PowerHA 系统构架在生产系统中的运用

《企业级管道完整性管理系统》中高可用体系的建设与运维

SACC2011

演讲人介绍：文平

- 独立技术顾问，从1995年起开始数据库领域、系统领域工作
- 职业历程
  - 程序员…
  - 分析员…
  - 管理员…
  - 数据库优化顾问…
  - 系统优化顾问…
  - DTCC 2011数据库技术大会 演讲嘉宾

- 历年工程实践：应用开发、系统设计、系统构建、系统优化、技术培训

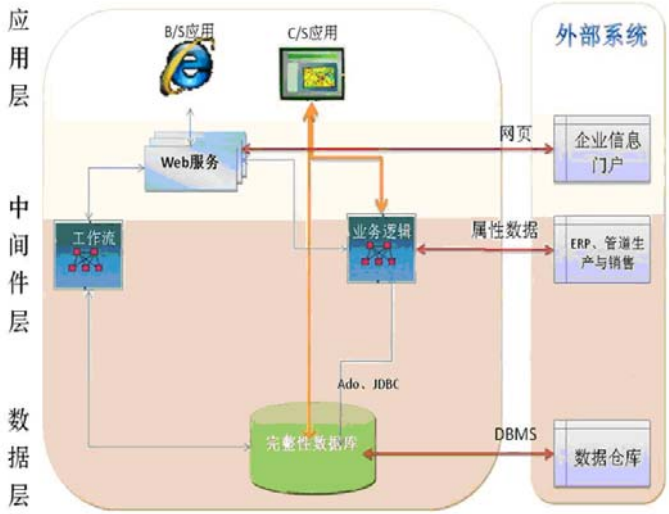
- 历年出版技术专著
  - 《AIX UNIX系统管理•维护与高可用集群建设》 机械工业出版社 2011
  - 《Oracle大型数据库在AIX / UNIX上的实战详解》电子工业出版社 2010
  - 《Sybase数据库在UNIX、Windows上的实施和管理》 电子工业出版社 2009
  - 《PowerBuilder 开发中的数据库设计》 汕头大学出版社 2001
  - 《Oracle 系统开发与管理：iAS配置、管理与开发》汕头大学出版社 1999

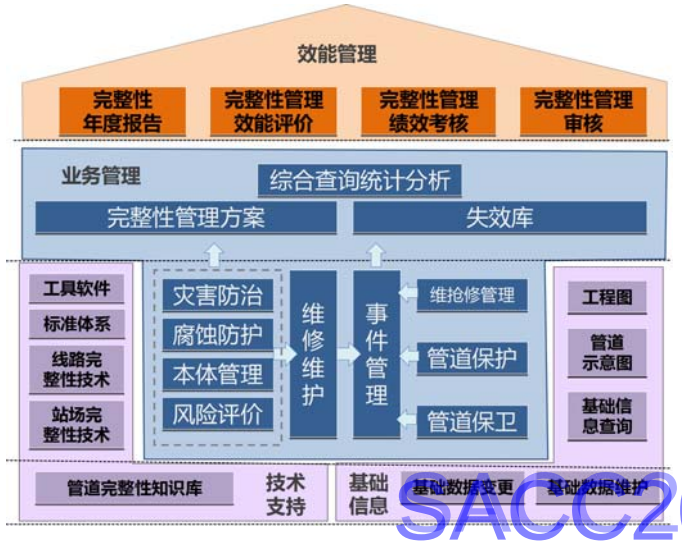
演讲主题

- 项目实施背景、用户需求和构架设计
- PowerHA / HACMP在系统架构中提供的能力
- PowerHA / HACMP的工程应用
- PowerHA / HACMP的后期运维

项目背景：PIS（Pipeline Integrity System）

- 管道完整性管理系统是一个实现管道完整性相关管理的企业级集成信息系统平台；
- 该平台具有对数据采集、高后果区分析、风险评价、完整性评价、维修维护、效能评价等完整性管理的各个环节进行信息化管理的能力，实现管道完整性管理与管道日常管理的有机融合；
- 该系统是保障管道安全运营的必要支撑条件。



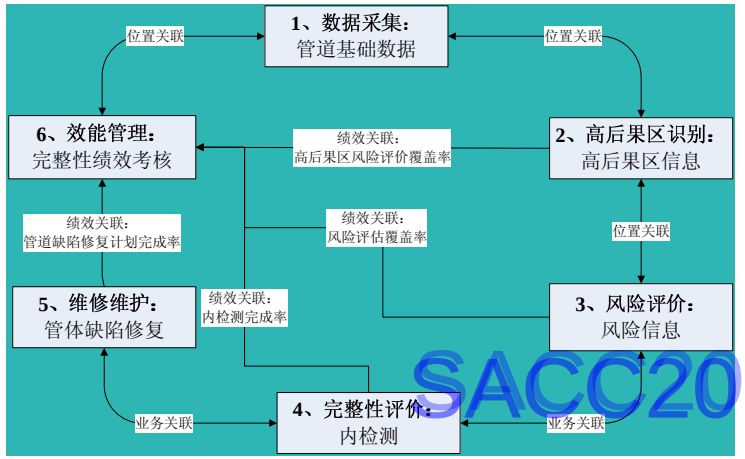


项目背景：PIS（Pipeline Integrity System）

- 管道完整性管理系统是一个实现管道完整性相关管理的企业级集成信息系统平台；
- 该平台具有对数据采集、高后果区分析、风险评价、完整性评价、维修维护、效能评价等完整性管理的各个环节进行信息化管理的能力，实现管道完整性管理与管道日常管理的有机融合；
- 该系统是保障管道安全运营的必要支撑条件。

- 管道业务办理
- 管道完整性数据信息管理
- 管道完整性管理技术集成
- 管道管理决策支持

应用示例：管道业务关联分析



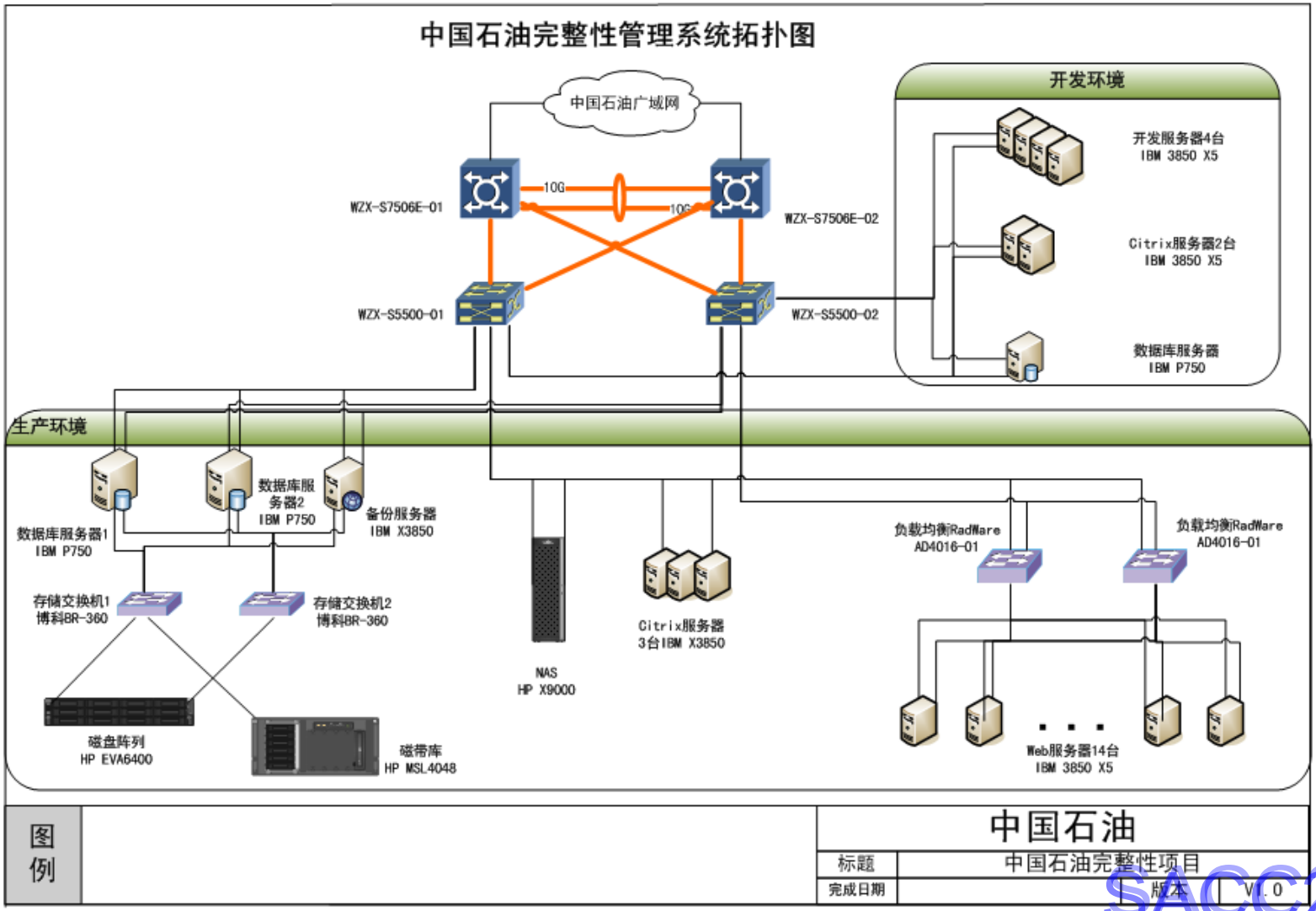
PIS 项目中的服务器可靠性指标

- **2011年4月1日管道完整性管理系统在中石油天然气与管道公司全面上线；**

- **该系统的上线标志着中国石油（天然气与管道公司）辖属的4家地区公司、42家分公司和386个基层站队2700余用户的完整性管理工作已全部纳入信息化管理流程；**

- **该系统的可靠性指标、可用性指标直接影响着西气东输、西部管道、北京天然气等管道业务，已成为中石油管道管理者重要的日常业务平台，如下几点是“你懂得”的：**
  - 7*24系统综合可用性保证是“必须的”
  - 业务数据无错备份和容灾是“必须的”
  - 性能保证乃至高效性运行是“必须的”

PIS 项目的拓扑结构图



PIS 项目中服务器结构定义和选型

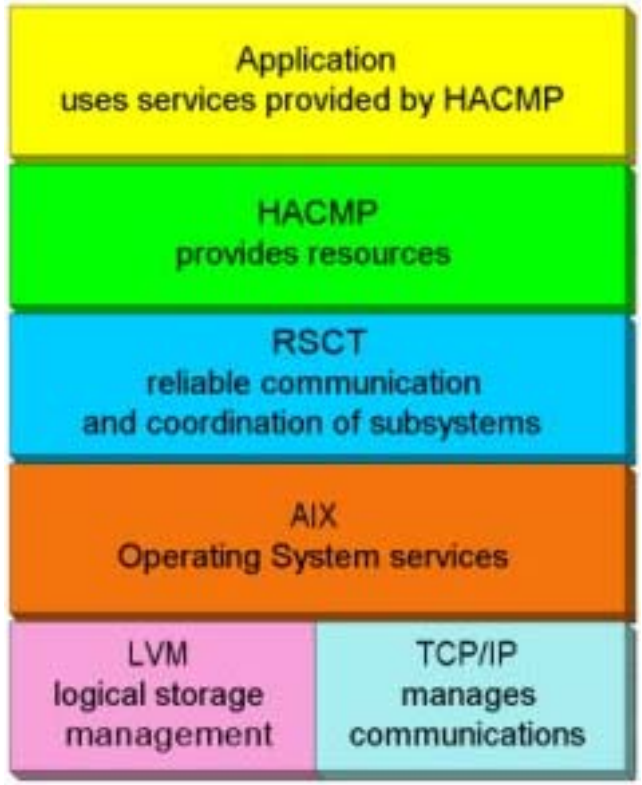
- IBM p系列服务器作为主数据节点
  - P750以其性价比在中高端用户中著称
- 采用PowerHA来作为集群基础结构
  - **PowerHA/HACMP的易用性、可靠性使用户感受深刻**
- 采用Oracle （11.2）RAC来进行数据库集群化
  - 实现均衡负载、实现7*24；
  - 实现计算节点可插拔（P&P：GRID、SCAN、GNS、ASM…）

PowerHA/HACMP在系统结构中的作用

- PowerHA/HACMP是什么？
  - IBM PowerHA 是以前的 IBM High Availability Cluster Multiprocessing (HACMP) 产品的新名称。
  - 改名是为了让 HACMP 与新的 IBM Power Systems Software 计划保持一致。PowerHA for AIX V5.5 是 HACMP 5.4 的后续版本
- 功能一HA：High Availability
  - 系统可用性或运行时间最大化
  - 系统宕机时间最小化
- 功能二CMP：multi-processing:
  - 一个cluster里的各个节点上可以运行多个应用
  - 共享数据或并发访问数据.
- PowerHA/HACMP的目的
  - 消除单点故障(SPOF)，实现高可用

PowerHA/HACMP涉及的模块

- **Application**
  - 高可用系统的服务目标：Oracle、Sybase…
- **PowerHA/HACMP**
  - 为应用提供高可用服务
  - 集群各节点间资源运行的协同
- **RSCT**
  - 节点间通信管理
  - 子系统间的协同
- **AIX**
  - 提供操作系统服务
- **LVM**
  - 提供逻辑卷管理服务
- **TCP/IP**
  - 逻辑层面的通信管理



PowerHA/HACMP和Oracle RAC的关联

- **PowerHA可以提供的集群基本服务**
  - 共享卷组
  - 共享逻辑卷
  - 节点监测
- **与Oracle11gR2可以配合使用**
  - 用于Grid的基础环境
  - 提供VotingDisk、OCR的存储
  - 独立于数据存储区域（ASM）

在传统的RAC部署中，HACMP起到的作用是久经考验的！

未选用PowerHA/HACMP做容灾的原因

- 不能提供无间断服务保护
  - 节点间的失败转移需要必要的时间，哪怕是几十秒钟
- 对结构不稳定的系统不适用
  - 集群，各组件成网状关联，牵一发而动全身
  - 集群成员的变化会导致整体拓扑的变化
- 在失败恢复中极有可能需要人工的介入：
  - 设备调整、服务状态调整… …
  - PowerHA/HACMP不是容错系统

未选用PowerHA/HACMP做容灾的原因

- **应用必须能够容忍资源转移中的停、启：**
  - 节点停止是需要关闭资源的使用：
    - 可能导致内存丢失
    - 可能导致进程状态信息丢失（会话信息丢失）
  - 节点接管需要重新启动资源
    - 应用需要从可能的写失败中恢复

演讲主题

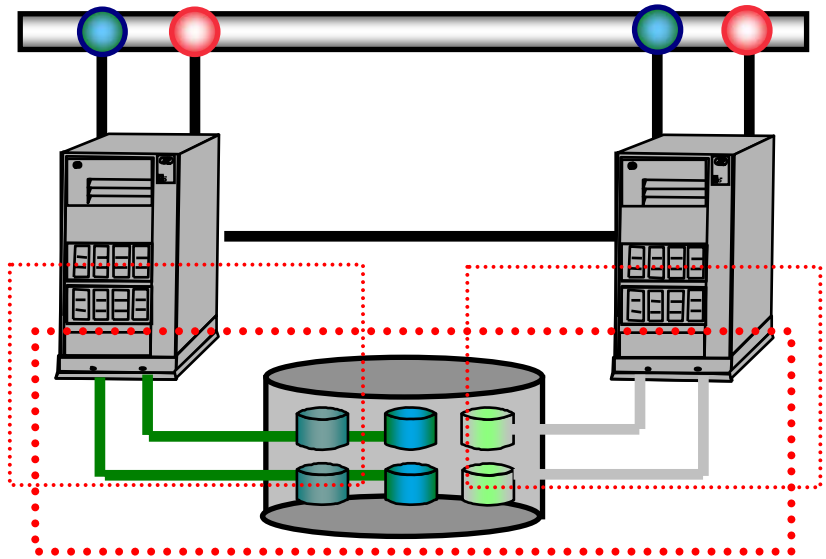
- 项目实施背景、用户需求和构架设计
- PowerHA / HACMP在系统架构中提供的能力
- 中石油对PowerHA / HACMP的工程应用
- 中石油对PowerHA / HACMP的后期运维

PowerHA / HACMP集群组件：集群节点

- 服务器被分区化，成为逻辑服务器
- 多节点（分区）集成，合为一个集群
- 在主备模式下，其中的一个节点作为备用节点
- 在并行模式下，所有节点都可以作为应用节点
- 如果是备用节点，则可以选择低资源分配分区
  - 理论上这个备用点需要能够承载所有主节点的载荷
  - 但事实上其他主节点的失败该路很低
- 缺点是系统“有点复杂”
- 缺点是服务器本身会成为SPoF薄弱点

PowerHA / HACMP集群组件：共享存储

- 支持两种方式的共享:
  - 多节点并发共享
  - 多节点独占共享
- 共享情况下卷组可以迁移
- 应用数据放置在共享存储上，在需要的情况下迁移；
- 应用代码（二进制文件）的放置是不受限制的。

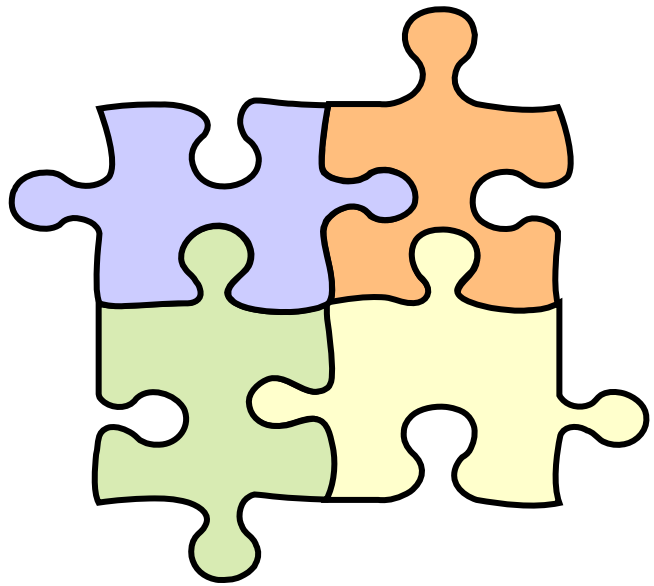


PowerHA / HACMP集群组件：应用服务

- 应用服务器，是集群的支撑目标
- **PowerHA/HACMP就是为应用服务而存在的，应用服务会反应为几个Shell脚本:**
  - 启动应用服务
  - 关闭应用服务
  - 监控应用服务 (可选)
  - 重置应用服务 (可选)
- 应用服务应能够从某种未知状态下启动
- 应用服务应能够从某种未知状态下关闭

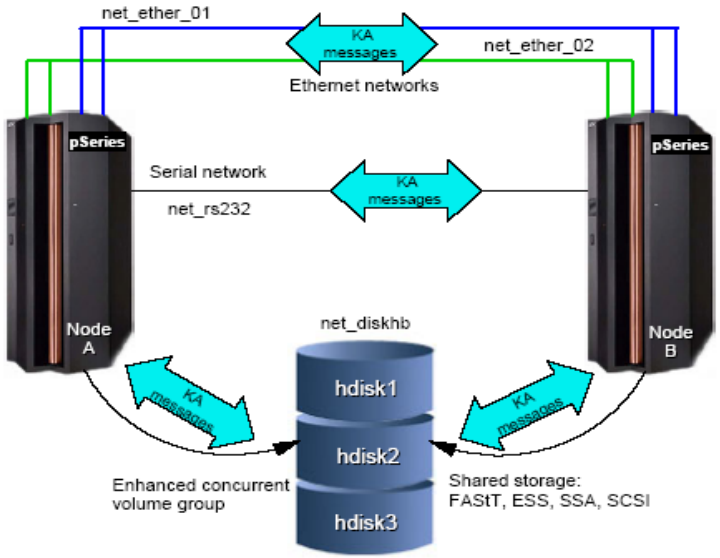
PowerHA / HACMP集群组件：资源组

- 为应用运行而需准备的若干逻辑资源组
- 这个资源组由**PowerHA/HACMP**集中管理，可以在各个节点间漂移
- 这个资源组在集群中的某些指定节点上部署
  - 存在节点间的部署优先级问题
  - 存在本地优先级指定问题
- 资源组可能会在节点间漂移:
  - 在某集群节点启动
  - 失败转移到某个集群节点
  - 节点重聚后资源组的回退
  - 其他动态分配策略



PowerHA / HACMP集群组件：心跳选择

- 由网卡构成的心跳环路
- 由磁盘构成的心跳环路 (Fibre Channel）
- 由串行网络构成的心跳环路



PowerHA / HACMP集群部署：操作系统

- 一般来说，在一个**cluster**中，涉及到的应用软件版本一致，这样易于管理；

- **PowerHA/HACMP**对应用软件没有严格的限制，用户可以根据实际需求选择需要加入**cluster**的应用软件：
  - 本项目采用了Oracle10、Oracle11等若干版本
  - 本项目采用了RAC、DataGuard、Stream等若干模块

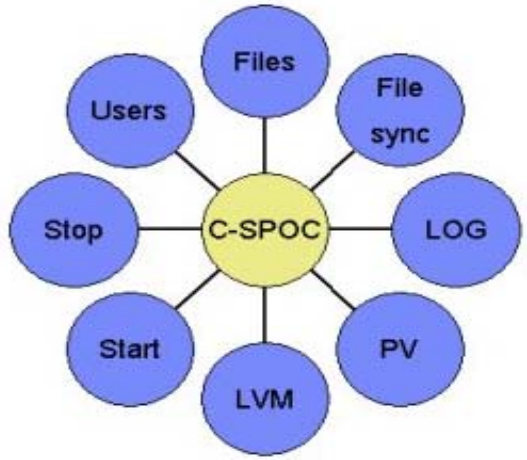
- 用户需要自己的脚本来管理应用服务

PowerHA / HACMP集群管理：Smitty 环境

```
                        Add a Resource Group (extended)

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

                                                  [Entry Fields]
* Resource Group Name                            [rsg2]
* Participating Nodes (Default Node Priority)    [lpar1 lpar2]                +

  Startup Policy                                  Online On All Available N> +
  Fallover Policy                                 Fallover To Next Priority> +
  Fallback Policy                                 Never Fallback             +

F1=Help          F2=Refresh       F3=Cancel        F4=List
Esc+5=Reset      Esc+6=Command    Esc+7=Edit       Esc+8=Image
Esc+9=Shell      Esc+0=Exit       Enter=Do
```

PowerHA/HACMP集群管理：C-SPOC

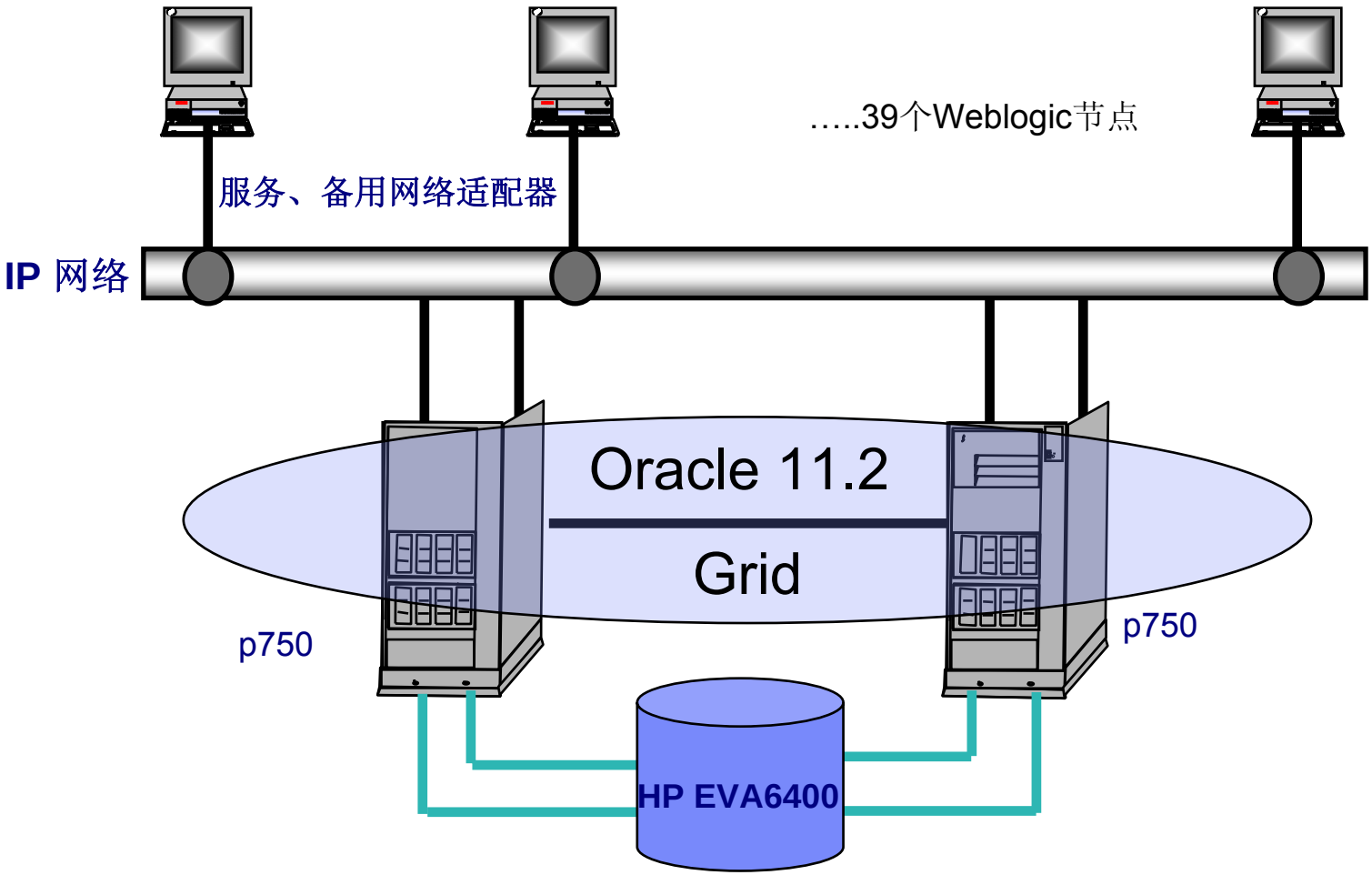
- **C-SPOC：Cluster-Single Point of Control， HACMP的管理工具**
- 集群内单点管理全部：
  - users
  - file systems
  - logical volumes
  - physical volumes
  - Start、stop
  - log files
  - Resource Group State
  - …



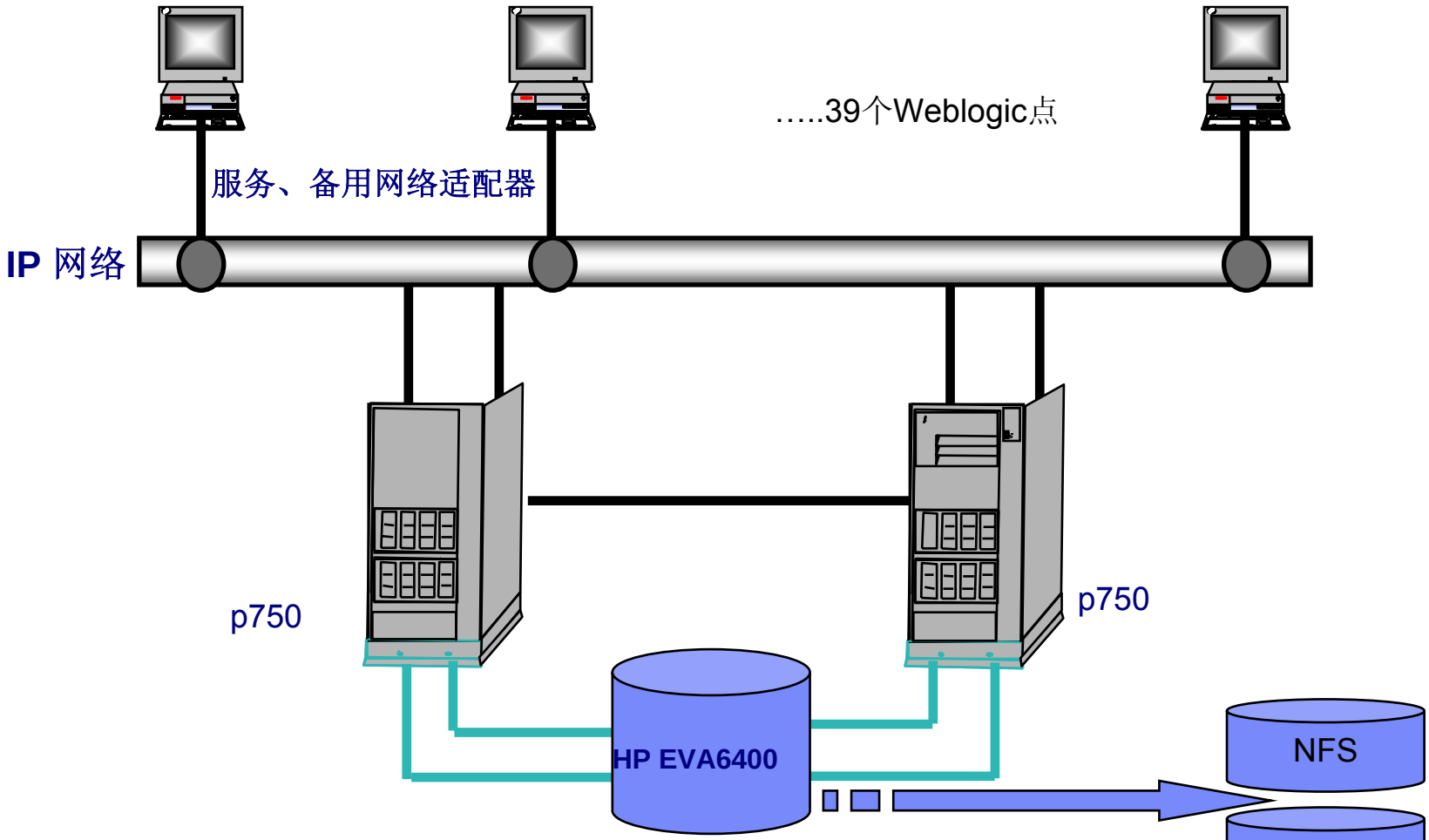
演讲主题

- 项目实施背景、用户需求和构架设计
- PowerHA / HACMP在系统架构中提供的能力
- 中石油对PowerHA / HACMP的工程应用
- 中石油对PowerHA / HACMP的后期运维

中石油PowerHA现场应用：存储共享访问



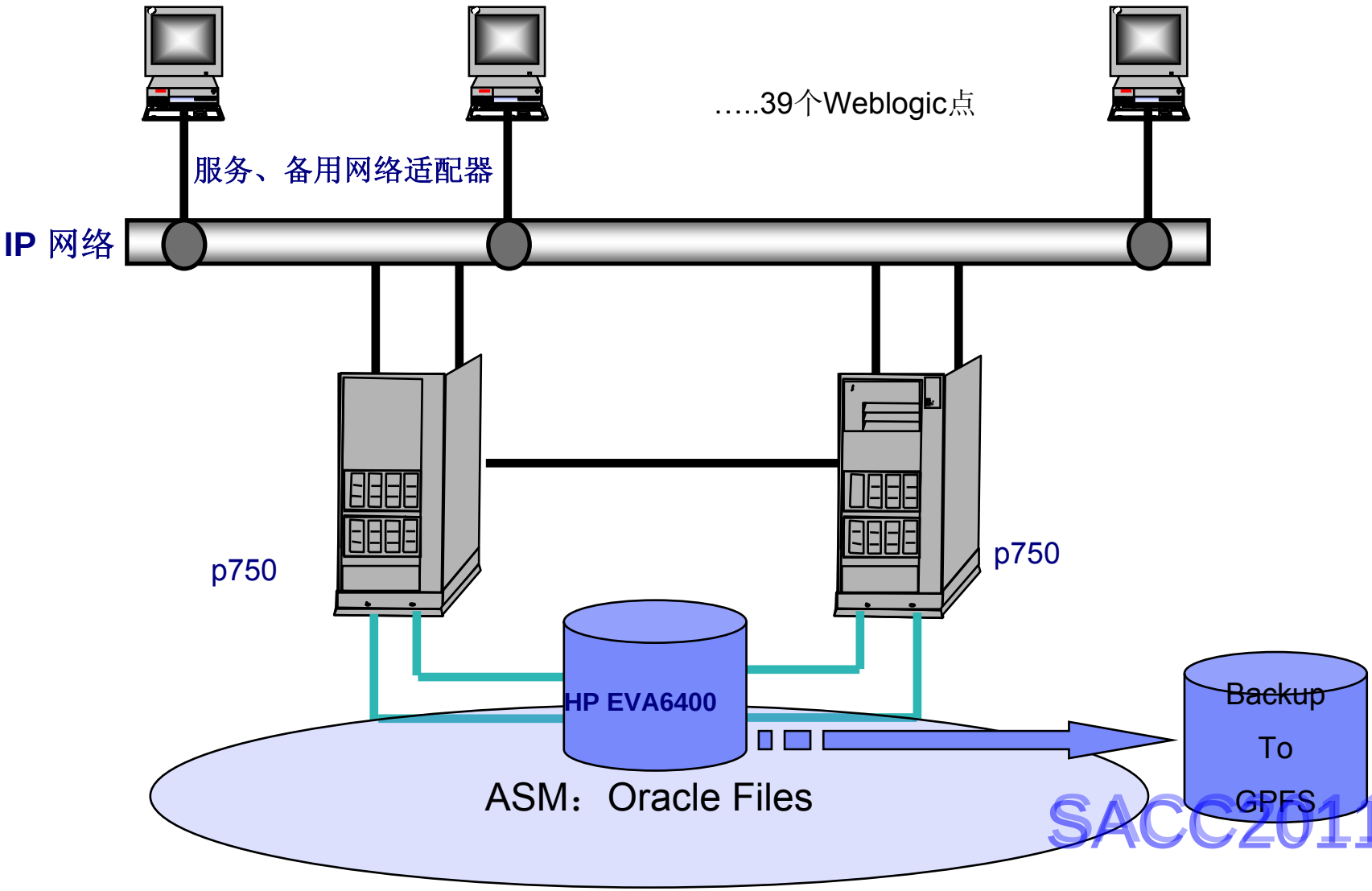
PowerHA现场应用：结合NFS、GPFS



GPFS： Oracle ArchiveLogs

Oracle Backup Sets

PowerHA现场应用：数据库存放

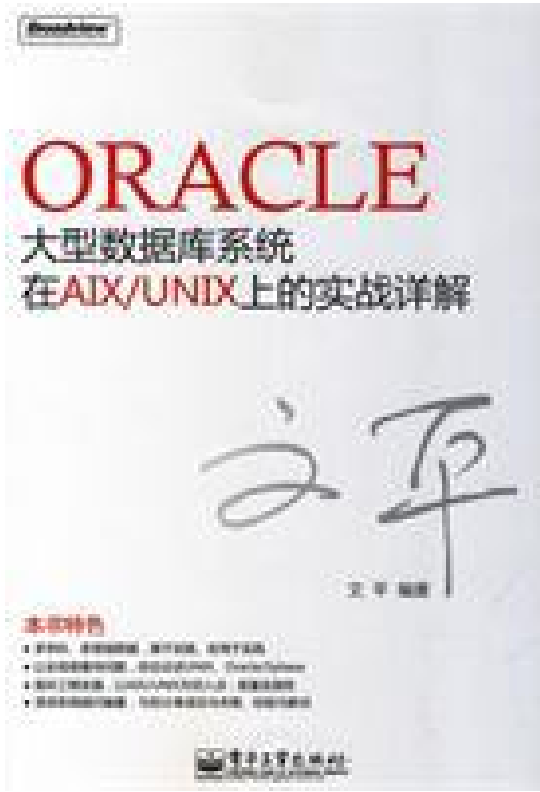


PowerHA现场应用：配置过程

- HACMP配置前的准备工作
  - 配置IP地址
  - 编辑/etc/hosts文件
  - 创建vg和文件系统
  - 准备串口设备及磁盘心跳设备
- HACMP的Standard配置过程
  - 添加Cluster和节点
  - 配置Cluster资源
  - 创建Cluster资源组
  - 同步HACMP的配置
- HACMP的Extended配置过程
  - 添加心跳
  - 定制Cluster资源
- Grid安装与准备
  - Shell Limit调整
  - Kernel Parameter调整
  - Users & Groups 创建
  - Disks 设定
  - Grid 安装
  - DNS、VIP、SCAN指定
- Oracle安装与准备
  - 安装集群数据库
- 数据库创建和客户连接
  - DBCA
  - SCAN

PowerHA现场应用：过程参考

- 有关Oracle RAC在 AIX PowerHA上的部署，请参见如下技术专著：
  - Oracle大型数据库系统在AIX/UNIX上的实战详解



演讲主题

- 项目实施背景、用户需求和构架设计
- PowerHA / HACMP在系统架构中提供的能力
- 中石油对PowerHA / HACMP的工程应用
- 中石油对PowerHA / HACMP的后期运维

PowerHA/HACMP运行维护中的状态监控

- 监控 **Cluster Log Files**
- 监控 **Cluster Daemons**
- 监控 **Cluster:**
  - clstat/xclstat
  - check log files
  - check daemons by lssrc -g cluster or ps -ef
  - lsvg -o
  - ifconfig -a
  - netstat -in
  - lslpp -l cluster.*
- 应用修补（**patches**）

PowerHA/HACMP运行维护中的状态监控

- **/tmp/clstrmgr.debug**
  - **clstrmgrES 守护进程产生的带有时间信息的日志**
- **/tmp/cspoc.log**
  - HACMP C-SPOC 命令执行相关的带有时间信息的日志
- **/tmp/emuhacmp.out**
  - HACMP Event 生成的带有时间信息的日志，并会被汇总到 **/tmp/emuhacmp.out**日志.
- **/tmp/hacmp.out**
  - HACMP 内部脚本生成的带有时间信息的日志。这是HACMP的主日志。
- **/usr/es/adm/cluster.log**
  - HACMP脚本和守护进程生成的带有时间信息的日志。该日志带有集群的状态信息，可用于分析集群一级的问题。

PowerHA/HACMP运行维护中的性能相关调整

- **I/O pacing**
  - 每当系统内有其它应用在做大量的I/O操作时，用户可能会碰到如交互性能受到严重影响等问题，能够通过调整系统的I/O pacing，以使系统在大量的磁盘读写操作期间的资源分配更加均衡。
    - 可以使用smitty chgsys 去设置I/O pacing 到high-water 和low-water，缺省值为“0”（disable I/O pacing），一般情况下high-water 设置为“33”而low-water设置为“24”。
- **syncd 频率**
  - 编辑/sbin/rc.boot 文件去增加syncd 频率，可以从缺省的60 秒到30、20、10 秒。
    - 增加此频率可在繁重的I/O传输期间促使更频繁的 I/O flush 和减少触发deadman switch的可能性。

PowerHA/HACMP运行维护中的性能相关调整

- 文件系统缓存设置
  - 当系统存在大量的文件操作时，系统可能面临文件系统缓存使用过量的问题。minperm 和 maxperm 是两个最基本的分页替换可调参数。这两个可调参数用于指出 AIX 内核应该使用多少内存来缓存非计算性的分页。maxperm 可调参数指出应该用于缓存非计算性分页的最大内存量。
    - 可以使用AIX的vmo参数minperm、maxperm等参数限定缓存的使用。
- 换页操作的倾向性设置
  - 当 numperm 在 minperm 和 maxperm 之间的时候，如果 lru_file_repage 可调参数设置为 1，那么 AIX 分页替换守护进程将根据其内部重新分页表来确定选择何种类型的分页进行操作。

```
# vmo -p -o minperm%=3
# vmo -p -o maxperm%=90
# vmo -p -o maxclient%=90
# vmo -p -o lru_file_repage=0
# vmo -p -o strict_maxclient=1
# vmo -p -o strict_maxperm=0
```

```
-/etc/security/limits
default:
fsize = -1
data = -1
core = -1
fsize_hard = -1
cpu_hard = -1
data_hard = -1
stack_hard = 65536
core_hard = -1Shell
```

PowerHA/HACMP运行维护中的性能相关调整

- 网络参数的相关性设置

udp_sendspace	65536
udp_recvspace	655360(udp_sendspace* 10)
tcp_sendspace	65536
tcp_recvspace	65536
rfc1323	1
sb_max	4194304
ipqmaxlen	512

- 改变故障检测速率

设置 I/O pacing、延长syncd频率

改变故障检测速率到–“slow”，延长心跳时间间隔。

PowerHA/HACMP运行维护中的问题与解决

- **HACMP运行和Oracle有关系吗？可能由于AIX配置不当造成HACMP的故障，节点间出现错误失败转移，见下例：**

```
#iostat:

Disks:        % tm_act     Kbps      tps    Kb_read   Kb_wrtn
hdisk0          100.0    934.0    199.5       1576       292
hdisk1          100.0   1046.0    228.5       1800       292

.....................
#vmo -p -o lru_file_repage=0 -o maxclient%=20 -o maxperm%=20 -o minperm%=5
# sar -d 5 2

AIX p570 3 5 00C80E6D4C00    06/09/11
System configuration: lcpu=8 drives=8
15:57:00     device     %busy     avque     r+w/s     Kbs/s     avwait     avserv
15:57:05     hdisk0      21        0.0        49       405        0.1        4.9
             hdisk1       2        0.0         4        17        0.0        5.6

 .....................
```

- （注：**AIX5.3、6.1等版本可采用lru_file_repage参数解决上述问题**）

PowerHA/HACMP运行维护中的问题与解决

- **HACMP造成了系统多项任务的错误，通过设置AIX的IO调步、心跳频率设置等方式，有效降低了错误发生的频率，见下例：**

864D2CE3 0226235311 P S topsvcs NIM thread blocked

```
                    Change / Show Characteristics of Operating System

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

[TOP]                                                [Entry Fields]
System ID                                           0X80000CD4FC500000
Partition ID                                        0X80000CD4FC500001
Maximum number of PROCESSES allowed per user        [2048]                +#
Maximum number of pages in block I/O BUFFER CACHE   [20]                  +#
Maximum Kbytes of real memory allowed for MBUFS     [0]                   +#
Automatically REBOOT system after a crash           true                  +
Continuously maintain DISK I/O history              false                 +
HIGH water mark for pending write I/Os per file     [33]                  +#
LOW water mark for pending write I/Os per file      [24]                  +#
Amount of usable physical memory in Kbytes          14090240
State of system keylock at boot time                normal
Enable full CORE dump                               false                 +
Use pre-430 style CORE dump                         false                 +
Pre-520 tuning compatibility mode                   disable               +
Maximum login name length at boot time              [9]                   +#
Stack Execution Disable (SED) Mode                  select                +
NFS4 ACL Compatibility Mode                         secure                +
ARG/ENV list size in 4K byte blocks                 [6]                   +#
CPU Guard                                           enable                +
Processor capacity increment                        1.00
Partition is capped                                 true
Partition is dedicated                              true
[MORE...4]

F1=Help          F2=Refresh       F3=Cancel        F4=List
Esc+5=Reset      Esc+6=Command    Esc+7=Edit       Esc+8=Image
Esc+9=Shell      Esc+0=Exit       Enter=Do
```

3C81E43F 1204235610 P U topsvcs Late in sending heartbeat

3C81E43F 1127235710 P U topsvcs Late in sending heartbeat……

项目总结：PIS已进入稳定运行阶段

- 该系统的可靠性直接影响着西气东输、西部管道、北京天然气等油气业务，是中石油最重要的业务平台之一；
- 2011年4月1日管道完整性管理系统在中石油天然气与管道公司全面上线，并已完成：
  - 7*24的系统运行状态保证
  - 数据采用了最成熟的存储技术
  - 数据备份实现了本地化、异地化
  - 数据备份具有实时性
  - 数据备份具有完备的检测手段
  - 系统具有一定的容灾能力
  - 系统在高效状态下运行
  - 系统在有效的管理框架下运行

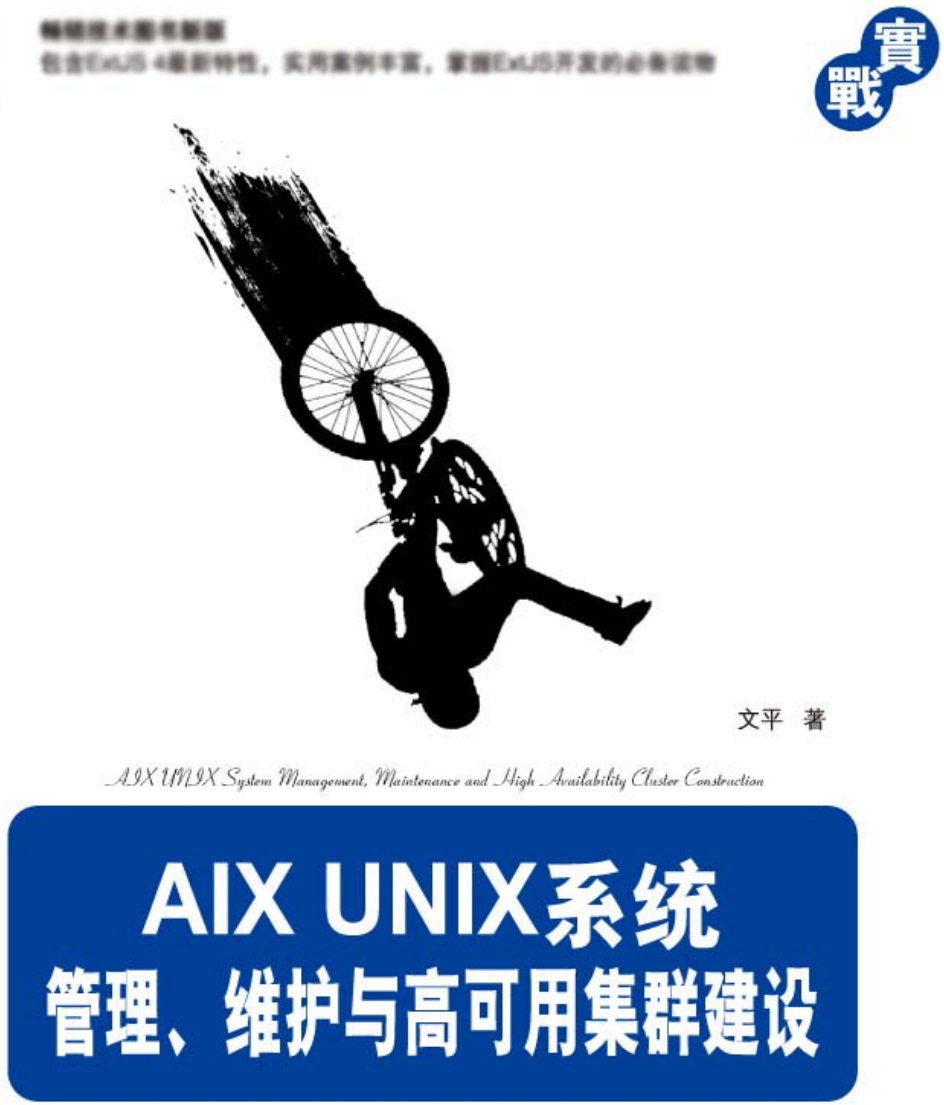
参考：

- 企业级管道完整性管理系统的研究
  - 周利剑、李祎、余海冲、杨宝龙，中国石油管道科技研究中心
- 企业级完整性管理平台的建设及应用
  - 周利剑、郭磊、余海冲、李祎、贾韶辉，中国石油管道科技研究中心
- 管道完整性管理系统部署与优化方案
  - 文平

谢谢大家！

更为详细的内容，请参见即将出版的**AIX**技术专著：

AIX UNIX系统管理、维护与高可用集群建设



SACC2011

谢谢大家！

Q & A