр (tail) файла /tmp/hacmp.out на любом узле. Дополнительные сведения по этой теме см. в разделе «Утилиты проверки состояния кластера».

7.2.3 Остановка служб кластера

Ниже приведен список действий, представляющих процедуру остановки служб кластера на одном узле, на нескольких узлах или на всех узлах кластера с использованием утилиты C-SPOC на одном из узлов кластера. C-SPOC останавливает узлы последовательно, а не параллельно. Если какой-либо из заданных узлов неактивен, операция завершения работы на этом узле прерывается. Как и при запуске служб, для остановки служб кластера следует всегда использовать SMIT. Экран SMIT остановки кластера представлен на рис. 7.2.

Чтобы остановить службы кластера:

1. Введите быстрый путь **smit cl_admin.**

2. В SMIT выберите **Manage HACMP Services (Управление службами HACMP)** -> **Start Cluster Services (Запустить службы кластера)** и нажмите **Enter**.

3. Введите следующие значения полей:

- **Stop now, on system restart or both (Остановить сейчас, при перезапуске системы или в обоих случаях).**

```
                              Stop Cluster Services

  Type or select values in entry fields.
  Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
  * Stop now, on system restart or both                  now +
    Stop Cluster Services on these nodes                [Maddi,Melany] +
    BROADCAST cluster shutdown?                          true +
  * Shutdown mode                                        graceful +


  F1=Help             F2=Refresh          F3=Cancel           F4=List
  F5=Reset            F6=Command          F7=Edit             F8=Image
  F9=Shell            F10=Exit            Enter=Do
```

Рис. 7.2. Меню остановки служб кластера

Определяет, когда следует останавливать службы кластера: в данный момент (**now**), при перезагрузке операционной системы (**restart**) или в обоих случаях (**both**). При выборе значений **restart** или **both** в файле **/etc/inittab** происходит удаление записи, запускающей службы кластера. Службы кластера больше не будут автоматически запускаться после перезагрузки.

- **BROADCAST cluster shutdown? (Отправлять широковещательное сообщение при завершении работы кластера?).** Указывает, требуется ли отправлять широковещательное сообщение пользователям перед остановкой служб кластера. Если задано значение **true**, выполняется широковещательная рассылка сообщения на все узлы кластера.
- **Shutdown mode (Режим завершения работы).** Указывает тип остановки:
  - **graceful (постепенная).** Завершение работы после запуска скрипта **/usr/es/sbin/cluster/events/node_down_complete** на узле для освобождения ресурсов кластера. Другие узлы кластера не выполняют перехват ресурсов остановленного узла.
  - **graceful with takeover (постепенная с передачей ресурсов на резервные узлы).** Завершение работы после запуска скрипта **/usr/es/sbin/cluster/events/node_down_complete** для освобождения ресурсов кластера. Другие узлы выполняют перехват ресурсов остановленного узла.
  - **forced (принудительная).** Немедленное завершение работы. Узел сохраняет управление всеми своими ресурсами. Эту опцию можно использовать для отключения узла при выполнении обслуживания или внесении изменений в конфигурацию кластера, в частности при добавлении сетевой карты.

Однако после остановки служб кластера больше не обеспечивается высокая доступность приложений. При возникновении отказа восстановление выполняться не будет.

Примечание. При использовании групп томов с расширенным одновременным доступом в любом из режимов одновременного доступа для мониторинга пульса через диски или в режиме быстрого перехвата дисков опция forced недоступна.

Наблюдения

В прежних версиях HACMP после остановки диспетчера кластера (Cluster Manager) больше не обеспечивалась высокая доступность приложений. При возникновении отказа восстановление приложений не обеспечивалось. В HACMP 5.3 диспетчер кластера (Cluster Manager) продолжает работать даже после остановки служб кластера. Скрипты rc.cluster и clstop выдают диспетчеру кластера команды IPC™. Диспетчер кластера выполняет скрипт обработки событий для запуска стека RSCT (rc.cluster).

Кроме того, в HACMP 5.3 node_down_complete вызывает процедуру остановки стека RSCT. Диспетчер кластера реагирует на команды stopsrc немедленной остановкой без запуска события node_down, что необходимо перед установкой, обслуживанием или заменой демона clstrmgr или зависимых библиотек (например, libclstr.a).

Важно! Никогда не используйте команду kill – 9 для остановки диспетчера кластера или демонов RSCT. Это вызывает аварийное завершение. SRC запускает скрипт clexit.rc и немедленно останавливает систему. Это, в свою очередь, инициирует перемещение при сбое на других узлах.

7.3 Управление группами ресурсов и приложениями

В этом разделе обсуждаются следующие темы:

- перевод группы ресурсов в отключенное состояние;
- перевод группы ресурсов в подключенное состояние;
- перемещение группы ресурсов;
- приостановка/возобновление мониторинга приложения.

Понимание этих вопросов так же важно, как и понимание остановки и запуска служб кластера, так как эти операции часто применяются в периоды обслуживания.

В следующих разделах мы начинаем с предположения, что службы кластера выполняются, группы ресурсов подключены, приложения работают и кластер стабилен. Если кластер находится в нестабильном состоянии, то операции с группами ресурсов невозможны.

Все три обсуждаемые нами операции с группами ресурсов могут быть выполнены с использованием команды **clRGmove**. Однако в наших примерах мы будем применять C-SPOC. Все эти операции имеют похожие экраны SMIT и списки для выбора. Чтобы сократить объем главы, мы представим только по одному экрану SMIT в каждом из последующих разделов.

При выполнении любой из операций с группами ресурсов важно понимать действие параметра **priority override location (расположение, отменяющее приоритет**). Это настолько важно, что мы посвятили этому целый раздел – «Расположение, отменяющее приоритет».

7.3.1 Перевод группы ресурсов в отключенное состояние через SMIT

Чтобы перевести группу ресурсов в отключенное состояние:

1. Введите быстрый путь **smit cl_admin.**

2. В SMIT выберите **HACMP Resource Group and Application Management (Управление группами ресурсов и приложениями HACMP)** -> **Bring a Resource Group Offline (Перевести группу ресурсов в отключенное состояние)**. Выводится список, как показано на рис. 7.3. Он отображает только группы ресурсов, находящиеся в подключенном состоянии или в состоянии ERROR (Ошибка) на всех узлах в кластере.

3. Выберите требуемую группу ресурсов из списка и нажмите **Enter**. После выбора группы ресурсов появляется другой список, **Select a Destination Node (Выбор целевого узла)**. Список будет содержать только активные на данный момент узлы кластера, участвующие в ранее выбранной группе ресурсов.

4. Выберите целевой узел из списка и нажмите **Enter**.

5. Появляется последнее меню SMIT с информацией, выбранной в предыдущих списках. Кроме того, необходимо задать значение еще одного дополнительного поля, **Persist across Cluster Reboot?**. Вообще говоря, следует оставить в этом поле зна-

```
                HACMP Resource Group and Application Management

 Move cursor to desired item and press Enter.

   Bring a Resource Group Online
   Bring a Resource Group Offline
   Move a Resource Group to Another Node / Site

   Suspend/Resume Application Monitoring
   Application Availability Analysis

   ___________________________________________________________________________
   |                          Select a Resource Group                        |
   |                                                                         |
   | Move cursor to desired item and press Enter.                            |
   |                                                                         |
   |   #                                                                     |
   |   # Resource Group                  State                Node(s) / Site|
   |   #                                                                     |
   |   Maddi_rg                          ONLINE               Maddi /        |
   |                                                                         |
   | F1=Help                 F2=Refresh              F3=Cancel               |
   | F8=Image                F10=Exit                Enter=Do                |
 F1| /=Find                  n=Find Next                                     |
 F9|_________________________________________________________________________|
```

Рис. 7.3. Список групп ресурсов

чение по умолчанию или false. Это поле непосредственно связано с параметром **POL**; дополнительные сведения см. в разделе «Расположение, отменяющее приоритет».

6. Просмотрите ранее заданные записи, после чего нажмите **Enter**, чтобы начать обработку отключаемой группы ресурсов.

После завершения обработки группа ресурсов будет отключена, тогда как службы кластера на узле останутся активными.

7.3.2 Перевод группы ресурсов в подключенное состояние через SMIT

Чтобы перевести группу ресурсов в подключенное состояние:

1. Введите быстрый путь **smit cl_admin.**

2. В SMIT выберите **HACMP Resource Group and Application Management (Управление группами ресурсов и приложениями HACMP)** -> **Bring a Resource Group Online (Перевести группу ресурсов в подключенное состояние)**. Выводится список. Он отображает только группы ресурсов, находящиеся в подключенном состоянии или в состоянии ERROR (Ошибка) на всех узлах в кластере.

3. Выберите требуемую группу ресурсов из списка и нажмите **Enter**. После выбора группы ресурсов появляется другой список, **Select a Destination Node (Выбор целевого узла)**. Список будет содержать только активные на данный момент узлы кластера, участвующие в ранее выбранной группе ресурсов.

```
                 HACMP Resource Group and Application Management

 Move cursor to desired item and press Enter.

   Bring a Resource Group Online
   ___________________________________________________________________________
   |                      Select a Destination Node                          |
   |                                                                         |
   | Move cursor to desired item and press Enter.                            |
   |                                                                         |
   |   # To choose the highest priority available node for the               |
   |   # resource group, and to remove any Priority Override Location        |
   |   # that is set for the resource group, select                          |
   |   # "Restore_Node_Priority_Order" below.                                |
   |  Restore_Node_Priority_Order                                            |
   |                                                                         |
   |   # To choose a specific node, select one below.                        |
   |   Maddi                                                                 |
   |   Melany                                                                |
   |                                                                         |
   | F1=Help                 F2=Refresh              F3=Cancel               |
   | F8=Image                F10=Exit                Enter=Do                |
 F1| /=Find                  n=Find Next                                     |
 F9|_________________________________________________________________________|
```

Рис. 7.4. Список целевых узлов

4. Выберите целевой узел из списка, как показано на рис. 7.4. После выбора этого узла он становится расположением, отменяющим приоритет (priority override location) для данной группы ресурсов.

5. Появляется последнее меню SMIT с информацией, выбранной в предыдущих списках. Кроме того, необходимо задать значение еще одного дополнительного поля, **Persist across Cluster Reboot?**. Вообще говоря, следует оставить в этом поле значение по умолчанию или false. Это поле непосредственно связано с параметром **POL**; дополнительные сведения см. в разделе «Расположение, отменяющее приоритет».

6. Просмотрите ранее заданные записи, после чего нажмите **Enter**, чтобы начать обработку подключаемой группы ресурсов.

После успешного завершения HACMP выводит сообщение о состоянии, расположении и типе расположения (постоянное или нет) группы ресурсов, которая была успешно остановлена на заданном узле.

7.3.3 Перемещение группы ресурсов через SMIT

Перемещение группы ресурсов включает постепенную остановку ресурсов на текущем узле-владельце с последующим выполнением обычной процедуры запуска группы ресурсов на целевом узле. При этом возникает короткий перерыв в обслуживании, во время которого доступ к приложению отсутствует.

В HACMP V5.3 была добавлена возможность перемещения группы ресурсов на другой сайт. Общая идея такая же, как и при перемещении между локальными узлами. В нашем случае мы будем выполнять перемещение на другой узел, а не на другой сайт.

Чтобы переместить группу ресурсов:

1. Введите быстрый путь **smit cl_admin.**

2. В SMIT выберите **HACMP Resource Group and Application Management (Управление группами ресурсов и приложениями HACMP)** -> **Move Resource Groups to Another Node (Переместить группу ресурсов на другой узел)**. Выводится список. Он отображает только группы ресурсов, находящиеся в подключенном состоянии или в состоянии ERROR (Ошибка) на всех узлах в кластере.

3. Выберите требуемую группу ресурсов из списка и нажмите **Enter**. После выбора группы ресурсов появляется другой список, **Select a Destination Node (Выбор целевого узла)**. Список будет содержать только активные на данный момент узлы кластера, участвующие в ранее выбранной группе ресурсов.

4. Выберите целевой узел из списка. После выбора этого узла он становится расположением, отменяющим приоритет (priority override location) для данной группы ресурсов.

5. Появляется последнее меню SMIT с информацией, выбранной в предыдущих списках. Кроме того, необходимо задать значение еще одного дополнительного поля, **Persist across Cluster Reboot?**. Вообще говоря, следует оставить в этом поле значение по умолчанию или false. Параметр **POL** устанавливается в любом случае; вопрос в том, нужно ли, чтобы он оставался установленным после остановки/запуска кластера. Дополнительные сведения см. в разделе «Расположение, отменяющее приоритет».

6. Просмотрите ранее заданные записи, после чего нажмите **Enter**, чтобы начать выполнять перемещение группы ресурсов.

После успешного завершения HACMP выводит сообщение о состоянии, расположении и типе расположения (постоянное или нет) группы ресурсов, которая была успешно перемещена на заданном узле, как показано на рис. 7.6.

```
                              Move a Resource Group

  Type or select values in entry fields.
  Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
    Resource Group to be Moved                         Maddi_rg
    Destination Node                                   Melany
    Persist Across Cluster Reboot?                     false+


  F1=Help             F2=Refresh          F3=Cancel           F4=List
  F5=Reset            F6=Command          F7=Edit             F8=Image
  F9=Shell            F10=Exit            Enter=Do
```

Рис. 7.5. Экран перемещения группы ресурсов

```
                                 COMMAND STATUS

Command: OK            stdout: yes           stderr: no

Before command completion, additional instructions may appear below.

[MORE...7]


Cluster Name: Testcluster

Resource Group Name: Maddi_rg
Priority Override Information:
    Primary Instance POL:   Melany
Node                         State
---------------------------- ---------------
Maddi                        OFFLINE
Melany                       ONLINE

[BOTTOM]

F1=Help             F2=Refresh          F3=Cancel           F6=Command
F8=Image            F9=Shell            F10=Exit            /=Find
n=Find Next
```

Рис. 7.6. Состояние группы ресурсов с выводом параметра POL

Внимание! В показанном выше примере обратите внимание на то, что параметр POL установлен.

При каждом перемещении группы ресурсов на другой узел мониторинг приложений приостанавливается на время остановки приложения. После перезапуска приложения на целевом узле мониторинг приложения возобновляется. Дополнительные сведения см. в разделе «Приостановка/возобновление мониторинга приложения».

7.3.4 Расположение, отменяющее приоритет

В версиях до HACMP 5.1 можно было установить атрибут «sticky» для перемещенной группы ресурсов. В HACMP 5.1 атрибут «sticky» отсутствует. Вместо этого можно задать атрибут расположения, отменяющего приоритет (priority override location). Действие атрибута «sticky» в версиях до HACMP 5.1 эквивалентно действию атрибута постоянного расположения, отменяющего приоритет. Атрибут расположения, отменяющего приоритет, указывает целевой узел, на котором имеет место миграция группы ресурсов.

Параметр расположения, отменяющего приоритет (priority override location), в действительности «отменяет» параметры конфигурации перемещения при сбое и возврата после восстановления для группы ресурсов. Для этого параметра могут быть установлены значения **persistent (постоянный)** или **not-persistent (непостоянный)**. Этот параметр непосредственно связан с полем **Persist Across Cluster Reboot** в экране SMIT. Два возможных значения: **true** и **false**. По умолчанию для параметра POL установлено значение false.

Ниже приведена информация по работе атрибута POL в группах ресурсов без одновременного доступа при использовании различных флагов **clRGmove**.

Примечание. Одновременно с помощью clRGmove можно перемещать только одну группу ресурсов.

- При каждом перемещении группы ресурсов без одновременного доступа, при котором используется флаг **-n** для явного указания целевого узла вместо флага **-r** (restoring node priority, восстановление приоритета узла), целевой узел становится расположением, отменяющим приоритет (Priority Override Location). Расположение, отменяющее приоритет, действует до явного использования параметра **-r** для указания целевого узла вместо параметра **-n** при повторном перемещении группы ресурсов вручную.
- При отключении группы ресурсов группа ресурсов остается в отключенном состоянии до тех пор, пока она не будет вручную переведена в подключенное состояние. При ее переводе в подключенное состояние вручную с использованием флага **-n** при указании узла этот узел становится расположением, отменяющим приоритет.

- При переводе группы ресурсов обратно в подключенное состояние с использованием флага -r применяется активный узел с наивысшим приоритетом и расположение, отменяющее приоритет, удаляется из группы ресурсов.

  Для групп ресурсов с одновременным доступом имеет место следующее:

- При переводе группы ресурсов с одновременным доступом в отключенное состояние на всех узлах расположение, отменяющее приоритет, переходит в состояние OFFLINE для всех узлов в группе ресурсов. При переводе группы ресурсов с одновременным доступом в отключенное состояние только на одном узле состояние OFFLINE для группы ресурсов на узле добавляется в список расположений, отменяющих приоритет.
- При переводе группы ресурсов с одновременным доступом в подключенное состояние на всех узлах расположение, отменяющее приоритет, удаляется для всех узлов в группы ресурсов.
- При переводе группы ресурсов с одновременным доступом в подключенное состояние только на одном узле состояние OFFLINE для группы ресурсов на этом узле удаляется из списка расположений, отменяющих приоритет.
- Параметр POL групп ресурсов можно просмотреть с использованием команды clRGinfo -p. Выходные данные этой команды соответствуют результатам выполнения операции с группой ресурсов через SMIT, представленным на рис. 7.6.

7.3.5 Приостановка/возобновление мониторинга приложения

В периоды обслуживания приложения часто возникает необходимость перевести приложение в отключенное состояние, не останавливая службы кластера. Если включен мониторинг приложения, то перед остановкой приложения требуется приостановить мониторинг приложения. В противном случае, когда обнаружится, что приложение отключено, HACMP выполнит предписанные процедуры восстановления, что нежелательно во время обслуживания. Определение мониторов приложения описывается в разделе «Мониторинг приложения».

Чтобы приостановить мониторинг приложения:

1. Введите **smit cl_admin.**

2. В SMIT выберите **HACMP System Management (Управление системой HACMP) -> Suspend/Resume Application Monitoring (Приостановка/возобновление мониторинга приложения) -> Suspend Application Monitoring (Приостановить мониторинг приложения)** и нажмите **Enter**.

Система запросит вас выбрать сервер приложения, для которого этот монитор сконфигурирован. Если используется несколько мониторов приложения, все они приостанавливаются до тех пор, пока вы не возобновите их работу или пока не возникнет событие кластера, вызывающее автоматическое возобновление их работы, как описано выше.

Мониторинг будет приостановлен либо до его возобновления вручную, либо до остановки/перезапуска группы ресурсов.

Чтобы возобновить мониторинг приложения:

1. Введите **smit cl_admin.**

2. В SMIT выберите **HACMP System Management (Управление системой HACMP) -> Suspend/Resume Application Monitoring (Приостановка/возобновление мониторинга приложения) -> Resume Application Monitoring (Возобновить мониторинг приложения)** и нажмите **Enter**.

Выберите соответствующий сервер приложения, связанный с монитором приложения, работу которого требуется возобновить.

Мониторинг приложения остается активным либо до его приостановки вручную, либо до перевода группы ресурсов в отключенное состояние.

7.4 Сценарии

В этом разделе мы рассмотрим следующие распространенные сценарии:

- «горячая замена» сетевой карты PCI;
- загрузка исправлений AIX и HACMP;
- замена зеркального диска LVM;
- обслуживание приложения.

7.4.1 «Горячая замена» сетевой карты PCI

Этот раздел описывает процесс «горячей замены» (hot-plug replacement) сетевой карты PCI с использованием средства C-SPOC «PCI Hot Plug Replace a Network Interface Card».

Специальные аспекты

При выполнении «горячей замены» сетевой интерфейсной карты PCI необходимо учитывать следующее:

- Если сетевой интерфейс, для которого вы выполняете «горячую замену», представляет единственный доступный keepalive-путь на узле, где он находится, вы **должны** отключить HACMP на этом узле, чтобы не допустить разделение кластера при замене интерфейса. Этого можно избежать при наличии рабочей сети, отличной от IP, между узлами кластера.
- SMIT позволяет выполнить постепенное завершение работы (graceful shutdown) на этом узле. При этом можно выполнить «горячую замену» сетевой интерфейсной карты вручную.
- Поддерживается «горячая замена» сетевых интерфейсных карт Ethernet, Token-Ring, FDDI и ATM. Этот процесс не поддерживается на коммуникационных устройствах, отличных от IP.
- Следует вручную записать параметры IP-адреса для сетевого интерфейса, для которого выполняется замена, чтобы подготовиться к незапланированным отказам.
- Не следует пытаться изменять какие-либо параметры конфигурации в ходе выполнения «горячей замены».

SMIT-интерфейс упрощает процесс «горячей замены» сетевой интерфейсной карты PCI. HACMP поддерживает одновременное выполнение «горячей замены» только для одной сетевой карты PCI на узле.

> **Примечание.** Если сетевой интерфейс был в рабочем состоянии до начала процесса замены, то между началом и завершением «горячей замены» интерфейс, для которого выполняется замена, находится в режиме обслуживания. На это время приостанавливается мониторинг связи в сети, пока не будет завершен процесс замены.

Сценарий 1 (только для работающих NIC)

Необходимо следовать приведенной ниже процедуре при «горячей замене» таких компонентов, как:

- работающий сервисный сетевой интерфейс PCI в группе ресурсов с доступным несервисным интерфейсом;
- работающий сервисный сетевой интерфейс PCI не в группе ресурсов с доступным несервисным интерфейсом;
- доступный загрузочный сетевой интерфейс PCI с доступным несервисным интерфейсом.

1. Перейдите на узел, на котором требуется выполнить «горячую замену» сетевой интерфейсной карты PCI.

2. Введите **smit hacmp.**

3. В SMIT выберите **System Management (C-SPOC) -> HACMP Communication Interface Management (Управление коммуникационными интерфейсами HACMP) -> PCI Hot Plug Replace a Network Interface Card (Горячая замена сетевой интерфейсной карты PCI)** и нажмите **Enter**.

К этой панели также можно перейти с использованием быстрого пути **smitty cl_pcihp.**

SMIT отображает список доступных сетевых интерфейсов PCI с возможностью «горячей замены».

4. Выберите сетевой интерфейс, для которого требуется выполнить «горячую замену». Нажмите **Enter**. Сервисный адрес интерфейса PCI переносится на доступный несервисный интерфейс.

5. SMIT предложит физически заменить сетевую интерфейсную карту. После замены карты система запросит подтверждения выполнения замены.

Если вы выберете **Yes (Да)**, сервисный адрес будет перенесен обратно на сетевой интерфейс, для которого была выполнена «горячая замена». В сетях с синонимами сервисный адрес не будет перенесен обратно на первоначальный сетевой интерфейс, а останется синонимом на том же сетевом интерфейсе. «Горячая замена» завершена.

Если вы выберете **No (Нет)**, необходимо вручную установить первоначальные значения параметров интерфейса:

а) выполните команду **drslot**, чтобы вывести PCI-слот из удаленного состояния (removed state);

б) выполите **mkdev** на физическом интерфейсе.

в) используйте команду **ifconfig** вместо **smit chinet, cfgmgr** или **mkdev**, чтобы не допустить конфигурирования повторяющихся IP-адресов или нежелательного загрузочного адреса.

Сценарий 2 (только для работающих NIC)

При «горячей замене» работающего сервисного сетевого интерфейса PCI в группе ресурсов без доступного несервисного интерфейса необходимо следовать приведенной ниже процедуре. Действия пп. 1–3 совпадают с предыдущим сценарием, так что в этом сценарии мы начинаем с быстрого пути **smitty cl_pcihp**.

1. Выберите сетевой интерфейс, для которого требуется выполнить «горячую замену», и нажмите **Enter**.

SMIT предложит указать, следует ли перемещать группу ресурсов на другой узел в процессе замены, чтобы обеспечить ее доступность.

2. Если вы укажете, что это нужно сделать, SMIT предложит переместить группу ресурсов обратно на узел, на котором произошла «горячая замена» после завершения процесса замены.

Если вы не переместите группу ресурсов на другой узел, он будет отключен в продолжение процесса замены.

3. SMIT предложит физически заменить сетевую интерфейсную карту. После замены карты система запросит подтверждения выполнения замены.

Если вы выберите **Yes (Да)**, «горячая замена» будет завершена.

Если вы выберите **No (Нет)**, необходимо вручную установить первоначальные значения параметров интерфейса:

а) выполните команду **drslot**, чтобы вывести PCI-слот из удаленного состояния (removed state);

б) выполите **mkdev** на физическом интерфейсе;

в) используйте команду **ifconfig** вместо **smit chinet**, **cfgmgr** или **mkdev**, чтобы не допустить конфигурирования повторяющихся IP-адресов или нежелательного загрузочного адреса;

г) (если применимо) переместите группу ресурсов обратно на узел, с которого он был перемещен на этапе 2.

Сценарий 3 (только для неработающих NIC)

Необходимо следовать приведенной ниже процедуре при «горячей замене» таких компонентов, как:

- неработающий сервисный сетевой интерфейс PCI в группе ресурсов с доступным несервисным интерфейсом;
- неработающий сервисный сетевой интерфейс PCI не в группе ресурсов с доступным несервисным интерфейсом;
- неработающий загрузочный сетевой интерфейс PCI с доступным несервисным интерфейсом.

Как и в предыдущем сценарии, мы снова начинаем с быстрого пути **smitty cl_pcihp.**

1. Выберите сетевой интерфейс, для которого требуется выполнить «горячую замену», и нажмите **Enter**. SMIT предложит физически заменить сетевую интерфейсную карту.

2. После выполнения физической замены SMIT запросит подтверждение выполнения замены.

Если вы выберете **Yes (Да)**, «горячая замена» будет завершена.

Если вы выберете **No (Нет)**, необходимо вручную установить первоначальные значения параметров интерфейса:

а) выполните команду **drslot**, чтобы вывести PCI-слот из удаленного состояния (removed state);

б) выполите **mkdev** на физическом интерфейсе;

в) используйте команду **ifconfig** вместо **smit chinet**, **cfgmgr** или **mkdev**, чтобы не допустить конфигурирования повторяющихся IP-адресов или нежелательного загрузочного адреса.

«Горячая замена» сетевой интерфейсной карты ATM

Сетевые интерфейсные карты ATM поддерживают использование нескольких логических интерфейсов на одной сетевой интерфейсной карте. «Горячая замена» сетевого интерфейса ATM выполняется так же, как и для других сетевых интерфейсных карт, за исключением следующего:

- Все логические интерфейсы на заменяемой карте, не сконфигурированные для замены и управляемые HACMP, утрачиваются в процессе замены. Они не будут повторно сконфигурированы на новой установленной интерфейсной карте ATM. Все остальные логические интерфейсы на заменяемой сетевой интерфейсной карте ATM, сконфигурированные для замены и управляемые HACMP, восстанавливаются после завершения замены.
- Так как на сетевой интерфейсной карте ATM можно сконфигурировать несколько сервисных интерфейсов, а значит, и несколько групп ресурсов для одного сетевого интерфейса ATM, то при «горячей замене» сетевой интерфейсной карты ATM, через SMIT выполняется процесс поочередного перемещения каждой группы ресурсов на интерфейсе ATM.

Дополнительные сведения см. в гл. 12 руководства **High Availability Cluster Multi-Processing Administration Guide**, SC23-4862-06.

7.4.2 Исправления

Этот раздел описывает установку исправлений (APAR/PTFS) как в AIX, так и в HACMP. Мы рекомендуем загружать исправления и осуществлять обслуживание ежеквартально. Однако опрос клиентов показывает, что чаще эти операции выполняются два

раза в год в сезоны отпусков. В некоторых случаях приходится отклоняться от стандартной практики при возникновении серьезных проблем.

Некоторые исправления AIX можно загружать динамически без перезагрузки системы. Обновления ядра и драйвера устройств часто требуют перезагрузки, так как при их установке запускается **bosboot**. Чтобы определить, необходима ли перезагрузка системы, следует просмотреть файл .toc, создаваемый командой **inutoc** перед установкой исправлений. Файл содержит информацию о наборах файлов (filesets), подобную представленной в примере 7.1.

Пример 7.1. Просмотр файла .toc перед установкой исправлений

```
bos.64bit                    5.3.0.0                    I  b usr,root
#   Base Operating System 64 bit Runtime
bos.INed                     5.3.0.0                    I  N usr,root
#   INed Editor
```

В приведенном примере набор файлов bos.64bit требует перезагрузки, на что указывает символ **b** в четвертом столбце. Символ **N** указывает на то, что перезагрузка необязательна.

Применение исправлений HACMP подобно применению исправлений AIX. Наборы файлов, подлежащие обновлению, указывают, необходимо ли выполнять перезапуск кластера с использованием метода, указанного выше. Если есть неуверенность в последствиях загрузки тех или иных исправлений, следует проконсультироваться с группой поддержки.

При обновлении программного обеспечения AIX или HACMP рекомендуется выполнить следующие действия:

- Создать снимок кластера и сохранить его вне кластера.
- Выполнить резервное копирование операционной системы и данных до выполнения обновления. Подготовьте план возврата в случае возникновения проблем при обновлении.
- **Всегда** выполняйте первый запуск в тестовом кластере.
- Если возможно, используйте обновление дисков.
- Следуйте этим же общим правилам при применении исправлений приложения; следуйте также указаниям для приложения.

Общая процедура применения исправлений AIX и HACMP имеет следующий вид:

- Примените (apply), не фиксируя (commit), APAR на дежурном узле (standby node).
- Выполняйте перемещение при сбое (постепенную остановку с переносом ресурсов, graceful shutdown with takeover) на дежурном компьютере (standby machine).
- Примените (apply) APAR на основном узле (primary node).

Перед применением исправлений на дежурном узле (standby node) необходимо остановить службы кластера. После применения исправлений при необходимости нужно перезагрузить узел. Для реинтеграции узла в кластер в качестве дежурного узла следует перезапустить службы кластера.

Для того чтобы применить исправления на рабочих узлах, следует выполнить постепенную остановку служб кластера с переносом ресурсов (gracefully with takeover). После завершения переноса ресурсов службы кластера должны продолжить процесс остановки. После полной остановки служб кластера следует применить исправления, при необходимости перезагрузить узел и перезапустить службы кластера. В зависимости от политики перемещения при сбое для группы ресурсов, при реинтеграции узла в кластер, он может «подхватить» ресурсы. Если этого не произошло, можно использовать C-SPOC для перемещения группы ресурсов обратно на первоначальный узел.

7.4.3 Хранение

В большинстве современных общих сред хранения применяется определенный уровень технологии RAID для защиты и дублирования данных. При использовании устройств RAID (1, 5 или 10) отказы отдельных дисков обычно не требуют выполнения обслуживания AIX LVM. Все требуемые процедуры часто являются внешними по отношению к узлам кластера и не влияют на сам кластер. Однако если защита осуществляется с использованием зеркального отображения (mirroring) LVM, то необходимо выполнять процедуры обслуживания LVM.

C-SPOC содержит средство поддержки при замене отказавшего диска с зеркальным отображением LVM. Это средство имеет название **Cluster Disk Replacement** и выполняет все необходимые операции LVM по замене диска с зеркальным отображением LVM. Для использования этого средства необходимо следующее:

- наличие привилегий пользователя «root»;
- должно осуществляться зеркальное отображение отказавшего диска (и желательно всей группы томов);

требуемый резервный диск должен быть доступен для всех узлов; ему уже должен быть назначен PVID, который должен отображаться на всех узлах при вводе команды **lspv.**

При физической замене существующего диска следует удалить старый диск и установить на его место новый диск. Это, конечно же, предполагает, что привод допускает «горячую замену», что в настоящее время является стандартом.

Для замены диска с зеркальным отображением через C-SPOC проделайте следующее:

1. Найдите отказавший диск. Запишите PVID для физического тома (диска).

2. Введите **smitty cl_admin.**

3. В SMIT выберите **HACMP Physical Volume Management (Управление физическими томами HACMP) -> Cluster Disk Replacement (Замена дисков кластера)** и нажмите **Enter**.

SMIT выводит список дисков, входящих в группы томов, содержащихся в группах ресурсов кластера. Группа томов, в которой расположен отказавший диск, должна содержать два диска или более. Список включает группу томов, hdisk, PVID диска и опорный узел кластера (reference cluster node; обычно подразумевается узел кластера, на котором активизирована группа томов).

4. Выберите диск, для которого требуется выполнить замену (**исходный диск; source disk**) и нажмите **Enter**. SMIT выводит список доступных дисков, которым назначен PVID и которые можно использовать для замены (для замены отказавшего диска подходят только диски такого же или большего объема).

5. Выберите требуемый диск для замены **(целевой диск; destination disk)** и нажмите **Enter**. SMIT выводит выбранные вами параметры с двух предыдущих панелей.

6. Нажмите **Enter** для продолжения или **Cancel** для отмены процесса замены дисков. Появится предупреждение, сообщающее о том, что продолжение операции удалит всю информацию, хранящуюся на целевом диске.

7. Нажмите **Enter** для продолжения или **Cancel** для отмены. SMIT отображает панель состояния команды и отображает каталог восстановления **replacepv**. Если при отказе дисковой конфигурации требуется продолжить замену дисков, необходимо сконфигурировать целевой диск вручную. Учтите, что при отмене процедуры на данном этапе целевой диск может быть сконфигурирован на нескольких узлах в кластере.

Утилита **replacepv** выполняет обновление группы томов, используемой в процессе замены дисков (только на опорном узле).

> **Примечание.** В процессе выполнения команды SMIT сообщает имя каталога восстановления, который следует использовать при отказе replacepv. Запишите эту информацию, так как она необходима в процессе восстановления.

Выполняется конфигурирование целевого диска на всех узлах в группе ресурсов.

В случае неудачного импортирования обновленной группы томов узлом в группе ресурсов можно использовать средство C-SPOC **Import a Shared Volume Group**, как показано на рис. 8.20.

C-SPOC не удаляет информацию об отказавшем дисковом устройстве с узлов кластера. Это нужно сделать вручную командой **rmdev -dl <devicename>.**

Прочие операции C-SPOC, связанные с LVM, описываются в разделе 8.4, «Управление общим хранением».

7.4.4 Приложения

Конечно же, каждое приложение имеет свои особенности, однако в большинстве случаев для выполнения обслуживания приложения требуется перевести его в отключенный режим. Это можно сделать несколькими способами. Наиболее оптимальный метод для каждой конкретной среды зависит от общей конфигурации кластера.

В многоуровневой среде (multi-tier environment), где сервер приложения зависит от базы данных, и следует производить обслуживание базы данных обычно нужно останавливать как приложение, так и базу данных. Чаще всего хотя бы база данных находится в кластере. При использовании зависимостей группы ресурсов сервер приложения можно легко реализовать в этом же кластере.

Кроме того, для того чтобы сократить общее время простоя приложения, часто обслуживание приложения сначала выполняется на нерабочих узлах. Традиционно под

такими узлами подразумеваются дежурные узлы (standby nodes), однако очень редко резервные узлы (backup/fallover node) используются только как дежурные. В таких случаях следует учитывать текущую рабочую нагрузку или приложения, выполняющиеся на этом узле, чтобы свести к минимуму неблагоприятное воздействие процесса обслуживания. Рекомендуется прежде выполнить тестирование в тестовом кластере.

В большинстве случаев останавливать службы кластера не требуется. Можно просто перевести группу ресурсов в отключенное состояние, как описано в разделе «Перевод группы ресурсов в отключенное состояние через SMIT». Если общая группа томов должна быть подключена во время обслуживания, можно просто приостановить мониторинг приложения и выполнить скрипт остановки сервера приложения, чтобы перевести приложение в отключенное состояние. Однако при этом сервисный IP-адрес останется подключенным, что может быть нежелательным.

В среде, где несколько групп ресурсов и/или несколько приложений выполняются на одном узле, остановка служб кластера на локальном узле может быть невозможна. Помните о возможных последствиях остановки служб кластера на узле, на котором выполняется обслуживание приложения.

Если во время обслуживания возникнет фатальная ошибка, вызывающая отказ, произойдет перемещение при сбое. Это может быть нежелательным, если прежде не было выполнено обслуживание на резервных узлах и/или если не завершено обслуживание на локальном узле. Хотя такие ситуации возникают редко, вероятность их возникновения все-таки существует, и это нужно учитывать.

В случае отказа другого рабочего узла в процессе обслуживания может произойти успешное перемещение при сбое на локальном узле без неблагоприятных эффектов. Если это нежелательно и существует несколько групп ресурсов, может потребоваться сначала переместить другие группы ресурсов на другой узел и затем остановить службы кластера на локальном узле.

Если при использовании постоянных адресов выполняется остановка служба кластера, защита от переключения локального адаптера (local adapter swap protection) отключается. Хотя опять же такие ситуации редки, существует вероятность того, что в случае отказа содержащей адрес сетевой карты при использовании постоянного адреса для выполнения обслуживания подключение будет разорвано.

После выполнения обслуживания приложения **всегда** следует выполнять повторное тестирование кластера. В зависимости от способа остановки приложения потребуется либо перезапустить службы кластера, обратно подключить группу ресурсов через C-SPOC, либо вручную запустить скрипт запуска сервера приложения и при необходимости продолжить мониторинг приложения.

8

Управление кластером

В этой главе описывается управление кластером HACMP, а также советы и рекомендации по администрированию кластера, включая следующее:

- общее описание C-SPOC;
- наборы файлов (File collections);
- администрирование пользователей;
- управление общим хранилищем;
- синхронизация времени;
- верификация и синхронизация кластера;
- мониторинг кластера.

8.1 CSPOC DP

C-SPOC (Cluster Single Point of Control) – очень полезный инструмент, позволяющий осуществлять управление всем кластером с одного узла. Он содержит средства выполнения общих задач администрирования в масштабе кластера с любого активного узла в кластере. Использование C-SPOC позволяет сократить перерыв в обслуживании, связанный с администрированием кластера.

Кластеры высокой доступности требуют особого внимания к системному администрированию. Мы настоятельно рекомендуем придерживаться процедуры управления изменениями.

Прежде чем перейти к описанию подробностей управления кластером, мы хотим описать общие рекомендации в администрировании кластера.

- Там, где это возможно, используйте C-SPOC для внесения изменений в кластер.
- Документируйте стандартные рабочие процедуры (например, завершение работы, запуск, увеличение размера файловой системы).
- Доступ к паролю пользователя «root» должны иметь только опытные администраторы HACMP.
- Прежде чем вносить изменения, всегда создавайте снимок существующей конфигурации.
- Осуществляйте регулярный мониторинг кластера.

8.1.1 Общее описание C-SPOC

Функционирование C-SPOC осуществляется с использованием собственного набора команд администрирования кластера, доступных через меню SMIT. Команды располагаются в каталоге /usr/es/sbin/cluster/cspoc. Для выполнения команд на удаленных узлах используется демон коммуникаций кластера clcomdES. Если этот демон не запущен, выполнение команд невозможно, что вызывает отказ операции C-SPOC.

> **Примечание.** Начиная с HACMP 5.3 процесс диспетчера кластера (Cluster Manager; clstrmgrES) запускается из процесса init, так что он всегда выполняется, независимо от того, запущен ли кластер. Под активным узлом понимается узел, на котором помимо clstrmgrES запущены другие необходимые службы кластера.

Отказ операций C-SPOC происходит в тех случаях, когда на момент выполнения не работает целевой узел или недоступен требуемый ресурс. Кроме того, необходимо, чтобы кластер был правильно настроен, т. е. чтобы все узлы в кластере могли связаться друг с другом.

При возникновении отказа узла во время операции C-SPOC на терминал выводится ошибка; сообщения об ошибке и прочая информация о ней записываются в журнал C-SPOC (по умолчанию в /tmp/cspoc.log). При возникновении каких-либо проблем в работе C-SPOC следует просматривать этот журнал. Дополнительные сведения о журналах HACMP см. в разделе 8.7.5, «Файлы журналов».

8.1.2 Меню C-SPOC SMIT

Доступ в меню C-SPOC SMIT осуществляется путем вызова **smit hacmp > System Management (C-SPOC)** или с использованием быстрого пути **smit cl_admin**. Основные функции или подменю C-SPOC представлены ниже в том порядке, в котором они отображаются в главном меню SMIT C-SPOC.

- **Manage HACMP Services (Управление службами HACMP).** Эта часть содержит меню запуска, остановки кластера на одном узле или на выбранных узлах, а также меню вывода запущенных служб кластера на локальном узле. Дополнительные сведения см. в разделе 7.2, «Запуск и остановки кластера», и в разделе 8.7.2, «Утилиты проверки состояния и служб кластера».
- **HACMP Communication Interface Management (Управление коммуникационными интерфейсами HACMP).**
- **HACMP Resource Group and Application Manipulation (Управление группами ресурсов HACMP и приложениями).** Эта часть содержит меню и утилиты для управления группами ресурсов кластера и мониторинга приложений, а также инструменты определения доступности приложений. Дополнительные сведения об управлении группами ресурсов см. в разделе «Управление группами ресурсов и приложениями», о мониторинге приложений – в разделе 8.7.7, «Мониторинг приложений», и об инструменте анализа приложений – в разделе 8.7.8, «Управление доступностью приложения».
- **HACMP Log Viewing and Management (Просмотр и управление журналами HACMP).** Эта часть содержит утилиты для вывода содержимого файлов журналов и для установки значений некоторых параметров файлов журналов, в частности каталога расположения, уровня отладки и формата (стандартный формат – html). Дополнительные сведения см. в разделе 8.7.5, «Файлы журналов».
- **HACMP File Collection Management (Управление наборами файлов в архитектуры HACMP).** Эта часть содержит утилиты, обеспечивающие возможность синхронизации в масштабе кластера через функции наборов файлов, реализованные в C-SPOC. Утилита синхронизации основана на функции наборов файлов. Набор файлов (file collection) представляет собой определенное пользователем собрание файлов. Дополнительные сведения см. в разделе 8.2, «Наборы файлов».
- **HACMP Security and Users Management (Безопасность и управление пользователями HACMP).** Эта часть содержит меню и утилиты для установки различных параметров безопасности, а также для управления пользователями, группами и паролями в кластере. Дополнительные сведения о безопасности см. в главе 9, «Безопасность кластера», об управлении пользователями – в разделе 8.3, «Управление пользователями».
- **HACMP Logical Volume Management (Управление логическими томами HACMP).** Эта часть содержит утилиты, обеспечивающие управление общими группами томов, общими логическими томами и общими файловыми системами в масштабе кластера. Дополнительные сведения по этим темам см. в разделе 8.4.2, «Диспетчер логических томов C-SPOC».

- **HACMP Concurrent Logical Volume Management (Управление логическими томами HACMP с одновременным доступом).** Эта часть содержит утилиты, обеспечивающие управление группами томов с одновременным доступом, логическими томами с одновременным доступом и файловыми системами с одновременным доступом в масштабе кластера. Дополнительные сведения по этим темам см. в разделе 8.4.3, «Управление логическими томами с одновременным доступом через C-SPOC».
- **HACMP Physical Volume Management (Управление физическими томами HACMP).** Эта часть содержит утилиты для управления физическими томами в масштабе кластера, в частности для выполнения операций добавления, удаления и замены физических томов. Также она содержит поддержку устройств datapath и зеркального отображения через LVM (cross-LVM mirroring[1]). Дополнительные сведения по этим темам см. в разделе 8.4.4, «Управление физическими томами C-SPOC».
- **Open a SMIT Session on a Node (Открыть сеанс SMIT на узле).** Это средство обеспечивает возможность открытия базового окна SMIT на любом активном узле в кластере. Из локального меню SMIT можно инициировать любое действие SMIT для любого узла в кластере.

8.2 Наборы файлов

HACMP обеспечивает возможности синхронизации файлов через функции наборов файлов (file collections) C-SPOC в масштабе кластера. Утилита синхронизации файлов основана на функции наборов файлов. Набор файлов представляет собой определенное пользователем собрание файлов. Можно добавлять или удалять файлы из наборов файлов и задавать время синхронизации этих файлов в HACMP.

HACMP обеспечивает три способа распространения этих файлов:

- Вручную. Синхронизацию файлов вручную можно осуществлять в любое время. Файлы копируются с локального узла на удаленный узел.
- Автоматически во время верификации и синхронизации кластера. Распространение файлов осуществляется с узла, с которого запущена верификация HACMP.
- Автоматически при обнаружении изменений. HACMP периодически проверяет набор файлов на всех узлах, и при обнаружении изменений в файле выполняет синхронизацию файла в кластере. Можно установить таймер частоты проверки наборов файлов в HACMP.

HACMP сохраняет разрешения, владение и отметку времени файла и распространяет их на удаленные узлы. В набор файлов можно включать только обычные файлы; нельзя добавлять символические ссылки, каталоги, каналы, сокеты, файлы устройств (/dev/*), файлы из каталога /proc и ODM-файлы из каталогов /etc/objrepos/* и /etc/es/objrepos/*. Следует всегда использовать полные пути. Каждый файл может быть добавлен только в один набор файлов, за исключением тех файлов, которые автоматически добавляются в набор HACMP_Files. Файлы не обязательно должны существовать на удаленных узлах, HACMP создает их при первой синхронизации. Файлы нулевой длины или несуществующие файлы не распространяются с локального узла.

[1] Имеется в виду Cross-Site LVM Mirroring. *Примеч. науч. ред.*

HACMP создает резервную копию измененных файлов во время синхронизации на всех узлах. Эти резервные копии хранятся в каталоге /var/hacmp/filebackup. Сохраняется только одна последняя версия, и восстановление осуществляется только вручную.

Журналы функции наборов файлов хранятся в файле /var/hacmp/log/clutils.log.

> **Важно!** Вы сами должны обеспечить наличие актуальных и неповрежденных файлов на локальном узле (с которого запускается распространение файлов).

8.2.1 Предопределенные наборы файлов

По умолчанию HACMP содержит два набора файлов: **Configuration_Files** и **HACMP_Files**. Ни для одного из них автоматическая синхронизация не устанавливается по умолчанию. Ее можно включить, установив в SMIT в меню **Change/Show a file collection (Изменение/вывод набора файлов)** для одной из опций «Propagate files during cluster synchronization» («Распространять файлы во время синхронизации кластера») или «Propagate files automatically when changes are detected» («Распространять файлы при обнаружении изменений») значение **Yes (Да)**; см. также раздел «Изменение набора файлов».

Configuration_Files

Этот набор содержит основные файлы конфигурации AIX:

- /etc/hosts;
- /etc/services;
- /etc/snmpd.conf;
- /etc/snmpdv3.conf;
- /etc/rc.net;
- /etc/inetd.conf;
- /usr/es/sbin/cluster/netmon.cf;
- /usr/es/sbin/cluster/etc/clhosts;
- /usr/es/sbin/cluster/etc/rhosts.

Можно легко добавлять и удалять файлы из этих наборов. Дополнительные сведения см. в разделе «Добавление файлов в набор файлов».

HACMP_Files

Этот набор файлов автоматически собирает все пользовательские скрипты из конфигурации HACMP. При определении любого из нижеперечисленных файлов в конфигурации кластера файлы автоматически добавляются в набор файлов HACMP_Files.

- Скрипт запуска сервера приложения.
- Скрипт остановки сервера приложения.
- Скрипт уведомления о событии.

- Скрипт предобработки события.
- Скрипт постобработки события.
- Скрипт восстановления после ошибок событий.
- Скрипт уведомления монитора приложения.
- Скрипт очистки монитора приложения.
- Скрипт перезапуска монитора приложения.
- Файл текста сообщения на пейджер.
- Скрипты запуска и остановки каналов SNA.
- Скрипты запуска и остановки каналов X.25.
- Скрипт запуска поддержки HA Tape.
- Скрипт остановки поддержки HA Tape.
- Пользовательская программа восстановления после ошибок событий.
- Настраиваемый скрипт метода создания снимков.

Посмотрим пример работы этой функции. Наш кластер содержит сервер приложения DB2. Его скрипт запуска – /usr/ha/db2.start, скрипт остановки – /usr/ha/db2.stop. Также имеется настраиваемый скрипт постобработки события node_up под

```
                         Change/Show a File Collection

  Type or select values in entry fields.
  Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
    File Collection Name                               HACMP_Files
    New File Collection Name                          []
    File Collection Description                       [User-defined scripts >
    +-------------------------------------------------------------------+
    ¦                            Collection files                       ¦
    ¦                                                                   ¦
    ¦ The value for this entry field must be in the                     ¦
    ¦ range shown below.                                                ¦
    ¦ Press Enter or Cancel to return to the entry field,               ¦
    ¦ and enter the desired value.                                      ¦
    ¦                                                                   ¦
    ¦   /usr/ha/db2.start                                               ¦
    ¦   /usr/ha/db2.stop                                                ¦
    ¦   /usr/ha/post.node_up                                            ¦
    ¦                                                                   ¦
    ¦ F1=Help                 F2=Refresh              F3=Cancel         ¦
  F1¦ F8=Image                F10=Exit                Enter=Do          ¦
  F5¦ /=Find                  n=Find Next                               ¦
  F9+-------------------------------------------------------------------+
```

Рис. 8.1. Пример набора файлов HACMP_Files

названием /usr/ha/post.node_up. Эти три файла автоматически добавляются в набор файлов HACMP_Files при их определении во время конфигурирования HACMP. Можно выполнить следующие действия:

1. Откройте меню управления наборами файлов SMIT HACMP: **smit cm_filecollection_mgt.**

2. Выберите **Change/Show a File Collection (Изменение/вывод набора файлов)**.

3. Из всплывающего списка выберите **HACMP_Files** и нажмите Enter.

4. Перейдите в поле **Collection files (Файлы набора)** и нажмите F4. Как можно увидеть на скриншоте SMIT на рис. 8.1, скрипты запуска и остановки приложения, а также команда постобработки событий автоматически добавляются в этот набор файлов.

> **Внимание!** Нельзя добавлять (или удалять) файлы непосредственно в этот набор файлов. Если вы начинаете использовать набор файлов HACMP_Files, убедитесь в том, что скрипты будут корректно работать на всех узлах.

Если вы не хотите выполнять синхронизацию всех пользовательских скриптов или если они различаются на разных узлах, то следует отключить этот набор файлов и создать другой набор файлов, включающий только требуемые файлы.

8.2.2 Управление наборами файлов

В этом разделе описывается создание, изменение и удаление наборов файлов.

Добавление набора файлов

1. Запустите SMIT: **smitty->Communications Applications and Services (Коммуникационные приложения и службы)->HACMP for AIX (HACMP для AIX)**; выберите **System Management (CSPOC)->HACMP File Collection Management (Управление наборами файлов HACMP)**.

Средство управления наборами файлов HACMP также можно вызвать путем ввода **smit cm_filecollection_menu**

2. Выберите **Manage File Collections (Управление наборами файлов)**.

3. Выберите **Add a File Collections (Добавление наборов файлов)**.

4. Введите требуемую информацию (рис. 8.2):

- **File Collection Name (Имя набора файлов).** Уникальное имя для набора файлов.
- **File Collection Description (Описание набора файлов).** Короткое описание данного набора файлов.
- **Propagate files during cluster synchronization? (Распространять файлы во время синхронизации кластера?).** Если установлено значение **Yes (Да)**, HACMP распространяет этот набор файлов во время синхронизации кластера. Это решение является удобным для файлов, связанных с кластером, в частности происходит автоматическая синхронизация скриптов запуска приложений после внесения изменений в конфигурацию кластера.

```
                                Add a File Collection

  Type or select values in entry fields.
  Press Enter AFTER making all desired changes.

                                                          [Entry Fields]
  * File Collection Name                                 [app_files]
    File Collection Description                          [Application config fi>
    Propagate files during cluster synchronization?       yes +
    Propagate files automatically when changes are det    no +
    ected?



  F1=Help             F2=Refresh          F3=Cancel           F4=List
  F5=Reset            F6=Command          F7=Edit             F8=Image
  F9=Shell            F10=Exit            Enter=Do
```

Рис. 8.2. Добавление набора файлов

- **Propagate files automatically when changes are detected? (Автоматически распространять файлы при обнаружении изменений?).** Если установлено значение **Yes (Да)**, HACMP будет регулярно проверять файлы в этом наборе, и при обнаружении изменений в любом из них выполняется их повторное распространение.

Если для обеих опций установлено значение **No (Нет)**, то автоматическая синхронизация не выполняется.

Изменение набора файлов

1. Запустите средство HACMP File Collection Management (Управление наборами файлов HACMP) с использованием быстрого пути **smit cm_filecollection_menu.**

2. Выберите **Manage File Collections (Управление наборами файлов)**.

3. Выберите **Change/Show a File Collections (Изменение/вывод наборов файлов)**.

4. Выберите набор файлов из всплывающего списка.

5. Теперь можно изменять следующую информацию (см. экран SMIT на рис. 8.3):

- File collection name (Имя набора файлов).
- File Collection Description (Описание набора файлов).
- Propagate files during cluster synchronization? (Распространять файлы во время синхронизации кластера?).
- Propagate files automatically when changes are detected? (Автоматически распространять файлы при обнаружении изменений?).

```
                        Change/Show a File Collection

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
  File Collection Name                                 Configuration_Files
  New File Collection Name                            []
  File Collection Description                         [AIX and HACMP configu>
  Propagate files during cluster synchronization?      no +
  Propagate files automatically when changes are det   no +
  ected?
  Collection files +



F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Рис. 8.3. Изменение набора файлов

- **Collection files (Файлы набора).** Нажмите F4, чтобы просмотреть список файлов из этого набора.

  Описание полей см. в разделе «Добавление набора файлов».

Удаление набора файлов

1. Запустите средство HACMP File Collection Management (Управление наборами файлов HACMP) с использованием быстрого пути **smit cm_filecollection_menu.**

2. Выберите **Manage File Collections (Управление наборами файлов)**.

3. Выберите **Remove a File Collections (Удаление наборов файлов)**.

4. Выберите набор файлов из всплывающего списка.

5. Еще раз нажмите Enter, чтобы подтвердить удаление набора файлов.

Изменение таймера автоматического обновления

Этот раздел описывает настройку таймера, устанавливающего частоту выполнения проверок изменений в наборах файлов. Для набора файлов можно установить только один таймер.

1. Запустите средство HACMP File Collection Management (Управление наборами файлов HACMP) с использованием быстрого пути **smit cm_filecollection_menu.**

2. Выберите **Manage File Collections (Управление наборами файлов)**.

3. Выберите **Change/Show Automatic Update Time (Изменение/вывод времени автоматического обновления)**.

4. Выберите набор файлов из всплывающего списка.

5. Задайте значение поля **Automatic File Update Time (Время автоматического обновления файлов)** в минутах. Поле может принимать значения от 10 мин до 1 440 мин (один день).

Добавление файлов в набор файлов

1. Запустите средство HACMP File Collection Management (Управление наборами файлов HACMP) с использованием быстрого пути **smit cm_filecollection_menu.**

2. Выберите **Manage File Collections (Управление наборами файлов)**.

3. Выберите **Add Files to a File Collection (Добавление файлов в набор файлов)**.

4. Выберите набор файлов из всплывающего списка и нажмите Enter.

5. На панели SMIT можно выбрать текущий список файлов или добавить новые файлы (см. рис. 8.4):

- Чтобы получить список текущих файлов в данном наборе, нужно перейти в поле **Collection files (Файлы набора)** и нажать F4.
- Чтобы добавить новые файлы, нужно перейти в поле **New files (Новые файлы)** и ввести имя файла, который требуется добавить в набор файлов. Можно добавить только один файл зараз. Имя файла должно начинаться с «/». Здесь можно указывать только обычные файлы; нельзя добавить символическую ссылку, каталог, канал, сокет, файл устройства (/dev/*), файлы из каталога /proc и ODM-файлы из каталогов /etc/objrepos/* и /etc/es/objrepos/*.

```
                        Add Files to a File Collection

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
  File Collection Name                                 app_files
  File Collection Description                          Application Config Fi>
  Propagate files during cluster synchronization?      no
  Propagate files automatically when changes are det   no
  ected?
  Collection files +
* New files                                           [/db2/app.config] /



F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Рис. 8.4. Добавление файлов в набор файлов

Внимание! Нельзя добавить файлы в набор HACMP_Files.

Удаление файлов из набора файлов

1. Запустите средство HACMP File Collection Management (Управление наборами файлов HACMP) с использованием быстрого пути **smit cm_filecollection_menu.**

2. Выберите **Manage File Collections (Управление наборами файлов)**.

3. Выберите **Remove Files from a File Collection (Удаление файлов из набора файлов)**.

4. Выберите набор файлов из всплывающего списка и нажмите Enter.

5. Выберите один или несколько файлов из списка и нажмите Enter. (рис. 8.5).

6. Еще раз нажмите Enter для подтверждения.

```
                        Manage File in File Collections

 Move cursor to desired item and press Enter.

   Add Files to a File Collection
   Remove Files from a File Collection


   +-----------------------------------------------------------------------+
   ¦       Select one or more files to remove from this File Collection    ¦
   ¦                                                                       ¦
   ¦ Move cursor to desired item and press F7.                             ¦
   ¦     ONE OR MORE items can be selected.                                ¦
   ¦ Press Enter AFTER making all selections.                              ¦
   ¦                                                                       ¦
   ¦   /db2/appconfig1.txt                                                 ¦
   ¦   /db2/databases.conf                                                 ¦
   ¦   /db2/appdata.conf                                                   ¦
   ¦                                                                       ¦
   ¦ F1=Help                 F2=Refresh              F3=Cancel             ¦
   ¦ F7=Select               F8=Image                F10=Exit              ¦
 F1¦ Enter=Do                /=Find                  n=Find Next           ¦
 F9+-----------------------------------------------------------------------+
```

Рис. 8.5. Удаление файлов из наборов файлов

Внимание! Нельзя таким способом удалить файлы из набора файлов HACMP_Files.

Распространение файлов в наборах файлов вручную

Можно выполнить синхронизацию любого набора файлов вручную (рис. 8.6).

1. Запустите средство HACMP File Collection Management (Управление наборами файлов HACMP) с использованием быстрого пути **smit cm_filecollection_menu.**

2. Выберите **Propagate Files in File Collections (Распространение файлов в наборах файлов)**.

3. Выберите набор файлов из всплывающего списка и нажмите Enter.

4. Еще раз нажмите Enter для подтверждения.

```
                                COMMAND STATUS

Command: OK            stdout: yes           stderr: no

Before command completion, additional instructions may appear below.

Manual file collection propagation called.
The following file collections will be processed:
app_files
Starting file propagation to remote node p650n01.
Successfully propagated file /db2/appconfig1.txt to node p650n01.
Successfully propagated file /db2/databases.conf to node p650n01.
Successfully propagated file /db2/appdata.conf to node p650n01.
Total number of files propagated to node p650n01: 3




F1=Help             F2=Refresh          F3=Cancel           F6=Command
F8=Image            F9=Shell            F10=Exit            /=Find
n=Find Next
```

Рис. 8.6. Распространение файлов в наборах файлов вручную

8.3 Администрирование пользователей

В кластере HACMP необходимо постоянно осуществлять синхронизацию идентификаторов и паролей пользователей. Если идентификаторы пользователя и группы неодинаковы на всех узлах кластера, приложение не сможет работать и пользователи не смогут получить доступ к своим файлам в общем хранилище. Кроме того, мы рекомендуем также выполнять синхронизацию паролей; тогда в случае перемещения при сбое пользователи смогут без труда войти в систему, не тратя время на то, чтобы определить, какие у них пароли на резервном узле.

При синхронизации пользователей и паролей следует учитывать следующее:

- Использование C-SPOC: HACMP содержит набор утилит в C-SPOC, обеспечивающих простоту администрирования пользователей. В разделе «Администрирование

пользователей и групп через C-SPOC» описывается функция управления пользователями и группами через C-SPOC.

- Network Information Server (NIS). В настоящее время используется реже, чем несколько лет назад; представляет мощное решение для централизованного управления пользователями. Конфигурирование NIS в HACMP очень хорошо описано в руководстве **HACMP/ES Customization Examples**, SG24-4498.
- LDAP. Наилучшее решение для управления большим количеством пользователей в сложной среде. LDAP можно легко настроить на совместную работу с HACMP. Дополнительные сведения о LDAP см. в руководстве **Understanding LDAP-Design and Implementation**, SG24-4986.

8.3.1 Администрирование пользователей и групп через CSPOC

HACMP содержит инструменты C-SPOC, обеспечивающих простоту администрирования пользователей, групп и паролей. В них реализованы следующие функции:

- добавление пользователя;
- вывод списка пользователей;
- изменение атрибутов пользователя;
- удаление пользователя;
- изменение пароля;
- добавление, вывод списка, изменение и удаление групп.

Добавление пользователя

Чтобы добавить пользователя на всех узлах в кластере:

- 1. Запустите SMIT: **smit.**
- Выберите **Communications Applications and Services (Коммуникационные приложения и службы)**.
- Выберите **HACMP for AIX (HACMP для AIX)**.
- Выберите **System Management (C-SPOC)**.
- Выберите **HACMP Security and Users Management (Безопасность и управление пользователями HACMP)**.

  По-другому инструмент HACMP Security and Users Management (Безопасность и управление пользователями HACMP) можно запустить путем ввода быстрого пути **smit cl_usergroup.**

  2. Выберите **Users in a HACMP Cluster (Пользователи в кластере HACMP)**.

  3. Выберите **Add a User to the Cluster (Добавление пользователя в кластер)**.

4. Выберите узлы, на которых требуется создать пользователей. Если оставить поле **Select Nodes by Resource Group (Выбрать узлы по группе ресурсов)** пустым, пользователь будет создан на всех узлах. Если выбрать группу ресурсов, пользователь будет создан только в подмножестве узлов, относящемся к этой группе ресурсов. В случае кластера из двух узлов следует оставить это поле пустым.

Если кластер содержит более двух узлов, можно создать пользователей в подмножестве узлов или на всех узлах. В нашем примере кластера из четырех узлов, описанного в разделе 6.2, «Конфигурация из четырех узлов», имеется две группы ресурсов, привязанных только к узлам node1 и node2 (rg1 и rg2). Если вы хотите создать пользователя только на этих двух узлах (например, чтобы они могли применять приложение на узлах node1 и node2, но не могли выполнить вход в систему на узлах node3 и node4), то следует выбрать соответствующую группу ресурсов.

Табл. 8.1 представляет таблицу перекрестных ссылок между пользователями, группами ресурсов и узлами. Она показывает, что в нашем примере пользователь app1adm будет создан на всех узлах [поле Select Nodes by Resource Group (Выбрать узлы по группе ресурсов) оставлено пустым], тогда как пользователи «karesz» и «galamb» будут созданы только на узлах node1 и node2 [следует выбрать «rg2» в поле Select Nodes by Resource Group (Выбрать узлы по группе ресурсов)] (рис. 8.7).

Таблица 8.1. Таблица перекрестных ссылок пользователей, групп ресурсов и узлов

Группа ресурсов	Узлы	Пользователи
rg1	node1, node2	db2adm, db2inst, db2user
rg2	node2, node1	galamb, karesz
rg3	node3, node4	matyi, adrien
rg4	node4, node3, node2, node1	app1adm, app1user

```
                          Add a User to the Cluster

  Type or select a value for the entry field.
  Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
    Select nodes by Resource Group                    []                  +
     *** No selection means all nodes! ***


    +-------------------------------------------------------------------+
    ¦                     Select nodes by Resource Group                ¦
    ¦                  *** No selection means all nodes! ***            ¦
    ¦ Move cursor to desired item and press Enter.                      ¦
    ¦                                                                   ¦
    ¦   rg1                                                             ¦
    ¦   rg2                                                             ¦
    ¦   rg3                                                             ¦
    ¦   rg4                                                             ¦
    ¦                                                                   ¦
    ¦ F1=Help                 F2=Refresh              F3=Cancel         ¦
  F1¦ F8=Image                F10=Exit                Enter=Do          ¦
  F5¦ /=Find                  n=Find Next                               ¦
  F9+-------------------------------------------------------------------+
```

Рис. 8.7. Выбор узлов по группе ресурсов

5. Создайте пользователя. Следует ввести имя пользователя и прочую релевантную информацию, требуемую при создании обычного пользователя. Здесь можно задать идентификатор пользователя, после чего C-SPOC проверит доступность заданного идентификатора на всех узлах. Если поле User ID оставить пустым, создаваемому пользователю на всех будет назначен произвольный доступный идентификатор (рис. 8.8).

```
                              Add a User to the Cluster

    Type or select values in entry fields.
    Press Enter AFTER making all desired changes.

    [TOP]                                                   [Entry Fields]
      Select nodes by resource group
       *** No selection means all nodes! ***

    * User NAME                                           [matyi]
      User ID                                             [] #
      ADMINISTRATIVE USER?                                 false +
      Primary GROUP                                       [] +
      Group SET                                           [] +
      ADMINISTRATIVE GROUPS                               [] +
      Another user can SU TO USER?                         true +
      SU GROUPS                                           [ALL] +
      HOME directory                                      [/home/matyi]
      Initial PROGRAM                                     []
      User INFORMATION                                    [Mr. Matyi]
    [MORE...33]

    F1=Help             F2=Refresh          F3=Cancel           F4=List
    F5=Reset            F6=Command          F7=Edit             F8=Image
    F9=Shell            F10=Exit            Enter=Do
```

Рис. 8.8. Создание пользователя на всех узлах кластера

> **Внимание!** При создании пользователя, домашний каталог которого должен находиться в общей файловой системе, C-SPOC не выполняет проверку подключения файловой системы. В этом случае C-SPOC создает домашний каталог пользователя непосредственно под пустой точкой подключения (mount point) общей файловой системы. Это можно исправить путем перемещения домашнего каталога в общую файловую систему.
> Если домашний каталог пользователя находится на общем томе, пользователь может выполнить вход в систему только с узла, на котором файловая система уже подключена.

Вывод списка пользователей кластера

Для вывода списка пользователей кластера проделайте следующее:

1. Запустите HACMP Security and User Management (Безопасность и управление пользователями HACMP) путем ввода команды **smit cl_usergroup.**

2. Выберите **Users in a HACMP Cluster (Пользователи в кластере HACMP)**.

3. Выберите **List Users in the Cluster (Вывести список пользователей в кластере)**.

4. Выберите узлы, для которых требуется вывести список пользователей. Если оставить поле **Select Nodes by Resource Group (Выбрать узлы по группе ресурсов)** пустым, будет выведен список пользователей со всех узлов.

Если выбрать группу ресурсов, C-SPOC выведет только пользователей с узлов, относящихся к заданной группе ресурсов.

5. Нажмите Enter (рис. 8.9).

```
                                COMMAND STATUS

Command: OK            stdout: yes           stderr: no

Before command completion, additional instructions may appear below.

p650n01  root   0       /
p650n01  daemon 1       /etc
p650n01  bin    2       /bin
p650n01  sys    3       /usr/sys
p650n01  adm    4       /var/adm
p650n01  sshd   207     /var/empty
p650n01  matyi  302     /home/matyi
p650n01  adrien 305     /home/adrien
p650n01  karesz 307     /home/karesz
p650n01  db2adm 1000    /home/db2adm
p650n01  db2inst        1001    /home/db2inst
p650n01  db2user        1003    /home/db2user
p650n01  galam  312     /home/galamb
p650n02  root   0       /
p650n02  daemon 1       /etc
p650n02  bin    2       /bin
p650n02  sys    3       /usr/sys
p650n02  adm    4       /var/adm
p650n02  sshd   204     /var/empty
p650n02  matyi  302     /home/matyi
p650n02  adrien 305     /home/adrien
p650n02  karesz 307     /home/karesz
p650n02  db2adm 1000    /home/db2adm
p650n02  db2inst        1001    /home/db2inst
p650n02  db2user        1003    /home/db2user
p650n02  galam  312     /home/galamb

F1=Help             F2=Refresh          F3=Cancel           F6=Command
F8=Image            F9=Shell            F10=Exit            /=Find
n=Find Next
```

Рис. 8.9. Вывод списка пользователей в кластере

Изменение атрибутов пользователей

Для изменения атрибутов пользователей в кластере проделайте следующее:

1. Запустите HACMP Security and User Management (Безопасность и управление пользователями HACMP) путем ввода команды **smit cl_usergroup.**

2. Выберите **Users in a HACMP Cluster (Пользователи в кластере HACMP)**.

3. Выберите **Change/Show Characteristics of a User in the Cluster (Изменить/показать свойства пользователя в кластере)**.

4. Выберите узлы, на которых следует изменить атрибуты пользователей. Если оставить поле **Select Nodes by Resource Group (Выбрать узлы по группе ресурсов)** пустым, будет выполнено изменение атрибутов пользователя на всех узлах.

Если выбрать группу ресурсов, можно изменить атрибуты пользователя, относящегося к заданной группе ресурсов.

5. Введите имя пользователя или нажмите F4, чтобы выбрать пользователя из всплывающего списка.

6. Теперь можно изменить атрибуты пользователя (рис. 8.10):

```
                    Change / Show Characteristics of a User

  Type or select values in entry fields.
  Press Enter AFTER making all desired changes.

  [TOP]                                                   [Entry Fields]
  * User NAME                                            adrien
    User ID                                             [305] #
    ADMINISTRATIVE USER?                                 false +
    Primary GROUP                                       [staff] +
    Group SET                                           [staff] +
    ADMINISTRATIVE GROUPS                               [] +
    ROLES                                               [] +
    Another user can SU TO USER?                         true +
    SU GROUPS                                           [ALL] +
    HOME directory                                      [/home/adrien]
    Initial PROGRAM                                     [/usr/bin/ksh]
    User INFORMATION                                    [Mrs Adrien]
    EXPIRATION date (MMDDhhmmyy)                        [0]
    Is this user ACCOUNT LOCKED?                         false +
  [MORE...36]

  F1=Help             F2=Refresh          F3=Cancel           F4=List
  F5=Reset            F6=Command          F7=Edit             F8=Image
  F9=Shell            F10=Exit            Enter=Do
```

Рис. 8.10. Изменение атрибутов пользователя

Удаление пользователя

1. Запустите HACMP Security and User Management (Безопасность и управление пользователями HACMP) путем ввода команды **smit cl_usergroup.**

2. Выберите **Users in a HACMP Cluster (Пользователи в кластере HACMP)**.

3. Выберите **Remove a User from the Cluster (Удаление пользователя из кластера)**.

4. Выберите узлы, на которых требуется выполнить удаление пользователей. Если оставить поле **Select Nodes by Resource Group (Выбрать узлы по группе ресурсов)** пустым, пользователь будет удален на всех узлах.

Если выбрать группу ресурсов, C-SPOC удалит пользователя только на узлах, относящихся к заданной группе ресурсов.

5. Введите имя пользователя, которого следует удалить, или нажмите F4, чтобы выбрать пользователя из всплывающего списка.

6. **Remove AUTHENTICATION information (Удалить информацию аутентификации).** Если нужно удалить пароль пользователя и прочую информацию аутентификации, выберите **Yes (Да)**. Если нужно оставить пароль пользователя в файле /etc/security/passwd, выберите **No (Нет)**. По умолчанию установлено **Yes (Да)** (рис. 8.11).

```
                         Remove a User from the Cluster

     Type or select values in entry fields.
     Press Enter AFTER making all desired changes.

                                                          [Entry Fields]
       Select nodes by resource group
        *** No selection means all nodes! ***

     * User NAME                                         [suri] +
       Remove AUTHENTICATION information?                 Yes +



     F1=Help             F2=Refresh          F3=Cancel           F4=List
     F5=Reset            F6=Command          F7=Edit             F8=Image
     F9=Shell            F10=Exit            Enter=Do
```

Рис. 8.11. Удаление пользователя из кластера

Добавление группы пользователей в кластер

1. Запустите HACMP Security and User Management (Безопасность и управление пользователями HACMP) путем ввода команды **smit cl_usergroup**.

2. Выберите **Groups in a HACMP Cluster (Группы в кластере HACMP)**.

3. Выберите **Add a Group to the Cluster (Добавление группы в кластер)**.

4. Выберите узлы, на которых требуется создать группу пользователей. Если оставить поле **Select Nodes by Resource Group (Выбрать узлы по группе ресурсов)** пустым; группа будет создана на всех узлах.

Если выбрать группу ресурсов, группа будет создана только в подмножестве узлов, относящемся к заданной группе ресурсов. В случае кластера из четырех узлов следует оставить это поле пустым. В нашем примере кластера из четырех узлов, описанного в разделе 6.2, «Конфигурация из четырех узлов», имеется две группы ресурсов, привязанных только к узлам node1 и node2 (rg1 и rg2). Если вы хотите создать группу только на этих двух узлах, то следует выбрать соответствующую группу ресурсов.

Таблица 8.2. Таблица перекрестных ссылок групп, групп ресурсов и узлов

Группа ресурсов	Узлы	Группа
rg1	node1, node2	db2group
rg2	node2, node1	developers
rg3	node3, node4	itstaff
rg4	node4, node3, node2, node1	app1users

Табл. 8.2 представляет таблицу перекрестных ссылок между группами, группами ресурсов и узлами. Она показывает, что app1adm app1users существует на всех узлах [поле Select Nodes by Resource Group (Выбрать узлы по группе ресурсов) оставлено пустым], тогда как группа «db2group» будет создана только на узлах node1 и node2 [следует выбрать «rg1» в поле Select Nodes by Resource Group (Выбрать узлы по группе ресурсов)].

```
                          Add a Group to the Cluster

  Type or select values in entry fields.
  Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
    Select nodes by resource group
     *** No selection means all nodes! ***

  * Group NAME                                         [app1users]
    ADMINISTRATIVE group?                               false +
    Group ID                                           [] #
    USER list                                          [db2adm,db2user] +
    ADMINISTRATOR list                                 [] +


  F1=Help             F2=Refresh          F3=Cancel           F4=List
  F5=Reset            F6=Command          F7=Edit             F8=Image
  F9=Shell            F10=Exit            Enter=Do
```

Рис. 8.12. Добавление группы в кластер

5. Создайте группу (рис. 8.12). Следует ввести имя группы, список пользователей и прочую релевантную информацию, требуемую при создании обычной группы. Нажмите F4 для вывода списка доступных пользователей и внесите их в группу.

Здесь можно задать идентификатор группы, после чего C-SPOC проверит доступность заданного идентификатора на всех узлах. Если поле Group ID оставить пустым, создаваемой группе на всех будет назначен произвольный доступный идентификатор.

Вывод списка групп пользователей кластера

1. Запустите HACMP Security and User Management (Безопасность и управление пользователями HACMP) путем ввода команды **smit cl_usergroup.**

```
                                COMMAND STATUS

Command: OK            stdout: yes           stderr: no

Before command completion, additional instructions may appear below.

p650n01  system 0       true    root    files
p650n01  staff  1       false adrien,karesz,db2adm,db2inst,db2user files
p650n01  bin    2       true    root,bin        files
p650n01  sys    3       true    root,bin,sys    files
p650n01  adm    4       true    bin,adm files
p650n01  security       7       true    root    files
p650n01  cron   8       true    root    files
p650n01  shutdown       21      true    files
p650n01  sshd   205     false   sshd    files
p650n01  hacmp  206     false   files
p650n01  haemrm 207     false   files
p650n01  db2group       208     false   db2adm,db2user  root    files
p650n02  system 0       true    root    files
p650n02  staff  1       false adrien,karesz,db2adm,db2user files
p650n02  bin    2       true    root,bin        files
p650n02  sys    3       true    root,bin,sys    files
p650n02  adm    4       true    bin,adm files
p650n02  security       7       true    root    files
p650n02  cron   8       true    root    files
p650n02  shutdown       21      true    files
p650n02  sshd   202     false   sshd    files
p650n02  hacmp  203     false   files
p650n02  haemrm 204     false   files
p650n02  db2group       208     false   db2adm,db2user  root    files

F1=Help             F2=Refresh          F3=Cancel           F6=Command
F8=Image            F9=Shell            F10=Exit            /=Find
n=Find Next
```

Рис. 8.13. Список групп кластера

2. Выберите **Groups in a HACMP Cluster (Группы в кластере HACMP)**.

3. Выберите **List All Groups in the Cluster (Вывести список всех групп в кластере)**.

4. Выберите узлы, для которых требуется вывести список групп. Если оставить поле **Select Nodes by Resource Group (Выбрать узлы по группе ресурсов)** пустым, будет выведен список групп пользователей со всех узлов кластера.

Если выбрать группу ресурсов, C-SPOC выведет только группы с узлов, относящихся к заданной группе ресурсов.

5. C-SPOC выводит группы с соответствующими атрибутами для выбранных узлов (рис. 8.13).

Изменение группы в кластере

1. Запустите HACMP Security and User Management (Безопасность и управление пользователями HACMP) путем ввода команды **smit cl_usergroup.**

2. Выберите **Groups in a HACMP Cluster (Группы в кластере HACMP)**.

3. Выберите **Change / Show Characteristics of a Group in the Cluster (Изменить/показать свойства группы в кластере)**.

4. Выберите узлы, на которых следует выполнить изменение групп. Если оставить поле **Select Nodes by Resource Group (Выбрать узлы по группе ресурсов)** пустым, будет выполнено изменение групп на всех узлах. Если выбрать группу ресурсов, произойдет изменение групп на узлах, относящихся к заданной группе ресурсов.

5. Изменяйте атрибуты группы (рис. 8.14).

```
                    Change / Show Group Attributes on the Cluster

  Type or select values in entry fields.
  Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
    Select nodes by resource group

    Group NAME                                           db2group
    Group ID                                            [208] #
    ADMINISTRATIVE group?                                false +
    USER list                                           [db2adm,db2user] +
    ADMINISTRATOR list                                  [root] +


  F1=Help             F2=Refresh          F3=Cancel           F4=List
  F5=Reset            F6=Command          F7=Edit             F8=Image
  F9=Shell            F10=Exit            Enter=Do
```

Рис. 8.14. Изменение/вывод атрибутов группы в кластере

Удаление группы

1. Запустите HACMP Security and User Management (Безопасность и управление пользователями HACMP) путем ввода команды **smit cl_usergroup.**

2. Выберите **Groups in a HACMP Cluster (Группы в кластере HACMP)**.

3. Выберите **Remove a Group from the Cluster (Удалить группу из кластера)**.

4. Выберите узлы, на которых требуется выполнить удаление групп. Если оставить поле **Select Nodes by Resource Group (Выбрать узлы по группе ресурсов)** пустым, заданная группа будет удалена на всех узлах кластера.

Если выбрать группу ресурсов, C-SPOC удалит группу только на узлах, относящихся к заданной группе ресурсов.

5. Выберите группу, которую следует удалить. Нажмите F4, чтобы вывести список групп в кластере.

Замечания по управлению пользователями через C-SPOC

Ниже приведены некоторые замечания, касающиеся администрирования пользователей и групп через C-SPOS.

- Управление пользователями и паролями через C-SPOC требует, чтобы демон коммуникаций кластера был включен на всех узлах. Нельзя использовать C-SPOC, если какой-либо из узлов отключен. В этом случае возникнет сообщение об ошибке, подобное следующему: clhaver[152]: cannot connect to node p650n01 rc=-1 errno=0. Однако C-SPOC можно использовать вне зависимости от состояния кластера.
- Будьте осторожны при выборе узлов по группам ресурсов. Нужно выбрать именно те узлы, на которых существуют пользователи, которых следует удалить или изменить. Нельзя изменить или удалить пользователя или группу, если этот пользователь или группа не существует на заданном узле.
- При возникновении ошибок во время использования C-SPOC следует обратиться к файлу /tmp/cspoc.log за дополнительной информацией.
- Управление пользователями через C-SPOC нельзя применять совместно с NIS или LDAP.

8.3.2 Управление паролями

HACMP C-SPOC представляет удобный способ изменения пароля пользователя на всех узлах кластера одновременно. Когда кто-либо применяет утилиту **passwd** с любого узла, C-SPOC распространяет новый пароль по всем узлам.

Настройка управления паролями через C-SPOC

По умолчанию утилиты управления паролями через C-SPOC отключены. Ниже описана процедура их подключения.

1. Замените системную утилиту управления паролями на программу управления паролями в кластере. На автономном компьютере AIX для изменения пароля пользо-

вателя применяется команда /usr/bin/passwd. Необходимо заменить эту команду на /usr/es/sbin/cluster/utilities/clpasswd, которая также способна изменять пароль на удаленных узлах:

а) Запустите HACMP Security and User Management (Безопасность и управление пользователями HACMP) путем ввода команды **smit cl_usergroup**.

б) Выберите **Passwords in an HACMP cluster (Пароли в кластере HACMP)**.

в) Выберите **Modify System Password Utility (Изменить системную утилиту управления паролями)**.

г) Нажмите F4 и из всплывающего окна выберите **Link to Cluster Password Utility (Ссылка на утилиту управления паролями в кластере)** (рис. 8.15).

д) Выберите узлы, на которых следует изменить утилиту управления паролями. Чтобы выбрать все узлы, нужно оставить это поле пустым. Рекомендуем выполнить настройку утилиты управления паролями в кластере на всех узлах.

2. Создайте список пользователей, которые могут изменять свой пароль с любого узла кластера:

а) Запустите HACMP Security and User Management (Безопасность и управление пользователями HACMP) путем ввода команды **smit cl_usergroup.**

б) Выберите **Passwords in an HACMP cluster (Пароли в кластере HACMP)**.

```
                        Modify System Password Utility

  Type or select values in entry fields.
  Press Enter AFTER making all desired changes.

                                                      [Entry Fields]
  * /bin/passwd utility is                          [Link to Cluster Passw> +

    Select nodes by Resource Group                  []                      +
       *** No selection means all nodes! ***

    +-------------------------------------------------------------------+
    ¦                        /bin/passwd utility is                     ¦
    ¦                                                                   ¦
    ¦ Move cursor to desired item and press Enter.                      ¦
    ¦                                                                   ¦
    ¦   Original AIX System Command                                     ¦
    ¦   Link to Cluster Password Utility                                ¦
    ¦                                                                   ¦
    ¦ F1=Help                 F2=Refresh              F3=Cancel         ¦
  F1¦ F8=Image                F10=Exit                Enter=Do          ¦
  F5¦ /=Find                  n=Find Next                               ¦
  F9+-------------------------------------------------------------------+
```

Рис. 8.15. Изменение системной утилиты управления паролями

```
                    Manage List of Users Allowed to Change Password

   Type or select values in entry fields.
   Press Enter AFTER making all desired changes.

                                                          [Entry Fields]
     Users allowed to change password                   [adrien karesz] +
     cluster-wide


     F4 lists all users defined to all cluster nodes




   F1=Help             F2=Refresh          F3=Cancel           F4=List
   F5=Reset            F6=Command          F7=Edit             F8=Image
   F9=Shell            F10=Exit            Enter=Do
```

Рис. 8.16. Управление списком пользователей, которым разрешено изменять пароль в масштабе кластера

```
                  Manage List of Users Allowed to Change Password

  Type or select values in entry fields.
  Press Enter AFTER making all desired changes.

                                                          [Entry Fields]
    Users allowed to change password                    [adrien karesz]     +
    +-------------------------------------------------------------------+
    ¦                     Users allowed to change password              ¦
    ¦                               cluster-wide                        ¦
    ¦ Move cursor to desired item and press F7.                         ¦
    ¦     ONE OR MORE items can be selected.                            ¦
    ¦ Press Enter AFTER making all selections.                          ¦
    ¦                                                                   ¦
    ¦   ALL_USERS                                                       ¦
    ¦   sshd                                                            ¦
    ¦   adrien                                                          ¦
    ¦   karesz                                                          ¦
    ¦   db2adm                                                          ¦
    ¦   db2user                                                         ¦
    ¦                                                                   ¦
    ¦ F1=Help                 F2=Refresh              F3=Cancel         ¦
  F1¦ F7=Select               F8=Image                F10=Exit          ¦
  F5¦ Enter=Do                /=Find                  n=Find Next       ¦
  F9+-------------------------------------------------------------------+
```

Рис. 8.17. Выбор пользователей, которым разрешено изменять свой пароль в масштабе кластера

в) Выберите **Manage List of Users Allowed to Change Password (Управление списком пользователей, которым разрешено изменять пароль)**.

г) SMIT выводит пользователей, которым уже разрешено изменять свой пароль в масштабе кластера (рис. 8.16).

д) Чтобы добавить пользователей или изменить список пользователей, которым разрешено изменять пароль в масштабе кластера, нажмите F4 и выберите имена пользователей из всплывающего списка. Чтобы все текущие и будущие пользователи могли осуществлять управление паролями через C-SPOC, следует выбрать **ALL_USERS** (рис. 8.17).

Мы рекомендуем включить сюда только реально существующих пользователей и выполнять изменение паролей вручную для технических пользователей.

Внимание! Если включить утилиты управления паролями через C-SPOC для всех пользователей в кластере, в котором есть пользователи, созданные только на одном узле, будет выдаваться сообщение об ошибке, подобное следующему:

```
# passwd joe
Changing password for "joe"
joe's New password:
Enter the new password again:
p650n02: clpasswdremote: User joe does not exist on node p650n02
p650n02: cl_rsh had exit code = 1, see cspoc.log and/or clcomd.log for more information
```

Изменение пароля происходит, несмотря на сообщение об ошибке.

Изменение пароля пользователя через C-SPOC

1. Запустите HACMP Security and User Management (Безопасность и управление пользователями HACMP) путем ввода команды **smit cl_usergroup.**

2. Выберите **Passwords in an HACMP cluster (Пароли в кластере HACMP)**.

3. Выберите **Change a User's Password in the Cluster (Изменение пароля пользователя в кластере)**.

4. Выберите узлы, на которых следует изменить пароль пользователя. Чтобы выбрать все узлы, нужно оставить это поле пустым. Если выбрать группу ресурсов, CSPOC изменит пароль только на узлах, относящихся к заданной группе ресурсов.

5. Введите имя пользователя или нажмите F4, чтобы выбрать пользователя из всплывающего списка.

6. Установите в поле **User must change password on first login (Пользователь должен изменить пароль при первом входе в систему)** значение **true (да)** или **false (нет)**, в зависимости от ваших требований (рис. 8.18).

7. Нажмите Enter и введите новый пароль.

Замечание. Вы также можете использовать команду `passwd` для изменения пароля того или иного пользователя на всех узлах.

```
                    Change a User's Password in the Cluster

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
  Selection nodes by resource group
   *** No selection means all nodes! ***

* User NAME                                            [karesz] +
  User must change password on first login?             true +



F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Рис. 8.18. Изменение пароля пользователя в кластере

Изменение своего собственного пароля

1. Запустите HACMP Security and User Management (Безопасность и управление пользователями HACMP) путем ввода команды **smit cl_usergroup**.

2. Выберите **Passwords in an HACMP cluster (Пароли в кластере HACMP)**.

3. Выберите **Change Current Users Password (Изменение пароля текущего пользователя)**.

```
                         Change Current Users Password

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
  Selection nodes by resource group                   [] +
   *** No selection means all nodes! ***

  User NAME                                            adrien




F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Рис. 8.19. Изменение своего собственного пароля

4. Выберите узлы, на которых следует изменить ваш пароль. Чтобы выбрать все узлы, нужно оставить это поле пустым. Если выбрать группу ресурсов, C-SPOC изменит пароль только на узлах, относящихся к заданной группе ресурсов.

5. На экране SMIT выводится ваше имя пользователя (рис. 8.19).

6. Нажмите Enter и измените свой пароль.

Теперь ваш пароль изменен на всех узлах.

> **Замечание.** Вы также можете использовать команду passwd для изменения своего пароля на всех узлах.

8.4 Управление общим хранением

Утилита C-SPOC упрощает обслуживание общих компонентов LVM в кластерах. Команды C-SPOC выполняют в кластерной среде функции, подобные стандартным командам AIX 5L, работающим на автономном узле. Путем автоматизации повторяющихся задач на разных узлах в кластере C-SPOC устраняет потенциальный источник ошибок и ускоряет процесс обслуживания кластера. Хотя для выполнения этих процедур можно применять утилиты AIX 5L на каждом узле, мы рекомендуем использовать C-SPOC там, где это возможно.

Команды C-SPOC работают только с общими компонентами LVM и компонентами LVM с одновременным доступом, определенными в составе группы ресурсов HACMP. При использовании SMIT HACMP C-SPOC выполнение команд происходит на узле, являющемся владельцем компонента LVM (узле, на котором этот компонент активизирован).

> **Примечание.** Команды C-SPOC, изменяющие компоненты LVM, требуют использования имени группы ресурсов в качестве аргумента. Компонент LVM, для которого выполняется команда, должен быть сконфигурирован в заданной группе ресурсов. C-SPOC использует информацию из группы ресурсов, чтобы определить, на каких узлах следует выполнять заданную операцию.

Дополнительные сведения о подсистеме хранения см. в гл. 13, «Аспекты подсистемы хранения».

8.4.1 Обновление компонентов LVM

При изменении какого-либо определения любого общего компонента LVM в кластере (включая группы томов, логические тома и файловые системы) операция обновляет данные в AIX ODM, описывающие компонент на локальном узле и в области дескриптора группы томов (Volume Group Descriptor Area, VGDA) на всех дисках в группе томов. Это обновление ODM следует распространить на все узлы в кластере, чтобы обеспечить корректную работу кластера.

Если эти изменения в LVM производятся через C-SPOC, распространение изменений в ODM на все узлы кластера осуществляется автоматически.

Если эти изменения в LVM производятся с использованием команд AIX на локальном узле, необходимо вручную распространить изменения в ODM на другие узлы. Изменения в LVM связаны со структурой отдельной группы томов.

Импорт групп томов вручную

Обычная процедура распространения информации в ODM о группе томов на основе AIX на другие узлы для групп томов без одновременного доступа представлена в примере 8.1. Эти же действия можно использовать и для групп томов с расширенным одновременным доступом. Можно также использовать аналогичную команду AIX SMIT вместо командной строки.

Пример 8.1. Импорт групп томов AIX вручную

Задачи, выполняющиеся на локальном узле (там, где активизирована группа томов)

```
p630n03> lsvg -l applvg
applvg:
LV NAME             TYPE       LPs   PPs   PVs  LV STATE      MOUNT POINT
applvglog           jfs2log    1     2     2    open/syncd    N/A
applv               jfs2       200   400   4    open/syncd    /app1
p630n03> umount /app1
p630n03> varyoffvg applvg
p630n03> ls -l /dev/applvg
crw-r-----   1 root     system       90,  0 Jun 21 13:09 /dev/applvg
```

Задачи, выполняющиеся на остальных узлах

```
p630n02> lspv |grep applvg
vpath0          000685bf8595e225                    applvg
vpath1          000685bf8595e335                    applvg
vpath2          000685bf8595e445                    applvg
vpath3          000685bf8595e559                    applvg
p630n02> exportvg applvg
p630n02> importvg -y applvg -n -V90 vpath0
p630n02> chvg -a n applvg
p630n02> varyoffvg applvg
```

> **Внимание!** При экспорте и последующем импорте группы томов права владения и разрешения для устройств логических томов сбрасываются. После экспорта и импорта владельцем группы томов является root:system. Это может повлиять на приложения, в частности на некоторые серверы баз данных, использующие логические тома прямого доступа (raw logical volumes). Вы должны проверить права владения и разрешения, прежде чем экспортировать группу томов, и восстановить их вручную в том случае, если они отличны от root:system.

Вместо команды экспорта-импорта можно применять команду **importvg -L VGNAME HDISK** на других узлах, но следует помнить о том, что опцию -L нельзя использовать, если на основном узле был удален логический том или файловая система. Команда **importvg -L** сохраняет права владения логическими томами.

Начиная с HACMP 5.2 администрирование LVM может быть упрощено с использованием групп томов с расширенным одновременным доступом, так как некоторые изменения в LVM ODM непосредственно распространяются на все узлы кластера. Дополнительные сведения о группах томов с расширенным одновременным доступом см. в гл. 13, «Аспекты подсистемы хранения».

Lazy update

HACMP содержит специальное средство автоматической синхронизации информации в LVM ODM во время перемещения при сбое в случаях, когда записанные отметки времени VGDA различаются. Эта функция называется *ленивым обновлением* (*lazy update*). AIX 5L обновляет эту отметку времени при каждом изменении компонента LVM. Когда другой узел кластера пытается активизировать группу томов, HACMP сравнивает свою копию отметки времени с отметкой времени в VGDA на диске. Если значения различаются, программное обеспечение HACMP экспортирует и повторно импортирует группу томов перед ее активизацией. Если отметки времени совпадают, HACMP активизирует группу томов без экспорта и повторного импорта.

> **Примечание.** Мы рекомендуем периодически выполнять верификацию кластера и следовать рекомендованной процедуре администрирования, вместо того чтобы полагаться на функцию lazy update.

Однако lazy update может отказать в случае удаления логического тома или файловой системы с первоначального узла.

> **Примечание.** Начиная с версии 5.2 HACMP не требует обработки lazy update для групп томов с расширенным одновременным доступом, так как он обеспечивает обновление всех узлов кластера информацией LVM. Дополнительные сведения см. в разделе 13.1.1, «Группы томов с расширенным одновременным доступом».

Автоматический импорт групп томов

HACMP содержит дополнительное средство для автоматического импорта группы томов. Оно позволяет автоматически импортировать группы томов с возможностью общего доступа на всех целевых узлах в группе ресурсов. Это выполняется через меню **Extended Resource Configuration (Расширенное конфигурирование ресурсов).** Автоматический импорт позволяет создать группу томов и сразу же добавить ее в группу ресурсов, не импортируя ее вручную на всех целевых узлах в группе ресурсов.

Запустите **smit hacmp > Extended Configuration (Расширенное конфигурирование) > Extended Resource Configuration (Расширенное конфигурирование ресурсов) > HACMP Extended Resource Group Configuration (Расширенное конфигурирование групп ресурсов HACMP) > Change/Show Resources and Attributes for a Resource Group (Изменить/показать ресурсы и атрибуты группы ресурсов)**, после чего выберите группу ресурсов и установите для параметра **Automatically Import Volume Groups (Автоматически импортировать группы томов)** значение **true**.

Для того чтобы обеспечить импорт доступных групп томов в HACMP, должны быть выполнены следующие условия:

- имена группы томов должны совпадать на всех узлах кластера и быть уникальными в кластере;
- логические тома и файловые системы должны иметь уникальные имена;
- все физические диски должны быть видимы из AIX 5L и иметь назначенные PVID;
- физические диски, на которых располагается группа томов, должны быть доступны для всех узлов в группе ресурсов.

Перед этим следует установить для опции Auto Discover (Автоматическое обнаружение) значение true. Для этого следует запустить **smit hacmp > Extended Configuration (Расширенное конфигурирование) > Discover HACMP-related Information from Configured Nodes (Обнаружение информации, относящейся к HACMP, на сконфигурированных узлах)**.

Импорт групп томов с использованием C-SPOC

Импорт общей группы томов на всех узлах в кластере можно осуществлять с использованием утилиты C-SPOC. Для этого нужно запустить **smit hacmp > System Management (C-SPOC) > HACMP Logical Volume Management (Управление логическими томами HACMP) > Shared Volume Groups (Общие группы томов) > Import a Shared Volume Group (Импорт общей группы томов)** и выбрать группу томов из списка. Список групп ресурсов без одновременного доступа также включает группы томов с расширенным одновременным доступом. После выбора группы томов выводится список физических дисковых устройств. После выбора физического диска для импорта следует заполнить поля на экране импорта группы томов, как показано на рис. 8.20.

После выполнения этой операции можно запустить процесс обнаружения, чтобы включить новую группу томов в список для последующих действий.

```
 Import a Shared Volume Group

Type or select values in entry fields.
Press Enter AFTER making all desired changes.
                                                      [Entry Fields]
  Resource Group Name                                C10RG1
  VOLUME GROUP name                                  testvg
  Reference node                                     panther
  PHYSICAL VOLUME name                               vpath2
  Volume group MAJOR NUMBER                         [100]                +#
  Make this VG Concurrent Capable?                   no                   +
  Make default varyon of VG Concurrent?              no                   +


F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Рис. 8.20. Экран импорта группы томов в C-SPOC

8.4.2 Диспетчер логических томов C-SPOC

Меню диспетчера логических томов C-SPOC (Logical Volume Manager) позволяет выполнять команды LVM из меню и экранов, похожих на меню AIX LVM SMIT (вызываемых командой **smit lvm)**. Различие состоит в том, что в C-SPOC для выполнения большинства задач необходимо сначала выбрать локальный узел или соответствующий ресурс (группу ресурсов, группу томов). После этого обычно появляется экран SMIT, очень похожий на соответствующий экран AIX LVM. Администраторам AIX, знакомым с меню AIX LVM SMIT, будет очень легко использовать перечисленные ниже опции C-SPOC.

Еще одно преимущество C-SPOC состоит в том, что начиная с HACMP 5.3 все операции с дисками VPATH, поддерживаемые в AIX 5L, поддерживаются и в C-SPOC. Должен быть установлен пакет SDD 1.3.1.3 или пакет более поздней версии.

Меню LVM C-SPOC для управления логическими томами в масштабе кластера можно вызвать следующим образом: **smit cl_admin > HACMP Logical Volume Management (Управление логическими томами HACMP)**. Это меню содержит следующие опции:

- Shared Volume Groups (Общие группы томов):
  - List All Shared Volume Groups (Вывод списка всех общих групп томов);
  - Create a Shared Volume Group (Создание общей группы томов);
  - Create a Shared Volume Group with Data Path Devices (Создание общей группы томов с устройствами Datapath);
  - Set Characteristics of a Shared Volume Group (Назначение свойств общей группы томов);
  - Import a Shared Volume Group (Импорт общей группы томов);
  - Mirror a Shared Volume Group (Зеркальное отображение общей группы томов);
  - Unmirror a Shared Volume Group (Отмена зеркального отображения общей группы томов).
- Shared Logical Volumes (Общие логические тома):
  - List All Shared Logical Volumes by Volume Group (Вывод списка всех общих логических томов по группам томов);
  - Add a Shared Logical Volume (Добавление общего логического тома);
  - Set Characteristics of a Shared Logical Volume (Назначение свойств общего логического тома);
  - Show Characteristics of a Shared Logical Volume (Вывод свойств общего логического тома);
  - Change a Shared Logical Volume (Изменение общего логического тома);
  - Remove a Shared Logical Volume (Удаление общего логического тома).
- Shared File Systems (Общие файловые системы):
  - Journaled File Systems (Файловые системы JFS);

- Enhanced Journaled File Systems (Улучшенные файловые системы JFS[1]); ниже приведены подменю для этих типов файловых систем одинаковы:
  - Add an (Enhanced) Journaled File System [Добавление (улучшенной) файловой системы JFS];
  - Add an (Enhanced) Journaled File System on a Previously Defined Logical Volume [Добавление (улучшенной) файловой системы JFS на предварительно определенном логическом томе];
  - List All Shared File Systems (Вывод списка всех общих файловых систем);
  - Change/Show Characteristics of a Shared (Enhanced) Journaled File System [Изменение/вывод параметров общей (улучшенной) файловой системы JFS];
  - Remove a Shared File System (Удаление общей файловой системы).

- Synchronize Shared LVM Mirrors (Синхронизация общих зеркальных отображений LVM):
  - Synchronize by Volume Group (Синхронизация по группам томов);
  - Synchronize by Logical Volume (Синхронизация по логическим томам).
- Synchronize a Shared Volume Group Definition (Синхронизация определения общей группы томов).

Дополнительные сведения о выполнении задач см. в разделе 8.4.5, «Примеры».

8.4.3 Управление логическими томами с одновременным доступом через C-SPOC

Меню диспетчера логических томов в C-SPOC позволяет выполнять команды LVM в группах томов с одновременным доступом из меню и экранов, похожих на меню C-SPOC LVM, описанные в предыдущем разделе.

Меню LVM C-SPOC для управления логическими томами с одновременным доступом в масштабе кластера можно вызвать следующим образом: **smit cl_admin > HACMP Concurrent Logical Volume Management (Управление логическими томами HACMP)**. Это меню содержит три основные опции:

- Concurrent Volume Groups (Группы томов с одновременным доступом):
  - List All Concurrent Volume Groups (Вывод списка всех групп томов с одновременным доступом);
  - Create a Concurrent Volume Group (Создание группы томов с одновременным доступом);
  - Create a Concurrent Volume Group with Data Path Devices (Создание группы томов с одновременным доступом с устройствами Datapath);
  - Set Characteristics of a Concurrent Volume Group (Установка параметров группы томов с одновременным доступом);
  - Import a Concurrent Volume Group (Импорт группы томов с одновременным доступом).

[1] Enhanced JFS и JFS2 – одно и то же. *Примеч. науч. ред.*

  - Mirror a Concurrent Volume Group (Зеркальное отображение группы томов с одновременным доступом);
  - Unmirror a Concurrent Volume Group (Отмена зеркального отображения группы томов с одновременным доступом);
- Concurrent Logical Volumes (Логические тома с одновременным доступом):
  - List All Concurrent Logical Volumes by Volume Group (Вывод списка всех логических томов с одновременным доступом по группам томов);
  - Add a Concurrent Logical Volume (Добавление логического тома с одновременным доступом);
  - Set Characteristics of a Concurrent Logical Volume (Установка параметров тома с одновременным доступом);
  - Show Characteristics of a Concurrent Logical Volume (Вывод параметров логического тома с одновременным доступом);
  - Remove a Concurrent Logical Volume (Удаление логического тома с одновременным доступом).
- Synchronize Concurrent LVM Mirrors (Синхронизация зеркальных отображений LVM с одновременным доступом):
  - Synchronize by Volume Group (Синхронизация по группам томов);
  - Synchronize by Logical Volume (Синхронизация по логическим томам).

Дополнительные сведения о выполнении задач см. в разделе 8.4.5, «Примеры».

8.4.4 Управление физическими томами через C-SPOC

Меню LVM C-SPOC для управления физическими томами и управления виртуальными путями SDD можно вызвать следующим образом: **smit cl_admin > HACMP Physical Volume Management (Управление физическими томами HACMP)**. Это меню содержит следующие опции:

- Add a Disk to the Cluster (Добавление диска в кластер).
- Remove a Disk From the Cluster (Удаление диска из кластера).
- Cluster Disk Replacement (Замена диска кластера).
- Cluster Data Path Device Management (Управление устройствами Datapath в кластере):
  - Display Data Path Device Configuration (Вывод конфигурации устройства Datapath);
  - Display Data Path Device Status (Вывод состояния устройства Datapath);
  - Display Data Path Device Adapter Status (Вывод состояния адаптера устройства Datapath);
  - Define and Configure all Data Path Devices (Определение и конфигурирование всех устройств Datapath);
  - Add Paths to Available Data Path Devices (Добавление путей для доступных устройств Datapath);

- Configure a Defined Data Path Device (Конфигурирование определенного устройства Datapath);
- Remove a Data Path Device (Удаление устройства Datapath);
- Convert ESS hdisk Device Volume Group to an SDD VPATH Device Volume Group (Преобразование группы томов ESS hdisk в группу томов SDD VPATH);
- Convert SDD VPATH Device Volume Group to an ESS hdisk Device Volume Group (Преобразование группы томов SDD VPATH в группу томов ESS hdisk).

► Конфигурирование расположения дисков и сайтов для межсайтового зеркального отображения LVM (Cross-Site® LVM Mirroring). Дополнительные сведения о реализации межсайтового зеркального отображения вместе с примером см. в гл. 16, «HACMP с межсайтовым зеркальным отображением LVM».

Дополнительные сведения о выполнении задач см. в разделе 8.4.5, «Примеры».

8.4.5 Примеры

Этот раздел содержит несколько примеров и сценариев управления C-SPOC LVM. Приведены примеры выполнения следующих задач:

1. Вывод текущей конфигурации datapath.
2. Вывод состояния адаптера устройства.
3. Добавление группы томов с расширенным одновременным доступом на основе vpath.
4. Добавление группы томов в группу ресурсов.
5. Создание нового логического тома.
6. Создание нового логического тома jfslog2.
7. Создание новой файловой системы.
8. Добавление дополнительного vpath в группу томов.
9. Увеличение размера файловой системы.
10. Удаление файловой системы.
11. Синхронизация определения группы томов по кластерам узла.

Для демонстрации примеров используется кластер из трех узлов с перехватом IP-адресов посредством синонимов, применяющий также мониторинг пульса посредством синонимов. В качестве хранилища используется ESS с двумя путями, так что действия осуществляются на устройствах vpath. На рис. 8.21 представлена конфигурация тестового кластера.

Вывод текущей конфигурации datapath

Этот пример показывает, как вывести существующую конфигурацию vpath с информацией PVID на заданном узле.

Экран конфигурации datapath в C-SPOC вызывается следующим образом: **smit cl_admin > HACMP Physical Volume Management (Управление физическими томами HACMP) > Cluster Data Path Device Management (Управление уст-**

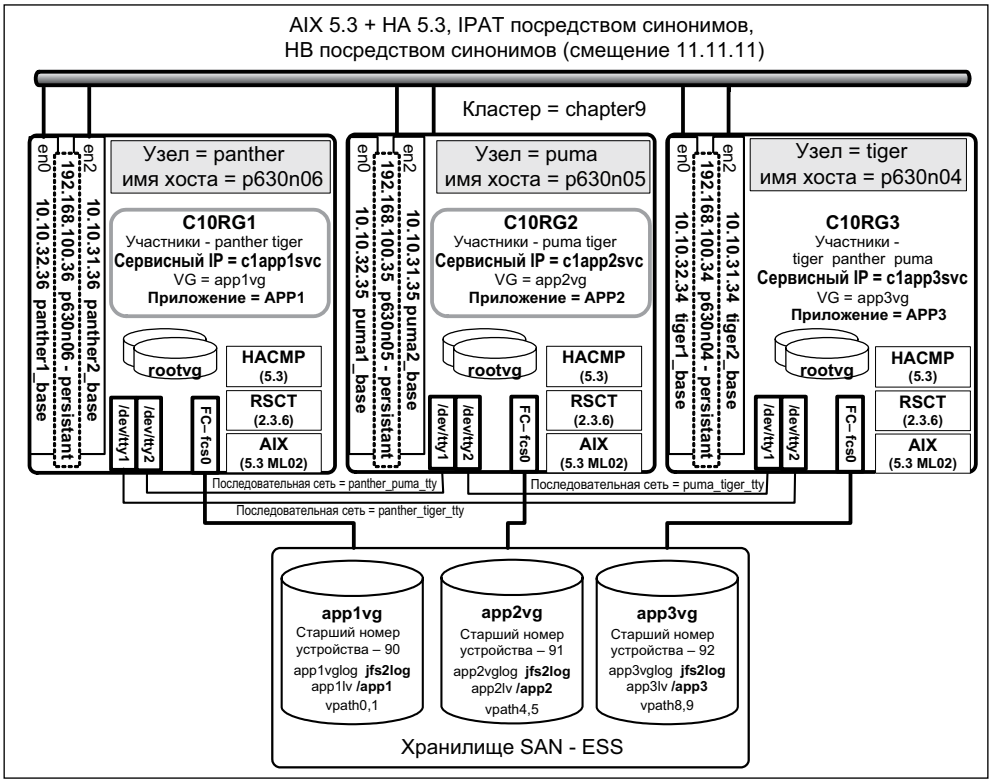

Рис. 8.21. Конфигурация тестового кластера C-SPOC LVM

ройствами Datapath в кластере) > Display Data Path Device Configuration (Вывод конфигурации устройства Datapath), после чего следует выбрать узел panther. Выводимая информация представлена в примере 8.2.

Пример 8.2. Экран конфигурации datapath в C-SPOC

```
PVID: 000685cf86a5dfe6
panther: vpath8 (Avail pv app3vg) 10822513 = hdisk11 (Avail ) hdisk23 (Avail )
PVID: 000685cf86a5e0f6
panther: vpath9 (Avail pv app3vg) 10922513 = hdisk12 (Avail ) hdisk24 (Avail )
PVID: 000685cf86a5e1fa
panther: vpath10 (Avail pv ) 10A22513 = hdisk13 (Avail ) hdisk25 (Avail )
PVID: 000685cf86a5e2fa
panther: vpath11 (Avail pv ) 10B22513 = hdisk14 (Avail ) hdisk26 (Avail )
PVID: 000685cf86aaa0a3
panther: vpath4 (Avail pv app2vg) 10422513 = hdisk7 (Avail ) hdisk19 (Avail )
PVID: 000685cf86aaa3a3
panther: vpath5 (Avail pv app2vg) 10522513 = hdisk8 (Avail ) hdisk20 (Avail )
PVID: 000685cf86aaa63e
```

Вывод состояния адаптера устройства

Этот пример показывает, как вывести информацию о состоянии адаптера FC на заданном узле.

Экран состояния адаптера устройства в C-SPOC вызывается следующим образом: **smit cl_admin > HACMP Physical Volume Management (Управление физическими томами HACMP) > Cluster Data Path Device Management (Управление устройствами Datapath в кластере) > Display Data Path Device Adapter Status (Вывод состояния адаптера устройства Datapath)**, после чего следует выбрать узел panther. Выводимая информация представлена в примере 8.3.

Пример 8.3. Экран datapath-адаптера в C-SPOC

```
panther:

Active Adapters :1

Adpt#     Name   State     Mode              Select     Errors  Paths  Active
    0   fscsi0  NORMAL   ACTIVE              109346          0     24       4
```

Добавление группы томов с расширенным одновременным доступом на основе vpath

Следующий пример показывает, как добавить новую группу томов с расширенным одновременным доступом в кластер.

Перед созданием общей группы томов для кластера через C-SPOC необходимо проверить соответствие следующим требованиям:

- все дисковые устройства должны быть должным образом подключены к узлам кластера;
- все дисковые устройства должны быть должным образом сконфигурированы на всех узлах кластера, а также устройства должны быть доступными со всех узлов;
- диски должны иметь PVID.

Добавление группы томов с расширенным одновременным доступом выполняется с использованием следующей команды: **smit cl_admin > HACMP Concurrent Logical Volume Management (Управление логическими томами HACMP) > Concurrent Volume Groups (Группы томов с одновременным доступом) > Create a Concurrent Volume Group with Data Path Devices (Создание группы томов с одновременным доступом с устройствами Datapath).**

Затем следует выбрать два узла: panther и puma. Мы выбираем эти два узла, так как мы собираемся добавить тестовую группу томов (testvg) в группу ресурсов C10RG1, в которой эти два узла участвуют. После этого следует выбрать соответствующие vpath-устройства в меню SMIT, как показано в примере 8.4. Список выбора PVID отображает только незанятые устройства на выбранных узлах.

Пример 8.4. Список выбора PVID

```
¦                              Physical Volumes                                ¦
¦                                                                              ¦
¦ Move cursor to desired item and press F7.                                    ¦
¦     ONE OR MORE items can be selected.                                       ¦
¦ Press Enter AFTER making all selections.                                     ¦
¦                                                                              ¦
¦ [MORE...1]                                                                   ¦
¦   #   panther: vpath11: hdisk14 hdisk26                                      ¦
¦   #   tiger: vpath11: hdisk15 hdisk27                                        ¦
¦ > 000685cf86af4e3c                                                           ¦
¦   #   panther: vpath3: hdisk6 hdisk18                                        ¦
¦   #   tiger: vpath3: hdisk7 hdisk19                                          ¦
¦ > 000685cf86af4bc7                                                           ¦
¦   #   panther: vpath2: hdisk5 hdisk17                                        ¦
¦   #   tiger: vpath2: hdisk6 hdisk18                                          ¦
¦ [MORE...3]                                                                   ¦
```

В экране, выводимом после выбора устройств vpath, следует ввести все необходимые поля для создания группы томов, как показано в примере 8.5.

Пример 8.5. Создание группы томов

```
 Create a Shared Volume Group with Data Path Devices

 Type or select values in entry fields.
 Press Enter AFTER making all desired changes.

 [TOP]                                                   [Entry Fields]
   Node Names                                          panther
   PVID                                                000685cf86af4bc7 0006>
   VOLUME GROUP name                                  [testvg]
   Physical partition SIZE in megabytes                128                    +
   Volume group MAJOR NUMBER                          [100]
   Enhanced Concurrent Mode                            true                   +
   Enable Cross-Site LVM Mirroring Verification        false                  +

   Warning:
   Changing the volume group major number may result
   in the command being unable to execute
   successfully on a node that does not have the
   major number currently available. Please check
   for a commonly available major number on all nodes
   before changing this setting.

 F1=Help             F2=Refresh          F3=Cancel           F4=List
 F5=Reset            F6=Command          F7=Edit             F8=Image
 F9=Shell            F10=Exit            Enter=Do
```

После создания группы томов с расширенным одновременным доступом testvg была выполнена проверка регистрации информации о новой группе томов на узлах tiger и panther. Как и ожидалось, информация о группе томов в ODM была успешно обновлена на обоих узлах.

Добавление группы томов в группу ресурсов

Так как команды C-SPOC LVM оперируют только с компонентами, которые определены в составе группы ресурсов HACMP, мы добавляем группу томов в группу ресурсов C10RG1.

Эти примеры иллюстрируют добавление группы томов в существующую группу ресурсов с использованием функции Automatically Import Volume Groups (Автоматически импортировать группы томов). Эта функция полезна при создании новых групп томов только на локальном узле и использовании функции Automatically Import Volume Groups (Автоматически импортировать группы томов) для импортирования группы томов на все остальные узлы, участвующие в выбранной группе ресурсов. При добавлении группы томов из C-SPOC выполняется синхронизация LVM ODM на всех узлах в кластере и функцию Automatically Import Volume Groups (Автоматически импортировать группы томов) использовать необязательно.

Мы добавляем группу томов в существующую группу ресурсов C10RG1 с использованием меню Change/Show Resources and Attributes for a Resource Group (Изменить/показать ресурсы и атрибуты группы ресурсов) в SMIT. Запускаем **smit hacmp > Extended Resource Configuration (Расширенное конфигурирование ресурсов) > HACMP Extended Resource Group Configuration (Расширенное конфигурирование групп ресурсов HACMP) > Change/Show Resources and Attributes for a Resource Group (Изменить/показать ресурсы и атрибуты группы ресурсов)**, после чего выделяем группу томов C10RG1 и заполняем поля, как показано в примере 8.6.

Пример 8.6. Добавление группы томов в группу ресурсов с использованием опции автоматического импорта

```
 Change/Show All Resources and Attributes for a Custom Resource Group
 Type or select values in entry fields.
 Press Enter AFTER making all desired changes.

 [TOP]                                                  [Entry Fields]
   Resource Group Name                                 C10RG1
   Participating Nodes (Default Node Priority)         cobra viper

   Startup Policy                                      Online On Home Node O>
   Fallover Policy                                     Fallover To Next Prio>
   Fallback Policy                                     Never Fallback

   Service IP Labels/Addresses                        [app1svc testvg]        +
   Application Servers                                [APP1]                  +

   Volume Groups                                      [app1vg ]               +
   Use forced varyon of volume groups, if necessary    false                  +
   Automatically Import Volume Groups                  true                   +
 [MORE...20]
```

```
F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

> **Примечание.** Функция Automatically Import Volume Groups (Автоматически импортировать группы томов) импортирует группу томов только на тех узлах, которые участвуют в группе ресурсов, в которой была добавлена новая группа томов. В нашем примере узел puma не участвует в группе ресурсов C10RG1, так что testvg импортируется только на узел tiger.

После добавления группы томов testvg в группу ресурсов C10RG1 выполняется синхронизация конфигурации в целях распространения изменений в группе ресурсов на другие узлы. Запускаем **smit hacmp > Extended Configuration (Расширенное конфигурирование) > Extended Resource Configuration (Расширенное конфигурирование ресурсов) > Extended Verification and Synchronization (Расширенная верификация и синхронизация)** и запускаем процесс синхронизации и верификации с использованием параметров по умолчанию. Дополнительные сведения о синхронизации и верификации см. в разделе 8.6, «Верификация и синхронизация кластера».

Создание нового логического тома

Следующий пример показывает, как создать новый логический том в выбранной группе томов, которая уже активна в составе группы ресурсов.

Мы добавляем логический том test01lv в группу томов testvg командой **smit cl_admin > HACMP Logical Volume Management (Управление логическими томами HACMP) > Shared Logical Volumes (Общие логические тома)**. Затем выбираем группу томов testvg из появившегося списка, показанного в примере 8.14. На следующем экране, выводящемся после этого, выбираем устройства для размещения логического тома, как показано в примере 8.7.

Пример 8.7. Создание нового логического тома – 1

```
¦                          Physical Volume Names                          ¦
¦                                                                         ¦
¦ Move cursor to desired item and press F7.                               ¦
¦     ONE OR MORE items can be selected.                                  ¦
¦ Press Enter AFTER making all selections.                                ¦
¦                                                                         ¦
¦   Auto-select                                                           ¦
¦   panther vpath2                                                        ¦
¦   panther vpath3                                                        ¦
```

После этого мы заполнили необходимые поля, как показано в примере 8.8.

Пример 8.8. Создание нового логического тома – 2

```
 Add a Shared Logical Volume

 Type or select values in entry fields.
 Press Enter AFTER making all desired changes.

 [TOP]                                                   [Entry Fields]
   Resource Group Name                                  C1ORG1
   VOLUME GROUP name                                    testvg
   Reference node
 * Number of LOGICAL PARTITIONS                         [10]                    #
   PHYSICAL VOLUME names
   Logical volume NAME                                  [test01lv]
   Logical volume TYPE                                  [jfs2]                 +
   POSITION on physical volume                           middle                +
   RANGE of physical volumes                             minimum               +
   MAXIMUM NUMBER of PHYSICAL VOLUMES                   []                      #
     to use for allocation
   Number of COPIES of each logical                      1                     +
     partition
 [MORE...11]

 F1=Help             F2=Refresh          F3=Cancel           F4=List
 F5=Reset            F6=Command          F7=Edit             F8=Image
 F9=Shell            F10=Exit            Enter=Do
```

Создается test01lv, после чего информация передается на узел puma. После этого мы выполнили верификацию кластера командой **smit hacmp > Problem Determination Tools (Инструменты определения проблем) > HACMP Verification (Верификация HACMP).**

Верификация прошла без ошибок.

Создание нового логического тома jfslog2

Для добавления нового логического тома jfs2log под названием testjlog2lv в группу томов testvg мы использовали процедуру, описанную выше в разделе «Создание нового логического тома». В C-SPOC в экране создания нового логического тома – 2, представленном в примере 8.8, выбираем тип логического тома jfs2log:

Logical volume TYPE [jfs2log]

После добавления логического тома jfs2log в группу томов testvg производится форматирование jfs2log на узле panther, на котором группа томов testvg является активной. Выполняется следующая команда:

p630n06 >logform /dev/testjloglv

logform: destroy /dev/rtestjloglv (y)?y

Создание новой файловой системы

Следующий пример показывает, как создать файловую систему jfs2 на предварительно созданном логическом томе. Эта задача выполняется следующим образом **smit cl_admin-> HACMP Logical Volume Management (Управление логическими томами HACMP) -> Shared File Systems (Общие файловые системы) -> Enhanced Journaled File Systems (Улучшенные файловые системы JFS) -> Add an Enhanced Journaled File System on a Previously Defined Logical Volume (Добавление улучшенной файловой системы JFS на предварительно определенном логическом томе).**

Затем выбирается предварительно созданный логический том из списка SMIT. После этого заполняются все необходимые поля, как показано в примере 8.9.

Пример 8.9. Создание файловой системы jfs2 в C-SPOC

```
 Add an Enhanced Journaled File System on a Previously Defined Logical Volume

 Type or select values in entry fields.
 Press Enter AFTER making all desired changes.

                                                      [Entry Fields]
   Node Names                                        panther,tiger,
   LOGICAL VOLUME name                               test01lv
 * MOUNT POINT                                       [/cltestfs]
   PERMISSIONS                                        read/write            +
   Mount OPTIONS                                     []                     +
   Block Size (bytes)                                 4096                  +
   Inline Log?                                        no                    +
   Inline Log size (MBytes)                          []                      #


 F1=Help             F2=Refresh          F3=Cancel           F4=List
 F5=Reset            F6=Command          F7=Edit             F8=Image
 F9=Shell            F10=Exit            Enter=Do
```

Происходит создание файловой системы[1]/cltestfs. Мы проверяем содержимое файла /etc/filesystems на узле tiger и видим, что информация о новой файловой системе была добавлена в конец файла /etc/filesystems, как и ожидалось.

Добавление дополнительного vpath в группу томов

Этот пример показывает, как добавить новый vpath в существующую общую группу томов. Эта задача в C-SPOC выполняется следующим образом: **smit cl_admin > HACMP Logical Volume Management (Управление логическими томами HACMP) > Shared Volume Groups (Общие группы томов) > Set Characteristics of a Shared Volume Group (Назначение свойств общей группы томов) > Add a Volume to a Shared Volume Group (Добавление тома в общую группу томов).**

1 В оригинале руководства опечатка – «/cltestfs file is created», т. е. создание файла /cltestfs. Правильно читать – «файловой системы». *Примеч. науч. ред.*

Затем мы выбираем группу томов в экране SMIT, представленном в примере 8.14. В следующем экране выбираем устройства virtual path, которые требуется добавить в группу томов testvg:

¦> panther vpath11

¦> panther vpath6

После этого нужно просто подтвердить выполнение операции на экране информации о выбранных параметрах, представленном в примере 8.10.

Пример 8.10. Добавление vpath в общую группу томов через C-SPOC

```
Add a Volume to a Shared Volume Group

Type or select values in entry fields.
Press Enter AFTER making all desired changes.
                                                      [Entry Fields]
  Resource Group Name                                C10RG1
  VOLUME GROUP name                                  testvg
  Reference node                                     panther
  VOLUME names                                       vpath11 vpath6

F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Проверим конфигурацию vpath на узле tiger на экране конфигурации datapath следующим образом: **smit cl_admin > HACMP Physical Volume Management (Управление физическими томами HACMP) > Cluster Data Path Device Management (Управление устройствами Datapath в кластере) > Display Data Path Device Configuration (Вывод конфигурации устройства Datapath).**

Затем мы выбираем узел tiger. Выходные данные команды показывают, что новые устройства vpath были корректно добавлены в группу томов testvg.

Увеличение размера файловой системы

Следующий пример показывает, как увеличить размер файловой системы через C-SPOC. Эта операция выполняется следующим образом: **smit cl_admin > HACMP Logical Volume Management (Управление логическими томами HACMP) > Shared File Systems (Общие файловые системы) > Enhanced Journaled File Systems (Улучшенные файловые системы JFS) > Change/Show Characteristics of a Shared Enhanced Journaled File System (Изменение/вывод свойств общей улучшенной файловой системы JFS).**

После этого выводится список файловых систем, в котором мы выбираем файловую систему /cltestfs, как показано в примере 8.11.

Пример 8.11. Изменение файловой системы через C-SPOC – выбор файловой системы

```
¦          Enhanced Journaled File System Name and Resource Group          ¦
¦                                                                          ¦
¦ Move cursor to desired item and press Enter.                             ¦
¦                                                                          ¦
¦   # Resource Group         File System                                   ¦
¦     C10RG1                 /app1                                         ¦
¦     C10RG1                 /cltestfs                                     ¦
¦     C10RG2                 /app2                                         ¦
¦     C10RG3                 /app3                                         ¦
¦                                                                          ¦
¦ F1=Help                 F2=Refresh              F3=Cancel                ¦
¦ F8=Image                F10=Exit                Enter=Do                 ¦
¦ /=Find                  n=Find Next                                      ¦
```

После выбора файловой системы вводим дополнительный размер в гигабайтах, как показано в примере 8.12.

Пример 8.12. Изменение файловой системы через C-SPOC – поля на экране SMIT

```
Change/Show Characteristics of a Shared Enhanced Journaled File System

Type or select values in entry fields.
Press Enter AFTER making all desired changes.
                                                        [Entry Fields]
  Resource Group Name                                  C10RG1
  File system name                                     /cltestfs
  NEW mount point                                     [/cltestfs]
  SIZE of file system                                 [+2G]
  Mount GROUP                                         []
  PERMISSIONS                                          read/write            +
  Mount OPTIONS                                       []                     +
  Start Disk Accounting?                               no                    +
  Block Size (bytes)
  Inline Log?
  Inline Log size (MBytes)

F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Удаление файловой системы

Этот пример показывает, как удалить общую файловую систему из кластера через C-SPOC. Прежде чем начать операцию удаления файловой системы, необходимо вручную отключить эту файловую систему с локального узла. Это делается следующей командой:

umount /cltestfs

После этого надо выполнить следующее: **smit cl_admin > HACMP Logical Volume Management (Управление логическими томами HACMP) > Shared File Systems (Общие файловые системы) > Enhanced Journaled File Systems (Улучшенные файловые системы JFS) > Remove a Shared File System (Удаление общей файловой системы).**

Затем мы выбираем файловую систему из списка, подобного представленному в примере 8.11. После этого требуется просто подтвердить выполнение операции на следующем экране. Этот экран представлен в примере 8.13.

Пример 8.13. Экран удаление файловой системы через C-SPOC

```
                         Remove a Shared File System

 Type or select values in entry fields.
 Press Enter AFTER making all desired changes.
                                                        [Entry Fields]
   Resource Group Name                                 C10RG1
 * FILE SYSTEM name                                    /cltestfs              +
   Remove Mount Point                                  yes                    +


 F1=Help             F2=Refresh          F3=Cancel           F4=List
 F5=Reset            F6=Command          F7=Edit             F8=Image
 F9=Shell            F10=Exit            Enter=Do
```

Синхронизация определения группы томов между узлами

При внесении изменений в определение группы томов, логического тома или файловой системы на локальном узле с использованием команд AIX LVM вместо C-SPOC требуется вручную выполнить распространение изменений в LVM ODM на другие участвующие узлы. Синхронизация информации LVM ODM между узлами кластера с использованием C-SPOC выполняется следующим образом: запускается **smit cl_admin > HACMP Logical Volume Management (Управление логическими томами HACMP) > Synchronize a Shared Volume Group Definition (Синхронизация определения общей группы томов)**, после чего выбирается группа томов из списка SMIT, представленного в примере 8.14.

Пример 8.14. Синхронизация определения группы томов через

```
¦   #Resource Group                        Volume Group                  ¦
¦    C10RG1                                app1vg                        ¦
¦   #Resource Group                        Volume Group                  ¦
¦    C10RG2                                app2vg                        ¦
¦   #Resource Group                        Volume Group                  ¦
¦    C10RG3                                app3vg                        ¦
¦   #Resource Group                        Volume Group                  ¦
¦    C10RG1                                testvg
```

8.5 Синхронизация времени

Мы настоятельно рекомендуем выполнять синхронизацию времени на узлах кластера. Это необходимо для работы некоторых приложений, и, кроме того, выполнение демона xntpd упрощает администрирование кластера.

Конфигурирование NTP-клиентов

Если в вашей сети уже установлен сервер синхронизации времени, можно настроить узлы кластера на получение точного времени с этого сервера.

1. Измените файл /etc/ntp.conf таким образом, чтобы он включал следующие три строки:

server 192.169.1.254 # здесь следует указать IP-адрес вашего ntp-сервера времени

driftfile /etc/ntp.drift

tracefile /etc/ntp.trace

2. Запустите демон xntpd:

startsrc -s xntpd

3. Настройте xntpd на автоматический запуск во время загрузки. Удалите символ комментария в начале следующей строки в файле /etc/rc.tcpip:

start /usr/sbin/xntpd “$src_running”

Установка сервера времени

Если в сети нет сервера времени, можно легко установить его в кластере. Без внешнего источника информации о времени значение времени на узлах будет неточным, но одинаковым.

1. Выберите один из узлов на роль NTP-сервера. Желательно, чтобы это был узел с наивысшим приоритетом в кластере.

2. Измените файл /etc/ntp.conf на NTP-сервере:

disable auth

server 127.127.1.1 prefer # использовать локальное время как предпочтительное

fudge 127.127.1.1 stratum 4

driftfile /etc/ntp.drift

3. Отредактируйте файл /etc/ntp.conf на других узлах:

server 10.10.1.1 # здесь указывается базовый IP-адрес ntp-сервера времени вашего кластера

driftfile /etc/ntp.drift

tracefile /etc/ntp.trace

Мы рекомендуем использовать базовый адрес NTP-сервера или постоянный интерфейс, если такой существует.

4. Отредактируйте файл /etc/rc.tcpip на всех узлах, чтобы xntpd запускался автоматически:

start /usr/sbin/xntpd “$src_running”

5. Запустите xntpd на всех узлах:

startsrc -s xntpd

8.6 Верификация и синхронизация кластера

Верификация и синхронизация кластера HACMP обеспечивает правильность конфигурирования всех ресурсов, используемых или управляемых HACMP, а также согласованность всех правил владения ресурсами и прочих параметров между всеми узлами кластера.

Кластер HACMP хранит информацию обо всех ресурсах кластера и топологии кластера, а также некоторые дополнительные параметры в объектных классах HACMP в ODM. Для обеспечения надлежащего функционирования кластера ODM-файлы HACMP должны быть согласованы между всеми узлами кластера. В процессе верификации кластера проверяется согласованность ODM-файлов HACMP между всеми узлами, а также проверяется согласованность информации HACMP ODM с информацией AIX ODM. Успешное выполнение верификации означает, что конфигурация кластера синхронизирована между всеми узлами. На активном узле синхронизация вступает в действие немедленно. Синхронизация кластера синхронизирует ODM по всем узлам путем применения информации ODM с узла, на котором синхронизация была инициирована, на другие узлы.

Внимание! Несогласованность топологии кластера, ресурсов или других параметров между узлами в кластере может вызвать неправильную работу кластера.

Рекомендуется выполнять верификацию конфигурации кластера при каждом конфигурировании, переконфигурировании или обновлении кластера либо при изменении параметров операционной системы, которые могут повлиять на ресурсы кластера. Также рекомендуется периодически выполнять верификацию кластера.

8.6.1 Верификация и синхронизация кластера через SMIT

В SMIT (при выполнении команды **smit hacmp**) существует три различных пути выполнения верификации и синхронизации кластера:

- путь верификации **Initialization and Standard Configuration (Инициализация и стандартное конфигурирование);**
- путь верификации **Extended Configuration (Расширенное конфигурирование);**
- путь верификации **Problem Determination Tools (Инструменты определения проблем).**

Путь верификации Initialization and Standard Configuration (Инициализация и стандартное конфигурирование)

Можно использовать путь верификации Initialization and Standard Configuration (Инициализация и стандартное конфигурирование) следующим образом:

smit hacmp > Initialization and Standard Configuration (Инициализация и стандартное конфигурирование) > Verify and Synchronize HACMP Configuration (Верификация и синхронизация конфигурации HACMP).

При использовании пути верификации Initialization and Standard Configuration (Инициализация и стандартное конфигурирование) синхронизация автоматически выполняется после успешной верификации. Меню SMIT не содержит каких-либо дополнительных опций. При использовании этого пути функция автоматического исправления ошибок всегда активна. Дополнительные сведения о функции автоматического исправления ошибок см. в разделе 8.6.4, «Автоматическое выполнение корректирующих действий во время верификации».

Путь верификации Extended Configuration (Расширенное конфигурирование)

При использовании пути Extended Configuration (Расширенное конфигурирование) можно выбрать один из нескольких типов верификации, а также можно указать, следует ли выполнять синхронизацию.

Для использования пути верификации Extended Configuration (Расширенное конфигурирование) в своем кластере следует выполнить следующие действия:

1. Запустите **smit hacmp -> Extended Configuration (Расширенное конфигурирование) -> Extended Verification and Synchronization (Расширенная верификация и синхронизация)**.

2. Измените параметры в полях (для обычного выполнения верификации и синхронизации следует оставить параметры по умолчанию, как показано в примерах) и нажмите Enter.

3. На рис. 8.22 показан экран SMIT при активном кластере (DARE – Dynamic Reconfiguration).

4. На рис. 8.23 показан экран SMIT при неактивном кластере.

5. После выполнения верификации можно увидеть, была ли она выполнена успешно.

Параметры пути верификации Extended Verification and Synchronization (Расширенная верификация и синхронизация) зависят от того, активен ли кластер на узле, на котором инициируется верификация. В активном кластере экран SMIT содержит следующие параметры:

1. Emulate or Actual (Эмуляция или действительное выполнение). Опция **Emulate** выполняет верификацию в режиме эмуляции без применения изменений, тогда как опция **Actual** применяет изменения.

2. Verify changes only (Выполнить только верификацию изменений). Опция **No** выполняет полную проверку топологии и ресурсов, тогда как опция **Yes** проверяет только изменения, произошедшие со времени последней верификации. Эта функция относится только к файлам HACMP ODM!

3. Logging (Ведение журналов). Опция **Verbose** направляет все выходные данные, которые обычно записываются в файл clverify.log, на консоль.

При инициации верификации в неактивном кластере экран SMIT содержит следующие параметры:

```
        HACMP Verification and Synchronization (Active Cluster Nodes Exist)

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
* Emulate or Actual                                    [Actual]            +
* Verify changes only?                                 [No]                +
* Logging                                              [Standard]          +


F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Рис. 8.22. Экран синхронизации и верификации – активный кластер

```
                    HACMP Verification and Synchronization

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
* Verify, Synchronize or Both                          [Both]              +
* Automatically correct errors found during            [No]                +
  verificatio

* Force synchronization if verification fails?         [No]                +
* Verify changes only?                                 [No]                +
* Logging                                              [Standard]          +


F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Рис. 8.23. Экран синхронизации и верификации – неактивный кластер

1. Verify, Synchronize or both (Верификация, синхронизация или обе операции). Опция **Verify** выполняет только верификацию, опция **Synchronize** выполняет только синхронизацию, опция **Both** выполняет как верификацию, так и синхронизацию в соответствии со значением опции **Force synchronization if verification fails (Принудительная синхронизация при отказе верификации)**.

2. Automatically correct errors found during verification (Автоматическое исправление ошибок, обнаруженных во время верификации). Подробное описание см. в разделе 8.6.4, «Автоматическое выполнение корректирующих действий во время верификации».

3. Force synchronization if verification fails (Принудительная синхронизация при отказе верификации). Опция **No** останавливает процедуру синхронизации/верификации после обнаружения ошибок верификации, тогда как опция **Yes** выполняет синхронизацию принудительно, независимо от результатов верификации. В целом мы не рекомендуем включать принудительную синхронизацию. В тех ситуациях, когда включение принудительной синхронизации необходимо, нужно понимать возможные последствия внесения изменений в конфигурацию кластера.

4. Verify changes only (Выполнить только верификацию изменений). Опция **No** выполняет полную проверку топологии и ресурсов, тогда как опция **Yes** проверяет только изменения, произошедшие со времени последней верификации. Эта функция относится только к файлам HACMP ODM!

5. Logging (Ведение журналов). Опция **Verbose** направляет все выходные данные, которые обычно записываются в файл clverify.log, на консоль.

> **Примечание.** Синхронизация может быть инициирована как на активном, так и на неактивном кластере. Если некоторые узлы неактивны, синхронизацию можно инициировать только на активном узле с использованием DARE (Dynamic Reconfiguration). Дополнительные сведения о DARE см. в разделе 8.6.2, «Динамическая реконфигурация кластера – DARE».

Путь верификации Problem Determination Tools (Инструменты определения проблем)

Можно использовать путь верификации Problem Determination Tools (Инструменты определения проблем) путем запуска **smit hacmp > Problem Determination Tools (Инструменты определения проблем) > HACMP Verification (Верификация HACMP) > Verify HACMP Configuration (Верификация кластера HACMP)**.

Использование пути верификации Problem Determination Tools (Инструменты определения проблем) предоставляет больше опций верификации, в частности определение настраиваемого метода верификации, однако не предоставляет возможности выбрать синхронизацию. Экран SMIT пути верификации Problem Determination Tools (Инструменты определения проблем) представлен на рис 8.24.

> **Примечание.** Верификацию с использованием пути Problem Determination Tools (Инструменты определения проблем) можно инициировать как с активного, так и с неактивного узла.

В случае отказа верификации необходимо исправить ошибки и повторить верификацию как можно скорее. Сообщения верификации указывают, где произошла ошибка (например, на узле, на устройства или в команде). Расположение и назначение журналов верификации описывается в разделе 8.6.3, «Файлы журналов верификации».

```
                         Verify HACMP Configuration

  Type or select values in entry fields.
  Press Enter AFTER making all desired changes.
                                                        [Entry Fields]
    HACMP Verification Methods                         Pre-Installed        +
            (Pre-Installed, none)
    Custom Defined Verification Methods               []                    +
    Error Count                                       []                    #
    Log File to store output                          []
    Verify changes only?                              [No]                  +
    Logging                                           [Standard]            +


  F1=Help             F2=Refresh          F3=Cancel           F4=List
  F5=Reset            F6=Command          F7=Edit             F8=Image
  F9=Shell            F10=Exit            Enter=Do
```

Рис. 8.24. Экран верификации при использовании пути верификации Problem Determination Tools (Инструменты определения проблем)

8.6.2 Динамическая реконфигурация кластера – DARE

HACMP позволяет вносить некоторые изменения как в топологию, так и в ресурсы кластера при работающем кластере. Эта функция называется *динамической реконфигурацией (dynamic reconfiguration, DARE)*. Можно выполнить несколько изменений ресурсов и топологии одной операцией динамической реконфигурации, что упрощает работу в целом, особенно при сложных изменениях конфигурации.

DARE поддерживает внесение изменений в ресурсы и топологию в одной операции. Начиная с HACMP 5.3 DARE поддерживается в конфигурациях HACMP/XD.

> **Внимание!** 1. Не следует вносить изменения в конфигурацию или выполнять какие-либо действия, влияющие на ресурсы, если для какого-либо узла установлено принудительное отключенное состояние (forced down).
> 2. Во время миграции кластера на новую версию HACMP динамическая реконфигурация не поддерживается.

В активном кластере можно динамически внести следующие изменения в ресурсы кластера:

- Добавить, удалить или изменить сервер приложения.
- Добавить, удалить или изменить мониторинг приложения.
- Добавить или удалить содержимое одного или нескольких групп ресурсов.
- Добавить, удалить или изменить ресурс накопителя на магнитной ленте.
- Добавить или удалить один или несколько групп ресурсов.
- Добавить, удалить или изменить порядок участвующих узлов в группе ресурсов.

- Изменить отношения узлов в группе ресурсов.
- Изменить порядок обработки группы ресурсов.
- Добавить, удалить или изменить политику таймера возврата после восстановления, связанную с группой ресурсов. Новый таймер возврата после восстановления не будет действовать до тех пор, пока группа ресурсов не будет подключена на другом узле.
- Добавить, удалить или изменить время установления (settling time) для групп ресурсов.
- Добавить или удалить политику распределения узлов (node distribution policy) для групп ресурсов.
- Добавить, удалить или изменить зависимости «родительский объект/дочерний объект» («parent/child») или зависимости расположения для групп ресурсов (здесь применяются некоторые ограничения).
- Добавить, удалить или изменить политику межсайтового управления (inter-site management policy) для групп ресурсов.
- Добавить, удалить или изменить скрипты пред- и постобработки событий.

Динамическую реконфигурацию можно инициировать только с активного узла, т. е. с узла, на котором запущены демоны кластера. Изменения необходимо вносить с работающего узла, чтобы можно было осуществлять синхронизацию кластера. Прежде чем вносить изменения в определение кластера, нужно проверить следующее:

- На всех узлах должна быть установлена одинаковая версия HACMP.
- Некоторые узлы должны быть включены и выполнять HACMP, а также быть способны вести обмен данными друг с другом. Ни для одного узла не должно быть установлено принудительное отключенное состояние (forced down).
- Кластер должен быть в стабильном состоянии, и файл hacmp.out не должен содержать недавних ошибок или событий config_too_long.

В зависимости от конфигурации вашего кластера и от изменений, которые вы планируете внести в среду своего кластера, возможно множество различных ситуаций и ограничений при использовании утилиты динамической реконфигурации. Необходимо понимать возможные последствия некоторых изменений в кластере, поэтому мы рекомендуем обратиться к руководству **HACMP for AIX Administration Guide** за подробной информацией, прежде чем начинать вносить динамические изменений в среду своего кластера.

8.6.3 Файлы журналов верификации

Во время верификации кластера HACMP осуществляет сбор данных конфигурации со всех узлов в ходе выполнения наборов проверок. Подробные выходные данные сохраняются в файле /var/hacmp/clverify/clverify.log. При этом осуществляется чередование файлов журналов.

Приведенные ниже выходные данные содержат список файлов журналов верификации, находящихся в каталоге /var/hacmp/clverify/:

```
[p650n01][/var/hacmp/clverify]> ls -l
total 2024
-rw----     1 root     system     99451 Jun 21 00:00 clverify.log
-rw----     1 root     system     99956 Jun 20 12:13 clverify.log.1
-rw----     1 root     system     98639 Jun 20 00:00 clverify.log.2
-rw----     1 root     system     98639 Jun 19 00:00 clverify.log.3
-rw----     1 root     system     98639 Jun 18 00:00 clverify.log.4
-rw----     1 root     system     98549 Jun 17 18:46 clverify.log.5
-rw----     1 root     system     98549 Jun 17 18:24 clverify.log.6
-rw----     1 root     system     98866 Jun 17 18:02 clverify.log.7
-rw----     1 root     system     98866 Jun 17 17:49 clverify.log.8
-rw----     1 root     system     99091 Jun 17 16:58 clverify.log.9
-rw----     1 root     system     9296 Jun 21 00:00 clverify_daemon.log
drwx---     4 root     system     256 Jun 17 11:42 fail
drwx---     4 root     system     256 Jun 21 00:00 pass
drwx---     4 root     system     256 Jun 20 12:12 pass.prev
```

На узле, на котором инициируется утилита верификации, подробная информация собирается в файлах журналов, в которых регистрируются все собранные данные и выполненные задачи. Эти файлы журналов записываются в следующие каталоги и используются обслуживающим персоналом для определения расположения ошибок:

- /var/hacmp/clverify/pass/nodename/ – при успешной верификации;
- /var/hacmp/clverify/fail/nodename/ – при отказе верификации.

Примечание. Для выполнения верификации требуется 4 Мб дискового пространства на каждом узле в файловой системе /var. Как правило, файлы /var/hacmp/clverify/clverify.log требуют дополнительно 1–2 Мб дискового пространства. В кластере из четырех узлов рекомендуется иметь 18 Мб [свободного] дискового пространства.

8.6.4 Автоматическое выполнение корректирующих действий во время верификации

HACMP 5.3 позволяет автоматически выполнять некоторые корректирующие действия во время верификации и синхронизации кластера. Активизация этой опции зависит от используемого пути верификации и синхронизации.

Функция автоматического выполнения корректирующих действий может исправлять только некоторые типы ошибок, обнаруживаемых во время верификации кластера. В следующем списке представлены ошибки, которые можно исправить с использованием этой функции.

- Неактуальные отметки времени общей группы томов HACMP на узле.
- Файл /etc/hosts на узле содержит не все IP-адреса, управляемые HACMP.

- Группы томов SSA с одновременным доступом требуют использования уникальных номеров узлов SSA.
- На узле не создана файловая система, хотя диски доступны.
- Диски доступны, однако группа томов не была импортирована на узел.
- На узле отсутствуют необходимые записи /etc/services.
- На узле отсутствуют необходимые записи HACMP snmpd.

Путь верификации Initialization and Standard Configuration (Инициализация и стандартное конфигурирование)

При использовании пути верификации Initialization and Standard Configuration (Инициализация и стандартное конфигурирование) функция автоматического исправления ошибок всегда активна и не допускает отключения.

> **Примечание.** Во время динамической реконфигурации автоматические корректирующие действия не выполняются.

Путь верификации Extended Configuration (Расширенное конфигурирование)

При использовании пути верификации Extended Configuration (Расширенное конфигурирование) возможность активизации функции автоматического исправления ошибок зависит от состояния кластера. Можно либо отключить эту функцию, либо запустить ее в одном из двух режимов:

- Interactively (Интерактивный; опция меню **Interactively**) – верификация определяет исправимое состояние, связанное с импортом группы томов или экспортом и реимпортом точек подключения и файловых систем, и запрашивает подтверждения выполнения корректирующего действия, прежде чем продолжить верификацию;
- Automatically (Автоматический; опция меню **Yes**) – верификация определяет, что имеет место одно из ошибочных состояний, перечисленных в разделе «Состояния, которые могут вызвать корректирующее действие», и инициирует корректирующее действие автоматически без запроса.

Если кластер неактивен, можно выбрать режим автоматического исправления ошибок непосредственно в меню пути верификации Extended Configuration (Расширенное конфигурирование), выполнив **smit hacmp -> Extended Configuration (Расширенное конфигурирование) -> Extended Verification and Synchronization** (**Расширенная верификация и синхронизация**), как показано на рис. 8.23. Изменение режима выполняется путем установки в поле Automatically correct errors found during verification (Автоматическое исправление ошибок, обнаруженных во время верификации) значений **Yes**, **No** или **Interactively**.

Если кластер активен, функция автоматического исправления ошибок включена по умолчанию. Можно изменить режим функции автоматического исправления ошибок в активном кластере непосредственно в меню SMIT кластера. Следует запустить **smit hacmp > System Management (C-SPOC) > Manage HACMP Services (Управление службами HACMP) > Start Cluster Services (Запустить службы кластера)** и выбрать одно из значений **Yes**, **No** или **Interactive**. Это установит режим автоматического исправления ошибок для следующих ситуаций:

1. Для пути верификации Extended Configuration (Расширенное конфигурирование).

2. Для автоматической верификации кластера при запуске служб кластера на узле или реинтеграции узла в кластер. Дополнительные сведения по этой теме см. в разделе 8.6.5, «Автоматическая верификация кластера».

3. Для периодически выполняемой автоматической верификации кластера. Дополнительные сведения по этой теме см. в разделе 8.6.5, «Автоматическая верификация кластера».

Путь верификации Problem Determination Tools (Инструменты определения проблем)

При использовании этого пути верификации активизация функции автоматического исправления ошибок невозможна.

8.6.5 Автоматическая верификация кластера

HACMP выполняет автоматическую верификацию в следующих случаях:

- при каждом запуске служб кластера на узле;
- при каждой реинтеграции узла в кластер;
- каждые 24 ч.

Во время автоматической верификации и синхронизации HACMP выполняет обнаружение и исправление некоторых общих ошибок конфигурации. При таком автоматическом исправлении ошибок, если в кластере не была выполнена верификация и синхронизация вручную перед запуском служб кластера, HACMP выполнит это автоматически. Автоматическую верификацию и синхронизацию часто называют просто верификацией.

Через меню SMIT можно установить параметры периодического выполнения утилиты **Automatic cluster verification checking**[1], выполнив **smit hacmp > Problem Determination Tools (Инструменты определения проблем) > HACMP Verification (Верификация HACMP) > Automatic Cluster Configuration Monitoring (Автоматический мониторинг конфигурации кластера)**. Экран SMIT содержит следующие поля:

- Automatic cluster configuration verification (Автоматическая верификация конфигурации кластера). Здесь можно включить или отключить утилиту, выбрав **Disable** или **Enable**.
- Node name (Имя узла). Здесь можно выбрать узлы, на которых должна выполняться утилита. Можно выбрать либо опцию **default**, которая обозначает выбор всех узлов, либо определенный узел.
- HOUR (00–23). Здесь определяется время запуска утилиты. По умолчанию установлено значение 00:00 (полночь), которое можно изменить на любое другое значение времени.

[1] Имеется в виду Automatic cluster configuration verification. *Примеч. науч. ред.*

```
                Automatic Cluster Configuration Monitoring

  Type or select values in entry fields.
  Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
  * Automatic cluster configuration verification        Enabled            +
    Node name                                           Default            +
  * HOUR (00 - 23)                                     [00]                +

  F1=Help             F2=Refresh          F3=Cancel           F4=List
  F5=Reset            F6=Command          F7=Edit             F8=Image
  F9=Shell            F10=Exit            Enter=Do
```

Рис. 8.25. Automatic Cluster Configuration Monitoring (Автоматический мониторинг конфигурации кластера

На рис. 8.25 показан экран Automatic Cluster Configuration Monitoring (Автоматический мониторинг конфигурации кластера).

smit clautover.dialog

Результат автоматической верификации кластера можно просмотреть в файлах журналов верификации. По умолчанию они находятся в каталоге /var/hacmp/clverify/. Дополнительные сведения о файлах журналов верификации см. в разделе 8.6.3, «Файлы журналов верификации».

8.7 Мониторинг HACMP

HACMP обеспечивает высокую доступность среды приложений путем преднамеренного маскирования или устранения отказов аппаратного и программного обеспечения в среде высокой доступности. Маскирование отказов означает перемещение активных ресурсов с отказавшего компонента на резервный компонент. Таким образом, все приложения высокой доступности продолжают работать и клиенты осуществляют к ним доступ и используют их, несмотря на отказ.

В результате вы можете даже и не знать об отказе компонента кластера. Опасность такой ситуации заключается в том, что, хотя HACMP может продолжать работу после одного или даже нескольких отказов, каждый отказ, ускользающий от вашего внимания, неблагоприятно влияет на способность кластера обеспечивать среду высокой доступности, так как сокращается избыточность компонентов кластера.

Во избежание такой ситуации мы рекомендуем выполнять регулярные проверки и мониторинг кластера. Кроме того, HACMP содержит различные средства, помогающие осуществлять мониторинг кластера, в частности следующие:

- Автоматическая верификация кластера. Дополнительные сведения см. в разделе 8.6.5, «Автоматическая верификация кластера».
- Утилиты проверки состояния кластера.
- Команды информации о группах ресурсов.
- Команды информации о топологии.

- Файлы журналов.
- Уведомления об ошибках.
- Мониторинг приложений.
- Измерение доступности приложений.
- Мониторинг кластеров с использованием инструментов системного администрирования и мониторинга на уровне предприятия (Tivoli, NetView®).

Для конфигурирования и управления кластером можно использовать либо ASCII SMIT, либо WebSMIT.

8.7.1 Утилиты проверки состояния кластера

clstat

/usr/es/sbin/cluster/clstat – очень полезный инструмент, который можно использовать для мониторинга состояния кластера. Он применяет процедуры библиотеки **clinfo** для вывода различной информации о кластере, включая имя и состояние узлов, интерфейсов и групп ресурсов.

Эта утилита требует, чтобы подсистема **clinfoES** была активна на всех узлах, на которых инициируется команда **clstat.**

Утилита clstat имеет два режима работы: режим ASCII и режим X Window. Режим ASCII может выполняться на любом физическом или виртуальном ASCII-терминале, включая xterm или aixterm. Если узел кластера работает в графическом режиме, **clstat** выводит выходные данные в графическом окне. Перед выполнением команды необходимо убедиться в том, что переменная DISPLAY была экспортирована на X-сервер и что разрешен доступ X-клиентов.

Синтаксис команды clstat представлен на рис. 8.26.

```
(single ASCII display)
  clstat -o [-c ID | -n name] [-s]

  (ASCII mode)
  clstat -a [-c ID | -n name] [-i] [-r interval][-s]

  (X mode)
   xclstat [-c ID | -n name] [-r interval] [-D debug_level][-s]
```

Рис. 8.26. Синтаксис команды clstat

clstat -a //запускает программу в режиме ASCII.

clstat -o //запускает программу в режиме ASCII и выполняет выход (полезно для получения выходных данных из shell-скрипта или задания cron).

clstat -s //выводит включенные и отключенные сервисные метки; в противном случае выводит только активные сервисные метки.

Пример 8.15. Выходные данные команды clstat -o

```
clstat - HACMP Cluster Status Monitor
                ------------------------------------

Cluster: migr1  (1120388255)
Wed Jul  6 14:13:15 CDT 2005
                State: UP               Nodes: 3
                SubState: STABLE

        Node: panther           State: UP
           Interface: panther1_base (0)         Address: 10.10.31.36
                                                State:   UP
           Interface: panther2_base (0)         Address: 10.10.32.36
                                                State:   DOWN
           Interface: tty1_patnh (1)            Address: 0.0.0.0
                                                State:   UP
           Interface: tty2_panth (2)            Address: 0.0.0.0
                                                State:   UP
           Interface: c1app1svc (0)             Address: 192.168.100.86
                                                State:   UP
           Resource Group: C1ORG1                       State:  On line
 Node: puma              State: UP
    Interface: puma1_base (0)            Address: 10.10.31.35
                                         State:   UP
    Interface: puma2_base (0)            Address: 10.10.32.35
                                         State:   DOWN
    Interface: tty1_puma (1)             Address: 0.0.0.0
                                         State:   UP
    Interface: tty_puma (3)              Address: 0.0.0.0
                                         State:   UP
    Interface: c1app2svc (0)             Address: 192.168.100.85
                                         State:   UP
    Resource Group: C1ORG2                       State:  On line

 Node: tiger             State: UP
    Interface: tiger1_base (0)           Address: 10.10.31.34
                                         State:   UP
    Interface: tiger2_base (0)           Address: 10.10.32.34
                                         State:   DOWN
    Interface: tty2_tiger (2)            Address: 0.0.0.0
                                         State:   UP
    Interface: tty_tiger (3)             Address: 0.0.0.0
                                         State:   UP
    Interface: c1app3svc (0)             Address: 192.168.100.84
                                         State:   UP
    Resource Group: C1ORG3                       State:  On line
```

Пример 8.15 показывает выходные данные команды **clstat -o** на нашем тестовом кластере.

cldump

Еще одной полезной утилитой является утилита **cldump (/usr/es/sbin/cluster/utilities/cldump)**. Она создает снимки основных компонентов кластера: самого кластера, узлов в кластере, сетевых интерфейсов, подключенных к узлам, а также состояние групп ресурсов на каждом узле.

Утилита cldump не имеет каких-либо опций, так что нужно просто запустить cldump из командной строки.

8.7.2 Утилиты проверки состояния кластера и служб

lssrc

Состояние подсистемы HACMP или RSCT можно проверить с помощью команды **lssrc** с опцией **-s** или **-g**. Она выводит имя, группу, PID и состояние (активное или нерабочее) подсистемы.

lssrc -s subsystem_name//выводит информацию об определенной подсистеме.

lssrc -g subsystem_group_name //выводит информацию обо всех подсистемах в определенной группе.

> **Внимание!** Начиная с HACMP 5.3 демон диспетчера кластера clstrmgrES инициируется из процесса init, так что он автоматически запускается при загрузке. Диспетчер кластера (Cluster Manager) должен быть запущен, прежде чем можно будет запустить какие-либо службы кластера на узле. Так как демон clstrmgr является постоянно выполняющимся процессом, нельзя использовать команду lssrc -s clstrmgrES для определения состояния кластера. Вместо этого следует использовать команду /usr/es/sbin/cluster/clstat или любую другую утилиту, описанную в этом разделе.

На рис. 8.27 показаны имена подсистем и имена групп для всех подсистем, используемых в HACMP.

```
Subsystem_name   group_name

RSCT subsystems used by HACMP:
 topsvcs          topsvcs
 grpsvcs          grpsvcs
 grpglsm          grpsvcs
 emsvcs           emsvcs
 emaixos          emsvcs
 ctrmc            rsct
HACMP subsystems:
 clcomdES         clcomdES
 clstrmgrES       cluster
optional HACMP subsystems:
 clinfoES         cluster
```

Рис. 8.27. Имена подсистем и имена групп, используемых в HACMP

clshowsrv

Существует еще один способ вывода информации о состоянии подсистем HACMP с использованием команды **clshowsrv (/usr/es/sbin/cluster/utilities/cldump)**. Она выводит состояние всех подсистем, используемых в HACMP, или состояние выбранной подсистемы. Формат выходных данных команды соответствует формату команды **lssrc -s**.

На рис. 8.28 представлен синтаксис команды clshowsrv.

```
clshowsrv [-a|-v] [clstrmgrES|clinfoES|clcomdES]
```

Рис. 8.28. Синтаксис команды clshowsrv

clshowsrv -a //выводит состояние подсистемы HACMP: clstrmgrES, clinfoES и clcomdES.

clshowsrv -v // выводит состояние подсистем HACMP и RSCT.

Пример 8.16 показывает выходные данные команды **clshowres** из нашего тестового кластера при запущенных службах кластера.

Пример 8.16. Выходные данные команды clshowres -v

```
Status of the RSCT subsystems used by HACMP:
Subsystem         Group            PID          Status
 topsvcs          topsvcs          22756        active
 grpsvcs          grpsvcs          21858        active
 grpglsm          grpsvcs                       inoperative
 emsvcs           emsvcs           24932        active
 emaixos          emsvcs           28982        active
 ctrmc            rsct             13430        active

Status of the HACMP subsystems:
Subsystem         Group            PID          Status
 clcomdES         clcomdES         15738        active
 clstrmgrES       cluster          26498        active

Status of the optional HACMP subsystems:
Subsystem         Group            PID          Status
 clinfoES         cluster          26260        active
```

Также можно выполнить команду **clshowsrv -v**, используя меню SMIT: **smit hacmp > System Management (C-SPOC) > Manage HACMP Services (Управление службами HACMP) > Show Cluster Services (Вывод служб кластера)**.

8.7.3 Команды информации о топологии

cltopinfo

Команда **cltopinfo (/usr/es/sbin/cluster/utilities/cltopinfo)** выводит информацию о топологии кластера в более простом формате.

На рис. 8.29 представлен синтаксис команды **cltopinfo**.

```
cltopinfo [-c] [-n] [-w] [-i]
```

Рис. 8.29. Синтаксис команды cltopinfo

Также можно использовать меню SMIT для вывода информации о топологии в различных форматах; для этого следует запустить **smit hacmp > Extended Configuration (Расширенное конфигурирование) > Extended Topology Configuration (Расширенное конфигурирование топологии) > Show HACMP Topology (Отображение топологии HACMP)** и выбрать требуемый формат. На рис. 8.30 показаны меню SMIT для вывода информации о топологии с различными опциями формата. Эти опции согласуются с опциями команды, представленными на рис. 8.29.

smit cm_show_menu

```
                                Show HACMP Topology

    Move cursor to desired item and press Enter.

      Show Cluster Topology
      Show Cluster Definition
      Show Topology Information by Node
      Show Topology Information by Network
      Show Topology Information by Communication Interface


    F1=Help             F2=Refresh          F3=Cancel           F8=Image
    F9=Shell            F10=Exit            Enter=Do
```

Рис. 8.30. Меню топологии кластера

Служба topsvcs

Можно выполнить команду **lssrc -ls topsvcs** для мониторинга активности пульса на основе службы топологии. Выходные данные деятельности демона **topsvcs** содержат всю информацию по мониторингу пульсации для всех активных сетевых путей. Информация об определенных параметрах активности пульса содержится в полях **Missed HBs (Пропущенные импульсы)**, **Packets sent (Отправленные пакеты)**, **Packets received (Полученные пакеты)** и **Errors (Ошибки)** в выходных данных того или иного сетевого пути. В примере 8.17 представлен фрагмент выходных данных команды **`lssrc -ls topsvcs`** в нашем тестовом кластере. Выделенные слова указывают интересующую информацию в выходных данных.

Пример 8.17. Выходные данные команды lssrc -ls topsvcs

```
Subsystem         Group            PID     Status
 topsvcs          topsvcs          811144  active
Network Name   Indx Defd Mbrs St Adapter ID      Group ID
migr1_eth_01_0 [ 0]    3    3  S 11.11.1.11      11.11.1.13
```

```
migr1_eth_01_0 [ 0] en2          0x42cc1f00      0x42cc1f1b
HB Interval = 1.000 secs. Sensitivity = 10 missed beats
Missed HBs: Total: 64 Current group: 64
Packets sent    : 8893 ICMP 17 Errors: 0 No mbuf: 0
Packets received: 10616 ICMP 31 Dropped: 0
NIM's PID: 774256
rs232_1        [ 3]    2    2  S 255.255.0.2     255.255.0.2
rs232_1        [ 3] tty1         0x82cc1f02      0x82cc1f05
HB Interval = 2.000 secs. Sensitivity = 5 missed beats
Missed HBs: Total: 0 Current group: 0
Packets sent    : 5841 ICMP 0 Errors: 0 No mbuf: 0
Packets received: 6250 ICMP 0 Dropped: 0
NIM's PID: 893118
```

8.7.4 Команды информации о группах ресурсов

clrginfo

Используя команду **clRGinfo (/usr/es/sbin/cluster/utilities/clRGinfo)**, можно вывести атрибуты группы ресурсов в кластере. Выходные данные команды представляют отчет о расположении и состоянии одной или нескольких заданных групп ресурсов. Выходные данные команды представляют как глобальное состояние, так и конкретное состояние группы ресурсов на локальном узле.

На рис. 8.31 представлен синтаксис команды **clRGinfo.**

```
clRGinfo [-h][-v][-a][-s|-c][-p][-t][-d][groupname1] [groupname2] ...
```

Рис. 8.31. Синтаксис команды clRGinfo

clRGinfo -v // выводит расположение, отменяющее приоритет (priority override location), и активные таймеры группы ресурсов.

clRGinfo -p // выводит параметры запуска, перемещения при сбое и возврата после восстановления для группы ресурсов.

clRGinfo -t // запрашивает диспетчер кластера (Cluster Manager) только на локальном узле.

clRGinfo -c // команда выводит выходные данные в формате с разделяющими двоеточиями.

clRGinfo -a // команда выводит информацию о том, какие операции с группами ресурсов выполняются во время текущего события кластера (если она выполняется во время обработки события кластера).

Существуют следующие состояния групп ресурсов:

- **Online**. Группа ресурсов в настоящий момент работает должным образом.
- **Offline**. Группа ресурсов в настоящий момент не работает в кластере и не находится в состоянии ошибки.
- **Acquiring**. Группа ресурсов в настоящий момент подключается на одном из узлов кластера. При нормальных условиях состояние должно измениться на Online.

- **Releasing**. Группа ресурсов находится в процессе освобождения от владения одним из узлов. При нормальных условиях состояние должно измениться на Offline.
- **Error**. Группа ресурсов находится в состоянии ошибки. Требуется вмешательство пользователя.
- **Unknown**. Текущее состояние группы ресурсов нельзя определить в связи с потерей связи из-за отказа какого-либо ресурса в группе ресурсов или в связи с несоответствием зависимости группы ресурсов.

Если на локальном узле не запущены службы кластера, команда определяет узел, на котором службы кластера активны, и получает информацию группы ресурсов от активного диспетчера кластера.

Вместо команды clrginfo можно использовать команду **clfindres**, представляющую собой ссылку на clRGinfo (**/usr/es/sbin/cluster/utilities/clfindres**).

В примере 8.18 представлены выходные данные команды clRGinfo в нашем тестовом кластере.

Пример 8.18. Выходные данные команды clRGinfo

```
-----------------------------------------------------------------------------
Group Name     Group State                  Node
-----------------------------------------------------------------------------
rg1            ONLINE                       p650n01
               OFFLINE                      p650n02

rg2            ONLINE                       p650n02
               OFFLINE                      p650n01
```

8.7.5 Файлы журналов

HACMP хранит всю информацию о кластере и записывает генерируемые и выводимые в консоль сообщения в несколько файлов журналов. Каждый файл журнала содержит отдельный набор типов сообщений, генерируемых HACMP, поэтому просмотр различных файлов журналов дает разное представление о состоянии кластера.

Использование утилиты C-SPOC позволяет выполнять следующие действия над файлами журналов:

- View/Save/Delete HACMP Event Summaries (Просмотр/сохранение/удаление обзоров событий HACMP). Эта опция позволяет выводить содержимое, сохранять или удалять обзоры событий кластера.
- View Detailed HACMP Log Files (Просмотр подробных файлов журналов HACMP). Эта опция позволяет вывести журнал скриптов HACMP (/tmp/hacmp.out), системный журнал HACMP (/usr/es/adm/cluster.log), системный журнал C-SPOC (/tmp/cspoc.log).
- Change/Show HACMP Log File Parameters (Изменение/вывод параметров файла журнала HACMP). Эта опция позволяет задать уровень отладки (высокий/низкий) и вариант форматирования (default, standard, html-low, html-high) для заданного узла.
- Change/Show Cluster Manager Log File Parameters (Изменение/вывод параметров файла журнала диспетчера кластера). Эта опция позволяет задать уровень отладки диспетчера кластера (standard/high)

- Change/Show a Cluster Log Directory (Изменение/вывод каталога журнала кластера). Это меню позволяет определить новый каталог для заданного файла журнала, как описано ниже.
- Collect Cluster log files for Problem Reporting (Сбор файлов журналов кластера для отчетов о проблемах). Эта функция используется для сбора мгновенных данных (snap data) о кластере (командой clsnap), необходимых для дополнительного определения и анализа проблем. Здесь можно выбрать опцию отладки, включить файлы журналов RSCT и выбрать узлы, включенные в этот набор данных. По умолчанию набор мгновенных данных хранится в каталоге /tmp/ibmsupt/hacmp/ (для команды clsnap) и, например, в каталоге /tmp/phoenix.snapOut (для узла phoenix).

Список всех журналов HACMP с описанием их назначения:

- strmgr.debug: генерируется демоном clstrmgrES, каталог по умолчанию – [/tmp].
- cluster.log: генерируется скриптами и демонами кластера, каталог по умолчанию – [/usr/es/adm].
- cluster.mmddyyyy: файлы истории кластера, генерируются ежедневно, каталог по умолчанию – [/usr/es/sbin/cluster/history].
- cl_sm.log: генерируется библиотекой кластера Shared Memory, каталог по умолчанию – [/tmp].
- cspoc.log: генерируется командами C-SPOC, каталог по умолчанию – [/tmp].
- dms_loads.out: генерируется при работе deadman's switch, каталог по умолчанию – [/tmp].
- emuhacmp.out: генерируется скриптами эмулятора событий, каталог по умолчанию – [/tmp].
- hacmp.out: генерируется при выполнении скриптов и утилит обработки событий, каталог по умолчанию – [/tmp].
- clavan.log: генерируется утилитой доступности приложения (Application Availability), каталог по умолчанию – [/var/adm].
- clverify.log: генерируется утилитой верификации кластера (Cluster Verification), каталог по умолчанию – [/var/hacmp/clverify].
- clcomd.log: генерируется демоном clcomd, каталог по умолчанию – [/var/hacmp/clcomd].
- clcomddiag.log: генерируется демоном clcomd, отладочная информация, каталог по умолчанию – [/var/hacmp/clcomd].
- clconfigassist.log: генерируется инструментом Two-Node Cluster Configuration Assistant, каталог по умолчанию – [/var/hacmp/log].
- clutils.log: генерируется утилитами кластера, каталог по умолчанию – [/var/hacmp/log].
- cl_testtool.log: генерируется инструментом тестирования кластера (Cluster Test Tool), каталог по умолчанию – [/var/hacmp/log].

Необходимо обеспечить достаточно пространства для всех файлов журналов в файловых системах. Необходимое количество пространства в файловой системе /var

зависит от количества узлов в кластере. Определить общее значение для каждого узла можно, исходя из следующих соображений:

- 2 Мб должно быть свободно для записи файлов clverify.log[0–9];
- 4 Мб на узел для записи данных верификации с узлов;
- 20 Мб для записи информации журнала clcomd;
- 1 Мб на узел для записи данных кеша ODM.

Например, для кластера из четырех узлов потребуется 2 + 4x4 + 20 + 4x1 = 42 Мб свободного пространства в файловой системе /var.

Определить необходимое количество пространства в каталоге /tmp очень сложно, так как в этом каталоге находятся некоторые файлы с отладочной информацией, а также некоторые нечередующиеся файлы журналов. Размер этих журналов зависит от операций, конфигурации и состояния кластера. Из практического опыта мы рекомендуем выделить около 50 Мб свободного места в каталоге /tmp.

Через меню SMIT можно изменить стандартный каталог для определенного файла журнала; для этого нужно запустить **smit hacmp > System Management (C-SPOC) > HACMP Log Viewing and Management (Просмотр и управление журналами HACMP) > Change/Show a Cluster Log Directory (Изменение/вывод каталога журнала кластера)**, после чего выбрать определенный файл журнала (быстрый путь SMIT: **smit clusterlog_redir.select)**. Затем выполняется изменение стандартного каталога журналов для всех узлов в кластере. После изменения параметров журнала необходимо выполнить синхронизацию кластера.

> **Внимание!** Мы рекомендуем в качестве нового расположения для журналов использовать локальные файловые системы, а не общие файловые системы или файловые системы NFS. Хранение журналов на общих файловых системах или файловых системах NFS может вызвать проблемы, если потребуется отключить файловую систему во время перемещения при сбое. Кроме того, перенаправление журналов в общие файловые системы или файловые системы NFS может вызвать сбой при запуске служб кластера во время реинтеграции узла.

Кроме того, кластер генерирует некоторые файлы отладки. (Они находятся в каталоге /tmp. Содержимое этих файлов зависит от установленного уровня отладки):

- clinfo.debug – записываются выходные данные, генерируемые при работе скриптов обработки событий;
- clsmuxtrmgr.debug – файл журнала функции smux peer;
- clstlrmgr.debug – содержит отформатированные сообщения с отметками времени, генерируемые clstrmgrES.

8.7.6 Уведомление об ошибках

Средство уведомления об ошибках (Error Notification) AIX 5L можно использовать для добавления дополнительного уровня высокой доступности в среде HACMP. Можно добавить уведомления об отказах ресурсов, для которых HACMP не выполняет восстановление по умолчанию.

Дополнительные сведения об автоматических уведомлениях об ошибках, использовании и конфигурировании уведомлений об ошибках с некоторыми примерами см. в разделе 12.3, «Уведомления об ошибках».

8.7.7 Мониторинг приложения

После запуска служб кластера на определенном узле и подключения всех ресурсов рекомендуется проверить, все ли приложения запущены и все ли службы, предоставляемые приложениями, доступны. Также можно проверить, все ли процессы приложения выполняются и все ли другие ресурсы (в частности, файловые системы), необходимые для работы приложения, доступны.

Начиная с HACMP 5.2 можно конфигурировать несколько мониторов приложений и связывать их с одним или несколькими серверами приложений. Каждому монитору через SMIT можно назначить уникальное имя.

Однако кластер HACMP содержит возможность автоматического мониторинга требуемого приложения. В том случае, если HACMP обнаруживает прекращение работы процесса или отказ приложения, он пытается их перезапустить. Существует два способа выполнения мониторинга приложения:

- система мониторинга процессов приложения обнаруживает прекращение работы одного или нескольких процессов приложения с использованием RSCT Resource Monitoring and Control (RMC);
- настраиваемый метод мониторинга приложения регулярно проверяет работу приложения через интервал, заданный пользователем.

Конфигурирование мониторинга приложения через меню SMIT выполняется путем запуска **smit hacmp -> Extended Resource Configuration**[1] **(Расширенное конфигурирование ресурсов) -> Extended Resource Configuration (Расширенное конфигурирование ресурсов) -> HACMP Extended Resources Configuration (Расширенное конфигурирование ресурсов HACMP) -> Configure HACMP Applications (Конфигурирование приложений HACMP) -> Configure HACMP Application Monitoring (Конфигурирование мониторинга приложений HACMP)**, после чего выбирается либо меню **Configure Process Application Monitors (Конфигурирование мониторов процессов приложений)**, либо меню **Configure Custom Application Monitors (Конфигурирование настраиваемых мониторов приложений)**. Кроме того, можно использовать быстрый путь **smit cm_cfg_appmon**.

Мониторинг процессов приложений

Средство мониторинга процессов приложений использует встроенную функцию мониторинга, реализованную в RSCT, и не требует каких-либо дополнительных скриптов. Оно обнаруживает только прекращение работы процесса приложения и неспособно обнаружить какое-либо другое нарушение в работе приложения.

Когда HACMP обнаруживает прекращение работы какого-либо процесса приложения, он пытается перезапустить приложение на текущем узле заданное количество раз.

---

[1] Здесь и далее в этой главе ошибка в путях меню smit: smit hacmp -> Extended Configuration (потом – по книге). *Примеч. науч. ред.*

Для добавления нового монитора приложения следует запустить **smit hacmp > Extended Resource Configuration (Расширенное конфигурирование ресурсов) > Extended Resource Configuration (Расширенное конфигурирование ресурсов) > HACMP Extended Resources Configuration (Расширенное конфигурирование ресурсов HACMP) > Configure HACMP Applications (Конфигурирование приложений HACMP) > Configure HACMP Application Monitoring (Конфигурирование мониторинга приложений HACMP) > Configure Process Application Monitors (Конфигурирование мониторов процессов приложений) > Add a Process Application Monitor (Добавить монитор процессов приложения)** либо использовать быстрый путь **smit cm_cfg_process_appmon**. На рис. 8.32 показан экран SMIT с полями конфигурирования монитора приложения нашего тестового процесса.

```
 Add a Process Application Monitor

Type or select values in entry fields.
Press Enter AFTER making all desired changes.
                                                      [Entry Fields]
* Monitor Name                                       [APP1_monitor]
* Application Server(s) to Monitor                    APP1                +
* Monitor Mode                                       [Long-running monito>+
* Processes to Monitor                               [app1d app1testd]
* Process Owner                                      [root]
  Instance Count                                     [1]                 #
* Stabilization Interval                              [120]               #
* Restart Count                                      [5]                 #
  Restart Interval                                   [600]               #
* Action on Application Failure                      [notify]            +
Notify Method                                       [/usr/local/App1Mon.sh>
  Cleanup Method                                     [/usr/local/App1Stop]
  Restart Method                                     [/usr/local/App1Start]

F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Рис. 8.32. Экран добавления монитора процесса приложения

Мы определили монитор приложения **APP1_monitor** для монитора приложения **APP1**. Мы использовали стандартный режим мониторинга: **Long-running monitoring (Долгосрочный мониторинг)**. В этом режиме монитор приложения периодически проверяет, работает ли сервер приложения. Проверка начинается по истечении заданного стабилизационного интервала **(Stabilization interval)**. Альтернативные варианты – **Startup Monitoring (Мониторинг при запуске)** и **both (оба режима)**. При выборе опции **Startup Monitoring (Мониторинг при запуске)** монитор приложения проверяет успешность запуска сервера приложения в заданном стабилизационном интервале. Мы установили для параметра **Stabilization interval (Стабилизационный интервал)** значение 120 с.

Мы определили процессы app1d и app1testd, владельцем которых является root, и для которых будет осуществляться мониторинг с использованием данного монитора приложений.

Мы можем определить два разных варианта работы монитора в том случае, если отказ не был устранен после всех попыток, в соответствии со значением поля **Restart Interval (Интервал перезапуска)**. Эти варианты устанавливаются через поле **Action on Application Failure (Действие при отказе приложения)**. Возможные значения поля: **notify** и **fallover**. Если выбрано значение notify, после выполнения метода уведомления никаких действий не предпринимается. Если выбрано значение fallover, группа ресурсов, содержащая приложение, для которого осуществляется мониторинг, перемещается на другой узел в кластере.

Параметр **Notification method (Метод уведомления)** определяет скрипт, выполняющийся при каждом перезапуске приложения, полном отказе или перемещении на следующий узел в кластере. Настоятельно рекомендуется сконфигурировать этот метод.

Cleanup Method (Метод очистки) и **Restart Method (Метод перезапуска)** определяют скрипты остановки и запуска приложения после обнаружения отказа. Заданные по умолчанию скрипты запуска и остановки используются как определенные в конфигурации сервера приложения.

Настраиваемый мониторинг приложения

Настраиваемый монитор приложения представляет другой вариант мониторинга доступности приложения с использованием настраиваемых скриптов, которые могут имитировать доступ клиента к службам, предоставляемым приложением. Он использует встроенную функцию мониторинга, обеспечиваемую RSCT, и не требует настраиваемого скрипта[1]. В зависимости от кода завершения этого скрипта, монитор определяет, доступно ли приложение. Если скрипт завершается с кодом 0, это означает, что приложение доступно. Любой другой код завершения означает, что приложение недоступно.

Можно добавить новый настраиваемый монитор приложения с использованием SMIT путем запуска **smit hacmp > Extended Resource Configuration (Расширенное конфигурирование ресурсов) > Extended Resource Configuration (Расширенное конфигурирование ресурсов) > HACMP Extended Resources Configuration (Расширенное конфигурирование ресурсов HACMP) > Configure HACMP Applications (Конфигурирование приложений HACMP) > Configure HACMP Application Monitoring (Конфигурирование мониторинга приложений HACMP) > Configure Custom Application Monitors (Конфигурирование настраиваемых мониторов приложений) > Add a Custom Application Monitor (Добавление настраиваемого монитора приложения)** или используя быстрый путь **smit cm_cfg_custom_appmon**. Экран SMIT и его опции добавления метода в конфигурацию кластера подобны экрану добавления монитора процесса приложения, как показано на рис. 8.32.

---

[1] Здесь авторы ошиблись. Для использования этого метода необходимы скрипты (или программы), разработанные специально для мониторинга состояния конкретного приложения в конкретной среде. *Примеч. науч. ред.*

В меню конфигурирования настраиваемых мониторов приложения отличаются следующие поля:

- **Monitor Method (Метод мониторинга).** Указывает полный путь к скрипту, определяющему метод проверки состояния приложения. Если приложением является база данных, этот скрипт может подключиться к базе данных и выполнить SQL-запрос select для заданной таблицы базы данных. Корректный результат SQL-запроса select означает, что база данных работает нормально.
- **Monitor Interval (Интервал мониторинга).** Определяет интервал (в секундах) периодического выполнения метода монитора.
- **Hung Monitor Signal (Сигнал при зависании монитора).** Определяет сигнал, отправляемый для остановки метода монитора, если ответ от него не был получен за время, заданное параметром **Monitor Interval (Интервал мониторинга)**. По умолчанию отправляется сигнал SIGKILL(9).

Приостановка/возобновление мониторинга приложения

После конфигурирования монитора приложения необходимо его активизировать. Это можно сделать через меню **Resume Application Monitoring (Возобновление мониторинга приложения)** путем запуска **smit cl_admin > HACMP Resource Group and Application Management (Управление группами ресурсов и приложениями HACMP) > Suspend/Resume Application Monitoring (Приостановка/возобновление мониторинга приложения) > Resume Application Monitoring (Возобновить мониторинг приложения)**, после чего следует выбрать сервер приложения, связанный с монитором, который требуется активизировать. Пример 8.19 содержит выходные данные после успешного возобновления монитора приложения на сервере приложения APP1 в нашем тестовом кластере.

Пример 8.19

```
Jul  6 2005 18:00:17 cl_RMupdate: Completed request to resume monitor(s) for
applic ation APP1.
Jul  6 2005 18:00:17 cl_RMupdate: The following monitor(s) are in use for
applicati on APP1:
test
```

Можно приостановить либо возобновить мониторинг приложения в любое время. Это действие не влияет на доступность сервера приложения, однако влияет на статистические результаты, выводимые инструментом анализа доступности приложения (application availability analysis tool). Дополнительные сведения об этом инструменте см. в разделе 8.7.8, «Измерение доступности приложения».

8.7.8 Измерение доступности приложения

С помощью инструмента анализа доступности приложения (application availability analysis tool) можно выполнять измерение количества времени общей доступности для приложений высокой доступности. Программное обеспечение HACMP осуществляет сбор и регистрацию следующей информации в формате с отметками времени:

- запуски, остановки и отказы приложения;
- отказы узлов, завершения работы и возобновления работы на узлах, а также запуск и завершение работы служб кластера;
- отключение или перемещение группы ресурсов;
- приостановка или возобновление мониторинга приложения.

На основе информации, собранной инструментом анализа доступности приложения (application availability analyze tool), можно выбрать время для периода измерения, после чего инструмент выводит статистику времени работы и времени простоя для определенного приложения за этот период. Используя SMIT, можно вывести следующие показатели:

- процент времени работы;
- количество времени работы;
- самый долгий период времени работы;
- процент времени простоя;
- количество времени простоя;
- самый долгий период времени простоя;
- процент времени приостановки мониторинга приложения.

Инструмент анализа доступности приложения (application availability analyze tool) выдает информацию о доступности приложения с точки зрения кластера HACMP. Он может анализировать только те приложения, которые были правильно настроены в конфигурации кластера.

Этот инструмент выводит только статистику, отражающую доступность сервера приложения HACMP, группы ресурсов и монитора приложения (если сконфигурировано). Он не может определить внутренний отказ приложения, обнаруживаемый конечным пользователем, если этот отказ не определяется монитором приложения.

Использование инструмента анализа доступности приложения

Инструмент анализа доступности приложения можно использовать сразу же после определения серверов приложения, так как инструмент не требует какой-либо дополнительной настройки и автоматически собирает статистику по всем серверам приложений.

Можно вывести определенные статистические показатели приложения, генерируемые инструментом анализа доступности приложения (Application Availability Analysis tool) из меню SMIT, выполнив **smit hacmp > System Management (C-SPOC) > Resource Group and Application Management (Управление группами ресурсов и приложениями HACMP) > Application Availability Analysis (Анализ доступности приложения)**. На рис. 8.33 показан экран SMIT инструмента анализа доступности приложения в среде нашего тестового кластера.

smit cl_app_AAA.dialog

В меню SMIT инструмента анализа доступности приложения нужно только указать требуемый сервер приложения, ввести начальное и конечное время показа статистических показателей и запустить инструмент. Пример 8.20 содержит выходные данные инструмента анализа доступности приложения для нашего тестового кластера.

```
                        Application Availability Analysis

  Type or select values in entry fields.
  Press Enter AFTER making all desired changes.
                                                        [Entry Fields]
  * Select an Application                             [APP1]              +
  * Begin analysis on YEAR (1970-2038)                [2005]               #
  * MONTH (01-12)                                     [07]                 #
  * DAY (1-31)                                        [06]                 #
  * Begin analysis at HOUR (00-23)                    [16]                 #
  * MINUTES (00-59)                                   [22]                 #
  * SECONDS (00-59)                                   [00]                 #
  * End analysis on YEAR (1970-2038)                  [2005]               #
  * MONTH (01-12)                                     [07]                 #
  * DAY (1-31)                                        [06]                 #
  * End analysis at HOUR (00-23)                      [17]                 #
  * MINUTES (00-59)                                   [42]                 #
  * SECONDS (00-59)                                   [00]                 #

  F1=Help             F2=Refresh          F3=Cancel           F4=List
  F5=Reset            F6=Command          F7=Edit             F8=Image
  F9=Shell            F10=Exit            Enter=Do
```

Рис. 8.33. Экран анализа доступности приложения

Пример 8.20. Выходные данные инструмента анализа доступности приложения

```
Analysis begins:                Wednesday, 06-July-2005, 16:20
Analysis ends:                  Wednesday, 06-July-2005, 17:42
Application analyzed:           APP1

Total time:                     0 days, 1 hours, 22 minutes, 0 seconds

Uptime:
        Amount:                 0 days, 1 hours, 16 minutes, 51 seconds
        Percentage:             93.72%
        Longest period:         0 days, 1 hours, 10 minutes, 35 seconds

Downtime:
        Amount:                 0 days, 0 hours, 5 minutes, 9 seconds
        Percentage:             6.28%
        Longest period:         0 days, 0 hours, 5 minutes, 9 seconds

Log records terminated before the specified ending time was reached.

Application monitoring was suspended for 75.87% of the time period analyzed.

Application monitoring state was manually changed during the
time period analyzed.

Cluster services were manually restarted during the time period analyzed.
```

9

Безопасность кластера

Эта глава описывает средства безопасности HACMP и способы настройки безопасности кластера. В этой главе мы обсудим следующие темы:

- безопасность кластера и демона clcomd;
- управление пользователями и паролями;
- зашифрованный обмен сообщениями в кластере (Encrypted cluster messaging);
- защищенное удаленное выполнение команд;
- безопасность WebSMIT;
- HACMP и брандмауэры;
- безопасность RSCT.

9.1 Безопасность кластера и демона clcomd

Демон коммуникаций кластера (Cluster Communication Daemon) представляет реализацию надежного, защищенного транспортного уровня для HACMP. Clcomd осуществляет управление почти всеми межкластерными коммуникациями, в частности C-SPOC, верификацией и синхронизацией кластера, наборами файлов. Clcomd также улучшает производительность кластера: он имеет кеш для файлов HACMP ODM, что позволяет ускорить доставку, и может повторно использовать существующие подключения сокетов.

Демон диспетчера кластера и мониторинг пульса основаны на RSCT (см. раздел 9.6, «Безопасность RSCT»), тогда как clinfo использует протокол SNMP.

Существует два режима безопасности для аутентификации подключения в HACMP:

- **Стандартный (Standard).** Выполняется проверка входящего подключения со списком доступа, и разрешается только минимальный требуемый доступ. В этой главе рассматривается только стандартный режим безопасности кластера.
- **Kerberos:.** Все коммуникации в кластере защищаются с использованием инфраструктуры PSSP Kerberos. Этот режим безопасности работает только на системах SP с установленным программным обеспечением PSSP.

Начиная с HACMP 5.1 файл .rhosts не применяется. Безопасность HACMP основана на аутентификации подключений по принципу минимума привилегий. Демон коммуникаций кластера использует внутренний список доступа для авторизации других узлов для удаленного выполнения команд. При получении узлом запроса на удаленное выполнение Clcomd ищет входящий IP-адрес подключения среди IP-адресов в следующих расположениях:

1. ODM-класс HACMPnode.
2. ODM-класс HACMPadapter.
3. Файл /usr/es/sbin/cluster/etc/rhosts.

Демон коммуникаций кластера использует принцип минимума привилегий. При удаленном выполнении команды разрешается только минимальный объем требуемых привилегий. Это обеспечивает удаленное выполнение только для доверенных команд HACMP, причем, если возможно, команды выполняются под учетной записью **nobody**. Утилиты кластера разделены на две группы:

- доверенные команды, для которых допускается выполнение под учетной записью **root;**
- остальные команды, выполняющиеся под учетной записью **nobody**.

> **Ограничение.** Нельзя использовать аутентификацию на основе Clcomd для собственных скриптов (скриптов запуска и остановки приложений, настраиваемых скриптов обработки событий и т. д.); в этом случае следует применять rsh или SSH.

9.1.1 Файл /usr/es/sbin/cluster/etc/rhosts

Файл /usr/es/sbin/cluster/etc/rhosts должен содержать список всех IP-адресов всех узлов кластера для повышенной безопасности. Этот файл автоматически получает информацию о подключении узла (базовые адреса узлов) от HACMP в процессе обнаружения, а также верификации и синхронизации кластера. Кроме того, можно вручную отредактировать файл /usr/es/sbin/cluster/etc/rhosts, включив в него IP-адреса узлов.

> **Важно!** Файл /usr/es/sbin/cluster/etc/rhosts должен иметь следующие разрешения:
> - владелец: root;
> - группа: system;
> - разрешения: 0600.

Изначальная настройка кластера

При начальной настройке кластера файл /usr/es/sbin/cluster/etc/rhosts пуст. Обычно HACMP записывает в него базовые адреса одноранговых узлов при первом запуске процессов обнаружения и синхронизации кластера. При первоначальном конфигурировании безопасности кластера рекомендуем добавить IP-адреса кластера на всех узлах в файл /usr/es/sbin/cluster/etc/rhosts, прежде чем начать конфигурирование HACMP.

> **Внимание!** При изначальной настройке, пока файл /usr/es/sbin/cluster/etc/rhosts на узле HACMP пуст, первым может произойти подключение с нежелательного хоста, вследствие чего в файл rhosts будет записан IP-адрес этого хоста. В этом случае другие одноранговые узлы не смогут подключиться до тех пор, пока файл /usr/es/sbin/cluster/etc/rhosts не будет исправлен вручную.

9.1.2 Отключение демона коммуникаций кластера

Для повышения безопасности можно отключить clcomd. При этом следует учитывать, что без clcomd не смогут работать следующие функции:

- верификация и синхронизация HACMP;
- C-SPOC, включая LVM, управление пользователями и паролями;
- наборы файлов;
- аутентификация и шифрование сообщений.

Отключение демона коммуникаций кластера выполняется командой **stopsrc -s clcomdES.** Перезапуск clcomd выполняется командой **startsrc -s clcomdES.**

9.1.3 Дополнительные средства безопасности кластера

Файлы HACMP ODM хранятся в каталоге /etc/es/objrepos. В целях повышения безопасности для них задается владелец **root** и группа **hacmp**. Кроме того, для них устанавливается разрешение 0640, кроме файла HACMPdisksubsystem, для которого устанавливается разрешение 0600. Для всех утилит кластера, предназначенных для открытого использования, включен hacmp setgid, что позволяет им читать файлы HACMP ODM. Если группа hacmp не существует, она создается во время установки HACMP.

9.2 Использование зашифрованных межузловых коммуникаций

HACMP поддерживает зашифрованный обмен сообщениями (encrypted messaging) между узлами кластера. Шифрование может осуществляться для трафика диспетчера кластера (Cluster Manager), демона коммуникаций кластера (Cluster Communication Daemon) и RSCT.

Шифрование основано на схеме симметричного ключа, обеспечиваемой службами безопасности кластера (Cluster Security Services, CtSec). CtSec является частью RSCT. См. также раздел 9.6, «Безопасность RSCT».

Можно использовать следующие дополнительные модули шифрования RSCT:

- Data Encryption Standard (md5_des);
- Triple DES (md5_3des);
- Advanced Encryption Standard (md5_aes).

Шифрование вызывает дополнительную нагрузку на процессоры, однако на продолжительность перемещения при сбое оно не влияет.

9.2.1 Управление ключами шифрования

Модули шифрования RSCT используют симметричные ключи. Это означает, что все узлы кластера используют один ключ для шифрования и дешифрования. На каждом узле ключи безопасности расположены в каталоге /usr/es/sbin/cluster/es. Имя файла соответствует выбранному метод шифрования:

- key_md5_des;
- key_md5_3des;
- key_md5_aes.

Генерировать ключ следует при первой настройке шифрования сообщений, при изменении конфигурации безопасности кластера или в соответствии с правилами безопасности вашей компании. Один ключ можно использовать столько, сколько потребуется.

HACMP содержит механизм распространения ключей по узлам:

- Автоматическое распространение ключей: новый ключ распространяется по узлам кластера автоматически после того, как он был сгенерирован. Этот метод удобен, однако менее безопасен, так как копирование ключей осуществляется через сеть.
- Распространение ключей вручную: необходимо вручную копировать файл ключей на каждый узел. Можно использовать SFTP или SCP, однако безопаснее всего применять дискету для распространения ключей. Мы рекомендуем использовать этот метод.

Внимание! Убедитесь, что владельцем файла ключей является пользователь root, группа system и установлены права доступа 0400. Если кто-либо сможет скопировать ваш ключ или перехватить его по сети, безопасность кластера будет в серьезной опасности.

9.2.2 Настройка шифрования сообщений

Основные этапы конфигурирования:

1. Установка требуемого модуля шифрования RSCT.

2. Включение автоматического распространения ключей (только при автоматическом распространении ключей).

3. Включение аутентификации и шифрования сообщений.

4. Генерирование и распространение ключей.

5. Активизация ключей.

6. Синхронизация кластера.

7. Отключение автоматического распространения ключей (только при автоматическом распространении ключей).

В целях безопасности рекомендуется использовать распространение ключей вручную.

Выполнение команд всегда следует осуществлять с одного и того же узла, за исключением этапов 2 и 7, требующих выполнения на всех узлах.

Важно!

- Убедитесь, что параметры аутентификации и шифрования согласованы по всему кластеру. Все узлы должны содержать одинаковые файлы ключей. В противном случае HACMP не сможет осуществлять обмен данными с другими узлами.
- Во время настройки шифрования сообщений в кластере не выполняйте какие-либо другие задачи конфигурирования HACMP.
- Убедитесь, что ваш кластер находится в рабочем состоянии, что в нем недавно выполнялись процедуры синхронизации и верификации и что узлы могут осуществлять связь друг с другом.

- 1. **Обязательные условия.** Следует установить требуемую библиотеку шифрования с диска AIX 5L Expansion Pack CD. Используя **smit install_latest**, установите следующие наборы файлов:
- **rsct.crypt.des** – для использования md5_des;
- **rsct.crypt.3des** – для применения md5_3des;
- **rsct.crypt.aes256** – для использования md5_aes.

Если кластер уже работает, необходимо перезапустить clcomd:

stopsrc -s clcomdes

startsrc -s clcomdes

2. **Включение автоматического распространения ключей на всех узлах. (**Этот этап следует выполнять, только если требуется использовать автоматическое распространение ключей.):

а) перейдите в SMIT **HACMP Cluster Security (Безопасность кластера HACMP)**:

запустите **smit**;

- выберите **Communications Applications and Services (Коммуникационные приложения и службы);**
- выберите **HACMP for AIX (HACMP для AIX);**
- выберите **System Management (C-SPOC);**
- выберите **HACMP Security and Users Management (Безопасность и управление пользователями HACMP);**
- выберите **HACMP Cluster Security (Безопасность кластера HACMP)** либо используйте быстрый путь **smit cm_config_security;**

б) выберите **Configure Message Authentication Mode and Key Management (Конфигурирование режима аутентификации сообщений и управления ключами)**;

в) выберите **Enable/Disable Automatic Key Distribution (Включение-отключение автоматического распространения ключей)**;

г) установите для параметра **Enable/Disable Automatic Key Distribution (Включение-отключение автоматического распространения ключей)** значение **Enabled** (рис. 9.1);

д) нажмите Enter для подтверждения.

Повторите эти действия на всех узлах.

3. **Включение или изменение метода аутентификации и шифрования сообщений:**

- а) перейдите в SMIT **HACMP Cluster Security (Безопасность кластера HACMP)**; выполните **smit cm_config_security**;

```
                    Enable/Disable Automatic Key Distribution

  Type or select values in entry fields.
  Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
  * Enable/Disable Key Distribution                    Enabled            +


    +-----------------------------------------------------------------------+
    ¦                     Enable/Disable Key Distribution                   ¦
    ¦                                                                       ¦
    ¦ Move cursor to desired item and press Enter.                          ¦
    ¦                                                                       ¦
    ¦   Enabled                                                             ¦
    ¦   Disabled                                                            ¦
    ¦                                                                       ¦
    ¦ F1=Help                 F2=Refresh              F3=Cancel             ¦
  F1¦ F8=Image                F10=Exit                Enter=Do              ¦
  F5¦ /=Find                  n=Find Next                                   ¦
  F9+-----------------------------------------------------------------------+
```

Рис. 9.1. Включение автоматического распространения ключей

б) выберите **Configure Message Authentication Mode and Key Management (Конфигурирование режима аутентификации сообщений и управления ключами)**;

в) выберите **Configure Message Authentication Mode (Конфигурирование режима аутентификации сообщений)**;

г) нажмите F4 и в разделе **Message Authentication Mode (Режим аутентификации сообщений)** выберите требуемый режим аутентификации (рис. 9.2):

- **md5_des,**
- **md5_3des,**
- **md5_aes,**
- **None**: не использовать ни аутентификацию, ни шифрование сообщений;

д) установите в поле **Enable Encryption (Включить шифрование)** значение **Yes (Да)**;

е) нажмите Enter.

4. **Генерирование и распространение ключей.**

а) перейдите в SMIT **HACMP Cluster Security (Безопасность кластера HACMP)**: выполните **smit cm_config_security.**

```
                    Configure Message Authentication Mode

 Type or select values in entry fields.
 Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
 * Message Authentication Mode                         md5_aes             +
 * Enable Encryption                                   Yes                 +


   +-------------------------------------------------------------------+
   ¦                       Message Authentication Mode                 ¦
   ¦                                                                   ¦
   ¦ Move cursor to desired item and press Enter.                      ¦
   ¦                                                                   ¦
   ¦   md5_des                                                         ¦
   ¦   md5_3des                                                        ¦
   ¦   md5_aes                                                         ¦
   ¦   none                                                            ¦
   ¦                                                                   ¦
   ¦ F1=Help                 F2=Refresh              F3=Cancel         ¦
 F1¦ F8=Image                F10=Exit                Enter=Do          ¦
 F5¦ /=Find                  n=Find Next                               ¦
 F9+-------------------------------------------------------------------+
```

Рис. 9.2. Конфигурирование режима аутентификации сообщений

б) выберите **Configure Message Authentication Mode and Key Management (Конфигурирование режима аутентификации сообщений и управления ключами)**.

в) выберите **Generate/Distribute a Key (Сгенерировать/распространить ключ)**.

г) нажмите F4 и выберите **Type of Key to Generate (Тип генерируемого ключа)**. Здесь следует выбрать такой же ключ, как и на этапе 3, «Включение или изменение метода аутентификации и шифрования сообщений» (рис. 9.3).

д) выберите **Distribute a Key (Распространить ключ)**:

- **Yes (Да)**: при использовании автоматического распространения ключей;
- **No (Нет)**: при использовании распространения ключей вручную.

е) Нажмите Enter.

```
                          Generate/Distribute a Key

  Type or select values in entry fields.
  Press Enter AFTER making all desired changes.

                                                      [Entry Fields]
  * Type of Key to Generate                          md5_aes +
  * Distribute a Key                                 No +




  F1=Help             F2=Refresh          F3=Cancel           F4=List
  F5=Reset            F6=Command          F7=Edit             F8=Image
  F9=Shell            F10=Exit            Enter=Do
```

Рис. 9.3. Генерирование ключа

Если выбрано автоматическое распространение ключей, то следует перейти к шагу 5, «Активизация ключа».

Теперь скопируйте файл ключа с этого узла на другие узлы. Можно использовать любой метод передачи, но мы рекомендуем применять один из следующих способов:

- SCP. Утилита защищенного зашифрованного удаленного копирования. Пример:
  scp /usr/es/sbin/cluster/etc/key_md5_aes \
  node2:/usr/es/sbin/cluster/etc/key_md5_aes
- Дискета. Следует скопировать соответствующий файл ключа из каталога /usr/es/sbin/cluster/etc на дискету. Затем следует скопировать файл с дискеты на все узлы.

Если у вас уже есть файл ключа, просто замените его новым файлом ключа.

Будьте осторожным с файлом ключа: не пересылайте его через сеть в незашифрованном виде (например, не следует использовать ftp или rcp) и обязательно установите для него разрешение 0400.

5. **Активизация ключа:**

а) перейдите в SMIT **HACMP Cluster Security (Безопасность кластера HACMP)**: выполните **smit cm_config_security**;

б) выберите **Configure Message Authentication Mode and Key Management (Конфигурирование режима аутентификации сообщений и управления ключами)**;

в) выберите **Activate the new key on all HACMP cluster nodes (Активизировать новый ключ на всех узлах кластера HACMP)**;

г) еще раз нажмите Enter для подтверждения.

> **Важно!** Не активизируйте новый ключ до тех пор, пока на всех узлах кластера не будет установлен одинаковый файл ключа.

6. **Синхронизация кластера. При возникновении ошибки, связанной с коммуникациями кластера, следует отключить аутентификацию и шифрование сообщений и начать эту процедуру сначала.**

7. **Отключение автоматического распространения ключей на всех узлах. (**Этот этап следует выполнять, только если вы используете автоматическое распространение ключей):

а) перейдите в SMIT **HACMP Cluster Security (Безопасность кластера HACMP)**: выполните **smit cm_config_security**;

б) выберите **Configure Message Authentication Mode and Key Management (Конфигурирование режима аутентификации сообщений и управления ключами)**;

в) выберите **Enable/Disable Automatic Key Distribution (Включение-отключение автоматического распространения ключей)**;

г) установите для параметра **Enable/Disable Automatic Key Distribution (Включить-отключение автоматического распространения ключей)** значение **Disabled**;

д) нажмите Enter для подтверждения.

Повторите эти действия на всех узлах.

> **Важно!** Всегда отключайте автоматическое распространение ключей после того, как были успешно сгенерированы и распространены новые ключи. В противном случае безопасность кластера нарушается.

Изменение ключа аутентификации

Ключ можно применять столько, сколько требуется, однако чем дольше вы используете один ключ, тем больше вероятность его перехвата и несанкционированного использования. Для обеспечения максимальной безопасности рекомендуется регулярно менять ключ, например ежемесячно.

1. **Включение автоматического распространения ключей на всех узлах. Этот этап следует выполнять, только если требуется использовать автоматическое распространение ключей:**

а) перейдите в SMIT **HACMP Cluster Security (Безопасность кластера HACMP)**: выполните **smit cm_config_security**;

б) выберите **Configure Message Authentication Mode and Key Management (Конфигурирование режима аутентификации сообщений и управления ключами)**;

в) выберите **Enable/Disable Automatic Key Distribution (Включение-отключение автоматического распространения ключей)**;

г) установите для параметра **Enable/Disable Automatic Key Distribution (Включение-отключение автоматического распространения ключей)** значение **Enabled**;

д) нажмите Enter для подтверждения.

Повторите эти действия на всех узлах.

2. **Генерирование и распространение ключей**.

а) Перейдите в SMIT **HACMP Cluster Security (Безопасность кластера HACMP)**: выполните **smit cm_config_security**.

б) Выберите **Configure Message Authentication Mode and Key Management (Конфигурирование режима аутентификации сообщений и управления ключами)**.

в) Выберите **Generate/Distribute a Key (Сгенерировать/Распространить ключ)**.

г) Нажмите F4 и выберите **Type of Key to Generate (Тип генерируемого ключа)**. Здесь следует выбрать такой же ключ, как и на этапе 3, «Включение или изменение метода аутентификации и шифрования сообщений» (рис. 9.3).

д) Выберите **Distribute a Key (Распространить ключ)**:

- **Yes (Да)**: при использовании автоматического распространения ключей;
- **No (Нет)**: при использовании распространения ключей вручную.

е) Нажмите Enter.

ж) Если вы используете распространение ключей вручную, тогда следует скопировать файл ключей с этого узла на другие узлы, заменив старые файлы.

3. **Активизация ключа:**

а) перейдите в SMIT **HACMP Cluster Security (Безопасность кластера HACMP)**: выполните **smit cm_config_security**;

б) выберите **Configure Message Authentication Mode and Key Management (Конфигурирование режима аутентификации сообщений и управления ключами)**;

в) выберите **Activate the new key on all HACMP cluster nodes (Активизировать новый ключ на всех узлах кластера HACMP)**;

г) еще раз нажмите Enter для подтверждения.

4. **Отключение автоматического распространения ключей на всех узлах. Этот этап следует выполнять, только если вы используете автоматическое распространение ключей:**

а) перейдите в SMIT **HACMP Cluster Security (Безопасность кластера HACMP)**: выполните **smit cm_config_security**;

б) выберите **Configure Message Authentication Mode and Key Management (Конфигурирование режима аутентификации сообщений и управления ключами)**;

в) выберите **Enable/Disable Automatic Key Distribution (Включение-отключение автоматического распространения ключей)**;

г) установите для параметра **Enable/Disable Automatic Key Distribution (Включение-отключение автоматического распространения ключей)** значение **Disabled**;

д) нажмите Enter для подтверждения.

Повторите эти действия на всех узлах.

9.2.3 Устранение неполадок аутентификации и шифрования сообщений

При возникновении ошибок коммуникации кластера (например, при отказе верификации кластера или если C-SPOC не может связаться с другими узлами) следует отключить и аутентификацию сообщений и шифрование сообщений:

1. Перейдите в SMIT **HACMP Cluster Security (Безопасность кластера HACMP)**: выполните **smit cm_config_security.**

2. Выберите **Configure Message Authentication Mode and Key Management (Конфигурирование режима аутентификации сообщений и управления ключами)**.

3. Выберите **Configure Message Authentication Mode (Конфигурирование режима аутентификации сообщений)**.

4. В разделе **Message Authentication Mode (Режим аутентификации сообщений)** выберите **None**.

5. Установите в поле **Enable Encryption (Включить шифрование)** значение **No (Нет)**.

6. Нажмите Enter.

7. Выполните синхронизацию кластера.

9.2.4 Проверка текущих параметров аутентификации сообщений

Текущие параметры аутентификации и шифрования сообщений можно проверить следующим образом:

1. Запустите **smit hacmp.**

2. Выберите **Extended Configuration (Расширенное конфигурирование)**.

3. Выберите **Extended Topology Configuration (Расширенное конфигурирование топологии)**.

```
                                    COMMAND STATUS

  Command: OK            stdout: yes           stderr: no

  Before command completion, additional instructions may appear below.

  Cluster Name: two2four
  Cluster Connection Authentication Mode: Standard
  Cluster Message Authentication Mode: md5_aes
  Cluster Message Encryption: Enabled
  Use Persistent Labels for Communication: No



  F1=Help             F2=Refresh          F3=Cancel           F6=Command
  F8=Image            F9=Shell            F10=Exit            /=Find
  n=Find Next
```

Рис. 9.4. Проверка параметров аутентификации сообщений

4. Выберите **Show HACMP Topology (Отображение топологии HACMP)**.

5. Выберите **Show Cluster Definition (Отображение определения кластера)**. Нажмите Enter для вывода параметров определения кластера (рис. 9.4).

9.3 Защищенное удаленное выполнение команд в HACMP

> **Примечание.** Хотя использование Secure Shell (SSH) в кластере HACMP не является обязательным, этот метод защиты удаленного выполнения команд является очень популярным в современных сетевых средах.

Скрипты запуска и остановки приложений, настраиваемые скрипты обработки событий и прочие скрипты могут требовать выполнения команд на удаленных узлах. Для этой цели все еще можно использовать rsh, однако это небезопасно, так как при этом аутентификация основана на файле .rhosts. Мы настоятельно рекомендуем использовать Secure Shell (SSH). Кроме того, функционирование DLPAR также требует применения SSH.

SSH и Secure Socket Layer (SSL) совместно обеспечивают аутентификацию, конфиденциальность и целостность данных. Аутентификация SSH основана на инфраструктуре открытых-закрытых ключей[1], тогда как SSL осуществляет шифрование сетевого трафика.

[1] Более распространен термин «Public Key Infrastructure, PKI» (инфраструктура открытых ключей), наличие закрытого (private) ключа подразумевается. *Примеч. науч. ред.*

SSH содержит следующие утилиты, заменяющие классические r-команды:

ssh	защищенная удаленная оболочка, подобная rsh или rlogin;
scp	защищенное удаленное копирование, подобное rcp;
sftp	утилита зашифрованной передачи файлов, подобная ftp.

9.3.1 Установка SSH

Весьма удобно, что IBM поставляет SSH и все его компоненты вместе с AIX. Ниже приведен список требуемых компонентов.

rpm.rte	поддержка rpm-пакетов; автоматически устанавливается вместе с базовым программным обеспечением AIX.
perl	автоматически устанавливается вместе с базовым программным обеспечением AIX.
openssl.rpm	поддержка Open Source Secure Socket Layer (AIX Toolbox for Linux Applications CD).
openssh.base.client	команды клиента Open Source Secure Shell (AIX Expansion Pack).
openssh.base.server	сервер Open Source Secure Shell (AIX Expansion Pack).
openssh.msg.*	каталог сообщений для SSH (AIX Expansion Pack).
openssh.man.*	документация (man pages) для Secure Shell (AIX Expansion Pack).
openssh.license	open Source License для SSH (AIX Expansion Pack).
prngd.rpm	генератор случайных чисел, необходимый только в AIX 5.1 и 5.2. (AIX Toolbox for Linux Applications CD).

Другой вариант состоит в том, чтобы скопировать наиболее актуальные пакеты из Интернета:

- OpenSSL с сайта IBM AIX Toolbox for Linux Applications (требует процедуры регистрации продолжительностью в несколько минут):

  http://www6.software.ibm.com/dl/aixtbx/aixtbx-p

- OpenSSH для AIX:

  http://www.sourceforge.net/projects/openssh-aix

  В нашем тесте мы использовали openssl-0.9.6m-1 и openssh.3.8.0.53.

  Установка состоит из следующих этапов:

  1. Убедитесь в том, что установлен rpm.rte: **lslpp -l rpm.rte**.

  2. Убедитесь в том, что установлен Perl: **lslpp -l perl.rte**.

  3. При использовании AIX 5.1 или 5.2 следует установить Prngd с диска AIX Toolbox for Linux Applications CD через SMIT.

  4. Установите OpenSSL. Можно применять **smit install_latest** для его установки с диска AIX Toolbox for Linux Applications CD. В противном случае следует использовать команду rpm:

  rpm -i openssl-0.9.6m-1.aix5.1.ppc.rpm

  5. Используйте **smit install_latest** для установки openssh.base, файла лицензии и соответствующих языковых наборов файлов.

  6. Запустите демон SSH: **startsrc -s sshd**.

9.3.2 Настройка SSH для удаленного выполнения команд без пароля

Некоторые скрипты могут требовать использования SSH и SCP в конфигурации без применения паролей. В этом случае SSH использует пары открытых и закрытых ключей для аутентификации. Закрытый ключ хранится на локальном хосте, тогда как открытый ключ пересылается на удаленные узлы и добавляются в их базу данных авторизованных ключей.

При выполнении команды SSH (например, **ssh node2 date**) коммуникация подписывается с использованием закрытого или секретного ключа узла-источника. Эту подпись можно дешифровать только с применением открытого ключа узла-источника. Если узел-получатель может успешно дешифровать подпись, это подтверждает подлинность узла-отправителя.

Для настройки удаленного выполнения команд через SSH без паролей следует выполнить следующие действия на всех узлах кластера:

1. Выполнить вход в систему под требуемой учетной записью.

2. Сгенерировать пару ключей аутентификации:

```
ssh-keygen -t rsa -f ~/.ssh/node1
```

Нажмите Enter для ввода пароля-фразы из нескольких слов. Эта команда генерирует два файла:

- ~/.ssh/node1: файл секретного ключа;
- ~/.ssh/node1.pub: файл открытого ключа.

3. Переименуйте файл секретного ключа, назначив ему имя **identity**:

```
mv ~/.ssh/node1 ~/.ssh/identity
```

4. Добавьте открытый ключ в файл **authorized_keys** на локальном узле, чтобы SSH мог работать на локальном хосте (localhost):

```
cat ~/.ssh/node1.pub >> ~/.ssh/authorized_keys
```

5. Скопируйте свой открытый ключ на все остальные узлы:

```
scp ~/.ssh/node1.pub nodeX:~/.ssh/node1.pub
```

Повторите эту команду для каждого узла в кластере.

6. Добавьте открытый ключ узла node1 в файл **authorized_keys** на удаленных узлах:

```
ssh nodeX "cat ~/.ssh/node1.pub >> ~/.ssh/authorized_keys"
```

Повторите эту команду для всех узлов в кластере.

7. Повторите действия пп. 1–6 на всех узлах.

Теперь SSH должен работать на всех узлах не запрашивая пароль.

> **Важно!** Будьте осторожны с файлами закрытого и открытого ключа: это ключи к вашей системе. Если какой-либо неавторизованный пользователь получит файлы, он сможет подключиться к вашей системе.

Дополнительные сведения об OpenSSL и OpenSSH можно найти в Интернете:

- веб-сайт проекта OpenSSL:
  http://www.openssl.org
- страница проекта OpenSSH:
  http://www.openssh.org
- OpenSSH на AIX:
  http://www.sourceforge.net/projects/openssh-aix

9.4 Безопасность WebSmit

WebSMIT является очень хорошим инструментом конфигурирования и мониторинга кластера. Однако WebSMIT осуществляет доступ к файлам HACMP и выполняет команды кластера под учетной записью пользователя root. Это является потенциально уязвимым местом в безопасности, однако при тщательном планировании и реализации безопасности риск можно полностью устранить. Этот раздел описывает требования безопасности и механизмы, используемые в средах WebSMIT.

Для обеспечения максимальной безопасности WebSMIT можно выполнить конфигурирование следующих элементов:

- шифрование коммуникаций между клиентом и сервером с использованием SSL;
- SSL-сертификат для аутентификации;
- аутентификация веб-сервера и AIX;
- обеспечение доступа к WebSMIT только для определенных пользователей AIX;
- обеспечение доступа через WebSMIT только к определенным панелям SMIT.

Для аутентификации WebSMIT требует SSL и SSL-сертификат. При установке Apache генерируется самоподписанный (self-signed) сертификат, который можно использовать в WebSMIT.

9.4.1 Параметры безопасности WebSMIT

Параметры безопасности WebSMIT хранятся в файле **/usr/es/sbin/cluster/wsm/wsm_smit.conf**. По умолчанию файл содержит максимальные установки параметров безопасности. В этом разделе описываются опции безопасности.

> **Внимание!** При изменении любого из параметров WebSMIT необходимо перезапустить HTTP-сервер.

Авторизованный порт

Параметр AUTHORIZED_PORT представляет порт TCP, используемый для коммуникаций WebSMIT. По умолчанию применяется порт 42267. Рекомендуем изменить значение этого параметра на другой номер, чтобы не допустить доступа неавторизованного пользователя к WebSMIT через стандартный порт. Если определить номер порта здесь, то WebSMIT будет использовать тот протокол (http или https), который связан с

этим портом в файле конфигурации HTTP-сервера. Например, если определить порт 42267 в качестве порта SSL в файле httpd.conf, то браузер будет использовать шифрование и в URL нужно будет применять «https://».

Принудительное использование защищенного http

Параметр REDIRECT_TO_HTTPS=1 позволяет использовать только защищенные https-коммуникации, чтобы все данные передавались по сети в зашифрованном виде. Если был определен параметр AUTHORIZED_PORT, то эта переменная не применяется. В противном случае будет выполнено перенаправление браузера на порт SSL вне зависимости от введенного URL. Отношение между параметрами AUTHORIZED_PORT и REDIRECT_TO_HTTPS см. в табл. 9.1.

Таблица 9.1. Отношение между параметрами AUTHORIZED_PORT и REDIRECT_HTTPS

	AUTHORIZED_PORT установлен	**AUTHORIZED_PORT не установлен**
REDIRECT_TO_HTTPS=0	Используются параметры безопасности из файла httpd.conf	Используется незащищенный протокол HTTP, без шифрования, без SSL
REDIRECT_TO_HTTPS=1	Используются параметры безопасности порта из файла httpd.conf	Используется протокол HTTPS, с SSL и с шифрованием

Замечание. Мы рекомендуем задать порт SSL TCP в файле httpd.conf и использовать этот порт в качестве авторизованного порта для WebSMIT.

Аутентификация AIX

WebSMIT может применять аутентификацию AIX вдобавок к аутентификации веб-сервера путем установки параметра REQUIRE_AUTHENTICATION=1. В этом случае WebSMIT запрашивает имя и пароль пользователя AIX. Так как при этом используются механизмы аутентификации AIX, отказы при входе в систему могут вызвать блокирование учетной записи.

Замечание. Для повышения уровня безопасности WebSMIT следует установить и аутентификацию веб-сервера и аутентификацию AIX.

Допустимые пользователи

Переменная ACCEPTED_USERS содержит список идентификаторов пользователей AIX, которым разрешен доступ к WebSMIT. Все пользователи, указанные в этой переменной, будут иметь в WebSMIT уровень привилегий пользователя root. По умолчанию задано значение ACCEPTED_USERS=”root”.

Замечание. Настоятельно рекомендуем создать отдельного пользователя AIX для доступа к WebSMIT.

Тайм-аут сеанса

Переменная SESSION_TIMEOUT определяет значение тайм-аута сеанса, по истечении которого пользователь WebSMIT должен выполнить повторный вход в систему вне зависимости от того, активен он или нет. По умолчанию задан тайм-аут в 20 мин.

Идентификатор пользователя веб-сервера

Переменная REQUIRED_WEBSERVER_UID содержит идентификатор пользователя, используемый в скриптах cgi-bin в WebSMIT. Здесь должен указываться тот же идентификатор, который применяется и в веб-сервере. Кроме того, следует убедиться в том, что этот идентификатор пользователя не имеет привилегий входа в AIX, чтобы никто не мог переключиться в эту учетную запись. По умолчанию употребляется учетная запись «nobody».

9.4.2 Разрешение или запрет на доступ к определенным панелям SMIT в WebSMIT

Можно сконфигурировать, к каким панелям SMIT можно получить доступ через WebSMIT. В качестве имени панели SMIT применяется быстрый путь SMIT к заданной па-

```
                        Extended Topology Configuration

     Move cursor to desired item and press Enter.

       Configure an HACMP Cluster
       Configure HACMP Nodes
       Configure HACMP Sites
       Configure HACMP Networks
       Configure HACMP Communication Interfaces/Devices
       Configure HACMP Persistent Node IP Label/Addresses
       Configure HACMP Global Networks
       Configure HACMP Network Modules
       Configure +-------------------------------------------------------+
       Show HACMP¦                     PRINT SCREEN                      ¦
                 ¦                                                       ¦
                 ¦ Press Enter to save the screen image                  ¦
                 ¦     in the log file.                                  ¦
                 ¦ Press Cancel to return to the application.            ¦
                 ¦                                                       ¦
                 ¦   Current fast path:                                  ¦
                 ¦       "cm_extended_topology_config_menu_dmn"          ¦
                 ¦                                                       ¦
                 ¦ F1=Help          F2=Refresh       F3=Cancel           ¦
     F1=Help     ¦ F8=Image         F10=Exit         Enter=Do            ¦
     F9=Shell    +-------------------------------------------------------+
```

Рис. 9.5. Быстрый путь к панели SMIT Extended Topology Configuration (Расширенное конфигурирование топологии)

нели; вывод этого имени выполняется нажатием F8 в SMIT (рис. 9.5). Например, панель SMIT Extended Topology Configuration (Расширенное конфигурирование топологии) имеет имя «cm_extended_topology_config_menu_dmn».Файл **/usr/es/sbin/cluster/wsm/wsm_smit.allow** содержит список панелей SMIT, к которым разрешен доступ из WebSMIT. Ко всем остальным панелям доступ запрещен.

В файле **/usr/es/sbin/cluster/wsm/wsm_smit.deny** можно перечислить панели SMIT, недоступные из WebSMIT. Если панель задана и в .allow- и в .deny-файле, приоритет имеет указание в .deny-файле.

Если панель SMIT указана в файле **/usr/es/sbin/cluster/wsm/wsm_smit.redirect**, WebSMIT выполняет перенаправление страницы на заданный URL. Этот файл уже содержит некоторые панели, которые не следует изменять.

> **Замечание.** По окончании конфигурирования кластера следует запретить доступ ко всем панелям SMIT, кроме страницы состояния кластера HACMP. (пример 9.1).

Пример 9.1. Пример файла /usr/es/sbin/cluster/wsm/wsm_smit.allow

```
##################################################
# webSMIT .allow file                            #
##################################################
cm_hacmp_main_menu_dmn
```

9.4.3 Журналы WebSMIT

WebSMIT регистрирует все свои операции в журналах, подобно SMIT. Файл **wsm_smit.log** расположен в каталоге ../log относительно скриптов cgi-bin. Например, при использовании скриптов cgi-bin из заданного по умолчанию каталога /usr/es/sbin/cluster/wsm/cgi-bin файл журнала находится в каталоге /usr/es/sbin/cluster/wsm/log. Кроме того, в файл **wsm_smit.script** записывается последняя команда WebSMIT.

HACMP не осуществляет управление этими файлами журналов, они могут неограниченно расти, подобно файлу smit.log. Команда **snap -e** может осуществлять сбор файлов журналов WebSMIT, только если они находятся каталоге по умолчанию /usr/es/sbin/cluster/wsm/log.

9.5 HACMP и брандмауэры

При размещении кластера HACMP за брандмауэром следует учитывать следующие аспекты:

- HACMP не требует наличия какого-либо открытого порта на брандмауэре; отсутствует внешний трафик от Clcomd, RSCT или от диспетчера кластера (Cluster Manager). Открытые порты требуются только для управления приложением и системой (например, SSH).
- Необходимо убедиться в том, что все сервисные IP-адреса могут осуществлять обмен данными с внешней сетью вне зависимости от их привязки. Следует учиты-

вать, что при отказе сети сервисный интерфейс перемещается с одного базового адаптера на другой. В случае перемещения при сбое сервисный адрес отказавшего узла перемещается на другой узел.

- Не следует помещать брандмауэр между узлами. В кластере HACMP/XD узлы могут подключаться через открытую сеть. В этом случае следует использовать виртуальные частные сети (Virtual Private Networks) или другое решение, прозрачное для демона коммуникаций кластера (Cluster Communication Daemon).
- При использовании файла netmon.cf для улучшенного обнаружения отказов сети убедитесь в том, что IP-адреса, указанные в этом файле, доступны через брандмауэр (с помощью команды ping).
- Если имеются клиенты Clinfo, работающие через брандмауэр, следует открыть порт clinfo_client: 6174/tcp.
- Решение на основе брандмауэра должно быть реализовано по принципу избыточности; в противном случае брандмауэр становится единой точкой отказа.

9.6 Безопасность RSCT

В этой главе описывается терминология и понятия, используемые в безопасности RSCT. Также описывается архитектура безопасности RSCT и механизмы, используемые в HACMP. Предполагается, что у вас есть знание RSCT и его базовых компонентов (таких, как RMC, диспетчеры ресурсов и т. д.). Важно понимать, что основная часть компонентов и механизмов, описанных в этой главе, не нуждается в конфигурировании; уровень безопасности RSCT работает сразу же. Кроме того, конфигурирование некоторых функций и файлов выполняется только после настройки аутентификации и шифрования сообщений HACMP (см. раздел 9.2, «Использование зашифрованных межузловых коммуникаций»). Некоторые из описанных функций не применяются в HACMP, но являются неотъемлемой частью безопасности RSCT.

9.6.1 RSCT и HACMP

HACMP использует RSCT в качестве уровня базовой инфраструктуры. Узлы HACMP связываются с одноранговыми узлами через компоненты RSCT, проверяя их состояние или запрашивая доступ к их ресурсам. Службы топологии (Topology Services) и службы групп (Group Services) реализуют мониторинг пульса HACMP через одноранговый домен RSCT. Подсистема RMC применяется для следующего:

- для настраиваемых событий;
- мониторинга приложений;
- динамического приоритета узлов;
- экспорта состояния сети для использования в Oracle RAC.

На рис. 9.6 показана связь между компонентами HACMP и RSCT.

Каждый раз, когда компонент HACMP (например, диспетчер кластера) одного сервера отправляет функциональный запрос к ресурсу RMC (локальному или уда-

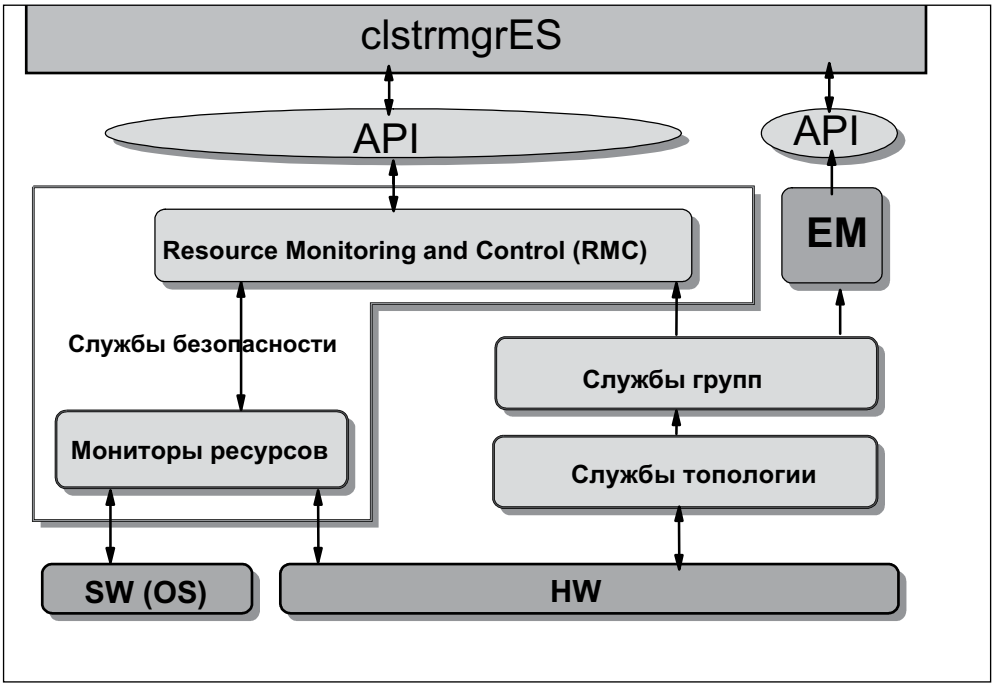

Рис. 9.6. RSCT и HACMP

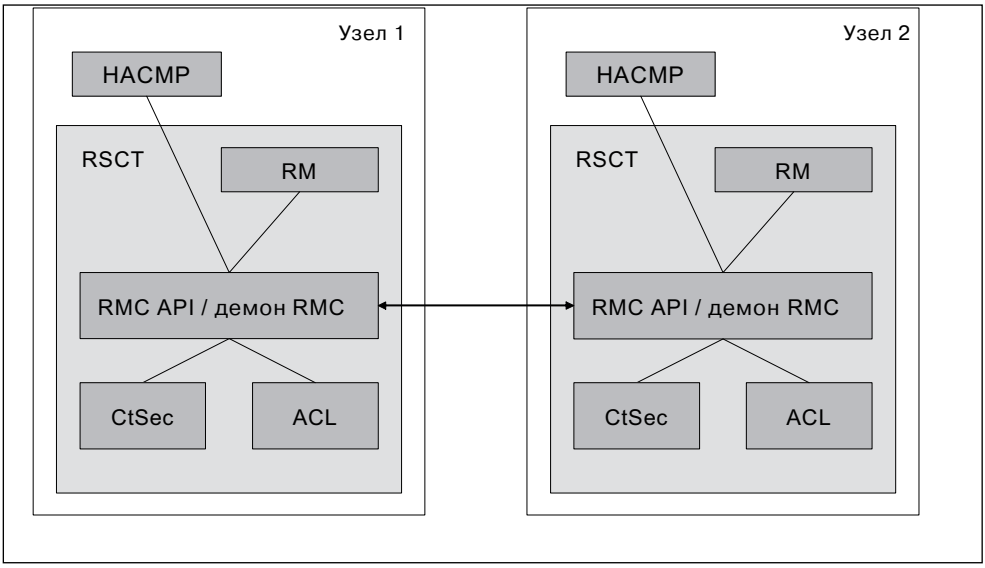

Рис. 9.7. Обзор базовых коммуникаций кластера

ленному), вызывается подсистема RSCT и запросы отправляются на узел, на котором находится ресурс. (рис. 9.7).

Так как удаленные подключения представляют собой подключения сокетов TCP/IP, их необходимо защищать, чтобы обеспечить подлинность запрашивающего узла и серверного узла в кластере. Здесь в действие вступают службы безопасности кластера (Cluster Security Services, CtSec).

Между узлами кластера осуществляется связь типа «клиент-сервер». В следующих разделах под **клиентом** подразумевается приложение, например локальный диспетчер ресурсов или функция HACMP, использующая клиентский API RMC.

Эти клиентские приложения подключены к общим библиотекам RMC и CtSec. Если эти клиенты запрашивают доступ к ресурсам на удаленном узле, демон RMC на этом удаленном узле будет **сервером** для этих запросов.

9.6.2 Обзор служб безопасности кластера (CtSec)

Службы безопасности кластера (Cluster Security Services, CtSec) интегрированы в подсистему RSCT и используются в RMC для определения подлинности клиента с узла. Этот процесс аутентификации создает контекст безопасности, применяемый RMC для связи между участвующими узлами в целях выполнения клиентских запросов.

> **Примечание.** Контекст безопасности реализуется на уровне «клиент-сервер», а не на уровне узла.

При создании этого контекста безопасности CtSec используют учетные данные для аутентификации. Эти учетные данные применяются для определения подлинности узла и клиентского приложения. Для доступа к ресурсам на удаленном узле, в процессе аутентификации оба узла отправляют и сравнивают учетные данные.

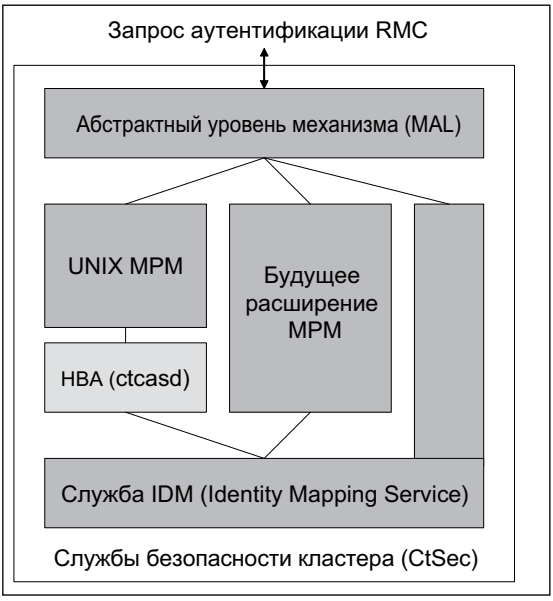

Рис. 9.8. Архитектура служб безопасности кластера

Этот процесс позволяет достичь следующих целей:

- клиент может представить о себе информацию, которую другие не смогут подделать;
- сервер может четко идентифицировать клиента по предоставленным учетным данным;
- клиент может быть уверен в том, что он получает данные с требуемого сервера.

Аутентификация на основе учетных данных использует другие компоненты для создания и идентификации учетных данных. Такая архитектура на основе компонентов, представленная на рис. 9.8, позволяет в будущем выполнять расширение уровня безопасности в RSCT.

В RMC службы CtSec отвечают только за аутентификацию. Вообще RMC отвечает за авторизацию с использованием таблицы управления доступом (access control list, ACL), разрешая или запрещая доступ к ресурсам в кластере.

9.6.3 Компоненты служб безопасности кластера (CtSec)

Библиотека CtSec содержит несколько компонентов, необходимых для обеспечения существующих и будущих функций безопасности.

Контекст безопасности, создаваемый CtSec, содержит учетные данные клиента и сервера, состояние аутентификации (выполнена или не выполнена) и информацию о сеансе (например, ключ сеанса или время завершения действия). Контекст безопасности создается компонентами CtSec и отправляется в RMC как результат аутентификации CtSec.

9.6.4 Абстрактный уровень механизма (MAL)

CtSec экспортирует интерфейс в приложения, требующие реализации безопасности кластера. Этот интерфейс называется абстрактным уровнем механизма (mechanism abstract layer, MAL). MAL обеспечивает общий, независимый от механизма интерфейс для базовых механизмов безопасности и отправляет эти общие инструкции на сконфигурированные модули безопасности.

Такие подключаемые модули механизма безопасности называются *подключаемыми модулями механизма (mechanism pluggable module, MPM)*. Результатом работы MAL и сконфигурированного MPM является контекст безопасности, содержащий учетные данные и ключ сеанса, используемые обеими сторонами, участвующими в процессе связи.

Если клиент отключает аутентификацию, установив значение переменной окружения CTSEC_CC_MECH=none, MAL не применяет MPM для аутентификации, а возвращает контекст безопасности без аутентификации как результат процесса аутентификации.

Конфигурирование MPM выполняется в файле /var/ct/cfg/ctcec.cfg. Этот файл содержит все MPM, которые службы CtSec должны использовать в процессе аутентификации. В будущих версиях компания IBM собирается разработать другие модули безопасности. Применение модулей основывается на предопределенных приоритетах.

Если клиент запрашивает доступ к ресурсам, MAL используется для инициирования контекста безопасности. В случае успешного выполнения процесса этот контекст кешируется уровнем MAL для последующих запросов и сохраняется в памяти до удаления приложением.

9.6.5 Подключаемый модуль механизма (MPM)

Каждый подключаемый модуль механизма (mechanism pluggable module, MPM) преобразует общие задачи, получаемые с уровня MAL, в необходимые задачи, используемые механизмом безопасности для выполнения запросов MAL.

MPM осуществляет сбор всех учетных данных, необходимых для выполнения процесса аутентификации каждым отдельным механизмом безопасности. MPM также осуществляет сопоставление сетевых идентификационных данных с локальными идентификационными данными [см. раздел 9.6.8, «Служба IDM (Identity Mapping Service)»].

В настоящее время поддерживается только UNIX MPM. Модуль UNIX MPM был создан для работы как в 32-, так и в 64-разрядных средах, в зависимости от ядра операционной системы. Благодаря модульной архитектуре MAL и MPM в последующих версиях могут быть добавлены другие механизмы безопасности. MPM представляют собой объектные модули, загружаемые MAL во время выполнения.

MPM расположены в каталоге /usr/sbin/rsct/lib. Каждый MPM должен иметь ссылку в каталоге /usr/lib/, указывающую на соответствующий файл в каталоге (пример 9.2).

Пример 9.2. Расположение MPM

```
ls -al /usr/sbin/rsct/lib/*.mpm*
-r--r--r--   1 bin   bin   192120 Sep 06 13:26 /usr/sbin/rsct/lib/unix.mpm
-r--r--r--   1 bin   bin   199952 Sep 24 11:27 /usr/sbin/rsct/lib/unix.mpm64
ls -al /usr/lib/*.mpm*
lrwxrwxrwx 1  root  system  27 Oct 11 10:22 /usr/lib/unix.mpm ->
                                            /usr/sbin/rsct/lib/unix.mpm
lrwxrwxrwx 1  root  system  29 Oct 11 10:22 /usr/lib/unix.mpm64 ->
                                            /usr/sbin/rsct/lib/unix.mpm64
```

9.6.6 Подключаемый модуль UNIX MPM

Ядром модуля UNIX MPM является клиентский API ctcasd (см. раздел 9.6.7, «Аутентификация на основе хостов с использованием ctcasd»), который создает учетные данные аутентификации на основе хостов (host-based authentication, HBA).

Компонент ctcasd вызывается только при использовании сокетов TCP/IP для удаленных подключений. Если запрос предназначен для локального хоста, UNIX MPM применяет сокеты домена UNIX в ядре. Безопасность ядра является доверенной, и не требует использования дополнительных функций безопасности.

9.6.7 Аутентификация на основе хостов с использованием ctcasd

Демон службы Cluster Technology Cluster Authentication Service (ctcasd) создает учетные данные на основе имени хоста и идентификационных данных клиента. Помните о том, что **клиентом** называется приложение, использующее клиентский API RMC.

Кроме того, ctcasd создает ключ сеанса, который может применяться в качестве симметричного ключа для обмена данными между узлами после аутентификации.

Этот ключ сеанса необходим для шифрования и дешифрования данных в RSCT с использованием алгоритма симметричного ключа, который имеет более высокое быстродействие, чем алгоритмы открытого и закрытого ключа (public and private key algorithms, PPK).

Чтобы обеспечить конфиденциальность данных о секретном ключе сеанса и целостность учетных данных хоста, ctcasd использует пару открытый/закрытый ключ для шифрования и дешифрования этой информации. В текущей реализации эта пара ключей генерируется с использованием алгоритма RSA512. Это можно изменить, задав 1024-разрядный ключ, путем редактирования файла конфигурации ctcasd (/usr/sbin/rsct/cfg/ctcasd.cfg).

Чтобы убедиться в том, что ctcasd использует корректный открытый ключ в процессе шифрования, открытые ключи в файле списка доверенных хостов (trusted host list, THL) связываются с именем хоста. Поэтому необходимо, чтобы все узлы в кластере HACMP выполняли разрешение имен идентичным образом.

Важно! Чтобы обеспечить разрешение имен хостов идентичным образом, все участники кластера должны использовать метод разрешения имен, дающий идентичные результаты на всех узлах в кластере HACMP. Метод и порядок разрешения имен может быть изменен в файле /etc/netsvc.conf (для систем AIX). Все хосты должны также использовать либо короткие, либо полные доменные имена хостов. Если кластер содержит узлы из разных доменов, необходимо применять полные доменные имена хостов.

Во-первых, ctcasd шифрует ключ сеанса с использованием открытого ключа целевого хоста из файла THL. Это гарантирует, что только целевой узел сможет дешифровать этот ключ сеанса своим закрытым ключом, что обеспечивает конфиденциальность данных.

Чтобы обеспечить подлинность учетных данных хоста, ctcasd выполняет шифрование всех структур учетных данных с применением закрытого ключа инициатора. Блок данных может быть дешифрован кем угодно с использованием открытого ключа отправителя, однако целевой узел может быть уверен, что он был отправлен требуемым узлом.

Распространение открытых ключей по всем узлам выполняет RSCT. При добавлении узла в одноранговый кластер RSCT выполняет команды для обмена открытыми ключами между управляющим сервером и его узлами.

Обмен открытыми ключами выполняется через сеть. Во время этого обмена сеть должна быть защищена от трассирования и подмены, так как ключи привязаны к узлу в кластере.

Внимание! Ключи ctcasd не идентичны ключам, используемым для шифрования сообщений HACMP.

9.6.8 Служба IDM (Identity Mapping Service)

Результат работы службы IDM (Identity Mapping Service) используется для авторизации. Служба IDM сопоставляет сетевые идентификационные данные с локальными идентификационными данными, если в файлах конфигурации (называемых сопоставлениями, maps) задано правило сопоставления.

Эти локальные идентификационные данные могут использоваться в RMC для извлечения разрешения из таблицы управления доступом RMC. Описание сбора подсистемой RMC разрешений доступа к ресурсам из таблиц управления доступом см. в разделе 9.6.9, «Таблица управления доступом RMC».

По умолчанию в кластере RSCT отсутствует общее пользовательское пространство, поэтому одним и тем же именам пользователей на различных узлах могут соответствовать разные разрешения доступа к ресурсам.

Как описывается в разделе 9.6.9, «Таблица управления доступом RMC», подсистема RMC осуществляет сбор разрешений доступа к ресурсам, используя сначала сетевые идентификационные данные клиента.

Например, если пользователь **david** должен существовать на всех узлах в кластере из четырех узлов и иметь доступ к ресурсам на определенном узле, файл ACL подсистемы RMC на этом узле должен содержать записи для всех сетевых идентификационных данных клиента. Пример 9.3 показывает выходные данные определенного ресурса в файле ACL RMC для четырех узлов.

Пример 9.3. Выходные данные /var/ct/cfg/ctrmc.acl для записи определенного ресурса

```
IBM.FileSystem
  david@node001   *   rw   # access for david from node 001
  david@node002   *   rw   # access for david from node 002
  david@node003   *   rw   # access for david from node 003
  david@node004   *   rw   # access for david from node 004
```

Это основная цель, для которой была разработана служба IDM. Служба IDM просто сопоставляет сетевые идентификационные данные в управляющем кластере с локальными идентификационными данными, чтобы избежать записи сотен строк в файл ACL RMC.

Файл конфигурации сопоставлений, /var/ct/cfg/ctsec_map.global, содержит отношения сопоставления между локальными и сетевыми идентификационными данными.

В нашем примере файл сопоставлений должен содержать только одну строку, чтобы сопоставить пользователя david на всех узлах кластера с локальными идентификационными данными mapped_david:

unix:david@<cluster>=mapped_david

Это обеспечивает сопоставление пользователя david с каждого узла в текущем активном кластере с локальными идентификационными данными mapped_david с использованием сопоставлений, инициируемых UNIX MPM. Эти локальные идентификационные данные не обязательно должны существовать в качестве реального

пользователя в операционной системе. Доступ к ресурсам осуществляется, даже если эти идентификационные данные не существуют локально (в /etc/passwd).

При использовании таких идентификационных данных с сопоставлением ресурс в файле ACL требует только одной записи, как показано в примере 9.4.

Пример 9.4. Файл ACL RMC, использующий локальные идентификационные данные

```
IBM.FileSystem
  mapped_david   *   rw   # access for david from node 001
```

В файле ACL можно легко различить сетевые идентификационные данные и локально сопоставленные идентификационные данные. Сетевые идентификационные данные всегда содержат @ и имя хоста или идентификатор узла, с которого подключается клиент. Локально сопоставленные идентификационные данные содержат только спецификатор, так как имя хоста уже было задано в файле сопоставлений.

Изначально глобальный файл сопоставлений содержит эти записи, как показано в примере 9.5 (например, unix: определяет MPM, используемый для аутентификации, а =root определяет локальные идентификационные данные).

Пример 9.5. Глобальный файл сопоставлений

```
cat /var/ct/cfg/ctsec_map.global
unix:root@<cluster>=root
unix:root@<any_cluster>=any_root
unix:*@LOCALHOST=*
```

В целом первая строка указывает на то, что для каждого запроса к RMC, исходящего от пользователя root с любого узла в текущем активном кластере, выполняется сопоставление с локальными идентификационными данными пользователя root. Для каждого запроса от пользователя root с узлов вне активного кластера выполняется сопоставление с any_root.

Каждый клиентский запрос локальных ресурсов с локального хоста от любого пользователя просто сопоставляется с локальным пользователем. Процесс просмотра файла сопоставлений выполняет поиск первого вхождения; это означает, что первый идентификатор, соответствующий заданным сетевым идентификационным данным, будет сопоставлен соответствующим локальным идентификационным данным, после чего процесс сопоставления останавливается.

9.6.9 Таблица управления доступом RMC

RMC использует таблицу управления доступом (access control list, ACL) для разрешения или запрета доступа к ресурсам в одноранговом домене RSCT. Таблица управления доступом обновляется HACMP во время процедур администрирования кластера (например, при добавлении и удалении узлов).

После успешной аутентификации клиента в RMC, выполняется поиск разрешений по заданным критериям в ACL (рис. 9.7). Файл ACL RMC имеет расположение /usr/sbin/rsct/cfg/ctrmc.acl.

Файл ACL состоит из разделов (stanzas), содержащих классы ресурсов и определенные разрешения пользователей и хостов в кластере. RMC применяет два типа идентификационных данных для извлечения разрешений, хранящихся в ACL. Эти типы идентификационных данных следующие:

- сетевые идентификационные данные клиента, представленные именем пользователя и именем хоста клиента, например root@node1.
- локально сопоставленные идентификационные данные, описанные в разделе 9.6.8, «Служба IDM (Identity Mapping Service)».

В процессе просмотра файла ACL сначала выполняется поиск сетевых идентификационных данных, а затем локальных идентификационных данных.

Часть 4

Расширенные задачи администрирования (с примерами)

Часть 4 охватывает следующие дополнительные темы:

- динамические LPAR (DLPAR) и виртуализация (VIO);
- расширение возможностей группы ресурсов;
- настройка событий;
- аспекты подсистемы хранения;
- сетевые функции.

10

Динамические LPAR (DLPAR) и виртуализация (VIO)

В этой главе обсуждаются следующие темы:

- реализация HACMP и DLPAR (на системах Power4);
- реализация HACMP и виртуализация (vio/vscsi/vlan).

Так как эти темы относятся к расширенным задачам администрирования, предполагается базовое понимание LPAR, DLPAR и виртуализации, а также соответствующей терминологии. Дополнительные сведения по этим темам см. в следующих книгах серии Redbooks:

- **Partitioning Implementations for IBM @server® p5 Servers**, SG24-7039;
- **Advanced POWER Virtualization on IBM @server® p5 Servers: Introduction and Basic Configuration**, SG24-7940[1].

Эти, а также многие другие книги можно скопировать по адресу

http://www.redbooks.ibm.com/

Представленные в этой главе сведения и шаги дополняют существующий кластер HACMP. Этапы конфигурирования кластера HACMP опущены, чтобы избежать повторения описаний в этой публикации. Сведения о конфигурировании базового кластера HACMP см. в гл. 4, «Сценарии установки кластера».

[1] Современное название книги – «Advanced POWER Virtualization on IBM System p5». Планируется ее выход на русском языке. *Примеч. науч. ред.*

10.1 Реализация DLPAR в HACMP

В этом разделе описываются следующие вопросы, относящиеся к DLPAR:

- требования;
- предоставление доступа к приложениям (application provisioning);
- определение DLPAR в HACMP;
- наша тестовая конфигурация;
- результаты тестирования.

Прежде чем реализовать LPAR и DLPAR, необходимо произвести надлежащее планирование соответствующей конфигурации. Важно понимать не только требования и способы осуществления, но и общее влияние каждого решения на реализацию в целом.

10.1.1 Требования

Для использования встроенных функций DLPAR и/или CUoD в HACMP на Power4 на всех узлах LPAR в кластере должны быть установлены **как минимум** следующие версии программного обеспечения:

- AIX 5.2;
- HACMP 5.2.0.1 (с IY58577 для поддержки DLPAR);
- APAR IY58497 (для поддержки CUoD);
- RSCT 2.3.3.1;
- OpenSSH 3.4p1.

Программное обеспечение OpenSSH можно получить из следующих источников:

- AIX 5.2 Bonus pack;
- AIX 5.3 Expansion pack;
- Linux Toolbox CD;
- скопировать с сайта

  http://sourceforge.net/projects/openssh-aix.

OpenSSH для AIX имеет собственные требования:

- rpm.rte;
- библиотека сжатия/распаковки zlib;
- демон генерирования псевдослучайных чисел (prngd);
- криптографические библиотеки OpenSSL.

Эти пакеты можно скопировать по адресу:

http://www-1.ibm.com/servers/aix/products/aixos/linux/download.html.

Пакет OpenSSL можно получить следующим образом: щелкнуть по ссылке «AIX Toolbox Cryptographic Content», зарегистрироваться и принять лицензионное соглашение.

Подключение HMC к LPAR необходимо для правильного управления и работы функций DLPAR. Для выполнения удаленных операций DLPAR, HMC должен быть подключен к общей сети с LPAR. Кроме того, на HMC должны быть установлены **как минимум** следующие версии программного обеспечения:

- HMC 3 Version 2.6;
- HMC build level/firmware 20040113.1 или выше.

> **Важно!** APAR IY69525 для HACMP V5.2 или APAR IY73051 для HACMP V5.3 необходимы для поддержки функций DLPAR, CUoD и CBU на системах Power5.

На момент написания данной книги APAR, необходимые для поддержки Power5, были недоступны. Уровни дополнительного программного обеспечения, необходимые для поддержки Power5, также еще не определены[1].

> **Внимание!** Основное требование к конфигурации заключается в том, что имя раздела LPAR, имя хоста AIX и имя узла HACMP должны совпадать. Это показано на рис. 10.1.

Прочие аспекты

При планировании кластера, включающего операции DLPAR, следует учитывать некоторые аспекты, в частности следующие:

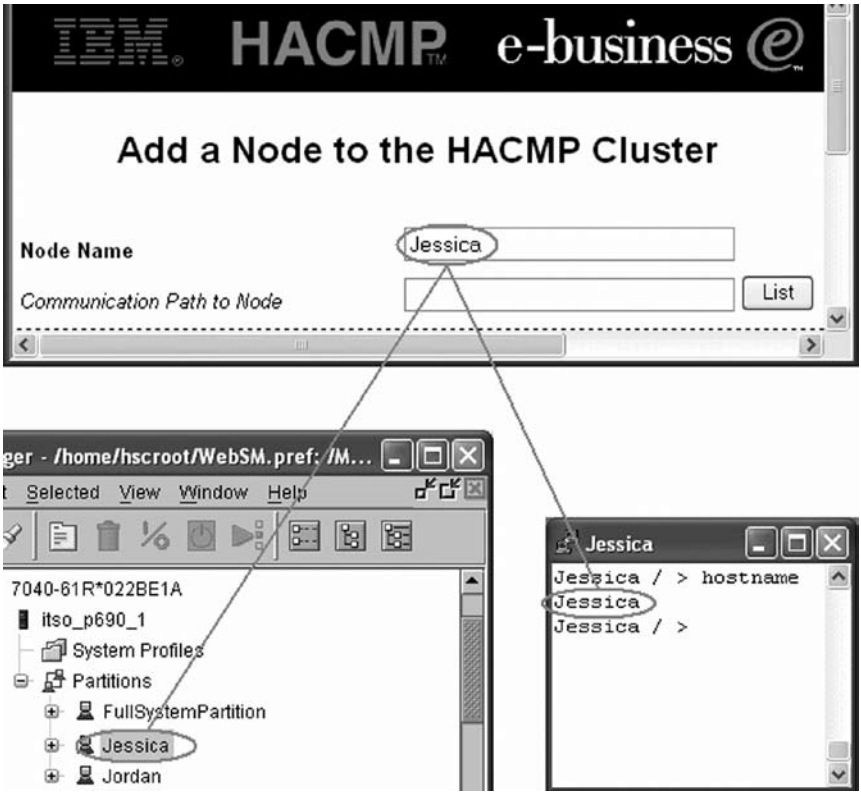

Рис. 10.1. Соответствие имени раздела, имени хоста AIX и имени узла HACMP

[1] Информацию о поддержке HACMP на системах IBM System p5 с DLPAR и APV вы можете найти в документации на сайте IBM. *Примеч. науч. ред.*

- во время событий DLPAR возможно возникновение сообщения config_too_long;
- сочетание разделов LPAR с другими системами (т. е. не с разделами);
- обеспечение ресурсами CUoD.

С появлением поддержки Power5 DLPAR/CUoD возможны следующие дополнительные конфигурации:

- сочетание Power4 и Power5 DLPAR;
- использование общих и/или выделенных процессоров;
- использование процессоров capped (с ограничениями) и/или uncapped (без ограничений).

Как и в любом кластере, конфигурацию следует тщательно протестировать. Это включает все, что только можно сделать для имитации или создания реальной рабочей нагрузки для максимальной реалистичности сценариев тестирования.

10.1.2 Обеспечение приложений ресурсами

Дополнительные сведения по этой теме см. в руководстве **High Availability Cluster Multi-Processing Administration Guide**, SC23-4862-06.

Этот раздел описывает последовательность действий в кластере HACMP, если сконфигурирована функция обеспечения приложения ресурсами (application provisioning) через DLPAR и CUoD. Также раздел содержит несколько примеров, иллюстрирующих выделение ресурсов в зависимости от требований к ресурсам.

Обзор

При конфигурировании LPAR в HMC (вне HACMP) указывается минимальные, желательные и максимальные значения количества процессоров и объема памяти. Эти значения можно получить при запуске команды **lshwres** в HMC. Указанные минимальные ресурсы должны быть доступны на момент запуска узла LPAR. Если в свободном пуле фрейма доступно больше ресурсов, LPAR может выделить желаемое (desired) количество ресурсов.

Во время операций динамического выделения ресурсов система не позволяет, чтобы значения количества процессоров и объема памяти выходили за пределы минимального и максимального значений, заданных для LPAR.

HACMP получает минимальные и максимальные значения для LPAR и использует их для выделения и освобождения процессоров и памяти при запуске и остановке серверов приложений на узле LPAR.

HACMP запрашивает выделение ресурсов для DLPAR в HMC перед запуском серверов приложений и освобождает ресурсы после остановки серверов приложений. Диспетчер кластера (Cluster Manager) ожидает завершения этих событий, прежде чем продолжить обработку событий в кластере.

HACMP осуществляет управление выделением и освобождением ресурсов для серверов приложений последовательно, даже если группы ресурсов обрабатываются параллельно. Это позволяет устранить конфликты между серверами приложений, возникающие при попытке выделить или освободить одни и те же ресурсы процессора

или памяти. Таким образом, необходимо выполнить тщательное конфигурирование кластера для корректной обработки всех запросов процессоров и памяти в LPAR.

Также важно учитывать следующие аспекты:

- когда HACMP получает дополнительные ресурсы для сервера приложения при перемещении сервера приложения на другой узел, HACMP освобождает только те ресурсы, которые больше не нужны для поддержки этого приложения на узле;
- HACMP **не** выполняет запуск и остановку узлов LPAR.

Можно создать настраиваемое событие или выполнить настройку скриптов запуска/остановки, чтобы останавливать узлы LPAR по требованию.

Получение ресурсов DLPAR и CUoD

При конфигурировании сервера приложения с указанием минимального и желаемого количества ресурсов (процессоров или памяти) HACMP определяет, нужно ли выделять дополнительные ресурсы для узла, и, если возможно, выделяет их.

В целом HACMP пытается выделить максимально возможное количество ресурсов, чтобы достичь оптимального количества ресурсов для приложения, и, если возможно, использует для этого CUoD.

Узел LPAR с минимальным количеством ресурсов

Если узлу доступно только лишь минимальное количество ресурсов, HACMP запрашивает дополнительные ресурсы через DLPAR и CUoD (если применимо).

В целом HACMP начинает подсчет дополнительных ресурсов, требуемых для работы приложения, относительно минимального количества. Другими словами, минимальное количество ресурсов используется для операций обеспечения функционирования самого узла и **не** используется для содержания приложения.

Узел LPAR с достаточным количеством ресурсов для содержания приложения

Узел LPAR, который готовится к тому, чтобы содержать приложение, должен уже иметь достаточно ресурсов (в дополнение к минимальному количеству ресурсов для LPAR), чтобы обеспечить оптимальное количество ресурсов для данного приложения.

В этом случае HACMP не выделяет какие-либо дополнительные ресурсы и приложение может быть успешно запущено на узле LPAR. Кроме того, HACMP определяет, имеет ли узел достаточно ресурсов, чтобы содержать это приложение в дополнение ко всем остальным серверам приложений, которые могут быть запущены на узле.

Запрос ресурсов в свободном пуле и в пуле CUoD

Если ресурсов в свободном пуле недостаточно для обеспечения общего количества ресурсов, запрошенного для выделения (минимальные требования для одного или нескольких приложений), HACMP запрашивает ресурсы в CUoD (если включено).

Если HACMP обеспечивает требование минимального количества ресурсов для сервера приложения, работа сервера приложения продолжается. Работа сервера приложения продолжается, даже если общее желаемое количество ресурсов (для одного или нескольких приложений) не обеспечивается или обеспечивается частично. В целом HACMP пытается получить желаемое количество ресурсов, запрашиваемое для приложения.

Если ресурсов недостаточно для содержания приложения, HACMP запускает операции восстановления группы ресурсов для перемещения группы ресурсов на другой узел.

Минимальное количество ресурсов, запрошенное для приложения, не может быть обеспечено

В некоторых случаях, даже после того как HACMP запрашивает использование ресурсов из пула CUoD, выделяемое количество ресурсов меньше минимального количества, заданного для приложения.

Если ресурсов все еще недостаточно для того, чтобы содержать приложение, архитектура HACMP запускает операции восстановления группы ресурсов для перемещения группы ресурсов на другой узел.

Узел LPAR содержит серверы приложений

Во всех случаях HACMP проверяет, содержит ли узел серверы приложений, требующие обеспечения приложения ресурсами (application provisioning), и не превышено ли максимальное значение для узла LPAR.

Во время последующих перемещений при сбое HACMP проверяет, не превышает ли минимальное количество запрашиваемых ресурсов еще для одного сервера приложения в сумме с уже выделенным количеством ресурсов для приложений, находящихся на узле, максимальное значение для LPAR. В этом случае HACMP пытается выполнить операции восстановления группы ресурсов для перемещения группы ресурсов на другой LPAR. Заметьте, что при конфигурировании требований DLPAR и CUoD для этого сервера приложения во время верификации кластера HACMP выдает предупреждение, если общее запрошенное количество ресурсов для всех приложений превышает максимальное количество ресурсов для LPAR.

Выделение ресурсов в кластере с несколькими приложениями

Если у вас есть несколько приложений в различных группах ресурсов кластера с узлами LPAR и несколько приложений потенциально могут запросить дополнительные ресурсы через функции DLPAR и CUoD, выделение ресурсов в кластере становится более сложным.

В зависимости от порядка обработки группы ресурсов некоторые группы ресурсов (а значит, и приложения) могут не запуститься. Более подробно это описывается в разделе «Примеры использования ресурсов DLPAR и CUoD».

Освобождение ресурсов DLPAR и CUoD

При остановке сервера приложения на узле LPAR (группа ресурсов перемещается на другой узел) HACMP освобождает только те ресурсы, которые больше не нужны для поддержки этого приложения на узле. Ресурсы освобождаются в свободный пул фрейма.

HACMP сначала освобождает ресурсы DLPAR или CUoD, которые были получены в последнюю очередь. Это означает, что ресурсы CUoD не всегда освобождаются раньше ресурсов DLPAR.

Свободный пул ограничен только одним фреймом. Другими словами, в кластерах, сконфигурированных на основе двух фреймов, HACMP не запрашивает из второго фрейма ресурсы для узла LPAR, расположенного в первом фрейме.

Кроме того, если LPAR-1 освобождает приложение, помещающее некоторые ресурсы DLPAR в свободный пул, LPAR-2, использующий ресурсы CUoD, не предпринимает попыток освободить ресурсы CUoD и получить свободные ресурсы DLPAR.

Остановка узлов LPAR

При принудительном отключении диспетчера кластера (Cluster Manager) на узле LPAR с последующим завершением работы LPAR (вне HACMP) происходит освобождение ресурсов процессора и памяти (без участия HACMP), которые становятся доступными для других групп ресурсов, запущенных на других LPAR. HACMP не следит за ресурсами процессора и памяти, выделенными для LPAR, и не сохраняет их для использования при реинтеграции узла LPAR в кластер.

> **Примечание.** Если используется лицензия On/Off для ресурсов CUoD и происходит завершение работы узла LPAR (вне HACMP), выполняется освобождение ресурсов CUoD (без участия HACMP) в свободный пул, однако лицензия On/Off остается включенной. Может потребоваться вручную отключить лицензию для ресурсов CUoD, находящихся в свободном пуле (это позволит вам не платить за ресурсы, не используемые в настоящее время).

Если LPAR не был остановлен после принудительного отключения диспетчера кластера на узле, ресурсы процессора и памяти остаются выделенными для LPAR и используются при реинтеграции LPAR в кластер.

Динамическое изменение ресурсов DLPAR и CUoD

Можно выполнить изменение требований к ресурсам DLPAR и CUoD для серверов приложений без остановки служб кластера. После внесения изменений необходимо выполнить синхронизацию кластера.

Новая конфигурация не отображается до возникновения следующего события, вызывающего освобождение приложения (а значит, и группы ресурсов) и его перехват другим узлом. Другими словами, изменение в требованиях к ресурсам процессора, памяти или обоих компонентов не вызывает перерасчета ресурсов DLPAR. HACMP не останавливает и не перезапускает серверы приложений только для того, чтобы внести изменения в систему обеспечения приложения ресурсами (application provisioning).

Если какое-либо **другое** изменение динамической реконфигурации (т. е. rg_move) вызывает освобождение и перехват групп ресурсов, новые требования к ресурсам для DLPAR и CUoD используются в конце события динамической реконфигурации.

Примеры использования ресурсов DLPAR и CUoD

Следующие примеры объясняют выделение и освобождение процессора. Процессы выделения и освобождения памяти очень похожи. Эти примеры лишь описывают механизм работы, тогда как реальные результаты из тестовой конфигурации приведены в разделе «Результаты тестирования».

Примечание. Помните о том, что после того как HACMP получит дополнительные ресурсы для сервера приложения, когда сервер снова перемещается на другой узел, он сохраняет свои ресурсы; другими словами, узел LPAR освобождает все полученные им дополнительные ресурсы.

Конфигурация представляет собой фрейм на восемь процессоров с кластером из двух узлов (каждый из которых представляет собой LPAR). Пул CUoD содержит два процессора, доступные посредством активизации CUoD. Узлы (разделы) имеют характеристики, представленные в табл. 10.1 и 10.2.

Таблица 10.1. Характеристики partition profiles

Имя узла	Минимум для LPAR	Максимум для LPAR
Longhorn	1	9
Hurricane	1	5

Определено три сервера приложений, каждый из которых относится к отдельной группе ресурсов.

Таблица 10.2. Требования приложений

Имя сервера приложения	Желаемое количество процессоров	Минимальное количество процессоров	Допускается использование CUoD
App1	1	1	да
App2	2	2	нет
App3	4	4	нет

Пример 1. Процессоры не добавляются при запуске, некоторые процессоры освобождаются при остановке

Параметры стартовой конфигурации:

- Longhorn имеет 3 выделенных процессора;
- Hurricane имеет 1 выделенный процессор;
- свободный пул содержит 4 выделенных процессора.

Серверы приложений запускаются в следующем порядке:

- Longhorn запускает App2. Не происходит выделения процессоров для выполнения требования наличия трех процессоров (3 процессора представляют сумму минимального количества процессоров для LPAR первого узла, равного единице, и желательного количества процессоров для App2, равного двум).
- Longhorn останавливает App2. Освобождается 2 процессора; остается 1 процессор, что соответствует минимальному требованию. (Так как отсутствуют другие запущенные серверы приложений, единственное требование – наличие минимального количества процессоров для LPAR Longhorn, равного единице).

Пример 2. Процессоры не добавляются из-за порядка обработки групп ресурсов

В этом примере мы начинаем с тех же параметров конфигурации, которые представлены в первом примере.

Серверы приложений запускаются следующим образом:

- Longhorn запускает App1; процессоры не выделяются, так как требование наличия двух процессоров выполняется. Longhorn запускает App3; выделяется 3 процессора для выполнения требования наличия шести процессоров. Теперь свободный пул содержит 1 процессор.
- Longhorn пытается запустить App2. После того как Longhorn подхватит App1 и App3, общее количество процессоров, требуемое узлу Longhorn для выполнения этих требований, равно шести (сумма минимума для LPAR Longhorn, равного единице, желательного количества для App1, равного единице, и желательного количества для App3, равного четырем).

Так как минимальное количество процессоров для App2 равно двум, то для получения App2 узлу Longhorn требуется выделить 2 дополнительных процессора, однако в свободном пуле остался только 1 процессор, что не позволяет выполнить минимальное требование для App2, составляющее 2 процессора. Группа ресурсов с App2 не подхватывается локально, так как свободный пул содержит только 1 процессор, и использование CUoD не разрешено. Если другие участвующие узлы отсутствуют, группа ресурсов переходит в состояние ERROR.

Если бы узел Hurricane был членом группы ресурсов App2 и был активным в кластере, то произошел бы вызов rg_move для переноса группы ресурсов на узел Hurricane.

Пример 3. Успешное выделение и освобождение ресурсов CUoD

Стартовая конфигурация имеет следующие параметры:

- Longhorn имеет 3 выделенных процессора;
- Hurricane имеет 1 выделенный процессор;
- свободный пул содержит 4 выделенных процессора.

Серверы приложений запускаются в следующем порядке:

- Longhorn запускает App3; выделяется 2 процессора, чтобы выполнить условие наличия пяти процессоров;
- Longhorn запускает App2; выделяется 2 процессора, чтобы выполнить условие наличия семи процессоров. В свободном пуле процессоры отсутствуют;
- Longhorn запускает App1, из CUoD выделяется 1 процессор, чтобы выполнить условие наличия восьми процессоров;
- Longhorn останавливает App3, освобождается 4 процессора, 1 из которых помещается обратно в пул CUoD.

Пример 4. Отказ группы ресурсов, невыполнение требования минимальных ресурсов

В этом примере при получении группы ресурсов происходит сбой в связи с тем, что минимальные требуемые ресурсы недоступны при достижении максимума в LPAR.

Конфигурация имеет следующие параметры:

- Longhorn имеет 1 процессор;
- Hurricane имеет 1 процессор;
- свободный пул содержит 6 процессоров.

Серверы приложений запускаются в следующем порядке:

- Hurricane запускает App3; выделяется 4 процессора для выполнения требования наличия пяти процессоров. В свободном пуле остается 2 процессора;
- Hurricane пытается запустить App2, однако App2 переходит в состояние ERROR, так как максимальное количество процессоров для Hurricane – 5 и Hurricane не может получить больше процессоров.

Примечание. Если бы в качестве минимума для App2 было установлено нулевое значение, а не единица, получение ресурсов прошло бы успешно, так как при этом не требуются дополнительные ресурсы.

Пример 5. Отказ группы ресурсов, минимум и максимум для LPAR одинаковы

Ниже приведен реальный пример, имевший место на ранних этапах тестирования. Он иллюстрирует прямое следствие неправильного планирования обеспечения приложений ресурсами.

Мы все еще используем фрейм на восемь процессоров, однако дополнительные серверы приложений и узлы неприменимы к данному примеру. Конфигурация LPAR для узла Longhorn представлена в табл. 10.3.

Таблица 10.3. Свойства LPAR для узла Longhorn

Минимум для LPAR	Желаемое значение для LPAR	Максимум для LPAR
4	4	4

Сервер приложения App1 имеет параметры, представленные в табл. 10.4:

Таблица 10.4. Требования приложений для App1

Минимальное количество процессоров	Максимальное количество процессоров
1	4

Стартовая конфигурация представлена ниже.

- Longhorn имеет 4 выделенных процессора;
- Свободный пул содержит 4 процессора;

Сервер приложения App1 запускается локально на узле Longhorn. В процессе получения проверяется минимум для LPAR, который прибавляется к минимуму сервера приложений, что в сумме дает 5. Эта сумма превышает максимум для LPAR, вследствие чего группа ресурсов переходит в состояние ERROR.

Хотя технически LPAR может иметь достаточно ресурсов для содержания приложения, сочетание параметров вызывает отказ. Вообще говоря, минимум и максимум у вас не будут одинаковы.

Такой ситуации можно было избежать одним из трех способов:

- измените минимум для LPAR на 3 или меньшее значение;
- измените максимум для LPAR на значение больше четырех;
- измените минимальное количество процессоров для App1 на 0.

10.1.3 Конфигурирование DLPAR в HACMP

Следующие сведения частично взяты из существующего официального документа (whitepaper), созданного перед интеграцией DLPAR с HACMP. Этот документ был посвящен настройке скриптов событий HACMP для использования DLPAR. Однако он также включал и другие основные этапы подготовки. Этот документ был предназначен для внутреннего пользования компанией IBM и ее бизнес-партнерами.

Мы рассмотрим следующие этапы конфигурирования DLPAR в кластере HACMP:

- разрешение имен;
- установка ssh на узлах HACMP;
- конфигурирование доступа к HMC через ssh;
- определение HMC и управляемых систем в HACMP;
- определение ресурсов DLPAR в HACMP (т. е. настройка обеспечения приложений ресурсами).

Разрешение имен

Одна из распространенных проблем состоит в несогласованности разрешения имен между всеми системами. Если разрешение имен не сконфигурировано корректно, функцию DLPAR использовать нельзя. Базовая инфраструктура RSCT (Reliable Scalable Cluster Technology) предполагает использование одинакового разрешения имен хостов на всех участвующих узлах. В противном случае RSCT не сможет осуществлять обмен данными должным образом.

Убедитесь в том, что все узлы и HMC настроены одинаково, просмотрев следующий список. Словосочетание «все системы» включает все узлы HACMP и HMC.

- Все системы должны выполнять разрешение имен и IP-адресов участвующих хостов идентичным образом. Это относится и к обратному разрешению имен.
- Все системы должны использовать один и тот же способ разрешения имен, либо короткое, либо длинное разрешение имен.
- Все системы должны использовать один порядок разрешения имен, либо локальный, либо удаленный. Для этого следует просмотреть следующие файлы:
  - /etc/hosts на всех системах;
  - /etc/netsvc.conf на всех узлах AIX;
  - /etc/host.conf на HMC.

Просмотр файлов на системах AIX не представляет сложности. Однако выполнение этой операции в HMC требует дополнительного рассмотрения.

Убедитесь в том, что HMC видит LPAR хоста. Имена и IP-адреса хостов должны быть указаны в списке хостов HMC. Мы рекомендуем конфигурировать информацию хостов через консоль HMC, так как каждая версия кода HMC ограничивает опции командной строки. В Power4 HMC проверка включает следующие действия: **HMC Maintenance (Обслуживание HMC) -> System Configuration (Конфигурирование системы) -> Customize Network Settings (Настройка параметров сети) -> Hosts (Узлы)**. Если адреса и имена не указаны, их можно добавить, щелкнув по кнопке **New (Новый)**. Конфигурация нашего файла узлов HMC представлена на рис. 10.2.

> **Примечание.** При использовании Power5 HMC нужно выбрать HMC Management (Управление HMC) -> HMC Configuration (Конфигурирование HMC) -> Customize Network Settings (Настройка параметров сети).

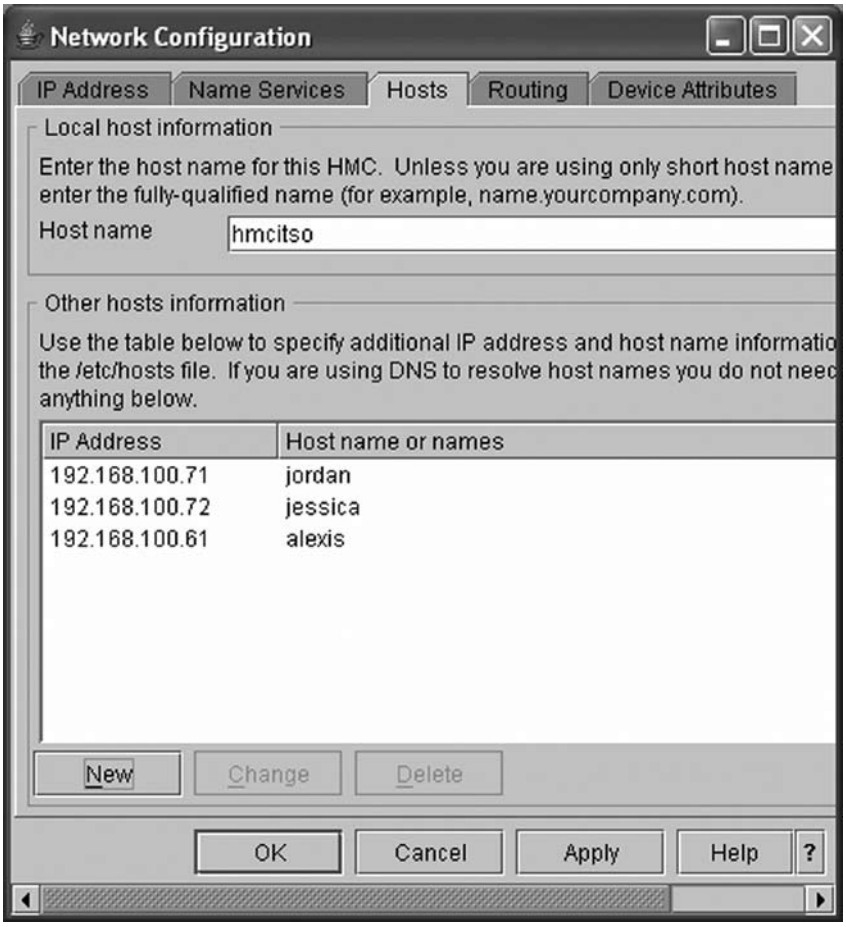

Рис. 10.2. Узлы HMC

Установка и конфигурирование SSH на узлах HACMP

Для использования операций удаленного выполнения команд в HMC нужно, чтобы на узлах HACMP был установлен SSH. HMC должен быть сконфигурирован таким образом, чтобы разрешать доступ из этих разделов.

В этом разделе мы рассмотрим установку пакетов ssh, включая порядок их установки. Важно соблюдать порядок установки этих пакетов. В большинстве случаев при нарушении порядка установки выдается сообщение об ошибке, указывающее, какой пакет следует установить в первую очередь.

Для каждой версии SSH и HMC эти этапы могут несколько различаться. Мы описали процессы, которые мы использовали при успешной реализации своей среды. Информацию по установке SSH в каждой версии AIX можно найти по адресу

http://www-1.ibm.com/support/docview.wss?uid=isg1pTechnote0707

Установка SSH

В разделе «Требования» мы рассмотрели, какие пакеты необходимы и где их можно получить. Перечисленные ниже действия предполагают, что эти пакеты уже были скопированы на узлы HACMP. Мы решили поместить все свои образы в общий каталог установки **/usr/sys/inst.images**.

Пакет **rpm.rte** должен быть установлен до установки дополнительных пакетов rpm. Для его установки можно использовать либо **installp**, либо **smit install_all**. Успешность его установки можно проверить командой **lslpp -l rpm.rte** (пример 10.1).

Пример 10.1. Проверка установки rpm

```
Jordan / > lslpp -l rpm.rte
  Fileset                      Level  State      Description
  ----------------------------------------------------------------------
Path: /usr/lib/objrepos
  rpm.rte                   3.0.5.36  COMMITTED  RPM Package Manager
Path: /etc/objrepos
  rpm.rte                   3.0.5.36  COMMITTED  RPM Package Manager
```

Теперь можно установить остальные необходимые пакеты с использованием команды **rpm**. Мы установили эти пакеты, как показано в примере 10.2.

Пример 10.2. Установка пакетов, необходимых для работы openSSH

```
rpm -i zlib-1.2.1-2.aix5.1.ppc.rpm
rpm -i prngd-0.9.23-3.aix4.3.ppc.rpm
rpm -i openssl-0.9.7d-2.aix5.1.ppc.rpm
rpm -i openssl-devel-0.9.7d-2.aix5.1.ppc.rpm
rpm -i openssl-doc-0.9.7d-2.aix5.1.ppc.rpm
```

Это выполняет требования, необходимые для установки SSH. В AIX 5.1 и более поздних версиях пакет openSSH имеет формат **installp**. Если образ был извлечен из tar-пакета, можно теперь его установить с использованием **smitty install_all**. Необходимо установить три основных набора файлов. Эти наборы файлов и результаты нашей установки представлены в примере 10.3.

Пример 10.3. Установка SSH

```
smitty install_all
> openssh.base                                                    ALL ¦
  ¦    + 3.8.0.5202  Open Secure Shell Commands                       ¦
  ¦    + 3.8.0.5202  Open Secure Shell Server                         ¦
  ¦ > openssh.license                                             ALL ¦
  ¦    + 3.8.0.5202  Open Secure Shell License                        ¦
  ¦                                                                   ¦
  ¦ > openssh.man.en_US                                           ALL ¦
  ¦    + 3.8.0.5202  Open Secure Shell Documentation - U.S. English   ¦

Name                 Level           Part        Event       Result
-------------------------------------------------------------------
openssh.license      3.8.0.5202      USR         APPLY       SUCCESS
openssh.base.client  3.8.0.5202      USR         APPLY       SUCCESS
openssh.base.server  3.8.0.5202      USR         APPLY       SUCCESS
openssh.base.client  3.8.0.5202      ROOT        APPLY       SUCCESS
openssh.base.server  3.8.0.5202      ROOT        APPLY       SUCCESS
openssh.man.en_US    3.8.0.5202      USR         APPLY       SUCCESS
```

Замечание. Необходимо выбрать Yes в поле принятия лицензионного соглашения.

После установки SSH необходимо выполнить конфигурирование узлов HACMP для доступа к HMC без паролей для выполнения удаленных операций DLPAR.

Конфигурирование доступа к HMC через SSH

Документ «Managing the Hardware Management Console» («Управление консолью управления оборудованием») содержит раздел, описывающий этапы настройки доступа к Remote Secure Shell. Это руководство можно скопировать с веб-сайта IBM по адресу

http://publib.boulder.ibm.com/infocenter/iseries/v1r2s/en_US/info/iphai/iphai.pdf

Ниже приведены действия, выполнявшиеся нами при настройке для включения доступа к SSH с узлов HACMP.

- Включение доступа к HMC через SSH.
- Генерирование ключей SSH на узлах HACMP.
- Включение беспарольного доступа к HMC через файл **authorized_keys2**.

Во-первых, следует убедиться в том, что консоль HMC настроена на выполнение удаленных операций; для этого нужно выполнить следующие действия:

1. В зоне навигации выберите **HMC Maintenance (Обслуживание HMC)**.

2. В зоне навигации выберите **System Configuration (Конфигурирование системы)**.

3. В зоне содержания выберите **Enable/Disable Remote Command Execution (Включение-отключение удаленного выполнения команд)**.

4. Установите флажок включения ssh.

> **Примечание.** Обычно рекомендуется создать отдельного пользователя HMC для удаленного выполнения команд, однако HACMP использует hscroot.

Необходимо создать каталог $HOME/.ssh для пользователя root, чтобы хранить ключи аутентификации. HACMP будет выполнять удаленные операции DLPAR ssh под учетной записью root. По умолчанию создается каталог с именем **/.ssh**, которые мы и применили.

Для генерирования открытых и закрытых ключей следует выполнить на каждом узле HACMP команду

/usr/bin/ssh-keygen -t rsa

При этом в каталоге **/.ssh** будут созданы следующие файлы:

закрытый ключ: id_rsa

открытый ключ: id_rsa.pub

Биты записи для группы и остальных пользователей отключаются. Убедитесь в том, что закрытый ключ имеет разрешение 600.

Открытый ключ HMC должен быть записан в файл **known_hosts** на каждом узле HACMP, и наоборот. Это легко делается путем выполнения **ssh** на HMC с каждого узла HACMP. При первом выполнении выводится запрос на вставку ключа в файл. Ответьте **Yes (Да)** для продолжения, после чего система запросит ввод пароля. Это связано с тем, что настройка беспарольного доступа к ssh еще не завершена (пример 10.4).

Пример 10.4. SSH в HMC

```
Jordan /tmp > ssh -l hscroot 192.168.100.69
The authenticity of host '192.168.100.69 (192.168.100.69)' can't be
established.
 RSA key fingerprint is 2d:50:3f:03:d3:51:96:27:5a:5e:94:f4:e3:9b:e7:78
 Are you sure you want to continue connecting (yes/no)?yes
Warning: Permanently added '192.168.100.69' (RSA) to the list of known
hosts.
```

При использовании двух консолей HMC, как в нашей тестовой конфигурации, необходимо повторить этот процесс для каждой консоли HMC. Также может потребоваться сделать это для всех узлов-участников, чтобы обеспечить выполнение операций ssh между ними (т. е. scp, sftp и ssh).

Чтобы обеспечить беспарольный доступ к ssh, нужно поместить открытые ключи каждого узла HACMP в файл **authorized_keys2** в HMC. Это можно сделать несколькими способами, однако ниже мы приведем действия, которые предприняли мы.

1. Создание файла **authorized_keys2** в консоли HMC.

2. Копирование (с использованием **scp**) открытого ключа со всех узлов на один компьютер.

3. Объединение (**cat**) всех файлов ключей в файл authorized_keys2.

4. Копирование (**scp**) объединенного файла через HMC /home/hscrtoot/.ssh.

На первом этапе мы вручную создали файл authorized_keys2. Это можно сделать как из командной строки в HMC, так и удаленно с клиента. Мы выполнили эту операцию удаленно с клиента. Мы решили сначала создать файл-макет (dummy file), после чего скопировать (scp) содержимое объединенных файлов ключей в HMC, заменяя, таким образом, наш файл-макет. Чтобы создать файл-макет мы выполнили с одного клиента следующую команду:

ssh hscroot@hmc "mkauthkeys --add '*' "

В консоли HMC убедитесь, что файл authorized_keys2 существует в каталоге .ssh. Для этого мы выполнили с клиента следующую команду:

ssh hscroot@hmc "ls -al .ssh/"

> **Примечание.** Можно выполнить команду mkauthkeys через ssh с каждого клиента AIX и добавлять по одному ключу зараз, как описано в руководстве Managing the Hardware Management Console. Однако с точки зрения синтаксиса это требует добавления строки ключа вручную. Мы считаем это обременительным, особенно при выполнении на нескольких системах.

Затем из каталога /.ssh в LPAR AIX мы создали копию открытого ключа и переименовали ее таким образом, чтобы имя файла включало имя локального узла. После этого мы скопировали (с использованием scp) открытый ключ каждого компьютера (Jessica и Alexis) на один узел (Jordan). Затем мы выполнили команду **cat** для создания файла authorized_keys2, содержащего информацию об открытых ключах для всех узлов HACMP. Команды, выполняемые на каждом узле, представлены в примере 10.5:

Пример 10.5. Защищенное копирование файла authorized_keys2 в HMC

```
Alexis /.ssh > cp id_rsa.pub id_rsa.pub.alexis
Alexis /.ssh > scp id_rsa.pub.alexis jordan:/.ssh/id_rsa.pub.alexis

Jessica /.ssh > cp id_rsa.pub id_rsa.pub.jessica
Jessica /.ssh > scp id_rsa.pub.jessica jordan:/.ssh/id_rsa.pub.jessica

Jordan /.ssh > cp id_rsa.pub id_rsa.pub.jordan
Jordan /.ssh > cat id_rsa.pub.alexis id_rsa.pub.jessica id_rsa.pub.jordan
>authorized_keys2

Jordan /.ssh > ls -al
total 64
drwx------   2 root     system          256 Jul 18 22:27 .
drwxr-xr-x  21 root     system         4096 Jul 14 02:11 ..
-rw-r--r--   1 root     system          664 Jun 16 16:31 authorized.keys2
```

```
-rw-------   1 root     system          883 Jun 16 14:12 id_rsa
-rw-r--r--   1 root     system          221 Jun 16 14:12 id_rsa.pub
-rw-r--r--   1 root     system          221 Jun 16 16:30 id_rsa.pub.alexis
-rw-r--r--   1 root     system          222 Jun 16 15:20 id_rsa.pub.jessica
-rw-r--r--   1 root     system          221 Jun 16 16:27 id_rsa.pub.jordan
-rw-r--r--   1 root     system         1795 Jul 14 04:08 known_hosts

Jordan/.ssh > scp authorized.keys2 hscroot@192.168.100.69:.ssh/authorized_keys2
hscroot@192.168.100.69's password:
authorized_keys2                                  100% 664 0.7KB/s   00:00
```

При выполнении команды копирования (**scp**) в консоль HMC система запрашивает пароль пользователя hscroot. После его ввода происходит копирование файла authorized_key2. Затем можно протестировать, работает ли беспарольный доступ с каждого узла путем выполнения команды ssh, как было показано в примере 10.4. Однако на этот раз вы должны будете попасть в приглашение командной строки оболочки HMC, как показано в примере 10.6.

Пример 10.6. Тестирование беспарольного доступа в ssh

```
Alexis /.ssh > ssh -l hscroot 192.168.100.
Last login:Thur Jun 16 22:46:51 2005 from 192.168.100.61
hscroot@hmcitso:~>
```

После проверки беспарольного доступа в HMC путем выполнения команды ssh этот этап завершается и верификация подключений к HMC в HACMP должна пройти успешно.

Определение имен HMC и управляемой системы

Для каждого узла HACMP должны быть заданы IP-адреса HMC, которые будут использовать DLPAR. В нашем примере каждый узел HACMP соответствует LPAR. Каждый LPAR назначается **управляемой системе**. Управляемыми системами называются системы, физически подключенные к HMC и управляемые HMC. Эти управляемые системы также должны быть определены в HACMP.

Имена управляемых систем выводятся в разделе навигации консоли HMC. В качестве имени управляемой системы может применяться имя, заданное пользователем, либо имя по умолчанию, содержащее тип и серийный номер компьютера.

Чтобы определить подключение HMC для каждого узла HACMP сделайте следующее:

1. В smit hacmp выберите **Extended Configuration (Расширенное конфигурирование) -> Extended Resource Configuration (Расширенное конфигурирование ресурсов) -> HACMP Extended Resources Configuration (Расширенное конфигурирование ресурсов HACMP) -> Configure HACMP Applications (Конфигурирование приложений HACMP) -> Configure HACMP for Dynamic LPAR and CUoD Resources (Конфигурирование ресурсов динамических LPAR и CUoD в HACMP) -> Configure Communication Path to HMC (Конфигурирование пути для связи с HMC) -> Add HMC IP Address for a Node (Добавить IP-адрес HMC для узла)** и нажмите Enter.

Появляется экран **Add HMC IP Address (Добавление IP-адреса HMC)**.

> **Замечание.** Можно использовать быстрый путь smit cladd_apphmc.dialog.

2. Заполните следующие поля:

- **Node Name (Имя узла).** Выберите имя узла для сопоставления с одним или несколькими IP-адресами HMC и управляемой системой.
- **HMC IP Address(es) (IP-адрес(а) HMC).** Введите один или несколько IP-адресов с разделяющими пробелами для HMC. При добавлении адресов нескольких консолей HMC HACMP пытается связаться с каждой консолью HMC, пока не будет найден рабочий путь для связи. После установления пути для связи HACMP использует этот путь для выполнения команд динамических логических разделов в данной консоли HMC.
- **Managed System Name (Имя управляемой системы).** Введите имя управляемой системы, в которой выполняется LPAR, представляющий узел. Максимальная длина составляет 32 символа.

3. Нажмите Enter.

На рис. 10.3 показан экран добавления информации HMC, описанной выше. Также представлена информация об имени управляемой системы HMC, используемом в нашей тестовой конфигурации.

> **Примечание.** Приложение, посвященное DLPAR/CUoD, в руководстве HACMP Administration Guide сообщает, что имя управляемой системы не может содержать символы подчеркивания. Однако в нашем примере мы использовали символы подчеркивания, и все работало нормально. Обсуждения с группой разработчиков подтвердили, что это ограничение не имеет места.

В процессе верификации кластера HACMP проверяет доступность HMC, выдавая **ping** на заданный IP-адрес. Если HMC реагирует, то HACMP проверяет способность каждого заданного узла HACMP поддерживать DLPAR, выдавая команду **lssycfg** через ssh в консоли HMC.

Конфигурирование обеспечения приложений ресурсами

Конфигурирование ресурсов динамических LPAR и CUoD для каждого сервера приложений, который может использовать выделенные ресурсы DLPAR или CUoD, состоит из следующих действий:

1. В smit hacmp выберите **Extended Configuration (Расширенное конфигурирование) -> Extended Resource Configuration (Расширенное конфигурирование ресурсов) -> HACMP Extended Resources Configuration (Расширенное конфигурирование ресурсов HACMP) -> Configure HACMP Applications (Конфигурирование приложений HACMP) -> Configure HACMP for Dynamic LPAR and CUoD Resources (Конфигурирование ресурсов динамических LPAR и CUoD в HACMP) -> Configure Dynamic LPAR and CUoD Resources for Applications (Конфигурирование ресурсов динамических LPAR и CUoD для прило-**

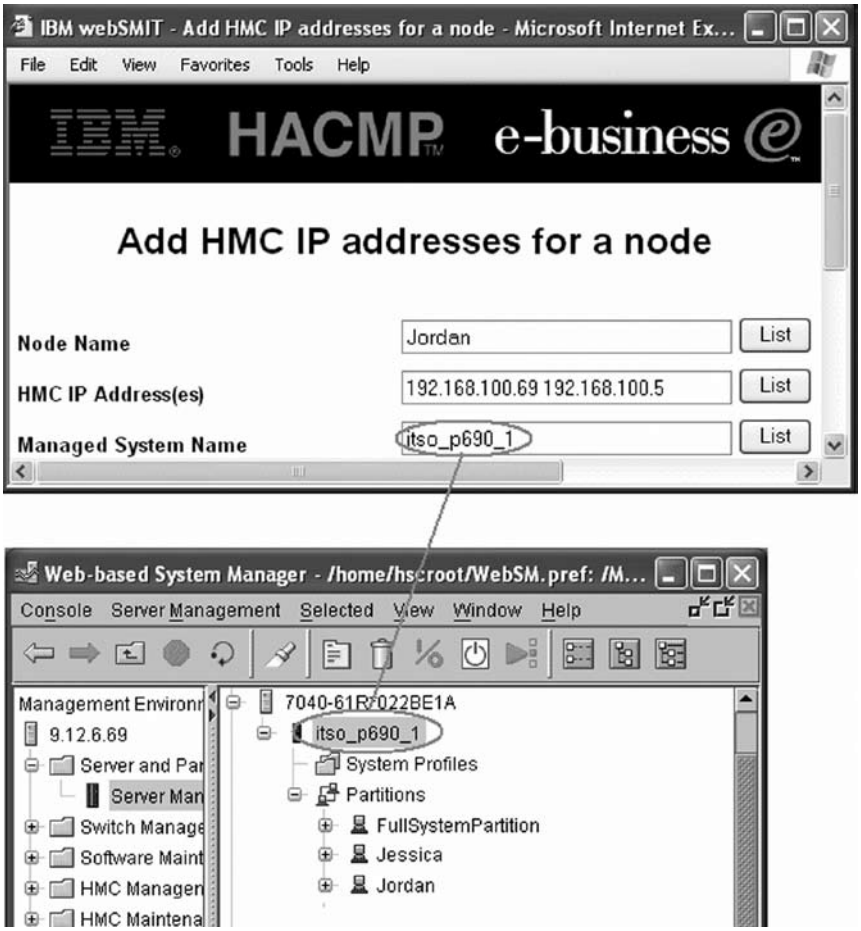

Рис. 10.3. Определение HMC и управляемой системы в HACMP

жений) -> Add Dynamic LPAR and CUoD Resources for Applications (Добавление ресурсов динамических LPAR и CUoD для приложений) и нажмите Enter.

Выводится список сконфигурированных серверов приложений.

> **Замечание.** Можно использовать быстрый путь smit cladd_appdlpar.dialog

2. Выберите сервер приложения из списка и нажмите Enter. Появляется экран указания требований для сервера приложения. Дополнительные сведения см. в справочных экранах и в разделе «Предоставление доступа к приложениям».

3. Заполните следующие поля:

- **Application Server Name (Имя сервера приложения).** Здесь указывается сервер приложений, для которого выполняется конфигурирование предоставления ресурсов динамического LPAR и CUoD, выбранных в предыдущем меню.

- **Minimum Number of CPUs (Минимальное количество процессоров).** Введите минимальное количество процессоров, которое следует получить при запуске сервера приложения. По умолчанию задано значение 0. Для осуществления обеспечения приложений ресурсами HACMP проверяет, сколько процессоров сверх минимального значения узел LPAR имеет на данный момент, сравнивает это число с минимумом, заданным в этом поле, и на основании этого запрашивает больше процессоров, если это необходимо.
- **Number of CPUs (Количество процессоров[1]).** Введите максимальное количество процессоров, которое HACMP попытается выделить на узле перед запуском этого приложения на данном узле. По умолчанию задано значение 0.
- **Minimum Amount of Memory (Минимальный объем памяти).** Введите объем памяти, который требуется получить при запуске сервера приложения. Значение должно быть кратно 256.
- **Use CUoD if resources are insufficient? (Использовать CUoD при нехватке ресурсов?).** По умолчанию установлено значение No. Выберите Yes, чтобы HACMP использовал CUoD (Capacity Upgrade on Demand), чтобы получить достаточно ресурсов для обеспечения запрошенного минимального количества. Использование CUoD требует ввода лицензионного ключа (кода активизации) в консоли управления оборудованием (Hardware Management Console, HMC), что может вызвать дополнительные затраты в связи с использованием лицензии CUoD.
- **I agree to use CUoD resources (Я согласен использовать ресурсы CUoD).** По умолчанию установлено значение No. Выберите Yes, чтобы подтвердить, что вы понимаете, что использование CUoD может вызвать дополнительные затраты. HACMP записывает ответ в файлы syslog и smit.log.

4. Нажмите Enter.

Когда приложение требует выделения дополнительных ресурсов на заданном узле, HACMP определяет, достаточно ли будет запросить только ресурсы DLPAR из свободного пула фрейма, чтобы обеспечить требуемое количество ресурсов, или же необходимо также запросить ресурсы CUoD для сервера приложения. После этого HACMP продолжает запрашивать желаемый объем памяти и оптимальное количество процессоров, если эти параметры были выбраны.

В процессе верификации HACMP выполняет проверку на то, чтобы введенные значения были ниже максимальных значений объема памяти и количества процессоров для LPAR. В противном случае HACMP выдает сообщение об ошибке с описанием этих требований.

HACMP также выполняет проверку на то, чтобы суммарное количество требуемых ресурсов для ВСЕХ серверов приложений, которые могут одновременно выполняться в LPAR, было меньше максимального значения для LPAR. При несоблюдении данного требования HACMP выдает предупреждение. Заметьте, что такая ситуация может произойти при последующих перемещениях при сбое. Другими словами, если узел LPAR уже содержит серверы приложений, требующие ресурсы DLPAR и CUoD, то при

---

[1] На рис.10.4 видно, что это поле называется Desired number of CPUs, т. е. «Желательное количество процессоров». *Примеч. науч. ред.*

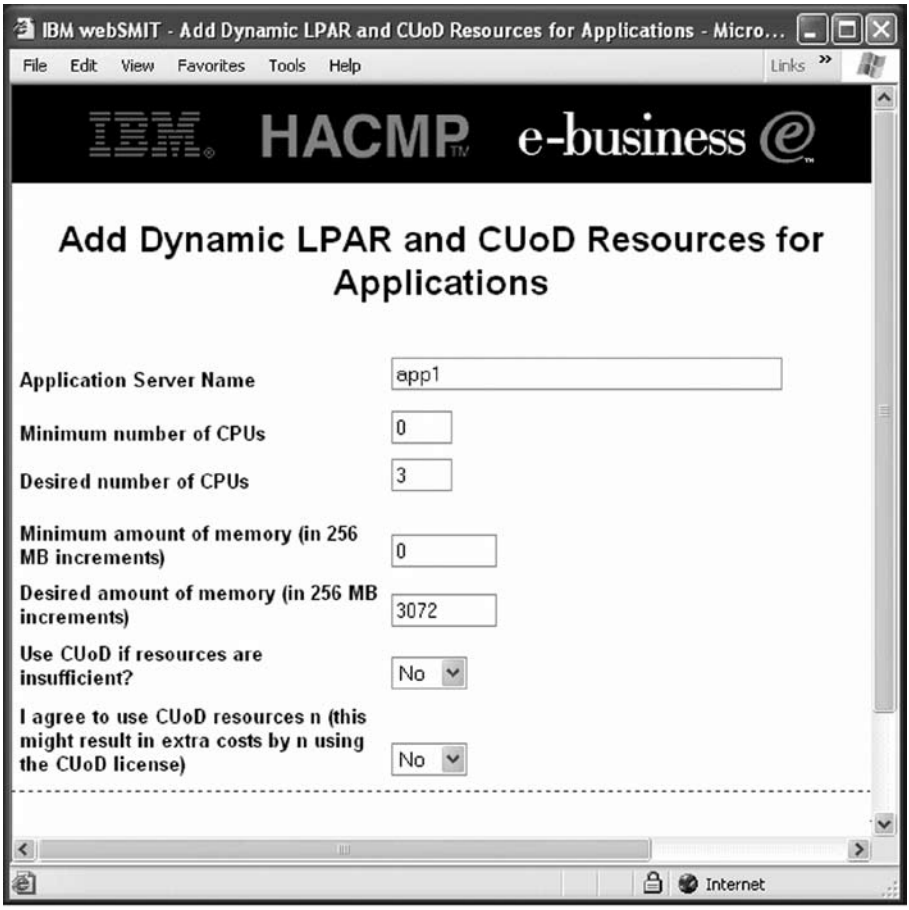

Рис. 10.4. Настройка обеспечения приложения ресурсами в HACMP

получении еще одного сервера приложения, LPAR может оказаться неспособным получить какие-либо дополнительные ресурсы, превышающие его максимум. HACMP выполняет соответствующую проверку и выдает предупреждение.

Пример настройки обеспечения приложения ресурсами в нашей тестовой конфигурации представлен на рис. 10.4.

После добавления подключений к HMC и предоставления доступа к приложениям необходимо выполнить синхронизацию кластера.

10.1.4 Устранение ошибок верификации HMC

В этом разделе описываются некоторые ошибки, которые могут возникнуть в процессе верификации, а также возможные причины их возникновения. Хотя некоторые сообщения об ошибках понятны без разъяснения, мы считаем, что любые замечания по устранению неполадок будут полезны.

Пример 10.7. Недоступность HMC в процессе верификации

```
ERROR: The HMC with IP label 192.168.100.69 configured on node jordan is not
reachable. Make sure the HMC IP address is correct, the HMC is turned on and
connected to the network, and the HMC has OpenSSH installed and setup with the
public key of node jordan.
```

В примере 10.7 сообщение об ошибке само по себе содержит предполагаемые причины проблемы. Ниже описывается, что можно сделать для определения источника проблемы.

- выполнить ping-опрос IP-адреса HMC;
- вручную подключиться к HMC с использованием команды **ssh -l hscroot hmcip.**

Если команда ssh выполняется неуспешно или если она запрашивает пароль, это указывает на неправильную конфигурацию ssh.

Пример 10.8. Ошибка верификации «Узел не поддерживает DLPAR»

```
ERROR: An HMC has been configured for node jordan, but the node does
not appear to be DLPAR capable.
```

Возникновение сообщения, представленного в примере 10.8, указывает на то, что доступ к HMC работает, однако определение LPAR, соответствующее определенному узлу, не сообщает о том, что оно поддерживает DLPAR. Это можно проверить вручную из командной строки HMC, как показано в примере 10.9.

Пример 10.9. Проверка поддержки DLPAR в LPAR

```
hscroot@hmcitso:~> lssyscfg -r lpar -m itso_p690_1 -n Jordan
Name     id   DLPAR  State    Profile       OpPanel
Jordan   001  NO     Running  Jordan_Prod
```

> **Примечание.** Синтаксис команд HMC может варьироваться в зависимости от уровня и типа кода HMC.

Это сообщение может быть вызвано, например, тем, что на узле выполняется версия AIX ниже 5.2, которая необходима для операций DLPAR. Оно также может быть вызвано некорректным обновлением RMC.

В процессе нашего тестирования возникло несколько событий за очень короткие периоды времени. В определенный момент наш LPAR выдал сообщение о том, что он не поддерживает DLPAR. Спустя некоторое время все снова работало нормально. Мы считаем, что это было вызвано нарушением синхронизации информации RMC между логическими разделами и HMC.

10.1.5 Конфигурация тестового кластера

Наша тестовая конфигурация состоит из трех LPAR на двух системах p690s: двух рабочих LPAR (Jordan, Jessica) на одной системе p690 (itso_p690_1) и одного дежурного LPAR (Alexis) на второй системе p690 (itso_p690_2). Каждая система p690 содержит восемь процессоров и 8 Гб памяти и подключена к двум консолям HMC в целях обеспечения избыточности.

В каждом разделе установлены следующие версии программного обеспечения:

- AIX 5.2 ML5;
- HACMP 5.2.0.3;
- RSCT 2.3.5.0;
- rpm-3.0.5-37.aix5.1;
- OpenSSH 3.8.1p1 (и следующие обязательные пакеты):
  - zlib 1.2.1-2;
  - prngd 0.9.23-3;
  - openssl 0.9.7d-2.

В каждом разделе установлены следующие адаптеры:

- (2) IBM Gigabit FC Adapter (#6228);
- (2) 10/100 Ethernet (#2975).

Оба HMC имеют:

- HMC 3 version 3.5;
- HMC build level/firmware 20050320.1.

Наши тестовые LPAR и соответствующие фреймы см. на рис. 10.5.

В качестве общего хранилища мы использовали ESS с восьмью LUN на 10 Гб, четыре из которых назначены рабочим группам ресурсов. Для тестирования общее хранилище не было важным; оно предназначалось только для настройки более полной конфигурации кластера с использованием мониторинга пульса через диски.

Для этих LPAR сконфигурированы параметры разделов, представленные в табл. 10.5.

Таблица 10.5. Профильные параметры разделов

Имя LPAR	Минимальное значение	Желаемое значение	Максимальное значение
Jordan	1 процессор – 1 Гб	1 процессор – 1 Гб	4 процессора – 4 Гб
Jessica	1 процессор – 1 Гб	1 процессор – 1 Гб	2 процессора – 2 Гб
Alexis	1 процессор – 1 Гб	1 процессор – 1 Гб	6 процессоров – 6 Гб

У нас сконфигурировано две группы ресурсов (app1_rg и app2_rg), каждая из которых содержит свои серверы приложений – app1 и app2 соответственно. Каждая группа ресурсов конфигурируется как подключающаяся на домашнем узле (online on home node). App1_rg содержит участвующие узлы Jordan и Alexis. App2_rg содержит участвующие узлы Jessica и Alexis. В итоге получается кластер с конфигурацией 2+1, где узел Alexis является дежурным (standby node).

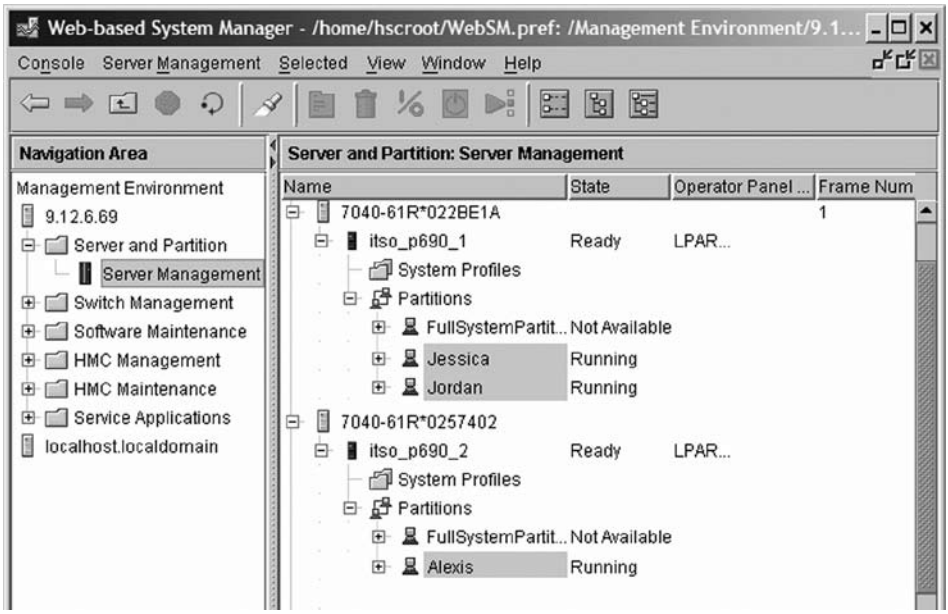

Рис. 10.5. Тестовые LPAR

Для каждого узла связи с HMC были настроены таким образом, чтобы включать обе консоли HMC с IP-адресами 192.168.100.69 и 192.168.100.5.

Параметры конфигурации DLPAR для серверов приложений представлены в табл. 10.6.

Таблица 10.6 Параметры конфигурации DLPAR для серверов приложений

Сервер приложения	Минимальное значение	Желаемое значение
app1	0	3
app2	0	2

Мы специально установили минимальные, нулевые, значения, чтобы всегда можно было получить группу ресурсов.

10.1.6 Результаты тестирования

Сценарий 1: получение группы ресурсов

В этом сценарии мы начинаем со следующей конфигурации:

- на узле Jordan выделен 1 процессор/1 Гб памяти;
- свободный пул содержит 7 процессоров/7 Гб памяти.

При запуске служб кластера на узле Jordan app1 запускается локально и пытается получить оптимальное количество ресурсов. Так как свободный пул содержит достаточно ресурсов, он получает еще 3 процессора и 3 Гб памяти, как показано на рис. 10.6.

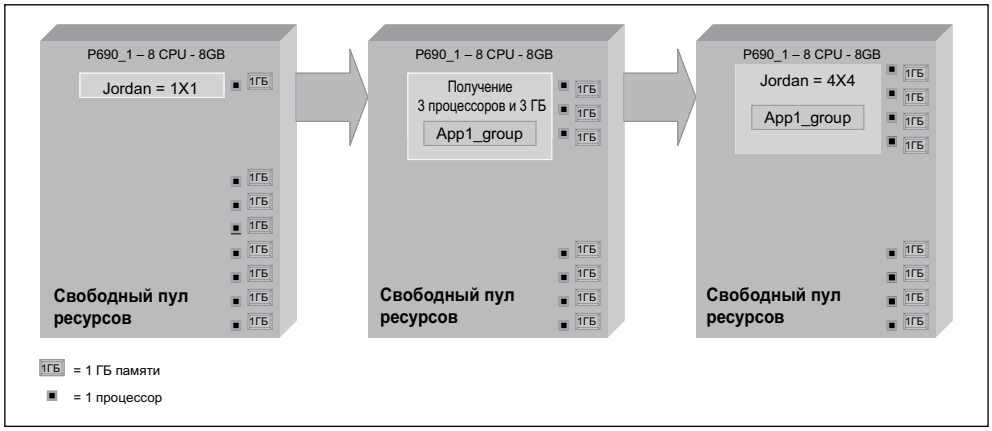

Рис. 10.6. Получение ресурсов DLPAR

Сценарий 2: освобождение группы ресурсов

В этом сценарии узел Jessica подключен в кластере с запущенным сервером приложения app2 при максимальных значениях: 2 процессора и 2 Гб.

При остановке служб кластера на узле Jessica сервер приложения app2 останавливается локально и освобождает ресурсы обратно в свободный пул. При этом HACMP не освобождает больше ресурсов, чем изначально было получено.

На рис. 10.7 представлено освобождение ресурсов и их перемещение обратно в свободный пул.

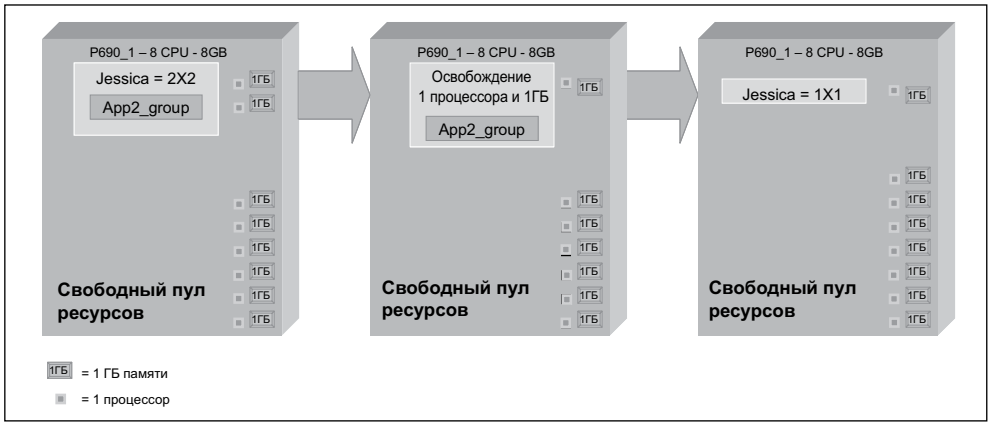

Рис. 10.7. Освобождение ресурсов DLPAR

Сценарий 3: последовательное перемещение при сбое для каждого LPAR

Этот сценарий состоит из двух частей, при котором выполняется перемещение при сбое для каждого раздела: сначала для узла Jordan, затем для узла Jessica. Это демонс-

трирует, что во время перемещения при сбое получение ресурсов выполняется подобно получению локальной группы ресурсов.

Также мы покажем различия между этим сценарием и сценарием 4, «Перемещение при сбое для рабочих LPAR в обратном порядке», в общем количестве ресурсов на дежурном узле.

В первой части инициируется отказ узла Jordan командой **reboot -q**. При этом происходит перемещение при сбое на узле Alexis. Alexis получает группу ресурсов app1 и выделяет желаемое количество ресурсов, как показано на рис. 10.8.

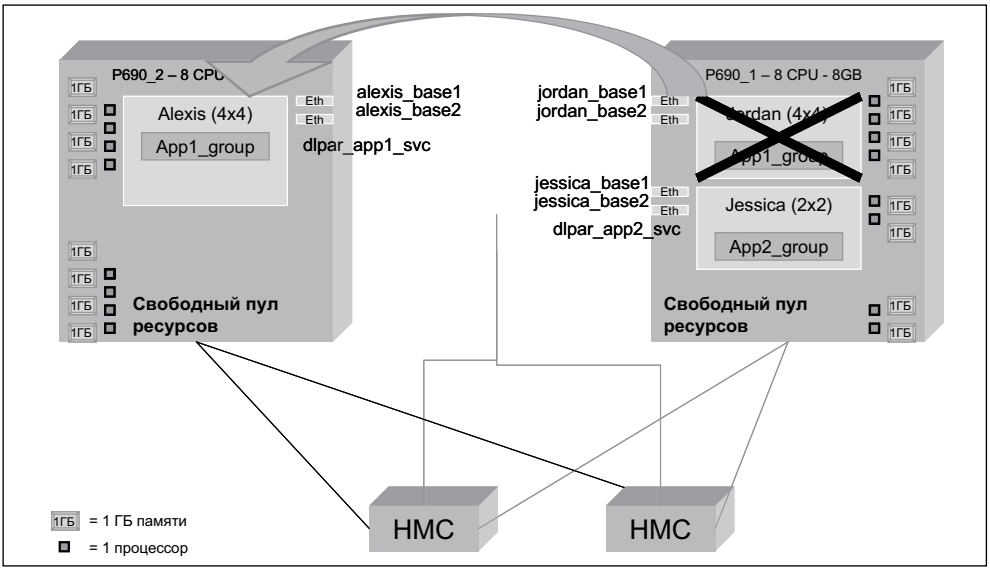

Рис. 10.8. Перемещение при сбое для первого рабочего LPAR

Узел Alexis теперь имеет такое же количество ресурсов, как и первый отказавший узел.

Вторая часть этого сценария начинается со следующей конфигурации:

- узел Jordan отключен;
- узел Jessica имеет 2 процессора и 2 Гб памяти;
- узел Alexis имеет 4 процессора и 4 Гб памяти;
- свободный пул (фрейм 2) содержит 4 процессора и 4 Гб памяти.

Теперь инициируется отказ узла Jessica командой **reboot -q**. Узел Alexis перехватывает группу ресурсов app2 и получает желаемое количество ресурсов, как показано на рис. 10.9.

В конечном итоге Alexis остается с максимальными параметрами раздела: 6 процессоров и 6 Гб памяти.

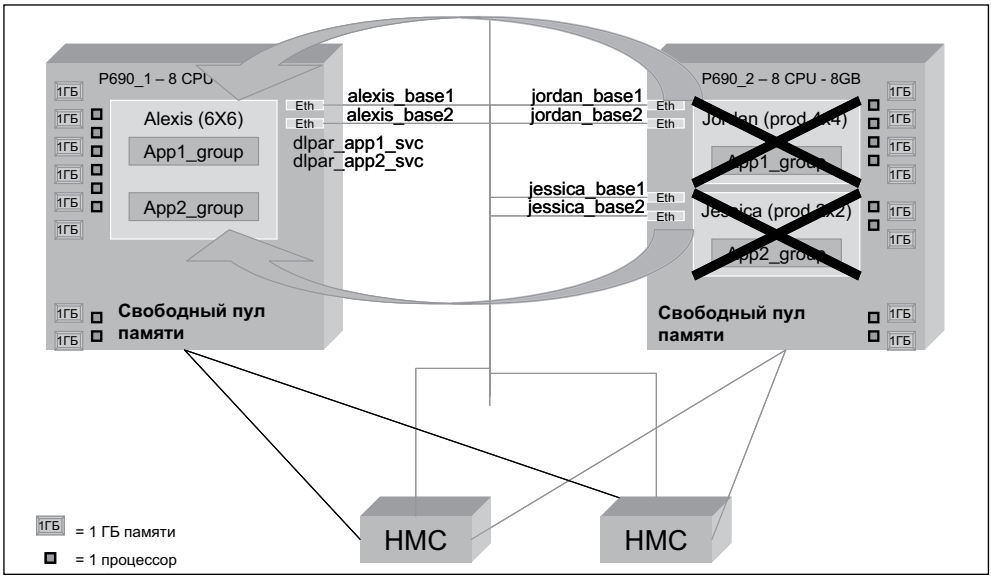

Рис. 10.9. Перемещение при сбое для второго рабочего LPAR

Сценарий 4: перемещение при сбое для рабочих LPAR в обратном порядке

Этот сценарий также состоит из двух частей. В начале мы имеем точно такую же конфигурацию, как и в сценарии 3 изначально службы кластера запущены на всех узлах и для каждого из них назначены следующие ресурсы:

- Jordan (фрейм 1) имеет 4 процессоров и 4 Гб памяти;
- Jessica (фрейм 1) имеет 2 процессора и 4 Гб памяти;
- Alexis (фрейм 2) имеет 1 процессор и 1 Гб памяти;
- свободный пул (фрейм 2) содержит 7 процессоров и 7 Гб памяти.

На этот раз мы сначала инициируем отказ узла Jessica своим предпочтительным методом или командой **reboot -q**. Это приводит к тому, что узел Alexis получает группу ресурсов app2 и оптимальное количество ресурсов, как показано на рис. 10.10.

Здесь необходимо отметить, что теперь Alexis имеет 3 процессора и 3 Гб памяти. Обычно группа ресурсов app2 имеет только 2 процессора и 2 Гб памяти на узле Jessica. Технически это может превышать необходимое количество ресурсов. Как видим, при предоставлении доступа к приложениям в конце может использоваться другое количество ресурсов, в зависимости от того, на каком LPAR/профиле раздела группа ресурсов оказывается в конечном итоге.

Вторая часть сценария начинается со следующей конфигурации:

- Jordan (фрейм 1) имеет 4 процессора и 4 Гб памяти;
- узел Jessica отключен;

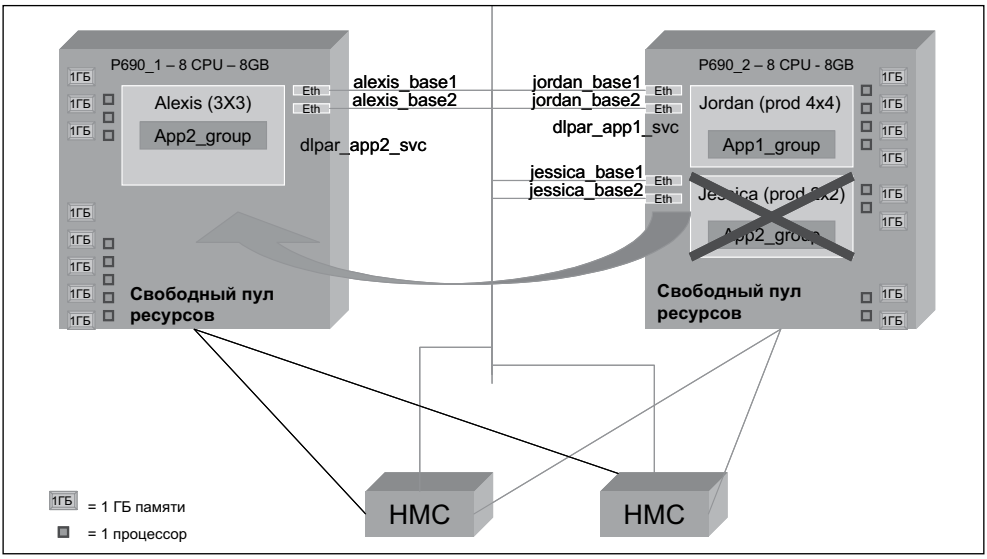

Рис. 10.10. Перемещение при сбое для второго рабочего LPAR сначала

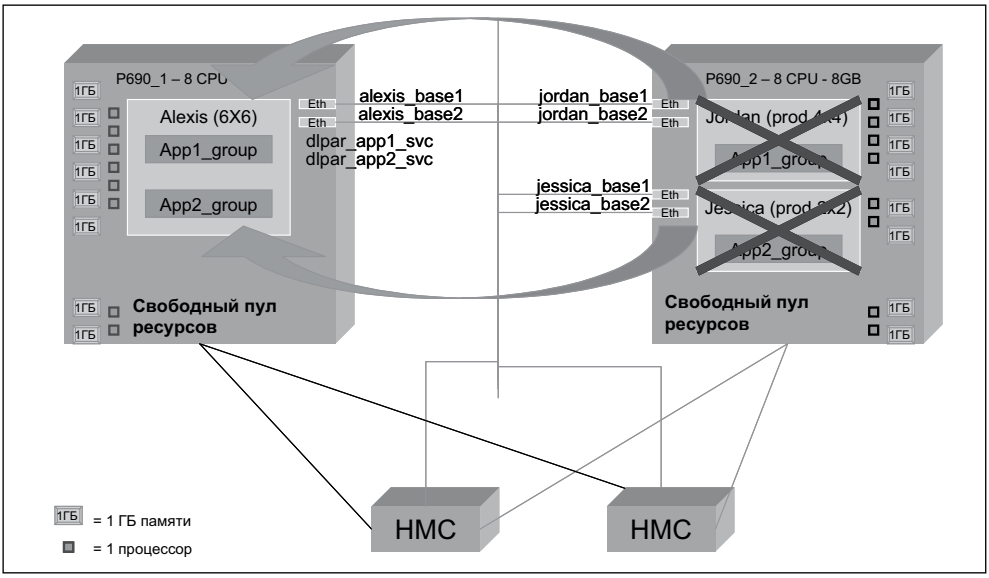

Рис. 10.11. Результаты второго перемещения при сбое

- Alexis (фрейм 2) имеет 3 процессора и 3 Гб памяти;
- свободный пул (фрейм 2) содержит 5 процессоров и 5 Гб памяти.

Теперь мы инициируем отказ узла Jordan командой **reboot -q.** Узел Alexis перехватывает группу ресурсов app1 и получает желаемые ресурсы, как показано на рис. 10.11. В итоге получаем то же, что и в предыдущем сценарии.

Сценарий 5: повторное получение группы ресурсов через rg_move

В этом сценарии мы начинаем с конфигурации, полученной в результате выполнения сценария 4 после перезапуска узлов Jordan и Jessica в кластере. Стартовая конфигурация имеет следующий вид:

- Jordan (фрейм 1) имеет 1 процессор и 1 Гб памяти;
- Jessica (фрейм 1) имеет 1 процессор и 1 Гб памяти;
- Alexis (фрейм 2) имеет 6 процессоров и 6 Гб памяти;
- свободный пул (фрейм 1) содержит 6 процессоров и 6 Гб памяти.

Узел Alexis на данный момент содержит две группы ресурсов. Мы выполняем rg_move для перемещения группы ресурсов app2_rg с узла Alexis обратно на домашний узел Jessica. Узел Alexis освобождает 2 процессора и 2 Гб памяти, тогда как узел Jessica получает **только** 1 процессор и 1 Гб памяти, как показано на рис. 10.12. Это опять же является прямым следствием сочетания настроек обеспечения приложений ресурсами и параметров LPAR.

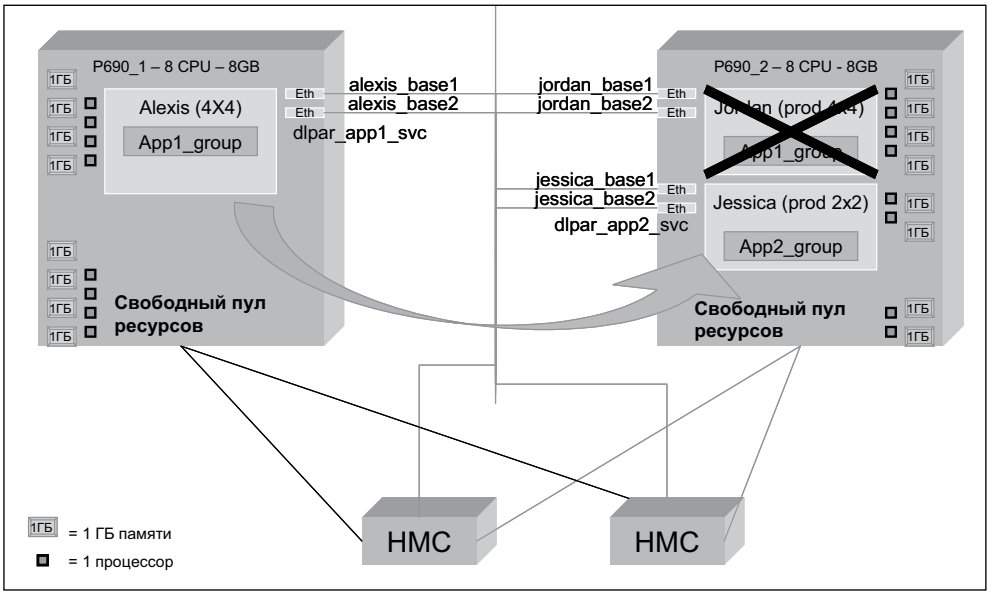

Рис. 10.12. Освобождение группы ресурсов и ее получение путем после rg_move

Сценарий 6: тестирование избыточности HMC

В этом сценарии мы тестируем избыточность HMC путем физического отключения сетевого подключения на одной из консолей HMC. В самом начале службы кластера выполняются на всех узлах и имеют следующие выделенные ресурсы:

- Jordan (фрейм 1) имеет 4 процессора и 4 Гб памяти;
- Jessica (фрейм 1) имеет 2 процессора и 4 Гб памяти;
- Alexis (фрейм 2) имеет 1 процессор и 1 Гб памяти;

- свободный пул (фрейм 2) содержит 7 процессоров и 7 Гб памяти.

Мы физически отключили кабель Ethernet из первой указанной консоли HMC с адресом 192.168.100.69. После этого мы инициировали отказ узла Jordan, чтобы вызвать перемещение при сбое. Это показано на рис. 10.13.

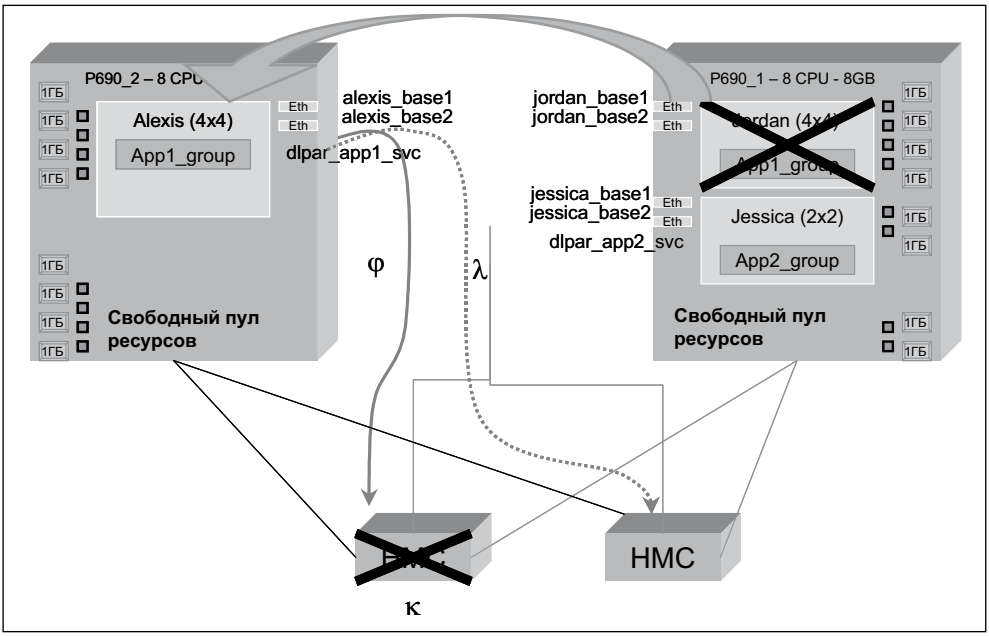

Рис. 10.13. Тестирование избыточности HMC

В процессе перемещения при сбое и попыток получения доступа к HMC происходит следующее:

1. HACMP выдает ping-запрос на первую консоль HMC.

2. Первая консоль HMC отключена и не реагирует.

3. HACMP выдает ping-запрос на вторую консоль HMC, которая работает нормально и продолжает обрабатывать операции командной строки DLPAR.

Операции тестирования HMC можно просмотреть в файле /tmp/hacmp.out с использованием утилиты **clhmcexec**.

10.2 HACMP и виртуализация

Во время написания данной книги было сделано официальное сообщение о поддержке виртуализации в HACMP. Подробности поддержки мы рассмотрим в последующих разделах, тогда как текст сообщения можно найти по следующему адресу:

http://w3-1.ibm.com/sales/systems/portal/_s.155/254?navID=f220s240&geoID= All&prodID=IBM%20eServer%20And%20TotalStorage%20Products&docID=hacmpvio063005

10.2.1 Требования

Для использования встроенных функций виртуализации и/или CUoD в HACMP в системах Power5 на всех узлах LPAR в кластере должны быть установлены как минимум следующие версии программного обеспечения:

- AIX 5.3 Maintenance Level 5300-002 с APAR IY70082 и eFIX IY72974;
- HACMP:
  - 5.1 с APAR IY66556 (или выше);
  - 5.2 с APAR IY68370 (или выше) и APAR IY68387;
  - 5.3.0.1 с APAR IY73051 RSCT;
  - rsct.basic.hacmp.2.4.2.1;
  - rsct.basic.rte.2.4.2.2;
  - rsct.compat.basic.hacmp.2.4.2.0.
- OpenSSH 3.4p1

  Программное обеспечение OpenSSH можно получить из следующих источников:
- AIX 5.3 Expansion pack, Linux Toolbox CD – или скопировать с сайта

  http://sourceforge.net/projects/openssh-aix

  OpenSSh для AIX имеет собственные требования; см. раздел 10.1.1, «Требования».
- Virtual I/O Server Version 1.1.2 с VIOS fixpack 6.2 и eFIX IY71303.062905.epkg.Z

  Fixpack 6.2 доступен по адресу

  http://techsupport.services.ibm.com/server/vios/download/home.html

  eFIX IY72974 доступен по адресу ftp.software.ibm.com/.../efixes
- Криптографические библиотеки OpenSSL:

  http://www-1.ibm.com/servers/aix/products/aixos/linux/download.html

Для правильного управления и обеспечения функций DLPAR необходимо подключение HMC к LPAR. В консоли HMC также должны быть установлены как минимум следующие версии программного обеспечения:

HMC Version 4 Release 5 Build 20050519.1 или выше.

> **Важно.** APAR IY73051 для HACMP V5.3 необходим для поддержки микроразделов, функций CUoD и CBU на системах Power5.

10.2.2 Обеспечение приложений ресурсами

Все аспекты, рассмотренные в разделе «Обеспечение приложений ресурсами», относятся и к конфигурированию микроразделов (micropartitioning).

HACMP может осуществлять управление использованием физических процессоров в качестве выделенных процессоров, а также применением виртуальных процессоров в конфигурации общих процессоров. Ниже представлены некоторые возможные сценарии.

При использовании микроразделов HACMP работает с виртуальными процессорами, а не с физическими процессорами. В смешанной среде HACMP определяет возможность добавления ресурсов, учитывая максимальное значение. В режиме выделенных процессоров максимальной единицей измерения является физический процессор. В режиме общих процессоров максимальной единицей измерения является виртуальный процессор (рис. 10.14). Ресурсы из свободного пула определяются путем суммирования значений оптимальной выделенной мощности (capacity entitlement, CE) для каждого раздела в режиме общих процессоров и физического процессора в режиме выделенных процессоров.

Вы не сможете использовать выделенные мощности (с точностью до 1/10 мощности процессора) в качестве значения для определения обеспечения приложений ресурсами в меню HACMP.

HACMP не проверяет, является ли раздел capped или uncapped. В зависимости от конфигурации необходимо проверить возможные варианты самостоятельно.

Максимальным значением для одного виртуального процессора является один физический процессор (рис. 10.14). Таким образом, количество виртуальных процессоров, назначенных для раздела, определяют максимум вычислительной мощности для этого раздела.

Поэтому можно достичь хороших результатов с использованием виртуальных процессоров с HACMP, вместо того чтобы настраивать количество виртуальных процессоров в разделе на случай наихудшего развития событий (когда множество приложений выполняет перемещение при сбое на один узел). В этом случае количество

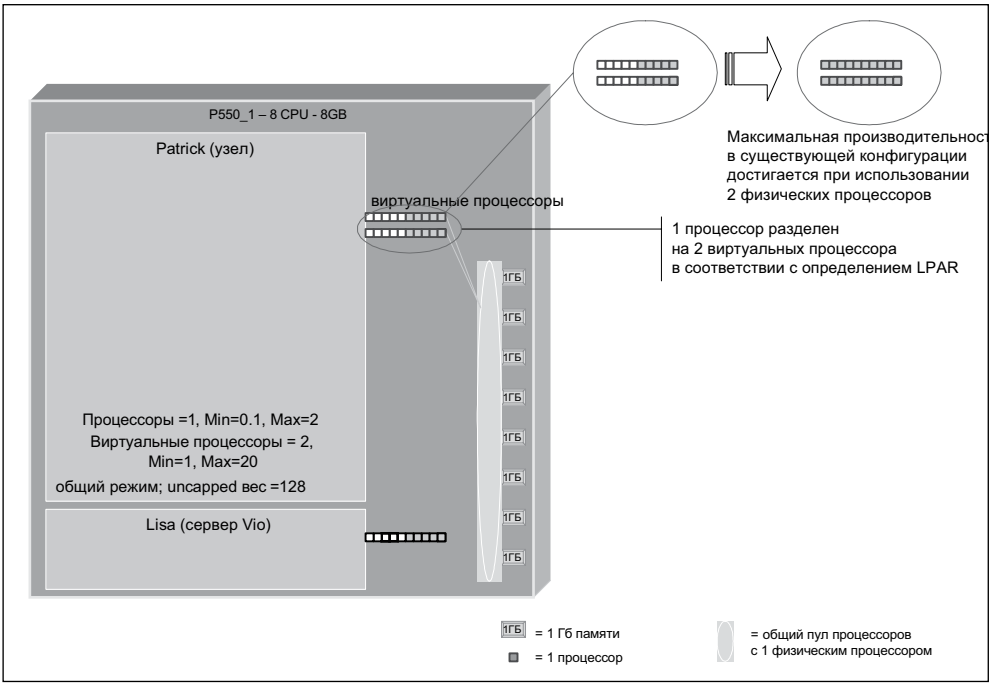

Рис. 10.14. Назначение выделенной мощности разделу

виртуальных процессоров не оптимизировано (с большим использованием вычислительной мощности в гипервизоре).

При работе в capped-среде следует заранее прогнозировать количество виртуальных процессоров в разделе для обеспечения требуемой производительности для своего приложения. В действительности, если установить выделенную мощность в один процессор и создать четыре виртуальных процессора, всегда будет назначена выделенная мощность в один процессор, что соответствует одному физическому процессору, а не четырем. В uncapped среде такая проблема не возникает (рис. 10.15). Единственный способ ограничить раздел состоит в ограничении количества виртуальных процессоров и веса.

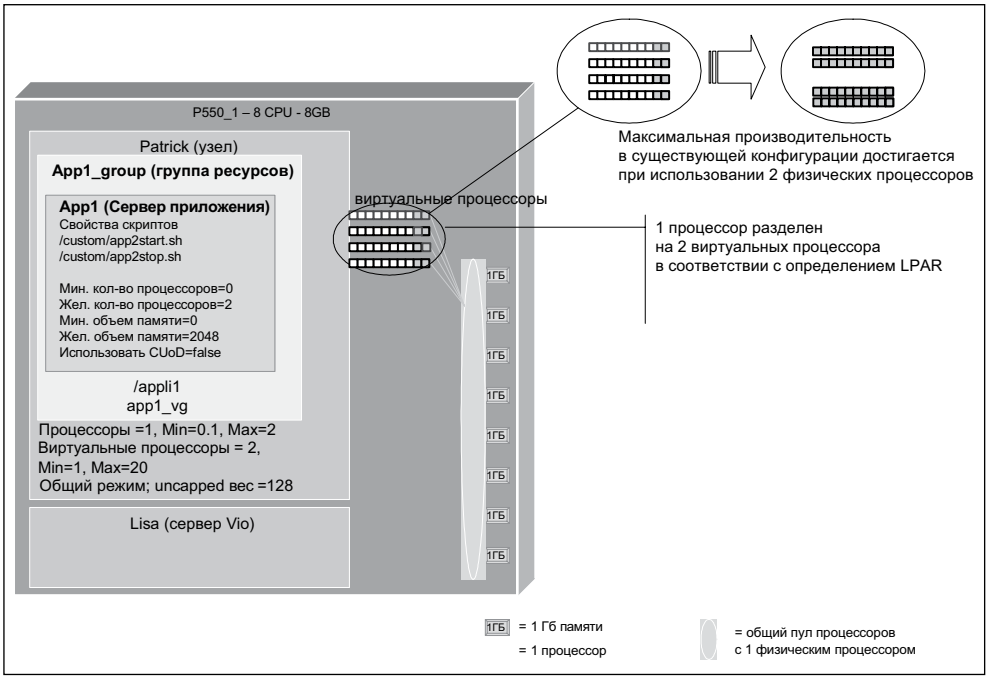

Рис. 10.15. Активизация группы ресурсов

Таблица 10.7. Значения параметров LPAR

Имя LPAR	Значения LPAR, мин/жел/макс	Количество виртуальных процессоров, мин/жел/макс	Значения HACMP	Процессоры	Режим процессоров
patrick	0.1 / 1.0 / 2.0	1 / 2 / 20	app1 / 0 / 2		Ообщий
lisa	0.1 / 0.5 / 2.0	1 / 2 / 20	Неприменимо		Общий
maelle	1.0 / 1.0 / 1.0	Неприменимо	Неприменимо	1	Выделенный
shawn	0.1 / 0.3 / 4.0	1 / 3 / 40	app2 / 0 / 1		Общий
lee	0.1 / 0.5 / 2.0	1 / 2 / 20			Общий
smokey	0.1 / 0.5 / 1.0	1 / 1 / 10			Общий

Ниже приведен пример микроразделов, позволяющий понять, как все работает. Мы используем только информацию о процессорах. В этом примере минимальные значения для предоставления доступа к приложениям HACMP равны нулю. Рекомендуется, чтобы это значение оставалось равным нулю, чтобы можно было запустить сервер приложений даже при отсутствии ресурсов.

Табл. 10.8 содержит действительные активные значения.

Таблица 10.8. Значения, используемые для определения

Имя LPAR	Значение LPAR, (1)	Количество виртуальных процессоров, мин/жел/макс	Значения HACMP	Процессоры
patrick	1	2	app1 / 0 / 2	
lisa	0.5	2	Неприменимо	
maelle	1.0	Неприменимо	Неприменимо	1
shawn	0.3	3	app2/ 0 / 1	
lee	0.5			
smokey	0.5	1		

Примечание. Здесь представлен образец вычислений при использовании версии программного обеспечения, доступной на момент написания этого материала. Проверьте PTF1, прежде чем выполнять реализацию конфигурации такого типа в рабочей среде.

Чтобы определить количество свободных ресурсов, следует отнять 2,8 [сумму столбца (1) для одного компьютера] от общего количества процессоров на компьютере.

Свободные ресурсы на компьютере patrick: 4 – 2,8 = 1,2.

Свободные ресурсы на компьютере shawn: 4 – 1,3 = 2,7.

Вычисления с одним ресурсом на узле Patrick:

1 VP (минимум для LPAR) + 2 VP (желательное количество) = 3 VP, активных при запущенном app2.

Вычисления по добавлению еще одного приложения app1:

1 VP (желательное количество) + 3 VP (активные на данный момент) – 1 VP (минимум для LPAR) – 2 VP = 1 VP (нужно добавить). Если это значение меньше количества свободных ресурсов (1,2), то HACMP добавляет его.

Внимание! Если требуется, чтобы приложение запускалось каждый раз, даже при недостаточном количестве ресурсов на целевом компьютере, нужно в качестве значения минимального количества ресурсов (процессоров и памяти) указать 0 в меню определения DLPAR для сервера приложения.
Это относится к DLPAR с выделенным режимом процессоров.

Код HACMP допускает настройку оптимального размера LMB. Размер LMB берется из HMC. Однако меню smit позволяет добавлять только по 256 Мб.

Количество свободных ресурсов определяется путем суммирования количества выделенных процессоров и выделенной мощности, заданной как желательная для каждого конфигурируемого LPAR.

10.3 Сценарии конфигурирования HACMP и виртуализации

Первый сценарий предназначен для того, чтобы помочь вам в базовом конфигурировании HACMP в виртуальной среде.

Второй сценарий предназначен для того, чтобы показать функциональные возможности виртуализации. Они дополняют механизм обеспечения высокой доступности, реализованный в HACMP.

- Сценарий 1 – конфигурация с двумя узлами HACMP со взаимным перехватом (mutual takeover). Иллюстрирует базовое конфигурирование HACMP.
- Сценарий 2 – конфигурация с двумя кластерами HACMP, каждый из которых содержит два узла.

10.3.1 Сценарий 1

Обзор HACMP и виртуальных компонентов

HACMP можно использовать как обычно в виртуальной среде. Конечно же, нужно учитывать вышеперечисленные ограничения. Схема на рис. 10.16 иллюстрирует компоненты так, как они представлены в HACMP. В целях избыточности мы определили два виртуальных сервера ввода-вывода (Virtual I/O server) на компьютер. Раздел клиента AIX указывает через виртуальный адаптер на раздел VIOS. При использовании Virtual SCSI для доступа к дискам (vscsi0, vscsi1, ...) автоматически применяется MPIO (Multi-path I/O). Поэтому существует два пути доступа к общим дискам. Команда

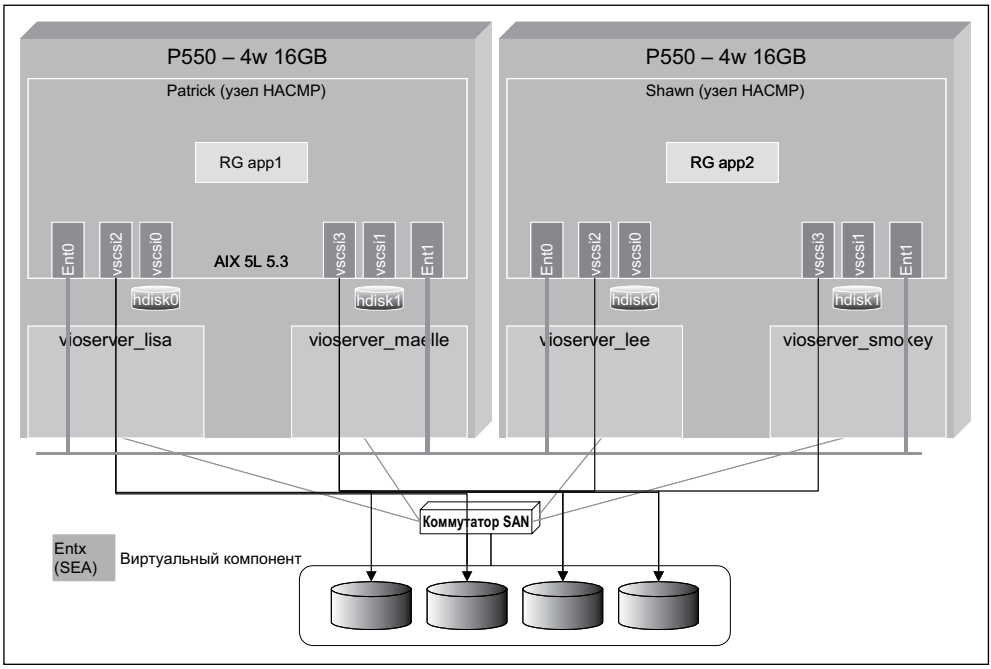

Рис. 10.16. Логическая схема HACMP

```
patrick / > lspath
Enabled hdisk1 vscsi1
Enabled hdisk3 vscsi2
Enabled hdisk4 vscsi2
Enabled hdisk5 vscsi2
Enabled hdisk6 vscsi2
Enabled hdisk6 vscsi3
Enabled hdisk5 vscsi3
Enabled hdisk4 vscsi3
Enabled hdisk3 vscsi3
Enabled hdisk2 vscsi0
Enabled hdisk0 vscsi0
```

Рис. 10.17. Команда lspath

lspath выводит путь, управляемый MPIO (рис. 10.17). Виртуальный адаптер не выдает ошибок оборудования в AIX. Однако HACMP использует RSCT для передачи информации о состоянии всех компонентов, участвующих в обмене данными.

Пример 10.10 показывает конфигурацию HACMP в нашем тестовом кластере.

Пример 10.10. Использование утилиты cldump

```
patrick / > cldump

Cluster Name: app2_cluster
Cluster State: UP
Cluster Substate: UNSTABLE

Node Name: patrick              State: UP

  Network Name: net_diskhb_01      State: UP

    Address:                 Label: patrick2shawn      State: UP

  Network Name: net_ether_01       State: UP

    Address: 10.10.5.2       Label: patrick_base1      State: UP
    Address: 10.10.6.2       Label: patrick_base2      State: UP
    Address: 192.168.101.143 Label: vio_svc2           State: UP

Node Name: shawn                State: UP

  Network Name: net_diskhb_01      State: UP

    Address:                 Label: shawn2patrick      State: UP

  Network Name: net_ether_01       State: UP

    Address: 10.10.5.1       Label: shawn_base1        State: UP
    Address: 10.10.6.1       Label: shawn_base2        State: UP
```

```
    Address: 192.168.101.142 Label: vio_svc1          State: UP

Cluster Name: app2_cluster

Resource Group Name: app2_group
Startup Policy: Online On Home Node Only
Fallover Policy: Fallover To Next Priority Node In The List
Fallback Policy: Fallback To Higher Priority Node In The List
Site Policy: ignore
Priority Override Information:
    Primary Instance POL:
Node                          Group State
----------------------------- ---------------
shawn                         ONLINE
patrick                       OFFLINE

Resource Group Name: app1_group
Startup Policy: Online On Home Node Only
Fallover Policy: Fallover To Next Priority Node In The List
Fallback Policy: Fallback To Higher Priority Node In The List
Site Policy: ignore
Priority Override Information:
    Primary Instance POL:
Node                          Group State
----------------------------- ---------------
patrick                       ONLINE
shawn                         OFFLINE
```

Необходимо сконфигурировать виртуальный сервер ввода-вывода (Virtual I/O, VIO), который бы предлагал виртуальные компоненты для использования в разделе.

На следующей схеме (рис. 10.18) представлены подробные сведения об используемых аппаратных компонентах. Условные обозначения позволяют отличить виртуальные компоненты от физических компонентов. Обратите внимание на то, что в нашем тесте работа AIX и HACMP основана только на виртуальных компонентах. На каждом узле используется четыре виртуальных SCSI, два из которых предназначены для зеркального отображения системного диска, а еще два для общих дисков. Весь доступ распределен по двум VIO-серверам.

Общие диски для обоих клиентских разделов должны быть определены как диски мониторинга пульса (hdisk) в целевом определении на VIO-сервере.

Настройка виртуализации

Во-первых, как сказано в книге **Advanced POWER Virtualization on IBM @server® p5 Servers: Introduction and Basic Configuration**, SG24-7940, необходимо планировать свои операции. Ниже перечислены операции, которые мы выполнили для реализации своей конфигурации виртуализации.

1. **Планирование операций (рис. 10.19).**

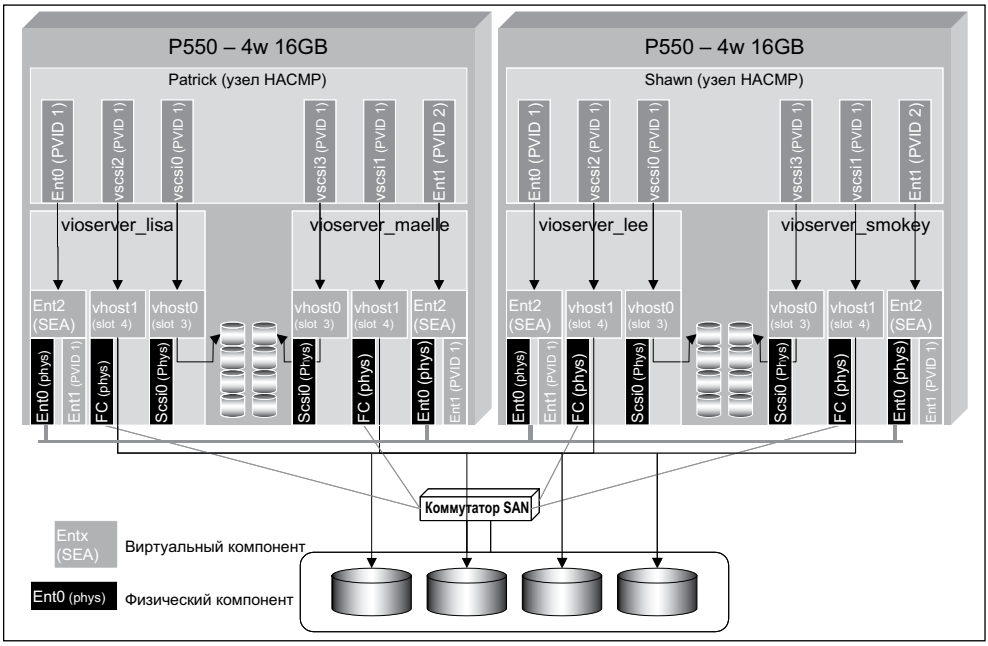

Рис. 10.18. Схема архитектуры виртуализации

Machine	Client Name	type SCSI or Ethernet	intrface id	Slot		remote partition	Remote Partition Vslot/PVID	VIOS name	type SCSI or Ethernet	interface id	Slot number	remote partition	remote slot / PVID
p550-1	**patrick**	ent	0	**2**	10.10.5.2		1	**lisa**	ent	4	6		**1**
p550-1	**patrick**	ent	1	**4**	10.10.6.2		2	**maelle**	ent		2		**2**
p550-1	**patrick**	scsi		**7**		maelle	7	**maelle**	scsi		7	patrick	**7**
p550-1	**patrick**	scsi		**6**		lisa	4	**lisa**	scsi		4	patrick	**6**
p550-1	**patrick**	scsi		**5**		maelle	5	**maelle**	scsi		5	patrick	**5**
p550-1	**patrick**	scsi		**3**		lisa	3	**lisa**	scsi		3	patrick	**3**
p550-2	**shawn**	ent	1	**4**	10.10.6.1		2	**smokey**	ent		2		**2**
p550-2	**shawn**	ent	0	**2**	10.10.5.1		1	**smokey**	ent		3		**1**
p550-2	**shawn**	scsi		**7**		smokey	7	**smokey**	scsi		7	shawn	**7**
p550-2	**shawn**	scsi		**5**		smokey	5	**smokey**	scsi		5	shawn	**5**
p550-2	**shawn**	scsi		**6**		lee	6	**lee**	scsi		6	shawn	**6**
p550-2	**shawn**	scsi		**3**		lee	3	**lee**	scsi		3	shawn	**3**

Рис. 10.19. Инструмент планирования операций виртуализации в среде Excel

2. **Создание конфигурации LPAR в консоли HMC.** Каждый раздел VIOS определяется с одним адаптером Virtual Ethernet и с двумя адаптерами Virtual SCSI. Адаптер Virtual Ethernet определяется как транковый (trunk) адаптер (в дальнейшем – ent1). Оба адаптера Virtual SCSI определены как серверы (в дальнейшем – vhost0 и vhost1).

3. **Установка раздела VIOS с последующим его обновлением командой updateios**.

4. **Установка общего адаптера Ethernet (Shared Ethernet Adapter, SEA). (Пример 10.11).**

Пример 10.11. Установка Shared Ethernet Adapter

```
mkvdev -sea ent0 -vadapter ent1 default ent1 -defaultid 1

mktcpip -hostname vioserver_lisa -inetaddr 192.168.100.220 -netmask
255.255.255.0 -gateway 192.168.100.60 -start
```

5. **Создание виртуальных дисков.** В нашем кластере мы создаем два типа дисков для клиента. На узле shawn мы создаем оба целевых диска в разделе VIOS как логический том (см. пример 10.2). На узле patrick мы создаем оба целевых диска в разделе VIOS как диск hdisk (см. пример 10.12).

Пример 10.12. Создание виртуальных дисковых устройств

```
mkvg -f -vg shawnvg hdisk1
mklv -lv shawn_lv shawnvg 30G
mkvdev -vdev shawn_lv -vadapter vhost0 -dev vshawn_disk
mkvdev -vdev hdisk1 -vadapter vhost0 -dev vpatrick_disk
```

6. **Установка системы и HACMP в клиентском разделе.**

7. **Зеркальное отображение системных дисков и изменение списка загрузочных устройств.**

8. **Если вы хотите использовать функцию DLPAR или CUoD, следует установить и сконфигурировать SSH, как описано в разделе «Установка и конфигурирование SSH на узлах HACMP».**

Результаты тестирования

Мы выполнили несколько тестов и получили следующие результаты:

Зеркальное отображение rootvg

При завершении работы VIOS пропадает доступ к части дисков. Зеркальное отображение позволяет продолжить работу. На имеющемся уровне AIX для восстановления доступности дисков необходимо убрать диск из определения rootvg, после чего добавить такой же диск и выполнить зеркальное отображение. Затем необходимо выполнить команду **bosboot**.

Замена адаптера

HACMP использует RSCT для слежения за состоянием коммуникационных интерфейсов или устройств. На рис. 10.20 показано выявление отказа на виртуальном адаптере. В этом тесте мы отключаем кабель Ethernet для имитации отказа физической сети.

В этом примере до 18:33 сетевое подключение было доступно. В 18:33:30 мы отключили кабель Ethernet. В 18:33:59 службы топологии определяют отсутствие пульса, указывающее на отключение адаптера en2. В 18:34:02 HACMP выполняет операцию **swap_adapter**.

```
07/05 18:32:45.542: Received a SEND MSG command. Dst: 10.10.5.2.
07/05 18:33:05.542: Received a SEND MSG command. Dst: 10.10.5.2.
07/05 18:33:25.548: Received a SEND MSG command. Dst: 10.10.5.2.
07/05 18:33:40.789: Heartbeat was NOT received. Missed HBs: 1. Limit: 10
07/05 18:33:42.789: Heartbeat was NOT received. Missed HBs: 2. Limit: 10
07/05 18:33:42.789: Starting sending ICMP ECHOs.
07/05 18:33:42.789: Invoking netmon to find status of local adapter.
07/05 18:33:44.789: Heartbeat was NOT received. Missed HBs: 3. Limit: 10
07/05 18:33:46.789: Heartbeat was NOT received. Missed HBs: 4. Limit: 10
07/05 18:33:47.497: netmon response: Adapter seems down
...
07/05 18:33:58.809: Heartbeat was NOT received. Missed HBs: 10. Limit: 10
07/05 18:33:58.809: Local adapter is down: issuing notification for local adapter
07/05 18:33:58.809: Adapter status successfully sent.
07/05 18:33:58.945: Dispatching netmon request while another in progress.
07/05 18:33:58.945: Received a SEND MSG command. Dst: 10.10.5.2.
07/05 18:33:59.497: netmon response: Adapter seems down
07/05 18:33:59.497: Adapter status successfully sent.
07/05 18:41:42.982: netmon response: Adapter is up
07/05 18:41:42.982: Bind to broadcast address succeeded.
07/05 18:41:42.982: Adapter status successfully sent.
07/05 18:41:45.849: Received a SEND MSG command. Dst: 10.10.5.2.
07/05 18:41:46.859: Received a SEND MSG command. Dst: 10.10.5.2.
07/05 18:41:46.859: Received a STOP MONITOR command.
07/05 18:41:46.861: Received a SEND MSG command. Dst: 10.10.5.2.
07/05 18:41:46.861: Received a SEND MSG command. Dst: 10.10.5.2.
07/05 18:41:46.861: Received a START HB command. Destination: 10.10.5.2.
07/05 18:41:46.861: Received a SEND MSG command. Dst: 10.10.5.2.
07/05 18:41:46.861: Received a START MONITOR command.
07/05 18:41:46.861: Address: 10.10.5.2 How often: 2000 msec Sensitivity: 10 Configuration Instance: 11
07/05 18:41:46.862: Received a SEND MSG command. Dst: 10.10.5.2.
07/05 18:41:46.862: Received a SEND MSG command. Dst: 10.10.5.2.
07/05 18:41:47.660: netmon response: Adapter is up
07/05 18:41:51.871: Received a SEND MSG command. Dst: 10.10.5.2.
07/05 18:41:57.871: Received a SEND MSG command. Dst: 10.10.5.2.
```

Рис. 10.20. nim.topsvcs.en2.app2_cluster

Отключение VIO-сервера

В этом тесте происходит отключение сервера vioserver_lisa. Все операции продолжают выполняться с использованием второго пути. MPIO выполняет свои функции, сохраняя доступ к данным через второй VIO-сервер, vioserver_maelle. Конечно же, rootvg всегда активен на диске с зеркальным отображением (см. рис. 10.21). Команда **lspath** показывает, что на одном пути произошел отказ доступа к общим дискам.

Как описывается в разделе «Зеркальное отображение rootvg», когда VIO-сервер включен, необходимо выполнять операцию вручную. Команда **lspath** показывает состояние каждого пути (рис. 10.13). Можно выполнить команду **chpath** или использовать быстрый путь **smit mpiopath_enable_all**; на рис. 10.21 представлено состояние после выполнения этой команды. Адаптер Virtual Ethernet подключается автоматически событием **join_interface**.

Пример 10.13 отображает информацию об отсутствующих путях для виртуальных дисков.

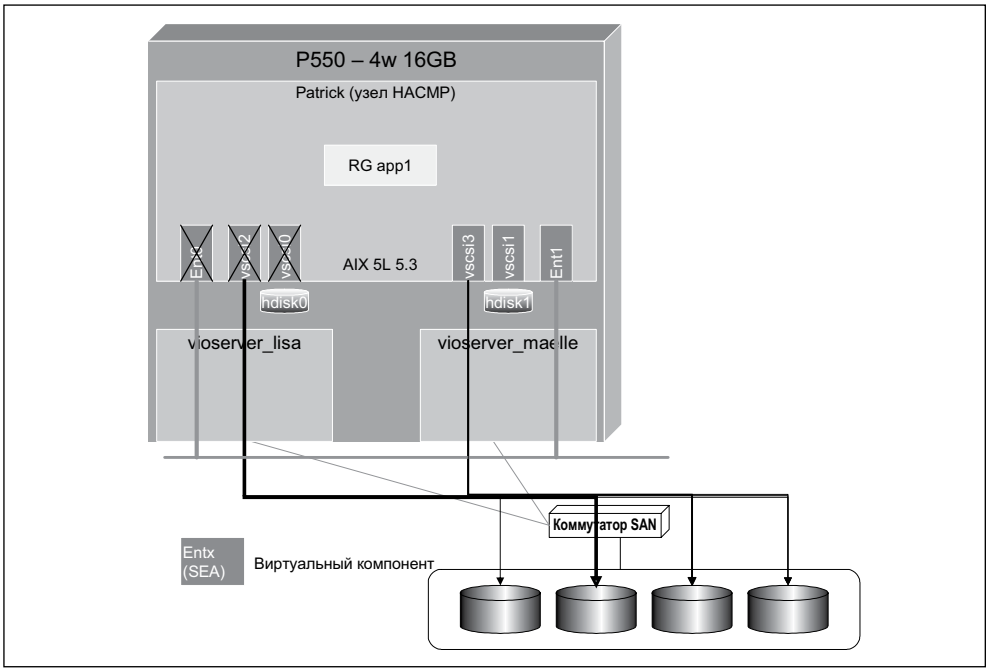

Рис. 10.21. Отказ VIO-сервера

Пример 10.13. Команда lspath – вывод отсутствующих путей

```
patrick / > lspath
Enabled hdisk1 vscsi1
Missing hdisk3 vscsi2
Missing hdisk4 vscsi2
Missing hdisk5 vscsi2
Missing hdisk6 vscsi2
Failed  hdisk6 vscsi3
Failed  hdisk5 vscsi3
Failed  hdisk4 vscsi3
Failed  hdisk3 vscsi3
Enabled hdisk2 vscsi0
Missing hdisk0 vscsi0
```

При отключении одного VIO-сервера команда lspath выдает выходные данные, подобные представленным в примере 10.14.

Пример 10.14. Пример выполнения команды lspath при отключении одного VIO-сервера

```
patrick / > lspath
Enabled hdisk1 vscsi1
Missing hdisk3 vscsi2
Missing hdisk4 vscsi2
Missing hdisk5 vscsi2
Missing hdisk6 vscsi2
Enabled hdisk6 vscsi3
Enabled hdisk5 vscsi3
Enabled hdisk4 vscsi3
Enabled hdisk3 vscsi3
Enabled hdisk2 vscsi0
Missing hdisk0 vscsi0
```

Перемещение при сбое для одного узла

В итоге ситуация после выполнения этих тестов одинакова:

- два VIO-сервера отключены;
- каждый Ethernet-адаптер отключен от обоих VIO-серверов;
- прекращение работы одного узла.

  HACMP завершил логическую обработку.

10.3.2 Аспекты производительности и архитектуры (сценарий 2)

Мы пойдем дальше и попытаемся добавить еще один кластер в ту же инфраструктуру. Этот сценарий включает два кластера HACMP. Мы определяем еще два сервера Virtual SCSI (рис. 10.22).

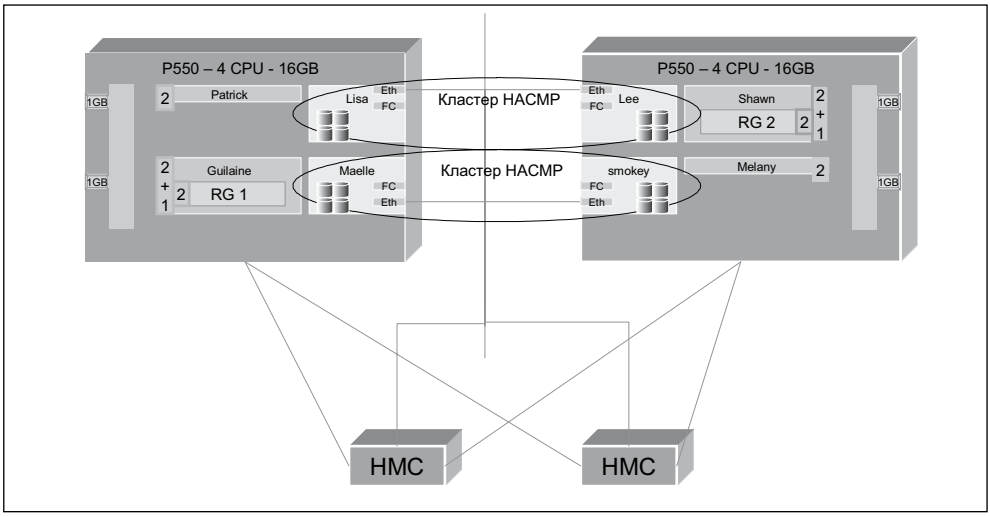

Рис. 10.22. Два кластера с двумя CEC и двумя HMC

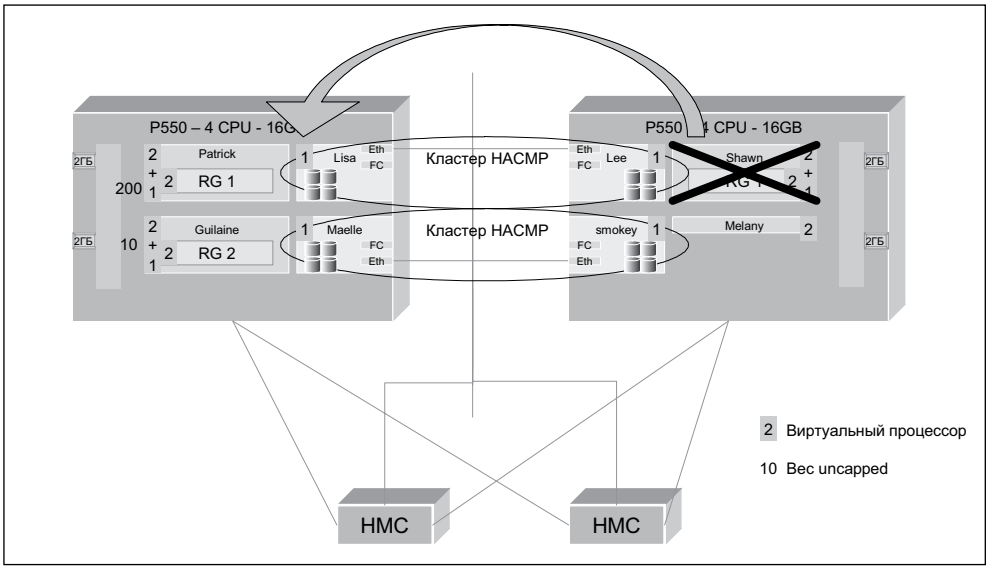

Рис. 10.23. Перемещение при сбое на одном компьютере с преимущественной обработкой RG1

Преимущество такой архитектуры состоит в том, что вы управляете двумя простыми кластерами с двумя компьютерами. В действительности принцип использования дежурного узла для защиты рабочей среды позволяет ограничить риск ее нарушения. Каждое приложение имеет собственный кластер. Администрирование упрощается, что позволяет улучшить управление средой. Конфигурирование виртуализации позволяет управлять обеспечением приложений ресурсами. Конфигурирование этого кластера с использованием HACMP config assist отнимает минимальное количество времени.

Группа ресурсов RG1 представляет рабочую группу ресурсов, требующую больше ресурсов процессора. Сконфигурировав значение веса uncapped на каждом разделе, можно осуществлять поддержку среды. HACMP активизирует ресурсы, связанные с сервером приложения (количество виртуальных процессоров). Механизм вычисления такой же, как и в разделе 10.1.2, «Обеспечение приложений ресурсами».

В этой конфигурации мы используем две функции: функцию обеспечения приложений ресурсами в HACMP и функцию микроразделов, входящую в APV (Advance Power Virtualization). Вместо остановки другого раздела в целях освобождения некоторых ресурсов мы используем функцию веса uncapped, чтобы сохранить приоритет за более важным разделом.

Параметры микроразделов (табл. 10.9) не зависят от параметров HACMP.

Ниже описывается тест, который можно выполнить. Инициируем перемещение при сбое с одного узла на другой. В то же время осуществляем сбор некоторых данных производительности в каждом разделе. Значение веса uncapped для LPAR patrick составляет 200, а для LPAR guilaine – 10; это означает, что раздел patrick имеет более высокий приоритет.

Вес uncapped представляет собой число в диапазоне от 0 до 255, устанавливаемое для каждого uncapped раздела в пуле общих процессоров; 255 – наивысший вес.

Таблица 10.9. Список параметров конфигурации кластера

Имя LPAR	CE, мин/жел/макс	HACMP DLPAR мин/жел	Вес uncapped	Виртуальные процессоры, мин/жел/макс
Patrick	0.5 / 2.0 / 4.0	app2 0 / 2	200	1 / 2 / 40
Guilaine	0.5 / 2.0 / 4.0	app1 0 / 2	10	1 / 2 / 40
Maelle	0.1 / 0.5 / 1	Неприменимо	128	1 / 2 / 10
Lisa	0.1 / 0.5 / 1	Неприменимо	128	1 / 2 / 10

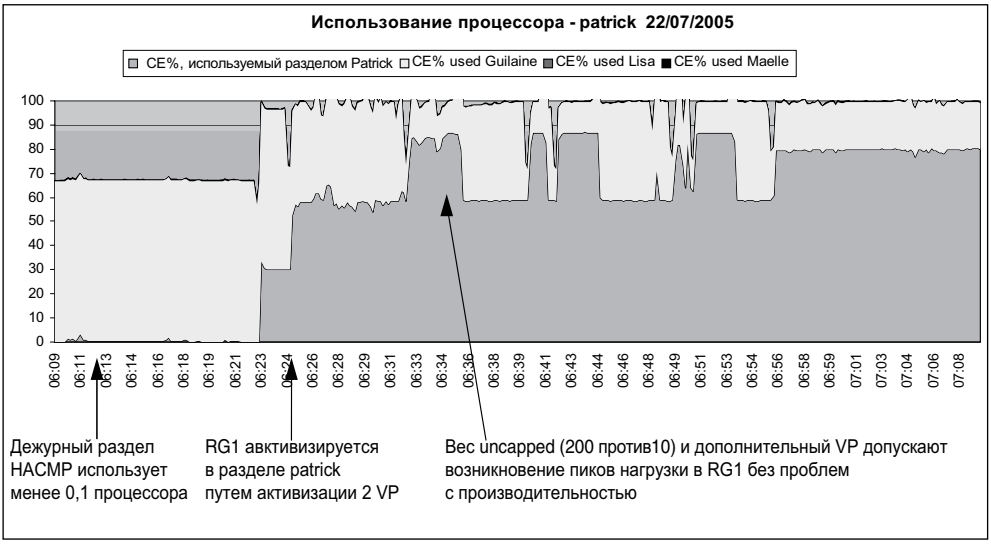

Рис. 10.24. График активизации процессоров в каждом разделе

Доступная неиспользуемая мощность распределяется между разделами пропорционально значениям веса uncapped.

График на рис. 10.24 представляет количество процессоров, активизируемых в каждом разделе на одном компьютере конфигурации.

11

Расширение возможностей группы ресурсов

В этой главе описывается (путем определения ресурсов и политик групп ресурсов), каким образом HACMP выполняет различные отдельные требования. Мы обсудим, как настроить эти опции, а также поделимся опытом проведения тестирований с использованием этих опций в HACMP 5.3. Управление выполнением запуска группы ресурсов, ее перемещения при сбое и возврата после восстановления определяется следующими политиками и атрибутами (табл. 11.1).

Таблица 11.1. Соотношение между атрибутами группы ресурсов и ее функциями

Атрибут	Запуск	Перемещение при сбое	Возврат после восстановления
Время установления (Setting Time)	Да		
Политика распределения узлов	Да		
Динамический приоритет узлов		Да	
Расположение, отменяющее приоритет (Priority Override Location)	Да		Да
Таймер отсроченного возврата после восстановления (Delayed Fallback Timer)			Да
Зависимость групп ресурсов – «родительский объект/дочерний объект»	Да	Да	Да
Зависимость групп ресурсов – расположение	Да	Да	Да

11.1 Время установления

Поведение отключенной группы ресурсов при запуске состоит в том, чтобы запуститься на первом доступном узле с наивысшим приоритетом, интегрируемом в кластер. Опция времени установления позволяет отсрочить старт группы ресурсов, чтобы в случае интеграции в кластер узла с более высоким приоритетом можно было подключить группу ресурсов на этом узле.

Работа опции времени установления

- Если эта опция сконфигурирована, она влияет на поведение при запуске всех групп ресурсов в кластере, для которых было установлено подключение на первом доступном узле (Online on First Available Node).
- Этот атрибут игнорируется только в тех случаях, когда узел, интегрируемый в кластер, является узлом с наивысшим приоритетом. В этой ситуации получение группы ресурсов происходит немедленно.
- Если группа ресурсов находится в состоянии ERROR, HACMP ожидает заданное время установления, прежде чем попытаться подключить группу ресурсов.
- Текущее значение времени установления остается активным, пока группа ресурсов не переместится на другой узел или не перейдет в отключенное состояние. Выполнение операции DARE может вызвать освобождение и повторное получение группы ресурсов; в этом случае новые значения времени установления вступают в действие немедленно.

Конфигурирование времени установления для групп ресурсов

Чтобы настроить время установления для групп ресурсов, нужно выполнить следующие действия:

- 1. Введите **smit hacmp.**

2. Выберите Extended Configuration (Расширенное конфигурирование) > Extended Resource Configuration (Расширенное конфигурирование ресурсов) > Configure a Resource Group Run-Time Policies (Конфигурирование политик времени выполнения для группы ресурсов) > Configure Settling Time for Resource Group (Конфигурирование времени установления для группы ресурсов) и нажмите Enter.

3. Выберите поле Settling Time (sec) [Время установления (с)] и введите в него поле любое положительное целое число. По умолчанию задается нулевое значение. Если задано это значение и узел, интегрируемый в кластер, не является узлом с наивысшим приоритетом, группа ресурсов будет ожидать в течение интервала времени установления. По истечении этого времени происходит получение группы ресурсов на узле, имеющем наивысший приоритет из списка узлов, интегрированных в кластер за интервал времени установления. Помните о том, что это относится только к группам ресурсов, использующим политику запуска Online on First Available Node (Подключение на первом доступном узле).

Вывод текущего времени установления

Чтобы вывести текущее значение времени установления в кластере, можно выполнить команду **clsettlingtime list**.

#/usr/es/sbin/cluster/utilities/clsettlingtime list

#SETTLING_TIME

120

Во время активизации групп ресурсов при запуске кластера значение времени установления можно также вывести командой **clRGinfo -t** (пример 11.1).

Пример 11.1. Вывод времени установления для группы ресурсов

```
#/usr/es/sbin/cluster/utilities/clRGinfo -t
-----------------------------------------------------------------------------
Group Name     Group State                  Node            Delayed
Timers
-----------------------------------------------------------------------------
settling_rg1   OFFLINE                      cobra          120 Seconds
               OFFLINE                      python         120 Seconds

settling_rg2   OFFLINE                      viper          120 Seconds
               OFFLINE                      python         120 Seconds
```

Обратите внимание на то, что это значение выводится только во время активизации; выходные данные команды **clRGinfo** будут пустыми по истечении периода времени установления и после активизации группы ресурсов на соответствующем узле.

Сценарий тестирования времени установления

Чтобы протестировать эту функцию, мы сконфигурировали кластер из трех узлов, использующий две группы ресурсов. При тестировании мы пытались проверить два утверждения:

1. После применения периода времени установления не происходит активизации группы ресурсов при запуске узла (если только этот узел не имеет наивысшего приоритета), пока не истечет время установления.

2. Если за период времени установления в кластер входит узел с наивысшим приоритетом, активизация группы ресурсов происходит немедленно, не ожидая окончания периода установления.

Мы начинаем выполнение своего теста с того, что задаем параметру времени установления значение 10 мин. (или 600 с). Две наши группы ресурсов (Settling1_RG и Settling2_RG) были настроены на использование политики запуска Online on First Available Node (Подключение на первом доступном узле). Мы задаем порядок в списке узлов для каждой группы ресурсов, чтобы перемещение при сбое для узлов **kaitlyn** и **thrish** выполнялось на узел **mike**. Рис. 11.1 содержит схему нашей конфигурации, а также последовательность поочередного запуска HACMP на узлах для тестирования параметра времени установления.

Наше тестирование включало следующие действия:

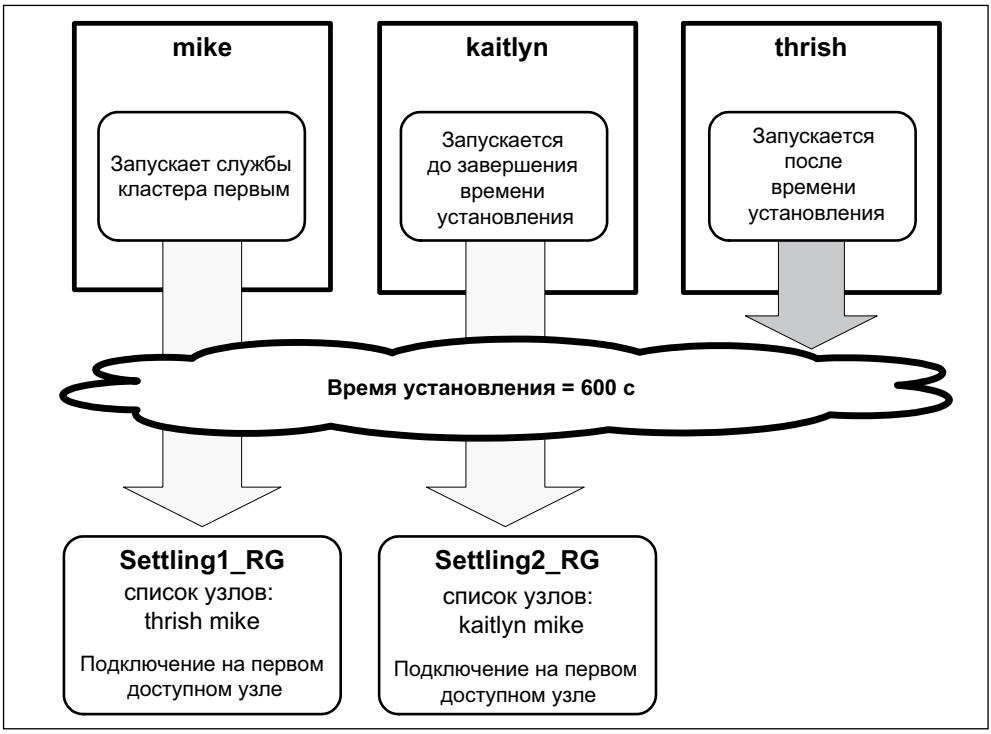

Рис. 11.1. Среда тестирования параметра времени установления

1. При неактивном HACMP на всех трех узлах, было определено значение времени установления 600 с.

2. Была выполнена синхронизация кластера. В процессе верификации/синхронизации была записана следующая информация:

The Resource Group Settling time value is: 120 secs.

The Resource Group(s) affected by the settling time are:

settling_rg1

settling_rg2

3. Были запущены службы кластера на узле **mike**. Мы запустили службы кластера HACMP на этом узле, так как он был последним в списке узлов для обеих групп ресурсов. После запуска демонов ни одна группа ресурсов не была активизирована на узле. Выполнение команды **clRGinfo -t** выдавало время установления, равное 600 с.

В файл hacmp.out была записана информация, представленная в примере 11.2:

4. Были запущены службы кластера на узле **kaitlyn**. Службы кластера были запущены через 2 мин. после начала периода времени установления. Сразу же после запуска узла была активизирована группа ресурсов Settling2_RG. Нам не пришлось ждать оставшиеся 8 мин. для получения и подключения группы ресурсов. Таких результатов мы и ожидали.

Пример 11.2. Проверка времени установления в /tmp/hacmp.out

```
#more /tmp/clstrmgr.debug
Wed Jul 13 10:40:53 NodeList::RmcComputeNodePriority: Using resource
attribute IBM.Host.PctTotalTimeIdle
Wed Jul 13 10:40:53 NodeList::RmcComputeNodePriority: for nodes 3, 2.
Wed Jul 13 10:40:53 NodeList::RmcComputeNodePriority: using values ,
0.0000, 0.0000.
Wed Jul 13 10:40:53 NodeList::RmcComputeNodePriority: condition is
DNP_largest
Wed Jul 13 10:40:53 In largest_comparison
```

5. Ожидание завершения периода времени установления. По истечении 600-секундного периода ожидания группа ресурсов Settling1_RG была активизирована на узле **mike**. Так как первый узел в списке узлов (**thrish**) не стал доступным в период времени установления, группа ресурсов была получена следующим узлом в списке узлов (**mike**). Опять же, таких результатов мы и ожидали.

6. Были запущены службы кластера на узле **thrish**. Мы запустили службы кластера на последнем узле по истечении периода времени установления; узлом не были получены ресурсы.

В целом наш тест подтвердил, что опция времени установления работает, как и ожидалось, и что группы ресурсов были распределены требуемым образом между узлами.

11.2 Политика распределения узлов

Одной из политик запуска, которую можно задать для группы ресурсов в кластере, является политика Online Using Node Distribution (Подключение с применением распределения узлов). При использовании этой политики распределения распространение групп ресурсов происходит таким образом, что во время запуска узел получает только одну группу ресурсов. Это позволяет сбалансировать приложения с интенсивным использованием процессоров на разных узлах.

В HACMP 5.3 поддерживается только политика распределения на основе узлов.

> **Примечание.** При обновлении с предыдущей версии, которая допускала использование распределения на основе сетей, эта конфигурация автоматически заменяется распределением на основе узлов.

Если при интеграции узла в кластер две или больше групп ресурсов отключены, политика сортирует группы ресурсов в следующем порядке:

1. Группа ресурсов с наименьшим количеством участвующих узлов.

2. Алфавитная сортировка имен групп ресурсов.

Если одна или несколько групп ресурсов являются родительскими группами ресурсов, HACMP отдает предпочтение родительской группе ресурсов. Дополнительные сведения по этой теме см. в разделе «Зависимости групп ресурсов».

11.2.1 Конфигурирование политики распределения групп ресурсов на основе узлов

Чтобы установить этот тип политики распределения, необходимо выполнить следующие действия (пример 11.3):

1. Введите **smit hacmp**.

2. В SMIT выберите Extended Configuration (Расширенное конфигурирование) > Extended Resource Configuration (Расширенное конфигурирование ресурсов) > HACMP Extended Resource Group Configuration (Расширенное конфигурирование групп ресурсов HACMP) > Add a Resource Group (Добавить группу ресурсов) и нажмите Enter.

3. Введите имя группы ресурсов.

4. Выберите политику запуска Online Using Distribution Policy (Подключение с использованием политики распределения) и нажмите Enter.

Пример 11.3. Конфигурирование политики распределения групп ресурсов на основе узлов

```
 Add a Resource Group (extended)

 Type or select values in entry fields.
 Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
 * Resource Group Name                                 []
 * Participating Nodes (Default Node Priority)         []

   Startup Policy                                      Online On Home Node O>
   Fallover Policy                                     Fallover To Next Prio>
   Fallback Policy                                     Fallback To Higher Pr>
 +-------------------------------------------------------------------------+
 |                            Startup Policy                               |
 |                                                                         |
 | Move cursor to desired item and press Enter.                            |
 |                                                                         |
 |   Online On Home Node Only                                              |
 |   Online On First Available Node                                        |
 |   Online Using Distribution Policy                                      |
 |   Online On All Available Nodes                                         |
 |                                                                         |
 | F1=Help                 F2=Refresh              F3=Cancel               |
 | F8=Image                F10=Exit                Enter=Do                |
 | /=Find                  n=Find Next                                     |
 +-------------------------------------------------------------------------
```

11.2.2 Сценарий тестирования политики распределения на основе узлов

В документации говорится, что при использовании групп ресурсов с этой политикой только одна группа ресурсов на узле будет получена при запуске. В этом сценарии тестирования мы хотели доказать, что это ограничение не повлияет на дополнительные группы ресурсов в кластере, использующие другие политики запуска, и что для них выполнение активизации не изменится.

Для выполнения тестирования политики распределения на основе узлов мы сконфигурировали кластер из двух узлов с пятью группами ресурсов:

- три группы ресурсов, использующие политику Online Using Distribution Policy (Подключение с использованием политики распределения);
- две группы ресурсов, применяющие политику Online On Home Node Only Policy (Подключение только на домашнем узле).

Это было сделано для того, чтобы увеличить количество групп ресурсов, использующих политику Online Using Node Distribution (Подключение с использованием распределения узлов), которые можно подключить в кластере. Рис. 11.2 содержит схему нашей конфигурации.

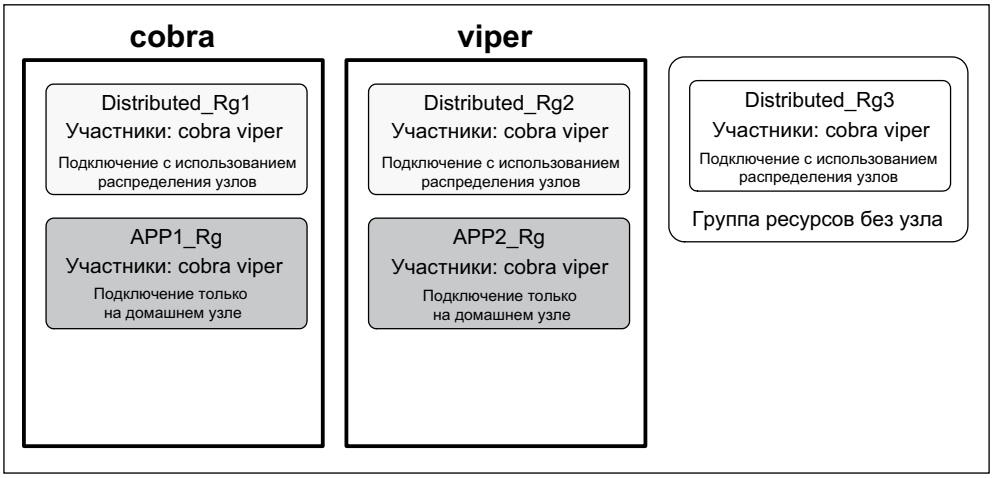

Рис. 11.2. Сценарий тестирования политики Online Using Node Distribution (Подключение с использованием распределения узлов)

После поочередного запуска служб кластера на каждом узле получение было выполнено только для групп ресурсов Distributed_Rg1 и Distributed_Rg2 из групп ресурсов с политикой запуска Online Using Node Distribution (Подключение с использованием распределения узлов). Третья группа ресурсов, Distributed_Rg3, осталась без узла, в состоянии OFFLINE. Однако ограничение не распространялось на оставшиеся группы ресурсов, использующие другую политику запуска. Группы ресурсов APP1_Rg и APP2_Rg были подключены на узлах **cobra** и **viper**, соответственно.

Результаты нашего тестирования подтвердили, что использование политики распределения на основе узлов ограничивает получение групп ресурсов этого типа на узле при запуске только одной группой ресурсов. Это не влияет на получение групп ресурсов, использующих другие политики.

> **Примечание.** Единственный метод перевода группы ресурсов Distributed_Rg3 в состояние ONLINE в этой среде состоит в том, чтобы установить опцию подключения группы ресурсов при каждом запуске кластера через панели HACMP или командой clRGmove.

11.3 Динамический приоритет узлов (DNP)

По умолчанию порядок приоритетов узлов для группы ресурсов соответствует порядку в списке участвующих узлов. Включение динамического приоритета узлов для групп ресурсов позволяет изменить стандартный механизм перемещения при сбое в HACMP и определить целевой узел для перемещения группы ресурсов при сбое на основании следующих предопределенных атрибутов RMC:

cl_highest_free_mem - узел с наибольшим процентным показателем свободной памяти

cl_highest_idle_cpu - узел с наименьшим использованием процессора

cl_lowest_disk_busy - узел с наименьшим использованием дисков

Для обеспечения эффективности DNP необходимо отметить следующее:

- DNP не предназначен для использования в кластере, содержащем меньше трех узлов;
- DNP не предназначен для групп ресурсов с одновременным доступом;
- DNP наиболее полезен в кластере, в котором все узлы имеют одинаковую вычислительную мощность и объем памяти.

> **Внимание!** Определение узла с наибольшим объемом свободной памяти выполняется на основании количества выполняющихся операций подкачки. При этом не учитывается, что объем физической памяти на одном узле может быть меньше, чем на другом.

На момент создания этой публикации сочетание устройств vpath с политикой cl_lowest_disk_busy не поддерживалось. Поддержка такой конфигурации, возможно, будет добавлена в версию позднее, в форме обновления.

11.3.1 Конфигурирование политики динамического приоритета узлов

Для того чтобы установить DNP для группы ресурсов, эта группа ресурсов не должна уже содержать какие-либо ресурсы. Назначение политики динамического приоритета узлов должно происходить при создании группы ресурсов. Для того чтобы группа ресурсов использовала одну из трех политик DNP, необходимо выполнить следующие действия (пример 11.4).

1. Введите **smit hacmp.**

2. В SMIT выберите Extended Configuration (Расширенное конфигурирование) > Extended Resource Configuration (Расширенное конфигурирование ресурсов) > HACMP Extended Resource Group Configuration (Расширенное конфигурирование групп ресурсов HACMP) > Add a Resource Group (Добавить группу ресурсов) и нажмите Enter.

Пример 11.4. Добавление группы ресурсов с использованием DNP

```
Add a Resource Group (extended)

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                    [Entry Fields]
* Resource Group Name                              []
* Participating Nodes (Default Node Priority)      []                    +

  Startup Policy                                    Online On Home Node O> +
  Fallover Policy                                   Fallover To Next Prio> +
  Fallback Policy                                   Fallback To Higher Pr> +
```

Установите в поле Fallover Policy (Политика перемещения при сбое) значение Dynamic Node Priority (Динамический приоритет узлов).

3. Назначьте ресурсы группе ресурсов, выбрав Extended Configuration (Расширенное конфигурирование) > Extended Resource Configuration (Расширенное конфигурирование ресурсов) > Extended Resource Group Configuration (Расширенное конфигурирование групп ресурсов) > Change/Show Resources and Attributes for a Resource Group (Изменить/показать ресурсы и атрибуты группы ресурсов), и нажмите Enter (пример 11.5).

Пример 11.5. Выбор используемой политики динамического приоритета узлов

```
Change/Show All Resources and Attributes for a Resource Group

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

[TOP]                                                 [Entry Fields]
  Resource Group Name                                DNP_test1
  Participating Nodes (Default Node Priority)        alexis jessica jordan
* Dynamic Node Priority Policy                       []                    +

  Startup Policy                                     Online On Home Node O>
  Fallover Policy                                    Fallover Using Dynami>
  Fallback Policy                                    Fallback To Higher Pr>
```

4. Выберите одну из трех доступных политик из выпадающего списка:

- cl_highest_free_mem
- cl_highest_idle_cpu;

- cl_lowest_disk_busy.

Затем следует выбрать ресурсы, которые будут составлять группу ресурсов.

5. Выполните верификацию и синхронизацию кластера.

Для просмотра текущей политики DNP существующей группы ресурсов в вашей конфигурации можно выполнить следующую команду:

```
#odmget -q group=APP1_RG HACMPresource | more
HACMPresource:
group = "APP1_RG"
name = "NODE_PRIORITY_POLICY"
value = "cl_highest_free_mem"
id = 1
```

> **Примечание.** Использование информации, получаемой непосредственно из ODM, предназначено только для информационных целей, так как формат разделов (stanzas) может быть различным в разных обновлениях и/или в новых версиях. Таким образом, жесткое кодирование ODM-запросов в пользовательских приложениях не поддерживается и его следует избегать.

11.3.2 Изменение существующей группы ресурсов для использования политики DNP

Нельзя изменить политику перемещения при сбое на DNP, если какие-либо ресурсы на данный момент являются частью группы ресурсов. При попытке изменения группы ресурсов через SMIT выдается ошибка, если прежде не удалить ресурсы. Для изменения политики можно сделать одно из двух:

1. Удалить группу ресурсов и заново создать ее, выбрав политику перемещения при сбое Dynamic Node Priority (Динамический приоритет узлов).
2. Перейти в экран Change/Show Resources and Attributes for a Resource Group (Изменить/показать ресурсы и атрибуты группы ресурсов), обнулить все ресурсы, входящие в группу ресурсов и нажать Enter. Затем следует выполнить переход Extended Configuration (Расширенное конфигурирование) > Extended Resource Configuration (Расширенное конфигурирование ресурсов) > Extended Resource Group Configuration (Расширенное конфигурирование групп ресурсов) > Change a Resource Group (Изменение группы ресурсов) и установить политику перемещения при сбое DNP. Затем можно выполнить назначение ресурсов в группу ресурсов и синхронизацию кластера, чтобы изменения вступили в действие.

11.3.3 Принцип работы динамического приоритета узлов

Начиная с HACMP 5.2 определение динамического приоритета узлов больше не опрашивает службы управления событиями (Event Management Services, emsvcs). Вместо этого **clstrmgrES** каждые 2 мин. опрашивает демон Resource Monitoring and Control (ctrmc) и ведет таблицу, в которой хранятся показатели текущего объема памяти, загруженности процессора и дискового ввода-вывода для каждого узла.

Мониторы ресурсов, содержащие информацию по каждой политике, следующие:

- IBM.PhysicalVolume,
- IBM.Host.

Каждый из этих мониторов может быть опрошен во время функционирования путем выполнения команд, представленных в примере 11.6.

Пример 11.6. Верификация показателей ресурсов

```
#lsrsrc -Ad IBM.Host | grep TotalPgSpFree
        TotalPgSpFree       = 123076
        PctTotalPgSpFree    = 93.8995

#lsrsrc -Ad IBM.Host | grep PctTotalTimeIdle
        PctTotalTimeIdle    = 99.6649

#lsrsrc -Ap IBM.PhysicalVolume
resource 1:
        Name              = "vpath7"
```

Текущую таблицу, которую ведет `clstrmgrES`, можно вывести путем выполнения команд, представленных в примере 11.7.

Пример 11.7. Значения DNP, выводимые диспетчером кластера

```
#lssrc -ls clstrmgrES
Current state: ST_STABLE
sccsid = "@(#)36   1.135.1.37   src/43haes/usr/sbin/cluster/hacmprd/main.C,
hacmp.pe, 51haes_r530, r5300525a 6/20/05 14:13:01"
i_local_nodeid 1, i_local_siteid -1, my_handle 2
ml_idx[1]=0     ml_idx[2]=1     ml_idx[3]=2
There are 0 events on the Ibcast queue
There are 0 events on the RM Ibcast queue
CLversion: 8
local node vrmf is 5300
cluster fix level is "0"
The following timer(s) are currently active:
Current DNP values
DNP Values for NodeId - 1  NodeName - cobra
    PgSpFree = 130771  PvPctBusy = 0  PctTotalTimeIdle = 99.947917
DNP Values for NodeId - 2  NodeName - python
    PgSpFree = 130741  PvPctBusy = 0  PctTotalTimeIdle = 99.879167
DNP Values for NodeId - 3  NodeName - viper
    PgSpFree = 124489  PvPctBusy = 0  PctTotalTimeIdle = 99.941667
```

Значения в таблице употребляются для определения DNP в случае перемещения при сбое. Если в процессе опроса текущего состояния демоном clstrmgrES происходит перемещение при сбое, для определения DNP используется значение, имевшее место, когда кластер последний раз был в стабильном состоянии.

11.3.4 Сценарий тестирования динамического приоритета узлов

Для тестирования функций динамического приоритета узлов мы установили кластер из трех узлов с двумя группами ресурсов, использующими различные политики DNP. Первая группа ресурсов A (APP1_RG) использует значение cl_highest_idle_cpu, тогда как группа ресурсов B (APP2_RG) применяет значение cl_highest_free_mem для определения узла для перемещения при сбое. Для каждой группы ресурсов мы выполнили тестирование различных сценариев перемещения при сбое и задокументировали результаты. На рис. 11.3 представлена схема используемой конфигурации.

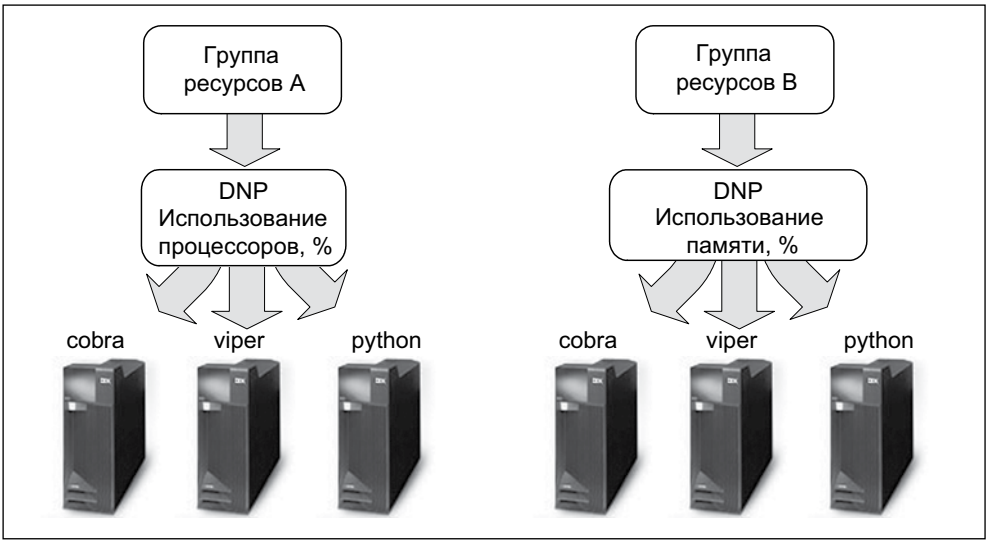

Рис. 11.3. Сценарий тестирования динамического приоритета узлов

В нашей среде мы установили для каждой группы ресурсов использование следующих политик запуска/перемещения при сбое/возврате после восстановления:

Startup = Online Using Distribution Policy

Fallover = Fallover Using Dynamic Node Priority

Fallback = Never Fallback

Мы установили следующий порядок узлов для обеих групп ресурсов:

cobra viper python

> **Примечание.** Мы установили для каждой группы ресурсов одинаковый список узлов, чтобы показать, что перемещение при сбое не следует порядку в списке. Вместо этого выполняется определение узла с наименьшим использованием ресурсов процессора, памяти или дисков (в зависимости от заданной политики DNP) для содержания групп ресурсов.

Наша топология содержит сеть Ethernet с применением IP-синонимов. Это означает, что в случае перемещения при сбое на одном узле могли располагаться несколько групп ресурсов. Это может произойти только в том случае, если определение

узла с наименьшим использованием процессора и определение узла с наименьшим употреблением памяти указывают на один и тот же узел.

> **Примечание.** Чтобы избежать перемещения нескольких групп ресурсов на один узел при использовании DNP и IP-синонимов, можно установить зависимость расположения группы ресурсов Online on Different Nodes (Подключение на разных узлах). Альтернативный вариант заключается в применении перехвата IP-адреса посредством замены, чтобы после перемещения при сбое хост не мог содержать несколько сервисных IP-адресов.

Этапы сценария тестирования DNP

В целях упрощения тестирования мы не использовали реальные рабочие нагрузки приложений для перегрузки процессоров или памяти узла. Вместо этого мы запустили цикл, чтобы загрузить два процессора, и применили команду **rmss**, чтобы логически сократить количество памяти, видимой в системе, инициируя выполнение операций подкачки.

Ниже приведен перечень действий, которые мы выполнили, чтобы убедиться в применении политик DNP.

1. Мы запустили службы кластера на узле **cobra**. Только первая группа ресурсов (APP1_RG) была подключена в связи с установленной политикой Online Using Distribution Policy (Подключение с использованием политики распределения).

2. Мы перешли в smit hacmp > System Management (C-SPOC) > HACMP Resource Group and Application Management (Управление группами ресурсов и приложениями HACMP) > Bring a Resource Group Online (Перевести группу ресурсов в подключенное состояние), после чего выбрали APP2_RG, затем выбрали Restore_Node_Priority_Order и нажали Enter. Это вызвало подключение второй группы ресурсов (APP2_RG) на узле **cobra**, как показано в примере 11.8:

Пример 11.8. Информация о группах ресурсов

```
#/usr/es/sbin/cluster/utilities/clRGinfo
APP1_RG        ONLINE                    cobra
               OFFLINE                   viper
               OFFLINE                   python

APP2_RG        ONLINE                    cobra
               OFFLINE                   viper
               OFFLINE                   python
```

Это позволило освободить узлы (**viper и python**), что дает нам возможность применять два узла для выполнения перемещения при сбое. Это также позволило нам протестировать определение показателей неиспользуемой памяти и процессора на следующем этапе.

3. Мы выполнили такую последовательность команд на узле **viper**, чтобы уменьшить количество доступной памяти и инициировать операции подкачки, как показано в примере 11.9:

Пример 11.9. Инициирование операций подкачки

```
cobra-#rmss -p
Simulated memory size is 8192 Mb.

viper-#rmss -p
Simulated memory size is 8192 Mb.

viper-#rmss -c 2000
Simulated memory size changed to 2000 Mb.

viper-#lptest 80 1000000000 > /app1/garbage_file.out

viper-#lsps -s
Total Paging Space   Percent Used
      512MB               7%

viper-#lsrsrc -Ad IBM.Host | grep PctTotalPgSpFree
      PctTotalPgSpFree    = 93.8995
```

Сначала мы сократили объем памяти на узле **viper** до 2 Гб, уменьшив, таким образом, объем доступной памяти. Затем мы использовали команду **lptest**, чтобы сгенерировать запись большого количества символов в файл, что инициирует операции подкачки. Мы осуществляли мониторинг изменений значения `TotalPgSpFree` в классе IBM.Host. В результате повышенной нагрузки на память узла **viper** узел **python** стал следующим логичным вариантом при определении DNP для группы ресурсов `APP1_RG`.

4. Мы выполнили два скрипта ksh, чтобы сгенерировать цикл для загрузки двух процессоров на узле **viper** (пример 11.10).

Пример 11.10. Инициирование нагрузки на процессор

```
#vi cpu_loop1
while true
do
:
done

#./cpu_loop1
#./cpu_loop2
```

В результате выполнения двух экземпляров скрипта в фоновом режиме два процессора оказались перегружены. Результаты приведены в примере 11.11:

Пример 11.11. Проверка нагрузки на процессор

```
#sar -P ALL 5
AIX p630n02 3 5 000685BF4C00    07/08/05
System configuration: lcpu=4

11:56:46 cpu    %usr    %sys    %wio   %idle
11:56:51  0        0       0       0     100
          1      100       0       0       0
          2      100       0       0       0
          3        0       0       0     100
          -       50       0       0      50
```

Мы перегрузили два процессора на узле **viper**, так как он был следующим узлом в списке узлов группы ресурсов. Таким образом, в результате уменьшение количества неиспользуемых процессоров узел **python** стал следующим логичным вариантом при определении DNP для группы ресурсов `APP2_RG`.

5. Мы выполнили постепенную остановку HACMP с передачей ресурсов на резервные узлы (graceful with takeover) на узле **cobra**.

После остановки служб кластера на узле было выполнено определение DNP по значениям в текущей таблице clstrmgrES. Значения на момент перемещения при сбое представлены в табл. 11.2.

Таблица 11.2. Определение целевого узла DNP для перемещения при сбое по первому сценарию

Узлы кластера	Наибольший показатель бездействия процессора	Наибольший показатель свободной памяти
viper	93.8995	93.8995
python	93.8995	93.8995
Наилучший целевой узел	python	python

Как видим из результатов[1], наилучшим целевым узлом является узел **python**. Процесс выполнения определения DNP можно просмотреть в файле /tmp/clstrmgr.debug во время перемещения при сбое (пример 11.12).

Пример 11.12. Сообщения отладки диспетчера кластера во время перемещения при сбое

```
#more /tmp/clstrmgr.debug
Wed Jul 13 10:40:53 NodeList::RmcComputeNodePriority: Using resource
attribute IBM.Host.PctTotalTimeIdle
Wed Jul 13 10:40:53 NodeList::RmcComputeNodePriority: for nodes 3, 2.
Wed Jul 13 10:40:53 NodeList::RmcComputeNodePriority: using values ,
```

[1] Из результатов в табл. 11.2 и 11.3 и примера 11.12 этого не видно. Скорее всего, в тексте опечатка. Мы не будем приводить какие-либо другие значения загруженности процессора и памяти, но предлагаем вам самостоятельно предположить, что, например, в табл. 11.2 показатели бездействия процессора и свободной памяти на узле python выше, чем на узле viper. *Примеч. науч. ред.*

```
0.0000, 0.0000.
Wed Jul 13 10:40:53 NodeList::RmcComputeNodePriority: condition is
DNP_largest
Wed Jul 13 10:40:53 In largest_comparison
Wed Jul 13 10:40:53 NodeList::RmcComputeNodePriority: Computed node order -
3, 2.
Wed Jul 13 10:40:53 For Resource Group APP1_RG, BestNode got node order
Wed Jul 13 10:40:53 NodeList::showNodeList: Got the following 2 node IDs:
Wed Jul 13 10:40:53 NodeList::showNodeList: 3 2
Wed Jul 13 10:40:53 The best node for group APP1_RG is python.
Wed Jul 13 10:40:53 RGPA got viper as highest priority node.
............
............
Wed Jul 13 10:40:53 NodeList::RmcComputeNodePriority: Using resource
attribute IBM.Host.TotalPgSpFree
Wed Jul 13 10:40:53 NodeList::RmcComputeNodePriority: for nodes 3, 2.
Wed Jul 13 10:40:53 NodeList::RmcComputeNodePriority: using values ,
0.0000, 0.0000.
Wed Jul 13 10:40:53 NodeList::RmcComputeNodePriority: condition is
DNP_largest
Wed Jul 13 10:40:53 In largest_comparison
Wed Jul 13 10:40:53 NodeList::RmcComputeNodePriority: Computed node order -
3, 2.
Wed Jul 13 10:40:53 For Resource Group APP2_RG, BestNode got node order
Wed Jul 13 10:40:53 NodeList::showNodeList: Got the following 2 node IDs:
Wed Jul 13 10:40:53 NodeList::showNodeList: 3 2
Wed Jul 13 10:40:53 The best node for group APP2_RG is python.
Wed Jul 13 10:40:53 RGPA got viper as highest priority node.
```

6. Мы выполнили реинтеграцию узла **cobra** в кластер.

7. Мы внесли изменения в параметры памяти и процессора во второй раз. На этот раз настройка узлов **cobra** и **viper** делается таким образом, чтобы при определении DNP выполнялось распределение двух групп ресурсов на узле **python** между двумя разными узлами (пример 11.13).

Пример 11.13. Инициирование нагрузки на процессор и память

```
viper-#rmss -r
Simulated memory size is 8192 Mb.

viper-#./cpuloop1
viper-#./cpuloop2

viper-#sar -P ALL 5
AIX p630n02 3 5 00065BF4C00 07/08/05
System configuration: lcpu=4

18:03:20 cpu    %usr    %sys    %wio   %idle
18:03:25  0        0       0       0     100
```

```
          1      100       0       0       0
          2      100       0       0       0
          3        0       0       0     100
          -       50       0       0      50

viper-#lsps -s
Total Paging Space   Percent Used
      512MB               1%


cobra-#rmss -c 2000
Simulated memory size changed to 2000 Mb.

cobra-#lptest 80 1000000000 > /app1/garbage_file.out

cobra-#lsps -s
Total Paging Space   Percent Used
      512MB             6%
```

Мы перегрузили два процессора на узле **viper** и инициировали операции подкачки на узле **cobra**, чтобы политика DNP в каждой группе ресурсов распределила их по разным узлам.

8. Мы выполнили **halt -q** на узле **python**. Это вызвало перемещение группы ресурсов APP1_RG на узел **cobra**, так как на нем были доступны все процессоры. Группа ресурсов APP2_RG была перемещена на узел **viper**, так как на нем не выполнялись операции подкачки и он имел больше свободной памяти. Табл. 11.3 показывает текущие значения clstrmgrES на момент перемещения при сбое.

Таблица 11.3. Определение целевого узла DNP для перемещения при сбое по второму сценарию

Узел кластера	Наибольший показатель бездействия процессора	Наибольший показатель свободной памяти
viper	93.8995	93.8995
python	93.8995	93.8995
Наилучший целевой узел[1]	cobra	viper

Результаты тестирования сценария DNP

В результате выполнения сценария тестирования мы смогли доказать, что во время перемещения при сбое порядок в списке узлов по умолчанию игнорировался. После реинтеграции отказавшего узла мы изменили нагрузку на память и процессоры на двух бездействующих узлах и выполнили тестирование отказа узла, содержащего группы ресурсов. При определении DNP во второй раз было успешно выполнено распределение групп ресурсов по двум разным дежурным узлам.

[1] Та же ошибка, что и в таблице 11.2. *Примеч. науч. ред.*

В нашей среде мы использовали практичные команды для изменения нагрузки на процессоры и память систем. Мы считаем, что результаты будут такими же, когда нагрузка будет вызвана выполняющимся приложением. Во время настройки своей среды DNP обязательно протестируйте все возможные сценарии перемещения при сбое и выполните команды **lsrscr -Ad и lssrc -ls clstrmgrES**, описанные в этом руководстве, для мониторинга текущей нагрузки на процессоры, память и дисковый ввод-вывод на узлах кластера.

11.4 Расположение, отменяющее приоритет (POL)

Понятие расположения, отменяющего приоритет (priority override location), появилось в HACMP 5.1 в качестве замены для прежнего атрибута «sticky». Одно важное отличие от более ранних версий состоит в том, что политика POL теперь всегда задается неявным образом при перемещении группы ресурсов вручную. Считается, что если вы явным образом перемещаете группу ресурсов куда-либо, значит, вы хотите, чтобы она там и оставалась.

При установке значения POL выполняется привязка группы ресурсов к этому узлу. Политика действует до перезагрузки всех узлов кластера или выполнения другой операции перемещения группы ресурсов с указанием опции Restore_Node_Priority_Order. Существует еще один параметр – Persist Across Cluster Reboot. Если он включен, POL сохраняется даже после перезагрузки всех узлов кластера.

> **Примечание.** При каждом явном перемещении группы ресурсов следует либо сбрасывать атрибут POL, либо помнить, что он установлен, так как это косвенным образом изменяет выполнение перемещения при сбое.

Атрибут POL также оказывает действие при отключении или подключении группы ресурсов. Если вы решили отключить свою группу ресурсов, атрибут POL будет отражать состояние OFFLINE. При повторном подключении группы ресурсов с явным указанием целевого узла в качестве POL будет установлен этот узел.

Перемещение группы ресурсов

Обратите внимание на то, что эта операция недоступна для групп ресурсов без одновременного доступа[1].

Чтобы переместить группу ресурсов проделайте следующее:

1. Перейдите в smit hacmp > System Management (C-SPOC) > HACMP Resource Group and Application Management (Управление группами ресурсов и приложениями HACMP) > Move a Resource Group to Another Node / Site (Перемещение группы ресурсов на другой узел/сайт) > Move Resource Groups to Another Node (Перемещение группы ресурсов на другой узел), после чего следует выбрать группу ресурсов и нажать Enter.

2. Выберите одну из следующих опций:

---

[1] Здесь ошибка – эта операция доступна только для групп ресурсов без одновременного доступа, а не наоборот. *Примеч. науч. ред.*

- Restore_Node_Priority_Order (Восстановление порядка приоритетов узлов)

  или

- Destination node (Целевой узел).

3. В следующем экране выберите соответствующее значение для следующего параметра:

- Persist Across Cluster Reboot?

По умолчанию задано значение false. Если установить true, то атрибут Priority Override Location сохраняется после полной перезагрузки кластера. Если установить false, то атрибут Priority Override Location не сохраняется после полной перезагрузки кластера, вследствие чего группа ресурсов возвращается к заданному по умолчанию режиму работы.

4. После перемещения группы ресурсов можно убедиться в успешности перемещения и в том, что задан атрибут POL, путем выполнения команды **clRGinfo -p**. Образец результата выполнения данной команды представлен в примере 11.14.

Пример 11.14. Проверка расположения, отменяющего приоритет

```
# /usr/es/sbin/cluster/utilities/clRGinfo -p

Cluster Name: migration2

Resource Group Name: C1ORG2
Priority Override Information:
    Primary Instance POL:
Node                          State
----------------------------- ---------------
viper                         ONLINE
cobra                         OFFLINE

Resource Group Name: C1ORG1
Priority Override Information:
    Primary Instance POL:   viper
Node                          State
----------------------------- ---------------
cobra                         OFFLINE
viper                         ONLINE
```

Если установлен атрибут POL, на обоих узлах генерируется файл /usr/es/sbin/cluster/etc/clpol, который остается на этих узлах до перезагрузки кластера или перемещения группы ресурсов с установленной опцией Restore_Node_Priority_Order. Файл содержит представление всех расположений, отменяющих приоритет, и его можно интерпретировать в следующем формате: [RG id] [node id] [pol] [per?]

3 2 2 1 // RG 3 on node 2 is OFFLINE persistent

3 1 2 1 // RG 3 on node 1 is OFFLINE persistent

1 1 1 0 // RG 1 on node 1 is ONLINE non-persistent

Данные файла имеют числовой формат и не предназначены для просмотра или изменения конечным пользователем.

> **Внимание.** Удаление файла clpol вручную НЕ сбрасывает атрибут POL в активном кластере.

Подключение или отключение группы ресурсов

Существует несколько дополнительных операций, при которых можно неявным образом установить расположение, отменяющее приоритет для группы ресурсов. При подключении или отключении группы ресурсов с использованием следующих опций меню и выборе целевого узла происходит установка атрибута POL:

- smit hacmp > System Management (C-SPOC) > HACMP Resource Group and Application Management (Управление группами ресурсов и приложениями HACMP) > Bring a Resource Group Online (Перевести группу ресурсов в подключенное состояние);
- smit hacmp > System Management (C-SPOC) > HACMP Resource Group and Application Management (Управление группами ресурсов и приложениями HACMP) > Bring a Resource Group Offline (Перевести группу ресурсов в отключенное состояние).

Если при подключении группы ресурсов установить опцию Restore_Node_Priority_Order, группа ресурсов будет подключена на узле, для которого установлен наивысший приоритет (определяемый на основании списка узлов), и атрибут POL не будет установлен.

При отключении группы ресурсов с использованием метода, приведенного выше, следует помнить о том, что устанавливается другой тип POL. Группа ресурсов будет выводить для атрибута POL состояние OFFLINE. Это означает, что последующие попытки перезапуска служб кластера не позволят вам выполнить подключение группы ресурсов. Способ сброса атрибута POL состоит в перемещении или подключении группы ресурсов с использованием опции Restore_Node_Priority_Order.

Сброс расположения, отменяющего приоритет

После установки POL для сброса этого атрибута и восстановления заданного по умолчанию режима работы группы ресурсов необходимо инициировать перемещение другой группы ресурсов и выбрать опцию Restore_Node_Priority_Order. В примере 11.15 показано, как выглядит эта опция меню.

При выборе этой опции:

- группа ресурсов перемещается на доступный узел с наивысшим приоритетом
- удаляется ранее установленное расположение, отменяющее приоритет
- если группа ресурсов уже находится на узле с наивысшим приоритетом, атрибут POL сбрасывается без выполнения перемещения.

При использовании групп ресурсов с политикой запуска Online Using Distribution Policy (Подключение с использованием политики распределения) опция меню будет выглядеть по-другому. Вместо опции Restore_Node_Priority_Order для сброса

Пример 11.15. Сброс расположения, отменяющего приоритет

```
 Move a Resource Group to Another Node / Site

 Move cursor to desired item and press Enter.

   Move Resource Groups to Another Node
   Move Resource Groups to Another Site

   +-------------------------------------------------------------------------+
   |                        Select a Destination Node                        |
   |                                                                         |
   | Move cursor to desired item and press Enter.                            |
   |                                                                         |
   |   # To choose the highest priority available node for the               |
   |   # resource group, and to remove any Priority Override Location        |
   |   # that is set for the resource group, select                          |
   |   # "Restore_Node_Priority_Order" below.                                |
   |   Restore_Node_Priority_Order                                           |
   |                                                                         |
   |   # To choose a specific node, select one below.                        |
   |   viper                                                                 |
   |                                                                         |
   | F1=Help                 F2=Refresh              F3=Cancel               |
   | F8=Image                F10=Exit                Enter=Do                |
 F1| /=Find                  n=Find Next                                     |
```

атрибута POL следует использовать опцию Reset_Any_Priority_Overrides. Для группы ресурсов данного типа восстановление приоритета узла не должно вызывать действительного перемещения ресурсов, так как в этом случае отсутствует понятие узла с наивысшим приоритетом.

Рекомендации по использованию атрибута расположения, отменяющего приоритет

При управлении расположением своих групп ресурсов следует быть особенно внимательным к тому, где расположен атрибут POL и каково его текущее значение. Ниже мы перечислим ситуации, в которых наиболее высока вероятность возникновения путаницы или при которых не восстановится заданный по умолчанию режим работы.

1. Подключение группы ресурсов. При попытке подключения группы ресурсов на узле с более высоким приоритетом можно случайно установить POL, выбрав имя узла и не установив опцию Restore_Node_Priority_Order.

2. Отключение группы ресурсов. При отключении группы ресурсов также происходит установка атрибута POL. Группа ресурсов будет оставлена в состоянии OFFLINE. При последующем запуске служб кластера не происходит повторное подключение группы ресурсов, если только не будет повторно выдан запрос на получение и установлена опция Restore_Node_Priority_Order.

3. Перемещение группы ресурсов обратно на узел с наивысшим приоритетом. Нужно быть особенно внимательным, если при установленной опции Never Fallback (Без выполнения возврата после восстановления) или Online Using Distribution Policy (Подключение с использованием политики распределения) вручную выполняется перемещение группы ресурсов обратно на узел с наивысшим приоритетом. Если вместо того чтобы выбрать Restore_Node_Priority_Order, выбирается определенное имя узла, атрибут POL будет установлен на этом узле. В такой ситуации легко забыть, что установлена политика и что при выполнении последующих операций группа ресурсов может работать не так, как ожидается.

4. Установка для параметра Persist Across Cluster Reboot значения Yes. Используйте эту опцию с осторожностью. При управлении расположением группы ресурсов следует всегда запоминать, установлена ли эта опция. Тот, кто не знаком с работой расположения, отменяющего приоритет, может забыть о том, что эта опция установлена, и получить смешанные результаты.

11.5 Таймер отсроченного возврата после восстановления

Этот параметр позволяет настраивать выполнение возврата после восстановления для группы ресурсов таким образом, чтобы оно происходило в заданное время: ежедневно, еженедельно, ежемесячно, ежегодно или в заданную дату и время. Это полезно для того, чтобы планировать возвраты после восстановления на нерабочее время. схема на рис. 11.5 показывает три различных этапа при использовании таймеров отсроченного возврата после восстановления.

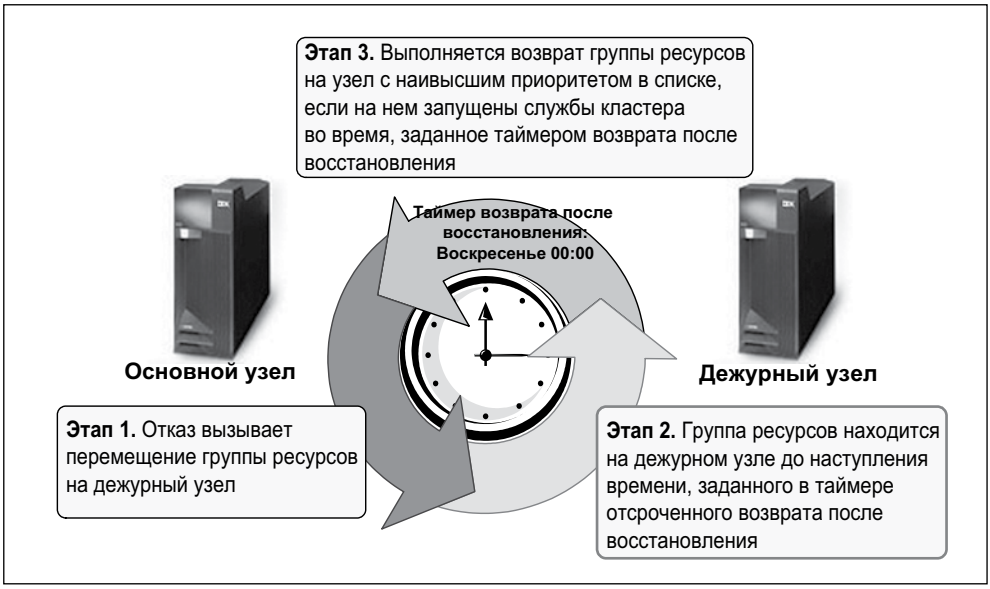

Рис. 11.4. Этапы таймеров отсроченного возврата после восстановления

В случае отказа происходит перемещение группы ресурсов на дежурный узел. Группа ресурсов остается там, пока не наступит время, заданное таймером отсроченного возврата после восстановления. Если в это время на основном узле активны службы кластера, группа ресурсов выполняет перемещение на узел с наивысшим приоритетом. Если узел с более высоким приоритетом в это время будет недоступен, таймер возврата после восстановления будет переустановлен и возврат после восстановления будет отсрочен до следующего повторения этого значения времени.

Принцип работы таймера отсроченного возврата после восстановления

Для эффективной реализации среды, использующей таймеры отсроченного возврата после восстановления, следует учитывать следующие аспекты:

- Таймер отсроченного возврата после восстановления используется только для групп ресурсов, применяющих политику Fallback To Higher Priority Node In The List (Возврат после восстановления на узел с более высоким приоритетом в списке).
- Если на промежуток времени, заданный таймером, узел с более высоким приоритетом недоступен, возврат группы ресурсов не выполняется.
- При использовании определенного значения даты в таймере возврата после восстановления повторная проверка доступности узла с более высоким приоритетом не выполняется. Только использование других значений перезапустит таймер и позволит выполнить проверку при следующей итерации.
- Если на определенном узле для группы ресурсов задан атрибут расположения, отменяющего приоритет (priority override location, POL), то в момент времени, когда таймер возврата после восстановления выполняет проверку, группа ресурсов будет перемещена на узел, заданный параметром POL.
- При использовании набора Same Node Dependency, если одна группа ресурсов в наборе имеет таймер возврата после восстановления, он применяется ко всему набору. При использовании таймера возврата после восстановления для групп ресурсов, применяющих политику Same Site Dependency, этот таймер должен быть одинаковым для всех групп ресурсов в наборе. Дополнительные сведения по этой теме см. в разделе «Зависимости групп ресурсов».
- Нельзя удалить таймер возврата после восстановления, если группа ресурсов его в данный момент использует.
- Нельзя сконфигурировать таймеры отсроченного возврата после восстановления через экраны Initialization and Standard Configuration (Инициализация и стандартное конфигурирование). Их можно сконфигурировать только с использованием пути Extended Configuration (Расширенное конфигурирование).

Конфигурирование таймера возврата после восстановления для групп ресурсов

Для конфигурирования этой функции нужно выполнить следующие действия:

1. Введите smit hacmp.

2. Выберите Extended Configuration (Расширенное конфигурирование) > Extended Resource Configuration (Расширенное конфигурирование ресурсов) > Configure Re-

source Group Run-Time Policies (Конфигурирование политик времени выполнения для группы ресурсов) > Configure Delayed Fallback Timer Policies (Конфигурирование политик таймера отсроченного возврата после восстановления) > Add a Delayed Fallback Timer Policy (Добавить политику таймера отсроченного возврата после восстановления) и нажмите Enter.

3. Выберите политику из списка: daily (ежедневно), weekly (еженедельно), monthly (ежемесячно), yearly (ежегодно) – либо укажите определенную дату.

4. Введите значения следующих полей:

- Name of Fallback Policy (Имя политики возврата после восстановления). Укажите имя политики длиной не более 32 символов. Используйте только алфавитно-цифровые символы и знаки подчеркивания. Не применяйте цифры в начале имени или зарезервированные слова.
- Policy selected (Выбранная политика). Выберите соответствующие значения для выбранной политики. Поля будут содержать день, час, минуты, год или их сочетание.

Пример 11.16. Определение политики таймера возврата после восстановления в группе ресурсов

```
 Change/Show All Resources and Attributes for a Resource Group

 Type or select values in entry fields.
 Press Enter AFTER making all desired changes.

 [TOP]                                                   [Entry Fields]
   Resource Group Name                                  fallback1_rg
   Participating Nodes (Default Node Priority)          cobra viper python

   Startup Policy                                       Online On Home Node Only
   Fallover Policy                                      Fallover To Next Priority>
   Fallback Policy                                      Fallback To Higher Priori>
   Fallback Timer Policy (empty is immediate)           []                         +

   Service IP Labels/Addresses                          []                         +
   Ap+-------------------------------------------------------------------------+ +
     ¦                Fallback Timer Policy (empty is immediate)               ¦
   Vo¦                                                                         ¦ +
   Us¦ Move cursor to desired item and press Enter.                            ¦ +
   Au¦                                                                         ¦ +
   Fi¦   july12                                                                ¦ +
   Fi¦   july1_test                                                            ¦ +
 [MOR¦                                                                         ¦
     ¦ F1=Help                 F2=Refresh              F3=Cancel               ¦
 F1=H¦ F8=Image                F10=Exit                Enter=Do                ¦
 F5=R¦ /=Find                  n=Find Next                                     ¦
 F9=S+-------------------------------------------------------------------------+
```

5. Чтобы назначить таймер группе ресурсов, нужно выбрать smit hacmp > Extended Configuration (Расширенное конфигурирование) > Extended Resource Configuration (Расширенное конфигурирование ресурсов) > HACMP Extended Resource Group Configuration (Расширенное конфигурирование групп ресурсов HACMP), после чего выбрать требуемую группу ресурсов из списка и нажать Enter.

6. Нажмите клавишу F4, чтобы выбрать одну из политик, сконфигурированных на предыдущих этапах. Появится следующий экран (пример 11.16).

7. Выберите требуемую политику таймера возврата после восстановления из списка и нажмите Enter.

8. Добавьте требуемые дополнительные ресурсы в группу ресурсов и нажмите Enter.

9. Запустите процесс верификации и синхронизации в кластере для распространения изменений на всех узлах.

Вывод таймеров отсроченного возврата после восстановления в группе ресурсов

При работе в среде с запущенным кластером можно проверить наличие существующих политик возврата после восстановления для групп ресурсов путем просмотра выходных данных команды **clshowres**, как показано в примере 11.7:

Пример 11.17. Ресурсы кластера

```
#/usr/es/sbin/cluster/utilities/clshowres
Resource Group Name                          fallback1_rg
Participating Node Name(s)                   cobra viper python
Startup Policy                               Online On Home Node Only
Fallover Policy                              Fallover To Next Priority Node In
The List
Fallback Policy                              Fallback To Higher Priority Node
In The List
Delayed Fallback Timer                       july12
```

> **Примечание.** Помните, что атрибут Delayed Fallback Timer (Таймер отсроченного возврата после восстановления) будет выводиться в меню, только если установлена политика Fallback To Higher Priority Node In The List (Возврат после восстановления на узел с более высоким приоритетом в списке).

Для вывода значений существующей политики таймера возврата после восстановления можно выбрать smit hacmp > Extended Configuration (Расширенное конфигурирование) > Extended Resource Configuration (Расширенное конфигурирование ресурсов) > Configure Resource Group Run-Time Policies (Конфигурирование политик времени выполнения для группы ресурсов) > Configure Delayed Fallback Timer Policies (Конфигурирование политик таймера отсроченного возврата после восстановления) > Change/Show a Delayed Fallback Timer Policy (Изменить/вывести политику отсроченного таймера возврата после восстановления) и выбрать политику из списка.

Альтернативный вариант состоит в том, чтобы опросить объектный класс HACMPtimer путем выполнения следующего кода:

```
#odmget HACMPtimer
        HACMPtimer:
                policy_name = "july12"
                recurrence = "once"
                year = 105
                month = 6
                day_of_month = 12
                week_day = 0
                hour = 17
                minutes = 47
```

Внимание! Использование информации, получаемой непосредственно из ODM, предназначено только для информационных целей, так как формат разделов (stanzas) может быть различным в разных обновлениях и/или в новых версиях. Таким образом, жесткое кодирование ODM-запросов в пользовательских приложениях не поддерживается и его следует избегать.

Сценарий тестирования таймеров отсроченного возврата после восстановления

Для того чтобы протестировать работу этой функции, мы сконфигурировали кластер из двух узлов с одной группой ресурсов, использующей таймер возврата после восстановления. В целях выполнения своего теста мы применили таймер возврата после восстановления, содержащий определенную дату.

Ниже перечислены атрибуты используемой нами группы ресурсов.

- Имя группы ресурсов: fallback1_rg.
- Участвующие узлы: cobra viper.
- Политика запуска: Online On Home Node Only (Подключение только на домашнем узле).
- Политика перемещения при сбое: Fallover To Next Priority Node In The List (Перемещение при сбое на узел из списка со следующим приоритетом).
- Политика возврата после восстановления: Fallback To Higher Priority Node In The List (Возврат после восстановления на узел с более высоким приоритетом в списке).
- Таймер отсроченного возврата после восстановления: 12 июля.

В процессе выполнения верификации/синхронизации нашего кластера мы заметили следующее сообщение в выходных данных clverify:

Resource Group 'fallback1_rg' is configured to use 'july12' fallback timer policy

После установления политики мы имитировали три этапа, представленные на рис. 11.5.

Ниже перечислены предпринятые нами действия.

1. Мы запустили службы кластера на обоих узлах: **cobra viper.**

2. Мы имитировали отказ (этап 1 на рис. 11.4) основного узла **cobra**. Для этого мы выполнили на узле постепенную остановку HACMP с передачей ресурсов на резервные узлы (graceful with takeover). Мы решили выполнить перемещение группы ресурсов таким способом, чтобы избежать назначения POL при выполнении операции перемещения группы ресурсов.

3. Мы убедились в том, что группа ресурсов была подключена на узле **viper**.

4. Была выполнена реинтеграция узла **cobra** (узла с более высоким приоритетом) в кластер. Когда мы запустили службы кластера на узле, мы заметили следующее сообщение в файле `/tmp/hacmp.out`:

No action taken on resource group 'fallback1_rg'

The Resource Group 'fallback1_rg' has been configured

to fallback using 'july12' Timer Policy.

Несмотря на то что в кластер был интегрирован узел с более высоким приоритетом, политика таймера была включена и для группы ресурсов не был выполнен немедленный возврат после восстановления. На этапе 2, представленном на рис. 11.4, как видим, группа ресурсов остается на дежурном узле до тех пор, пока не наступит период возврата после восстановления.

5. При наступлении периода времени, на который была настроена политика таймера, вызывается операция перемещения группы ресурсов и группа ресурсов fallback1_rg перемещается обратно на основной узел (**cobra**). Так как основной узел для группы ресурсов был доступен на момент проверки таймера возврата после восстановления (этап 3 на рис. 11.4), было выполнено немедленное перемещение группы ресурсов на этот узел.

Результаты тестирования таймера отсроченного возврата после восстановления

В ходе теста мы доказали, что политика таймера возврата после восстановления была реализована точно и что группа ресурсов не выполняла возврата на основной узел до наступления заданного времени. Следует сообщить об одном нашем наблюдении, связанном с методом определения времени после реализации политики возврата после восстановления при переходе на летнее время. При включении перехода на летнее время на компьютерах определение времени возврата после восстановления было неточным, что отразилось в файле журнала hacmp.out. Мы обнаружили это при мониторинге применения политики таймера возврата после восстановления. Ниже представлен пример обнаруженного нами несоответствия.

#date

Wed Jul 13 17:18:09 EST 2005

#tail -f /tmp/hacmp.out

No action taken on resource group 'fallback1_rg'

The Resource Group 'fallback1_rg' has been configured

to fallback on 'Wed Jul 13 18:25:00 2005'

Кластер сообщил, что возврат после восстановления произойдет в 18:25:00, тогда как мы установили политику таймера возврата после восстановления на 17:25:00, как показано ниже.

```
#odmget HACMPtimer
HACMPtimer:
policy_name = "july13"
recurrence = "once"
year = 105
month = 6
day_of_month = 13
week_day = 0
hour = 17
minutes = 25
```

Тот же тест с отключенной опцией перехода на летнее время работал без несоответствий определения времени.

> **Примечание.** Несоответствие при включенной опции перехода на летнее время было обнаружено в ранней версии кода HACMP 5.3 и будет разрешено в последующих исправлениях для этой версии.
> Между тем, если вы собираетесь установить таймер отсроченного возврата после восстановления при включенной опции перехода на летнее время, проверьте свою среду на подобные несоответствия, чтобы избежать возникновения возврата после восстановления в другое время, отличное от заданного вами.

11.6 Зависимости групп ресурсов

Появившееся в HACMP 5.2 понятие зависимости типа «родительский объект/дочерний объект» (parent/child dependency) для групп ресурсов дает администраторам больше контроля над многоуровневыми приложениями (multi-tiered applications), когда одно приложение зависит от успешного запуска другого.

Такие зависимости могут быть полезны в бизнес-конфигурациях, использующих многоуровневые приложения. Например, в среде, в которой база данных должна быть подключена до сервера приложения, в случае отказа базы данных и ее перемещения на другой узел группа ресурсов, содержащая сервер приложения, также должна быть отключена и перемещена на любой из доступных узлов кластера. При неуспешном перемещении родительской группы ресурсов обе группы ресурсов (родительская и дочерняя) переходят в состояние ERROR и остаются отключенными.

Зависимость расположения, добавленная в HACMP 5.3, дает контроль над типом политики распространения группы ресурсов при ее получении и освобождении.

Ниже описываются новые понятия, связанные с зависимостями.

- **Родительская группа ресурсов.** Во время получения группы ресурсов сначала выполняется получение родительской группы ресурсов.
- **Дочерняя группа ресурсов.** Дочерняя группа ресурсов зависит от родительской и не подключается, если родительская группа ресурсов недоступна. В случае перемещения при сбое или отключения родительской группы ресурсов дочерняя группа ресурсов также отключается и перемещается за родительской группой ресурсов.

- **Зависимость дочернего объекта.** Зависимость дочернего объекта (child dependency) позволяет связывать группы ресурсов в иерархическую структуру до трех уровней в глубину. Получение и освобождение набора групп ресурсов, составляющих зависимость типа «родительский объект/дочерний объект», всегда происходит совместно. Кроме того, можно также настроить зависимость расположения для управления совместным размещением групп ресурсов.
- **Зависимость расположения.** Эта опция, впервые появившаяся в HACMP 5.3, представляет собой расширение для управления группами ресурсов, позволяющее задавать политику, определяющую способ распространения групп ресурсов по узлам во время событий получения и перемещения при сбое. Можно установить совместное размещение набора групп ресурсов на одном узле или же их распределение по узлам кластера.

11.6.1 Зависимость дочернего объекта для группы ресурсов

Реализация зависимости дочернего объекта для группы ресурсов повышает гибкость конфигурирования приложений, зависящих друг от друга. Новым сотрудникам, осуществляющим управление кластером или пытающимся понять конфигурацию кластера, всегда следует сообщать об этих зависимостях, так как использование этой функции изменяет режим работы стандартных политик HACMP.

Можно вывести информацию об установленных зависимостях с использованием команды **clrgdependency**.

Планирование зависимостей дочерних объектов для групп ресурсов

При подготовке к использованию зависимостей групп ресурсов следует учитывать множество различных аспектов конфигурирования.

- Запланируйте, какие группы ресурсов будут содержать те или иные приложения. Убедитесь в том, что приложения, требующие последовательного запуска, расположены в различных группах ресурсов. После разделения приложений можно выполнить создание зависимостей между группами ресурсов.
- Помните о следующих ограничениях:
  - зависимости могут иметь не больше трех уровней в глубину;
  - нельзя задавать циклические зависимости между группами ресурсов.
- Для всех приложений, которые планируется включить в зависимые группы ресурсов, требуется сконфигурировать серверы приложений и мониторы приложений. В целом мы рекомендуем выполнить конфигурирование монитора, проверяющего работу приложения в дочерней группе ресурсов, а также монитора, проверяющего работу приложения в родительской группе ресурсов.

В родительской группе ресурсов также рекомендуется сконфигурировать монитор запуска приложения, чтобы убедиться в успешном выполнении запуска. Это обеспечивает подключение дочерней группы ресурсов после получения родительской группы ресурсов.

- Чтобы свести к минимуму потерю данных в процессе остановки и перезапуска приложений, следует настроить скрипты сервера приложений таким образом, чтобы информация о незавершенных (uncommitted) операциях на время сохранялась на общий диск в процессе остановки приложения и затем заново считывалась в приложение в процессе перезапуска приложения.

Конфигурирование зависимости дочернего объекта для группы ресурсов

Помните о том, что конфигурируемые зависимости:

- определяются явным образом через интерфейс SMIT;
- устанавливаются в масштабе кластера, а не только на локальном узле;
- обязательно учитываются в кластере.

Ниже перечислены действия по конфигурированию зависимости дочернего объекта для группы ресурсов.

1. Введите **smit hacmp**.

2. В SMIT выберите Extended Configuration (Расширенное конфигурирование) > HACMP Extended Resource Configuration (Расширенное конфигурирование ресурсов HACMP) > Configure Resource Group Run-Time Policies (Конфигурирование политик времени выполнения для группы ресурсов) > Configure Dependencies between Resource Groups (Конфигурирование зависимостей между группами ресурсов) > Configure Parent/Child Dependency (Конфигурирование зависимости типа «родительский объект/дочерний объект») > Add Parent/Child Dependency between Resource Groups (Добавить зависимость типа «родительский объект/дочерний объект» между группами ресурсов) и нажмите Enter.

3. Заполните следующие поля:

- **Parent Resource Group (Родительская группа ресурсов).** Выберите родительскую группу ресурсов из списка. Во время получения группы ресурсов HACMP получает родительскую группу ресурсов до получения дочерней группы ресурсов.
- **Child Resource Group (Дочерняя группа ресурсов).** Выберите дочернюю группу ресурсов из списка и нажмите Enter. Во время освобождения HACMP отключает дочернюю группу ресурсов перед родительской группой ресурсов.

  HACMP не позволяет задать циклическую зависимость.

4. Используйте опцию SMIT Verify and Synchronize HACMP Configuration (Верификация и синхронизация конфигурации HACMP), чтобы убедиться в возможности реализации требуемой конфигурации при заданных зависимостях, а также для распространения изменений на другие узлы в кластере.

11.6.2 Зависимость расположения для группы ресурсов

В HACMP 5.3 были реализованы зависимости расположения для групп ресурсов. Их назначение состоит в том, чтобы контролировать расположение подключения зависимых групп ресурсов. Возможные варианты включают размещение всех зависимых ресурсов на одном узле и их размещение на разных узлах. При использовании сайтов можно применить политику совместного размещения групп ресурсов на одном сайте.

Ниже представлены три политики зависимости расположения.

- Online On Same Node (Подключение на одном узле).
- Online On Same Site (Подключение на одном сайте).
- Online On Different Nodes (Подключение на разных узлах).

Возможно совместное использование зависимости дочернего объекта для группы ресурсов и зависимости расположения. При этом можно указать, чтобы набор групп ресурсов всегда подключался на одном узле или чтобы набор групп ресурсов всегда подключался на разных узлах.

Планирование зависимости подключения на одном узле

Для эффективной реализации этой политики нужно помнить следующее:

- Все группы ресурсов, составляющие одну зависимость, должны иметь одинаковый список узлов (содержащий одинаковый порядок участвующих узлов).
- Все группы ресурсов без одновременного доступа, входящие в одну зависимость, должны иметь одинаковые политики запуска/перемещения при сбое/возврата после восстановления:
  - Не допускается использование политики запуска Online Using Node Distribution (Подключение с использованием распределения узлов).
  - При использовании динамического приоритета узлов в качестве политики перемещения при сбое все группы ресурсов в зависимости должны использовать одинаковую политику DNP.
  - Если для одного ресурса установлен таймер возврата после восстановления, его действие распространяется на весь набор групп ресурсов в зависимости. Для всех групп ресурсов в наборе должен быть установлен таймер возврата после восстановления.

> **Примечание.** Обязательно пересмотрите описание планирования для каждой из этих политик, прежде чем пытаться реализовать их в своей конфигурации.

Конфигурирование зависимости подключения на одном узле

Зависимость расположения на одном узле позволяет установить для набора групп ресурсов получение на одном узле. Конфигурирование этой политики осуществляется следующим образом:

1. Введите smit hacmp.

2. В SMIT выберите Extended Configuration (Расширенное конфигурирование) > HACMP Extended Resource Configuration (Расширенное конфигурирование ресурсов HACMP) > Configure Resource Group Run-Time Policies (Конфигурирование политик времени выполнения для группы ресурсов) > Configure Dependencies between Resource Groups (Конфигурирование зависимостей между группами ресурсов) > Configure Online on the same node Dependency (Конфигурирование зависимости подключения на одном узле) > Add Online on the same node Dependency between Resource

Groups (Добавить зависимость подключения на одном узле между группами ресурсов) и выберите группы ресурсов, которые должны входить в этот набор.

Не забудьте убедиться в том, чтобы все списки участвующих узлов в каждой группе ресурсов были одинаковыми; в противном случае эта операция выдаст ошибку и произойдет отказ.

3. Для того чтобы распространить изменения по всем узлам кластера, необходимо выполнить верификацию и синхронизацию кластера.

Планирование зависимости подключения на разных узлах

Для эффективной реализации этой политики необходимо помнить о следующих правилах и ограничениях:

- Допускается реализация только одной зависимости Online On Different Nodes (Подключение на разных узлах) в кластере.
- Каждый набор групп ресурсов должен использовать отдельный домашний узел для запуска.
- При использовании этой политики можно устанавливать три различных значения приоритета:
  - High (Высокий);
  - Intermediate (Средний);
  - Low (Низкий).

Группы ресурсов с более высоким приоритетом имеют предпочтение перед группами ресурсов с более низким приоритетом при запуске, перемещении при сбое и возврате после восстановления:

- Если на узле подключена группа ресурсов с высоким приоритетом, то на этом узле не сможет подключиться ни одна другая группа ресурсов в другом наборе узлов зависимости.
- Если группа ресурсов в этом наборе подключена, но при этом группа ресурсов с более высоким приоритетом выполняет перемещение при сбое или возврат после восстановления на этот узел, то последняя группа ресурсов будет подключена, а группа ресурсов с более низким приоритетом будет отключена или перемещена на другой узел, если это возможно.
- Группы ресурсов с одинаковым приоритетом не могут быть подключены на одном узле. Приоритет групп ресурсов из одного набора, имеющих одинаковый уровень приоритета, определяется по алфавитному порядку групп.
- Группы ресурсов не могут вызвать перенос групп ресурсов с таким же приоритетом в результате перемещения при сбое или возврата после восстановления.
- Если задана зависимость типа «родительский объект/дочерний объект», дочерняя группа ресурсов не может иметь более высокий приоритет, чем родительская группа ресурсов.

Конфигурирование зависимости подключения на разных узлах

Конфигурирование зависимости расположения такого типа осуществляется следующим образом:

1. Введите smit hacmp.

2. В SMIT выберите Extended Configuration (Расширенное конфигурирование) > HACMP Extended Resource Configuration (Расширенное конфигурирование ресурсов HACMP) > Configure Resource Group Run-Time Policies (Конфигурирование политик времени выполнения для группы ресурсов) > Configure Dependencies between Resource Groups (Конфигурирование зависимостей между группами ресурсов) > Configure Online on the same node Dependency (Конфигурирование зависимости подключения на одном узле) > Add Online on Different Nodes Dependency between Resource Groups (Добавить зависимость подключения на разных узлах между группами ресурсов) и нажмите Enter.

3. Заполните следующие поля (и нажмите Enter):

- **High Priority Resource Group(s) (Группы ресурсов с высоким приоритетом).** Выберите группы ресурсов в наборе, получение и подключение которых должно происходить перед группами ресурсов с более низким приоритетом.

  При перемещении при сбое и возврате после восстановления эти группы ресурсов обрабатываются одновременно и подключаются на разных целевых узлах до обработки других групп. Если другие целевые узлы недоступны для перемещения при сбое или возврата после восстановления, эти группы (с одинаковым уровнем приоритета) могут оставаться на одном узле.

  Наивысший относительный приоритет в этом списке имеет группа, указанная первой (слева) в списке узлов.

- **Intermediate Priority Resource Group(s) (Группы ресурсов со средним приоритетом).** Выберите группы ресурсов в наборе, получение и подключение которых должно происходить после групп ресурсов с высоким приоритетом и перед группами ресурсов с низким приоритетом.

  При перемещении при сбое и возврате после восстановления эти группы ресурсов обрабатываются одновременно и подключаются на разных целевых узлах до обработки групп ресурсов с низким приоритетом. Если другие целевые узлы недоступны для перемещения при сбое или возврата после восстановления, эти группы (с одинаковым уровнем приоритета) могут оставаться на одном узле.

  Наивысший относительный приоритет в этом списке имеет группа, указанная первой (слева) в списке узлов.

- **Low Priority Resource Group(s) (Группы ресурсов с низким приоритетом).** Выберите группы ресурсов в наборе, получение и подключение которых должно происходить после групп ресурсов с более высоким приоритетом.

  При перемещении при сбое и возврате после восстановления эти группы ресурсов подключаются на разных целевых узлах после обработки групп ресурсов с более высоким приоритетом. Перемещение групп ресурсов с более высоким приоритетом на узел может вызвать перемещение или отключение этих групп.

4. Продолжайте конфигурирование политик времени выполнения для других групп ресурсов либо выполните верификацию и синхронизацию кластера.

Планирование зависимости подключения на одном сайте

При настройке двух или больше групп ресурсов на использование зависимости расположения эти группы ресурсов относятся к набору, связанному с определенной зависимостью. На зависимости сайтов распространяются следующие правила и ограничения:

- Все группы ресурсов в наборе зависимости подключения на одном сайте должны иметь одинаковую политику межсайтового управления (inter-site management policy), однако могут иметь различные политики запуска/перемещения при сбое/возврата после восстановления. Если используются таймеры возврата после восстановления, они должны быть одинаковыми для всех групп ресурсов в наборе.
- Все группы ресурсов в наборе зависимости подключения на одном сайте должны быть сконфигурированы таким образом, чтобы узлы, которые могут владеть группами ресурсов, были назначены в одних основных и дополнительных сайтах.
- Поддерживается политика Online Using Node Distribution (Подключение с использованием распределения узлов).
- Поддерживается использование групп ресурсов как с одновременным доступом, так и без одновременного доступа.
- В кластере можно устанавливать несколько зависимостей подключения на одном сайте.
- Все группы ресурсов в наборе зависимости подключения на одном сайте, являющиеся активными (в состоянии ONLINE), должны обязательно быть подключены на одном сайте, даже если некоторые группы ресурсов на том же сайте находятся в состоянии OFFLINE или ERROR.
- При добавлении группы ресурсов из набора зависимости подключения на одном узле в набор зависимости подключения на одном сайте необходимо добавить все остальные группы ресурсов из набора зависимости подключения на одном узле в набор зависимости подключения на одном сайте.

Конфигурирование зависимости подключения на одном сайте

Для того чтобы настроить группы ресурсы на использование зависимости подключения на одном сайте, необходимо выполнить следующие действия:

1. Введите smit hacmp.

2. В SMIT выберите Extended Configuration (Расширенное конфигурирование) > HACMP Extended Resource Configuration (Расширенное конфигурирование ресурсов HACMP) > Configure Resource Group Run-Time Policies (Конфигурирование политик времени выполнения для группы ресурсов) > Configure Dependencies between Resource Groups (Конфигурирование зависимостей между группами ресурсов) > Configure Online on the same node Dependency (Конфигурирование зависимости подключения на одном узле) > Add Online on the same Site dependency between Resource Groups (Добавить зависимость подключения на одном сайте между группами ресурсов) и нажмите Enter.

3. Выберите из списка группы ресурсов, которые следует включить в этот набор. При получении эти группы ресурсов будут подключены на одном сайте в соответствии с политикой запуска сайта и узла, заданной в группе ресурсов. При перемещении при сбое или возврате после восстановления группы ресурсов обрабатываются одновременно и подключаются на одном сайте.

4. Выполните верификацию и синхронизацию кластера.

11.6.3 Ограничения на сочетание зависимостей

Существуют следующие ограничения конфигураций совмещения зависимостей. Верификация выдаст ошибку, если не следовать следующим инструкциям:

- только одна группа ресурсов может одновременно входить в набор зависимости подключения на одном узле и в набор зависимости подключения на разных узлах;.
- если группа ресурсов одновременно входит в набор зависимости подключения на одном узле и в набор зависимости подключения на разных узлах, все узлы в наборе зависимости подключения на одном узле будут иметь такой же приоритет, как и общая группа ресурсов;
- только группы ресурсов с одинаковым приоритетом в наборе зависимости подключения на разных узлах могут входить в набор зависимости подключения на одном сайте.

11.6.4 Вывод зависимостей группы ресурсов

Если при получении уже существующего кластера HACMP требуется вывести существующие зависимости, можно использовать команду **clrgdependency**. Ниже приведено несколько примеров возможных выходных данных.

```
#clrgdependency -?
usage: clrgdependency -t <ANTICOLLOCATION> -u -hp <high priority RG list> -ip
<intermediate priority RG list> -lp <low priority RG list>
usage: clrgdependency -t <PARENT_CHILD | NODECOLLOCATION | SITECOLLOCATION |
ANTICOLLOCATION > -sl
#clrgdependency -t PARENT_CHILD -sl
        #Parent                         Child
        DB2_1Rg                         Child_1Rg
        DB2_2Rg                         Child_2Rg
```

Во время создания этой публикации не существовало страницы электронного руководства для команды **clrgdependency**.

Альтернативный вариант состоит в том, чтобы просмотреть разделы (stanzas) ODM-классов HACMPrg_loc_dependency и HACMPrgdependency с помощью команды odmget:

```
#odmget HACMPrgdependency
HACMPrgdependency:
```

id = 0
group_parent = “DB2_1Rg”
group_child = “Child_1Rg”
dependency_type = “PARENT_CHILD”
dep_type = 0

> **Внимание.** Использование информации, получаемой непосредственно из ODM, предназначено только для информационных целей, так как формат разделов (stanzas) может быть различным в разных обновлениях и/или в новых версиях. Таким образом, жесткое кодирование ODM-запросов в пользовательских приложениях не поддерживается и его следует избегать.

11.6.5 Сценарий тестирования зависимостей групп ресурсов

При использовании зависимостей групп ресурсов существует множество возможных комбинаций администрирования групп ресурсов. Для того чтобы протестировать некоторые из них, мы установили конфигурацию, использующую комбинацию различных доступных политик. Заметьте, что применяемая нами конфигурация выходит за рамки стандартного режима работы HACMP и предполагает, что вы знакомы с принципом работы зависимостей группы ресурсов.

Мы решили использовать кластер из трех узлов, включающий несколько экземпляров баз данных DB2, где дочерние группы ресурсов содержат приложения WebSphere. Мы сделали третий узел дежурным узлом, содержащим приложение Tivoli для создания резервных копий, и группу ресурсов, содержащую тестируемое приложение. Используемая конфигурация представлена на рис. 11.5:

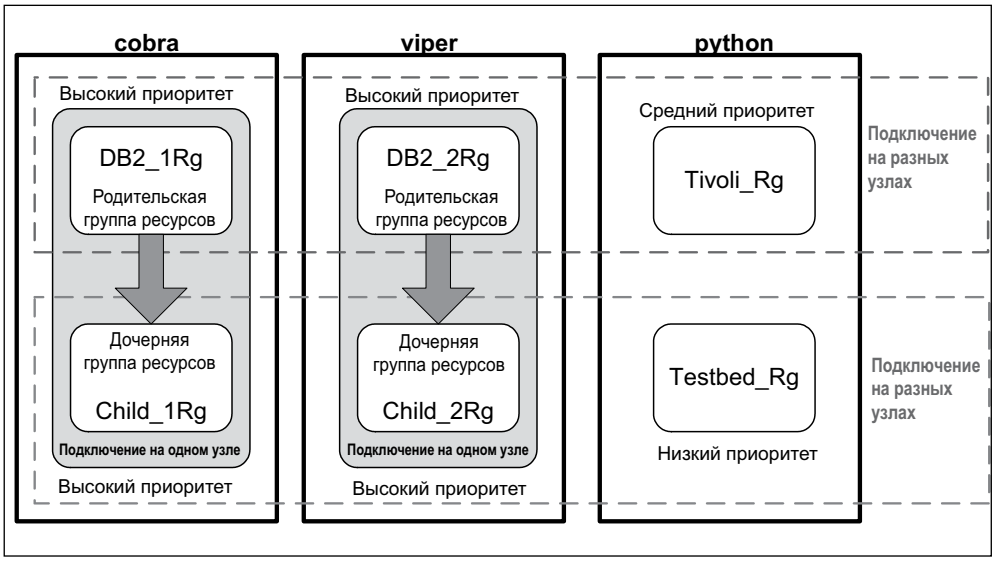

Рис. 11.5. Сценарий тестирования зависимости группы ресурсов

Мы реализовали шесть различных зависимостей, включая: зависимость «родительский объект/дочерний объект» (2 набора), зависимость подключения на одном узле (2 набора) и зависимость подключения на разных узлах (2 набора). Эти зависимости также описаны в табл. 11.4. Мы реализовали конфигурацию, в которой родительская группа ресурсов всегда подключается на одном узле с соответствующей дочерней группой ресурсов, но родительские группы ресурсов и группа ресурсов Tivoli всегда подключаются на разных узлах, что позволяет распределить нагрузку. Мы включили зависимость расположения на разных узлах для дочерних групп ресурсов и группы ресурсов Testbed_Rg, чтобы в случае перемещения рабочих ресурсов на дежурный узел **python** группа ресурсов с низким приоритетом, содержащая тестовое приложение, была отключена.

В табл. 11.5 описываются различные атрибуты групп ресурсов и политики запуска/перемещения при сбое/возврата после восстановления, используемые в наших группах ресурсов. Если вы незнакомы с сокращениями, указываемыми для каждой

Таблица 11.4. Сценарий тестирования политик зависимостей и распространения групп ресурсов

	Группы ресурсов, определенные в кластере					
Политика	**DB2_1Rg**	**Child_1Rg**	**DB2_2Rg**	**Child_2Rg**	**Testbed_Rg**	**Tivoli_Rg**
Родительский объект/дочерний объект	Да	Да				
Подключение на одном узле	Да	Да				
Родительский объект/дочерний объект			Да	Да		
Подключение на одном узле			Да	Да		
Подключение на разных узлах		Высокий		Высокий	Низкий	
Подключение на разных узлах	Высокий		Высокий			Средний

Таблица 11.5. Атрибуты групп ресурсов, используемых при тестировании

	Группы ресурсов, определенные в кластере					
Атрибуты	**DB2_1Rg**	**Child_1Rg**	**DB2_2Rg**	**Child_2Rg**	**Testbed_Rg**	**Tivoli_Rg**
Запуск	OHNO	OHNO	OHNO	OHNO	OHNO	OHNO
Перемещение при сбое	FNPNL	FNPNL	FNPNL	FNPNL	FNPNL	FNPNL
Возврат после восстановления	NF	NF	NF	NF	NF	NF
Участвующие узлы	cobra, python, viper	cobra, python, viper	viper, python, cobra	viper, python, cobra	python, cobra, viper	python, cobra, viper
Сервисная IP-метка	app1svc		app2svc		app3svc	app4svc
Сервер приложения	db2_1	db2_child1	db2_2	db2_child2	testbed_app	tivoli_app

политики группы ресурсов, просмотрите в этом руководстве раздел, посвященный терминам. Также мы включили участвующие узлы, сервисные IP-метки и серверы приложений, используемые в каждой группе ресурсов.

Операции и результаты сценария тестирования

При тестировании мы пытались воссоздать отказы и операции групп ресурсов, имеющие место в рабочей среде. Мы обнаружили, что методами, фактически применяющими политики зависимостей, являются завершение работы узла или постепенная остановка с передачей ресурсов (graceful stop with takeover). Так как в нашей среде на всех узлах были расположены группы ресурсов с высоким приоритетом, выполнение операций перемещения групп ресурсов не разрешалось; это будет рассмотрено ниже при описании результатов тестирования.

Наше тестирование содержало следующие действия:

1. Мы выполнили постепенную остановку HACMP с передачей ресурсов на узел **viper**.

После остановки служб кластера с передачей ресурсов группы ресурсов были успешно перемещены на дежурный узел **python**. Была применена политика размещения на разных узлах для дочерних групп ресурсов и группы ресурсов Testbed_Rg, и тестовая группа ресурсов была переведена в состояние OFFLINE. Группа ресурсов Tivoli_Rg была оставлена в состоянии ONLINE, так как был установлен средний приоритет. Выходные данные команды **clRGinfo** содержали следующее:

```
#/usr/es/sbin/cluster/utilities/clRGinfo
DB2_2Rg OFFLINE viper
                ONLINE                          python
                OFFLINE                         cobra
Child_2Rg       OFFLINE                         viper
                ONLINE                          python
                OFFLINE                         cobra
Testbed_Rg      OFFLINE due to lack of node     python
                ERROR                           cobra
                OFFLINE                         viper
Tivoli_Rg       ONLINE                          python
                OFFLINE                         cobra
                OFFLINE                         viper
```

2. Мы попытались снова подключить группу ресурсов Testbed_Rg. В этом состоянии мы попытались подключить тестовую группу ресурсов через меню HACMP. Мы обнаружили, что не было доступных узлов для подключения группы ресурсов. Затем мы попытались выбрать опцию Restore_Node_Priority_Order, чтобы подключить группу ресурсов; эта команда выполнилась без ошибок, однако, как и ожидалось, не было самого действия.

3. Мы выполнили реинтеграцию узла **viper** в кластер. Мы интегрировали узел обратно в кластер, и группа ресурсов Testbed_Rg была снова подключена на нем.

Примечание. Это произошло потому, что группа ресурсов Testbed_Rg имела низкий приоритет и не была расположена на каком-либо узле.

4. Мы попытались переместить группу ресурсов DB2_2Rg с узла **python** обратно на первоначальный узел **viper**.

Хотя группа ресурсов находилась в состоянии после перемещения при сбое, мы попытались переместить первоначальные ресурсы узла **viper** обратно на домашний узел через меню HACMP. Эта операция не показывала, что этот узел доступен для выполнения перемещения. Единственный способ, который позволил нам переместить группу ресурсов обратно на узел **viper**, состоял в постепенном отключении узла **python** с передачей ресурсов.

5. Мы выполнили реинтеграцию узла **python** в кластер.

Мы обратили внимание на задержку при реинтеграции этого узла обратно в кластер, однако в целом выполнение операции прошло успешно, и группа ресурсов Testbed_Rg возвратилась на узел **python**. Во время задержки в файлы hacmp.out и clstrmgr.debug ничего не было записано. Группа ресурсов Tivoli_Rg осталась подключенной на узле **viper**, и мы решили оставить ее на этом узле.

В целом тестирование зависимостей групп ресурсов прошло успешно. Основным нашим наблюдением было то, что операции перемещения групп ресурсов были недоступны для всех групп ресурсов после реинтеграции узла, для которого прежде было выполнено перемещение при сбое. Это распространялось на группы ресурсов с высоким и низким приоритетом.

Во время тестирования мы не смогли протестировать операции DARE для изменений в политиках зависимостей групп ресурсов из-за отсутствия соответствующих функций. Однако поддержка динамических изменений в политиках зависимостей групп ресурсов будет реализована в последующих обновлениях этой версии.

12

Настройка событий

Эта глава содержит информацию о настраиваемых событиях и использовании функции уведомления об ошибках в AIX для обнаружения аппаратных или программных ошибок вне области действия HACMP. Мы рассмотрим два способа обнаружения и обработки ошибок, не связанные с кластером:

- **Команды пред- и постобработки событий HACMP.** Предназначены для управления дополнительными аспектами, которые HACMP не обрабатывает по умолчанию.
- **Уведомление об ошибках HACMP.** Использует функцию журналов ошибок AIX для перехвата ошибок и соответствующего реагирования.

12.1 Написание скриптов для настраиваемых событий

Настраиваемые события HACMP и средство уведомления об ошибках требуют написания скриптов.

Необходимо учитывать следующее:

- нужно протестировать все возможные входные параметры;
- надо протестировать условные переходы, т. е. все переходы if и case;
- необходимо выполнить обработку кодов завершения и возвращаемых значений для всех внешних команд;
- следует обеспечить правильное возвращаемое значение: 0 для успешного выполнения, любое другое число для неуспешного выполнения;
- выполнение скрипта должно занимать разумное количество времени;
- необходимо протестировать максимально возможное количество сценариев;
- при отказе скрипта происходит отказ кластера;
- программа восстановления должна быть способна выполнить восстановление после отказа события, в противном случае произойдет отказ кластера;
- нужно сохранять свои скрипты в отдельном каталоге, например в каталоге /usr/ha.

Важно! Кластер не будет продолжать обработку событий, пока не завершится выполнение настраиваемых скриптов пред- и постобработки событий.

12.2 Команды пред- и постобработки событий HACMP

Конфигурация может содержать специальные аппаратные и программные компоненты, для которых HACMP не сможет осуществлять управление по умолчанию. В таких случаях можно добавлять собственные скрипты пред- и постобработки событий. Типичным применением является управление удаленным копированием с диска на диск на подсистемах хранения сторонних производителей.

Для всех предопределенных событий HACMP можно определить скрипт предобработки, скрипт постобработки, метод уведомления и команду восстановления:

- **Скрипт предобработки событий.** Выполняется перед выполнением события HACMP.
- **Скрипт постобработки событий.** Выполняется после выполнения события HACMP.
- **Метод уведомления.** Выполняется до и после события HACMP. Он отправляет системному администратору сообщение о начале или завершении события.
- **Команда восстановления.** Выполняется только при отказе события HACMP. При успешном выполнении метода восстановления событие также будет считаться успешным, независимо от результата скрипта обработки события[1].

[1] Здесь ошибка: при успешном выполнении команды восстановления основной скрипт события запускается повторно. См. также Recovery Counter (ниже). *Примеч. науч. ред.*

HACMP передает несколько аргументов в скрипты событий, которые вы можете использовать в своей программе. Табл. 12.1 содержит аргументы скриптов пред- и постобработки событий.

Таблица 12.1. Аргументы скриптов пред- и постобработки событий

Скрипт	Аргументы
Предобработки события	Имя события, завершающие аргументы
Постобработки события	Имя события, код завершения события, завершающие аргументы
Метод уведомления	Имя события, ключевое слово: start или complete; код завершения (для ключевого слова complete), завершающие аргументы
Команда восстановления	Имя события, завершающие аргументы

Завершающие аргументы. Аргументы предопределенного скрипта обработки события HACMP. Они определяются в заголовке командного файла события. Предопределенные командные файлы событий HACMP хранятся в каталоге /usr/es/sbin/cluster/events. Например, просмотрим файл события swap_adapter (пример 12.1):

Пример 12.1. Файл события /usr/es/sbin/cluster/events/swap_adapter

```
/usr/es/sbin/cluster/events# more swap_adapter
#                                                                    #
#       Arguments:      nodename network ip_address1 ip_address2     #
#                                                                    #
#                       ip_address1 - the new available address      #
#                       this script swaps the service adapter to     #
#                       ip_address2 - the failed address             #
```

Как видим, событие имеет следующие аргументы:

- nodename – имя узла;
- network – имя сети;
- ip_address1 – новый доступный адрес;
- ip_address2 – отказавший адрес.

Те же завершающие аргументы передаются в пользовательские скрипты пред- и постобработки событий.

> **Внимание!** Не изменяйте встроенные файлы событий: это не поддерживается и небезопасно. Всегда используйте скрипты пред- и постобработки событий.

Выбор события

По умолчанию HACMP обрабатывает группы ресурсов параллельно, что уменьшает количество событий кластера. Некоторые события выполняются только один раз, тогда как другие вообще не выполняются.

Некоторые события выполняются, только если для группы ресурсов установлен последовательный режим обработки. Мы рекомендуем выполнить несколько тестов

запуска/остановки кластера и перемещения при сбое и просмотреть файл /tmp/hacmp.out. Это даст вам представление о конфигурации/сценарии, в которых выполняется событие, которое вы собираетесь изменить.

12.2.1 Установка скриптов пред- и постобработки

Мы покажем, как установить скрипт пред- и постобработки событий, на примере. У нас имеется дисковая подсистема стороннего производителя, которая требует дополнительного конфигурирования, прежде чем мы сможем активизировать группу томов. Нам требуется выполнить скрипт предобработки события **get_disk_vg_fs**. Также нужно отправить уведомление системному администратору при запуске или завершении события.

> **Замечание.** HACMP поддерживает конфигурирование скриптов пред-/постобработки событий в подключенном состоянии.

1. Напишите и тщательно протестируйте скрипт обработки события. Скопируйте файлы на все узлы с использованием одинакового пути и имени. Рекомендуется также поместить созданные файлы под контроль наборов файлов HACMP (5.2 и выше).

В нашем примере мы создали два shell-скрипта: скрипт предобработки события (/usr/ha/pre_get_disk_vg_fs, см. пример 12.2) для выполнения конфигурирования OEM-дисков и команды уведомления (/usr/ha/notify_get_disk_vg_fs, пример 12.3).

Пример 12.2. Образец скрипта предобработки события (/usr/ha/pre_get_disk_vg_fs)

```
#!/bin/ksh
# Checking the arguments
# Event name
EVENT=$1
# Check the event name
if [ "$EVENT" != "get_disk_vg_fs" ]
then
        echo Ops! We were not supposed to run in event $EVENT!
        exit 0
fi
# Run the OEM disk setup command
/oem_setup_disk
if [ $ ?!= 0 ]
then
        echo Error setting up OEM disk!
        exit -1
fi
exit 0
```

Этот скрипт выполняет проверку аргументов события, после чего вызывает программу конфигурирования дисковой подсистемы стороннего производителя.

Пример 12.3. Образец метода уведомления (/usr/ha/notify_get_disk_vg_fs)

```
#!/bin/ksh
# Sample event notification command
# Processing the arguments
# Event name
EVENT=$1
#Event status: start or complete
STATUS=$2
# Result of the event, if status=complete
RESULT=$3
# Notify root that the has event started
if [ "$STATUS" == "start" ]
then
mail -s "Event notification" root <<EOF1
Event $EVENT started.
EOF1
fi
# Notify the root that the event has completed and send the return code
if [ "$STATUS" == "complete" ]
then
mail -s "Event notification" root <<EOF2
Event $EVENT completed, the result is $RESULT
EOF2
fi
exit 0
```

Этот образец программы уведомления отправляет сообщение электронной почты пользователю root с кодом завершения события get_disk_vg_fs.

2. Определите команды пред- и постобработки события:

а) Выполните **smit hacmp**.

б) Выберите **Extended Configuration (Расширенное конфигурирование)**.

в) Выберите **Extended Event Configuration (Расширенное конфигурирование событий)**.

г) Выберите **Configure Pre/Post-Event Commands (Конфигурирование команд пред- и постобработки событий)**.

д) Выберите **Add a Custom Cluster Event (Добавление настраиваемого события кластера)**.

е) Введите следующую информацию:

- **Cluster Event Name (Имя события кластера).** Короткое имя события; в дальнейшем используется это имя, а не имя файла. Наш образец предобработки события имеет название pre_get_disk_vg_fs.
- **Cluster Event Description (Описание события кластера).** Краткое описание вашего скрипта.
- **Cluster Event Script Filename (Имя файла скрипта обработки события кластера).** Имя файла скрипта пред-/постобработки события с полным путем. См. экран SMIT на рис. 12.1.

```
                          Add a Custom Cluster Event

  Type or select values in entry fields.
  Press Enter AFTER making all desired changes.

                                                [Entry Fields]
  * Cluster Event Name                          [pre_get_disk_vg_fs]
  * Cluster Event Description                   [Configure OEM disk]
  * Cluster Event Script Filename               [/usr/ha/pre_get_disk_vg_fs]



  F1=Help             F2=Refresh          F3=Cancel           F4=List
  F5=Reset            F6=Command          F7=Edit             F8=Image
  F9=Shell            F10=Exit            Enter=Do
```

Рис. 12.1. Добавление настраиваемого события кластера

Здесь можно определить только скрипты пред- и постобработки событий. Скрипты уведомления и восстановления должны определяться отдельно через панель SMIT **Change/Show Pre-Defined HACMP Events (Изменить/вывести предопределенные события HACMP)**. Дополнительные сведения см. на этапе 3.

3. Свяжите скрипт пред-/постобработки с предопределенным событием HACMP:

а) Запустите **smit hacmp**.

б) Выберите **Extended Configuration (Расширенное конфигурирование)**.

в) Выберите **Extended Event Configuration (Расширенное конфигурирование событий)**.

г) Выберите **Change/Show Pre-Defined HACMP Events (Изменить/вывести предопределенные события HACMP)**.

д) Выберите событие, которое требуется изменить (рис. 12.2).

е) Введите следующие значения (рис. 12.3):

- **Notify Command (Команда уведомления).** (Необязательно). Полное имя (включая путь) скрипта уведомления, если он имеется. В нашем примере используется /usr/ha/notify_get_disk_vg_fs.
- **Pre-event Command (Команда предобработки события).** (Необязательно). Имя настраиваемого события кластера, которое требуется выполнить в качестве скрипта предобработки события. Нажмите F4 для вывода списка уже определенных настраиваемых событий кластера. В нашем примере используется pre_get_disk_vg_fs.
- **Post-event Command (Команда постобработки события).** (Необязательно). Имя настраиваемого события кластера, которое требуется выполнить в качестве скрипта постобработки события. Нажмите F4 для вывода списка уже определенных настраиваемых событий кластера.

```
 Extended Event Configuration

Mo+-------------------------------------------------------------------------+
  |                        Select Event Name to Change                      |
  |                                                                         |
  | Move cursor to desired item and press Enter.                            |
  |                                                                         |
  | [MORE...4]                                                              |
  |   config_too_long                                                       |
  |   event_error                                                           |
  |   external_resource_state_change                                        |
  |   external_resource_state_change_complete                               |
  |   fail_interface                                                        |
  |   fail_standby                                                          |
  |   get_aconn_rs                                                          |
  |   get_disk_vg_fs                                                        |
  |   intersite_fallover_prevented                                          |
  |   join_interface                                                        |
  | [MORE...68]                                                             |
  |                                                                         |
  | F1=Help                 F2=Refresh              F3=Cancel               |
  | F8=Image                F10=Exit                Enter=Do                |
F1| /=Find                  n=Find Next                                     |
F9+-------------------------------------------------------------------------+
```

Рис. 12.2. Выбор изменяемого события

```
                         Change/Show Cluster Events

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                [Entry Fields]
  Event Name                                    get_disk_vg_fs

  Description                                   Script run to acquire>

* Event Command                                [/usr/es/sbin/cluster/>

  Notify Command                                [/usr/ha/notify_get_disk_vg_fs]
  Pre-event Command                             [pre_get_disk_vg_fs]    +
  Post-event Command                            []                      +
  Recovery Command                              []
* Recovery Counter                              [0]                      #

F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Рис. 12.3. Изменение предопределенного события кластера

- **Recovery Command (Команда восстановления).** (Необязательно). Полное имя (включая путь) скрипта восстановления.
- **Recovery Counter (Счетчик восстановления).** Количество попыток выполнения команды восстановления. По умолчанию установлено значение 0. Если установлено значение больше нуля, то при успешном выполнении команды восстановления событие запускается повторно.

4. Выполните верификацию и синхронизацию кластера.

> **Замечание.** При настройке автоматического обновления набора файлов HACMP_Files автоматически выполняется распространение настраиваемых файлов событий по узлам кластера. Дополнительные сведения о наборах файлов см. в разделе 8.2, «Наборы файлов».

12.3 Уведомление об ошибках

Средство уведомления об ошибках в HACMP является отличным инструментом мониторинга приложений и устройств, поддерживающим и использующим журналы ошибок AIX. При возникновении ошибки в приложении или устройстве, выполняется отправка сообщения в демон ведения журнала ошибок (error logger daemon, errdemon) с соответствующей информацией: датой, именем устройства или приложения, типом ошибки, описанием ошибки и прочей отладочной информацией. Демон errdemon записывает эти данные в файл журнала ошибок. Если уведомление об ошибках в HACMP настроено на этот тип ошибок, errdemon выполняет определенный пользователем метод уведомления/восстановления.

12.3.1 Автоматическое уведомление об ошибках

HACMP может сконфигурировать уведомление об ошибках и действие восстановления для некоторых типов ресурсов и ошибок:

- диски Rootvg;
- все диски, определенные в группе ресурсов HACMP;
- SCSI-адаптеры, используемые ресурсами HACMP или rootvg;
- адаптеры Fibre Channel, используемые ресурсами HACMP;
- адаптер SP Switch.

Для серьезных, необратимых ошибок настраивается инициирование перехвата с использованием команды **/usr/es/sbin/cluster/diag/cl_failover**. При некритических ошибках **/usr/es/sbin/cluster/diag/cl_logerror** отправляет сообщение электронной почты пользователю root и записывает ошибку в файл /tmp/hacmp.out.

Рассмотрение мониторинга дисков

HACMP осуществляет мониторинг только для «традиционных» автономных дисков, например для встроенных SCSI-дисков или SSA-хранилищ. При обнаружении ошибок на дисках этого типа используются стандартные метки журнала ошибок AIX, такие,

как DISK_ERR1 или SCSI_ERR3. Кроме того, HACMP может осуществлять мониторинг групп томов с полным зеркальным отображением, независимо от типа диска. В этом случае при обнаружении потери кворума (запись LVM_SA_QUORCLOSE в журнале ошибок) HACMP может инициировать перемещение.

ESS, FAStT, DS4000 Series и другие современные дисковые подсистемы имеют встроенные средства обеспечения высокой доступности, так что они содержат встроенные механизмы восстановления после ошибок, например аппаратные функции RAID и multipath-подключения. Так как они имеют собственные драйверы устройств, при возникновении аппаратных ошибок они используют другие метки журнала ошибок. В разделе 12.3.3, «Мониторинг общих дисков с уведомлением об ошибках HACMP», представлен пример мониторинга такой дисковой подсистемы.

Настройка автоматического уведомления об ошибках

> **Внимание.** Нельзя сконфигурировать автоматическое уведомление об ошибках при работающем кластере.

1. Запустите **smit hacmp**.

2. Выберите **Problem Determination Tools (Инструменты определения проблем)**.

3. Выберите **HACMP Error Notification (Уведомления об ошибках HACMP)**.

4. Выберите **Configure Automatic Error Notification (Конфигурирование автоматического уведомления об ошибках)**.

5. Выберите **Add Error Notify Methods for Cluster Resources (Добавить методы уведомления об ошибках для ресурсов кластера)**.

HACMP добавляет автоматическое уведомление об ошибках на всех узлах.

> **Внимание.** Необходимо повторно добавлять автоматическое уведомление об ошибках при каждой верификации и синхронизации кластера.

Вывод методов автоматического уведомления об ошибках

1. Запустите **smit hacmp**.

2. Выберите **Problem Determination Tools (Инструменты определения проблем)**.

3. Выберите **HACMP Error Notification (Уведомления об ошибках HACMP)**.

4. Выберите **Configure Automatic Error Notification (Конфигурирование автоматического уведомления об ошибках)**.

5. Выберите **List Error Notify Methods for Cluster Resources (Вывести методы уведомления об ошибках для ресурсов кластера)**.

На рис. 12.4 представлен образец вывода методов автоматического уведомления об ошибках.

```
                                COMMAND STATUS

Command: OK            stdout: yes           stderr: no

Before command completion, additional instructions may appear below.

p650n01:
p650n01: HACMP Resource         Error Notify Method
p650n01:
p650n01: hdisk0         /usr/es/sbin/cluster/diag/cl_failover
p650n01: scsi0          /usr/es/sbin/cluster/diag/cl_failover
p650n01: fcs0           /usr/es/sbin/cluster/diag/cl_logerror
p650n02:
p650n02: HACMP Resource         Error Notify Method
p650n02:
p650n02: hdisk0         /usr/es/sbin/cluster/diag/cl_failover
p650n02: scsi0          /usr/es/sbin/cluster/diag/cl_failover
p650n02: fcs0           /usr/es/sbin/cluster/diag/cl_logerror

F1=Help             F2=Refresh          F3=Cancel           F6=Command
F8=Image            F9=Shell            F10=Exit            /=Find
n=Find Next
```

Рис. 12.4. Список методов автоматического уведомления об ошибках

Удаление методов автоматического уведомления об ошибках

1. Запустите **smit hacmp**.

2. Выберите **Problem Determination Tools (Инструменты определения проблем)**.

3. Выберите **HACMP Error Notification (Уведомления об ошибках HACMP)**.

4. Выберите **Configure Automatic Error Notification (Конфигурирование автоматического уведомления об ошибках)**.

5. Выберите **Remove Error Notify Methods for Cluster Resources (Удалить методы уведомления об ошибках для ресурсов кластера)**.

6. Нажмите Enter для подтверждения.

12.3.2 Использование уведомления об ошибках

Можно вручную добавить объекты уведомления об ошибках для всех возможных типов ресурсов и ошибок. При создании объекта уведомления определяется набор правил или критериев типа ошибки, для которого следует осуществлять мониторинг. Эти правила могут включать метку для журнала ошибок, имя ресурса и т. д. При возникновении ошибки errlogger сопоставляет критерии, заданные в ODM-классе уведомления об ошибках, и в случае соответствия errdemon выполняет требуемый метод уведомления.

Добавление метода уведомления

1. Запустите **smit hacmp**.

2. Выберите **Problem Determination Tools (Инструменты определения проблем)**.

3. Выберите **HACMP Error Notification (Уведомления об ошибках HACMP)**.

4. Выберите **Add a Notify Method (Добавить метод уведомления)**.

5. Определите объект уведомления:

- **Notification Object Name (Имя объекта уведомления).** Заданное пользователем имя, идентифицирующее объект уведомления об ошибках.
- **Persist across system restart? (Сохранять после перезапуска системы?). Yes** – уведомление об ошибках будет использоваться постоянно; **No** – уведомление об ошибках будет использоваться только до следующей перезагрузки.
- **Process ID for use by Notify Method (Идентификатор процесса, используемый методом уведомления).** Уведомление об ошибках отправляется от имени выбранного идентификатора процесса. По умолчанию применяется 0 (root)[1]; мы рекомендуем оставить здесь 0. При использовании ненулевого кода процесса следует установить для параметра Persist across system restart? (Сохранять после перезапуска системы?) значение No.
- **Select Error Class (Выбор класса ошибки). None** – класс ошибок для сопоставления отсутствует; **All** – выполнять сопоставление для всех классов ошибок, **Hardware** – аппаратные ошибки, **Software** – программные ошибки, **Errlogger** – уведомления и сообщения оператора из программы **`errlogger`**.
- **Select Error Type (Выбор типа ошибки). None** – тип ошибок для сопоставления отсутствует, **All** – выполнять сопоставление для всех типов ошибок, **PEND**– предстоящая потеря доступности, **PERF** – снижение производительности, **PERM** – постоянные ошибки, **TEMP** – временные ошибки, **UNKN** – неизвестный тип ошибки.
- **Match Alertable errors? (Сопоставлять ошибки с возможностью оповещения?).** Это поле предназначено для использования агентами оповещения приложений управления системами. **None** – игнорировать этот параметр, **All** – оповещение для всех ошибок, **TRUE** – сопоставлять ошибки с возможностью оповещения, **FALSE** – сопоставлять ошибки без возможности оповещения. Если у вас нет приложения удаленного управления, оставьте в этом поле значение **None**.
- **Select Error Label (Выбор метки ошибки).** Нажмите F4, чтобы выбрать метку ошибки. Краткое описание меток ошибок см. в файле /usr/include/sys/errids.h.
- **Resource Name (Имя ресурса).** Имя отказавшего ресурса. Для класса аппаратных ошибок представляет имя устройства. Для класса программных ошибок

---

[1] Здесь ошибка – имеется в виду PID процесса (Process ID), используемый для метода уведомления. Как правило, это поле оставляется пустым (выбирается свободный, произвольный PID). При включенном флаге «Persist across system restart» оно должно быть оставлено пустым. *Примеч. науч. ред.*

представляет имя отказавшего исполняемого файла. Выберите **All** для сопоставления всех типов ресурсов.

- **Resource Class (Класс ресурса).** Для класса аппаратного ресурса представляет класс устройства. Неприменимо к программным ошибкам. Выберите **All** для сопоставления всех классов ресурсов.
- **Resource Type (Тип ресурса).** Тип устройства, под которым ресурс известен в объекте устройств. Применимо только к аппаратным ошибкам. Выберите **All** для сопоставления всех классов ресурсов.
- **Notify Method (Метод уведомления).** Полное имя (включая путь) программы для запуска при регистрации ошибки, соответствующей критериям, определенным выше. В исполняемый файл можно передать следующие переменные:

  $1 – порядковый номер ошибки в журнале;

  $2 – идентификатор ошибки;

  $3 – класс ошибки;

  $4 – тип ошибки;

  $5 – флаг оповещения;

  $6 – имя ресурса отказавшего устройства;

  $7 – тип ресурса отказавшего устройства;

  $8 – класс ресурса отказавшего устройства;

  $9 – метка для регистрации в журнале ошибок.

```
                          Add a Notify Method

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                      [Entry Fields]
* Notification Object Name                           [disk_error]
* Persist across system restart?                     No +
  Process ID for use by Notify Method                [] +#
  Select Error Class                                  Hardware +
  Select Error Type                                   PERM +
  Match Alertable errors?                             None +
  Select Error Label                                 [VPATH_PATH_REMOVED] +
* Resource Name                                      [All] +
* Resource Class                                     [All] +
* Resource Type                                      [All] +
* Notify Method                                      [/usr/ha/vpath_error]

F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Рис. 12.5. Добавление объекта уведомления об ошибках

6. Нажмите Enter для создания объекта уведомления об ошибках.

На рис. 12.5 показано, как добавить метод уведомления через SMIT.

Изменение/вывод метода уведомления

1. Запустите **smit hacmp**.

2. Выберите **Problem Determination Tools (Инструменты определения проблем)**.

3. Выберите **HACMP Error Notification (Уведомления об ошибках HACMP)**.

4. Выберите **Change/Show a Notify Method (Изменение/вывод метода уведомления)**.

5. Выберите изменяемый метод уведомления из выпадающего списка.

6. Измените объект уведомления. Описание полей см. в разделе «Добавление метода уведомления».

Удаление метода уведомления

1. Запустите **smit hacmp**.

2. Выберите **Problem Determination Tools (Инструменты определения проблем)**.

3. Выберите **HACMP Error Notification (Уведомления об ошибках HACMP)**.

4. Выберите **Remove a Notify Method (Удаление метода уведомления)**.

5. Выберите удаляемый метод уведомления из выпадающего списка.

6. Нажмите Enter для удаления.

Тестирование метода уведомления

Вы можете легко протестировать свои объекты уведомления. HACMP может эмулировать запись в журнале ошибок с использованием заданной меткой ошибки. Метка ошибки выводится в журнале ошибок, и метод уведомления запускается демоном errdemon.

1. Запустите **smit hacmp**.

2. Выберите **Problem Determination Tools (Инструменты определения проблем)**.

3. Выберите **HACMP Error Notification (Уведомления об ошибках HACMP)**.

4. Выберите **Emulate Error Log Entry (Эмуляция записи в журнал ошибок)**.

5. Выберите метку ошибки или метод уведомления из выпадающего списка. Здесь выводятся только те методы уведомления, для которых определены метки ошибок.

6. SMIT выводит метку ошибки, имя объекта уведомления и метод уведомления. Нажмите Enter для подтверждения эмуляции записи в журнал ошибок.

12.3.3 Мониторинг общих дисков с использование функции уведомления об ошибках HACMP

Здесь на конкретном примере мы покажем, как можно использовать функцию уведомления об ошибках для мониторинга общих дисков. Этот пример основан на нашей аппаратной конфигурации. Вы можете легко изменить этот пример для работы с вашей дисковой подсистемой.

Наши общие диски находятся в дисковой подсистеме ESS 2105-800 с подключением FC. Мы используем драйвер SDD, группы томов располагаются на дисковых устройствах vpath. Наша цель состоит в осуществлении мониторинга устройств vpath. При отказе ссылки vpath (hdiskX) требуется отправить уведомление системному администратору. При отказе дискового устройства vpath (vpathX) требуется выполнить перехват HACMP.

В нашем примере у нас есть общая группа томов под названием db2vg. Она расположена на устройстве vpath1, имеющем два базовых пути: hdisk3 и hdisk11 (один LUN, через два разных HBA).

Выбор метки ошибки

Сначала посмотрим, что произойдет, если на несколько минут отключить кабели Fibre Channel из адаптеров. Конечно же, ввод-вывод останавливается и драйвер SDD регистрирует ошибки в журнале ошибок AIX.

Теперь посмотрим на содержимое журнала ошибок (командой **errpt**); (пример 12.4).

> **Важно!** Не рекомендуется выполнять этот тест преднамеренно, особенно при работе приложения с реальными данными. Несмотря на то что HACMP и SDD способны справиться с таким отказом, необходимо, чтобы и приложение тоже было способно справиться с отказом.

Пример 12.4. Журнал ошибок AIX

```
#errpt
3074FEB7   0624164705 T H fscsi0        ADAPTER ERROR
3074FEB7   0624164705 T H fscsi0        ADAPTER ERROR
D7B7FF7E   0624164705 I O SYSJ2         USER DATA I/O ERROR
D7B7FF7E   0624164705 I O SYSJ2         USER DATA I/O ERROR
613E5F38   0624164705 P H LVDD          I/O ERROR DETECTED BY LVM
F4D25312   0624164705 P H vpath1        UNABLE TO COMMUNICATE WITH DEVICE
A7212C7B   0624164705 P H hdisk11       DEVICE ACCESS PROBLEM
A7212C7B   0624164705 P H hdisk3        DEVICE ACCESS PROBLEM
3074FEB7   0624164705 T H fscsi0        ADAPTER ERROR
3074FEB7   0624164705 T H fscsi0        ADAPTER ERROR
3074FEB7   0624164705 T H fscsi0        ADAPTER ERROR
3074FEB7   0624164705 T H fscsi0        ADAPTER ERROR
B8113DD1   0624164705 T H fcs0          LINK ERROR
AFA89905   0624163505 I O grpsvcs       Group Services daemon started
```

Существует два типа записей в журнале ошибок (см. Пример 12.4):

- vpath1: **unable to communicate with device (отсутствует связь с устройством)**. Эта ошибка возникает при отказе всех путей, принадлежащих к определенному устройству vpath. На устройстве vpath также происходит отказ, вследствие чего группа томов становится недоступной. Эта ошибка очень серьезно влияет на кластер, поэтому требуется, чтобы в этом случае произошло перемещение при сбое.
- hdisk11 и hdisk13: **device access problem (проблема с доступом к устройству)**: Ошибка возникает при отказе пути для связи с устройством vpath. Устройство vpath все еще доступно, однако отказ влияет на производительность и доступность. В этом случае адекватной реакцией является уведомление системного администратора.

Все эти ошибки произошли по одному разу и указывали на отказавшее устройство. Другие родственные сообщения об ошибках возникали по нескольку раз и не указывали на действительную причину проблемы. Теперь давайте рассмотрим эти записи журнала ошибок подробно. Запись см. в VPATH_OUT_SERVICE примере 12.5.

Пример 12.5. Записи журнала ошибок Vpath

```
#errpt -a|more
LABEL:          VPATH_OUT_SERVICE
IDENTIFIER:     F4D25312
Date/Time:       Fri Jun 24 16:47:29 CDT
Sequence Number: 222
Machine Id:      000197BA4C00
Node Id:         p650n02
Class:           H
Type:            PERM
Resource Name:   vpath1
Resource Class:  disk
Resource Type:   vpath
Location:

Description
UNABLE TO COMMUNICATE WITH DEVICE

Probable Causes
DISK
SCSI ADAPTER
SCSI CABLE

Failure Causes
DISK
SCSI ADAPTER
CABLE LOOSE OR DEFECTIVE

Recommended Actions
PERFORM PROBLEM DETERMINATION ON SCSI TARGET DEVICE
PERFORM PROBLEM DETERMINATION ON HOST SCSI ADAPTER
REPLACE SCSI CABLE

Detail Data
SENSE DATA
021F 4664 0000 0000 0029 0001 0000 0004 0000 0000 0000 0000 0000 0001
```

Запись VPATH_DEVICE_OFFLIN см. в журнале ошибок в примере 12.6.

Пример 12.6. Запись журнала ошибок Vpath

```
#errpt -a|more
LABEL:          VPATH_DEVICE_OFFLIN
IDENTIFIER:     A7212C7B

Date/Time:       Fri Jun 24 16:47:29 CDT
Sequence Number: 221
Machine Id:      000197BA4C00
Node Id:         p650n02
Class:           H
Type:            PERM
Resource Name:   hdisk11
```

Из выходных данных команды **errpt -a** мы можем получить всю требуемую информацию (в нашем примере они выделены полужирным шрифтом) для конфигурирования уведомления об ошибках HACMP.

Определение объектов уведомления об ошибках

Установите в SMIT уведомления об ошибках – отказах vpath – на всех узлах кластера. Мы используем значения команды **errpt -a**, представленные в примере 12.5.

1. Запустите **smit hacmp**.
2. Выберите **Problem Determination Tools (Инструменты определения проблем)**.
3. Выберите **HACMP Error Notification (Уведомления об ошибках HACMP)**.
4. Выберите **Add a Notify Method (Добавить метод уведомления)**.
5. Определите метод уведомления со следующими значениями (рис. 12.6):

- Notification Object Name (Имя объекта уведомления): **vpath_failed.**
- Persist across system restart? (Сохранять после перезапуска системы?): **Yes.**
- Process ID for use by Notify Method (Идентификатор процесса, используемый методом уведомления): **0** (root) подойдет.
- Select Error Class (Выбор класса ошибки): **Hardware** (см. «Class: H»).
- Select Error Type (Выбор типа ошибки): **Permanent** (см. «Тип: PERM»).
- Match Alertable errors? (Сопоставлять ошибки с возможностью оповещения?): **None.**
- Select Error Label (Выбор метки ошибки): **VPATH_OUT_OF_SERVICE.**
- Resource Name (Имя ресурса): **All**. Требуется установить уведомления об ошибках для всех устройств vpath.
- Resource Class (Класс ресурса): **disk** (см. «Resource Class: disk»).
- Resource Type (Тип ресурса): **vpath** (см. «Resource Type: vpath»).

```
                                Add a Notify Method

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
* Notification Object Name                            [disk_error]
* Persist across system restart?                      No +
  Process ID for use by Notify Method                 [] +#
  Select Error Class                                   Hardware +
  Select Error Type                                    PERM +
  Match Alertable errors?                              None +
  Select Error Label                                  [VPATH_PATH_REMOVED] +
* Resource Name                                       [All] +
* Resource Class                                      [All] +
* Resource Type                                       [All] +
* Notify Method                                       [/usr/ha/vpath_error]

F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Рис. 12.6. Конфигурирование уведомления об ошибках для ошибок vpath

- Notify Method (Метод уведомления): **/usr/es/sbin/cluster/diag/cl_failover $6 $9**. Этот скрипт предоставляется HACMP и запускает событие-ошибку HACMP **errnotify**, что осуществляет перемещение при сбое. Аргумент $6 представляет имя устройства, $9 представляет метку ошибки.

Установите через SMIT уведомление об ошибках для ошибок связи с vpath на всех узлах.

Мы используем значения из команды **errpt -a**, представленные в примере 12.6.

1. Запустите **smit hacmp**.

2. Выберите **Problem Determination Tools (Инструменты определения проблем)**.

3. Выберите **HACMP Error Notification (Уведомления об ошибках HACMP)**.

4. Выберите **Add a Notify Method (Добавить метод уведомления)**.

5. Определите метод уведомления со следующими значениями (рис. 12.6):

- Notification Object Name (Имя объекта уведомления): **path_offline.**
- Persist across system restart? (Сохранять после перезапуска системы?): **Yes.**
- Process ID for use by Notify Method (Идентификатор процесса, используемый методом уведомления): **0** (root) подойдет.
- Select Error Class (Выбор класса ошибки): **Hardware** (см. «Class: H»).
- Select Error Type (Выбор типа ошибки): **Permanent** (см. «Тип: PERM»).

- Match Alertable errors? (Сопоставлять ошибки с возможностью оповещения?): **None.**
- Select Error Label (Выбор метки ошибки): **VPATH_DEVICE_OFFLIN.**
- Resource Name (Имя ресурса): **All**. Требуется установить уведомления об ошибках для всех устройств vpath.
- Resource Class (Класс ресурса): **disk** (см. «Resource Class: disk»).
- Resource Type (Тип ресурса): **2105** (см. «Resource Type: 2105»).

Notify Method (Метод уведомления): **/usr/ha/notify_root $6 $9**. Исходный текст скрипта /usr/ha/notify_root представлен в примере 12.7. Мы передаем аргументы $6 и $9: имя устройства и метку ошибки. Скрипт отправляет сообщение электронной почты пользователю root, содержащее сведения о возникшей ошибке.

Пример 12.7. Скрипт /usr/ha/notify_root

```
#!/bin/ksh
# Processing the arguments
# device name
DEVICE=$1
# Error log label
ERRLABEL=$2
mail -s "Vpath communication error" root <<EOF
The $DEVICE device encountered an $ERRLABEL error. The disk subsystem
performance and availability is degraded. Perform problem determination!
EOF
exit 0
```

Тестирование уведомления об ошибках

Существует два способа тестирования объекта уведомления об ошибках:

1. Поочередное отключение кабелей FC от адаптеров HBA. Это в действительности позволяет протестировать решение; если все настроено правильно, HACMP инициирует перемещение при сбое.

2. Имитация записей в журнале ошибок:

а) Запустите **smit hacmp**.

б) Выберите **Problem Determination Tools (Инструменты определения проблем)**.

в) Выберите **HACMP Error Notification (Уведомления об ошибках HACMP)**.

г) Выберите **Emulate Error Log Entry (Эмуляция записи в журнал ошибок)**.

д) Выберите метку ошибки **VPATH_OUT_OF_SERVICE** или **VPATH_DEVICE_OFFLIN** из выпадающего списка.

е) SMIT выводит метку ошибки, имя объекта уведомления и метод уведомления. Нажмите Enter для подтверждения эмуляции записи в журнал ошибок (рис. 12.7).

```
                         Emulate Error Log Entry

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
  Error Label Name                                    VPATH_OUT_SERVICE
  Notification Object Name                            vpath_failed
  Notify Method                                       /usr/es/sbin/cluster/>


F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Рис. 12.7. Эмуляция записи VPATH_OUT_SERVICE журнала ошибок

Теперь HACMP отправляет запрошенную запись журнала ошибок демону журналов ошибок AIX и запускает метод уведомления. Если все настроено правильно, HACMP инициирует перемещение при сбое. Сведения о происходящих действиях записываются в файл /tmp/hacmp.out.

13

Аспекты подсистемы хранения

Эта глава содержит более подробные сведения по следующим вопросам:

- типы групп томов;
- резервирование дисков;
- принудительная активизация групп томов;
- быстрый перехват дисков;
- мониторинг пульса через диски.

13.1 Типы групп томов

Важно иметь представление о различных типах групп томов, а также об использовании каждого типа в HACMP. Мы рассмотрим следующие типы групп томов:

- с расширенным одновременным доступом (Enhanced concurrent);
- без одновременного доступа (Non-concurrent);
- с одновременным доступом (Concurrent);
- с одновременным доступом RAID (RAID concurrent).

Существуют другие дополнительные атрибуты групп томов, которые могут совместно использоваться с перечисленными выше атрибутами (например, «big» и «scalable»). Вообще говоря, можно совмещать эти атрибуты с типами, перечисленными выше. Например, можно использовать большую группу томов с расширенным одновременным доступом (big enhanced concurrent volume group). Однако эти дополнительные атрибуты не влияют на способ активизации групп томов в HACMP. HACMP активизирует большую группу томов с расширенным одновременным доступом, используя такой же метод, как если бы это была простая группа томов с расширенным одновременным доступом.

13.1.1 Группы томов с расширенным одновременным доступом

Группы томов с расширенным одновременным доступом (Enhanced concurrent) впервые появились в AIX 5.1. В отличие от групп томов с одновременным доступом, которые предназначены для использования только в SSA, эти группы томов поддерживали использование в любой дисковой подсистеме, поддерживаемой в общей конфигурации, AIX, pSeries, HACMP. В AIX 5.2 и выше группы томов с расширенным одновременным доступом представляют единственный доступный тип групп томов с одновременным доступом.

Группы томов с расширенным одновременным доступом используют демон Group Services Concurrent Logical Volume Manager daemon (gsclvmd), который осуществляет обмен данными с другими узлами кластера через IP.

Использование gsclvmd позволяет выполнять большинство изменений в LVM динамически, даже из командной строки. Для того чтобы эти динамические изменения работали корректно, необходимо, чтобы при выполнении обслуживания были запущены gsclvmd, topsvcs, groupsvcs и emsvcs. Это легко сделать при запущенном и работающем кластере HACMP с группами томов, подключенными в режиме одновременного доступа.

> **Примечание.** Все администрирование LVM кластера рекомендуется осуществлять через C-SPOC. В отличие от командной строки в данном случае не требуется, чтобы на узлах кластера были запущены службы кластера (за исключением clcomdES).

Группы томов с расширенным одновременным доступом можно использовать как в средах с одновременным доступом, так и в средах без одновременного доступа. От этого зависят дополнительные функции HACMP (т. е. быстрый перехват дисков и мониторинг пульса через диски).

Внимание! При конфигурировании групп томов с расширенным одновременным доступом в кластере убедитесь, что существует несколько сетей (IP и отличных от IP) для связи между узлами в кластере, чтобы избежать разделения кластера (cluster partitioning). При использовании быстрого перехвата дисков не устанав-ливается обычный бит резервирования SCSI, чтобы избежать подключения нескольких узлов к группе томов.

Существующие группы томов без одновременного доступа можно преобразовать в группы томов с расширенным одновременным доступом без потери какого-либо дополнительного пространства для хранения. Для изменения группы томов она должна быть подключена; изменение выполняется командой **chvg -c vgname**. Для того чтобы это изменение вступило в силу на других узлах, группу томов необходимо отключить и затем либо экспортировать и реимпортировать, либо использовать опцию **learn** команды importvg: **importvg -L vgname pvname.**

Для создания новой группы томов с расширенным одновременным доступом на локальном узле из командной строки нужно просто выполнить команду **mkvg -C vgname pvname.**

Для того чтобы проверить, имеет ли том расширенный одновременный доступ, можно выполнить команду **lsvg vgname** и посмотреть поле «Concurrent:»; оно должно содержать значение «Enhanced Capable», как показано на рис. 13.1.

```
Maddi / > lsvg applvg
VOLUME GROUP:   applvg                  VG IDENTIFIER: 0022be2a00004c48
VG STATE:       active                  PP SIZE:        16 megabyte(s)
VG PERMISSION:  read/write              TOTAL PPs:     1190
MAX LVs:        256                     FREE PPs:      1180
LVs:            0                       USED PPs:       10
OPEN LVs:       0                       QUORUM:         2
TOTAL PVs:      2                       VG DESCRIPTORS: 3
STALE PVs:      0                       STALE PPs:      0
ACTIVE PVs:     2                       AUTO ON:        no
Concurrent:     Enhanced-Capable        Auto-Concurrent: Disabled
VG Mode:        Non-Concurrent
MAX PPs per PV: 1016                    MAX PVs:        32
LTG size:       128 kilobyte(s)         AUTO SYNC:      no
HOT SPARE:      no                      BB POLICY:      relocatable
```

Рис. 13.1. Пример группы томов с расширенным одновременным доступом

13.1.2 Группы томов без одновременного доступа

При создании новой группы томов по умолчанию создается группа томов без одновременного доступа (Non-concurrent). Она также называется *стандартной группой томов*. Сущность групп томов без одновременного доступа состоит в том, что несколько систем не могут одновременно получить доступ к группе томов. Полный доступ для чтения-записи возможен только для системы, активизировавшей группу томов командой varyonvg vgname.

Группы томов без одновременного доступа не являются выделенным типом групп томов для LVM. Имеется в виду выделение режима работы, в котором группа томов будет использоваться. При выполнении команды lsvg для группы томов без одновременного доступа, выходные данные команды не содержат поле «Concurrent», показанное на рис. 13.1.

13.1.3 Группы томов с одновременным доступом

> **Примечание.** Версия AIX 5.1 с 32-разрядным ядром является последней версией, поддерживающей группы томов с одновременным доступом (Concurrent). AIX 5.2 и выше используют группы томов с расширенным одновременным доступом, описываемые в разделе 13.1.1, «С расширенным одновременным доступом».

Группы томов этого типа, также называемые *группами томов с возможностью одновременного доступа (Concurrent Capable),* используются в HACMP с дисками SSA в конфигурации «одновременного доступа». Эта комбинация являлась первым типом группы томов с настоящим режимом одновременного доступа.

Уникальная последовательная связь дисков SSA позволяет осуществлять доступ через так называемый *скрытый канал (covert channel)*. Этот скрытый канал используется диспетчером логических томов с одновременным доступом (Concurrent Logical Volume Manager, CLVM). CLVM способен осуществлять автоматическую синхронизацию информации ODM, связанной с LVM, используя демон CLVM (clvmd). Это позволяет осуществлять обслуживание LVM групп томов в подключенном режиме.

Clvmd автоматически запускается при активизированных группах томов в режиме одновременного доступа с использованием команды **varyonvg -c**.

13.1.4 Группы томов с одновременным доступом RAID

> **Примечание.** Группы томов с одновременным доступом RAID в настоящее время используются редко и считаются устаревшими.

Группа томов с одновременным доступом RAID (RAID concurrent) представляет группу томов без одновременного доступа, добавляемую в группу ресурсов HACMP с одновременным доступом [в HA 5.2 и выше такие группы ресурсов называются *группами ресурсов с подключением на всех узлах (online on all nodes)*]. В LVM нет выделения, соответствующего типу «RAID concurrent». Тип «RAID concurrent» в действительности представляет внутреннее выделение HACMP.

При подключении группы ресурсов с одновременным доступом выполняется проверка типа группы томов. Если она является группой томов без одновременного доступа, HACMP должен определить, из устройств какого типа она состоит. Очень важно знать тип устройств, так как эта конфигурация поддерживала только некоторые устройства хранения.

Поддерживаемые устройства хранятся в файле /usr/es/sbin/cluster/diag/clconraid.dat. Если HACMP находит тип устройства, продолжается подключение группы томов

в режиме полного доступа для чтения-записи для каждого участвующего узла с использованием команды **convaryonvg**. В старой документации и файлах журналов (например, в /tmp/hacmp.out) этот режим имеет название «mode 3».

После подключения группы томов с одновременным доступом RAID в режиме одновременного доступа невозможно внести изменения, связанные с LVM (включая C-SPOC). Для внесения каких-либо изменений в LVM необходимо отключить группе томов на всех узлах, затем подключить ее только на одном узле, внести требуемые изменения и снова реимпортировать группу томов на все остальные участвующие узлы. Это, конечно же, неудобно, так как внесение практически каждого изменения в LVM вызывает простой.

13.2 Резервирование дисков

При активизации группы томов в AIX для каждого участвующего диска устанавливается бит резервирования. Это гарантирует, что другие системы не смогут получить доступ к этим дискам, что позволяет обеспечить целостность данных.

Такое резервирование часто называется *резервированием SCSI*, так как оно основано на стандартах SCSI. Большинство новых дисков FC используют протокол FSCSI, и при этом все еще применяют резервирование SCSI.

Стандарты SCSI определяют два различных типа резервирования:

- SCSI-2 – «традиционное» резервирование;
- SCSI-3 – постоянное резервирование (persistent reservation, PR).

Резервирование SCSI-2 позволяет осуществлять доступ только с использованием одного пути, поэтому этот тип резервирования нельзя применять для многопутевого доступа к хранилищу. Резервирование SCSI-2 не является постоянным, так как оно не восстанавливается после перезагрузки узла.

SCSI-3 PR (persistent reservation, постоянное резервирование) поддерживает доступ к устройству для нескольких узлов, в то же время блокируя доступ для других узлов. Резервирование SCSI-3 PR является постоянным и восстанавливается после сброса шины SCSI и перезагрузки узлов, а также поддерживает использование нескольких путей от узла к диску.

SCSI-3 PR применяет понятие регистрации и резервирования. Некоторые участвующие системы регистрируют «ключ» для устройства SCSI-3. Каждая система регистрирует собственный ключ. Зарегистрированные системы затем могут выполнить резервирование. С использованием этого метода блокировка доступа для записи требует только лишь удаления регистрации с устройства. Когда одной системе требуется отключить доступ другой системе, она выдает команду «pre-empt and abort», которая отключает доступ другой системе. После отключения доступа узел не имеет зарегистрированного ключа и не может отключить доступ другим системам. Этот метод позволяет избежать возникновения ситуаций «split brain».

Еще одно преимущество метода SCSI-3 PR состоит в том, что, так как узел регистрирует один и тот же ключ для всех путей, отключение одного ключа блокирует все пути ввода-вывода для этого узла. Технология SCSI-3 PR, в частности, реализована в системах EMC Symmetrix, Sun™ T3 и Hitachi Storage. Системы ESS SDD используют

постоянное резервирование, тогда как команды LVM употребляют «традиционное» резервирование.

Резервирование обычно действует до отключения бита резервирования при деактивизации (varying off) группы томов. Даже при отключении системы AIX, если включено питание дисков, резервирование обычно сохраняется. По этой причине HACMP должен отключать резервирование дисков во время перемещения при сбое, чтобы подключить группу томов на дежурном узле.

Резервирование дисков не используется для групп томов в режиме одновременного доступа или в режиме расширенного одновременного доступа, применяемых при подключении групп томов в режиме одновременного доступа. Это также относится к использованию групп томов с расширенным одновременным доступом в конфигурации с быстрым перехватом дисков.

Дополнительные сведения о быстром перехвате дисков см. в разделе 13.4, «Быстрый перехват дисков».

13.3 Принудительная активизация групп томов

Эта функция очень важна при использовании групп томов с зеркальным отображением. Она реализована в виде атрибута команды varyonvg, представленного флагом -f. Использование этого флага позволяет подключить группу томов даже при отсутствии кворума дисков. В процессе активизации выполняется проверка всех логических томов; для успешной активизации должна существовать как минимум одна полная копия каждого логического тома.

Отключение настроек кворума при конфигурировании группы томов с зеркальным отображением является стандартной практикой. При отключенном кворуме группа томов может оставаться подключенной даже после отказа диска. Пока доступен хотя бы один диск, группа томов будет оставаться подключенной. Однако для первоначальной активизации группы томов должны быть доступны все диски. Для активизации группы томов, в которой доступны не все участвующие диски, необходимо использовать атрибут принудительной активизации.

Эта функция чаще всего используется при зеркальном отображении между подсистемами хранения и/или при зеркальном отображении между разными расположениями через межсайтовое зеркальное отображение LVM (cross-site LVM mirroring). Это позволяет обеспечить дублирование сайта, чтобы в случае простоя сайта (сайт состоит из сервера и одной копии хранилища) сервер на удаленном сайте мог активизировать группу томов с локальной копии LVM. Дополнительные сведения о межсайтовом зеркальном отображении LVM см. в гл. 16.

В HACMP 5.1 и более поздних версиях пользователь может установить этот атрибут в определении группы ресурсов. До версии 5.1 пользователи могли устанавливать переменную окружения `HACMP_MIRROR_VARYON=true` в /etc/environment или применять настраиваемые скрипты обработки событий.

13.4 Быстрый перехват дисков

Этот раздел описывает следующие аспекты быстрого перехвата дисков:

- требования;
- принцип работы быстрого перехвата дисков;
- включение быстрого перехвата дисков;
- преимущества;
- недостатки.

13.4.1 Требования

Ниже перечислены требования, выполнение которых необходимо для реализации быстрого перехвата дисков:

- HACMP V5.1 или выше;
- Cluster.es.clvm.rte (компонент HACMP CRM);
- AIX 5.2 или выше;
- Bos.clvm.enh 5.2.0.11 или выше;
- группы томов с расширенным одновременным доступом в группах ресурсов без одновременного доступа.

13.4.2 Принцип работы быстрого перехвата дисков

Традиционно перехват дисков предполагает отключение резервирования SCSI на каждом дисковом устройстве последовательным образом. Количество времени, занимаемое отключением резервирования, зависит от типа диска. В большой среде с сотнями дисков оно может значительно увеличить продолжительность перемещения при сбое.

Быстрый перехват дисков сокращает общее время перемещения при сбое, обеспечивая более быстрое получение дисков без отключения резервирования SCSI. При этом используются группы ресурсов с расширенным одновременным доступом и дополнительные усовершенствования LVM, реализованные в AIX 5.2.

В AIX 5.2 была реализована возможность активизации группы томов с расширенным одновременным доступом в двух различных режимах:

- активный режим;
- пассивный режим.

Использование активного режима подобно активизации группы томов без одновременного доступа с использованием простой команды **varyonvg**. Активный режим обеспечивает полный доступ для чтения-записи для всех логических томов, файловых систем и поддерживает все операции LVM.

Пассивный режим представляет своего рода ограждение группы томов на уровне LVM. Пассивный режим обеспечивает только возможность чтения VGDA и первых 4 Кб каждого логического тома. Он **не** позволяет осуществлять доступ для чтения-записи в файловые системы или логические тома. Он также не поддерживает операции LVM.

При подключении группы ресурсов, содержащей группу томов, сначала выполняется ее активизация в пассивном режиме, а уже затем выполняется активизация в активном режиме. Активный режим применяется только к текущей группе ресурсов, владеющей узлом. При подключении другого узла группы ресурсов выполняется активизация группы томов в пассивном режиме.

При отказе узла-владельца (домашнего узла) резервный узел просто изменяет состояние группы томов с пассивного режима на активный режим через LVM. Это изменение занимает около 10 с и выполняется на уровне группы томов. Оно может занимать больше времени при использовании нескольких групп томов по нескольку дисков в каждой. Тем не менее продолжительность выполнения данной процедуры минимальна по сравнению с предыдущим методом отключения резервирования SCSI.

Флаги активного и пассивного режима для команды varyonvg не документированы, так как они не предназначены для использования вне среды HACMP. Однако их легко можно найти в журнале hacmp.out.

```
Melany / > lspv |grep vpath
vpath0          0022be2a8607249f            app2vg
vpath1          0022be2a8617133e            None
vpath2          0022be2a86607918            app1vg          concurrent
vpath3          0022be2a8662c1a4            None
vpath4          0022be2a8662ce0e            app2vg
vpath5          0022be2a8662dfa8            None
vpath6          0022be2a8662f794            None
vpath7          0022be2a86630978            app1vg          concurrent

Melany / > lsvg -o
rootvg

Melany / > lsvg app1vg
VOLUME GROUP:   app1vg                  VG IDENTIFIER:  0022be2a00004c48
VG STATE:       active                  PP SIZE:        16 megabyte(s)
VG PERMISSION:  passive-only            TOTAL PPs:      1190
MAX LVs:        256                     FREE PPs:       1180
LVs:            0                       USED PPs:       10
OPEN LVs:       0                       QUORUM:         2
TOTAL PVs:      2                       VG DESCRIPTORS: 3
STALE PVs:      0                       STALE PPs:      0
ACTIVE PVs:     2                       AUTO ON:        no
Concurrent:     Enhanced-Capable        Auto-Concurrent: Disabled
VG Mode:        Concurrent
Node ID:        6                       Active Nodes:
MAX PPs per PV: 1016                    MAX PVs:        32
LTG size:       128 kilobyte(s)         AUTO SYNC:      no
HOT SPARE:      no                      BB POLICY:      relocatable
```

Рис. 13.2 Пассивный режим группы томов

Команда активизации в активном режиме:

varyonvg -n -c -A app2vg

Команда активизации в пассивном режиме:

varyonvg -n -c -P app2vg

> **Важно.** Не выполняйте эти команды, если не запущены службы кластера

Чтобы определить, в каком режиме подключена группа томов (активном или пассивном), посмотрите значение поля «VG PERMISSION» в выходных данных команды **lsvg**, представленных на рис. 13.2.

Существуют другие характерные элементы состояния LVM для групп томов, используемых в конфигурации быстрого перехвата дисков. Например, через команду **lspv** группа томов отображает подключение в режиме одновременного доступа для каждого активного узла. Однако команда **lsvg -o** выведет только группу томов, подключенных к узлу, на котором она активизирована в активном режиме. Пример вывода пассивного режима представлен на рис. 13.2.

13.4.3 Включение быстрого перехвата дисков

В конфигурации кластера HACMP нет специальной опции или флага, устанавливающих быстрый перехват дисков. Быстрый перехват дисков устанавливается в результате определенного сочетания параметров кластера.

Общие группы томов должны представлять собой группы томов с расширенным одновременным доступом. Эти группы томов затем добавляются как ресурсы в группу ресурсов в режиме неодновременного доступа. Сочетание этих двух параметров и определяют использование метода быстрого перехвата дисков для получения группы томов в HACMP.

При подключении группы ресурсов с неодновременным доступом HACMP проверяет один из дисков группы томов, является ли он группой томов с расширенным одновременным доступом. HACMP определяет это с помощью команды **lqueryvg -p devicename -X**. Если команда возвращает 0, значит, группа томов является обычной группой томов без одновременного доступа. Если возвращается значение 32, тогда группа томов является группой томов с расширенным одновременным доступом.

На рис. 13.3 диск hdisk0 входит в группу томов rootvg без одновременного доступа. Vpath0 представляет собой диск группы томов с расширенным одновременным доступом.

В AIX 5.1 было три возможных возвращаемых значения. Значение 16 возвращалось при использовании групп томов SSA с одновременным доступом. В AIX 5.2 и выше этот тип группы томов считается устаревшим.

```
Maddi / > lqueryvg -p hdisk0 -X
0
Maddi / >lqueryvg -p vpath0 -X
32
```

Рис. 13.3. Пример определения типа группы томов в HACMP

13.4.4 Преимущества

Существует по меньшей мере два преимущества использования быстрого перехвата дисков:

- уменьшение продолжительности получения дисков;
- синхронизация ODM LVM.

Мы уже описывали первое преимущество в разделе «Принцип работы быстрого перехвата дисков». Второе преимущество, синхронизация ODM LVM, напрямую связано с использованием групп томов с расширенным одновременным доступом и с выполнением демона gsclvmd со службами кластера.

Когда все узлы группы томов с расширенным одновременным доступом подключены в активном кластере, изменения в LVM, вносимые на домашнем узле (через командную строку или через SMIT), автоматически синхронизируются по всем остальным узлам. Это значительно уменьшает вероятность несогласованности информации группы томов на разных узлах.

Однако само по себе это преимущество не является рекомендованным методом работы в среде HACMP. Рекомендуется использовать C-SPOC для синхронизации информации ODM LVM между узлами кластера. Синхронизация в C-SPOC имеет следующие преимущества в сравнении с автономной синхронизацией ODM LVM:

1. Независима от состояния узлов кластера (активный/неактивный).
2. Может использоваться для изменений в JFS.
3. Обновляет файлы отметок времени VGDA на узлах кластера.

13.4.5 Известные проблемы

Один из недостатков использования быстрого перехвата дисков состоит в невозможности остановить службы кластера с использованием опции **forced**. При остановке кластера через **smit clstop** HACMP проверяет, содержит ли локальный узел какую-либо группу ресурсов с быстрым перехватом дисков. Если да, то опция forced не будет выводиться в меню.

Причина в том, что при этом группы томов оставляются подключенными и зависят от gsclvmd. Gsclvmd зависит от служб групп. При остановке служб кластера происходит остановка служб, необходимых для обеспечения согласованности группы томов. Это означает потенциальную уязвимость групп томов.

В последних версиях HACMP опция принудительного отключения используется не настолько часто. Однако в некоторые периоды обслуживания может быть желательно ее применять, чтобы сократить общее время простоя служб приложений.

Это ограничение является хорошо известным, и разработчики ищут способы его устранения в будущем. Однако нет никакой гарантии, что оно когда-либо будет устранено.

13.5 Мониторинг пульса через диски

Этот раздел описывает следующие аспекты мониторинга пульса через диски:

1. Обзор.
2. Требования.
3. Аспекты производительности.
4. Конфигурирование (на примере кластера из двух узлов).
5. Тестирование.
6. Мониторинг.

13.5.1 Обзор

Мониторинг пульса через диски представляет еще один вариант мониторинга пульса, отличный от IP, который использует существующие общие диски любого типа. Эта функция впервые была реализована в HACMP V5.1 и быстро стала предпочтительным методом мониторинга пульса, отличного от IP, так как он не требует использования последовательных кабелей и/или 8-портовых асинхронных адаптеров. При использовании среды SAN этот метод позволяет легко справиться с большими расстояниями между узлами.

Для обеспечения требуемого доступа к диску с каждого узла данная функция требует применения групп томов с расширенным одновременным доступом. Она использует специальную зарезервированную область на дисках для чтения и записи данных пульса. Так как эта функция применяет зарезервированную область, возможно использование существующих групп томов данных без потери дискового пространства. Однако следует помнить об аспектах производительности, описываемых в соответствующем разделе этой работы.

Возможно использование выделенного диска/LUN для мониторинга пульса через диски. Однако так как мониторинг пульса через диски применяет зарезервированное пространство, остальная область хранения данных не используется. Чем больше объем диска/LUN, выделенного для этой цели, тем больше пространства расходуется впустую. Однако при необходимости его можно использовать позже в качестве дополнительного дискового пространства.

Как и другие сети, отличные от IP, сеть пульса через диски имеет тип «точка-точка». Если кластер содержит больше двух узлов, потребуется как минимум **N** сетей пульса, отличных от IP, где **N** представляет количество узлов в кластере. Например, кластер из трех узлов требует по меньшей мере использования трех сетей пульса, отличных от IP.

13.5.2 Требования

Ниже перечислены требования к программному и аппаратному обеспечению.

- HACMP V5.1 или выше.
- Cluster.es.clvm.rte (компонент HACMP CRM).

- AIX 5.2 или выше (рекомендуется ML2).
- RSCT 2.3.1.1 или выше (рекомендуется 2.3.3.1).
- Bos.clvm.enh (необходимо для поддержки групп томов с расширенным одновременным доступом).
- Общие диски (конфигурируются как группы томов с расширенным одновременным доступом).

13.5.3 Аспекты производительности

Большинство современных дисков, не входящих в RAID-массивы, могут выполнять около 100 операций поиска в секунду. Секторы, используемые для мониторинга пульса через диски, являются частью VGDA. Область VGDA расположена на внешнем крае диска и может находиться далеко от данных приложений. Это означает, что при каждом импульсе мониторинга пульса выполняется операция поиска. Мониторинг пульса через диски обычно (при заданных по умолчанию параметрах) требует четыре (4) операции поиска в секунду. Другими словами, каждый из двух узлов будет выполнять запись на диск и чтение с диска один раз в секунду, что в сумме составляет 4 IOPS. При выборе диска, применяемого для мониторинга пульса, рекомендуется использовать диск с менее чем 60 операциями поиска в секунду. Для мониторинга активности операций поиска на диске можно применять инструмент **filemon**.

Если вы решите использовать диск, нагрузка ввода-вывода в котором превышает приведенное выше рекомендуемое значение, тогда рекомендуется изменить скорость обнаружения отказов (failure detection rate) в сети мониторинга пульса через диски на **slow**. Дополнительные сведения см. в разделе 14.2.

Указанные рекомендации относятся к хранению без использования RAID (т. е. JBOD). Технология дисковой подсистемы влияет на общие рекомендации. Например:

- Если диск является частью подсистемы хранения промышленного класса с большим объемом кеша записи (например, ESS), то количество операций поиска может быть значительно выше.
- Если диск, используемый для мониторинга пульса, входит в массив или подмассив RAID с небольшим кешем или без кеша, диск будет поддерживать меньшее количество операций поиска, в связи с дополнительной активностью, вызываемой операциями RAID. Свяжитесь с производителем, чтобы выяснить, сколько операций поиска то или иное устройство способно поддерживать.

13.5.4 Конфигурирование мониторинга пульса через диски

В рассматриваемом здесь примере употребляются кластер из двух узлов (Justen и Christie) со стандартными устройствами ESS vpath (vpath0 и vpath3 соответственно), предназначенными для использования в качестве устройства мониторинга пульса. Оба vpath-устройства уже сконфигурированы как диски, входящие в группу томов с расширенным одновременным доступом.

Существует два различных метода конфигурирования устройства мониторинга пульса через диски:

- с использованием метода обнаружения;
- с использованием метода предопределенных устройств.

В данном примере мы будем применять метод предопределенных устройств. При использовании этого метода необходимо сначала создать сеть diskhb, после чего назначить сети устройства пары диск-узел.

Прежде чем продолжить применение метода предопределенных устройств, необходимо точно знать имена устройств на каждом узле. Имена не обязательно должны совпадать, как показано в нашем примере. Сопоставление устройств может выполняться с использованием pvid на каждом узле с применением команды **lspv** в каждой системе.

> **Примечание.** При использовании метода обнаружения HACMP осуществляет автоматическое сопоставление устройств и выводит список для выбора.

Создайте сеть diskhb следующим образом. Выберите **smit hacmp → Extended Configuration (Расширенное конфигурирование) → Extended Topology Configuration (Расширенное конфигурирование топологии) → Configure HACMP Networks (Конфигурирование сетей HACMP) → Add a Network to the HACMP cluster (Добавить сеть в кластер HACMP)** → выберите **diskhb**. Введите требуемое имя сети (по умолчанию net_diskhb_01), как показано на рис. 13.4.

```
                    Add a Serial Network to the HACMP Cluster

  Type or select values in entry fields.
  Press Enter AFTER making all desired changes.

                                                      [Entry Fields]
  * Network Name                                     [net_diskhb_01]
  * Network Type                                      diskhb
```

Рис. 13.4. Добавление сети diskhb

Теперь нужно добавить два коммуникационных устройства (по одному для каждого узла) в сеть пульса через диски, созданную на предыдущем этапе.

Выберите **smit hacmp → Extended Configuration (Расширенное конфигурирование) → Extended Topology Configuration (Расширенное конфигурирование топологии) → Configure HACMP Communication Interfaces/Devices (Конфигурирование коммуникационных интерфейсов/устройств HACMP) → Add Communication Interfaces/Devices (Добавление коммуникационных интерфейсов/устройств) → Add Pre-Defined Communication Interfaces and Devices (Добавить предопределенные коммуникационные интерфейсы и устройства) → Communication Devices (Коммуникационные устройства)** → выберите diskhb, созданный на предыдущем этапе (net_diskhb_01), и нажмите Enter.

В поле Device Name (Имя устройства) следует указать уникальное имя, которое можно выбрать для описания устройства. Устройство будет отображаться в тополо-

```
                          Add a Communication Device

  Type or select values in entry fields.
  Press Enter AFTER making all desired changes.

                                                      [Entry Fields]
  * Device Name                                      [jc_disk_hb]
  * Network Type                                      diskhb
  * Network Name                                      net_diskhb_01
  * Device Path                                      [/dev/vpath0]
  * Node Name                                        [Justen]          +
```

Рис. 13.5. Добавление отдельных коммуникационных устройств diskhb

гии под этим именем, подобно тому как в прошлом отображались устройства последовательного мониторинга пульса и терминалы.

В поле Device Path (Путь к устройству) введите /dev/vpath0. Затем выберите соответствующий узел для данного устройства.

После создания первого устройства в любой сети, отличной от IP, может возникнуть предупреждение, представленное в примере 13.1.

Пример 13.1. Предупреждение Diskhb

```
WARNING: Serial network [net_name] has 1 communication device(s) configured.
Two devices are required for a serial network.
```

После повторения этой процедуры для другого узла (Christie) и другого устройства (vpath3) предупреждение не будет выводиться, так как требование наличия двух устройств выполнено.

13.5.5 Тестирование связи в сети пульса через диски

После создания определений устройств и сетей рекомендуется протестировать связь. Если группа томов активизируется в нормальном режиме на любом из узлов, тест вряд ли выполнится успешно.

Для тестирования подключения diskhb используется утилита **/usr/sbin/rsct/bin/dhb_read**. Команда dhb_read имеет следующий синтаксис (табл. 13.1).

Таблица 13.1. Синтаксис команды dhb_read

Команда	Действие
`dhb_read -p devicename`	Вывод содержимого сектора diskhb
`dhb_read -p devicename -r`	Получение данных через сеть diskhb
`dhb_read -p devicename -t`	Передача данных через сеть diskhb

Для тестирования связи в сети diskhb настроим один узел (Justen) на получение, а другой (Christie) на передачу.

На узле Justen выполняем команду;

dhb_read -p rvpath0 -r

На узле Christie выполняем команду;

dhb_read -p rvpath3 -t

> **Примечание.** Заданное устройство представляет собой устройство прямого доступа, на что указывает «r» перед именем устройства. Для дисков hdisk утилита dhb_read выполняет автоматическое преобразование этого имени в соответствующее имя устройства прямого доступа. Для всех остальных устройств требуется указывать имя устройства явным образом.

```
Justen /usr/sbin/rsct/bin >./dhb_read -p rvpath0 -r
Receive Mode:
Waiting for response . . .
Link operating normally
Justen /usr/sbin/rsct/bin >


Christie /usr/sbin/rsct/bin >./dhb_read -p rvpath0 -t
Transmit Mode:
Detected remote utility in receive mode.  Waiting for response . . .
Link operating normally
Christie /usr/sbin/rsct/bin >
```

Рис. 13.6. Тестирование связи мониторинга пульса через диски

Если связь между узлами находится в рабочем состоянии, оба узла выводят «Link operating normally», как показано на рис. 13.6.

В большинстве случаев устройство diskhb входит в группу общей группы томов данных. Если группа томов еще не входит в группу ресурсов, то следует добавить ее в группу ресурсов и выполнить синхронизацию кластера.

13.5.6 Мониторинг

После запуска служб кластера можно осуществлять мониторинг (обычно всех) импульсов через диски, используя команду **lssrc -ls topsvcs**. В основном нужно следить за полем Missed HBs (Пропущенные импульсы). Если сумма постоянно растет, это указывает на то, что выбранный диск не является оптимальным для применения в сети diskhb. Следует либо переместить diskhb на другой диск, либо изменить скорость обнаружения отказов (failure detection rate) в сети diskhb на значение slow.

Фрагмент информации о сети diskhb представлен на рис. 13.7.

```
Justen /usr/sbin/rsct/bin >lssrc -ls topsvcs

NIM's PID: 286930
diskhb_1        [ 3]    2    1  S 255.255.10.1    255.255.10.1
diskhb_1        [ 3] rvpath0      0x82b759bd      0x82b81f74
HB Interval = 2.000 secs. Sensitivity = 4 missed beats
Missed HBs: Total: 0 Current group: 0
Packets sent    : 25934 ICMP 0 Errors: 0 No mbuf: 0
Packets received: 25934 ICMP 0 Dropped: 0
NIM's PID: 282856
```

Рис. 13.7. Мониторинг diskhb

По умолчанию устанавливается 60-секундная отсрочка начала обработки. При выполнении данной команды сразу же после запуска кластера информация мониторинга пульса не будет выводиться до тех пор, пока не закончится отсрочка.

14

Сетевые функции

В этой главе описаны некоторые новые сетевые возможности, доступные в HACMP 5.3. К новым функциям относятся политика распределения сервисных IP-адресов и автоматическое создание файла clhosts. В этой главе также содержится описание технологии EtherChannel, а также ее использование в кластере HACMP. Кроме того, в главе описано назначение файлов netmon.cf, clhosts и clinfo.rc. Также приведены рекомендации по поводу того, в каких средах следует использовать эти файлы.

14.1 Etherchannel

EtherChannel (EC) представляет собой метод агрегирования портов, при котором до восьми Ethernet-адаптеров определяются как один EtherChannel. Удаленные системы видят EtherChannel как один IP- и MAC-адрес, в результате при использовании одной сети пропускная способность потенциально может быть увеличена в восемь раз.

Трафик распределяется между адаптерами стандартным способом (адресный алгоритм – address algorithm) или на основе циклического (round robin) обслуживания. При отказе адаптера трафик автоматически пересылается на следующий доступный адаптер в EtherChannel, не нарушая пользовательские подключения. Если активно только одно подключение в основном EtherChannel, тест на отказ начинает немедленное обнаружение/перемещение (в течение 2-4 с) на требуемый резервный адаптер без нарушения пользовательских подключений. Возможно проведение двух тестов: на отказ физической связи адаптера с сетью и на отказ TCP/IP-пути к узлу, заданному пользователем. При обнаружении отказа на резервном адаптере активизируются MAC- и IP-адреса. При восстановлении хотя бы одного адаптера в основном канале происходит повторная активизация адресов в основном канале.

Режим конфигурации Network Interface Backup (NIB), реализованный в AIX V5.1, был заменен и усовершенствован в AIX V5.2. Новый метод состоит в использовании одного адаптера EtherChannel с резервным адаптером, обеспечивая приоритет (возврат после восстановления связи) между основными и резервными каналами, что в предыдущей версии не было реализовано. Усовершенствование функции динамического членства адаптеров (dynamic adapter membership, DAM) в AIX V 5.2 позволяют осуществлять динамическое реконфигурирование адаптеров в EtherChannel без нарушения работающего подключения.

> **Примечание.** В HACMP не заявлена поддержка DAM, так как эта функция работает ниже уровня, на котором HACMP осуществляет мониторинг. Поэтому заявление о поддержке не требуется.

Все адаптеры, состоящие из нескольких каналов, требуют использования специальной конфигурации порта EtherChannel или IEEE 802.3ad в сетевом коммутаторе. В большинстве случаев коммутатор настраивается для применения в режиме EtherChannel. Однако если коммутатор не поддерживает EC или если корпорация в качестве стандарта использует IEEE 802.3ad, то следует сконфигурировать 802.3ad и в коммутаторе и в AIX. С другой стороны, подключения с одним адаптером не требуют специального конфигурирования на уровне сетевого коммутатора. Это включает EtherChannel с одним адаптером и подключение резервного адаптера.

EtherChannel имеет следующие преимущества:

- Более высокая пропускная способность и возможности балансировки нагрузки:
  - Каналы с несколькими адаптерами агрегируют пропускную способность.
  - Возможность использования нескольких вариантов направления трафика через адаптеры канала, настраиваемых пользователем.

- Встроенные функции обеспечения доступности:
  - Автоматическая обработка отказов адаптера, связей и сети.
  - Использование резервного адаптера для устранения единой точки отказа (single point of failure, SPOF) на уровне коммутатора сети. (Необязательно).
  - Методы проектирования для устранения единых точек отказа.
- Простое, гибкое решение и возможности масштабирования:
  - Один MAC- и IP-адрес Ethernet для всей агрегированной конфигурации (включая резервный адаптер).
  - Легко приспосабливается к будущим требованиям к пропускной способности.
  - Пользователь может добавлять, удалять и реконфигурировать адаптеры динамически (не прерывая обслуживания).
- Несколько вариантов взаимодействия с сетевым коммутатором;
  - Каналы с несколькими адаптерами для коммутаторов с поддержкой EtherChannel и 802.3ad.
  - Каналы с одним адаптером и резервные связи адаптеров прозрачны для сетевого коммутатора.
  - Опция подключения резервного адаптера канала (к другому сетевому коммутатору, чтобы избежать единой точки отказа).
  - При прямой связи двух систем канал работает без коммутатора (напрямую; однако в среде HACMP это неприменимо).
- Эта технология является бесплатной (при условии, что у вас уже установлены коммутаторы с поддержкой EC). Включена в AIX и регулярно улучшается, начиная с версии AIX v4.3.3.

14.1.1 Реализация EtherChannel в среде HACMP

HACMP официально поддерживает использование EtherChannel. Заявление о поддержке можно найти по адресу

http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/FLASH10284

Интеграция технологии EtherChannel в кластер осуществляется сравнительно просто и может значительно упростить адресацию в сети и требования к подсетям. Очень часто все адреса конфигурируются на одном логическом интерфейсе.

В нашем примере мы рассмотрим только то, что относится к совместному использованию HACMP и EtherChannel. Чтобы избежать повторения материала в этой книге, описание базового конфигурирования кластера HACMP опущено; предполагается, что эти знания у вас уже есть. Мы не будем приводить пошаговые инструкции по работе с меню HACMP. Рекомендуется также настроить сети пульса, отличные от IP, а также использовать коммутатор с поддержкой EC вместо кроссоверных кабелей. Следующая информация основана на предыдущей статье, посвященной комбинации этих технологий, вышедшей в мае 2004 г., однако все еще актуальной в настоящее время. Оригинальную версию этого документа можно найти по адресу

http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/TD101785

Конфигурация тестовой среды

Наша тестовая среда была построена с использованием следующих компонентов:

- две системы pSeries p630 (с именами **neo** и **trinity**);
- AIX V5.2 ML3;
- HACMP V5.1;
- сетевые подключения Ethernet ent0 – ent6:
  - ent0 и ent5 (не используется) представляют интегрированные адаптеры 10/100;
  - ent1, ent2, ent3, ent4 (не используется) представляют единый 4-портовый адаптер 10/100;
  - ent6 – EtherChannel (состоит из ent2, ent3 и ent0);
  - три кроссоверных кабеля UTP Ethernet.

На рис. 14.1 представлена схема используемой конфигурации кластера.

В этом тесте мы успешно реализовали «сеть с одним адаптером», использующую перехват IP-адреса (IP Address Takeover, IPAT) с функцией EtherChannel, включенной в AIX V 5.2. EtherChannel отвечает за переключение локального адаптера, расположенного за пределами HACMP. HACMP не знает о существовании EtherChannel и полностью независим. Хотя сеть с одним адаптером не является идеальным вариантом, в EtherChannel ее использование считается нормальным, так как в действительности одно псевдоустройство EtherChannel содержит несколько физических адаптеров. Таким образом, можно игнорировать предупреждения о недостаточном количестве адаптеров, выводимые в процессе синхронизации кластера.

Наша конфигурация содержала ротационную группу ресурсов и сеть с одним адаптером, использующую IP-синонимы. Наше тестирование подтвердило эффективность технологии в упрощении настройки HACMP. Мы реализовали подключение EtherChannel без сетевого коммутатора, подключив две тестовые системы напрямую с примечанием кроссоверных кабелей. Это было сделано только в целях тестирования. Как правило, в среде HACMP для полноценного использования этих адаптеров они подключаются к коммутатору с поддержкой Etherchannel.

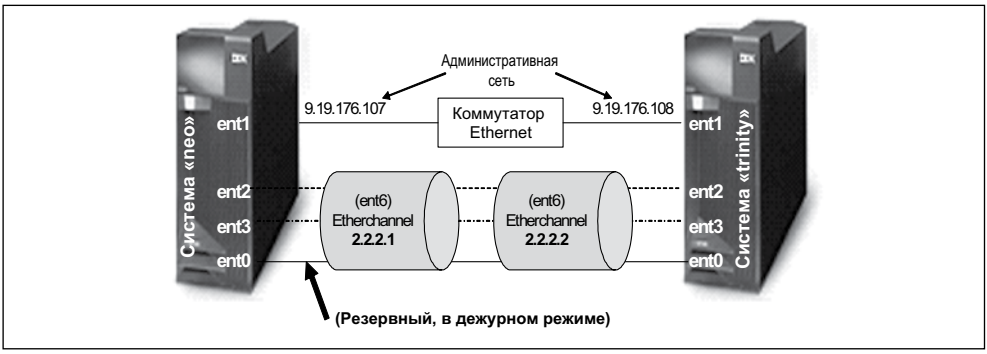

Рис. 14.1. Тестовая среда Etherchannel и HACMP

В настоящее время производители коммутаторов требуют подключения отдельных связей в EtherChannel к одному сетевому коммутатору. Для обеспечения дополнительной избыточности коммутаторов можно подключить резервный адаптер к отдельному коммутатору.

Тщательно подбирайте адаптеры для EtherChannel. Цель заключается в том, чтобы устранить единую точку сбоя. В тестовой среде мы использовали интегрированный адаптер Ethernet и один 4-портовый адаптер Ethernet в каждой системе, поэтому мы решили настроить интегрированный адаптер в качестве резервного, чтобы устранить единую точку сбоя на 4-портовом адаптере.

Процедуры конфигурирования

Перечислим основные этапы настройки кластера. Подробное описание каждого этапа для системы **neo** приведено ниже.

1. Проверка конфигурации адаптера Ethernet и кабелей адаптера.
2. Создание интерфейса EtherChannel.
3. Конфигурирование IP-адресов на новом интерфейсе (en6) через TCP/IP.
4. Добавление загрузочных и сервисных IP-адресов в топологию HACMP.
5. Создание группы ресурсов и назначение ей сервисного IP-адреса.
6. Синхронизация кластера.
7. Запуск служб кластера.
8. Тестирование избыточности NIC и проверка того, что HACMP не детектирует сбоев.

В начале у нас должны быть несконфигурированные адаптеры, желательно подключенные к коммутатору с поддержкой EC, где коммутатор уже настроен на конфигурацию EC. Наши адаптеры были предварительно сконфигурированы, так что мы удалили определения интерфейсов в ODM с использованием команды **smitty inet**. Мы выполнили эти основные действия на обеих системах, применяя IP-интерфейсы и IP-адреса, представленные на рис. 14.1.

Этап 1. Проверка конфигурации адаптера Ethernet

Адаптеры, которые планируется использовать в EtherChannel, должны быть настроены на одинаковую скорость и дуплексный режим. Мы настроили ent0, ent2 и ent3 на 100 Мбит/с и полный дуплекс через экран **smitty eadap -> Change/Show Characteristics of an Ethernet Adapter (Изменить/показать свойства адаптера Ethernet)**, представленный на рис. 14.2.

> **Примечание.** При использовании гигабитовых адаптеров, большинство из которых настроено на автоматическое определение скорости передачи, этот параметр не регулируется.

> **Замечание.** На данном этапе рекомендуется протестировать эти связи путем конфигурирования IP-адресов на каждой стороне. Только не забудьте удалить эти настройки, прежде чем перейти к следующему этапу.

```
                Change / Show Characteristics of an Ethernet Adapter

  Type or select values in entry fields.
  Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
    Ethernet Adapter                                   ent2
    Description                                        IBM 4-Port 10/100 Bas>
    Status                                             Available
    Location                                           12-08
    TRANSMIT queue size                                [8192]
    HARDWARE RECEIVE queue size                        [256]
    RECEIVE buffer pool size                           [384]
    Media Speed                                         100_Full_Duplex
    Inter-Packet Gap                                   [96]
    Enable ALTERNATE ETHERNET address                   no
    ALTERNATE ETHERNET address                         [0x000000000000]
    Enable Link Polling                                 no
    Time interval for Link Polling                     [500]
    Apply change to DATABASE only                       no

  F1=Help             F2=Refresh          F3=Cancel           F4=List
  Esc+5=Reset         F6=Command          F7=Edit             F8=Image
  F9=Shell            F10=Exit            Enter=Do
```

Рис. 14.2. Параметры адаптера Ethernet

Этап 2. Конфигурирование Etherchannel

Следует выполнить конфигурирование EtherChannel через быстрый путь **`smitty etherchannel`** **-> Add an Etherchannel/Link Aggregation (Добавить Etherchannel/Агрегирование связей)** и выбрать требуемые адаптеры, нажав клавишу F7. В нашей конфигурации ent2 и ent3 составляют основной канал, тогда как ent0 представляет резервный адаптер. При работе со следующим меню, представленном на рис. 14.3, происходит создание нового интерфейса EtherChannel (ent6).

Мы выбрали режим циклического обслуживания, чтобы обе связи можно было использовать в этой конфигурации. Просмотрите документацию по EtherChannel для получения сведений о различных режимах и выберите такой режим, который лучше всего подходит для вашей конфигурации.

> **Важно!** При использовании одного интерфейса с одним резервным адаптером (в прежних версиях – network interface backup или NIB) и указании значений параметров Number of Retries (Количество попыток) и Retry Timeout (Тайм-аут попыток) убедитесь в том, что они не превышают параметры скорости обнаружения отказов в NIM. Рекомендуется, чтобы значения этих параметров составляли как минимум половину значений параметров в HACMP.

```
                    Add An EtherChannel / Link Aggregation

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                      [Entry Fields]
  EtherChannel / Link Aggregation Adapters            ent2,ent3          +
  Enable Alternate Address                            no                 +
  Alternate Address                                   []                 +
  Enable Gigabit Ethernet Jumbo Frames                no                 +
  Mode                                                round_robin +
  Hash Mode                                           default            +
  Backup Adapter                                      ent0               +
       Automatically Recover to Main Channel          yes                +
  Internet Address to Ping                            []
  Number of Retries                                   []                 +
  Retry Timeout (sec)                                 []                 +

F1=Help             F2=Refresh          F3=Cancel           F4=List
Esc+5=Reset         F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Рис. 14.3 Меню добавления Etherchannel

Этап 3. Конфигурирование IP на устройстве Etherchannel

Следует выполнить конфигурирование IP-интерфейса (en6) в EtherChannel, используя быстрый путь **smitty chinet -> выбрать en6**, как показано на рис. 14.4.

Мы повторили эту процедуру на узле **trinity**, применяя IP-адрес 2.2.2.2.

> **Примечание.** Помните, что выполнение команд TCP/IP следует осуществлять для нового псевдоинтерфейса (en6), а не для отдельных интерфейсов.

Этап 4. Конфигурирование топологии HACMP

При тестировании мы решили использовать IP-синонимы для определения сети HACMP (`channet`). Мы настроили наши загрузочные IP-адреса на каждом устройстве EtherChannel (neo_boot 2.2.2.1, trinity_boot 2.2.2.2). Затем мы определили свой сервисный IP-адрес (привязанный к нескольким узлам) 192.168.43.4 и постоянные IP-адреса 192.168.43.10 для neo и 192.168.43.20 для trinity. Наша конфигурация топологии представлена на рис. 14.5 в выходных данных команды **cllsif**.

```
                    Change / Show a Standard Ethernet Interface

 Type or select values in entry fields.
 Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
   Network Interface Name                              en6
  INTERNET ADDRESS (dotted decimal)                   [2.2.2.1]
   Network MASK (hexadecimal or dotted decimal)       [255.255.255.0]
   Current STATE                                        up                  +
   Use Address Resolution Protocol (ARP)?               yes                 +
   BROADCAST ADDRESS (dotted decimal)                  []
   Interface Specific Network Options
      ('NULL' will unset the option)
     rfc1323                                          []
     tcp_mssdflt                                      []
     tcp_nodelay                                      []
     tcp_recvspace                                    []
     tcp_sendspace                                    []
   Apply change to DATABASE only                        no                  +

 F1=Help             F2=Refresh          F3=Cancel           F4=List
 Esc+5=Reset         F6=Command          F7=Edit             F8=Image
 F9=Shell            F10=Exit            Enter=Do
```

Рис. 14.4. Конфигурирование IP на устройстве EtherChannel

```
Adapter Type Network Net Type Attribute Node IP Addr Hardware Addr Interface Name Global Name Netmask
neo_boot1     boot    channet ether public neo     2.2.2.1       en6   255.255.255.0
neoec_srv     service channet ether public neo     192.168.43.4        255.255.255.0
trinity_boot1 boot    channet ether public trinity 2.2.2.2       en6   255.255.255.0
neoec_srv     service channet ether public trinity 192.168.43.4        255.255.255.0
```

Рис. 14.5. Конфигурация EtherChannel в выходных данных команды cllsif

Этап 5. Конфигурирование группы ресурсов HACMP

Мы сконфигурировали одну каскадную группу ресурсов с одной сервисной IP-меткой. Так как наша цель состояла в тестировании избыточности NIC, мы упростили конфигурацию, опустив дополнительные ресурсы в группе ресурсов. Определение нашей группы ресурсов представлено на рис. 14.6.

> **Примечание.** Хотя в нашем примере это опущено, в рабочей среде необходимо использовать как минимум одну последовательную сеть, отличную от IP.

```
                Change/Show Resources for a Rotating Resource Group

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
  Resource Group Name                                  testec_rg
  Participating Node Names (Default Node Priority)     neo trinity
* Service IP Labels/Addresses                          [neoec_svc]        +
  Volume Groups                                        []                 +
  Filesystems (empty is ALL for VGs specified)         []                 +
  Application Servers                                  []                 +
```

Рис. 14.6. Группа ресурсов Etherchannel

Этап 6. Синхронизация кластера

Хотя и предполагается, что вы знакомы с работой в HACMP, мы бы хотели показать предупреждение, выводимое при конфигурировании в топологии HACMP сети с одним адаптером. При синхронизации кластера мы получили следующее предупреждение (рис. 14.7):

```
WARNING: There may be an insufficient number of communication interfaces
defined on node: neo, network: channet. Multiple communication interfaces
are recommended for networks that will use IP aliasing.

WARNING: There may be an insufficient number of communication interfaces
defined on node: trinity, network: channet. Multiple communication
interfaces are recommended for networks that will use IP aliasing.
```

Рис. 14.7. Предупреждение при конфигурировании сети с одним адаптером

Так как в действительности мы сконфигурировали только один интерфейс в топологии HACMP, это предупреждение не является для нас неожиданным. EtherChannel при необходимости обеспечивает дублирование и переключение локального адаптера.

> **Важно.** При реализации подобной конфигурации сети с одним адаптером для корректного определения сети HACMP необходимо выполнить конфигурирование файла netmon.cf. Дополнительные сведения см. в разделе «Общее представление о файле netmon.cf».

Этап 7. Запуск служб кластера

Следует выполнить **smitty clstart** на каждом узле и подождать события node_up_complete.

Этап 8. Тестирование

Наше тестирование представляло физическое отключение кабелей, после чего мы смотрели, как на это реагирует система, чтобы убедиться, что в HACMP это прошло незамеченным. При выполнении каждого теста мы выполняли ping-опрос загрузочных IP-адресов и сервисного IP-адреса с внешнего клиентского узла.

1. Мы отключили кабель из ent3. В результате произошла передача обслуживания на ent2. Это было проверено командами netstat и entstat, а также выполнением ping-опросов с клиента. AIX регистрирует это в отчете об ошибках. Однако HACMP не знает о возникновении отказа. Ошибки из отчета об ошибках представлены на рис. 14.8.

```
F77ECAC2 0624145904 T H ent3 ETHERNET NETWORK RECOVERY MODE
8650BE3F 0624145904 I H ent6 ETHERCHANNEL RECOVERY
F77ECAC2 0624145904 T H ent2 ETHERNET NETWORK RECOVERY MODE
```

Рис. 14.8. Ошибки Etherchannel в отчете об ошибках AIX

2. Мы отключили кабель из ent2. Это вызвало перехват обслуживания дежурным адаптером ent0. Как и в первом тесте, AIX зарегистрировал отказ в отчете об ошибках, однако в HACMP отказ не был замечен. Так как мы использовали кроссоверные кабели, этот отказ имел двойное действие, вызывая подобные ошибки и переключения на обоих узлах.

3. Затем мы отключили кабель из последнего работающего адаптера ent0. Это, в свою очередь, вызвало полный отказ EtherChannel, что было зарегистрировано в HACMP как отказ сети.

Выводы по поводу технологии EtherChannel

Наши общие впечатления от реализации EtherChannels в среде HACMP были очень положительными. Хотя конфигурация требует дополнительного начального планирования, она прошла очень быстро и была простой в настройке. Мы были особенно довольны временем восстановления при тестировании; восстановление происходило почти мгновенно и не оказывало воздействия на работающий кластер. Также мы были довольны тем, что реализация этой модели позволяет избежать удаления маршрутов в событиях HACMP, связанных с переключениями локального адаптера, что сокращает время простоя и упрощает устранение неполадок.

В целом благодаря своей простоте и общим преимуществам модель EtherChannel является очень перспективой альтернативой при планировании новой среды, требующей обеспечения высокой доступности с использованием HACMP, а также масштабирования пропускной способности сети и избыточности. Динамическая масштабируемость и возможности обеспечения еще более высокой избыточности являются дополнительными стимулами для перехода на конфигурацию этого типа.

14.2 Варианты размещения сервисных IP-синонимов

При использовании IP-синонимов с конфигурированием нескольких сервисных IP-адресов HACMP анализирует общее количество синонимов, как определенных,

так и не определенных в HACMP, и назначает каждый сервисный адрес интерфейсу с наименьшей загруженностью. HACMP 5.1 и более поздние версии позволяют осуществлять дополнительный контроль над их размещением и определять варианты размещения сервисных IP-синонимов.

Этот атрибут сети можно использовать для настройки балансировки нагрузки сервисных IP-меток HA, учитывая, что все постоянные IP-метки уже сконфигурированы. Выбранный вариант размещения поддерживается при запуске кластера и последующих событиях кластера. Вариант размещения поддерживается до тех пор, пока в кластере доступны подходящие сетевые интерфейсы. Однако HACMP всегда поддерживает сервисные IP-метки в активном состоянии, даже при невозможности обеспечения требуемого варианта размещения.

Существует четыре различных политики размещения:

- **Без совместного размещения (Anti-collocation).** Используется по умолчанию. HACMP распределяет все сервисные IP-метки по всем базовым IP-адресам с использованием выбора «наименьшей загруженности».
- **С совместным размещением (Collocation).** HACMP размещает все сервисные IP-синонимы на одной сетевой интерфейсной карте (NIC).
- **Без совместного размещения и с постоянной меткой (Anti-Collocation with persistent).** HACMP размещает все сервисные IP-синонимы по всем активным физическим интерфейсам, не содержащим постоянную IP-метку. HACMP размещает сервисные IP-синонимы на интерфейсе, содержащем постоянную метку, только если недоступен другой сетевой интерфейс. Если не сконфигурированы постоянные IP-метки, HACMP позволяет выбрать вариант размещения Anti-Collocation with Persistent, однако выдает предупреждение и по умолчанию использует обычный вариант без совместного размещения.
- **С совместным размещением и с постоянной меткой (Collocation with persistent).** Все сервисные IP-синонимы размещаются на одной сетевой карте, содержащей постоянную IP-метку. Это может быть полезно в средах с VPN и брандмауэром, где только один интерфейс имеет внешний выход и где все IP-метки (постоянные и сервисные) должны размещаться на одной интерфейсной карте. Если не были сконфигурированы постоянные IP-адреса, HACMP позволяет выбрать вариант размещения Collocation with Persistent, однако выдает предупреждение и по умолчанию использует обычный вариант с совместным размещением.

14.2.1 Конфигурирование политики размещения сервисных IP-адресов

Вариант размещения можно устанавливать и изменять динамически. Ниже перечислены действия по конфигурированию политики размещения этого типа.

1. Введите `smit hacmp`.

2. В SMIT выберите Extended Configuration (Расширенное конфигурирование) > Extended Resource Configuration (Расширенное конфигурирование ресурсов) > HACMP Extended Resources Configuration (Расширенное конфигурирование ресурсов HACMP) > Configure Resource Distribution Preferences (Конфигурирование вариантов

размещения ресурсов) > Configure Service IP labels/addresses Distribution Preferences (Конфигурирование вариантов размещения сервисных IP-меток/адресов) и нажмите Enter. HACMP выводит только сети, использующие перехват IP-адресов посредством синонимов.

3. Выберите сеть, для которой требуется задать политику, и нажмите Enter.

4. На экране Configure Service IP Labels/Address Distribution Preference (Конфигурирование вариантов размещения сервисных IP-меток/адресов) выберите требуемый вариант размещения.

5. Нажмите Enter, чтобы принять выбранные значения и обновить ODM HACMP на локальном узле.

6. Для того чтобы изменения вступили в действие и были распространены на все узлы, требуется выполнить синхронизацию кластера. Перейдите в меню Initialization and Standard Configuration (Инициализация и стандартное конфигурирование) или Extended Configuration (Расширенное конфигурирование) и выберите Verification and Synchronization (Верификация и синхронизация). Это инициирует событие динамической реконфигурации.

> **Примечание.** При конфигурировании политики размещения сервисных IP-адресов мы получили следующие сообщения:
> cldare: Detected changes to service IP label app1svc. Please note that changing parameters of service IP label via a DARE may result in releasing resource group <name>

Просмотр варианта размещения сервисных IP-меток

У вас должна быть возможность выводить текущий вариант размещения для каждой сети с использованием команд **cltopinfo** или **cllsnw**.

Выходные данные команды **cltopinfo -w** имеют следующий вид:

```
# /usr/es/sbin/cluster/utilities/cltopinfo -w
NODE cobra:
Network net_diskhb_01
cobra_vpath0 /dev/vpath0
        Network net_ether_01
                app1svc 192.168.100.83
                app2svc 192.168.100.82
                cobraa 10.10.31.33
                cobrab 10.10.32.33
NODE viper:
        Network net_diskhb_01
                viper_vpath0 /dev/vpath0
        Network net_ether_01
                app1svc 192.168.100.83
```

app2svc 192.168.100.82
viperb 10.10.32.32
vipera 10.10.31.32

Network net_ether_01 is using the following distribution preference for service labels:

Collocation with persistent - service label(s) will be mapped to the same interface as the persistent label

Ниже приведен пример выходных данных команды **cllsnw -c**, выводящей вариант размещения сервисной метки (service label distribution preference, sldp) для определенной сети.

```
#/usr/es/sbin/cluster/utilities/cllsnw -c
#netname:attr:alias:monitor_method:sldp:
net_ether_01:public:true:default:ppstest::sldp_collocation_with_persistent
```

Эксперименты с политикой размещения сервисных меток

При тестировании мы могли изменять политику размещения сервисных IP-меток при отключенных службах кластера на всех узлах, после чего при запуске кластера происходило распределение IP-меток и постоянных IP-адресов в соответствии с установленной политикой. Это отображалось в выходных данных команды netstat -i:

```
python-# more /etc/hosts
10.10.31.31 pythona          # base address 1
10.10.32.31 pythonb          # base address 2
192.168.100.31 p630n01 n1    # python persistent address
192.168.100.82 app1svc       # cobra service address
192.168.100.83 app2svc       # viper service address
python-# netstat -i
```

Name	Mtu	Network	Address	Ipkts Ierrs	Opkts Oerrs	Coll
en0	1500	link#2	0.2.55.4f.c4.ab	5044669 0	1828909 0	0
en0	1500	10.10.31	pythona	5044669 0	1828909 0	0
en0	**1500**	**192.168.100**	**p630n01**	**5044669 0**	**1828909 0**	**0**
en0	**1500**	**192.168.100**	**app1svc**	**5044669 0**	**1828909 0**	**0**
en0	**1500**	**192.168.100**	**app2svc**	**5044669 0**	**1828909 0**	**0**
en3	1500	link#3	0.20.35.e2.7f.8d	3191047 0	1410806 0	0
en3	1500	10.10.32	pythonb	3191047 0	1410806 0	0
lo0	16896	link#1		1952676 0	1957548 0	0
lo0	16896	127	localhost	1952676 0	1957548 0	0
lo0	16896		localhost	1952676 0	1957548 0	0

Примечание. В представленных выше выходных данных узел python имел группы ресурсов для узлов cobra и viper, а также соответствующие сервисные IP-адреса. Была установлена политика распределения Collocation with persistent (С совместным размещением и с постоянной меткой).

При тестировании динамического изменения этой политики не произошло перемещения какой-либо метки после синхронизации. Во время синхронизации кластера после изменения политики размещения IP-адресов было зарегистрировано следующее сообщение:

Verifying additional pre-requisites for Dynamic Reconfiguration...

cldare: Detected changes to service IP label app1svc. Please note that changing parameters of service IP label via a DARE may result in releasing resource group APP1_RG .

cldare: Detected changes to service IP label app2svc. Please note that changing parameters of service IP label via a DARE may result in releasing resource group APP2_RG .

> **Примечание.** В данном случае зарегистрированное сообщение является обычным, и выдается только потому, что было обнаружено изменение. Если это изменение является единственным, отключения ресурсов не произойдет.

Изменение политики размещения сервисных IP-адресов применяется только при вызове события переключения или при остановке и перезапуске HACMP на узле вручную. Обратите внимание на то, что такое выполнение функции является преднамеренным и предназначено для того, чтобы не допустить потенциального нарушения связи с этими IP-адресами. Остальные узлы кластера не применяют политику, если только на них не будет выполнена остановка и перезапуск служб кластера.

14.3 Общее представление о файле netmon.cf

В HACMP можно создать файл конфигурации netmon.cf со списком дополнительных сетевых адресов. Эти адреса используются только для отправления ECHO-запросов ICMP службами топологии в целях определения состояния адаптера при определенных обстоятельствах.

Таким образом, реализация этого файла не является обязательной, однако рекомендуется для использования в конфигурациях кластера с одной сетевой картой на каждом узле или в кластере, в котором из-за отказов остался один сетевой адаптер. В таких ситуациях для HACMP может быть сложным точно определить отказ адаптера, так как службы топологии не могут принудительно направлять трафик через один адаптер для проверки его функционирования.

Усовершенствования в сетевом мониторе (netmon – сетевой компонент служб топологии RSCT) позволяют более точно определить отказ сервисного адаптера. Эту функцию можно применять в конфигурации, требующей использования одного сервисного адаптера на сеть.

Файл должен существовать при запуске кластера, так как службы топологии RSCT сканируют файл netmon.cf при инициализации. Когда сетевому монитору требуется проверить сеть, чтобы убедиться в функционировании адаптера, он направляет запросы ICMP ECHO на каждый IP-адрес в файле. После отправления запроса на все адреса сетевой монитор проверяет счетчик входящих пакетов, прежде чем определить, произошел ли отказ адаптера.

Создание файла netmon.cf

Файл netmon.cf должен находиться в каталоге /usr/es/sbin/cluster на всех узлах кластера.

Требования к созданию файла:

- Файл должен содержать по одному IP-адресу или IP-метке на кабель.
- Файл должен содержать удаленные IP-метки/адреса, не входящие в конфигурацию кластера, к которым можно получить доступ через интерфейсы HACMP. Мы рекомендуем использовать IP-адрес маршрутизатора.
- В файле netmon.cf может быть определено не более 30 IP-адресов/меток.
- Включите все IP-адреса и соответствующие им метки в файл /etc/hosts.

> **Примечание.** Когда процесс NIM (из служб топологии RSCT) пытается определить состояние локальных адаптеров, он может попытаться использовать разрешение имен хостов. Если IP-адрес и соответствующая метка не записаны в файле `/etc/hosts`, то при возникновении проблемы или при задержке при попытке разрешения адреса может быть продлено общее время обнаружения отказа, что замедляет операции перемещения при сбое.

Содержание файла netmon.cf может иметь примерно следующий вид:

```
/usr/es/sbin/cluster/netmon.cf
192.168.100.76
p690_1_lpar3
192.168.100.35
router_lan1
/etc/hosts (corresponding entries)
192.168.100.76      node365       #client node running oracle
192.168.100.21      p690_1_lpar3  #client node hosting application 4
192.168.100.35      node367       #client node running db2
192.168.100.200     router_lan1   #router hosting production lan
```

Рекомендации и дополнительные замечания

Основное правило состоит в том, чтобы реализовать файл netmon.cf в конфигурации кластера из двух узлов, используя одну IP-сеть, вне зависимости от реализации последовательной сети пульса, отличной от IP. Это необходимо для определения службами топологии отказа адаптера, если он переходит в состояние единственного адаптера, при котором в кластере остается только один узел.

Другие сценарии могут включать среды с использованием одного логического интерфейса EtherChannel, состоящего из нескольких связей Ethernet. В такой среде происходит прозрачная обработка отказов связей логикой EtherChannel, однако полный отказ канала приводит к возникновению проблем в случае, если службы топологии не имеют файла netmon.cf. Реализация EtherChannel в среде HACMP подробно описывается в разделе «Etherchannel».

14.4 Общее представление о файле clhosts

Файл clhosts содержит информацию об IP-адресах, позволяющую установить связь между демонами мониторинга на клиентах и на узлах кластера HACMP. К инструментам, использующим этот файл, относятся: clinfoES, HAView и clstat. Файл находится на всех серверах и клиентах кластера HACMP в каталоге /usr/es/sbin/cluster/etc/.

При запуске демона мониторинга он считывает файл /usr/es/sbin/cluster/etc/clhosts, чтобы определить узлы, доступные для связи. Поэтому при использовании инструментов мониторинга с клиента вне кластера важно, чтобы эти файлы были в системах. При установке серверной части HACMP происходит обновление файла clhosts на узлах кластера с применением адреса loopback (127.0.0.1). На каждом узле кластера файл обычно содержит только следующую строку:

127.0.0.1 # HACMP/ES for AIX

Создание файла clhosts

В предыдущих версиях требовалось вручную создавать и обслуживать клиентский файл clhosts. В HACMP 5.3 можно автоматически генерировать файл clhosts, требуемый клиентами при выполнении верификации с включенной функцией автоматического исправления ошибок. При верификации на всех узлах кластера создается файл /usr/es/sbin/cluster/etc/clhosts.client.

Этот файл будет иметь приблизительно следующий вид:

```
# /usr/es/sbin/cluster/etc/clhosts.client Created by HACMP Verification / Synchronization Corrective Actions
# Date Created: 07/01/2005 at 12:45:29
192.168.100.102 #dlpar_app2_svc
192.168.100.101 #dlpar_app1_svc
192.168.202.204 #alexis_base1
192.168.202.205 #alexis_base2
192.168.100.61 #alexis
192.168.201.203 #jessica_base2
192.168.201.202 #jessica_base1
192.168.100.72 #jessica
192.168.200.200 #jordan_base1
192.168.200.201 #jordan_base2
192.168.100.71 #jordan
```

Обратите внимание на то, что указаны все адреса, включая загрузочную, сервисную и постоянную IP-метки. Перед использованием любой из утилит мониторинга с клиентского узла должно выполняться копирование файла clhosts.client на все клиенты в каталог /usr/es/sbin/cluster/etc/clhosts. Не забудьте удалить расширение .client при копировании файла на клиентские узлы.

Важно! Файл clhosts на клиенте не должен содержать 127.0.0.1, loopback или localhost.

Clstat на клиенте и файл clhosts

При запуске утилиты clstat с клиента демон clinfoES получает информацию о состоянии кластера от SNMP на стороне сервера и заполняет MIB HACMP на стороне клиента. Она не сможет вести обмен данными с демоном и сообщит о том, что она не может найти кластеры при отсутствии файла clhosts.

В среде такого типа необходимо создать файл clhosts на клиенте. Этот файл предоставит демону clinfoES адреса для связи с процессом SNMP, выполняющемся на узлах кластера HACMP.

> **Ограничение.** При использовании перехвата IP-адресов посредством замены не следует включать дежурные адреса в файл clhosts.

HAView и файл clhosts

HAView выполняет мониторинг состояния кластера в топологии сети, используя информацию кластера в файле /usr/es/sbin/cluster/etc/clhosts. Он должен существовать на управляющем узле Tivoli NetView. Убедитесь в том, что имя хоста и сервисная метка узлов Tivoli NetView полностью одинаковы (если они неодинаковы, следует добавить синоним в файл /etc/hosts, чтобы устранить различия в именах). Если файл clhosts содержит недопустимый IP-адрес, HAView не сможет выполнить мониторинг кластера. Убедитесь в том, что применяются допустимые IP-адреса и что файл не содержит посторонних символов.

14.5 Общее представление о файле clinfo.rc

HACMP может быть настроен на изменение MAC-адреса сетевого интерфейса путем реализации перехвата аппаратного адреса (hardware address takeover, HWAT). В коммутируемой сетевой среде Ethernet коммутатор может не всегда получать своевременную информацию о новом MAC-адресе. В результате коммутатор не сможет направлять требуемые пакеты на сетевой интерфейс.

Скрипт clinfo.rc используется HACMP для очистки ARP-кеша в системе, чтобы отразить изменения в сетевых IP-адресах. Он не обновляет кеш до тех пор, пока другой адрес не ответит на ping-запрос. Обычно при включенной функции переключения аппаратных адресов в HACMP очистка ARP-кеша необязательна. Это связано с тем, что функция переключения аппаратных адресов устанавливает связь между IP-адресом и аппаратным адресом.

> **Примечание.** HWAT поддерживается в HACMP только при использовании перехвата IP-адресов посредством замены (IPAT via replacement).

На клиентах, на которых не выполняется clinfoES, может потребоваться выполнить косвенное обновление локального ARP-кеша посредством ping-опроса клиента с узла кластера. Чтобы этого избежать, следует добавить имя или адрес клиентского хоста, которого следует известить, в переменную PING_CLIENT_LIST в скрипте clinfo.rc. Программа clinfoES вызывает скрипт /usr/es/sbin/cluster/etc/clinfo.rc при

возникновении события сети или узла. Используя переменную PING_CLIENT_LIST, записи файла clinfo.rc могут выполнить обновление ARP-кешей клиентов и сетевых устройств, в частности маршрутизаторов.

При возникновении события кластера, clinfo.rc выполняет следующую команду для каждого хоста, указанного в переменной PING_CLIENT_LIST:

```
ping -c1 $host
```

> **Примечание.** Подразумевается, что клиент подключен напрямую к одной из сетей кластера.

Конфигурирование файла clinfo.rc

Чтобы убедиться в том, что новый MAC-адрес передан коммутатору, нужно выполнить следующие действия:

1. Измените строку файла /usr/es/sbin/cluster/etc/clinfo.rc, которая на данный момент выглядит следующим образом:

PING_CLIENT_LIST=" "

2. Включите в эту строку имена или IP-адреса как минимум одного клиента из каждой подсети в коммутируемой сети Ethernet.

3. Запустите clinfoES на всех узлах в кластере HACMP, подключенных к коммутируемой сети Ethernet.

Не забудьте выполнить следующее:

- если вы обычно запускаете службы кластера HACMP, используя shell-скрипт /usr/es/sbin/cluster/etc/rc.cluster, укажите опцию -i;
- если вы обычно запускаете службы кластера HACMP через SMIT, укажите Yes в поле Start Cluster Information Daemon? (Запускать демон информации кластера?);
- копия скрипта /usr/es/sbin/cluster/etc/clinfo.rc должна существовать на всех серверах и клиентах в кластере, чтобы обеспечить обновление всех ARP-кешей.

Принцип работы clinfo и clinfo.rc

Вызов clinfo.rc из clinfo имеет следующий формат:

clinfo.rc {join,fail,swap} interface_name

Clinfo получает информацию об интерфейсах и их текущем состоянии и выполняет проверку на изменение состояния интерфейсов:

- при состоянии UP Clinfo вызывает clinfo.rc join interface_name;
- при состоянии DOWN Clinfo вызывает clinfo.rc fail interface_name;
- при получении события node_down_complete Clinfo вызывает clinfo.rc с параметром fail для всех интерфейсов, находящихся в состоянии UP;
- при получении события fail_network_complete Clinfo вызывает clinfo.rc с параметром fail для всех соответствующих интерфейсов;
- при получении события swap_complete Clinfo вызывает clinfo.rc swap interface_name.

Часть 5

Аварийное восстановление

Часть 5 охватывает вопросы, связанные с HACMP Extended Distance (HACMP/XD). Рассматриваемые в этой части темы включают:

- понятия и планирование HACMP Extended Distance;
- HACMP и Cross-Site LVM;
- сценарий аварийного восстановления в HAGEO;
- концепции и конфигурирование GLVM.

15

Концепции и планирование HACMP Extended Distance

В этой главе рассматриваются функции и возможности HACMP Extended Distance (HACMP/XD) и описаны установка и конфигурирование некоторых функций аварийного восстановления в HACMP/XD.

Мы рассмотрим следующие вопросы:

- компоненты HACMP/XD;
- аспекты аварийного восстановления;
- дополнительные сведения.

15.1 Компоненты HACMP/XD

Базовый программный продукт High Availability Cluster Multi-Processing for AIX (HACMP) лишь частично разрешает задачу обеспечения непрерывной работы. Он выполняет восстановление после отказа узла, адаптера или локальной сети в пределах вычислительного комплекса на одном сайте.

HACMP/XD расширяет базовые возможности HACMP, обеспечивая поддержку автоматического перемещения при сбое и возврата после восстановления для приложений в географически распределенных системах. Системы, работающие в различных местах, определяются как узлы HACMP, приписанные к сайтам, и HACMP осуществляет для них управление как для обычных узлов.

Основной функцией HACMP/XD является репликация данных между сайтами. Для выполнения этой функции HACMP/XD может использовать несколько компонентов:

- HAGEO,
- PPRC,
- GLVM.

15.1.1 HACMP/XD HAGEO

Программное обеспечение HAGEO/GeoRM является оригинальным решением для репликации данных по сетям TCP/IP. Первоначально оно создавалось как отдельный продукт, который можно было использовать либо как автономную версию (GeoRM) только для репликации данных, либо как интегрированную версию, при использовании совместно с HACMP, для обеспечения функций автоматического перемещения при сбое и возврата после восстановления на сайте для приложений, использующих реплицируемые данные. Начиная с HACMP 5.1 HAGEO входит в программное обеспечение HACMP/XD.

Стандартная схема кластера HACMP/XD High Availability Geographic Cluster (HAGEO) представлена на рис. 15.1.

Функции HAGEO:

- Конфигурирование кластера с географически распределенными сайтами. HAGEO расширяет HACMP таким образом, чтобы охватить два географически распределенных информационных центра или сайта. В результате такого расширения отдельный сайт перестает быть единой точкой отказа в кластере.

Процесс географического зеркального отображения (geo-mirroring) создает на каждом сайте обновленную копию важных данных.

Любой из сайтов может выполнять основные приложения, обеспечивая постоянную доступность критически важных вычислительных ресурсов на географически распределенных сайтах в случае, если отказ или авария выведет из строя один из сайтов.

- Автоматические обнаружение и оповещение об отказах. HAGEO работает с HACMP, обеспечивая автоматическое обнаружение отказа сайта или географической сети. Он инициирует процесс восстановления и оповещает системного администратора обо всех обнаруживаемых отказах и предпринимаемых ответных действиях.

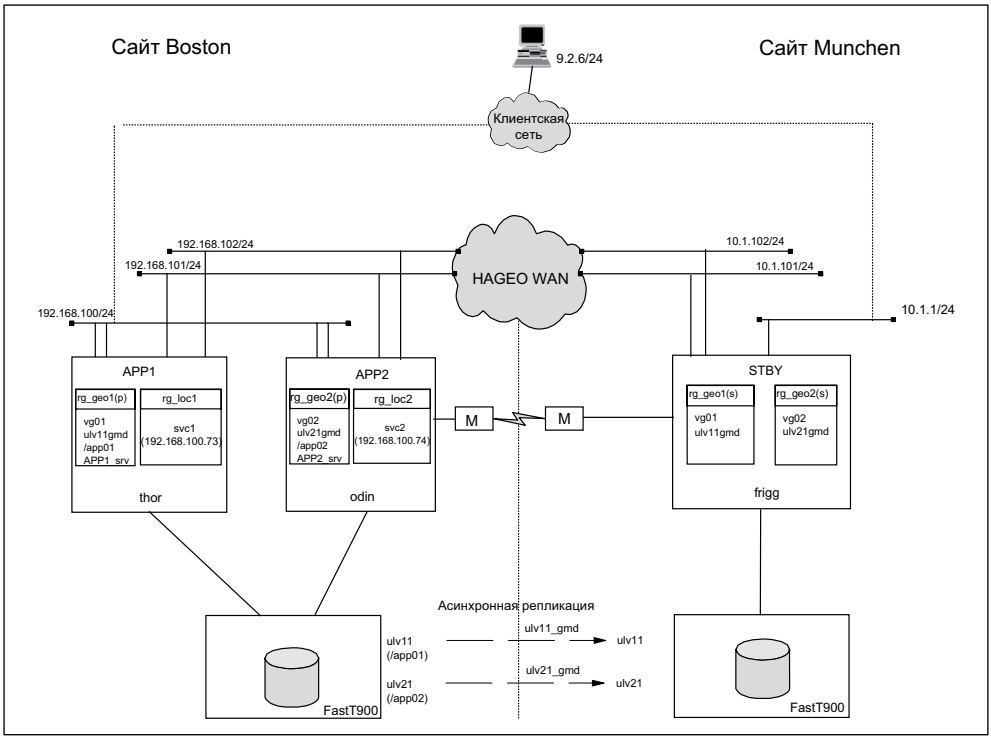

Рис. 15.1. Пример конфигурации HACMP/XD HAGEO

- Автоматическое перемещение при сбое. HAGEO включает скрипты обработки событий, предназначенные для выполнения восстановления после отказа сайта или географической сети. Эти скрипты интегрированы в стандартные скрипты обработки событий HACMP.

Можно настроить работу своей конфигурации путем добавления скриптов пред- или постобработки событий так же, как и в HACMP.

- Быстрое восстановление после аварии данных и приложений на рабочем сайте. Процесс географического зеркального отображения обеспечивает доступность данных на втором сайте при возникновении аварии. Восстановление обычно занимает несколько минут, не включая времени восстановления приложения.
- Автоматическая ресинхронизация данных при восстановлении сайта. HAGEO осуществляет ресинхронизацию зеркальных отображений на каждом сайте в ходе процесса восстановления сайта. Узлы на реинтегрируемом сайте автоматически обновляются с использованием данных, полученных, когда сайт находился в состоянии отказа.
- Надежное обеспечение целостности и согласованности данных. Компоненты географического зеркального отображения и географического обмена сообщениями в HAGEO в случае отказа сайта обеспечивают согласованность данных между оставшимся сайтом и отказавшим сайтом.

При реинтеграции отказавшего сайта в кластер HAGEO обновляет этот сайт актуальными данными с работающего сайта, что опять же обеспечивает согласованность данных.

- Поддержка широкого диапазона конфигураций, что позволяет настроить решение аварийного восстановления специально под ваши потребности. Кластер HAGEO может иметь до восьми узлов с различным количеством узлов на каждом сайте.

HAGEO не зависит от файловой системы и базы данных, так как устройство географического зеркального отображения работает так же, как и поддерживаемые им дисковые устройства. Так как зеркальное отображение является прозрачным, приложения, настроенные на использование географического зеркального отображения, не требуют каких-либо изменений.

Компоненты HAGEO

Программное обеспечение содержит три важных компонента:

- GeoMirror. Состоит из логического устройства и драйвера псевдоустройства, осуществляющего зеркальное отображение второго сайта; данные вводятся на одном сайте. Для транспортировки данных зеркального отображения используется TCP/IP. GeoMirror можно применять в синхронном и асинхронном режимах, в зависимости от пропускной способности связи между сайтами и объема транзакций приложений (определяющего количество измененных данных).
- GeoMessage. Обеспечивает надежную доставку данных и сообщений между устройствами GeoMirror на двух сайтах. GeoMessage представляет систему обмена сообщениями между процессами ядра, которую драйвер устройств GeoMirror использует для отправки и получения сообщений через IP-сети. GeoMessage может применять UDP или TCP в качестве транспортного протокола в IP-сети.
- Географическая топология. Обеспечивает логику для интеграции функций географического отображения с функциями HACMP для обеспечения автоматического обнаружения отказов и восстановления после событий, оказывающих влияние на целые сайты. Этот компонент включает:
  - скрипты и программы, встраивающие обработку GeoMirror и GeoMessage в события кластера, такие, как подключения и отказы узлов и сетей;
  - скрипты, встраивающие запуск и остановку функций GeoMirror и GeoMessage в скрипты запуска и остановки HACMP;
  - сообщения журнала ошибок, обеспечивающие регистрацию операций GeoMirror и GeoMessage.

Базовые конфигурации HACMP/XD HAGEO

Можно настроить любую конфигурацию кластера HAGEO, поддерживаемую базовым программным обеспечением HACMP. Это включает дежурные конфигурации, конфигурации с перехватом и конфигурации с одновременным доступом.

- Дежурные конфигурации. Дежурная конфигурация представляет традиционную конфигурацию с дублированием оборудования, где один или несколько узлов в кластере бездействуют до тех пор, пока не произойдет отказ серверного узла. В HAGEO это означает наличие бездействующего сайта. Сайт не является совсем бездействующим, так как он тоже может быть вовлечен в процесс географического зеркального отображения. Однако узлы на этом сайте не выполняют приложений.
- Конфигурации с перехватом:
  - конфигурации с перехватом внутри сайта (локальный);
  - конфигурации с удаленным односторонним перехватом;
  - конфигурации с удаленным взаимным перехватом.

  В конфигурации с перехватом все узлы осуществляют обработку; бездействующих узлов не существует.
- Конфигурации с одновременным доступом. В конфигурации с одновременным доступом все узлы на одном сайте имеют одновременный доступ к группе томов с одновременным доступом и владеют одинаковыми дисковыми ресурсами. Другой сайт настроен таким же образом. Выход узла из сайта не влияет на доступность ресурсов, так как на других узлах активизирована группа томов с одновременным доступом. При отказе сайта другой сайт обеспечивает одновременный доступ на узлах сайта. К приложению с одновременным доступом могут получить доступ все узлы в кластере. Диспетчер блокировки кластера (Cluster Lock Manager) должен быть запущен на всех узлах в кластере. Не все базы данных можно использовать для одновременного доступа к географически распределенным узлам.

15.1.2 Функция интеграции HACMP/XD PPRC

Эта функция впервые появилась одновременно в HACMP V4.5 PTF5 и в HACMP V5.1 и обеспечивает автоматическое перемещение при отказе сайта и активизацию удаленных копий данных приложения в среде, в которой на обоих сайтах используется IBM Enterprise Storage Server (ESS) и в которой зеркальное отображение дисковых томов обеспечивается функцией Peer to Peer Remote Copy (PPRC).

Стандартная конфигурация HACMP/XD PPRC представлена на рис. 15.2.

В случае отказа основного сайта данные должны быть доступны для использования на дополнительном сайте (реплицируются с использованием PPRC). Для использования копии данных на дополнительном сайте она должна быть активизирована.

Функция интеграции HACMP/XD PPRC обеспечивает автоматическое разделение копий в случае отказа основного сайта и автоматическую реинтеграцию при восстановлении основного сайта.

Дополнительные сведения см. в руководстве **HACMP/XD for ESS PPRC Version 5.3: Planning and Administration Guide**, SC23-4863.

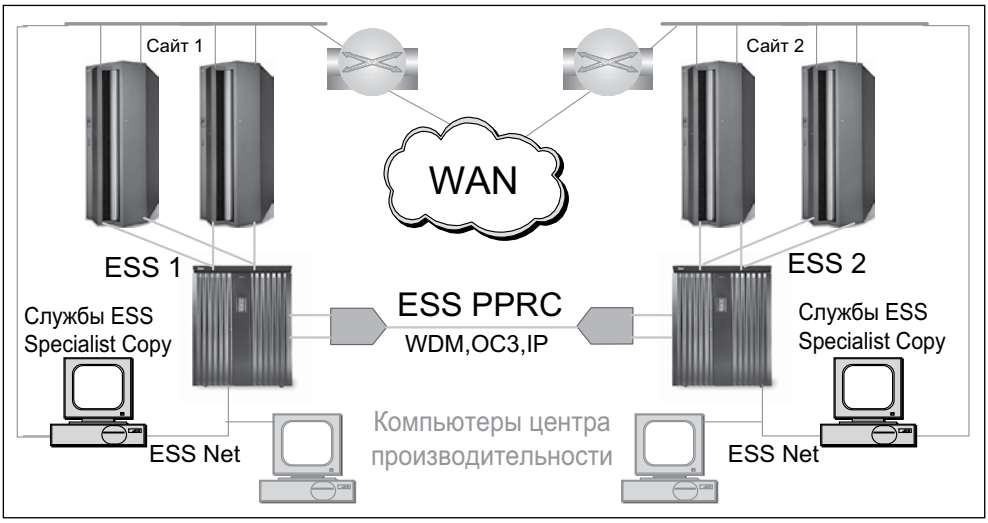

Рис. 15.2. Пример конфигурации HACMP/XD PPRC

15.2 Аспекты аварийного восстановления

Стратегии аварийного восстановления охватывают широкий диапазон от полной неготовности к восстановлению до автоматического восстановления с обеспечением высокой целостности данных. Стратегии восстановления данных должны учитывать следующие аспекты:

- Уровень готовности данных:
  - Уровень 0: неготовность. Аварийное восстановление не реализовано.
  - Уровень 1: периодическое резервное копирование. Данные, необходимые для восстановления на заданную дату, сохраняются и пересылаются на другой сайт.
  - Уровень 2: готовность к восстановлению с повтором всех завершенных транзакций. Помимо периодических резервных копий, на другой сайт также пересылаются журналы обновления данных. Транспортировка может осуществляться вручную или электронным путем. Можно осуществить восстановление до последнего набора данных журналов, хранящихся на сайте восстановления.
  - Уровень 3: восстановление с повтором всех завершенных транзакций. На дисках сайта восстановления поддерживается теневая копия данных. Журналы обновления данных принимаются и периодически применяются к теневой копии с использованием утилит восстановления.
  - Уровень 4: восстановление с повтором всех завершенных транзакций в реальном времени. Подобно восстановлению с повтором всех завершенных транзакций с тем исключением, что передача и применение обновлений происходит в то же время, когда они регистрируются на исходном сайте. Такая передача и применение данных журналов в реальном времени не влияет на время отклика транзакции на исходном сайте.

  - Уровень 5: удаленное обновление в реальном времени. И оригинальная и резервная копии данных обновляются перед отправкой отклика транзакции или перед выполнением задачи.

- Взаимное подключение между сайтами:
  - Уровень 0: отсутствует. Между сайтами отсутствует подключение и транспортировка данных.
  - Уровень 1: транспортировка вручную. Взаимное подключение отсутствует. Транспортировка данных между сайтами, управление передачей, отслеживанием и получением данных осуществляется вручную.
  - Уровень 2: удаленная магнитная лента. Данные транспортируются электронным путем на удаленную магнитную ленту. Передача и получение выполняются автоматически. Отслеживание может выполняться либо автоматически, либо вручную.
  - Уровень 3: удаленный диск. Данные транспортируются электронным путем на удаленный диск. Передача, получение и отслеживание выполняются автоматически.

- Готовность сайта восстановления:
  - Холодный сайт. Обычно представляет среду с требуемой инфраструктурой, но без оборудования для обработки данных. Это оборудование необходимо установить в первую очередь в процессе восстановления данных. Данные уровня 1 (периодическое резервное копирование) и уровня 2 (готовность к восстановлению с повтором всех завершенных транзакций) могут быть переданы из места хранения на этот сайт при возникновении аварии.
  - Теплый сайт. Содержит установленное и работающее оборудование обработки данных. До возникновения аварии это оборудование используется для выполнения других задач обработки данных. Ресурсы обработки данных можно использовать для хранения данных, например журналов. Восстановление начинается после завершения работы сайта и выполнения резервного копирования. Данные уровня 1 (периодическое резервное копирование) и уровня 2 (готовность к восстановлению с повтором всех завершенных транзакций) могут быть сохранены на этом сайте для ускорения аварийного восстановления.
  - Горячий сайт. Содержит установленное и работающее оборудование обработки данных, и восстановление данных может осуществляться либо непрерывно, либо регулярно в целях сокращения времени восстановления. Можно реализовать все уровни от уровня 3 (восстановление с повтором всех завершенных транзакций) до уровня 5 (удаленное обновление в реальном времени).

Программное обеспечение HACMP/XD обеспечивает наивысший уровень аварийного восстановления:

- Уровень готовности данных 5: HACMP/XD обеспечивает готовность удаленных данных обновления в реальном времени путем обновления как оригинальной копии данных, так и копии восстановления перед отправкой отклика транзакции или перед выполнением задачи.

- Уровень взаимного подключения между сайтами 3: HACMP/XD также обеспечивает подключение между сайтами с использованием удаленного диска путем электронной передачи данных на географически удаленный сайт, на котором обновление дисков и весь учет ведется автоматически.
- Горячая готовность сайта. Так как сайт восстановления содержит рабочее оборудование обработки данных вместе с актуальными данными, это позволяет свести время восстановления к минимуму. В случае использования функции HACMP PPRC осуществляется управление зеркальным отображением данных с помощью функции PPRC на уровне подсистемы хранения.

При использовании HACMP/XD сайт восстановления может активно обрабатывать данные и выполнять полезную работу. В действительности каждый сайт может представлять резервную копию другого сайта, что сокращает стоимость установки сайта восстановления для каждого исходного рабочего сайта.

15.3 Дополнительные сведения

Дополнительные сведения см. в следующих публикациях:

- **High Availability Cluster Multi-Processing XD (Extended Distance) for HA-GEO Technology: Concepts and Facilities**, SA22-7955.
- **HACMP/XD for ESS PPRC Version 5.3: Planning and Administration Guide,** SC23-4863.
- **HACMP/XD for Geographic LVM: Administration and Planning Guide**, SA23-1338.

16

HACMP и Cross-Site LVM

В этой главе описывается система аварийного восстановления на основе AIX и базового кластера HACMP. Она построена с использованием тех же компонентов, которые обычно применяются в локальных кластерных решениях с подключением к хранилищу SAN.

Межсайтовое зеркальное отображение Cross-Site LVM mirroring осуществляет репликацию данных между дисковыми подсистемами на каждом сайте.

16.1 Введение в межсайтовое зеркальное отображение Cross-Site LVM mirroring

Сеть хранения данных (storage area network, SAN) представляет собой высокоскоростную сеть, позволяющую реализовать прямые подключения между устройствами хранения и серверами. Максимальное расстояние подключений определяется ограничениями Fibre Channel. Это позволяет двум и более серверам, расположенным на разных сайтах, осуществлять доступ к одним и тем же физическим дискам.

Эти удаленные диски могут быть объединены в группу томов через AIX 5L Logical Volume Manager (LVM), после чего эту группу томов можно импортировать на узлах, расположенных на разных сайтах. Можно создать логические тома и установить на каждом сайте зеркальное отображение LVM с копией. HACMP поддерживает не больше двух активных сайтов в межсайтовом зеркальном отображении Cross-Site LVM mirroring.

Основное различие между обычными локальными кластерами и кластерными решениями с межсайтовым зеркальным отображением состоит в следующем. В локальных кластерах все узлы и подсистемы хранения расположены в одном месте. В кластере с межсайтовым зеркальным отображением узлы и подсистемы хранения располагаются в разных местах. Каждый сайт имеет как минимум один узел кластера и одну подсистему хранения со всей необходимой инфраструктурой IP-сети и SAN.

Такое решение обеспечивает автоматизацию зеркального отображения AIX LVM в дисковых подсистемах SAN между разными сайтами. Оно также обеспечивает автоматическую синхронизацию зеркальных отображений LVM и активизацию дисковых устройств, когда после отказа диска или сайта узел или диск становится недоступным.

Каждый узел в кластере Cross-Site LVM имеет доступ ко всем подсистемам хранения. Доступность данных обеспечивается посредством зеркального отображения LVM между томами, находящимися в различных подсистемах хранения на разных сайтах.

Рис. 16.1 представляет среду межсайтового отображения LVM с двумя сайтами, которую мы использовали при тестировании межсайтового зеркального отображения LVM.

В случае отказа сайта HACMP выполняет перехват ресурсов на дополнительном сайте в соответствии с установленной политикой кластера. Он активизирует все определенные группы томов с оставшейся зеркальной копии. В случае отказа одной подсистемы хранения доступ к данным не нарушается и приложения могут получить доступ к данным, расположенным на активной зеркальной копии в оставшейся дисковой подсистеме.

HACMP выполняет автоматическую синхронизацию зеркальных отображений LVM, и после интеграции отказавшего сайта в кластер он автоматически исправляет удаленные и отсутствующие тома (с состоянием PV **removed** и **missing**) и осуществляет синхронизацию данных. Автоматическая синхронизация в любом случае невозможна, однако вы можете использовать C-SPOC для синхронизации данных с рабочих зеркальных отображений на неактуальные (stale) зеркальные отображения после отказа диска или сайта.

16.1.1 Требования

Для обеспечения целостности данных и соответствующей реакции HACMP при отказе сайта или дисковой подсистемы необходимо выполнение следующих требований:

- атрибут принудительной активизации force varyon для группы ресурсов должен быть установлен в значение true;
- должна быть установлена политика размещения логических томов «super strict» (это обеспечит размещение копий логических томов на разных томах, а также размещение основной и дополнительной копии каждого логического раздела (LP) на дисках, расположенных на разных сайтах);
- зеркальные копии логических томов должны быть расположены на отдельных томах, расположенных в разных дисковых подсистемах (или на разных сайтах).

При увеличении размера файловой системы с зеркальным отображением необходимо обеспечить выделение нового логического раздела на различных томах и в разных дисковых подсистемах в соответствии с требованиями, приведенными выше. Для выполнения этой задачи необходимо сначала увеличить размер логического тома, выбрав соответствующий том, а затем увеличить размер файловой системы, желательно с использованием C-SPOC (в этом случае HACMP выполнит принудительное применение установленных параметров).

Перед конфигурированием среды зеркального отображения Cross-Site LVM mirroring проверьте выполнение следующих требований:

- сконфигурируйте сайты и группы ресурсов и запустите процесс обнаружения кластера HACMP;
- убедитесь в том, что оба сайта имеют копии логических томов и что для атрибута **forced varyon (принудительная активизация)** группы томов установлено значение **Yes**, если группа ресурсов уже содержит группу томов.

16.2 Аспекты инфраструктуры

В этом разделе описываются некоторые аспекты, связанные с установкой SAN, подключениями Fibre Channel и средой локальной сети. Аспекты и ограничения основаны на технологиях и протоколах, используемых для реализации кластера с межсайтовым зеркальным отображением.

SAN или сеть может быть расширена за пределами первоначального сайта с использованием передовых технологий. См. SG-245250-04[1].

Ниже перечислены технологии, которые можно использовать для такого расширения. Этот список не является исчерпывающим.

- Маршрутизатор FCIP.
- Устройства WDM (Wave Division Multiplexing). Эта технология включает:
  - CWDM (Coarse Wavelength Division Multiplexing) – менее дорогостоящий компонент технологии WDM;
  - DWDM (Dense Wavelength Division Multiplexing).

[1] Имеется в виду руководство (IBM Redbook) «IBM System Storage Solutions Handbook». *Примеч. науч. ред.*

16.3 Конфигурирование межсайтового зеркального отображения Cross-Site LVM mirroring

В этом разделе мы представим пример установки и конфигурирования среды зеркального отображения Cross-Site LVM mirroring. В целом можно выполнить конфигурирование зеркального отображения Cross-Site LVM mirroring при создании нового кластера. Также можно изменить существующий локальный кластер путем добавления зависимостей сайтов и функций Cross-Site LVM в конфигурацию кластера и их интеграции в межсайтовой среде.

16.3.1 Конфигурирование кластера Cross-Site LVM

Для тестирования зеркального отображения Cross-Site LVM мы решили выполнить конфигурирование новой кластерной среды. Через меню **Extended Configuration (Расширенное конфигурирование)** мы сначала определяем топологию кластера, включая определение кластера, узлы, сети, сетевые интерфейсы и сети мониторинга пульса через диски, отличные от IP, как в обычной кластерной среде. У нас есть узлы **koper** и **nantes**; **koper** расположен на сайте **Slovenia**, тогда как **nantes** расположен на сайте **France** (рис. 16.1).

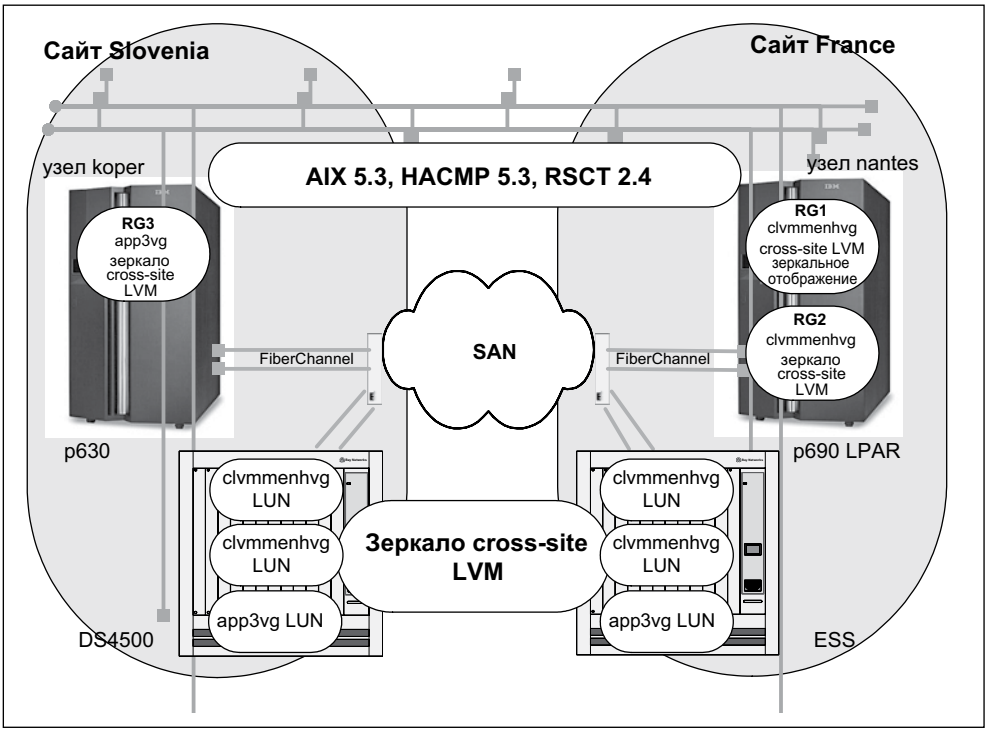

Рис. 16.1. Среда тестирования кластера с зеркальным отображением Cross-Site LVM mirroring

Важным элементом в межсайтовых средах являются пути для связи между узлами на обоих сайтах. Связь между узлами включает IP-подключения и подключения, отличные от IP. Подключения, отличные от IP, очень важны в межсайтовых кластерных системах, так как они позволяют не допустить изоляции узла или сайта («split brain»). Мы сконфигурировали сеть пульса, отличную от IP, с использованием функции мониторинга пульса через диски.

Следуя общим рекомендациям для сетей пульса, отличных от IP (см. раздел 3.8, «Планирование сетей кластера», мы определили две сети пульса, отличные от IP. Первая сеть использует дисковые устройства в хранилище ESS, расположенном на сайте **France**. Вторая сеть применяет дисковые устройства в хранилище FAStT/DS4xxx, расположенном на сайте **Slovenia**. Такое дублирование устанавливается для того, чтобы сохранить работоспособность отличных от IP сетей пульса в кластере в случае отказа одной дисковой подсистемы.

Рис. 16.1 представляет нашу тестовую среду для тестирования межсайтового зеркального отображения Cross-Site LVM mirroring. В следующих разделах подробно описываются этапы конфигурирования Cross-Site LVM для определенной конфигурации кластера.

16.3.2 Конфигурирование сайтов кластера

При конфигурировании межсайтового зеркального отображения LVM или любого из компонентов HACMP/XD следует использовать определения сайта в HAMCP.

Мы выполнили конфигурирование сайтов и добавили их в конфигурацию кластера, используя меню SMIT. Мы выбрали **smit hacmp > Extended Configuration (Расширенное конфигурирование) > Extended Topology Configuration (Расширенное конфигурирование топологии) > Configure HACMP Sites (Конфигурирование сайтов HACMP) > Add a Site (Добавление сайта)**. Мы добавили два сайта: France и Slovenia. Узел koper является частью сайта Slovenia, тогда как узел nantes является частью сайта France. На рис. 16.2 представлено меню создания сайта в SMIT.

```
                                Add Site

Type or select values in entry fields.
Press Enter AFTER making all desired changes.
                                                      [Entry Fields]
* Site Name                                          [France]          +
* Site Nodes                                          nantes           +
* Dominance                                          [Yes]             +
* Backup Communications                              [none]            +


F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Рис. 16.2. Добавление сайта в конфигурацию топологии

Для определения сайта в меню SMIT необходимо заполнить следующие поля:

- Site Name (Имя сайта). Определите имя сайта, используемое для назначения различных зависимостей сайта при последующем конфигурировании.
- Site Nodes (Узлы сайта). Для каждого сайта мы определяем список узлов, расположенных на сайте.
- Dominance (Доминирование). Выбрав **Yes**, мы определяем, что этот сайт должен быть доминирующим при изоляции сайтов.
- Backup Communications (Резервные связи). Следует определить тип резервных связей для обнаружения изоляции сайта (none, sgn, dbfs).

16.3.3 Конфигурирование зависимостей сайтов с зеркальным отображением Cross-Site LVM

После определения топологии кластера с зависимостями сайтов выполняется назначение определенных дисковых устройств соответствующим сайтам. На сайте **France** установлено хранилище ESS 2105-800, тогда как на сайте **Slovenia** установлено хранилище DS4500. Серверы используют MPIO для хранилища ESS и драйвер RDAC для хранилища DS4xxx. Конфигурация разделения SAN на зоны создается таким образом, чтобы каждый сервер мог осуществлять доступ к дисковой подсистеме ESS через адаптер Fiber Channel, тогда как доступ к хранилищу DS4xxx осуществляется через другой адаптер. В примере 16.1 представлен список дисков на узле **nantes**. Мы планируем использовать диски hdisk2 – hdisk6 и hdisk10 – hdisk14 для конфигурации зеркального отображения Cross-Site LVM, тогда как диски hdisks7 свободны для использования в группе томов без зеркального отображения.

Пример 16.1. Выходные данные lsdev -Cc на узле nantes

```
nantes /usr/es/sbin/cluster > lspv
hdisk0          0022be2ab1cd11ac                    rootvg          active
hdisk1          0022be2abc247c91                    altinst_rootvg
hdisk2          0022be2a8607249f                    clvmmenhvg      active
hdisk3          0022be2a0bfe1f60                    clvmmshvg       active
hdisk4          0022be2a86607918                    clvmmshvg       active
hdisk5          0022be2a08d4844e                    clvmmenhvg      active
hdisk6          0022be2a8662ce0e                    app3vg
hdisk7          0022be2a0bfe9e19                    None
hdisk8          00257400b4d32054                    None
hdisk9          0022be2a86630978                    None
hdisk10         0022be2a0bfe9eec                    clvmmenhvg      active
hdisk11         0022be2a0bfe9f8a                    clvmmenhvg      active
hdisk12         0022be2a0bfea026                    clvmmshvg       active
hdisk13         0022be2a0bfea0ca                    clvmmshvg       active
hdisk14         0022be2a11690be3                    app3vg
```

Мы используем SMIT для назначения зависимостей сайтов/дисков; запускаем **smit hacmp > System Management (C-SPOC) > HACMP Physical Volume Management (Управление физическими томами HACMP) > Configure Disk/Site Locations for Cross-Site LVM Mirroring (Конфигурирование расположения дисков/сай-**

тов при зеркальном отображении Cross-Site LVM) > Add Disk/Site Definition for Cross-Site LVM Mirroring (Добавление определения дисков/сайтов для зеркального отображения Cross-Site LVM). Для прямого доступа к меню **Configure Disk/Site Locations for Cross-Site LVM Mirroring (Конфигурирование расположения дисков/сайтов при зеркальном отображении Cross-Site LVM)** можно использовать быстрый путь **smit cl_xslvmm**.

> **Примечание.** Меню Configure Disk/Site Locations for Cross-Site LVM Mirroring (Конфигурирование расположения дисков/сайтов при зеркальном отображении Cross-Site LVM) функционирует корректно только в том случае, если файл обнаружения дисков (Disk Discovery File) отражает текущую конфигурацию дисков. Мы рекомендуем выполнить процесс обнаружения HACMP, прежде чем выполнять конфигурирование зависимостей дисков/сайтов.

> **Внимание!** Всем дискам должны быть назначены соответствующие PVID, прежде чем запускать Discover HACMP-related Information from Configured Nodes (Обнаружение информации, относящейся к HACMP, на сконфигурированных узлах), чтобы иметь полную информацию о дисках в файле обнаружения дисков. Это можно сделать, используя команду chdev -l hdiskX -a pv=yes на одном узле с последующим удалением и реконфигурированием дисковых устройств на других узлах (rmdev..., cfgmgr).

После использования меню **Add Disk/Site Definition for Cross-Site LVM Mirroring (Добавление определения дисков/сайтов для зеркального отображения Cross-Site LVM)** сначала следует выбрать сайт для своего определения, как показано в примере 16.2. После этого нужно выбрать диски, расположенные на выбранном сайте, как показано в примере 16.3.

Пример 16.2. Выбор сайта в меню Add Site/Disk Definition (Добавление определения дисков/сайтов)

```
| Move cursor to desired item and press Enter.                          |

|   France
|   Slovenia
```

Пример 16.3. Выбор дисков в меню Add Site/Disk Definition (Добавление определения дисков/сайтов)

```
| [MORE...10]                                                           |
|   0022be2a0bfea026        hdisk13 koper                               |
|   0022be2a0bfea0ca        hdisk14 koper                               |
| > 0022be2a8607249f        hdisk2  nantes                              |
| > 0022be2a0bfe1f60        hdisk3  nantes                              |
| > 0022be2a86607918        hdisk4  nantes                              |
| > 0022be2a08d4844e        hdisk5  nantes                              |
|   0022be2a8662ce0e        hdisk6  nantes                              |
|   0022be2a0bfe9e19        hdisk7  nantes                              |
| [MORE...6]
```

Впоследствии можно изменить зависимость сайтов/дисков, выполнив **smit cl_xslvmm** и выбрав **Change/Show Disk/Site Definition for Cross-Site LVM Mirroring (Изменение/вывод определения дисков/сайтов для зеркального отображения Cross-Site LVM)**. Также можно удалить зависимость сайтов/дисков, выполнив **smit cl_xslvmm** и выбрав **Remove Disk/Site Definition for Cross-Site LVM Mirroring (Удаление определения дисков/сайтов для зеркального отображения Cross-Site LVM)**.

16.3.4 Настройка групп томов зеркального отображения cross-site LVM

Для тестирования межсайтового зеркального отображения Cross-Site LVM мы создали три группы томов:

- clvmmshvg – общая группа томов без одновременного доступа (shared non-concurrent);
- clvmmenhvg – группа томов с расширенным одновременным доступом (enhanced concurrent);
- app3vg – группа томов с расширенным одновременным доступом (enhanced concurrent).

Мы создали свои группы томов через меню C-SPOC. Мы запустили **smit cl_admin**, после чего выбрали **HACMP Logical Volume Management (Управление логическими томами HACMP)** для общей группы томов или **HACMP Concurrent Logical Volume Management (Управление логическими томами HACMP с одновременным доступом)** для обеих групп томов с расширенным одновременным доступом.

После выбора участвующих узлов мы выбрали диски, как показано в примере 16.4. Затем мы заполняем все необходимые поля на экране создания группы томов и устанавливаем для опции **Enable Cross-Site LVM Mirroring Verification (Включить верификацию зеркального отображения Cross-Site LVM)** значение **true**. В примере 16.5 представлен экран создания группы томов.

Пример 16.4. Выбор диска для создания группы томов в меню SMIT

```
| [MORE...1]                                                          |
| > 0022be2a86607918        France                                    |
| > 0022be2a0bfe1f60        France                                    |
| > 0022be2a0bfea026        Slovenia                                  |
| > 0022be2a0bfea0ca        Slovenia                                  |
|   0022be2a86630978                                                  |
|   0022be2a0bfe9f8a        Slovenia                                  |
|   0022be2a0bfe9eec        Slovenia                                  |
|   0022be2a0bfe9e19
```

В примере 16.5 представлено создание общей группы томов.

Пример 16.5. Экран создания группы томов

```
Create a Shared Volume Group

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

[TOP]                                                   [Entry Fields]
  Node Names                                           koper,nantes
  PVID                                                 0022be2a86607918 0022>
  VOLUME GROUP name                                   [clvmmshvg]
  Physical partition SIZE in megabytes                 128                    +
  Volume group MAJOR NUMBER                           [44]                     #
  Enable Cross-Site LVM Mirroring Verification         true                   +

  Warning:
  Changing the volume group major number may result
  in the command being unable to execute
  successfully on a node that does not have the
  major number currently available.  Please check
  for a commonly available major number on all nodes
[MORE...2]
```

16.3.5 Конфигурирование группы ресурсов, содержащей группу томов с поддержкой зеркального отображения Cross-Site LVM

После определения в кластере всех компонентов топологии и LVM мы выполнили создание групп ресурсов. Для двух групп ресурсов узел **nantes** является основным узлом. Для третьей группы ресурсов основным узлом является **koper**. Каждая группа ресурсов владеет одной группой томов, для которой осуществляется межсайтовое зеркальное отображение.

Создание группы ресурсов в межсайтовой кластерной среде подобно обычной кластерной конфигурации. Для групп ресурсов с зависимостями сайтов в кластерной среде, в которой выполняется определение сайтов, устанавливается дополнительный параметр **Inter-Site Management Policy (Политика межсайтового управления)**. Этот параметр может принимать следующие значения:

- **Ignore (Игнорировать).** Это значение установлено по умолчанию и игнорирует параметры зависимостей сайта для группы ресурсов.
- **Prefer Primary Site (Предпочтительное использование основного сайта).** Для группы ресурсов можно установить перехват несколькими сайтами приоритетным образом. При отказе сайта ресурс перехватывает активный сайт с наивысшим приоритетом. При реинтеграции отказавшего сайта ресурс получает сайт с наивысшим приоритетом.
- **Online On Either Site (Подключение на любом сайте).** Группа ресурсов может быть получена любым сайтом в цепочке ресурсов. При отказе сайта перехват группы ресурсов выполняется на дежурном сайте с наивысшим приоритетом. При реинтеграции отказавшего сайта группа ресурсов остается у нового владельца.

- **Online On Both Sites (Подключение на обоих сайтах).** Получение группы выполняется на обоих сайтах. Этот вариант определяет группу ресурсов с возможностью одновременного доступа.

После определения групп ресурсов выполняется конфигурирование атрибутов группы ресурсов, таких, как группа томов, сервисный IP-адрес и сервер приложения. Важным параметром при добавлении группы томов с поддержкой межсайтового зеркального отображения в группу ресурсов является поле **Use forced varyon of volume groups, if necessary (При необходимости использовать принудительную активизацию групп томов)**. Необходимо установить для этого поля значение **true** в любой конфигурации межсайтового зеркального отображения Cross-Site LVM. Это обеспечивает возможность активизации определенной группы томов на другом узле только с одной (оставшейся) копией логического тома в случае отказа хранилища или сайта.

После определения групп ресурсов и назначения групп томов в группы ресурсов выполняется добавление логических томов и файловой системы в группы томов. Конфигурирование выполняется обычным способом с использованием C-SPOC. Дополнительные сведения о создании компонентов LVM см. в разделе 8.4, «Управление общим хранением».

Для каждой группы ресурсов мы определяем сервер приложения. Сервер приложения выполняет тестовое приложение, которое осуществляет интенсивную запись в файловые системы заданной группы ресурсов. При такой нагрузке мы можем достичь показателя использования заданного диска между 70 и 100 %.

После настройки кластерной среды мы выполнили активизацию автоматического уведомления об ошибках. Мы выполнили **smit hacmp > Problem Determination Tools (Инструменты определения проблем) > HACMP Error Notification (Уведомления об ошибках HACMP) > Configure Automatic Error Notification (Конфигурирование автоматического уведомления об ошибках) > Add Error Notify Methods for Cluster Resources (Добавить методы уведомления об ошибках для ресурсов кластера)**. Дополнительные сведения о функции уведомления об ошибках см. в разделе 12.3, «Уведомление об ошибках».

> **Примечание:** DARE не поддерживается в активном кластере с определенными сайтами. Можно использовать C-SPOC для некоторых изменений в конфигурации кластера. Все остальные изменения в конфигурации необходимо выполнять при неактивном кластере.

16.4 Тестирование межсайтового зеркального отображения Cross-Site LVM

После конфигурирования топологии и ресурсов кластера мы выполняем синхронизацию и верификацию кластера. Мы запустили кластер и проверили, все ли ресурсы и пути для связи активны. Утилита **cldump** выводит информацию об узлах кластера и сетевых интерфейсах, а также о состоянии групп ресурсов (включая политики каждой группы ресурсов). Пример 16.6 содержит выходные данные команды **cldump** из нашей тестовой кластерной среды.

Пример 16.6. Выходные данные команды cldump в кластерной среде с межсайтовым зеркальным отображением

```
koper /usr/es/sbin/cluster/utilities > cldump

Obtaining information via SNMP from Node: nantes...

_____________________________________________________________________
Cluster Name: crossitelvm
Cluster State: UP
Cluster Substate: STABLE
_____________________________________________________________________


Node Name: koper                State: UP

  Network Name: net_diskhb_01      State: UP

    Address:                 Label: diskhb_sloveniatonantes State: UP

  Network Name: net_ether_01       State: UP

    Address: 10.10.1.9       Label: koper_base1        State: UP
    Address: 10.10.2.9       Label: koper_base2        State: UP
    Address: 192.168.100.117 Label: app3svc            State: UP


Node Name: nantes               State: UP

  Network Name: net_diskhb_01      State: UP

    Address:                 Label: diskhb_francetoslovenia State: UP

  Network Name: net_ether_01       State: UP

    Address: 10.10.1.2       Label: nantes_base1       State: UP
    Address: 10.10.2.2       Label: nantes_base2       State: UP
    Address: 192.168.100.86  Label: app1svc            State: UP
    Address: 192.168.100.87  Label: app2svc            State: UP



Cluster Name: crossitelvm

Resource Group Name: RG1
Startup Policy: Online On Home Node Only
Fallover Policy: Fallover To Next Priority Node In The List
Fallback Policy: Fallback To Higher Priority Node In The List
Site Policy: ignore
Priority Override Information:
```

```
    Primary Instance POL:
Node                          Group State
----------------------------- ---------------
nantes                        ONLINE
koper                         OFFLINE

Resource Group Name: RG2
Startup Policy: Online On Home Node Only
Fallover Policy: Fallover To Next Priority Node In The List
Fallback Policy: Fallback To Higher Priority Node In The List
Site Policy: ignore
Priority Override Information:
    Primary Instance POL:
Node                          Group State
----------------------------- ---------------
nantes                        ONLINE
koper                         OFFLINE

Resource Group Name: RG3
Startup Policy: Online On Home Node Only
Fallover Policy: Fallover To Next Priority Node In The List
Fallback Policy: Fallback To Higher Priority Node In The List
Site Policy: ignore
Priority Override Information:
    Primary Instance POL:
Node                          Group State
----------------------------- ---------------
koper                         ONLINE
nantes                        OFFLINE
```

16.4.1 Протестированные сценарии

Постепенная остановка с передачей ресурсов

Сначала мы выполняем тестирование постепенной остановки с передачей ресурсов (graceful shutdown with takeover) на узле **nantes**. Службы кластера на узле **nantes** останавливают и перемещают группы ресурсов RG1 и RG2, активизируя их на узле **koper** сайта **Slovenia**, как и ожидалось. После тестирования мы запустили узел **nantes**0ы и кластер инициировал перемещение групп ресурсов RG1 и RG2 обратно на основной узел (**nantes**), как и было определено в политике группы ресурсов. Все ресурсы в группах ресурсов RG1 и RG2 стали доступны на узле **nantes**, после чего приложения стали активными.

Перемещение группы ресурсов на другой сайт

В следующем тесте мы выбрали узел **nantes**, владеющий двумя группами ресурсов. Используя функцию C-SPOC **Move a Resource Group to Another Node / Site (Перемещение группы ресурсов на другой узел/сайт)**, мы переместили только группу ресурсов RG1 на узел **koper**. Все ресурсы группы ресурсов RG1 были перемещены на узел **koper**, тогда как группа ресурсов RG2 осталась активной на узле **nantes** без прерывания обслуживания, как и ожидалось.

Отказ одной подсистемы хранения

В следующем тесте мы имитировали отказ (основного) хранилища ESS. Тестирование отказа хранилища производилось, когда все три тестовых приложения были активны и при интенсивной нагрузке на диски. Показатель использования дисков составлял около 100 % для всех тестовых файловых систем. Мы имитировали два различных типа отказов хранилища.

При первой имитации мы убрали назначения всех LUN, которыми владели наши группы томов с поддержкой межсайтового зеркального отображения. Это имитировало внутренний логический отказ хранилища. При имитации второго отказа мы отключили кабель Fibre Channel на узле **nantes** для имитации отказа подключения.

В обоих случаях приложения продолжали работать без прерывания и группы томов и файловые системы оставались доступными. После отказа мы проверили доступность диска и состояние синхронизации копии логического тома. Диски из хранилища ESS выводились с пометкой **missing** и с состоянием логического тома **stale**. Пример 16.7 содержит выходные данные для группы томов **clvmmenhvg** при недоступной дисковой подсистемы ESS.

Пример 16.7. Состояние дисков и логических томов после одного отказа подсистемы хранения

```
nantes > lsvg -p clvmmenhvg
clvmmenhvg:
PV_NAME          PV STATE         TOTAL PPs   FREE PPs    FREE DISTRIBUTION
hdisk11          active           79          78          16..15..15..16..16
hdisk10          active           79          59          16..00..11..16..16
hdisk2           missing          74          73          15..14..14..15..15
hdisk5           missing          74          54          15..00..09..15..15
nantes > lsvg -l clvmmenhvg
clvmmenhvg:
LV NAME            TYPE       LPs   PPs   PVs  LV STATE      MOUNT POINT
cenglog2lv         jfs2log    1     2     2    open/stale    N/A
enhtest1lv         jfs2       20    40    2    open/stale    /app1fs
```

Приложения продолжали работать с оставшейся копией логического тома. После этого теста мы восстановили доступ к хранилищу ESS. Мы использовали опцию CSPOC **Synchronize Shared LVM Mirrors (Синхронизация общих зеркальных отображений LVM)**, которая автоматически сделала доступными все устройства hdisk и выполнила синхронизацию всех логических томов. Мы выполнили **smit cl_admin -> HACMP Logical Volume Management (Управление логическими томами HACMP) -> Synchronize Shared LVM Mirrors (Синхронизация общих зеркальных отображений LVM) -> Synchronize by Volume Group (Синхронизация по группам томов)**, после чего выбрали требуемую группу томов.

Мы проверили доступность дисков и состояние синхронизации копии логического тома. Все диски во всех группах томов являются доступными, и все логические тома находятся в состоянии **synch**[1]. Пример 16.8 содержит данные для группы томов **clvmmenhvg** после реинтеграции хранилища.

[1] Имеется в виду состояние syncd (см. пример 16.8). *Примеч. науч. ред.*

Пример 16.8. Состояние дисков и логических томов после реинтеграции хранилища

```
nantes > lsvg -p clvmmenhvg
PV_NAME          PV STATE         TOTAL PPs   FREE PPs    FREE DISTRIBUTION
hdisk11          active           79          78          16..15..15..16..16
hdisk10          active           79          59          16..00..11..16..16
hdisk2           active           74          73          15..14..14..15..15
hdisk5           active           74          54          15..00..09..15..15
nantes > lsvg -l clvmmenhvg
clvmmenhvg:
LV NAME            TYPE       LPs   PPs   PVs  LV STATE      MOUNT POINT
cenglog2lv         jfs2log    1     2     2    open/syncd    N/A
enhtest1lv         jfs2       20    40    2    open/syncd    /app1fs
```

Отказ всех дисковых подключений на одном сайте

В следующем тесте выполняется разрыв подключений Fiber Channel на сайте **France** и узле **nantes**. После некоторой задержки (в пару минут) кластер обнаружил отказ хранилища для всех файловых систем в общей группе томов и выполнил перемещение групп ресурсов RG1 и RG2 на другой сайт **Slovenia**. Кластер активизирует обе группы ресурсов на узле **koper**, выполняет **varyonvg** для всех групп томов, подключает файловые системы и запускает приложения. Все ресурсы доступны на узле **koper** и группы томов из RG1 и RG2 активизированы.

Отказ сайта

В следующем тесте мы имитировали отказ сайта **France** посредством одновременного отказа узла **nantes** и подключений Fiber Channel к дисковой подсистеме ESS. Кластер обнаруживает отказ сайта, после чего происходит перемещение групп ресурсов RG1 и RG2 на сайт **Slovenia**. Кластер активизирует обе группы ресурсов на узле **koper**, выполняет команду **varyonvg** для всех групп томов, подключает файловые системы и запускает приложения. Все ресурсы доступны на узле **koper**, группы томов, относящиеся к RG1 и RG2, активизируются в оставшейся дисковой подсистеме (хранилище DS4500). Третья группа ресурсов RG3 (на узле **koper**) работает без перерыва.

После выполнения данного теста мы подключили дисковую подсистему ESS обратно, после чего дисковые ресурсы ESS опять стали доступными, и активизировали узел **nantes**. Кластер инициирует перемещение RG1 и RG2 обратно на основной узел (**nantes**) в соответствии с политикой группы ресурсов. Все ресурсы в RG1 и RG2 стали доступными на узле **nantes**, после чего были запущены приложения.

17

Сценарий аварийного восстановления в HAGEO

Эта глава содержит сценарий аварийного восстановления на основе HACMP/XD HAGEO. Мы представляем подробное описание конфигурирования среды HAGEO. Данная глава охватывает следующие темы:

- описание сценария;
- планирование конфигурации HAGEO;
- установка и конфигурирование HACMP/XD HAGEO;
- аспекты перемещения при сбое сайта и возврата после восстановления сайта

17.1 Описание сценария и планирование

В нашем сценарии используется три узла на двух сайтах: Boston и Munchen. На рис. 17.1 подробно изображено расположение узлов и путь для связи между двумя сайтами. Внешний клиент способен осуществлять доступ к открытой сети на каждом сайте.

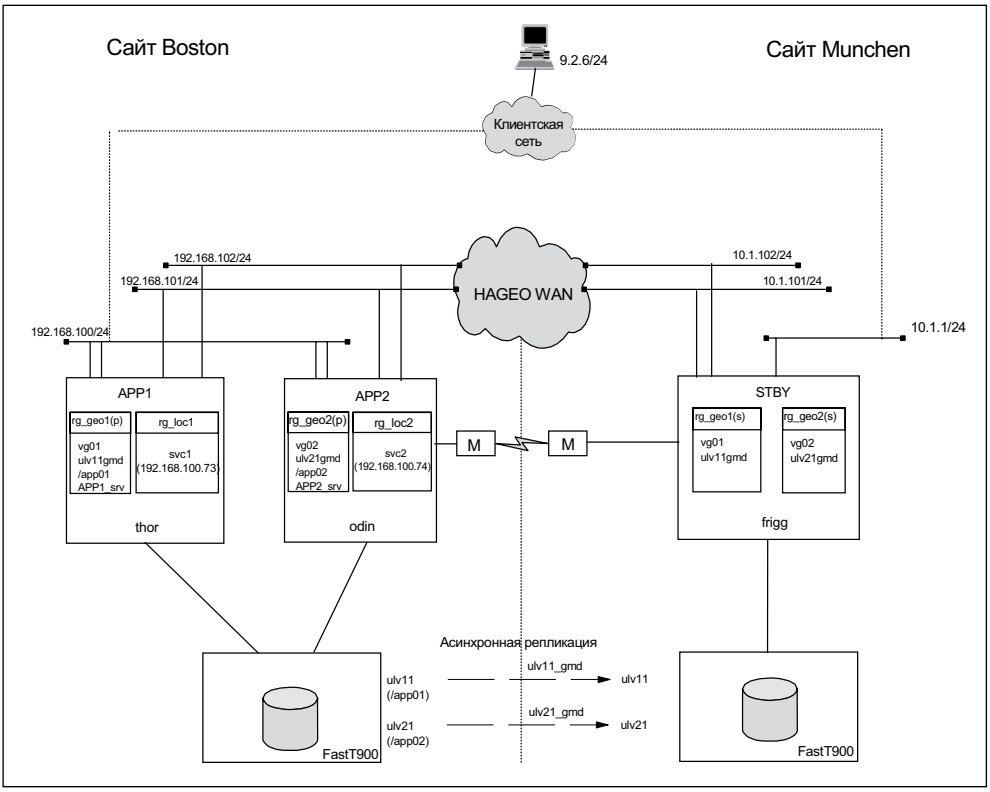

Рис. 17.1. Сценарий HAGEO

Мы используем узлы thor и odin в конфигурации со взаимным перехватом на сайте Boston и узел frigg в качестве дежурного узла на сайте Munchen. Каждый узел использует два интерфейса Ethernet для репликации данных между сайтами.

Две географические сети обеспечивают высокую доступность и балансировку нагрузки для механизма репликации. Последовательное подключение RS232 используется в качестве сетей пульса, отличных от IP, между сайтами. Имеется две группы ресурсов, построенные вокруг приложений APP01 и APP02, выполняющихся поверх устройств GeoMirror.

17.1.1 Планирование конфигурации сети

Мы определили следующие сети HACMP вместе с их подсетями:

- Открытая сеть на сайте Boston для доступа клиентов к приложениям APP01 и APP02: 172.1.1.0/24. Мы реализовали одну подсеть для локальной сети сайта Boston, так как у нас используется мониторинг пульса через синонимы для этой сети.

Наша подсеть пульса – 172.16.100.0/24.

На сайте Munchen нет сети HACMP для клиентского доступа, так как к этой сети подключен один узел с одним интерфейсом Ethernet. Подсеть, используемая для этой сети, – 10.1.1.0/24.

- Географические сети репликации:
  - geo1: 192.168.101.0/24 (сайт Boston) и 10.1.101.0/24 (сайт Munchen);
  - geo2: 192.168.102.0/24 (сайт Boston) и 10.1.102.0/24 (сайт Munchen).
- Сеть пульса через диски на сайте Boston, используемая в качестве сети пульса, отличной от IP, для связи между узлами odin и thor.
- Вторая географическая сеть пульса последовательного типа RS232 между узлами odin (Boston) и frigg (Munchen).

Табл. 17.1 отображает конфигурацию IP-адресов узлов в кластере.

Таблица 17.1. Конфигурация интерфейсов узлов

Имя хоста	Имя интерфейса	IP-адрес/маска сети	Интерфейс AIX	Назначение
thor	thor_boot1	172.1.1.73/24	en0	Загрузочный
	thor_boot2	172.1.1.75/24	en1	Загрузочный
	thor	192.168.100.73/24	Неприменимо	Постоянный
	thor_svc	192.168.100.75/24	Неприменимо	Сервисный
	thor_geo1	192.168.101.73/24	en2	Geo_primary
	thor_geo2	192.168.102.73/24	en3	Geo_primary
odin	odin_boot1	172.1.1.74/24	en0	Загрузочный
	odin_boot2	172.1.1.77/24	en1	Загрузочный
	odin	192.168.100.74/24	Неприменимо	Постоянный
	odin_svc	192.168.100.77/24	Неприменимо	Сервисный
	odin_geo1	192.168.101.74/24	en2	Geo_primary
	odin_geo2	192.168.102.74/24	en3	Geo_primary
frigg	frigg_geo1	10.1.101.192/24	en0	Geo_primary
	frigg_geo2	10.1.102.192/24	en2	Geo_primary
	frigg	10.1.1.192/24	en1	Загрузочный

Мы применяем одну подсеть для локальной сети сайта Boston, так как у нас используется мониторинг пульса посредством синонимов.

17.1.2 Планирование конфигурации логических томов

Определение устройства Geo-Mirror требует создания логических томов с одинаковыми именами на обоих сайтах. В реплицируемой группе ресурсов HACMP используется имя группы томов, содержащее устройства GeoMirror; оно также должно быть одинаковым на обоих сайтах.

Помимо логического тома, которому оно сопоставляется, каждое устройство GeoMirror использует логический том для карты состояния (state map logical volume) для регистрации несинхронизированных данных локальных и удаленных хостов. При создании state map logical volume необходимо учитывать размер логического тома, с которым он связан, используя следующую формулу:

размер statemap = макс. размер LV/(размер региона x 2).

Максимальный размер логического тома представляет приблизительную оценку максимального объема логического тома. Размер региона представляет размер блока данных на логическом томе, отображенного 4-битовой структурой данных на statemap logical volume. По умолчанию размер региона составляет 32768 байт (32 Кб).

Например, мы используем ulv11 с максимальным размером 10 Гб (160 физических разделов по 64 Мб). Действительный размер statemap составляет

10*1024*1024 Кб/(32 Кб x 2) = 160 Мб.

Размер statemap logical volume следует округлить в большую сторону до получения числа, кратного размеру физического раздела, так что в действительности для логического тома выделяется 192 Мб (три физических раздела). Табл. 17.2 представляет конфигурацию логических томов, используемых нами в конфигурации HAGEO.

Таблица 17.2. Логические тома на основном и дополнительном сайтах

Логический том	Группа томов	Размер (PP = 128 Мб)	Сайты
ulv11	vg01	160	Boston, Munchen
ulv11_sm	vg01	3	Boston, Munchen
ulv11_log	vg01	1	Boston, Munchen
ulv11_log_sm	vg01	1	Boston, Munchen
ulv21	vg02	160	Boston, Munchen
ulv21_sm	vg02	3	Boston, Munchen
ulv21_log	vg02	1	Boston, Munchen
ulv21_log_sm	vg02	1	Boston, Munchen

Для каждого логического тома определен statemap logical volume. Имя и размер логического тома должны быть одинаковы на обоих сайтах.

> **Примечание.** В целях повышения производительности мы рекомендуем размещать statemaps и логические тома данных на разных физических томах.

17.1.3 Определение GMD

В своей конфигурации мы определили четыре GMD, соответствующие двум файловым системам. APP01 и APP02 представляют собой два обычных приложения, использующих данные в файловых системах /app01 и /app02. Каждый логический том имеет соответствующий statemap. Табл. 17.3 представляет конфигурацию устройств GeoMirror.

Таблица 17.3. Определение GMD

Имя GMD	Младший номер	Том statemap	Логический том	Режим устройства	Файловая система
ulv11_gmd	10	ulv11_sm	ulv11	async	/app01
ulv11_log_gmd	11	ulv11_log_sm	ulv11_log	async	Неприменимо
ulv21_gmd	20	ulv21_sm	ulv21	async	/app02
ulv21_log_gmd	21	ulv21_log_sm	ulv21_log	async	Неприменимо

17.2 Установка и конфигурирование HAGEO

Мы использовали в своем сценарии следующие программные компоненты:

- AIX 5.3 ML02, RSCT 2.4.2;
- HACMP 5.3;
- HAGEO 5.3.

Установка программного обеспечения HAGEO требует наличия установленных наборов файлов HACMP.

Были установлены следующие пакеты HACMP/XD HAGEO:

- cluster.xd.license,
- hageo.doc.en_US,
- hageo.gmdsizing,
- hageo.man.en_US,
- hageo.manage,
- hageo.message,
- hageo.mirror.

Пример 17.1 содержит список наборов файлов hageo, использовавшихся в нашей конфигурации.

Пример 17.1. Список установленных наборов файлов hageo

```
  Fileset                      Level  State  Type  Description (Uninstaller)
  ----------------------------------------------------------------------------
hageo.doc.en_US.data       5.3.0.0    C     F    HAGEO Product Manuals - U.S.
                                                    English
hageo.gmdsizing            5.3.0.0    C     F    GMD Sizing Demonstration Tool
hageo.man.en_US.message.data
                             5.3.0.0    C     F    HAGEO GeoMessage Man Pages -
                                                    U.S. English
hageo.man.en_US.mirror.data
                             5.3.0.0    C     F    HAGEO GeoMirror Man Pages -
                                                    U.S. English
hageo.manage.utils         5.3.0.0    C     F    HAGEO GeoManage Utilities
hageo.message.ext          5.3.0.0    C     F    HAGEO GeoMessage Device Driver
hageo.message.utils        5.3.0.0    C     F    HAGEO GeoMessage Utilities
hageo.mirror.ext           5.3.0.0    C     F    HAGEO GeoMirror Device Driver
hageo.mirror.utils         5.3.0.0    C     F    HAGEO GeoMirror Utilities
```

Дополнительные сведения об установке наборов файлов HAGEO см. в руководстве **High Availability Cluster Multi-Processing XD (Extended Distance) for HAGEO Technology: Planning and Administration**, SA22-7956.

Конфигурирование IP-адресов адаптеров

Мы настроили загрузочные IP-адреса на узлах в соответствии с табл. 17.1. Помимо подсетей, представленных в этой таблице, мы используем выделенную подсеть для мониторинга пульса посредством синонимов: 172.16.100.1/24.

Сервисный адрес и адрес мониторинга пульса представляют синонимы, работающие через загрузочные IP-метки, активизируемые службами кластера. Постоянный IP-адрес представляет синоним, привязанный к узлу, который остается активным на этом узле даже после остановки служб кластера или перезагрузки системы.

Определение логических томов

На обоих сайтах мы создает похожие конфигурации логических томов. Предполагая, что на основном сайте логические тома и файловые системы уже определены, мы определяем логические тома на резервном сайте Munchen. В примере 17.2 мы создаем реплицируемые логические тома на узле frigg сайта Munchen. Обратите внимание на то, что используемый нами тип логического тома **statemap** представляет собой описательный атрибут логического тома без какой-либо функциональной роли.

Пример 17.2. Создание томов и файловых систем на удаленном сайте Munchen

```
frigg:/# mkvg -y vg01 -f -c hdisk1
frigg:/# varyonvg vg01
frigg:/# mklv -y ulv11_log -t jfs2log vg01 1
frigg:/# logform /dev/ulv11_log
logform: destroy /dev/rulv11_log (y)?y

frigg:/# mklv -y ulv11 -t jfs2 vg01 160

frigg:/# mkvg -y vg02 -f -c hdisk2
frigg:/# varyonvg vg02
frigg:/# mklv -y ulv21_log -t jfs2log vg02 1
frigg:/# logform /dev/ulv21_log
logform: destroy /dev/rulv21_log (y)?y

frigg:/# mklv -y ulv21 -t jfs2 vg02 160

frigg:/# mklv -y ulv11_sm -t statemap vg01 3
frigg:/# mklv -y ulv21_sm -t statemap vg02 3
frigg:/# mklv -y ulv11_log_sm -t statemap vg01 1
frigg:/# mklv -y ul211_log_sm -t statemap vg02 1
```

Примечание. Поддерживайте согласованность определений логических томов и групп томов на обоих сайтах в целях интеграции в реплицируемые группы ресурсов в HACMP. Например, при создании группы томов с расширенным одновременным доступом vg01 на сайте Boston необходимо создать группу томов с расширенным одновременным доступом с таким же именем на сайте Munchen.

Определение топологии HACMP

Прежде чем приступить к конфигурированию устройств зеркального отображения, необходимо сконфигурировать топологию кластера. В нашем сценарии мы описываем действия, выполняемые для определения топологии кластера.

1. **Определение имени кластера.**

Cluster name = itso

2. **Конфигурирование узлов вместе с путями для связи.**

Node names: thor, odin, frigg

Communication path: thor_geo1, odin_geo1, frigg_geo1

3. **Выполнение процесса обнаружения для получения информации об IP-адресах и дисках со всех узлов в кластере.**

4. **Конфигурирование сайтов HACMP.**

Мы выполнили конфигурирование сайтов Boston и Munchen, как показано в примерах 17.3 и 17.4.

Пример 17.3. Определение сайта Boston

```
Site Name                                    [Boston]
* Site Nodes                                  odin thor
* Dominance                                  [Yes]
* Backup Communications                      [sgn]
```

Поле Dominance (Доминирование) определяет, какой сайт будет остановлен в случае изоляции сайтов. Изоляция сайтов происходит при отказе всех географических сетей, когда по меньшей мере один узел на каждом сайте продолжает работать. В целях предотвращения несогласованности данных выполняется остановка недоминирующего узла.

Поле Backup Communications (Резервные связи) указывает альтернативный способ связи между сайтами. Возможные значения:

- sgn (Secondary Geographical Network);
- dbfs (Dial Back Fail Safe);
- none.

Настоятельно рекомендуется определить сеть резервной связи в кластере, так как это помогает избежать изоляции сайта в случае отказа основных географических сетей.

Пример 17.4. Определение сайта Munchen

```
* Site Name                                  [Munchen]
* Site Nodes                                  frigg
* Dominance                                  [No]
* Backup Communications                      [sgn]
```

```
There are 3 node(s) and 5 network(s) defined
NODE frigg:
        Network boston_diskhb_01
        Network boston_ether_01
        Network net_Geo_Primary_01
                frigg_geo1      10.1.101.192
        Network net_Geo_Primary_02
                frigg_geo2      10.1.102.192
        Network net_Geo_Secondary_01
                frigg_tty0      /dev/tty0
NODE odin:
        Network boston_diskhb_01
                odin_hdisk2     /dev/hdisk2
        Network boston_ether_01
                odin_boot2      172.1.1.77
                odin_boot1      172.1.1.74
        Network net_Geo_Primary_01
                odin_geo1       192.168.101.74
        Network net_Geo_Primary_02
                odin_geo2       192.168.102.74
        Network net_Geo_Secondary_01
                odin_tty0       /dev/tty0

NODE thor:
        Network boston_diskhb_01
                thor_hdisk2     /dev/hdisk2
        Network boston_ether_01
                thor_boot1      172.1.1.73
                thor_boot2      172.1.1.75
        Network net_Geo_Primary_01
                thor_geo1       192.168.101.73
        Network net_Geo_Primary_02
                thor_geo2       192.168.102.73
        Network net_Geo_Secondary_01
No resource groups defined
```

5. **Конфигурирование сетей HACMP.** В кластере HACMP выполняется определение следующих сетей:

- Клиентская локальная сеть для сайта Boston: boston_ether_01.

  Пример 17.5 содержит подробное описание сети boston_ether_01 в HACMP. Обратите внимание на то, что на этом этапе указывается сеть пульса через синонимы.

- Локальная сеть пульса через диски для сайта Boston: boston_diskhb_01.

  Сети пульса через диски используются в качестве сети пульса, отличной от IP, между узлами thor и odin на сайте Boston. В примере 17.6 представлено определение мониторинга пульса через диски.

Пример 17.5. Определение открытой сети сайта Boston

```
Add a GeoMirror Device

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
  Device Name                                         [ulv11gmd]
* Minor Device Number                                        [10]
* State Map Logical Volume                     [/dev/rulv11_sm]
* State Map Size (Number of Entries)                       [1024]
* State Map Region Size                                   [32768]
* Local Logical Volume                           [/dev/rulv11]
* Device Mode                                             async
* Device Role                                           primary
  High Water Mark                                            []
  Sync Concurrency Rate                                      []
* Remote Node, LV, and Statemap          [frigg@/dev/rulv11@/dev/rulv11_sm]
  Remote Node, LV, and Statemap                              []
  Remote Node, LV, and Statemap                              []
  Remote Node, LV, and Statemap                              []
  Remote Node, LV, and Statemap                              []
  Remote Node, LV, and Statemap                              []
  Remote Node, LV, and Statemap                              []
  Local Peer and State Map Device                     [thor@/dev/rulv11_sm]
  Local Peer and State Map Device                            []
  Local Peer and State Map Device                            []
  Local Peer and State Map Device                            []
  Local Peer and State Map Device                            []
  Local Peer and State Map Device                            []
```

Пример 17.6. Определение сети пульса через диски

```
GMD(s) for HACMP to start in parallel                [1]
Network Protocol                                     [TCP]
  Temporal Ordering Policy                           [SYSTEM]
  Autoset Network Parameters                         [Yes]
  TCP Send/Receive Space Size (KBytes)               [512]
```

- Основные географические сети репликации: net_Geo_Primary_01 и net_Geo_Primary_02.

HACMP использует сеть типа Geo_Primary для репликации данных с использованием HAGEO. Мы определяем сети Geo_Primary в примере 17.7.

По умолчанию сеть Geo_Primary создается с атрибутом открытой сети (public). Существует два варианта конфигурирования этой сети:

– Использовать открытую (public) сеть. На момент добавления коммуникационных интерфейсов в сети нет определенной сервисной IP-метки. Вам необхо-

димо определить сервисные IP-адреса, привязанные к узлу, через меню определения группы ресурсов.

– Использовать закрытую (private) сеть. В этом случае адреса, определяемые в топологии путем добавления коммуникационных интерфейсов в сеть, являются сервисными IP-метками.

Пример 17.7. Определение сетей Geo_Primary

```
* Network Name                                        [net_Geo_Primary_02]
* Network Type                                         Geo_Primary
* Netmask                                             [255.255.255.0]
* Enable IP Address Takeover via IP Aliases            No
```

В нашем сценарии мы используем закрытые сети HACMP, поэтому мы изменяем первоначальное определение сети с открытой на закрытую, как показано в примере 17.8.

Пример 17.8. Изменение атрибута типа сети

```
 Change/Show All Resources and Attributes for a Resource Group

 Type or select values in entry fields.
 Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
   Resource Group Name                                 thor_svc_rg
   Inter-site Management Policy                        ignore
   Participating Nodes from Primary Site               thor odin
   Participating Nodes from Secondary Site

   Startup Policy                        Online On Home Node Only
   Fallover Policy                       Fallover To Next Priority Node
   Fallback Policy                       Fallback To Higher Priority Node
   Fallback Timer Policy (empty is immediate)               []
   Service IP Labels/Addresses                      [thor_svc]
   Application Servers                                      []
   Volume Groups                                            []
   Use forced varyon of volume groups, if necessary    false
   Automatically Import Volume Groups                  false
   Filesystems (empty is ALL for VGs specified)             []
   Filesystems Consistency Check                       fsck
   Filesystems Recovery Method                         sequential
   Filesystems mounted before IP configured            false
   Filesystems/Directories to Export                        []
   Filesystems/Directories to NFS Mount                     []
   Network For NFS Mount                                    []
   Tape Resources                                           []
   Raw Disk PVIDs                                           []
   Primary Workload Manager Class                           []
   Secondary Workload Manager Class                         []
```

```
F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

> **Примечание.** Сеть Geo_Primary должна иметь сервисные IP-адреса, предназначенные для связи с устройствами GeoMirror.

- Дополнительная географическая сеть – сеть RS232 между узлами odin и frigg, соединяющая сайты. Ее определение представлено в примере 17.9.

Пример 17.9. Определение сети Geo_Secondary

```
* Network Name                                    [net_Geo_Secondary_01]
* Network Type                                     Geo_Secondary
```

6. **Добавление коммуникационных интерфейсов и устройств для определенных сетей.** На этом этапе выполняется заполнение ранее определенных сетей соответствующими интерфейсами.

- Для клиентской сети на сайте Boston.
- Для основных географических сетей мы добавляем IP-адреса, как в примере 17.10.

Пример 17.10. Определение интерфейса в сети Geo_Primary

```
* IP Label/Address                                [thor_geo1]
* Network Type                                     Geo_Primary
```

- Для дополнительной географической сети (пример 17.11).

Пример 17.11. Определение интерфейса в сети Geo_Secondary

```
* Device Name                                     [odin_tty0]
* Network Type                                     Geo_Secondary
* Network Name                                     itso_Geo_Secondary_01
* Device Path                                     [/dev/tty0]
* Node Name                                       [odin]                  +
```

- Для последовательной сети пульса через диски (пример 17.12).

Пример 17.12. Определение интерфейса последовательной сети пульса через диски

```
* Device Name                                     [odin_hdisk2]
* Network Type                                     diskhb
* Network Name                                     boston_diskhb_01
* Device Path                                     [/dev/hdisk2]
* Node Name                                       [odin]
```

7. **Конфигурирование постоянных IP-адресов (пример 17.13): odin, thor.** Мы определили интерфейсы с именами хостов как постоянные адреса в кластере. Узел frigg имеет один IP-интерфейс в открытой сети на сайте Munchen. Так как он является единственным узлом на сайте Munchen, мы не определяли интерфейс с именем хоста в HACMP.

Пример 17.13. Конфигурирование постоянных IP-адресов

```
* Node Name                                          thor
* Network Name                                      [boston_ether_01]
* Node IP Label/Address                             [thor]
```

8. **Синхронизация топологии кластера, определенной на данный момент.** Определенная нами конфигурация подробно представлена в выходных данных команды **cltopinfo** в примере 17.14.

Пример 17.14. Выходные данные команды cltopinfo

```
Cluster Name: itso
Cluster Connection Authentication Mode: Standard
Cluster Message Authentication Mode: None
Cluster Message Encryption: None
Use Persistent Labels for Communication: No
There are 3 node(s) and 5 network(s) defined
NODE frigg:
        Network boston_diskhb_01
        Network boston_ether_01
        Network net_Geo_Primary_01
                frigg_geo1      10.1.101.192
        Network net_Geo_Primary_02
                frigg_geo2      10.1.102.192
        Network net_Geo_Secondary_01
                frigg_tty0      /dev/tty0
NODE odin:
        Network boston_diskhb_01
                odin_hdisk2     /dev/hdisk2
        Network boston_ether_01
                odin_boot2      172.1.1.77
                odin_boot1      172.1.1.74
        Network net_Geo_Primary_01
                odin_geo1       192.168.101.74
        Network net_Geo_Primary_02
                odin_geo2       192.168.102.74
        Network net_Geo_Secondary_01
                odin_tty0       /dev/tty0

NODE thor:
        Network boston_diskhb_01
                thor_hdisk2     /dev/hdisk2
```

```
        Network boston_ether_01
                thor_boot1      172.1.1.73
                thor_boot2      172.1.1.75
        Network net_Geo_Primary_01
                thor_geo1       192.168.101.73
        Network net_Geo_Primary_02
                thor_geo2       192.168.102.73
        Network net_Geo_Secondary_01
No resource groups defined
```

Определение устройств GeoMirror

1. **Создание всех устройств GeoMirror.** Используя smitty hageo → Configure GeoMirror Devices (Конфигурирование устройств GeoMirror) → Configure a GeoMirror Device (Конфигурирование устройства GeoMirror) → Add a GeoMirror Device (Добавление устройства GeoMirror). В примере 17.15 мы определяем устройство ulv11gmd, соответствующее логическому тому ulv11, и определяем логический том ulv11_sm в качестве statemap device.

Пример 17.15. Определение GMD через SMIT

```
* Network Name                                        [boston_ether_01]
* Network Type                                         ether
* Netmask                                             [255.255.255.0]
* Enable IP Address Takeover via IP Aliases           [Yes]
  IP Address Offset for Heartbeating over IP Aliases [172.16.100.1]
Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
  Device Name                                         [ulv11gmd]
* Minor Device Number                                        [10]
* State Map Logical Volume                      [/dev/rulv11_sm]
* State Map Size (Number of Entries)                       [1024]
* State Map Region Size                                   [32768]
* Local Logical Volume                            [/dev/rulv11]
* Device Mode                                             async
* Device Role                                           primary
  High Water Mark                                            []
  Sync Concurrency Rate                                      []
* Remote Node, LV, and Statemap          [frigg@/dev/rulv11@/dev/rulv11_sm]
  Remote Node, LV, and Statemap                              []
  Remote Node, LV, and Statemap                              []
  Remote Node, LV, and Statemap                              []
  Remote Node, LV, and Statemap                              []
  Remote Node, LV, and Statemap                              []
  Remote Node, LV, and Statemap                              []
  Local Peer and State Map Device                     [thor@/dev/rulv11_sm]
  Local Peer and State Map Device                            []
  Local Peer and State Map Device                            []
  Local Peer and State Map Device                            []
  Local Peer and State Map Device                            []
  Local Peer and State Map Device                            []
```

2. **Синхронизация определения GMD между узлами.** Используем smitty hageo → Configure GeoMirror Devices (Конфигурирование устройств GeoMirror) → Synchronize GeoMirror Devices (Синхронизация устройств GeoMirror).

3. **Настройка свойств GeoMirror.** Используем smitty hageo → Configure GeoMirror Devices (Конфигурирование устройств GeoMirror) → Configure Global GeoMirror Properties (Настройка глобальных свойств GeoMirror). Мы используем в своем сценарии следующие параметры (пример 17.16):

Пример 17.16. Глобальные свойства GeoMirror

```
GMD(s) for HACMP to start in parallel                [1]
Network Protocol                                     [TCP]
  Temporal Ordering Policy                           [SYSTEM]
  Autoset Network Parameters                         [Yes]
  TCP Send/Receive Space Size (KBytes)               [512]
```

4. **Синхронизация свойств GeoMirror.** Используем smitty hageo → Configure GeoMirror Devices (Конфигурирование устройств GeoMirror) → Synchronize Global GeoMirror Properties (Синхронизация глобальных свойств GeoMirror).

5. **Верификация определения GMD.** Используем утилиты geo_verify или меню smitty hageo → Verify HAGEO configuration (Верификация конфигурации HAGEO).

6. **На каждом узле кластера связываем файловые системы /app01 и /app02 с устройствами GeoMirror.** Редактируем /etc/filesystems и заменяем логический том файловой системы и логический том журнала устройствами GMD, как показано в примере 17.17.

Пример 17.17. Определение файловых систем приложения в GMD

```
/app01:
        dev             = /dev/ulv11_gmd
        vfs             = jfs2
        log             = /dev/ulv11_log_gmd
        mount           = false
        check           = false
        account         = false
/app02:
        dev             = /dev/ulv21_gmd
        vfs             = jfs2
        log             = /dev/ulv21_log_gmd
        mount           = false
        check           = false
        options         = rw
        account         = false
```

7. **Тестирование созданных GMD и файловых систем.**

а) Загрузка расширения ядра Geo на узлах:

/usr/sbin/hageo/krpc/cfgkrpc -ci

б) Конфигурирование устройств GMD.

Активизируем группы томов на узлах и конфигурируем устройства GeoMirror с использованием команды cfggmd:

/usr/lib/methods/cfggmd -l <gmd_name>

в) Запуск устройств GMD.

На каждом узле основного сайта перед запуском gmd следует пометить устройство GeoMirror как отключенное на удаленном узле, используя команду **gmddown**. В нашем сценарии мы используем узел thor на основном сайте и помечаем устройство ulv11_gmd как отключенное на узле frigg.

На узле thor:

/usr/lib/methods/gmddown -l ulv11_gmd frigg

Запускаем устройства GMD на локальном узле:

/usr/lib/methods/startgmd -l ulv11_gmd

На удаленном узле помечаем соответствующий локальный узел thor как отключенный и запускаем устройства GeoMirror. В нашем примере ulv11_gmd активируется на узле thor. На узле frigg мы помечаем GMD как отключенное для узла odin и запускаем устройство:

/usr/lib/methods/gmddown -l ulv11_gmd odin

/usr/lib/methods/startgmd -l ulv11_gmd

г) Подключение файловых систем на основном узле.

д) Для освобождения устройств GeoMirror необходимо отключить файловые системы, затем остановить устройства GeoMirror на основном и дополнительном сайтах, используя на каждом сайте такую последовательность команд:

останавливаем устройство GeoMirror:

stopgmd -l <gmd_name>

отменяем конфигурирование GMD:

ucfggmd -l <gmd_name>

выгружаем расширение ядра:

/usr/sbin/hageo/krpc/cfgkrpc -u

> **Примечание.** Перед запуском устройства GeoMirror мы помечаем это устройство как отключенное на узлах, на которых оно сконфигурировано, но не запущено. Это позволяет предотвратить тайм-аут команды startgmd при связи с удаленным узлом.

Определение групп ресурсов HACMP/XD

Мы создали четыре группы ресурсов:

- Две реплицируемые группы ресурсов, содержащие группы томов, файловые системы и GMD, соответствующие приложениям APP01 и APP02, обычно работающие на сайте Boston, на узлах thor и odin соответственно. Они активизируются на обоих сайтах одновременно: основной экземпляр активируется на сайте Boston, а дополнительный экземпляр – на сайте Munchen.

- Две группы ресурсов, содержащие сервисные IP-метки (odin_svc и thor_svc), доступны только на сайте Boston.

Определение групп ресурсов

При конфигурировании групп ресурсов необходимо учитывать следующие аспекты:

- Нельзя смешивать ресурсы, зависящие от сайта, и межсайтовые ресурсы в одной группе ресурсов. В нашем сценарии сервисные IP-метки odin_svc и thor_svc доступны только на сайте Boston, поэтому их нельзя включить в одну группу ресурсов с устройствами GeoMirror.
- При определении зависимостей между двумя группами ресурсов убедитесь, что вы используете одинаковые узлы для обеих групп ресурсов. Зависимость группы ресурсов с использованием смешанных реплицируемых и нереплицируемых групп ресурсов не допускается.
- Можно учитывать порядок последовательного получения/освобождения для определения приоритета обработки групп ресурсов.

1. **Определение дополнительных ресурсов.**

- Сервисные IP-адреса для локальной сети на сайте Boston: odin_svc, thor_svc.
- Конфигурирование серверов приложений: app01_srv, app02_srv.

2. **Определение групп ресурсов.** Выполняется определение групп ресурсов, привязанных к сайту. Пример 17.18 представляет конфигурацию группы ресурсов thor_svc_rg, связанной с сервисным IP-адресом thor_svc на сайте Boston.

Пример 17.18. Добавление сервисных IP-адресов в группу ресурсов.

```
Change/Show All Resources and Attributes for a Resource Group

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
  Resource Group Name                                 thor_svc_rg
  Inter-site Management Policy                        ignore
  Participating Nodes from Primary Site               thor odin
  Participating Nodes from Secondary Site

  Startup Policy                         Online On Home Node Only
  Fallover Policy                        Fallover To Next Priority Node
  Fallback Policy                        Fallback To Higher Priority Node
  Fallback Timer Policy (empty is immediate)               []
  Service IP Labels/Addresses                       [thor_svc]
  Application Servers                                      []
  Volume Groups                                            []
  Use forced varyon of volume groups, if necessary    false
  Automatically Import Volume Groups                  false
  Filesystems (empty is ALL for VGs specified)             []
  Filesystems Consistency Check                       fsck
  Filesystems Recovery Method                         sequential
```

```
  Filesystems mounted before IP configured            false
  Filesystems/Directories to Export                       []
  Filesystems/Directories to NFS Mount                    []
  Network For NFS Mount                                   []
  Tape Resources                                          []
  Raw Disk PVIDs                                          []
  Fast Connect Services                                   []
  Communication Links                                     []
  Primary Workload Manager Class                          []
  Secondary Workload Manager Class                        []
  Miscellaneous Data                                      []
  GeoMirror Devices                                       []

F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Затем мы определяем ресурсы GeoMirror. Пример 17.19 представляет определение группы ресурсов для приложения APP01. Здесь мы добавляем дисковые ресурсы, которые включают устройства GeoMirror, определенные между сайтами.

Пример 17.19. Определение групп ресурсов GeoMirror

```
         Change/Show All Resources and Attributes for a Resource Group

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                      [Entry Fields]
  Resource Group Name                                 app01_rg
  Inter-site Management Policy                        Prefer Primary Site
  Participating Nodes from Primary Site               odin thor
  Participating Nodes from Secondary Site             frigg

  Startup Policy                                      Online On Home Node Only
  Fallover Policy                   Fallover To Next Priority Node In The List
  Fallback Policy                  Fallback To Higher Priority Node In The List
  Fallback Timer Policy (empty is immediate)         []
  Service IP Labels/Addresses                        [odin_svc]
  Application Servers                                [app01_srv]
  Volume Groups                                      [vg01]

  Use forced varyon of volume groups, if necessary    false
  Automatically Import Volume Groups                  false
  Filesystems (empty is ALL for VGs specified)       [/app01]
  Filesystems Consistency Check                       fsck
  Filesystems Recovery Method                         sequential
  Filesystems mounted before IP configured            false
  Filesystems/Directories to Export                  []
  Filesystems/Directories to NFS Mount               []
  Network For NFS Mount                              []
```

```
  Tape Resources                                    []
  Raw Disk PVIDs                                    []
  Fast Connect Services                             []
  Communication Links                               []
  Primary Workload Manager Class                    []
  Secondary Workload Manager Class                  []
  Miscellaneous Data                                []
  GeoMirror Devices                                 [ulv11_loggmd ulv11gmd]

F1=Help                 F2=Refresh                  F3=Cancel    F4=List
F5=Reset                F6=Command                  F7=Edit      F8=Image
```

3. **Синхронизация кластера HACMP.** Используем smitty hacmp → Extended Configuration (Расширенное конфигурирование) → Extended Verification and Synchronization (Расширенная верификация и синхронизация).

4. **Запуск служб кластера.** Используем **smitty clstart** для запуска служб кластера на узле. После запуска служб кластера каждый узел на основном сайте Boston получает группы ресурсов в соответствии с заданным приоритетом. Сервисные IP-адреса thor_svc и odin_svc активируются на узлах thor и odin соответственно.

Основной экземпляр групп ресурсов app01_rg и app02_rg находится на сайте Boston, где подключены файловые системы, а дополнительный экземпляр – на сайте Munchen. Устройства GeoMirror активируются на обоих сайтах при запуске служб кластера, что вызывает копирование данных, записываемых на основном сайте Boston, через сети Geo_Primary на узел frigg сайта Munchen.

Пример 17.20 отображает состояние группы ресурсов, когда службы кластера запущены на всех узлах.

Пример 17.20. Нормальное состояние группы ресурсов

```
---------------------------------------------------------------------------
Group Name     Type           State      Location
---------------------------------------------------------------------------
thor_svc_rg      non-concurrent  ONLINE     thor
                                 OFFLINE    odin
app01_rg         non-concurrent  ONLINE     thor
                                 OFFLINE    odin
                                 ONLINE SEC frigg

odin_svc_rg      non-concurrent  OFFLINE    thor
                                 ONLINE     odin

app02_rg         non-concurrent OFFLINE     thor
                               ONLINE      odin
                               ONLINE SEC  frigg
```

Локальное перемещение при сбое на сайте Boston

При отказе одного из узлов на сайте Boston второй узел получает свои ресурсы, как в случае обычного перемещения при сбое. Основной экземпляр реплицируемой группы ресурсов повторно активируется на работающем узле (рис. 17.2). Устройства

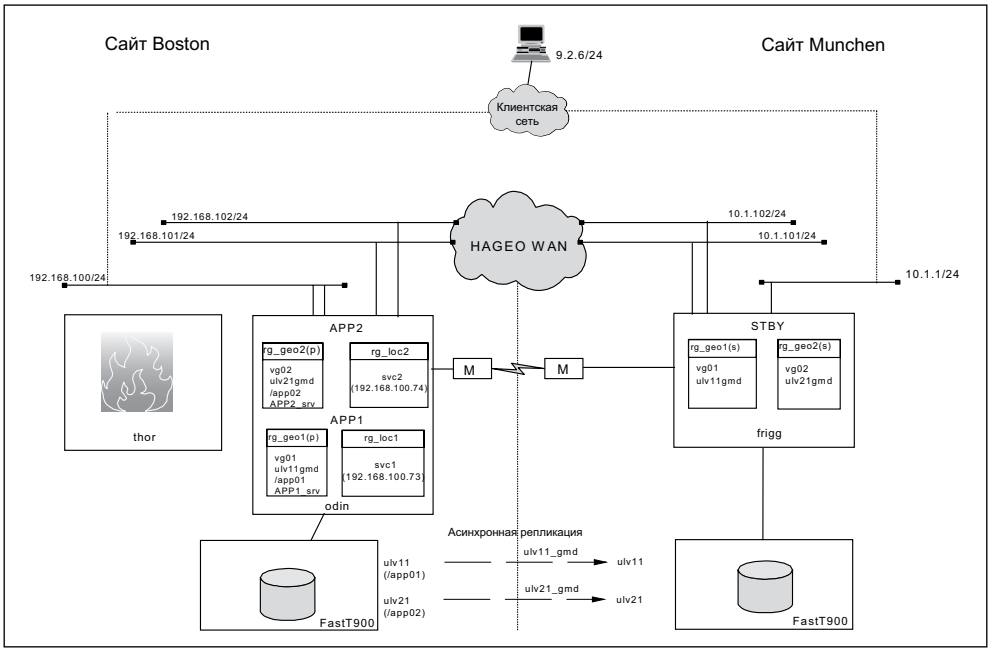

Рис. 17.2. Перемещение ресурсов при отказе узла на основном сайте

GeoMirror перезапускаются во время процесса восстановления, после чего репликация данных между сайтами возобновляется.

Мы протестировали локальное перемещение при сбое на сайте Boston, остановив службы кластера на узле thor с использованием опции постепенной остановки с передачей ресурсов. Группы ресурсов thor_svc_rg и app01_rg активируются на узле odin. Пример 17.21 показывает состояние групп ресурсов после перемещения при сбое.

Пример 17.21. Состояние группы ресурсов после перемещения при сбое

```
-----------------------------------------------------------------------------
Group Name     Type           State      Location
-----------------------------------------------------------------------------
thor_svc_rg      non-concurrent  OFFLINE    thor
                                 ONLINE     odin

app01_rg         non-concurrent  OFFLINE    thor
                                 ONLINE     odin
                                 ONLINE SEC frigg

odin_svc_rg      non-concurrent  OFFLINE    thor
                                 ONLINE     odin

app02_rg         non-concurrent OFFLINE     thor
                                ONLINE      odin
```

Перемещение при сбое сайта и возврат после восстановления сайта

В случае аварии на основном сайте оба узла с основного сайта становятся недоступными. Службы кластера на удаленном узле проверяют доступность узлов на основном сайте с использованием всех сетей Geo_Primary и Geo_Secondary. В случае использования системы DBFS (Dial Back Fail Safe) удаленный сайт вызывает основной сайт, чтобы проверить, работает ли какой-либо узел на основном сайте. Если с основного сайта нет ответа, дополнительный сайт перехватывает группы ресурсов. Рис. 17.3 показывает перемещение групп ресурсов после аварии на основном сайте.

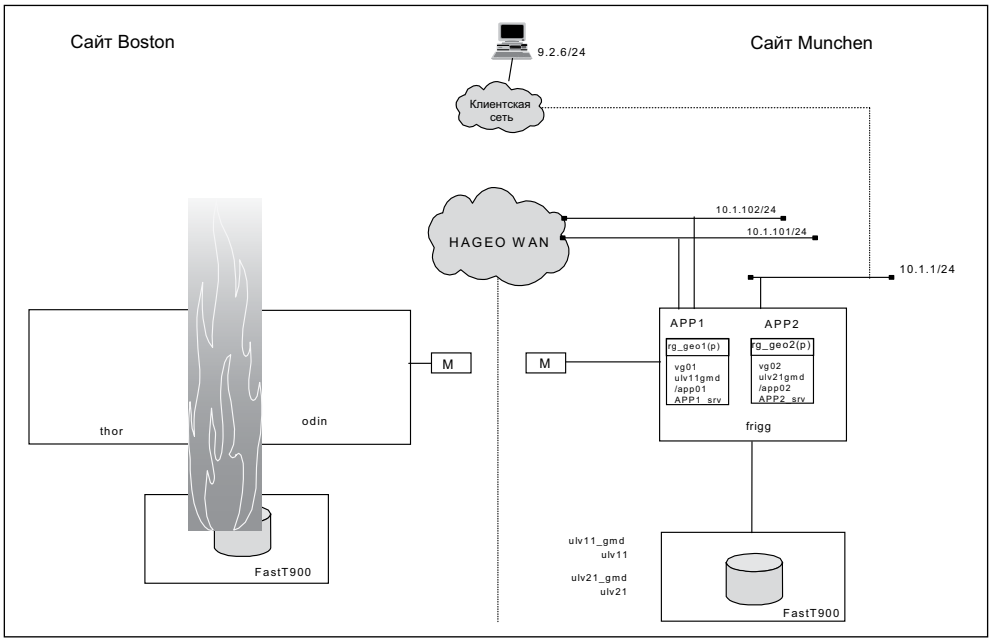

Рис. 17.3. Перемещение при сбое сайта

Обратите внимание на то, что теперь только реплицируемые группы ресурсов активны на сайте Munchen. Внешний клиент подключается к узлу frigg через локальную сеть, доступную на сайте Munchen. Так как мы используем асинхронные устройства GeoMirror, при перемещении при сбое сайта атрибуты устройств GeoMirror изменяются на атрибуты основной роли, вследствие чего устройства GeoMirror способны обрабатывать запросы записи от приложений.

Мы имитировали в своей среде перемещение при сбое сайта путем остановки служб кластера на обоих узлах, odin и thor, одновременно. После перемещения на сайт Munchen состояние реплицируемых ресурсов на узле frigg изменяется с ONLINE SECONDARY на ONLINE (пример 17.22).

При реинтеграции основного сайта в кластер сети Geo_Primary переходят в рабочее состояние, после чего запускается репликация данных в обратном направлении, с основных устройств GeoMirror на сайте Munchen на дополнительные устройства GeoMirror на сайте Boston (рис. 17.4).

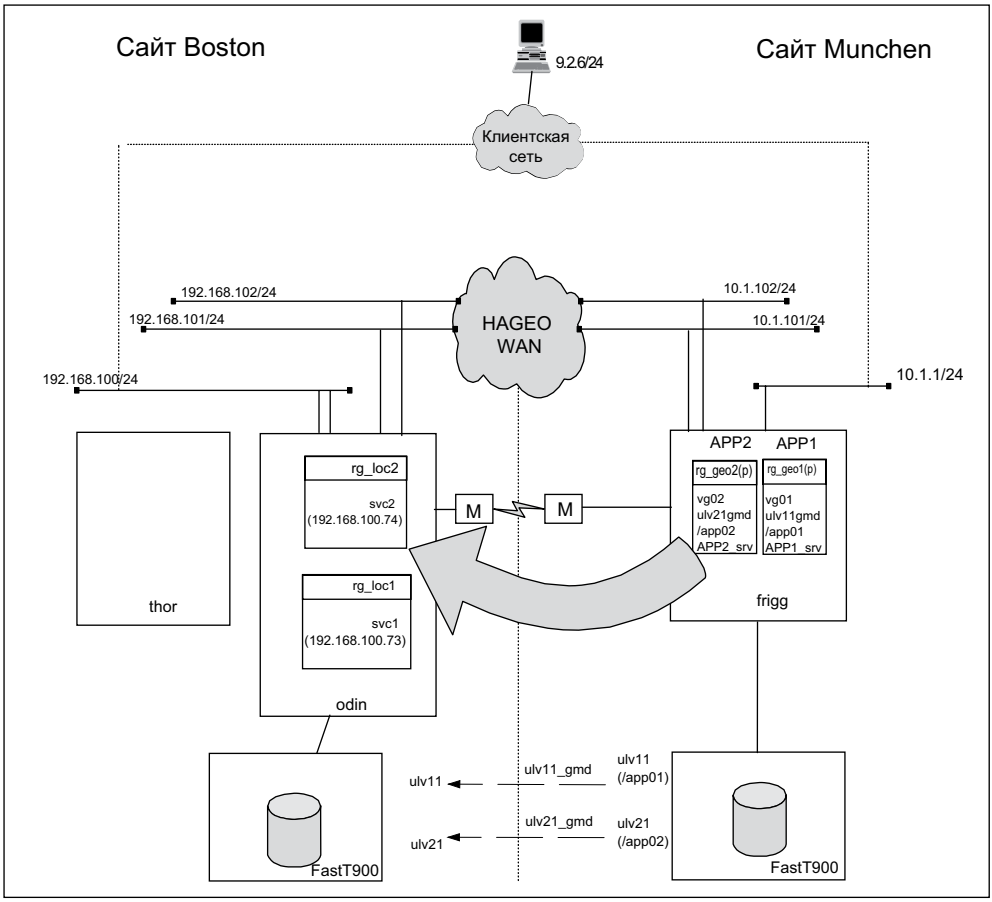

Рис. 17.4. Возврат после восстановления сайта

Пример 17.22. Состояние группы ресурсов после перемещения при сбое сайта

```
-----------------------------------------------------------------------------
Group Name     Type           State      Location
-----------------------------------------------------------------------------
thor_svc_rg      non-concurrent  OFFLINE    thor
                                 OFFLINE    odin

app01_rg         non-concurrent  OFFLINE    thor
                                 OFFLINE    odin
                                 ONLINE     frigg

odin_svc_rg      non-concurrent  OFFLINE    thor
                                 OFFLINE    odin
```

```
app02_rg         non-concurrent OFFLINE     thor
                                OFFLINE     odin
                                ONLINE      frigg
```

Мы используем в своем сценарии политику межсайтового управления Prefer Primary Site (Предпочтительное использование основного сайта). После выполнения синхронизации происходит повторное получение ресурсов на сайте Boston, а также возврат к первоначальным значениям атрибутов устройств GeoMirror: устройства GeoMirror на сайте Boston играют основную роль, а на сайте Munchen – второстепенную. Подключаются файловые системы /app01 и /app02, и оба приложения – APP01 и APP02 – запускаются на сайте Boston.

18

Понятия и конфигурирование GLVM

HACMP/XD Global Logical Volume Manager (GLVM) – новая функция высокой доступности, которая может осуществлять зеркальное отображение данных через стандартную IP-сеть и обеспечивать поддержку автоматического перемещения при сбое и возврата после восстановления для приложений с использованием данных с географическим зеркальным отображением.

GLVM осуществляет удаленное зеркальное отображение логических томов AIX с использованием функций AIX Logical Volume Manager (LVM), обеспечивая оптимальную производительность и простоту конфигурирования и обслуживания.

Эта глава рассматривает Geographic Logical Volume Manager (GLVM): общие понятия, установку, конфигурирование и миграцию.

18.1 HACMP/XD GLVM

GLVM выполняет примерно те же функции, что и HAGEO, однако использует упрощенный метод определения и выполнения репликации данных между сайтами. Он разработан с целью заменить HAGEO.

HACMP/XD GLVM обеспечивает две основные функции:

- удаленное зеркальное отображение данных;
- удаленное перемещение при сбое и возврат после восстановления.

Совместно эти функции обеспечивают поддержку высокой доступности приложений и данных через стандартную сеть TCP/IP к удаленному сайту.

HACMP/XD GLVM обеспечивает следующие возможности аварийного восстановления:

- позволяет осуществлять автоматическое обнаружение и реагирование на отказы сайтов и сетей в географическом кластере без вмешательства пользователя;
- выполняет автоматический перехват и восстановление сайта и обеспечивает высокую доступность критически важных приложений посредством перемещения при сбое и мониторинга приложений;
- позволяет упростить конфигурацию групп томов, логических томов и групп ресурсов;
- использует сеть TCP/IP для удаленного зеркального отображения на неограниченные расстояния;
- поддерживает логические тома максимального размера.

HACMP/XD:GLVM представляет HACMP Extended Distance с использованием географических логических томов для зеркального отображения данных на удаленном сайте, при этом:

- поддерживает кластеры с несколькими узлами на двух сайтах;
- осуществляет зеркальное отображение данных, обеспечивая локальное представление удаленного физического тома в LVM;
- локальная и удаленная системы хранения не обязательно должны иметь оборудование одинакового типа;
- цель состоит в том, чтобы упростить работу HACMP/XD:GeoRM и обеспечить зеркальное отображение на уровне LVM;
- в сравнении с HAGEO GLVM содержит меньше кода, поэтому он более эффективен и надежен;
- код был упрощен, поэтому GLVM более прост в поддержке и использовании;
- план разработчиков в том, чтобы GLVM в конечном итоге заменил GMD.

18.1.1 Определения и понятия

1. Удаленный физический том (remote physical volume, RPV). Псевдоустройство, обеспечивающее доступ к удаленным дискам, как если бы они имели локальное подключение. Удаленная система должна быть подключена через сеть TCP/IP, и на ней должен быть запущен HACMP. Расстояние между сайтами ограничено задержкой связующей сети.

RPV состоит из двух частей:

а) RPV-клиент. Представляет собой драйвер псевдоустройства, выполняющийся на локальном компьютере и позволяющий AIX LVM осуществлять доступ к удаленным физическим томам, как если бы они были локальными. RPV[1]-клиенты видимы в системе как устройства hdisk, которые являются логическим представлением удаленных физических томов.

Драйвер RPV-клиента представляется как обычный драйвер дискового устройства, например устройства RPV-клиента, hdisk8, и весь его ввод-вывод направляется на удаленный RPV-сервер. Он не имеет представления об узлах, сетях и т. п.

HACMP/XD 5.3 не поддерживает одновременный доступ для GLVM, так что при доступе к RPV-клиентам локальные аналоги RPV-серверов и удаленные RPV-клиенты должны быть определены.

При конфигурировании RPV-клиента определяются следующие параметры:

- Адрес RPV-сервера.
- Локальный адрес (определяет используемую сеть).
- Тайм-аут. Этот параметр используется главным образом при автономной установке GLVM, так как HACMP перезаписывает этот параметр значением config_too_long для кластера. В кластере HACMP это значение указывает на наихудший вариант развития событий, так как HACMP обнаруживает проблемы с удаленным узлом намного раньше.

В SMIT есть меню конфигурирования RPV-клиентов, smitty rpvclient.

б) RPV-сервер. Выполняется на удаленном компьютере; каждому реплицируемому физическому тому соответствует один RPV-сервер. RPV-сервер может прослушивать множество удаленных RPV-клиентов на различных хостах, обеспечивая перемещение при сбое.

RPV-сервер представляет собой экземпляр расширения ядра драйвера RPV с именем, подобным rpserver0, и не является действительным псевдоустройством.

При конфигурировании RPV-сервера определяются следующие параметры:

- PVID локального физического тома;
- адреса RPV-клиентов (с разделяющими запятыми).

2. Группа томов с географическим зеркальным отображением (geographically mirrored volume group, GMVG). Группа томов, состоящая из локального и удаленного физических томов. В GMVG применяются строгие правила, что значительно снижает вероятность отсутствия полной копии зеркального отображения на каждом сайте. По этой

[1] В тексте ошибка. Не RPC-клиенты, а RPV (Remote Physical Volume) -клиенты. *Примеч. науч. ред.*

причине требуется использование политики размещения «super strict» для каждого логического тома в GMVG. В HACMP/XD также ожидается зеркальное отображение каждого логического тома в GMVG.

В HACMP/XD осуществляется управление GMVG, где они распознаются как отдельный класс реплицируемых ресурсов и имеют собственные события. Верификация в HACMP/XD также выдает предупреждение при обнаружении групп ресурсов, содержащих ресурсы GMLV, для которых не установлен флаг принудительной активизации (forced varyon) и не отключен кворум.

В SMIT есть меню конфигурирования RPV-серверов, smitty rpvserver.

Важно! HACMP/XD требует, чтобы для каждого физического тома, входящего в группу томов с RPV-клиентами, было определена обратная связь. Как минимум каждая GMVG содержит два физических тома на каждом сайте – один локальный диск и логическое представление удаленного физического тома.

3. Утилиты GLVM. GLVM содержит меню SMIT для создания GMVG и логических томов. Они управляют расположением логических томов, чтобы обеспечить правильное размещение зеркальных копий (аналогичные меню SMIT выполняют эту же функцию неявным образом).

При использовании стандартных команд для конфигурирования GMVG рекомендуется использовать утилиту верификации GLVM.

Важно! Команды LVM и меню SMIT не имеют точного представления о схеме RPV, поэтому можно создавать группы томов GMVG, не имеющие полной копии данных на одном из сайтов.

4. В HACMP/XD GLVM добавлены новые определения:

а) XD_data Сеть, которую можно использовать только для репликации данных. Представляет аналог сети Geo_Primary. Эта сеть поддерживает переключение адаптера, но не поддерживает перемещение при сбое на другой узел. Пакеты пульса RSCT пересылаются через эту сеть.

Примечание. HACMP/XD 5.3 поддерживает только одну сеть XD_data. Также поддерживается Etherchannel.

б) XD_ip Сеть, которую можно использовать для мониторинга пульса RSCT через IP. Обычно имеет низкую пропускную способность и используется только для того, чтобы не допустить разделения кластера. Аналогична сетям Ethernet с тем исключением, что параметры пульса были настроены для работы на больших расстояниях. Не поддерживает перехват IP-адреса (IPAT) и не может использоваться для зеркального отображения данных.

в)XD_rs232 Сеть, которую можно использовать для последовательных коммуникаций. Аналогична сетям RS232 с тем исключением, что параметры пульса были настроены для работы на больших расстояниях. Подобна сетям Geo_Secondary в HACMP/XD:HAGeo. Может представлять выделенную линию или последовательную линию с драйвером линии.

Пакеты пульса RSCT пересылаются по всем сетям.

> **Примечание.** HACMP/XD:GLVM требует использования одной сети XD_data для репликации данных и одной сети XD_rs232 или XD_ip, чтобы иметь возможность отличить отказ удаленного узла от отказа сети XD_data.

Рис. 18.1 представляет пример конфигурации из двух сайтов с одним узлом на каждом сайте. Рассматривая репликацию с точки зрения узла Node1, мы видим, что целевой физический том hdisk3 узла Node2 на узле Node1 представлен как hdisk8.

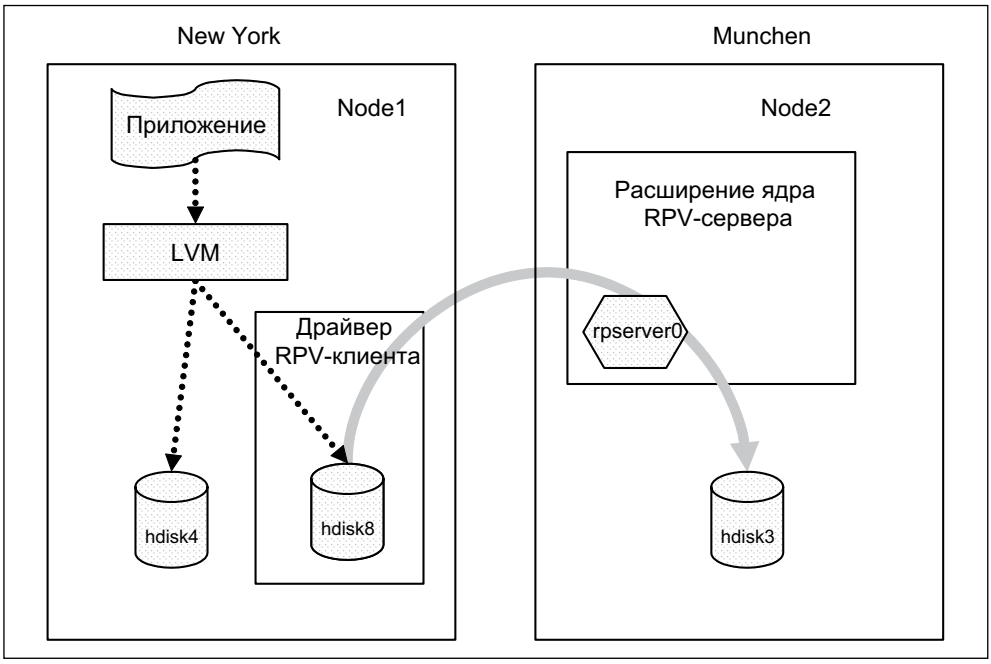

Рис. 18.1. RPV-клиент с точки зрения узла Node1

На узле Node2 имеются RPV-серверы для каждого физического тома. На узле Node1 имеются соответствующие RPV-клиенты для каждого RPV-сервера, которые представляются в LVM как локальные физические тома. Теперь на узле Node1 можно построить группу томов glvm_vg, представляющую зеркальное отображение локального физического тома (hdisk4) для локального RPV-клиента (hdisk8). RPV-клиент и RPV-сервер обеспечивают передачу всего ввода-вывода через сеть XD_data на физические тома на узле Node2.

Когда Node2 становится активным узлом, действие приобретает обратный характер, как видно на рис. 18.2. Удаленные физические тома на узле Node1 представляются через RPV-клиент как локальные физические тома на узле Node2.

> **Примечание.** Как и в случае физических томов, совместно используемых разными системами, нумерация hdisk может быть несогласованной, однако PVID в разных системах будет одинаковым.

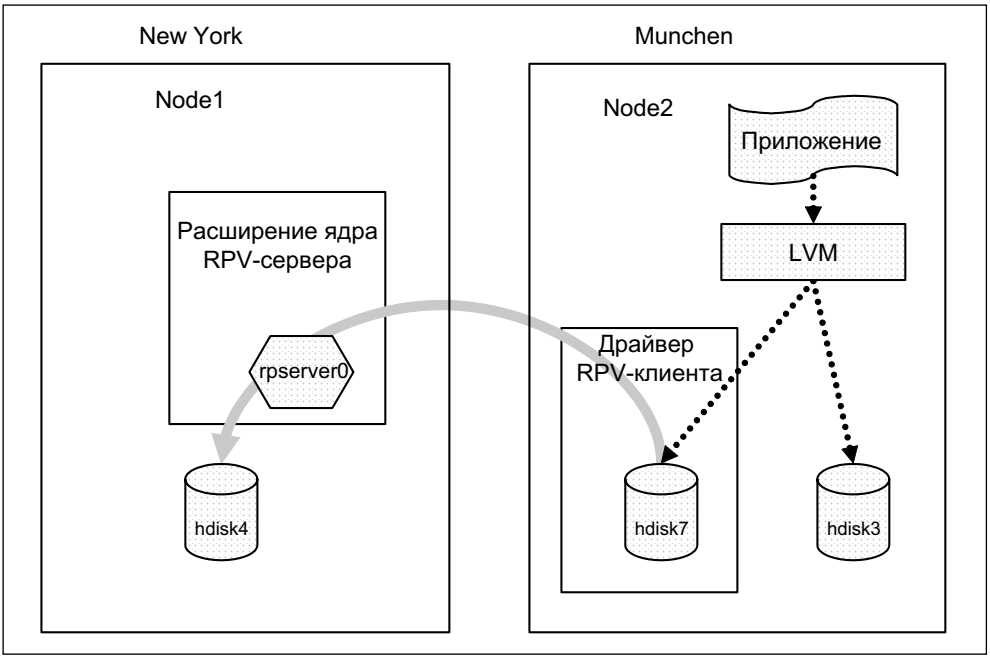

Рис. 18.2. Обратная конфигурация при перемещении на удаленный сайт

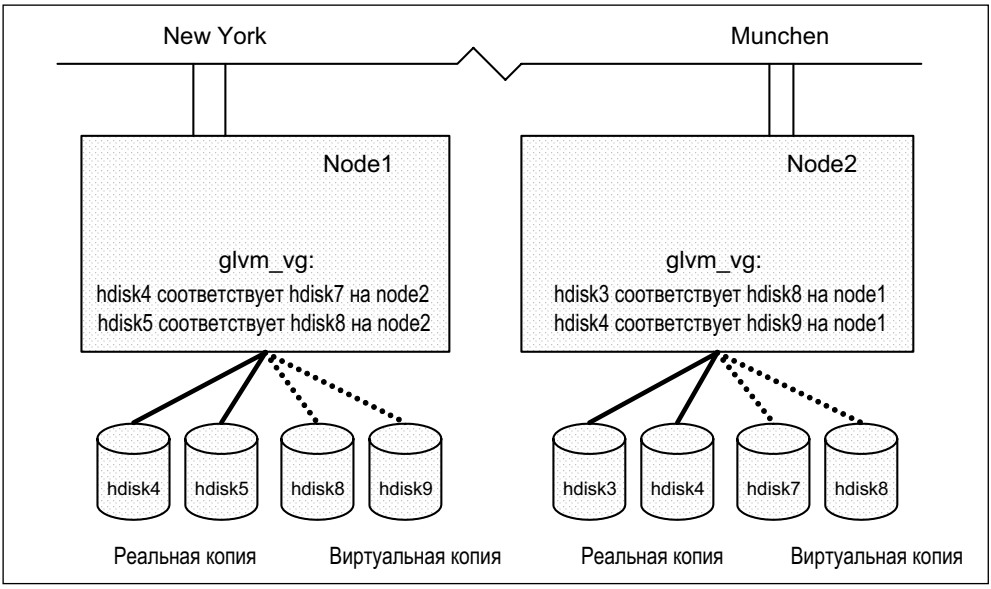

Рис. 18.3. Конфигурация RPV-клиента и сервера на обоих сайтах

Рис. 18.3 представляет конфигурацию обоих узлов с определением RPV-серверов и клиентов. Группа томов glvm_vg определена на обоих узлах и содержит два локальных физических тома и два локальных представления физических томов удаленного узла.

5. Конфигурирование GLVM под управлением HACMP. RPV-клиенты, RPV-серверы и GLVG должны быть сконфигурированы на всех узлах, прежде чем они могут быть определены в составе группы ресурсов HACMP. Каждому локальному диску должен соответствовать RPV-сервер, и каждому удаленному диску, входящему в GMVG, должен соответствовать RPV-клиент.

Табл. 18.1 содержит метки для конфигурации, представленной на рис 18.3.

Таблица 18.1. Имена RPV, используемые в нашей среде

	Локальные диски	RPV-серверы	RPV-клиенты
New York/Node1	hdisk4	rpvserver0	hdisk7 на сайте Munchen
	hdisk5	rpvserver1	hdisk8 на сайте Munchen
Munchen/Node2	hdisk3	rpvserver0	hdisk8 на сайте New York
	hdisk4	rpvserver1	hdisk9 на сайте New York

Можно получить доступ к GMVG с любого сайта, если доступны соответствующие серверы и клиенты. Другими словами, для конфигурирования GMVG на узле Node1 должны быть доступны RPV-клиенты на узле Node1 и RPV-серверы на удаленном узле.

Группы томов GMVG нельзя конфигурировать через CSPOC; их нужно создать на одном узле, затем деактивизировать и импортировать на другом узле. Повторите, пока они не станут определены на всех узлах в кластере.

После конфигурирования групп томов GMVG их нужно добавить в группу ресурсов, после чего HACMP/XD распознает их и вызовет соответствующие события для их обработки.

Общие рекомендации

Рекомендуется отключить кворум для групп томов GMVG. Если оставить кворум включенным, это позволит выполнять дальнейшую проверку, однако при этом HACMP/XD будет пытаться выполнить перемещение группы ресурсов при потере половины дисков, т. е. при отказе удаленного сайта. Часто такое поведение является нежелательным.

Рекомендуется установить флаг принудительной активизации (forced varyon) в атрибутах группы ресурсов. Это нужно для того, чтобы любой из сайтов мог подключить группу ресурсов, если недоступен удаленный сайт. При этом возможно использование устаревших данных, если в HACMP/XD не используется флаг «may be stale».

Сценарий использования устаревших данных

Если основной сайт работает, а сеть XD_data не работает, то данные на физических томах основного сайта более актуальны, чем данные на резервном сайте. Основной сайт знает, что резервный сайт работает и что данные не реплицируются, поэтому для групп томов GMVG на удаленном сайте устанавливается флаг «may be stale».

Если выполнить перемещение приложения на резервный сайт без установленного флага «may be stale», группа томов GMVG активизируется (принудительно) и при запуске приложения используются устаревшие данные. Флаг «may be stale» останавливает активизацию группы томов на данном этапе, позволяя администратору принять решение, что делать дальше.

Запуск кластера

После конфигурирования и синхронизации кластера следует запустить службы кластера HACMP на основном узле. При недоступном удаленном сервере локальные RPV-клиенты также не будут доступны, поэтому HACMP придется выполнять принудительную активизацию группы томов и всего ввода-вывода для локальных физических томов, что вызывает устаревание разделов на томах RPV-клиентов при обновлении локальных физических разделов (рис. 18.4).

Когда узел на удаленном сайте становится активным, HACMP активирует на этом узле RPV-серверы и запускает RPV-клиенты на основном сайте. HACMP/XD запускает событие, информирующее LVM о доступности RPV-клиентов, так что устаревшие разделы на RPV-клиентах обновляются путем копирования данных с локальных физических томов на удаленные физические тома (рис. 18.5).

Если имеет место перемещение приложения на резервный сайт или отключение основного сайта, возникает обратный процесс. Те физические тома на удаленном узле, которые контролировали RPV-серверы, теперь представляют собой локальные физические тома для группы томов, тогда как RPV-клиенты (указывающие на физи-

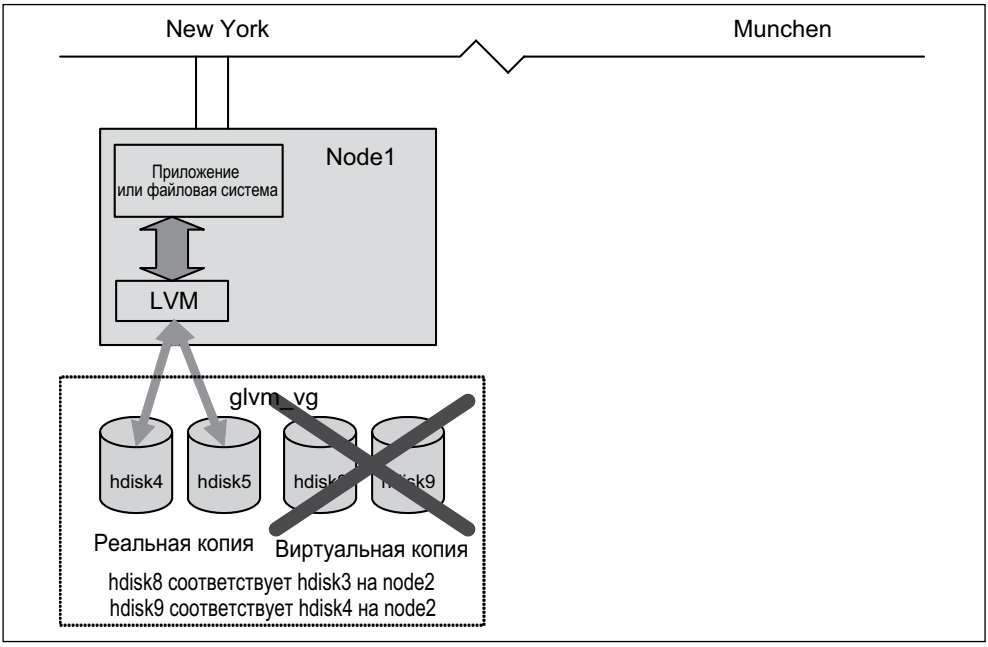

Рис. 18.4. Активный узел основного сайта при неработающем резервном сайте

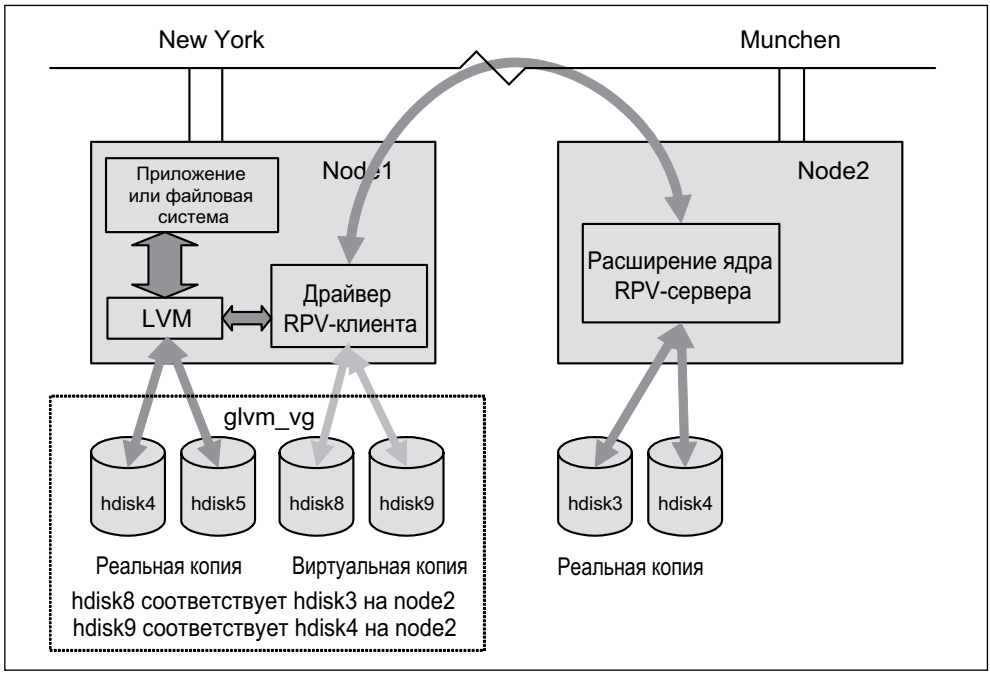

Рис. 18.5. Активный резервный сайт и репликация данных

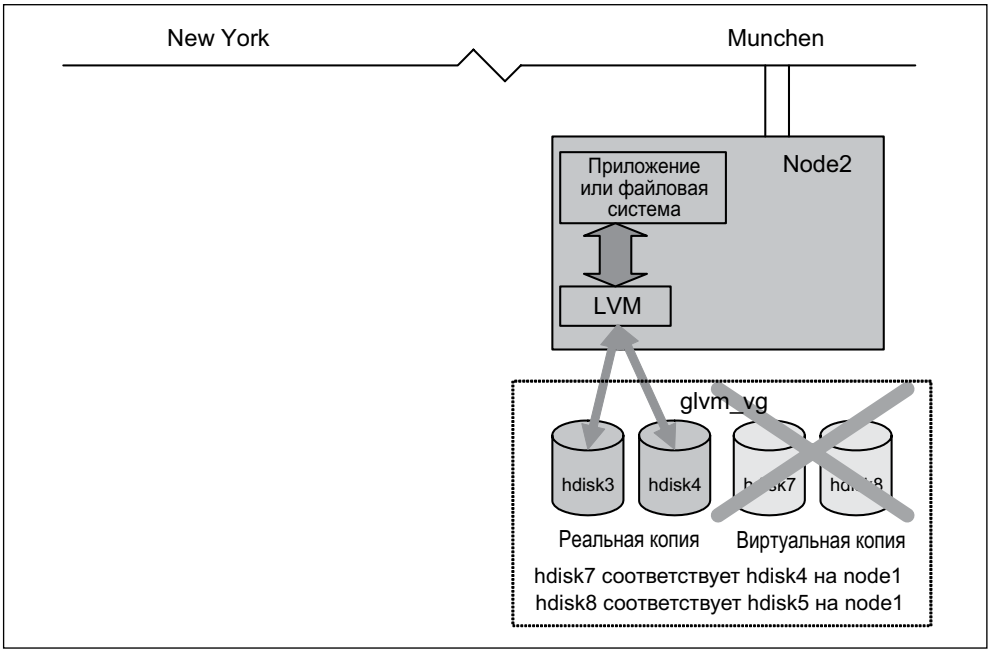

Рис. 18.6. Активный резервный узел при неработающем основном сайте

ческие тома на основном сайте) будут отключены до тех пор, пока не станет активным основной узел. Группу томов опять необходимо активизировать в принудительном режиме. См. Рисунок 18.6 на странице 670.

Когда основной сайт становится активным, HACMP/XD оставляет группу ресурсов на резервном сайте, где запускаются RPV-серверы, подключая RPV-клиентов на резервном сайте. LVM опять же обрабатывает репликацию данных на физических томах резервного сайта, для которых установлен флаг «stale», выполняя синхронизацию через RPV с физическими томами основного узла (рис. 18.7).

Однако если для группы ресурсов сконфигурирована политика Prefer Primary Site (Предпочтительное использование основного сайта), остановится работа резервного сайта и ресурсы перейдут на основной сайт. Это означает, что на основном сайте будет работать приложение, тогда как устаревшие данные будут синхронизированы обратно (с резервного сайта). Любая попытка чтения устаревшего раздела приложением приведет к чтению с RPV-сервера (рис. 18.5).

Вопрос использования политики предпочтительного использования сайта для группы ресурсов может оказывать неблагоприятное воздействие на производительность, так как обработку ввода-вывода осуществляет LVM, который не знает, что базовое устройство является удаленным физическим томом.

Например, при настройке для локальной группы томов GMVG зеркального отображения на двух физических томах и одном RPV-клиенте (рис. 18.8) увеличивается вероятность локальной доступности данных.

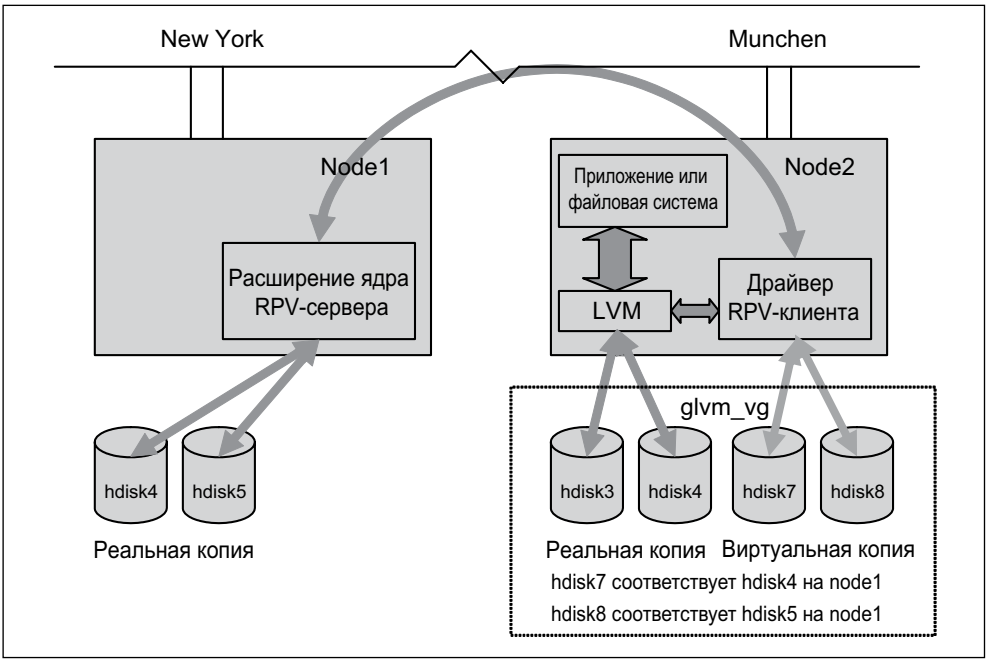

Рис. 18.7. Интеграция основного узла в кластер и запуск репликации

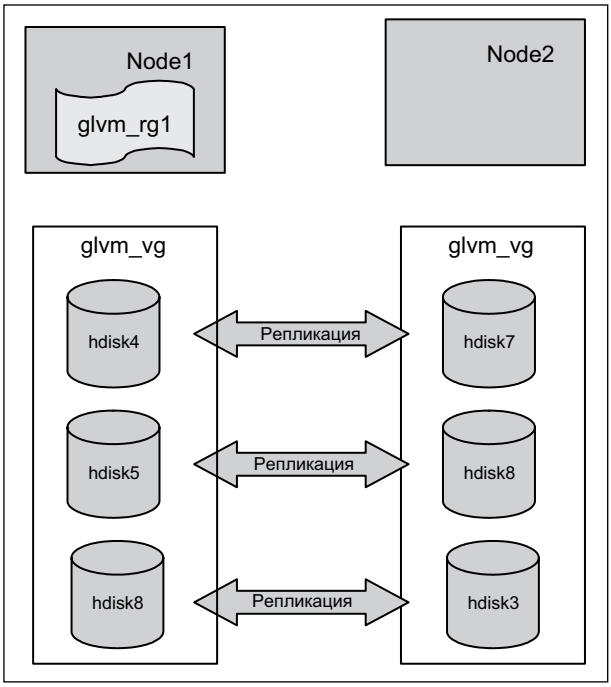

Рис. 18.8. Приложение активно на основном сайте с двумя физическими томами и с одним RPV-клиентом

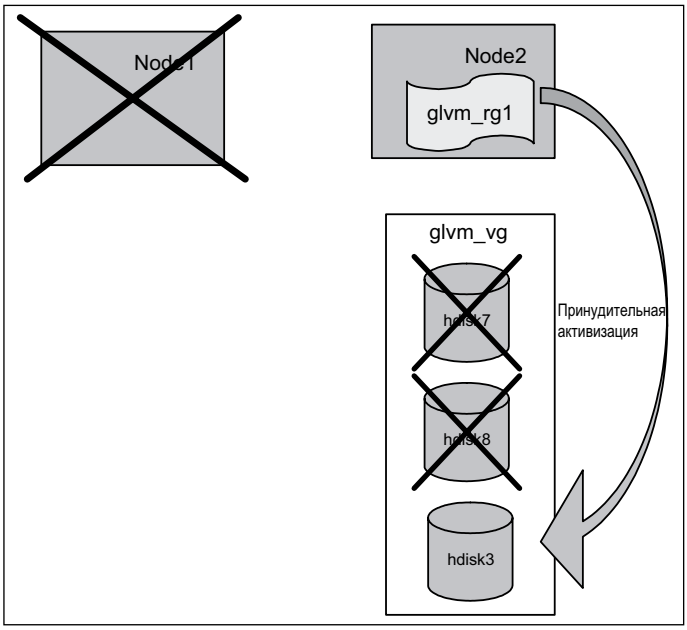

Рис. 18.9. Отключение основного сайта и приложение, использующее одну локальную копию группы томов

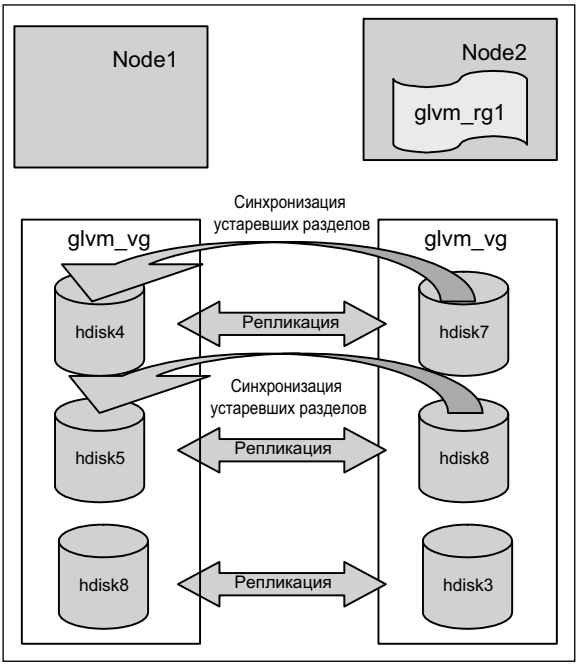

Рис. 18.10. Приложение не выполнило перемещения при интеграции основного сайта

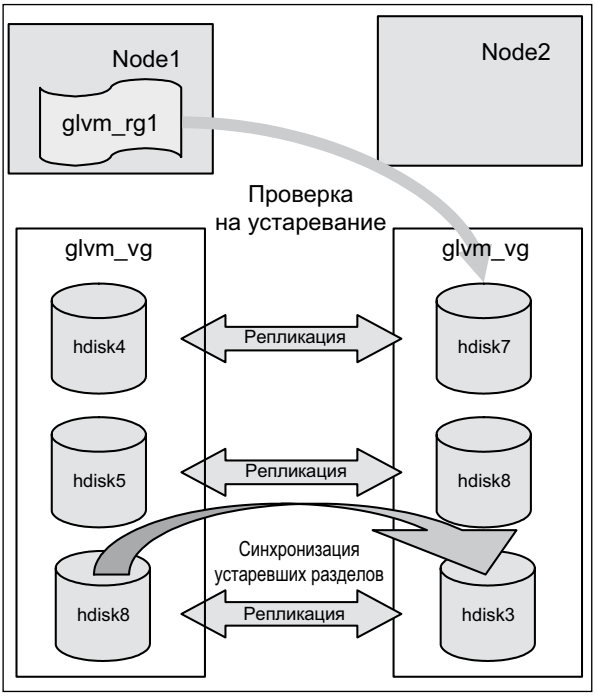

Рис. 18.11. Перемещение приложения при интеграции основного сайта

Однако в случае отключения основного сайта резервный сайт будет работать с группой томов, состоящей из одного локального физического тома и двух RPV-клиентов (рис. 18.9).

При таком сценарии политика предпочтительного использования сайта группой ресурсов оказывает большое влияние на производительность ресинхронизации данных.

Если группа ресурсов не имеет политики предпочтительного использования сайта, она будет подключена на резервном сайте и две копии устаревших данных будут пересланы через сеть XD_data – по одной для каждого RPV-сервера (рис. 18.10).

Однако при включенной политике предпочтительного использования основного сайта группой ресурсов группа томов будет активизирована на основном сайте с одним актуальным физическим томом (RPV-клиентом) и с двумя локальными физическими томами с устаревшими разделами. Таким образом, через сеть будет пересылаться только одна копия каждого раздела для перевода физических томов в синхронизированное состояние. Операции чтения с устаревших разделов выполняются дольше, так как они включают задержку сети, связанную с тем, что чтение выполняется с удаленных физических томов (рис.18.11).

Путь ввода-вывода

При нормальном вводе-выводе приложение передает запрос ввода-вывода в LVM, который передает запрос через драйвер дискового устройства и через драйвер адаптера на физический том.

При использовании GLVM приложение передает запрос ввода-вывода в LVM, который передает запрос на драйвер RPV-клиента, который пересылает запрос через TCP/IP-сеть на удаленный RPV-сервер. Удаленный RPV-сервер передает запрос через драйвер дискового устройства на удаленный узел, через драйвер адаптера на удаленный физический том. Ответ возвращается таким же образом.

Любая задержка в сети уменьшает производительность ввода-вывода, тогда как все ошибки возвращаются в LVM, как и при использовании драйвера физического тома.

В LVM RPV-клиент представляется как более медленный и менее надежный физический том – более медленный из-за более длинного пути ввода-вывода (в частности, из-за задержки сети) и менее надежный, так как в глобальных сетях с большим количеством промежуточных устройств вероятность отказов выше, чем при локальных операциях записи.

18.2 Миграция – переход от HAGeo к GLVM

Не существует автоматического механизма миграции с HACMP/XD:HAGeo на HACMP/XD: GLVM, однако HAGeo и GLVM могут совместно существовать в одном кластере. Поэтому можно выполнить пошаговую миграцию ресурсов GeoMirror с некоторым простоем.

Так как HACMP/XD:HAGeo не поддерживает динамическую реконфигурацию, весь кластер необходимо будет остановить для внесения изменений в топологию и ресурсы. Однако миграцию данных с GMD в GLVM можно выполнить при работающем приложении. Такая миграция требует повторного зеркального отображения всех данных на удаленном сайте.

Важно! Тщательное планирование необходимо, так как полная репликация каждого логического тома данных влияет на производительность основных и дополнительных узлов, а также вызывает использование значительной части пропускной способности сети. Размер логических томов и пропускная способность сети определяют продолжительность миграции, и ее выполнение, насколько это возможно, должно быть запланировано на определенный период.

В нашем примере миграции мы рассматриваем кластер с двумя узлами и двумя приложениями на основном сайте и с одним узлом на резервном сайте. Миграция состоит из следующих этапов:

- установка наборов файлов GLVM;
- остановка удаленного сайта и создание RPV-серверов на удаленном сайте;
- создание локального RPV-клиента и зеркального отображения;
- зеркальное отображение логических томов с локальными данными;
- создание локальных RPV-серверов и удаленных RPV-клиентов;
- изменение файла /etc/filesystems таким образом, чтобы он указывал на логические тома, а не на GMD (при использовании файловых систем);
- остановка кластера и изменение определений топологии и групп ресурсов;
- верификация и синхронизация кластера;
- запуск кластера.

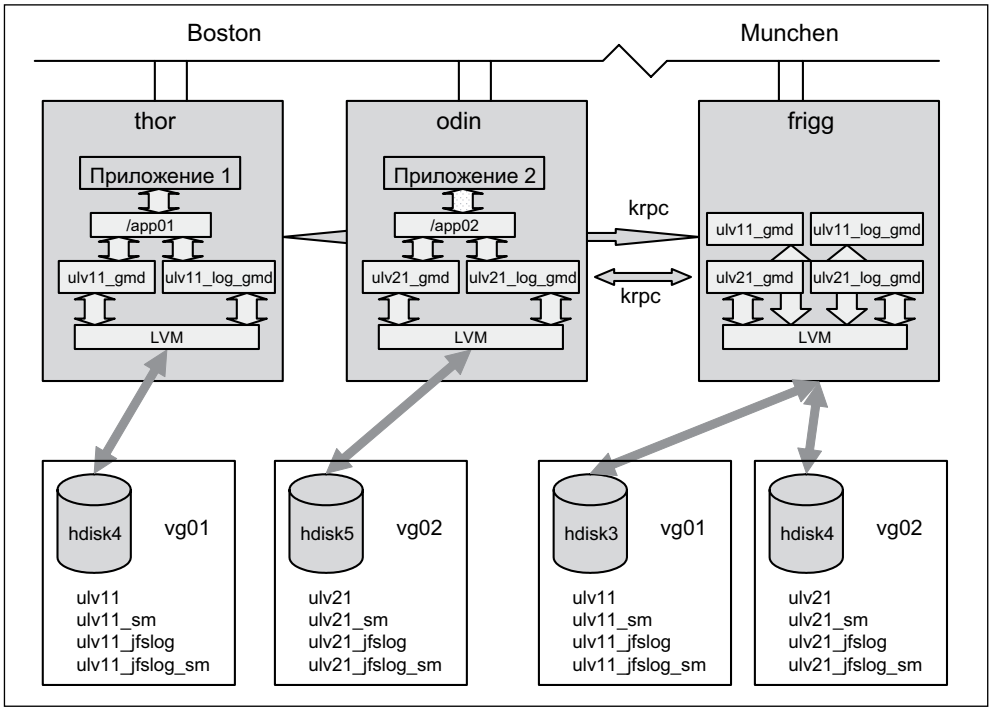

Рис. 18.12. Пример кластера HACMP/XD для миграции с GLVM

Рис. 18.12 изображает кластер с запущенным HACMP/XD:HAGeo. Узлы thor и odin находятся на сайте Boston; на каждом из узлов выполняется приложение. Каждое приложение использует один fleshiest с зеркальным отображением GeoMirror (зеркальное отображение выполняется и для базового логического тома, и для jfslog), который реплицируется на узел frigg сайта Munchen.

> **Примечание.** Описанные действия предполагают ограниченные возможности оборудования, поэтому резервная копия данных не будет доступна на протяжении всего периода миграции.

Мы подробно рассмотрим миграцию группы ресурсов только на узле thor, так как на всех узлах алгоритм работы одинаковый. Рис. 18.13 подробно изображает один реплицируемый ресурс.

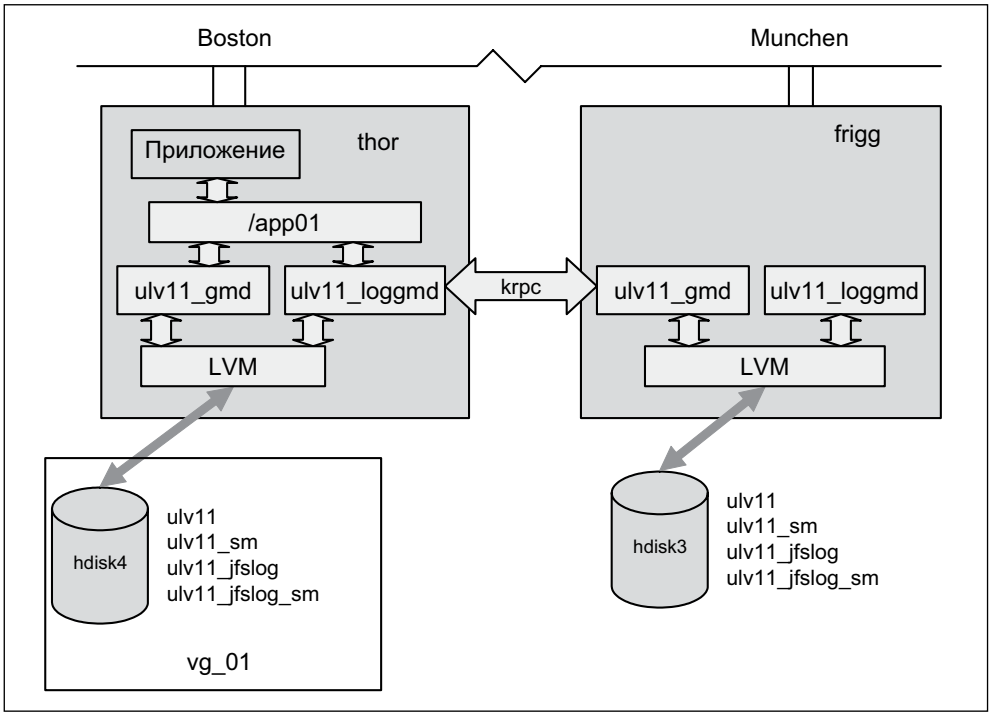

Рис. 18.13. Пример с одним ресурсом GMD

18.2.1 Установка наборов файлов GLVM и конфигурирование GLVM

Устанавливаем наборы файлов GLVM на каждом узле в кластере. Создаем снимок HACMP и HAGeo.

1. **Остановка удаленного сайта и создание RPV-серверов на удаленном сайте.** Цель миграции состоит в том, чтобы настроить использование GLVM при репликации данных без дополнительного оборудования. Недостаток этого процесса состоит в том,

что необходимо отключать удаленную копию GMD, а затем перезаписывать в качестве устройства GLVM, оставляя заказчика без резервной копии данных. Если это представляет проблему, то потребуется использование дополнительного оборудования.

В нашем примере происходит остановка реплицируемой копии группы ресурсов на резервном сайте (основное устройство GMD больше не выполняет синхронизацию данных), после чего происходит создание RPV-сервера, указывающего на физический том (рис. 18.14).

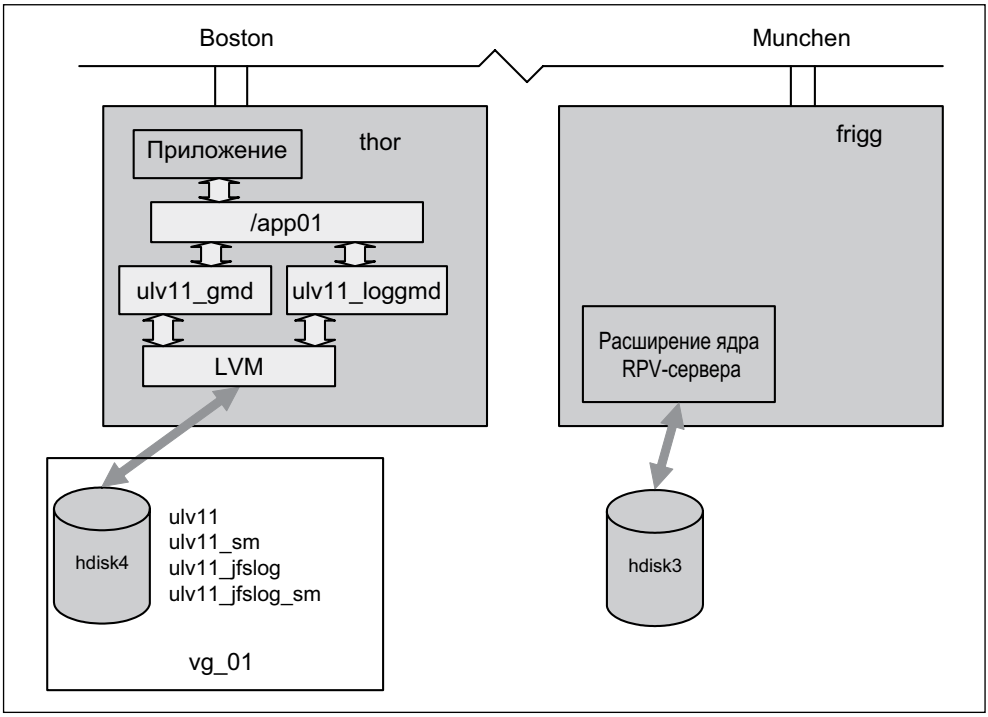

Рис. 18.14. Создание RPV-сервера

2. **Создание локального RPV-клиента и зеркального отображения.** Следующим этапом является создание локального RPV-клиента и добавление физического тома в группу томов GMD. Для всех логических томов данных и логических томов jfslog требуется установить политику размещения «super strict», после чего расширить путем добавления копии RPV-клиента.

Мы рекомендуем не выполнять зеркальное отображение логических томов statemap, так как после отключения GMD они будут не нужны.

После определения зеркальной копии логических томов можно осуществлять синхронизацию данных. При этом происходит репликация всех данных с основной копии даных на резервный сайт, что влияет на производительность обоих узлов и создает большую нагрузку на сеть (рис. 18.15).

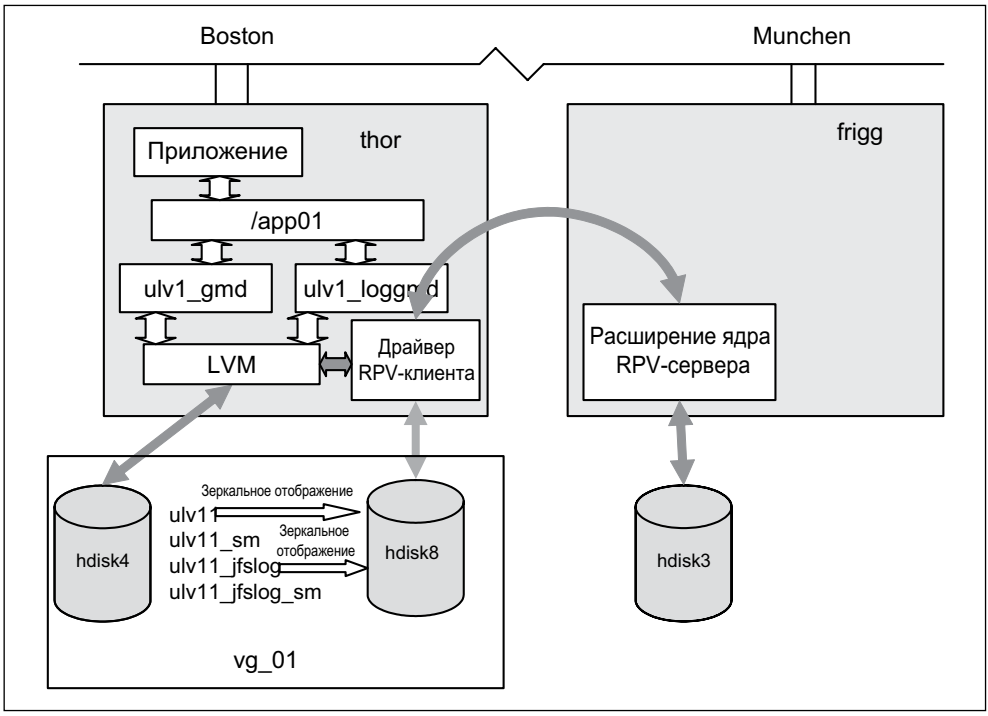

Рис. 18.15. Зеркальное отображение логических томов данных на RPV-клиенте

3. **Создание локальных RPV-серверов и удаленных RPV-клиентов.** Хотя они и не понадобятся до перемещения приложений при сбое, RPV-серверы и клиенты должны быть созданы на всех узлах, иначе верификация HACMP будет неуспешной.

4. **Изменение файла /etc/filesystems таким образом, чтобы он указывал на логические тома, а не на GMD.** Файловые системы теперь указывают на логический том, а не на GMD, поэтому на каждом узле необходимо внести изменения в файл /etc/filesystems. Все statemap или зеркальные логические тома следует удалить из GMVG, так как HACMP/XD выдаст ошибку при наличии нереплицируемых логических томов на любом физическом томе в GMVG.

5. **Остановка кластера и изменение определений топологии и групп ресурсов.** Так как HACMP/XD:HAGeo не поддерживает динамическую реконфигурацию, необходимо остановить кластер на всех узлах:

- для конфигурирования сети XD_Data;
- для удаления определений GMD из группы ресурсов;
- для настройки принудительной активизации для групп томов GLVM.

6. Верификация и синхронизация кластера. Необходимо провести верификацию и синхронизацию изменений в GLVM на каждом узле в кластере.

7. Запуск кластера. Теперь можно запустить кластер с измененной группой ресурсов, использующей устройства GLVM (рис. 18.16), а также остальными группами

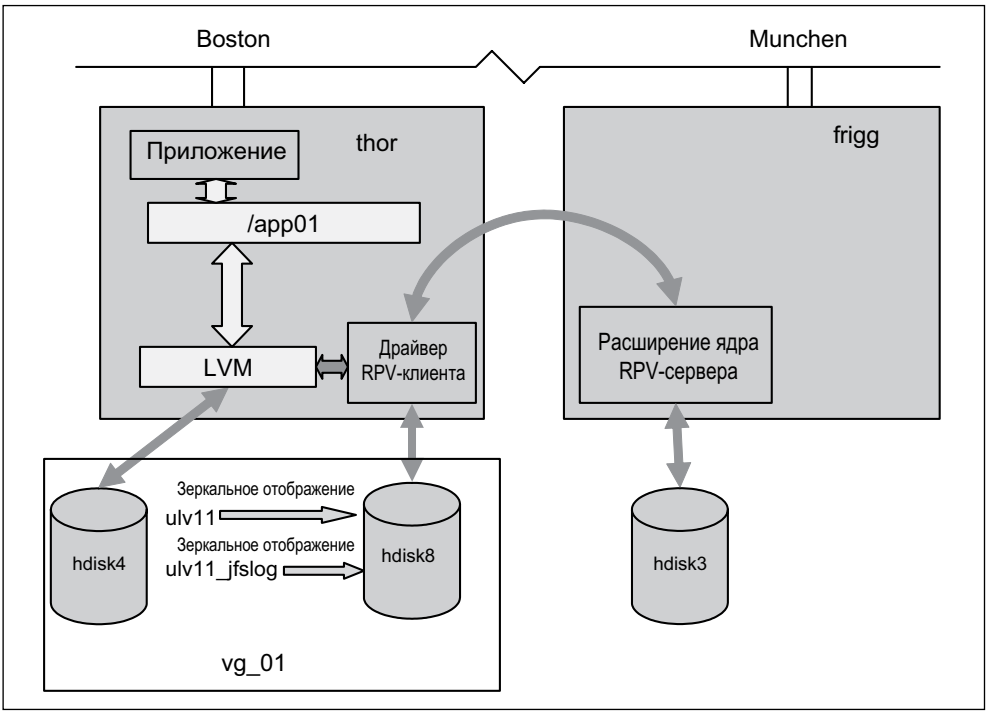

Рис. 18.16. Приложение использует устройства GLVM

ресурсов, использующими GMD. Неиспользуемые определения GMD теперь можно удалить.

Еще один вариант, который можно рассмотреть, заключается в том, чтобы остановить только группу ресурсов, содержащую устройства GMD, подлежащие замене, выполнить принудительное отключение кластера на всех узлах, внести изменения в конфигурации топологии и группы ресурсов, после чего перезапустить кластер. В этом случае простой возникнет только для изменяемой группы ресурсов.

18.2.2 Аспекты производительности

Хотя это и не относится к теме данной книги, при планировании конфигурации GMVG необходимо учитывать следующее:

1. Согласование операций записи зеркальных отображений. Можно отключить согласование операций чтения-записи зеркальных отображений для повышения производительности, однако при перезагрузке после отказа необходимо выполнить команду **syncvg -f**, прежде чем можно будет получить доступ к логическому тому. Логический том может быть установлен:

   а) Как активный. Используется по умолчанию для зеркального логического тома. Обеспечивает быстрое восстановление после отказа системы (не требует выполнения команды **syncvg -f** после перезагрузки). Это может вызвать проблемы с производительностью при выполнении операций записи.

б) Как пассивный. Отсутствуют проблемы с производительностью при выполнении операций записи и не требуется выполнение команды **syncvg -f** после перезагрузки. При этом выполняется фоновая ресинхронизация всех разделов, если обнаруживается, что система не была остановлена постепенно (gracefully).

2. Политики планирования LVM. Для зеркальных логических томов определено четыре политики чтения-записи:

   а) Параллельная. Выполняется балансировка операций чтения по физическим томам (они направляются на устройство с самой короткой очередью), операции записи направляются на физические устройства параллельно (т. е. одновременно).

   б) Последовательная. Операции чтения выполняются с основной копии, и операции записи выполняются последовательно (т. е. одна копия за другой).

   в) Параллельная запись, последовательное чтение. Операции чтения осуществляются с основной копии, а операции записи направляются на все физические тома параллельно.

   г) Параллельная запись, циклическое чтение. Операции чтения выполняются с каждой копии по очереди, а операции записи направляются на все физические тома параллельно.

3. Проверка записи. Возможны два варианта:

   а) Yes (Да). После всех операций записи на логический том выполняются операции чтения.

   б) No (Нет). Проверка операций записи не выполняется.

Для групп томов GMVG нужно учитывать следующее:

1. Согласование операций записи зеркальных отображений. Мы рекомендуем оставить согласование операций записи зеркальных отображений в активном состоянии, так как отказ узла вызовет синхронизацию всего логического тома. Однако если пропускная способность сети и размеры логических томов позволяют, следует рассмотреть вариант использования пассивного режима.

2. Политики планирования LVM. Рекомендуется использовать заданную по умолчанию параллельную политику, так как разработчики LVM сделали небольшое изменение для групп томов GMVG. Изменение состоит в том, что, если доступны физические тома, LVM пытается выполнить чтение с локальной копии, а не с RPV.

3. Проверку записи. Настоятельно рекомендуется оставить выключенной, как и задано по умолчанию.

18.2.3 Устранение неполадок

- В отличие от HAGEO журнал syslog содержит очень мало данных – одно трассируемое событие (trace hook 4A6).
- Снимок HACMP содержит выходные данные команд **lsrpvserver -H** и **lsrpvclient -H** в файле .info.
- snap -g содержит конфигурации RPV-сервера и клиента.

- general.snap – наборы файлов; атрибуты RPV-сервера и RPV-клиентов.
- CuAt – содержит информацию об имени удаленного сайта.

Пример 18.1 содержит свойства RPV-сервера.

Пример 18.1. Проверка свойств RPV-сервера

```
frigg:/# lsattr -El rpvserver0
auto_online n                               Configure at System Boot   True
client_addr 192.168.101.74                  Client IP Address          True
client_addr 192.168.101.73                  Client IP Address          True
rpvs_pvid   0022be2aa13f292e0000000000000000 Physical Volume Identifier True
frigg:/# lsattr -El hdisk7
io_timeout  180                             I/O Timeout Interval       True
local_addr  10.1.101.192                    Local IP Address           True
pvid        0022be2aa13dc0720000000000000000 Physical Volume Identifier True
server_addr none                            Server IP Address          True
```

Ниже также представлен пример фрагмента информации об ошибках RPV (пример 18.2).

Пример 18.2. Образец информации об ошибках RPV

```
odin:/# lsrpvserver -H
# RPV Server     Physical Volume Identifier       Physical Volume
# ---------------------------------------------------------------
  rpvserver0     0022be2aa13dc072                 hdisk2
odin:/# lsrpvclient -H
# RPV Client     Physical Volume Identifier       Remote Site
# ---------------------------------------------------------------
  hdisk6         0022be2aa13f292e                 Munchen

LABEL:          RPVC_IO_TIMEOUT
IDENTIFIER:     D034B795
Date/Time:       Thu Jul 14 15:48:03 2005
Sequence Number: 16314
Machine Id:      002574004C00
Node Id:         frigg
Class:           U
Type:            PERM
Resource Name:   hdisk7
Resource Class:  disk
Resource Type:   rpvclient
Location:
VPD:

Description
No response from RPV server within I/O timeout interval.

Probable Causes
RPV server is down or not reachable.
```

```
Failure Causes
There is a problem with the data mirroring network.
The node or site which hosts the RPV server is down.
RPV server is not configured in the Available state.

        Recommended Actions
        Correct the problem which has caused the RPV server to be down or
        not reachable.  Then, tell the RPV client to resume communication
        with the RPV server by running the command:
                chdev -l <device> -a resume=yes
        where <device> is the name of this RPV client device.

---------------------------------------------------------------------------
```

18.3 Этапы миграции с HAGEO на GLVM

Установка пакета для GLVM. Выберите следующие пакеты с установочного носителя:

- cluster.doc.en_US.glvm.html,
- cluster.doc.en_US.glvm.pdf,
- cluster.xd.glvm,
- glvm.rpv.client,
- glvm.rpv.server,
- glvm.rpv.util.

1. Начинаем с постепенной остановки служб кластера на узле frigg. Это остановит устройства GeoMirror на удаленном сайте Munchen:

smitty clstop

Подождите, пока службы кластера не будут остановлены на удаленном узле. Устройства GeoMirror будут находиться в состоянии «Defined».

Экспортируйте определение группы томов GMD на узле frigg:

exportvg vg01

Эта операция удаляет определение группы томов из ODM и удаляет разделы файловых систем из /etc/filesystems.

Сконфигурируйте среду RPV-сервера. Выполните следующие действия с RPV-сервера.

2. Назначение имени сайта удаленного зеркального отображения. На узле frigg выберите **smitty rpvserver -> Remote Physical Volume Server Site Name Configuration (Настройка имени сайта сервера удаленных физических томов) -> Define / Change / Show Remote Physical Volume Server Site Name (Определение / Изменение / Вывод имени сайта сервера удаленных физических томов)**. Определите имя сайта, как в определении сайта в HACMP.

Для определения сайта можно использовать команду **rpvsitename**:

/usr/sbin/rpvsitename -a ‘Munchen’

3. В меню Remote Physical Volume Servers (Серверы удаленных физических томов) выберите Add Remote Physical Volume Servers (Добавление серверов удаленных физических томов), чтобы определить RPV-серверы, связанные с целевыми дисками для зеркального отображения. После выбора целевых дисков укажите IP-адрес RPV-клиента, как в примере 18.3.

Пример 18.3. Добавление RPV-сервера

```
                    Add Remote Physical Volume Servers

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
  Physical Volume Identifiers                        0022be2aa13f292e
* Remote PV Client Internet Address     [192.168.101.73,192.168.101.74]
+
  Configure Automatically at System Restart?         [no] +
  Start New Devices Immediately?                     [yes] +


F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

При использовании командной строки используйте команду **mkdev**, как в примере 18.4:

Пример 18.4. Добавление RPV-сервера из командной строки

```
frigg:/# /usr/sbin/mkdev -c rpvserver -s rpvserver -t rpvstype \
>-a rpvs_pvid='0022be2aa13f292e' -a client_addr='192.168.101.73,\
192.168.101.74' -a auto_online='n'
rpvserver0 Available
```

Повторите действия пп. 1 и 2 для второго RPV.

Используйте **lsrpvserver** для вывода списка определенных RPV-серверов, как показано в примере 18.5.

Пример 18.5. Вывод списка RPV-серверов

```
frigg:/# lsrpvserver -H
# RPV Server     Physical Volume Identifier       Physical Volume
# ---------------------------------------------------------
  rpvserver0     0022be2aa13f292e                 hdisk1
frigg:/# lsattr -El rpvserver0
auto_online n                              Configure at System Boot   True
client_addr 192.168.101.73                 Client IP Address          True
client_addr 192.168.101.74                 Client IP Address          True
rpvs_pvid   0022be2aa13f292e0000000000000000 Physical Volume Identifier True
```

Сконфигурируйте RPV-клиенты. Выполните следующие действия на каждом клиенте.

4. Выберите **smitty rpvclient -> Add Remote Physical Volume Clients (Добавление клиентов удаленных физических томов)**.

Укажите IP-адрес RPV-сервера и локальный IP-адрес, используемые для репликации данных. Затем выберите удаленный диск из списка, как показано в примере 18.6.

Пример 18.6. Добавление RPV-клиента

```
                    Add Remote Physical Volume Clients

Type or select values in entry fields.

Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
  Remote Physical Volume Server Internet Address      10.1.101.192
  Remote Physical Volume Local Internet Address       192.168.101.74
  PV Identifiers                        0022be2aa13f292e0000000000000000
  I/O Timeout Interval (Seconds)                     [180] #
  Start New Devices Immediately?                     [yes] +

F1=Help              F2=Refresh           F3=Cancel            F4=List
F5=Reset             F6=Command           F7=Edit              F8=Image
F9=Shell             F10=Exit             Enter=Do
```

При использовании командной строки см. пример 18.7:

Пример 18.7. Добавление RPV-клиента из командной строки

```
thor:/# /usr/sbin/mkdev -c disk -s remote_disk -t rpvclient \
>-a pvid='0022be2aa13f292e' -a server_addr='10.1.101.192' \
>-a local_addr='192.168.101.73' -a io_timeout='180'
hdisk6 Available

thor:/# lsattr -El hdisk6
io_timeout  180                              I/O Timeout Interval       True
local_addr  192.168.101.73                   Local IP Address           True
pvid        0022be2aa13f292e0000000000000000 Physical Volume Identifier True
server_addr 10.1.1.192                       Server IP Address          True
```

К этому моменту на клиенте созданы дисковые устройства, которые можно использовать для интеграции в группу томов и определения зеркальных отображений логических томов. Используйте команду lsrpvclient для вывода списка определенных клиентских RPV.

На уровне операционной системы они определяются как обычные жесткие диски. Команды LVM, употребляемые для локальных томов, применяются и к RPV. Пример 18.8 содержит выходные данные команды **lsdev**.

Пример 18.8. Вывод списка физических томов, определенных на узле thor

```
thor:/# lspv
hdisk0          0022be2a80b97feb                    rootvg          active
hdisk1          none                                None
hdisk2          0022be2aa13dc072                    vg01            concurrent
hdisk3          0022be2aa13ea83e                    vg02            concurrent
hdisk4          none                                None
hdisk5          none                                None
hdisk6          0022be2aa13f292e                    None
```

Примечание. PVID RPV-клиента соответствует PVID удаленного диска.

Повторите действия пп. 1–3 для создания обратной пары RPV, связывающей RPV-сервер для локального диска на узле thor с RPV-клиентом на узле frigg.

Повторите те же действия на узле odin, используя odin_geo1 в качестве локального адреса связи.

Определение зеркального отображения LVM

5. Выполните расширение группы томов, содержащей основные данные с определенными RPV.

Используйте меню GLVM в SMIT для расширения группы томов. Выполните **smitty glvm_vg -> Add Remote Physical Volumes to a Volume Group (Добавление удаленных физических томов в группу томов)** или используйте команду **extendvg**:

extendvg vg01 hdisk6

6. Создайте зеркальное отображение группы томов, содержащей RPV.

Примечание. Перед миграцией логического тома необходимо установить политику размещения «super strict». Используйте команду chlv -s s < lv_name> -u <upper_bound>, чтобы изменить политику размещения на «super strict». Дополнительные сведения см. в документации по команде chlv.

В примере 18.9 представлено изменение логических томов ulv11_log и ulv11:

Пример 18.9. Изменение логических томов

```
thor:/# chlv -s s -u 2 ulv11_log
thor:/# lslv ulv11_log
LOGICAL VOLUME:     ulv11_log              VOLUME GROUP:   vg01
LV IDENTIFIER:      0022be2a00004c0000000104d52d0c6d.1 PERMISSION:
read/write
VG STATE:           active/complete        LV STATE:       opened/syncd
TYPE:               jfs2log                WRITE VERIFY:   off
MAX LPs:            512                    PP SIZE:        16 megabyte(s)
COPIES:             1                      SCHED POLICY:   parallel
LPs:                1                      PPs:            1
STALE PPs:          0                      BB POLICY:      relocatable
```

```
INTER-POLICY:       minimum                 RELOCATABLE:    yes
INTRA-POLICY:       middle                  UPPER BOUND:    2
MOUNT POINT:        N/A                     LABEL:          None
MIRROR WRITE CONSISTENCY: on/ACTIVE
EACH LP COPY ON A SEPARATE PV ?: yes (superstrict)
Serialize IO ?:     NO
thor:/# chlv -s s -u 2 ulv11
thor:/# lslv ulv11
LOGICAL VOLUME:     ulv11                   VOLUME GROUP:   vg01
LV IDENTIFIER:      0022be2a00004c0000000104d52d0c6d.2 PERMISSION:
read/write
VG STATE:           active/complete         LV STATE:       opened/syncd
TYPE:               jfs2                    WRITE VERIFY:   off
MAX LPs:            512                     PP SIZE:        16 megabyte(s)
COPIES:             1                       SCHED POLICY:   parallel
LPs:                10                      PPs:            10
STALE PPs:          0                       BB POLICY:      relocatable
INTER-POLICY:       minimum                 RELOCATABLE:    yes
INTRA-POLICY:       middle                  UPPER BOUND:    2
MOUNT POINT:        N/A                     LABEL:          /app01
MIRROR WRITE CONSISTENCY: on/ACTIVE
EACH LP COPY ON A SEPARATE PV ?: yes (superstrict)
Serialize IO ?:     NO
thor:/#
```

Выполните зеркальное отображение группы томов, выбрав smitty glvm_vg → Add a Remote Site Mirror Copy to a Logical Volume (Добавление зеркальной копии удаленного сайта в логический том). Можно использовать команду **mirrorvg** для зеркального отображения группы томов или команду **mklvcopy** для зеркального отображения логических томов, например:

/usr/sbin/mklvcopy -s's' ulv11_log 2 hdisk6

Проверьте состояние группы томов и логических томов с использованием команды **lsvg**, как показано в примере 18.10:

Пример 18.10. Использование команды lsvg для запроса состояния зеркальных отображений логических томов

```
thor:/# lsvg -p vg01
vg01:
PV_NAME           PV STATE      TOTAL PPs   FREE PPs    FREE DISTRIBUTION
hdisk2            active        639         476         128..00..92..128..128
hdisk6            active        639         478         128..02..92..128..128
thor:/# lsvg -l vg01
vg01:
LV NAME             TYPE       LPs   PPs   PVs  LV STATE      MOUNT POINT
ulv11_log           jfs2log    1     2     2    open/syncd    N/A
ulv11               jfs2       160   320   2    open/stale    N/A
ulv11_sm            statemap   1     1     1    open/syncd    N/A
ulv11_log_sm        statemap   1     1     1    open/syncd    N/A
```

7. Выполните постепенную остановку служб кластера на локальном узле через меню smitty clstop. Проверьте корректность остановки обработки ресурсов кластера. Используйте команду **lsgmd,** чтобы убедиться в том, что GMD находятся в состоянии «Defined».

8. На каждом узле в кластере измените определение файловой системы в файле /etc/filesystems, чтобы использовать обычные логические тома, а не GMD (пример 18.11).

Пример 18.11. Изменение файловых систем для работы с логическими томами

```
/app01:
        dev             = /dev/ulv11
        vfs             = jfs2
        log             = /dev/ulv11_log
        mount           = false
        check           = false
        account         = false
```

> **Важно!** Если изначально файловые системы создавались с использованием команды crfs, происходит обновление контрольного блока логического тома (logical volume control block, LVCB) информацией файловой системы, так что каждая команда importvg обновляет файл /etc/filesystems. Проверку данных LVCB можно выполнить с применением команды getlvcb -AT <lv_name>. Если был создан fleshiest over GMD с использованием команды mkfs, команда importvg не обнов-ляет информацию fleshiest в файле /etc/filesystem.

9. Изменение определений топологии и ресурсов HACMP для использования GLVM.

> **Примечание.** HACMP/XD HAGEO не поддерживает динамическую конфигурацию. Для изменения конфигурации кластера необходимо остановить службы кластера. HACMP/XD GLVM поддерживает динамическую реконфигурацию, если только не установлен HAGEO.

Для интеграции групп томов GLVM в HACMP требуется убедиться в том, что осуществляется репликация каждого логического тома. Если группы томов с географическим зеркальным отображением содержат нереплицируемые логические тома, HACMP выдает сообщение об ошибке.

10. Реконфигурация топологии кластера.

11. Измените тип сети с Geo_Primary на XD_data. На момент публикации этой книги использование двух сетей XD_data не поддерживалось. В кластере могут быть одновременно сконфигурированы GMD и RPV. Однако ресурсы GMD и RPV не могут входить в одну группу ресурсов. Если у вас имеется две сети Geo_Primary, можно оставить вторую сеть для непреобразованных GMD.

Пример 18.12 показывает преобразование первой сети Geo_Primary в тип XD_data.

Пример 18.12. Преобразование сети HAGEO в сеть XD_data

```
                Change/Show an IP-Based Network in the HACMP Cluster

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
* Network Name                                        net_Geo_Primary_01
  New Network Name                                      [XD_data_net_01]
* Network Type                                                 [XD_data]+
* Netmask                                               [255.255.255.0]+
* Enable IP Address Takeover via IP Aliases                          No+
  IP Address Offset for Heartbeating over IP Aliases                  []
* Network attribute                                             public+

F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

> **Примечание.** При изменении атрибута сети Geo_Primary с private на public, необходимо удалить и заново создать сеть.

12. Синхронизация топологии кластера.

13. Изменение групп ресурсов для интеграции RPV. Не требуется конфигурировать специальные ресурсы для использования RPV в кластере. На данном этапе следует удалить определения GMD из групп ресурсов (пример 18.13).

Пример 18.13. Определение группы ресурсов в HACMP

```
Change/Show All Resources and Attributes for a Resource Group

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
  Resource Group Name                                  app01_rg
  Inter-site Management Policy                         Prefer Primary Site
  Participating Nodes from Primary Site                thor odin
  Participating Nodes from Secondary Site              frigg
  Startup Policy                                       Online On Home Node Only
  Fallover Policy                                     Fallover To Next Priority
Node In The List
  Fallback Policy                                      Fallback To Higher
Priority Node In The Li>
  Fallback Timer Policy (empty is immediate)          []                    +
  Service IP Labels/Addresses                         []                    +
```

```
  Application Servers                                  [app01_srv]          +
  Volume Groups                                        [vg01 ]              +
  Use forced varyon of volume groups, if necessary      true                +
  Automatically Import Volume Groups                    false               +
  Filesystems (empty is ALL for VGs specified)         [/app01 ]            +
  Filesystems Consistency Check                         fsck                +
  Filesystems Recovery Method                           sequential          +
  Filesystems mounted before IP configured              false               +
  Filesystems/Directories to Export                    []                   +
  Filesystems/Directories to NFS Mount                 []
  Network For NFS Mount                                []                  +
  Tape Resources                                       []                  +
  Raw Disk PVIDs                                       []                  +
  Fast Connect Services                                []                  +
  Communication Links                                  []                  +
  Primary Workload Manager Class                       []                  +
  Secondary Workload Manager Class                     []                  +
  Miscellaneous Data                                   []
  GeoMirror Devices                                    []                  +

F1=Help         F2=Refresh          F3=Cancel           F4=List
F5=Reset        F6=Command          F7=Edit             F8=Image
F9=Shell        F10=Exit            Enter=Do
```

14. Синхронизация определения кластера по узлам.

15. Запуск кластера на узлах.

Часть 6

Приложения

A

Таблицы планирования на бумаге

Подробные таблицы планирования на бумаге приведены в прил. А руководства **HACMP 5.3 Planning and Installation Guide**.

Мы считаем, что было бы полезно привести эти таблицы к формату, соответствующему вашей среде. С этой целью мы включили в это приложение набор приспособленных таблиц, чтобы помочь в проектировании простого кластера.

Two-Node Cluster Configuration Assistant

Если вы планируете использовать Two-Node Cluster Configuration Assistant для конфигурирования первоначального кластера, следует использовать эту таблицу. Two-Node Cluster Configuration Assistant требует указания приведенной ниже информации для настройки простого кластера из двух узлов с одной группой ресурсов.

Таблица А-1. Two-Node Cluster Configuration Assistant

ТАБЛИЦА КЛАСТЕРА HACMP **TWO NODE CLUSTER CONFIGURATION ASSISTANT**	**ДАТА:**
Путь для связи с резервным узлом	
Сервер приложения	
Скрипт запуска приложения	
Скрипт остановки приложения	
Сервисная IP-метка	

Таблицы планирования узлов

Следующие таблицы планирования можно использовать для направления процесса планирования и реализации кластера HACMP. Как вы увидите, таблицы охватывают все важные аспекты конфигурации кластера и следуют логическому алгоритму планирования.

Таблица А-2. Обзор кластера

ТАБЛИЦА КЛАСТЕРА HACMP – ЧАСТЬ 1 из 11 **ОБЗОР КЛАСТЕРА**	**ДАТА:**
ИМЯ КЛАСТЕРА	
ОРГАНИЗАЦИЯ	
ИМЯ ХОСТА УЗЛА 1	
ИМЯ ХОСТА УЗЛА 2	
ИМЯ HACMP УЗЛА 1	
ИМЯ HACMP УЗЛА 2	
КОММЕНТАРИИ	

Таблица А-3. Оборудование кластера

ТАБЛИЦА КЛАСТЕРА HACMP – ЧАСТЬ 2 из 11 **ОБОРУДОВАНИЕ КЛАСТЕРА**		**ДАТА:**
КОМПОНЕНТ ОБОРУДОВАНИЯ	СПЕЦИФИКАЦИИ	КОММЕНТАРИИ
КОММЕНТАРИИ		

Таблица А-4. Программное обеспечение кластера

ТАБЛИЦА КЛАСТЕРА HACMP – ЧАСТЬ 3 из 11 **ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ КЛАСТЕРА**		**ДАТА:**
КОМПОНЕНТ ПО	ВЕРСИЯ	КОММЕНТАРИИ
AIX		
RSCT		
HACMP		
КОММЕНТАРИИ		

Таблица А-5. Ethernet-сети кластера

ТАБЛИЦА КЛАСТЕРА HACMP – ЧАСТЬ 4 из 11 **ETHERNET-СЕТИ КЛАСТЕРА**					**ДАТА:**
ИМЯ СЕТИ	ТИП СЕТИ	МАСКА СЕТИ	ИМЕНА УЗЛОВ	IPAT ЧЕРЕЗ СИНОНИМЫ	Смещение IP-адреса для мониторинга пульса через IP-синонимы
КОММЕНТАРИИ					

Таблица А-6. Последовательные сети кластера

ТАБЛИЦА КЛАСТЕРА HACMP – ЧАСТЬ 5 из 11 **ПОСЛЕДОВАТЕЛЬНЫЕ СЕТИ КЛАСТЕРА И СЕТИ ТИПА «ТОЧКА-ТОЧКА»**					**ДАТА:**
ИМЯ СЕТИ	ТИП СЕТИ	ИМЕНА УЗЛОВ	Устройство	ИМЯ ИНТЕРФЕЙСА	МЕТКА АДАПТЕРА
КОММЕНТАРИИ					

Таблица А-7. Интерфейсы

ТАБЛИЦА КЛАСТЕРА НАСМР – ЧАСТЬ 6 из 11 **ИНТЕРФЕЙСЫ**					**ДАТА:**
node01					
IP-метка	Предпочитаемое распределение IP-синонима	СЕТЕВОЙ ИНТЕРФЕЙС	ИМЯ СЕТИ	ФУНКЦИЯ ИНТЕРФЕЙСА	IP-АДРЕС/МАСКА
node02					
IP-метка	Предпочитаемое распределение IP-синонима	СЕТЕВОЙ ИНТЕРФЕЙС	ИМЯ СЕТИ	ФУНКЦИЯ ИНТЕРФЕЙСА	IP-АДРЕС/МАСКА
КОММЕНТАРИИ					

Таблица А-8. Общие диски

ТАБЛИЦА КЛАСТЕРА НАСМР – ЧАСТЬ 7 из 11 **ОБЩИЕ ДИСКИ**					**ДАТА:**
node01			node02		
VGNAME	VPATH	HDISK	HDISK	VPATH	VGNAME
КОММЕНТАРИИ					

Таблица А-9. Общие группы томов

ТАБЛИЦА КЛАСТЕРА HACMP – ЧАСТЬ 8 из 11 **ОБЩИЕ ГРУППЫ ТОМОВ (БЕЗ ОДНОВРЕМЕННОГО ДОСТУПА)**		**ДАТА:**
ГРУППА РЕСУРСОВ	ГРУППА ТОМОВ 1	ГРУППА ТОМОВ 1
КОММЕНТАРИИ	Создается общая группа томов на первом узле, после чего выполняется импорт на второй узел; #importvg -y app1vg -V 90 vpath0 (может быть, необходимо сделать pv доступным с использованием chdev -l vpath0 -a pv=yes); #chvg -an app1vg (отключает автоматическую активизацию для группы томов); #mount /app1 (для того, чтобы убедиться в возможности подключения файловой системы); #umount /app1; #varyoffvg app1vg (оставляет группу томов в отключенном режиме, оставляя HACMP управление группой)	

Таблица А-10. Таблица приложений

ТАБЛИЦА КЛАСТЕРА HACMP – ЧАСТЬ 9 из 11 **ТАБЛИЦА ПРИЛОЖЕНИЙ**				**ДАТА:**
APP1				
ЭЛЕМЕНТ	КАТАЛОГ	ФАЙЛОВАЯ СИСТЕМА	РАСПОЛОЖЕНИЕ	ДОСТУП
ИСПОЛНЯЕМЫЕ ФАЙЛЫ				
ФАЙЛЫ КОНФИГУРАЦИИ				
ФАЙЛЫ ДАННЫХ				
ФАЙЛЫ ЖУРНАЛОВ				
СКРИПТ ЗАПУСКА				
СКРИПТ ОСТАНОВКИ				
СТРАТЕГИЯ ПЕРЕМЕЩЕНИЯ ПРИ СБОЕ				
КОМАНДЫ И ПРОЦЕДУРЫ ОБЫЧНОГО ЗАПУСКА				
КОМАНДЫ И ПРОЦЕДУРЫ ВЕРИФИКАЦИИ				

Окончание табл. А-10.

ТАБЛИЦА КЛАСТЕРА НАСМР – ЧАСТЬ 9 из 11 **ТАБЛИЦА ПРИЛОЖЕНИЙ**				**ДАТА:**
КОМАНДЫ И ПРОЦЕДУРЫ ОБЫЧНОЙ ОСТАНОВКИ				
РЕИНТЕГРАЦИЯ УЗЛА				
APP2				
ЭЛЕМЕНТ	КАТАЛОГ	ФАЙЛОВАЯ СИСТЕМА	РАСПОЛОЖЕНИЕ	ДОСТУП
ИСПОЛНЯЕМЫЕ ФАЙЛЫ				
ФАЙЛЫ КОНФИГУРАЦИИ				
ФАЙЛЫ ДАННЫХ				
ФАЙЛЫ ЖУРНАЛОВ				
СКРИПТ ЗАПУСКА				
СКРИПТ ОСТАНОВКИ				
СТРАТЕГИЯ ПЕРЕМЕЩЕНИЯ ПРИ СБОЕ				
КОМАНДЫ И ПРОЦЕДУРЫ ОБЫЧНОГО ЗАПУСКА				
КОМАНДЫ И ПРОЦЕДУРЫ ВЕРИФИКАЦИИ				
КОМАНДЫ И ПРОЦЕДУРЫ ОБЫЧНОЙ ОСТАНОВКИ				
РЕИНТЕГРАЦИЯ УЗЛА				
КОММЕНТАРИИ				

Таблица А-11. Таблица мониторинга приложений

ТАБЛИЦА КЛАСТЕРА НАСМР – ЧАСТЬ 10 из 11 **МОНИТОРИНГ ПРИЛОЖЕНИЙ**	**ДАТА:**
APP1	
Можно ли осуществлять мониторинг приложения с использованием монитора процессов?	
Наблюдаемые процессы	
Владелец процесса	

Окончание табл. А-11.

ТАБЛИЦА КЛАСТЕРА НАСМР – ЧАСТЬ 10 из 11 **МОНИТОРИНГ ПРИЛОЖЕНИЙ**	**ДАТА:**
Счетчик экземпляров	
Стабилизационный интервал	
Счетчик перезапуска	
Интервал перезапуска	
Действие при отказе приложения	
Метод уведомления	
Метод очистки	
Метод перезапуска	
APP2	
Можно ли осуществлять мониторинг приложения с использованием монитора процессов?	
Наблюдаемые процессы	
Владелец процесса	
Счетчик экземпляров	
Стабилизационный интервал	
Счетчик перезапуска	
Интервал перезапуска	
Действие при отказе приложения	
Метод уведомления	
Метод очистки	
Метод перезапуска	

Таблица А-12. Группы ресурсов

ТАБЛИЦА КЛАСТЕРА НАСМР – ЧАСТЬ 11 из 11 **ГРУППЫ РЕСУРСОВ**		**ДАТА:**
ИМЯ РЕСУРСА		
Политика межсайтового управления		
Имена узлов-участников		
Политика запуска		
Политика перемещения при сбое		
Политика возврата после восстановления		
Таймер отсроченного возврата		
Время установления		
Политики времени выполнения		
Политика динамического приоритета узла		
Порядок обработки (параллельная, последовательная или настраиваемая)		

Окончание табл. А-11.

ТАБЛИЦА КЛАСТЕРА HACMP – ЧАСТЬ 11 из 11 **ГРУППЫ РЕСУРСОВ**		**ДАТА:**
Сервисная IP-метка		
Серверы приложений		
Группы томов		
Файловые системы		
Проверка согласованности файловой системы		
Метод восстановления файловой системы		
Файловые системы или каталоги для экспорта		
Файловые системы или каталоги для подключения NFS		
Сеть для подключения NFS		
Основной класс WLM		
Автоматический импорт групп томов		
Подключение файловых систем перед конфигурированием IP		
КОММЕНТАРИИ		

Список сокращений

ACL	Access Control List – Список контроля доступа
AIX	Advanced Interactive Executive
API	Application Programming Interface – Программный интерфейс приложения
ARP	Address Resolution Protocol – Протокол разрешения адресов
ATM	Asynchronous Transfer Mode
BOS	Base Operating System – Базовая операционная система
CA	Certificate Authority – Уполномоченный центр выдачи сертификатов
C-SPOC	Cluster Single Point Of Control – Единая точка управления кластером
CEC	Central Electronic Complex – Центральный электронный комплекс
CGI	Common Gateway Interface
CLI	Command Line Interface – Интерфейс командной строки
CLVM	Concurrent Logical Volume Manager – Диспетчер логических томов с одновременным доступом
CPU	Central Processing Unit – Центральный процессор
CSM	Cluster Systems Management
CWDM	Coarse Wave Division Multiplexing
CWOF	Cascading Without Fallback – Каскадирование без возврата после восстановления
DAC	Disk Array Controller – Контроллер дискового массива
DARE	Dynamic Reconfiguration – Динамическая реконфигурация
DBFS	Dial Back Fail Safe
DES	Data Encryption System
DLPAR	Dynamic LPAR – Динамический LPAR
DNP	Dynamic Node Priority – Динамический приоритет узлов
DNS	Domain Name Service – Служба доменных имен
DWDM	Dense Wave Division Multiplexing
ECM	Enhanced Concurrent Mode – Режим расширенного одновременного доступа
ESS	Enterprise Storage Server
FC	Fibre Channel
FCIP	Fibre Channel IP
FDDI	Fiber Distributed Data Interface

GLVM	Geographical LVM – Географический LVM
GMD	Geographic Mirror Device – Устройство GeoMirror
GPFS	General Parallel File System
HACMP	High Availability Cluster Multi-Processing
HACMP/ES	HACMP Enhanced Scalability
HACMP/XD	HACMP Extended Distance
HA-NFS	High Availability NFS
HBA	Host Bus Adapter – Хост-адаптер
HPS	High Performance Switch
HSC	Hardware Service Console – Консоль обслуживания оборудования
HWAT	Hardware Address Takeover – Перехват аппаратного адреса
IBM	International Business Machines Corporation
IHS	IBM Http Server
IPAT	IP Address Takeover – Перехват IP-адреса
ITSO	International Technical Support Organization
JBOD	Just a Bunch Of Disks – Простой массив дисков (часто – обычный диск)
JFS	Journal File System
JRE	Java Runtime Environment
LAA	Locally Administered Address – Локально администрируемый адрес
LAN	Local Area Network – Локальная сеть
LDAP	Lightweight Directory Application Protocol
LPAR	Logical Partition – Логический раздел[1]
LUN	Logical Unit Number – Номер логического устройства
LV	Logical Volume – Логический том
LVCB	Logical Volume Control Block – Контрольный блок логического тома
LVDD	Logical Volume Device Driver – Драйвер логического тома
LVM	Logical Volume Manager – Диспетчер логических томов
MAC	Media Access Control
MAL	Mechanism Abstraction Layer – Абстрактный уровень механизма
MIB	Management Information Base – База управляющей информации
MPIO	Multi-Path I/O – Многопутевой ввод-вывод
MPM	Mechanism Pluggable Module – Подключаемые модули механизма
MTU	Media Transmission Unit[2] – Максимальный размер блока данных

[1] Чтобы не путать с LP (Logical Partition – Логический раздел, компонент LVM), часто используют сокращение LPAR, если смысл явно не понятен из контекста. *Примеч. науч. ред.*

[2] Обычно MTU расшифровывается как Maximum Transmission Unit. *Примеч. науч. ред.*

NFS	Network File System – Сетевая файловая система
NIM	Network Interface Module – Сетевой интерфейсный модуль
POL	Priority Override Location – Расположение, отменяющее приоритет
PP	Physical Partition – Физический раздел[1]
PPK	Private-Public Key – Открытый/закрытый ключ
PPRC	Peer-to-Peer Remote Copy
PV	Physical Volume – Физический том
PVID	Physical Volume ID – Идентификатор физического тома[2]
RAC	Real Application Cluster
RAID	Redundant Array of Independent Disks
RDAC	Redundant Disk Array Controller – RAID-контроллер
RG	Resource Group – Группа ресурсов
RM	Resource Monitor – Монитор ресурсов
RMC	Resource Monitoring and Control – Служба мониторинга и управления ресурсами
RPV	Remote Physical Volume – Удаленный физический том
RSCT	Reliable Scalable Clustering Technology
SAN	Storage Area Network – Сеть хранения данных
SCSI	Small Computer System Interface
SDD	Subsystem Device Driver
SPOF	Single Point Of Failure – Единая точка отказа
THL	Trusted Host List – Список доверенных хостов
VG	Volume Group – Группа томов
VGDA	Volume Group Descriptor Area – Область дескриптора группы томов
VIO	Virtual I/O – Виртуальный ввод-вывод
VIOS	Virtual I/O Server – Сервер виртуального ввода-вывода
VLAN	Virtual LAN – Виртуальная локальная сеть
WLM	Workload manager – Диспетчер рабочей нагрузки

[1] Компонент LVM. Не надо путать с PPAR (Физический раздел сервера, см. LPAR). *Примеч. науч. ред.*

[2] В другом контексте PVID - Port VLAN ID. *Примеч. науч. ред.*

Рекомендуемая литература

Здесь перечислены публикации, содержащие более подробное обсуждение вопросов, рассматриваемых в этой книге.

Книги серии IBM Redbooks

(сведения о приобретении этих публикаций см. ниже, в подразделе «Как приобрести книги серии IBM Redbooks». Обратите внимание на то, что некоторые приведенные документы могут быть доступны только в электронном варианте).

- **HACMP/ES Customization Examples**, SG24-4498
- **Understanding LDAP - Design and Implementation**, SG24-4986
- **Advanced POWER Virtualization on IBM System p5**, SG24-7940[1]

Прочие публикации

(эти публикации также можно использовать в качестве дополнительных источников информации).

- **High Availability Cluster Multi-Processing Administration Guide**, SC23-4862
- **High Availability Cluster Multi-Processing for AIX 5L Planning and Installation Guide**, SC23-4861
- **High Availability Cluster Multi-Processing XD (Extended Distance) for HA-GEO Technology: Concepts and Facilities**, SA22-7955
- **HACMP/XD for ESS PPRC Version 5.3: Planning and Administration Guide**, SC23-4863
- **HACMP/XD for Geographic LVM: Administration and Planning Guide**, SA23-1338

Сетевые ресурсы

(нижеперечисленные веб-сайты и URL также можно использовать в качестве дополнительных источников информации).

- Веб-сайт проекта OpenSSL:

  http://www.openssl.org

[1] Эта книга готовится к изданию на русском языке. *Примеч. науч. ред.*

- Веб-страница проекта OpenSSH.

  http://www.openssh.org
- OpenSSH в AIX.

  http://www.sourceforge.net/projects/openssh-aix
- Рекомендуемые уровни обслуживания HACMP.

  http://www-912.ibm.com/eserver/support/fixes/fcgui.jsp
- Домашняя страница Availant; тестирование оборудования сторонних производителей (отличных от IBM).

  http://www.availant.com

Как приобрести книги серии IBM Redbooks

Можно выполнять поиск, просматривать и копировать книги серии Redbooks, документы серии Redpapers, публикации серии Hints and Tips, проекты публикаций и дополнительные материалы, а также заказать книги серии Redbooks в печатном варианте или на компакт-дисках через веб-сайт

ibm.com/redbooks

Поддержка от компании IBM

Поддержка IBM Support и загружаемые пакеты

ibm.com/support

Служба IBM Global Services:

ibm.com/services