

Эссе

Тема: «Кластеры информационных систем»

Алексеев Е.А.- группа ВБИо-305рсоб

Москва 2024

## Развитие многопроцессорных вычислительных систем

Развитие традиционных архитектур построения вычислительных систем, таких как SMP, MPP, векторных параллельных систем идет достаточно быстрыми темпами. Повышается производительность, растет надежность и отказоустойчивость. Однако у этих архитектур есть один недостаток - стоимость создаваемых систем, подчас недоступная для многих пользователей таких систем - образовательных и научно-исследовательских организаций. Она оказывает очень высокой из-за усложнения аппаратных и программных составляющих системы, которые требуются для обеспечения таких темпов роста производиельности. Однако потребность в вычислительных ресурсах в настоящее время очень высока во многих сферах научной и практической деятельности и для ее обеспечения не хватает ресурсов традиционных суперкомпьютерных систем.

Кластерные системы возникли как более дешевое решение проблемы недостатка вычислительных ресурсов, и основываются на использовании в своей архитектуре широко распространенных и относительно дешевых технологий, аппаратных и программных средств, таких как PC, Ethernet, Linux и т.д. Использование массовых технологии в кластерных системах стало возможным благодаря значительному прогрессу в развитии компонентов обычных вычислительных систем, таких как центральные процессоры, операционные системы, коммуникационные среды.

Так как кластерные системы архитектурно являются развитием систем с массовым параллелизмом MPP, то главную роль в их развитии является прогресс в области сетевых технологий. К настоящему времени появились недорогие, но эффективные коммуникационные решения. Это и предопределило быстрое появление и развитие кластерных вычислительных систем. Также прогрессу развития кластерных систем способствовали и другие факторы.

Производительность персональных компьютеров на базе процессоров Intel в последние годы также значительно выросла. Такие компьютеры стали создавать серьезную конкуренцию рабочим станциям на базе более дорогих и мощных RISC процессоров. Одновременно стала приобретать все большую популярность ОС Linux - бесплатно распространяемая версия UNIX. При этом в научных организациях и университетах, где и разрабатывается большинство кластерных систем, как правило, имеются специалисты по ОС Linux.

Высокую степень развития кластерных систем на сегоднящний день показывает тот факт, что в списке самых мощных суперкомпьютеров мира Top500 – числится 11 кластерных установок.

# Основные классы современных параллельных компьютеров

Кластерные системы являются развитием параллельных систем. Чтобы проказать место кластерных систем среди остальных типов параллельных архитектур вычислительных систем нужно привести их классификацию . Параллельные системы могут быть класифицированы по различным критериям.

С аппаратной точки зрения, основным параметром классификации паралелльных компьютеров является наличие общей (SMP) или распределенной памяти (MPP). Нечто среднее между SMP и MPP представляют собой NUMA-архитектуры, где память физически распределена, но логически общедоступна.

## Симметричные мультипроцессорные системы

SMP система состоит из нескольких однородных процессоров и массива общей памяти. Один из часто используемых в SMP архитектурах подходов для формирования масштабируемой, общедоступной системы памяти, состоит в однородной организации доступа к памяти посредством организации масштабируемого канала память-процессоры:

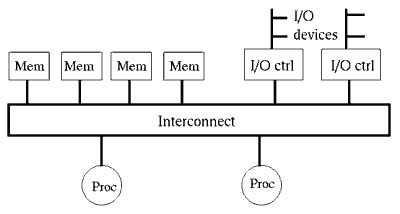


рис.1

Каждая операция доступа к памяти интерпретируется как транзакция по шине процессоры-память. Когерентность кэшей поддерживается аппаратными средствами.

В SMP каждый процессор имеет по крайней мере одну собственную кэш-память (а возможно, и несколько).

Можно сказать, что SMP система - это один компьютер с несколькими равноправными процессорами. Все остальное - в одном экземпляре: одна память, одна подсистема ввода/вывода, одна операционная система. Слово "равноправный" означает, что каждый процессор может делать все, что любой другой. Каждый процессор имеет доступ ко всей памяти, может выполнять любую операцию ввода/вывода, прерывать другие процессоры и т.д.

Недостатком данной архитектуры является необходимость организации канала процессоры-память с очень высокой пропускной способностью.

## Массивно-параллельные системы

Массивно-параллельная система MPP состоит из однородных вычислительных узлов, включающих в себя:

1. один или несколько центральных процессоров (обычно RISC)
2. локальную память (прямой доступ к памяти других узлов невозможен)
3. коммуникационный процессор или сетевой адаптер
4. жесткие диски и/или другие устройства В/В

К системе могут быть добавлены специальные узлы ввода-вывода и управляющие узлы. Узлы связаны через некоторую коммуникационную среду (высокоскоростная сеть, коммутатор и т.п.)

## Системы с неоднородным доступом к памяти NUMA

NUMA (nonuniform memory access) в отличие от привычной SMP архитектуры с разделяемой памятью представляет собой несколько отдельных процессоров, каждый из которых, кроме собственного кэша, обладает также локальной памятью:

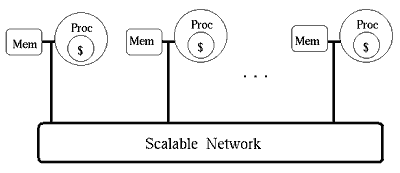


рис.2

В подобной архитектуре процессор и модули памяти тесно интегрированы, следовательно, скорость доступа к локальной памяти гораздо выше, чем к памяти “соседнего” процессора. Подсистемы ввода-вывода могут быть частью каждого узла или консолидированы на выделенных узлах ввода-вывода. Если во всей системе поддерживается когерентность кэшей, то такую архитектуру называют cc-NUMA.

Проще всего охарактеризовать NUMA-систему, представив себе большую систему SMP, разделенную на несколько частей, эти части связаны коммуникационной магистралью, подключенной к системным шинам, и каждая часть включает собственную основную память и подсистему ввода/вывода. Это и есть NUMA: большая SMP, разбитая на набор более мелких и простых SMP. Основной проблемой NUMA является обеспечение когерентности кэшей. Аппаратура позволяет работать со всеми отдельными устройствами основной памяти составных частей системы (называемых обычно узлами) как с единой гигантской памятью.

## Кластерная архитектура

Рассмотрим место кластерной архитектуры вычислительных систем в данной классификации.

Кластер - это связанный набор полноценных компьютеров, используемый в качестве единого ресурса. Под понятием "полноценный компьютер" понимается завершенная компьютерная система, обладающая всем, что требуется для ее функционирования, включая процессоры, память, подсистему ввода/вывода, а также операционную систему, подсистемы, приложения и т.д. Обычно для этого годятся персональные компьютеры или параллельные системы, которые могут обладать архитектурой SMP и даже NUMA. Кластеры являются слабосвязанными системами, связи узлов используется одна из стандартных сетевых технологий (Fast/Gigabit Ethernet, Myrinet) на базе шинной архитектуры или коммутатора. Поэтому они являются более дешевой в построении модификацией MPP архитектуры.

# Кластерная архитектура параллельных компьютеров

## Общие принципы

Как уже было сказано раньше вычислительный кластер — это совокупность компьютеров, объединенных в рамках некоторой сети для решения одной задачи (рис.3), которая для пользователя представляется в качестве единого ресурса. Такую концепцию кластера впервые предложила и реализовала в начале 80-х корпорация Digital Equipment, которая и по сей день развивает эту технологию

Понятие "единый ресурс" означает наличие программного обеспечения, дающего возможность пользователям, администраторам и прикладным программам считать, что имеется только одна сущность, с которой они работают - кластер. Например, система пакетной обработки кластера позволяет послать задание на обработку кластеру, а не какому-нибудь отдельному компьютеру. Более сложным примером являются системы баз данных. Практически у всех производителей систем баз данных имеются версии, работающие в параллельном режиме на нескольких машинах кластера. В результате приложения, использующие базу данных, не должны заботиться о том, где выполняется их работа. СУБД отвечает за синхронизацию параллельно выполняемых действий и поддержание целостности базы данных.

Компьютеры, образующие кластер, — так называемые узлы кластера — всегда относительно независимы, что допускает остановку или выключение любого из них для проведения профилактических работ или установки дополнительного оборудования без нарушения работоспособности всего кластера.

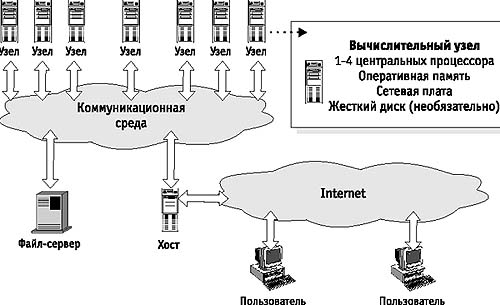


рис.3

В качестве вычислительных узлов в кластере обычно используются однопроцессорные персональные компьютеры, двух- или четырехпроцессорные SMP-серверы. Каждый узел работает под управлением своей копии операционной системы, в качестве которой чаще всего используются стандартные операционные системы: Linux, NT, Solaris и т.п. Состав и мощность узлов может меняться даже в рамках одного кластера, давая возможность создавать неоднородные системы. Выбор конкретной коммуникационной среды определяется многими факторами: особенностями класса решаемых задач, необходимостью последующего расширения кластера и т.п. Возможно включение в конфигурацию специализированных компьютеров, например, файл-сервера, и, как правило, предоставлена возможность удаленного доступа на кластер через Internet.

Из определения архитектуры кластерных систем следует, что она включает в себя очень широкий спектр систем. Рассматривая крайние точки, кластером можно считать как пару ПК, связанных локальной 10-мегабитной сетью Ethernet, так и вычислительную систему, создаваемую в рамках проекта Cplant в Национальной лаборатории Sandia: 1400 рабочих станций на базе процессоров Alpha, связанных высокоскоростной сетью Myrinet.

Таким образом видно, что различных вариантов построения кластеров очень много. При этом в архитектуре кластера большое значение имеют используемые коммуникационные технологии и стандарты. Они во многом определяют круг задач, для решения которых можно использовать кластеры, построенные на основе этих технологий.

## Коммуникационные технологии построения кластеров

Кластеры могут стоится как на основе специализированных высокоскоростных шин передачи данных, так и на основе массовых сетевых технологий. Среди массовых коммуникационных стандартов сейчас чаще всего используется сеть Ethernet или более ее производительный вариант - Fast Ethernet, как правило, на базе коммутаторов. Однако большие накладные расходы на передачу сообщений в рамках Fast Ethernet приводят к серьезным ограничениям на спектр задач, которые можно эффективно решать на таком кластере. Если от кластера требуется большая производительность и универсальность, то необходимо применять более скоростные и специализированные технологии. К ним относятся SCI, Myrinet, cLAN, ServerNet и др. Сравнительная характеристика параметров этих технологий приведена в  
таблице 1.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **SCI** | **Myrinet** | **CLAN** | **ServerNet** | **Fast Ethernet** |
| **Латентность (MPI)** | 5,6 мкс | 17 мкс | 30 мкс | 13 мкс | 170 мкс |
| **Пропускная способность(MPI)** | 80 Мбайт/c | 40 Мбайт/c | 100Мбайт/c | 180 Мбайт/c | 10 Мбайт/c |
| **Пропускная способность (аппаратная)** | 400 Мбайт/c | 160 Мбайт/c | 150 Мбайт/c | н/д | 12,5 Мбайт/c |
| **Реализация MPI** | ScaMPI | HPVM, MPICH-GM и др. | MPI/Pro | MVICH | MPICH |

Таблица 1.

Производительность коммуникационных сетей в кластерных системах определяется несколькими числовыми характеристиками. Основных характеристик две: латентность – время начальной задержки при посылке сообщений и пропускная способность сети, определяющая скорость передачи информации по каналам связи. При этом важны не столько пиковые характеристики, заявленные в стандарте, сколько реальные, достигаемые на уровне пользовательских приложений, например, на уровне MPI-приложений. В частности, после вызова пользователем функции посылки сообщения Send() сообщение последовательно пройдет через целый набор слоев, определяемых особенностями организации программного обеспечения и аппаратуры, прежде, чем покинуть процессор – поэтому существует существенный разбром по стандартам значений латентности. Наличие латентности приводит к тому, что максимальная скорость передачи по сети не может быть достигнута на сообщениях с небольшой длиной.

Скорость передачи данных по сети в рамках технологий Fast Ethernet и Scalable Coherent Interface (SCI) зависит от длины сообщения. Для Fast Ethernet характерна большая величина латентности – 160-180 мкс, в то время как латентность для SCI это величина около 5,6 мкс. Максимальная скорость передачи для этих же технологий 10 Мбайт/c и 80 Мбайт/с соответственно.

# Цели создания кластерных систем

Разработчики архитектур кластерных систем приследовали различные цели при их создании. Первой была фирма Digital Equipment с кластерами VAX/VMS. Целью создания этой машины было повышение надежности работы системы, обеспечение высокой готовности и отказоустойчивости системы. В настоящее время существует множество аналогичных по архитектуре систем от других производителей.

Другой целью создания кластерных систем является создание дешевых высокопроизводительных параллельных вычислительных систем. Один из первых проектов, давший имя целому классу параллельных систем – кластер Beowulf [2] – возник в центре NASA Goddard Space Flight Center для поддержки необходимыми вычислительными ресурсами проекта Earth and Space Sciences. Проект Beowulf начался летом 1994 года, и вскоре был собран 16-процессорный кластер на процессорах Intel 486DX4/100 МГц. На каждом узле было установлено по 16 Мбайт оперативной памяти и по 3 сетевых Ethernet-адаптера. Эта система оказалась очень удачной по отношению цена/производительность, поэтому такую архитектуру стали развивать и широко использовать в других научных организациях и институтах.

Для каждого класса кластеров характерны свои особенности архитекуры и применяемые аппаратные средства. Рассмотрим их более подробно.

# Отказоустойчивые кластеры

## Принципы построения

Для обеспечения надежности и отказоустойчивости вычислительных систем применяется множество различных аппаратурных и программных решений. Например, в системе может дублироваться все подверженные отказам элементы — источники питания, процессоры, оперативная и внешняя память. Такие отказоустойчивые системы с резервированием компонентов применяются для решения задач, в которых недостаточно надежности обычных вычислительных систем, оцениваемой в настоящий момент вероятностью безотказной работы 99%. В таких задачах требуется вероятность 99,999% и выше. Такую надежность можно достичь применяя отличные от приведенного выше методы повышения отказоустойчивости. В зависимости от уровня готовности вычислительной системы к использованию выделяют четыре типа надежности:

|  |  |  |
| --- | --- | --- |
| *Уровень готовности, %* | *Мaкс. время простоя* | *Тип системы* |
| 99,0 | 3,5 дня в год | Обычная (Conventional) |
| 99,9 | 8,5 часов в год | Высокая надежность (High Availability) |
| 99,99 | 1 час в год | Отказоустойчивая (Fault Resilient) |
| 99,999 | 5 минут в год | Безотказная (Fault Tolerant) |

Таблица 2.

В отличие от отказоустойчивых систем с избыточными компонентами, а также различных вариантов многопроцессорности, кластеры объединяют относительно независимые друг от друга машины, каждую из которых можно остановить для профилактики или реконфигурирования, не нарушая при этом работоспособности кластера в целом. Высокая производительность кластера и сведение к минимуму времени простоев приложений достигается благодаря тому, что:

1. в случае сбоя ПО на одном из узлов приложение продолжает функционировать или автоматически перезапускается на других узлах кластера;
2. выход из строя одного из узлов (или нескольких) не приведет к краху всей кластерной системы;
3. профилактические и ремонтные работы, реконфигурацию или смену версий программного обеспечения, как правило, можно осуществлять в узлах кластера поочередно, не прерывая работы других узлов.

Неотъемлемой частью кластера является специальное программное обеспечение, которое, собственно, и решает проблему восстановления узла в случае сбоя, а также решает другие задачи. Кластерное ПО обычно имеет несколько заранее заданных сценариев восстановления работоспособности системы, а также может предоставлять администратору возможности настройки таких сценариев. Восстановление после сбоев может поддерживаться как для узла в целом, так и для отдельных его компонентов — приложений, дисковых томов и т.д. Эта функция автоматически инициируется в случае системного сбоя, а также может быть запущена администратором, если ему, например, необходимо отключить один из узлов для реконфигурации.

Кластеры могут иметь разделяемую память на внешних дисках, как правило, на дисковом массиве RAID. Дисковый массив RAID — это серверная подсистема ввода- вывода для хранения данных большого объема. В массивах RAID значительное число дисков относительно малой емкости используется для хранения крупных объемов данных, а также для обеспечения более высокой надежности и избыточности. Подобный массив воспринимается компьютером как единое логическое устройство.

Восстановление после сбоев может поддерживаться как для узла в целом, так и для отдельных его компонентов — приложений, дисковых томов и т.д. Эта функция автоматически инициируется в случае системного сбоя, а также может быть запущена администратором, если ему, например, необходимо отключить один из узлов для реконфигурации.

Узлы кластера контролируют работоспособность друг друга и обмениваются специфической «кластерной» информацией, например, о конфигурации кластера, а также передавать данные между разделяемыми накопителями и координировать их использование. Контроль работоспособности осуществляется с помощью специального сигнала, который узлы кластера передают друг другу, для того чтобы подтвердить свое нормальное функционирование. Прекращение подачи сигналов с одного из узлов сигнализирует кластерному программному обеспечению о произошедшем сбое и необходимости перераспределить нагрузку на оставшиеся узлы. В качестве примера рассмотрим отказоустойчивый кластер VAX/VMS.

## Кластера VAX/VMS

Компания DEC первой анонсировала концепцию кластерной системы в 1983 году, определив ее как группу объединенных между собой вычислительных машин, представляющих собой единый узел обработки информации. По существу VAX-кластер представляет собой слабосвязанную многомашинную систему с общей внешней памятью, обеспечивающую единый механизм управления и администрирования.

VAX-кластер обладает следующими свойствами:

*Разделение ресурсов.* Компьютеры VAX в кластере могут разделять доступ к общим ленточным и дисковым накопителям. Все компьютеры VAX в кластере могут обращаться к отдельным файлам данных как к локальным.

*Высокая готовность.* Если происходит отказ одного из VAX-компьютеров, задания его пользователей автоматически могут быть перенесены на другой компьютер кластера. Если в системе имеется несколько контроллеров HSC и один из них отказывает, другие контроллеры HSC автоматически подхватывают его работу.

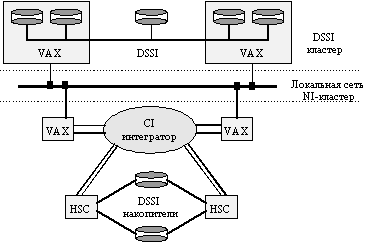
*Высокая пропускная способность*. Ряд прикладных систем могут пользоваться возможностью параллельного выполнения заданий на нескольких компьютерах кластера.

*Удобство обслуживания системы*. Общие базы данных могут обслуживаться с единственного места. Прикладные программы могут инсталлироваться только однажды на общих дисках кластера и разделяться между всеми компьютерами кластера.

*Расширяемость*. Увеличение вычислительной мощности кластера достигается подключением к нему дополнительных VAX-компьютеров. Дополнительные накопители на магнитных дисках и магнитных лентах становятся доступными для всех компьютеров, входящих в кластер.

Работа VAX-кластера определяется двумя главными компонентами. Первым компонентом является высокоскоростной механизм связи, а вторым - системное программное обеспечение, которое обеспечивает клиентам прозрачный доступ к системному сервису. Физически связи внутри кластера реализуются с помощью трех различных шинных технологий с различными характеристиками производительности.

Основные методы связи в VAX-кластере представлены на рис. 4.



*Рис. 4 VAX/VMS-кластер*

Шина связи компьютеров CI (Computer Interconnect) работает со скоростью 70 Мбит/с и используется для соединения компьютеров VAX и контроллеров HSC с помощью коммутатора Star Coupler. Каждая связь CI имеет двойные избыточные линии, две для передачи и две для приема, используя базовую технологию CSMA, которая для устранения коллизий использует специфические для данного узла задержки. Максимальная длина связи CI составляет 45 метров. Звездообразный коммутатор Star Coupler может поддерживать подключение до 32 шин CI, каждая из которых предназначена для подсоединения компьютера VAX или контроллера HSC. Контроллер HSC представляет собой интеллектуальное устройство, которое управляет работой дисковых и ленточных накопителей.

Компьютеры VAX могут объединяться в кластер также посредством локальной сети

Ethernet, используя NI - Network Interconnect (так называемые локальные VAX-кластеры), однако производительность таких систем сравнительно низкая из-за необходимости делить пропускную способность сети Ethernet между компьютерами кластера и другими клиентами сети.

Также кластера могут стоиться на основе шины DSSI (Digital Storage System Interconnect). На шине DSSI могут объединяться до четырех компьютеров VAX нижнего и среднего класса. Каждый компьютер может поддерживать несколько адаптеров DSSI. Отдельная шина DSSI работает со скоростью 4 Мбайт/с (32 Мбит/с) и допускает подсоединение до 8 устройств. Поддерживаются следующие типы устройств: системный адаптер DSSI, дисковый контроллер серии RF и ленточный контроллер серии TF. DSSI ограничивает расстояние между узлами в кластере 25 метрами.

**Системное программное обеспечение VAX-кластеров**

Для гарантии правильного взаимодействия процессоров друг с другом при обращениях к общим ресурсам, таким, например, как диски, компания DEC использует распределенный менеджер блокировок DLM (Distributed Lock Manager). Очень важной функцией DLM является обеспечение когерентного состояния дисковых кэшей для операций ввода/вывода операционной системы и прикладных программ. Например, в приложениях реляционных СУБД DLM несет ответственность за поддержание согласованного состояния между буферами базы данных на различных компьютерах кластера.

Задача поддержания когерентности кэш-памяти ввода/вывода между процессорами в кластере подобна задаче поддержания когерентности кэш-памяти в сильно связанной многопроцессорной системе, построенной на базе некоторой шины. Блоки данных могут одновременно появляться в нескольких кэшах и если один процессор модифицирует одну из этих копий, другие существующие копии не отражают уже текущее состояние блока данных. Концепция захвата блока (владения блоком) является одним из способов управления такими ситуациями. Прежде чем блок может быть модифицирован должно быть обеспечено владение блоком.

Работа с DLM связана со значительными накладными расходами. Накладные расходы в среде VAX/VMS могут быть большими, требующими передачи до шести сообщений по шине CI для одной операции ввода/вывода. Накладные расходы могут достигать величины 20% для каждого процессора в кластере.

# Высокопроизводительные кластеры

## Принципы построения

Архитектура высокопроизводительных кластеров появилась как развитие принципов построения систем MPP на менее производительных и массовых компонентах, управляемых операционной ситемой общего назначения. Кластеры также как и MPP системы состоят из слабосвязанных узлов, которые могут быть как однородными, так и, в отличие от MPP, различными или гетерогенными. Особое внимание при проектировании высокопроизводительной кластерной архутектуры уделяется обеспечению высокой эффективности коммуникационной шины, связывающей узлы кластера. Так как в кластерах нередко применяются массовые относительно низкопроизводительные шины, то приходится принимать ряд мер по исключению их низкой пропускной способности на производительность кластеров и организацию эффективного распараллеливания в кластере. Так например пропускная способность одной из самых высокоскоростных технологий Fast Ethernet на порядки ниже, чем у межсоединений в современных суперкомпьютерах МРР-архитектуры.

Для решения проблем низкой производительности сети применяют несколько методов:

* кластер разделяется на несколько сегментов, в пределах которых узлы соединены высокопроизводительной шиной типа Myrinet, а связь между узлами разных сегментов осуществляется низкопроизводительными сетями типа Ethernet/Fast Ethernet. Это позволяет вместе с сокращением расходов на коммуникационную среду существенно повысить производительность таких кластеров при решении задач с интенсивным обменом данными между процессами.
* применение так называемого «транкинга», т.е. объединение нескольких каналов Fast Ethernet в один общий скоростной канал, соединяющий несколько коммутаторов. Очевидным недостатком такого подхода является «потеря» части портов, задействованных в межсоединении коммутаторов.
* для повышения производительности создаются специальные протоколы обмена информацией по таким сетям, которые позволяют более эффективно использовать пропускную способность каналов и снимают некоторые ограничения накладываемые стандартными протоколами (TCP/IP,IPX). Такой метод часто используют в ситемах класса Beowulf.

Основным качеством, которым должен обладать высокопроизводительный кластер являтся горизонтальная масштабируемость, так как одним из главных преимуществ, которые предоставляет кластерная архитектура является возможность наращивать мощность существующей системы за счет простого добавления новых узлов в систему. Причем увеличение мощности происходит практически пропорционально мощности добавленных ресурсов и может производиться без остановки системы во время ее функционирования. В системах с другой архитектурой (в частности MPP) обычно возможна только вертикальная масштабируемость: добавление памяти, увеличение числа процессоров в многопроцессорных системах или добавление новых адаптеров или дисков. Оно позволяет временно улучшить производительность системы. Однако в системе будет установлено максимальное поддерживаемое количество памяти, процессоров или дисков, системные ресурсы будут исчерпаны, и для увеличеия производительности придется создавать новую систему или существенно перерабатывать старую. Кластерная система также допускает вертикальную масштабируемость. Таким образом, за счет вертикального и горизонтального масштабирования кластерная модель обеспечивает большую гибкость и простоту увеличения производительности систем.