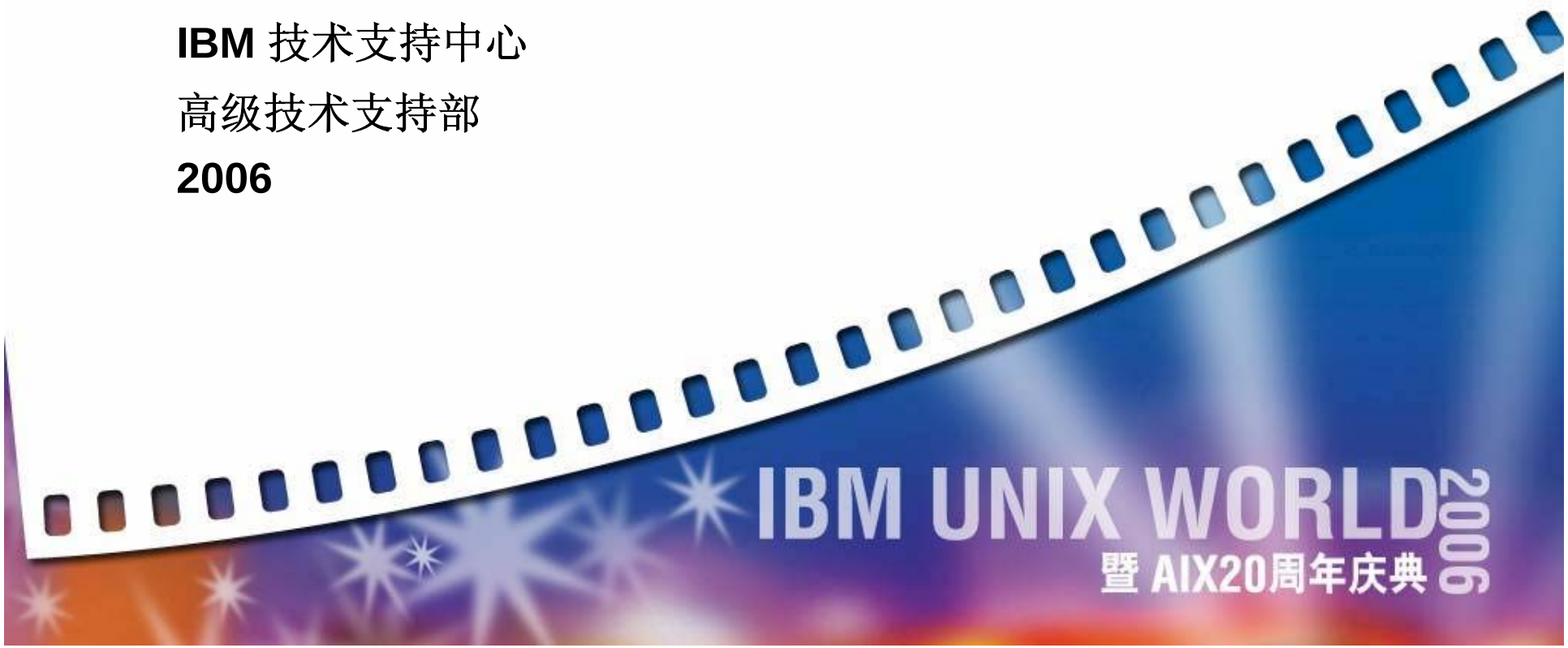
AIX系统环境下的
SAN存储规划、实施与配置的经验分享

黄大川

IBM 技术支持中心

高级技术支持部

2006



内容

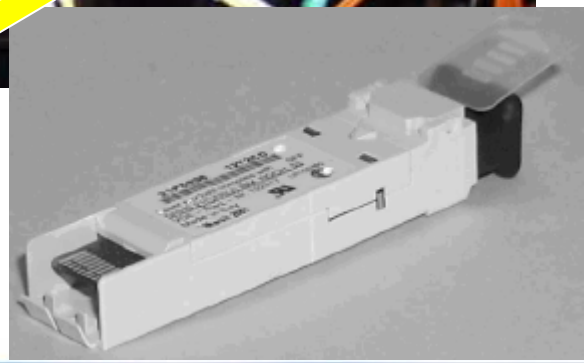
- （一）光纤存储网
- （二） **DS4000**
- （三） **AIX**
- （四）性能优化

（一）光纤存储网-内容

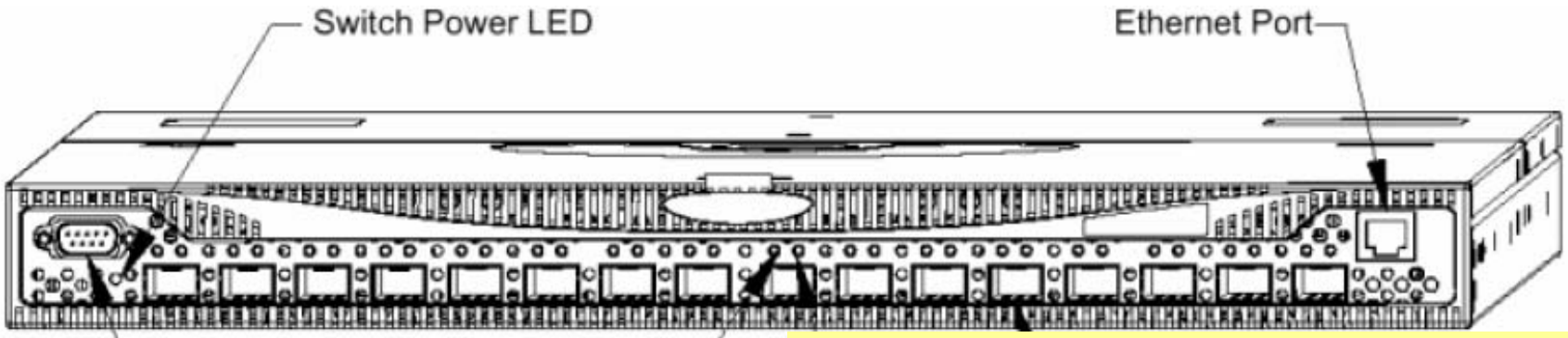
- 硬件安装
- 操作界面
- 集联组网
- 划分**zone**

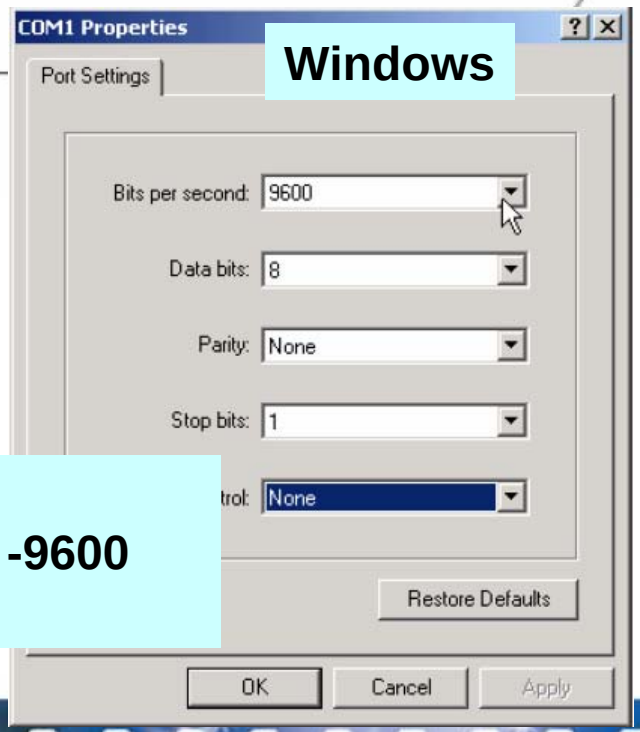
光纤存储网-硬件安装

网口，用于配置SAN



光纤存储网-交换机初始化-设置IP：ipAddrSet





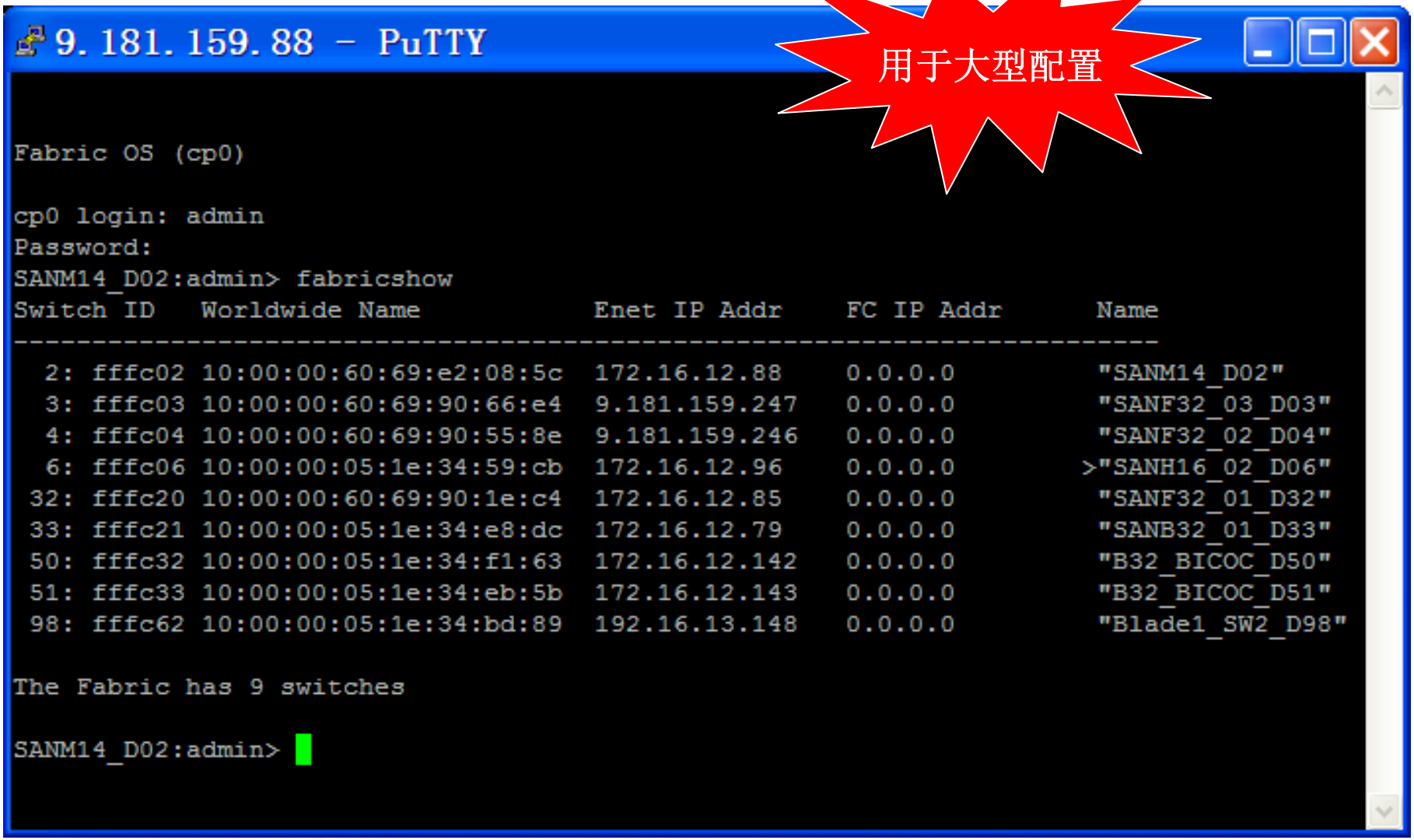
Windows

Unix
tip /dev/ttyb -9600

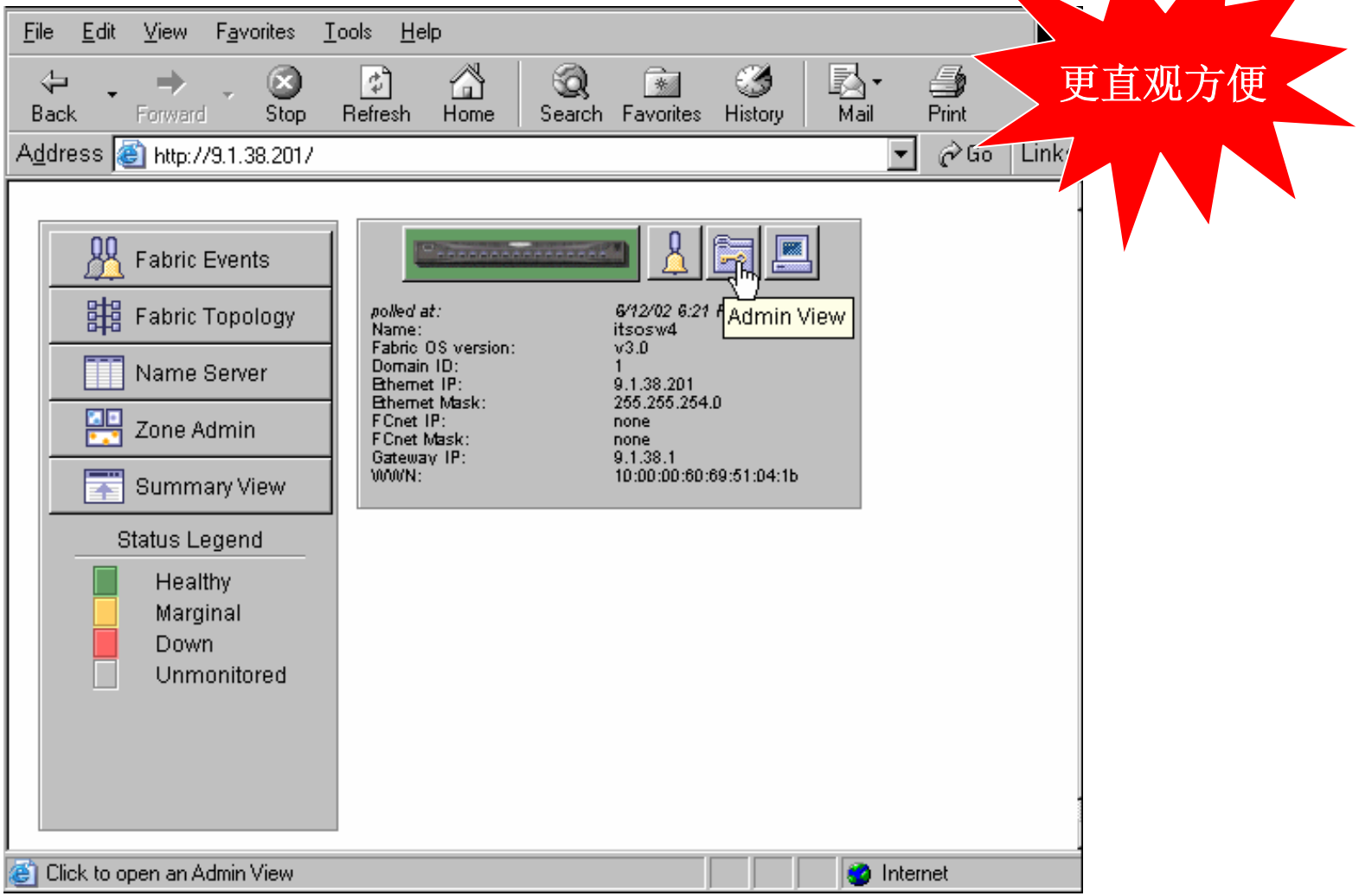
ID:admin/口令:password
ipAddrSet

Ethernet IP Address [10.77.77.77]:
Enter new ethernet IP address:
Ethernet Subnetmask [255.255.254.0]:
Enter new ethernet subnetmask:
Fibre Channel IP Address [none]:
Enter new Fibre Channel IP address if desired:
Fibre Channel Subnet Mask [none]:
Enter new Fibre Channel subnet mask if desired:
Gateway Address [none]:
Enter new gateway address:
Set IP address now? [y = set now, n = next reboot]:
Enter “y” to set now:

光纤存储网-操作界面-CLI(ID:admin/口令:password)

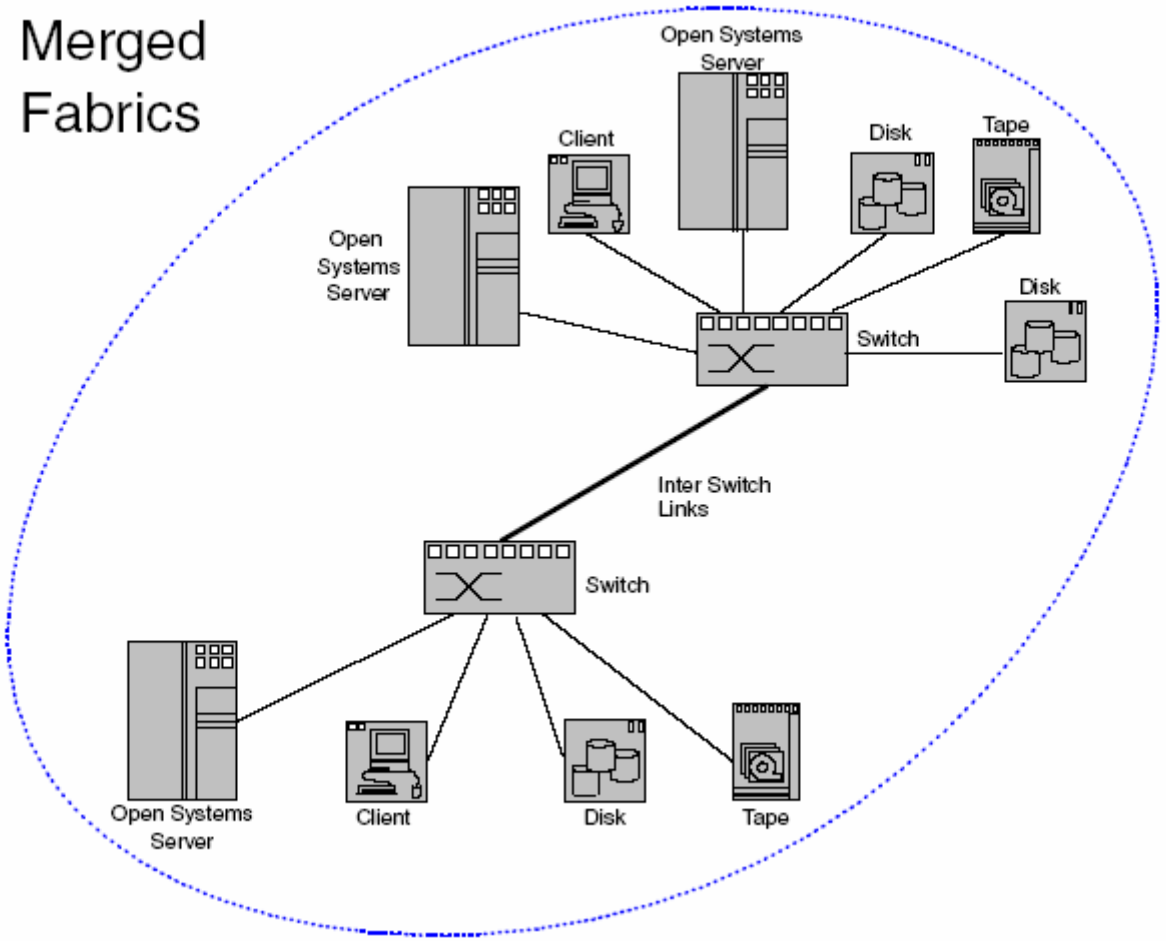


光纤存储网-操作界面-web界面



更直观方便

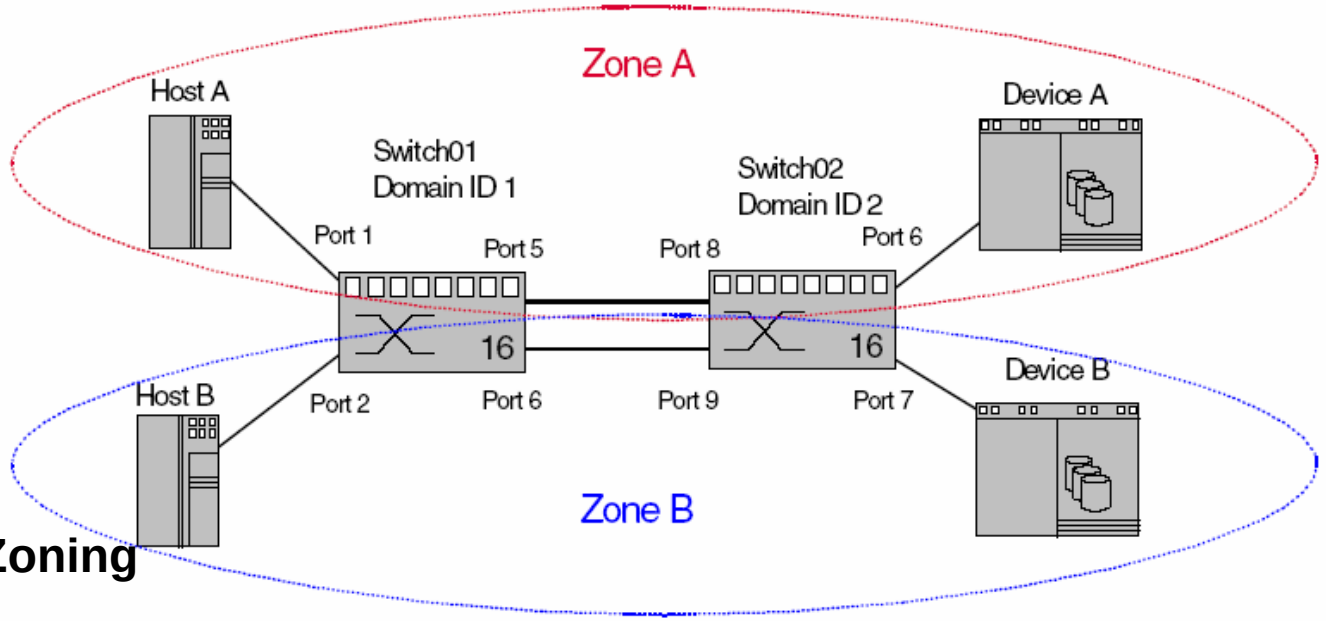
光纤存储网-集联组网


Merged
Fabrics
Open Systems
Server
Client
Disk
Tape
Open
Systems
Server
Disk
Switch
Inter Switch
Links
Switch
Open Systems
Server
Client
Disk
Tape

光纤存储网-划分zone

Zone概念：

•一种虚拟区域的概念



Zone分类：

•**Switch/Port Level Zoning**

•**WWN Level Zoning**

•**AL_PA Level Zoning**

•**Mixed Level Zoning**

光纤存储网-端口zone举例

```
zonecreate "z1","1,0;1,1;1,2;1,3"
zonecreate "z2","1,4;1,5;1,6;1,7"
zonecreate "z3","1,8;1,9;1,10;1,11"
zonecreate "z4","1,12;1,13;1,14;1,15"
cfgcreate "cfgbase","z1;z2;z3;z4"
cfgsave
cfgenable "cfgbase“
```

适用于：设备经常更换的场合

可以随意更换设备，特别是在更换光纤卡的情况下，不需要修改**SAN**配置，发挥优势…

光纤存储网-wwn zone举例

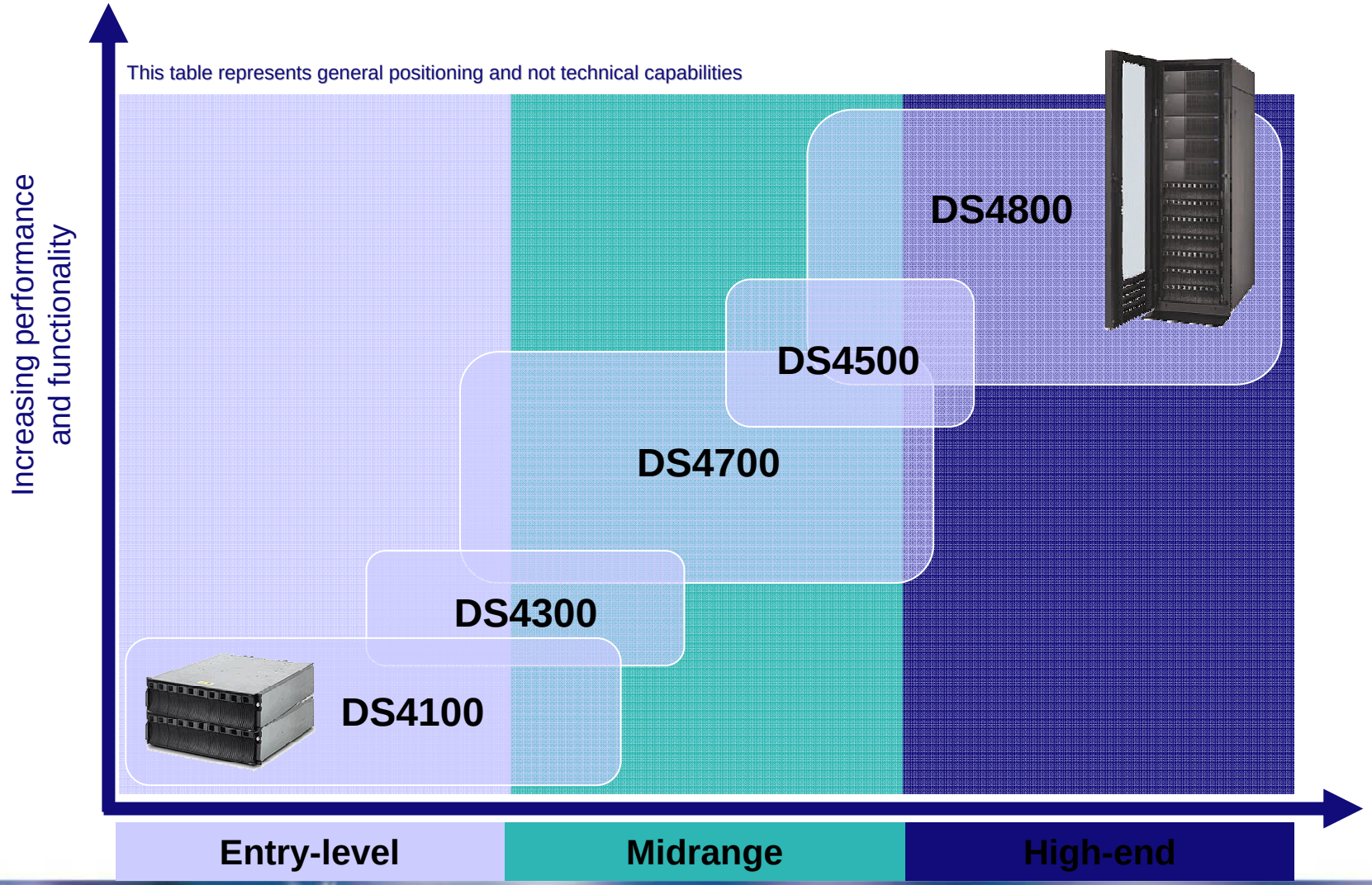
```
aliCreate "p570c_1","10:00:00:00:C9:46:08:96"
aliCreate "p570c_2","10:00:00:00:C9:46:11:F1"
aliCreate "p570d_1","10:00:00:00:C9:44:EA:AD"
aliCreate "p570d_2","10:00:00:00:C9:45:06:6E"

aliCreate "F600_1_A","20:06:00:a0:b8:16:7d:1d"
aliCreate "F600_1_B","20:07:00:a0:b8:16:7d:1d"
aliCreate "F600_2_A","20:06:00:a0:b8:13:a0:8d"
aliCreate "F600_2_B","20:07:00:a0:b8:13:a0:8d"

zoneCreate "z1_c_5","p570c_1;F600_1_A"
zoneCreate "z1_c_6","p570c_2;F600_1_B"
zoneCreate "z1_c_7","p570d_1;F600_1_A"
zoneCreate "z1_c_8","p570d_2;F600_1_B"

zoneCreate "z2_c_5","p570c_1;F600_2_A"
zoneCreate "z2_c_6","p570c_2;F600_2_B"
zoneCreate "z2_c_7","p570d_1;F600_2_A"
zoneCreate "z2_c_8","p570d_2;F600_2_B"

cfgadd "cfg0603","z1_c_5;z1_c_6;z1_c_7;z1_c_8;z2_c_5;z2_c_6;z2_c_7;z2_c_8"
cfgsave
cfgenable "cfg0603"
```

适用于：设备位置经常更换的场合特别是在将设备从一个物理位置移动到其他物理位置的时候，不需要修改**SAN**配置，发挥优势…

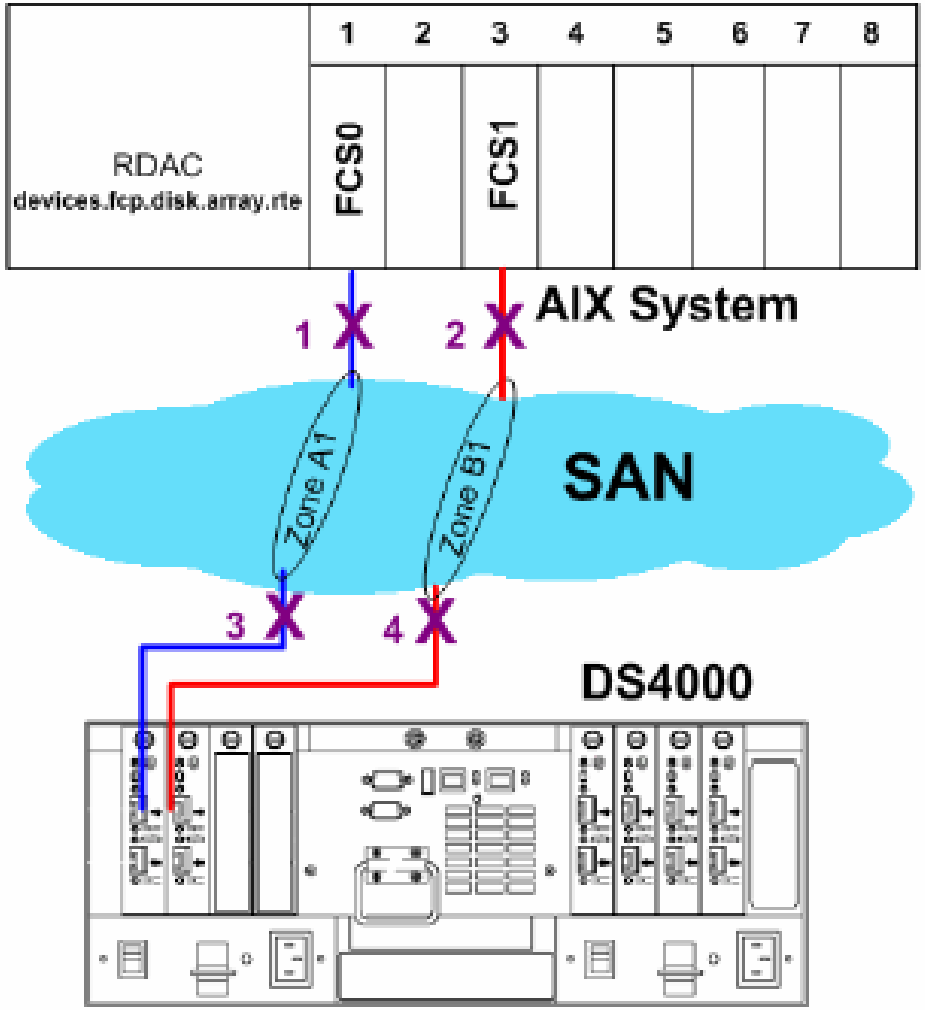
（二）DS4000-内容

- 硬件连接
- 划分盘
- 配置AIX主机

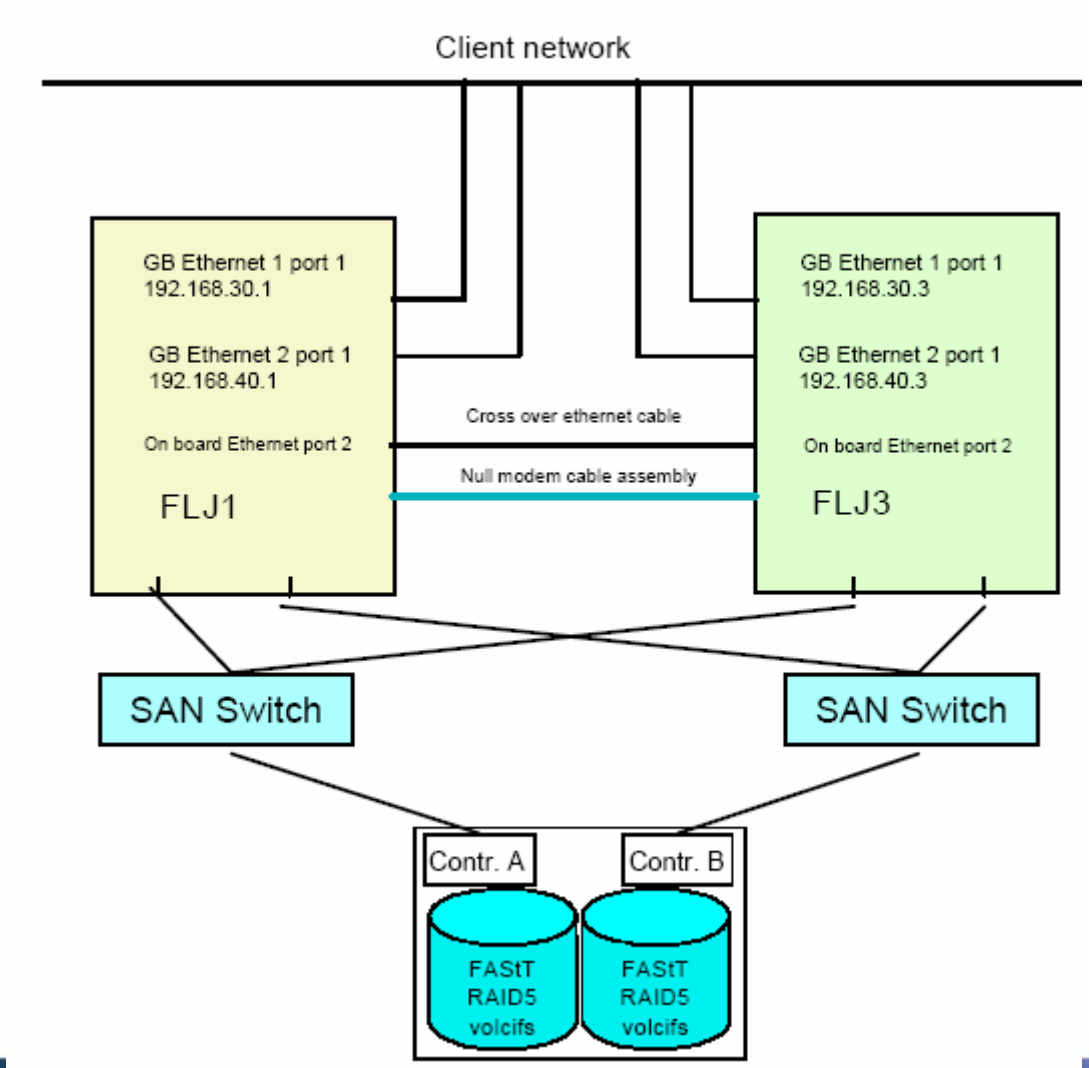
DS4000 Series Positioning



DS4000-硬件连接

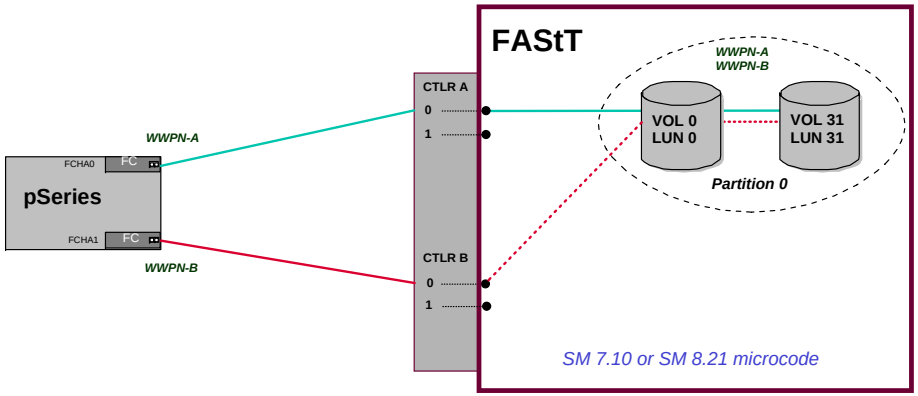

1
2
3
4
5
6
7
8
RDAC
devices.fcp.disk.array.rte
FCS0
FCS1
AIX System
1
2
Zone A1
Zone B1
SAN
3
4
DS4000

AIX连接DS4000-硬件连接- 推荐的HACMP连接方式

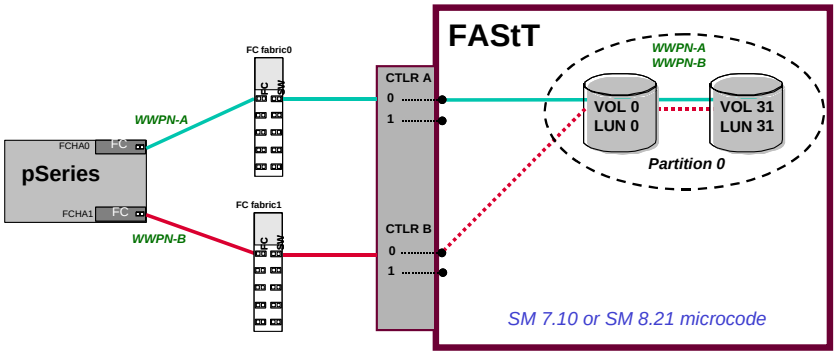


AIX连接DS4000-硬件连接-推荐的单机连接方式

- minimum config #1
- Direct-attach FC-AL
- 1 partition/server

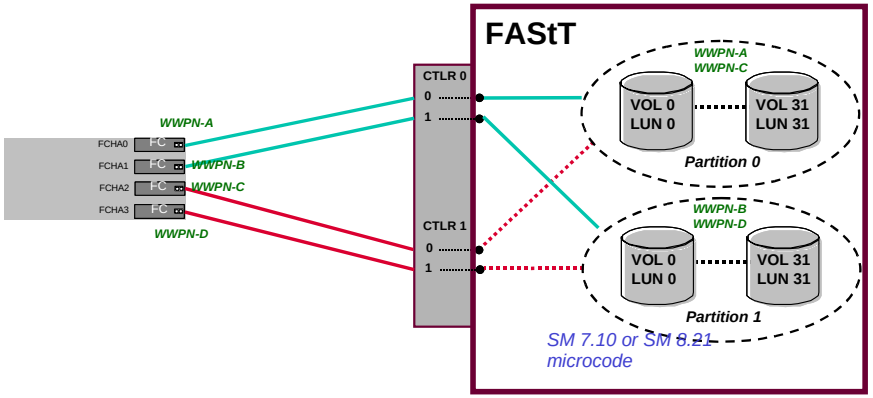


- minimum config #2
- Switch-fabric
- 1 partition/server

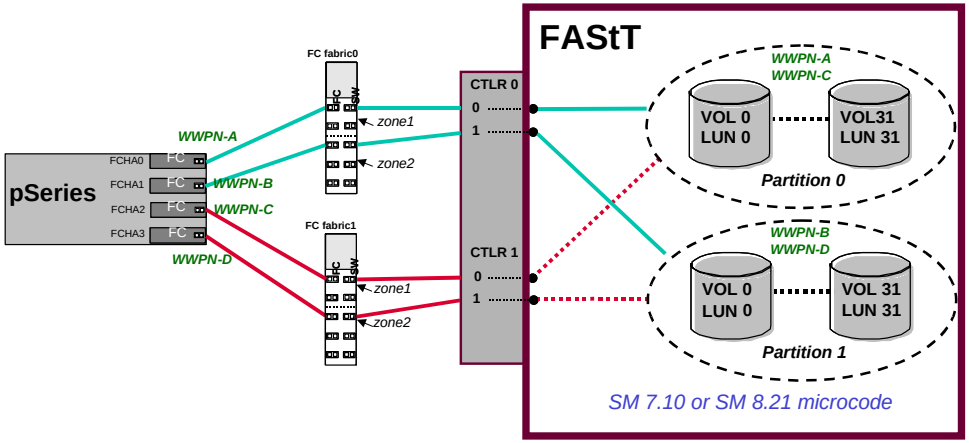


AIX连接DS4000-硬件连接-推荐的单机连接方式

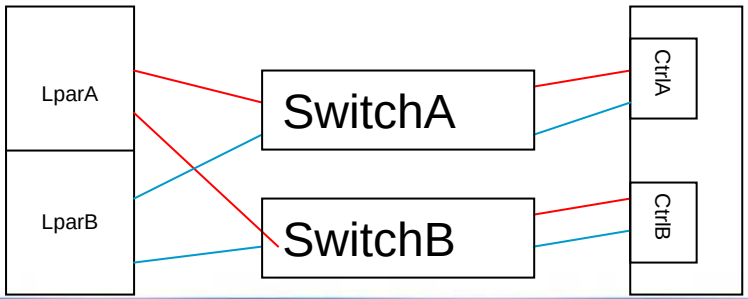
- minimum config #3
- Direct-attach FC-AL
- 2 partition/server



- minimum config #4
- Switch-fabric
- 2 partition/server

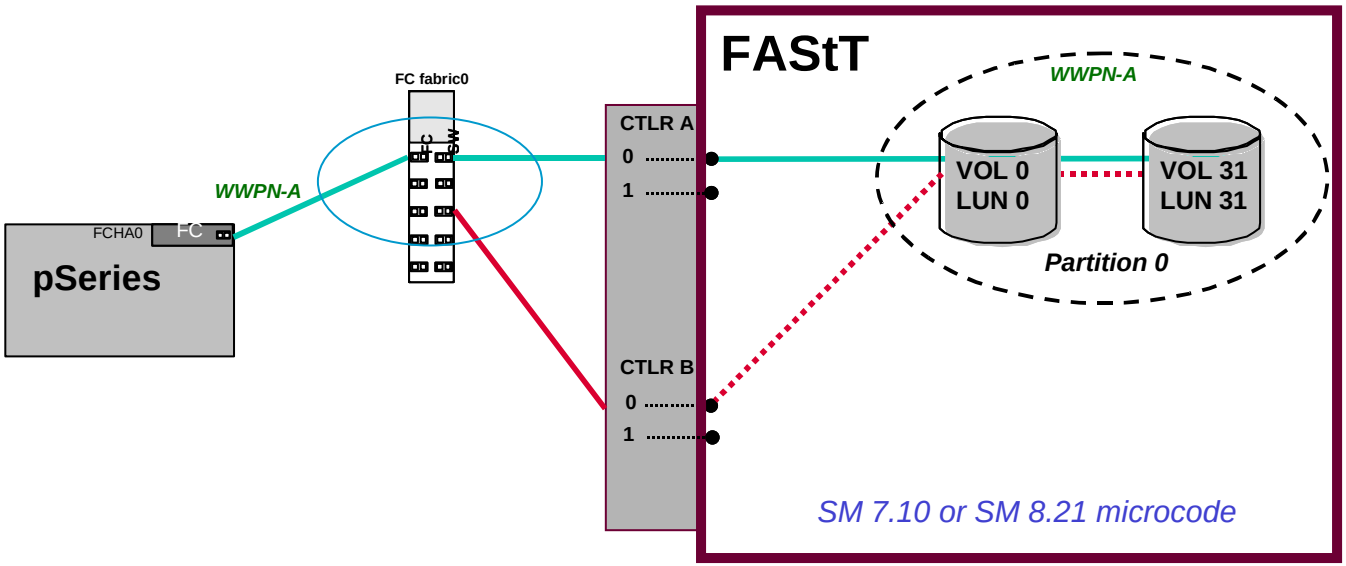


- Or



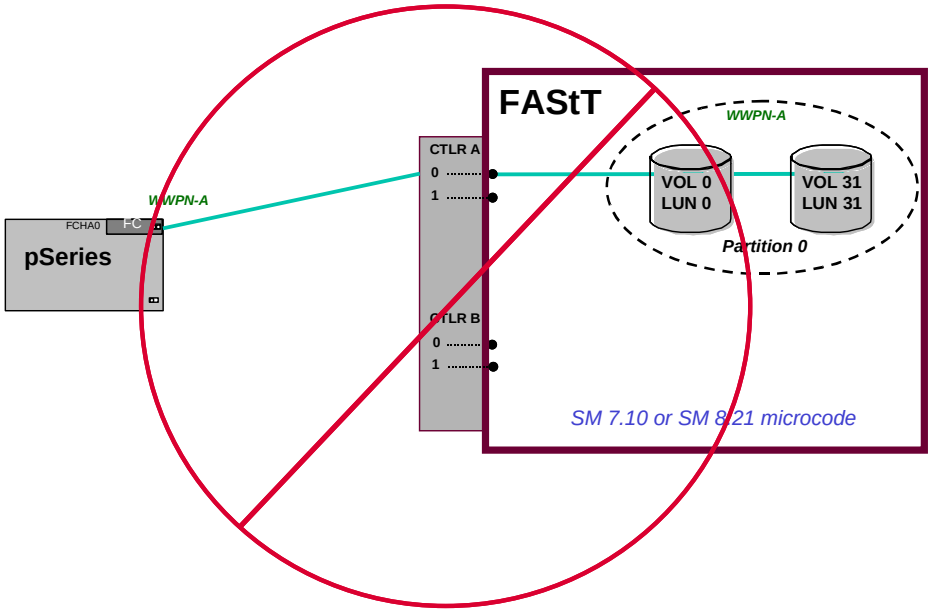
AIX连接DS4000-硬件连接-推荐的单机连接方式

- **Single HBA with Fan Switch Fanout config #5**
- **Switch-fabric**
- **1 partition/server**
- **Single-switch configurations are allowed, but each HBA and FAStT controller combination must be in a separate SAN zone.**

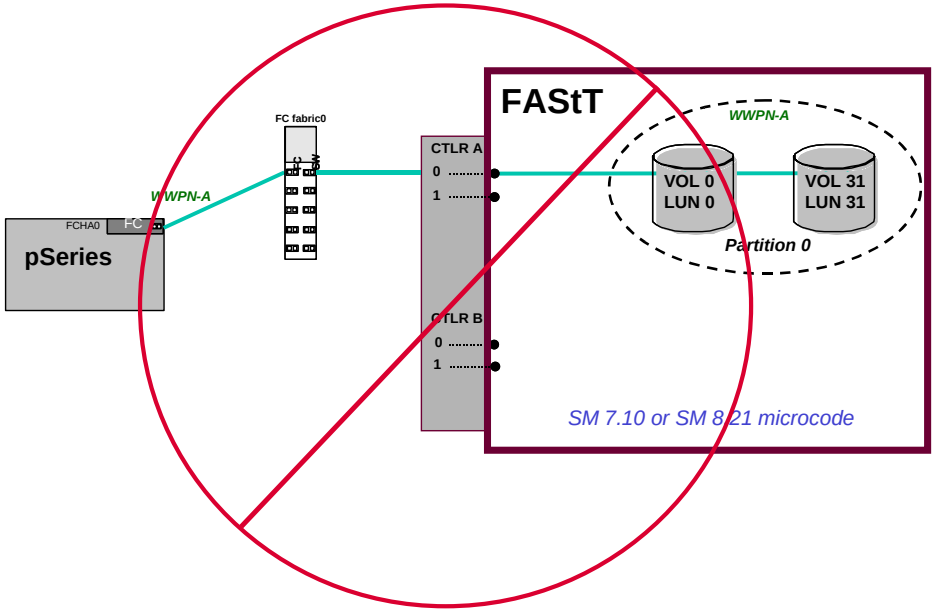


AIX连接DS4000-硬件连接-连接限制1

- Single HBA
- Direct-attach FC-AL
- 1 partition/server

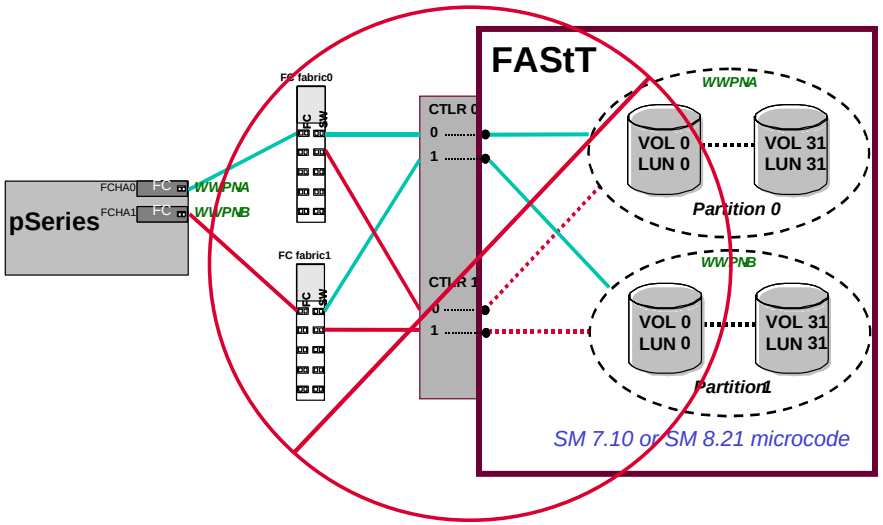


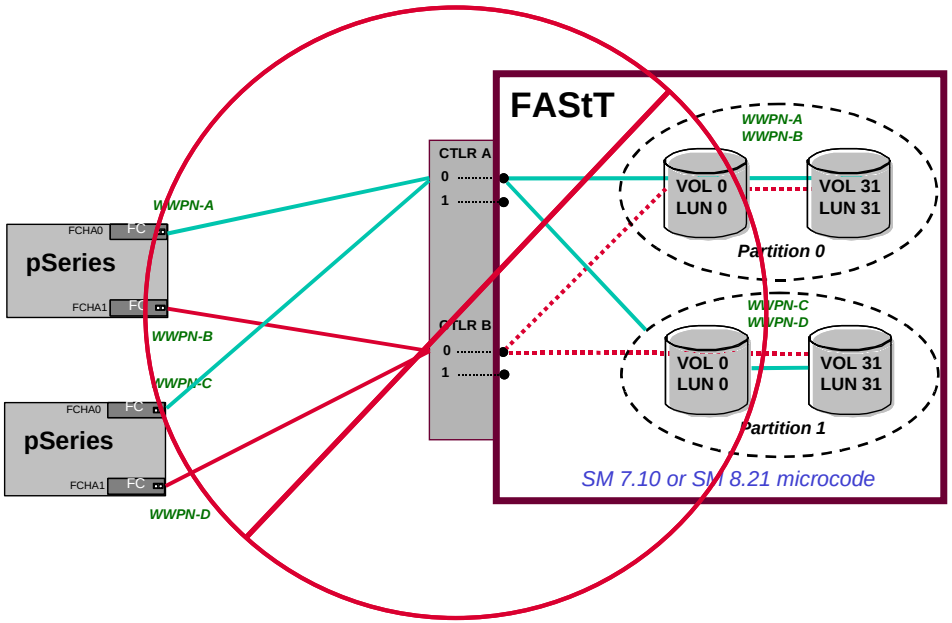
- Single HBA
- Switch-fabric
- 1 partition/server



AIX连接DS4000-硬件连接-连接限制2

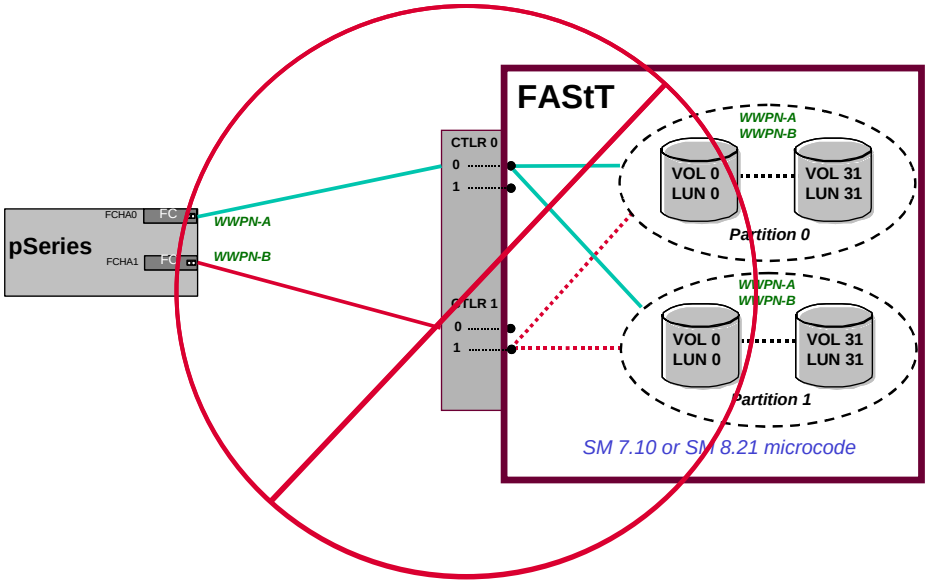
- Dual HBA
- Switch-fabric
- 2 partition/server with 2 HBAs
- Direct-attach FC-AL
- 2 servers on a mini-hub loop



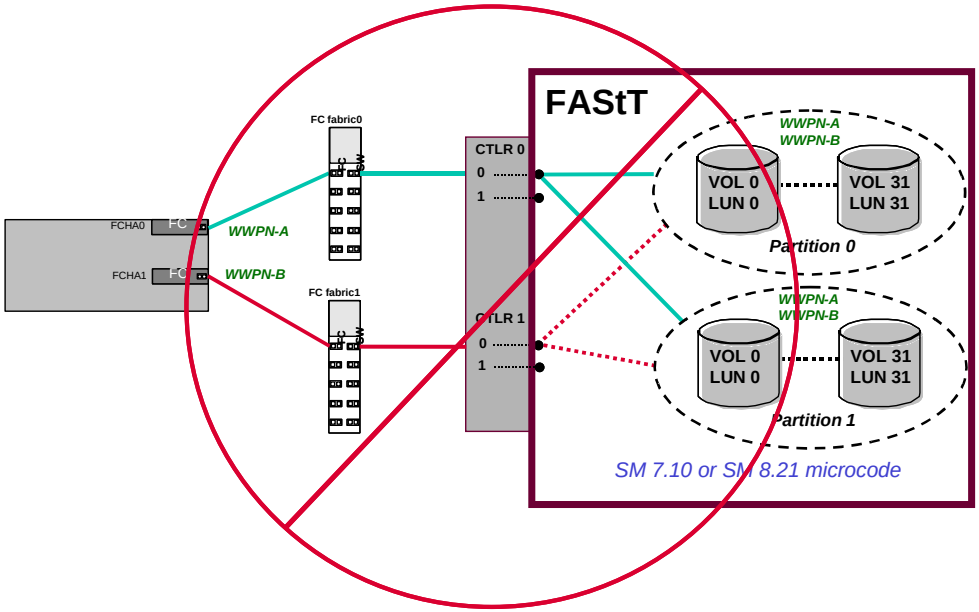


AIX连接DS4000-硬件连接-连接限制3

- Dual HBA
- Direct-attach FC-AL
- 2 partition/server with 2 HBAs
- 一台机器放到2个partition

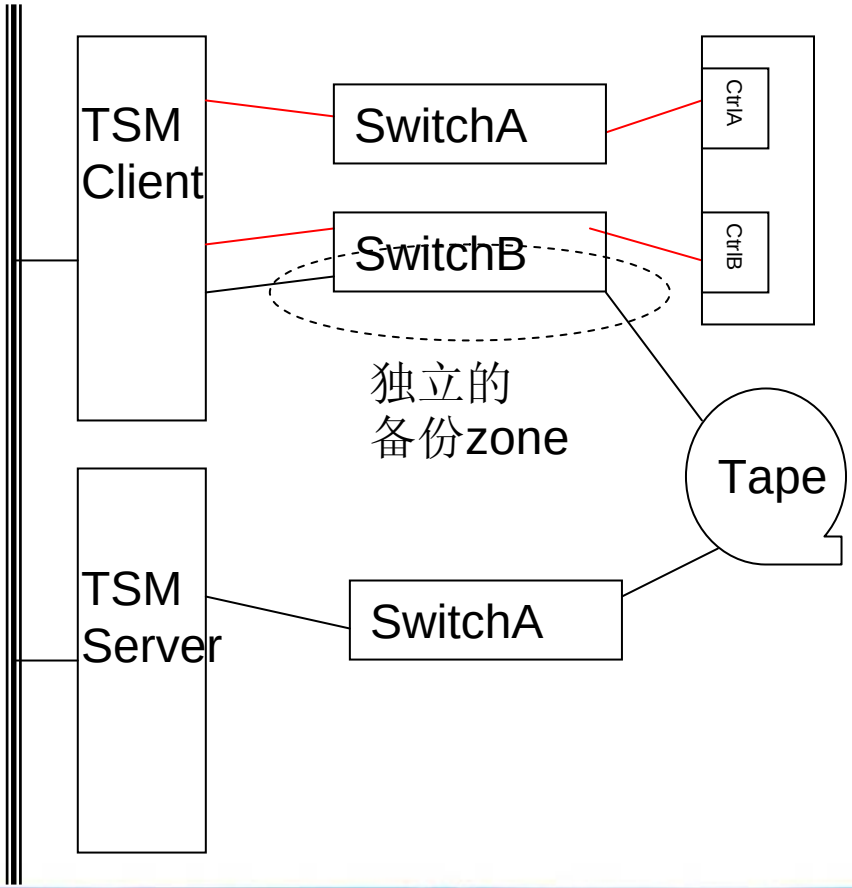


- Dual HBA
- Switch-fabric
- 2 partition/server with 2 HBAs
- 一台机器放到2个partition

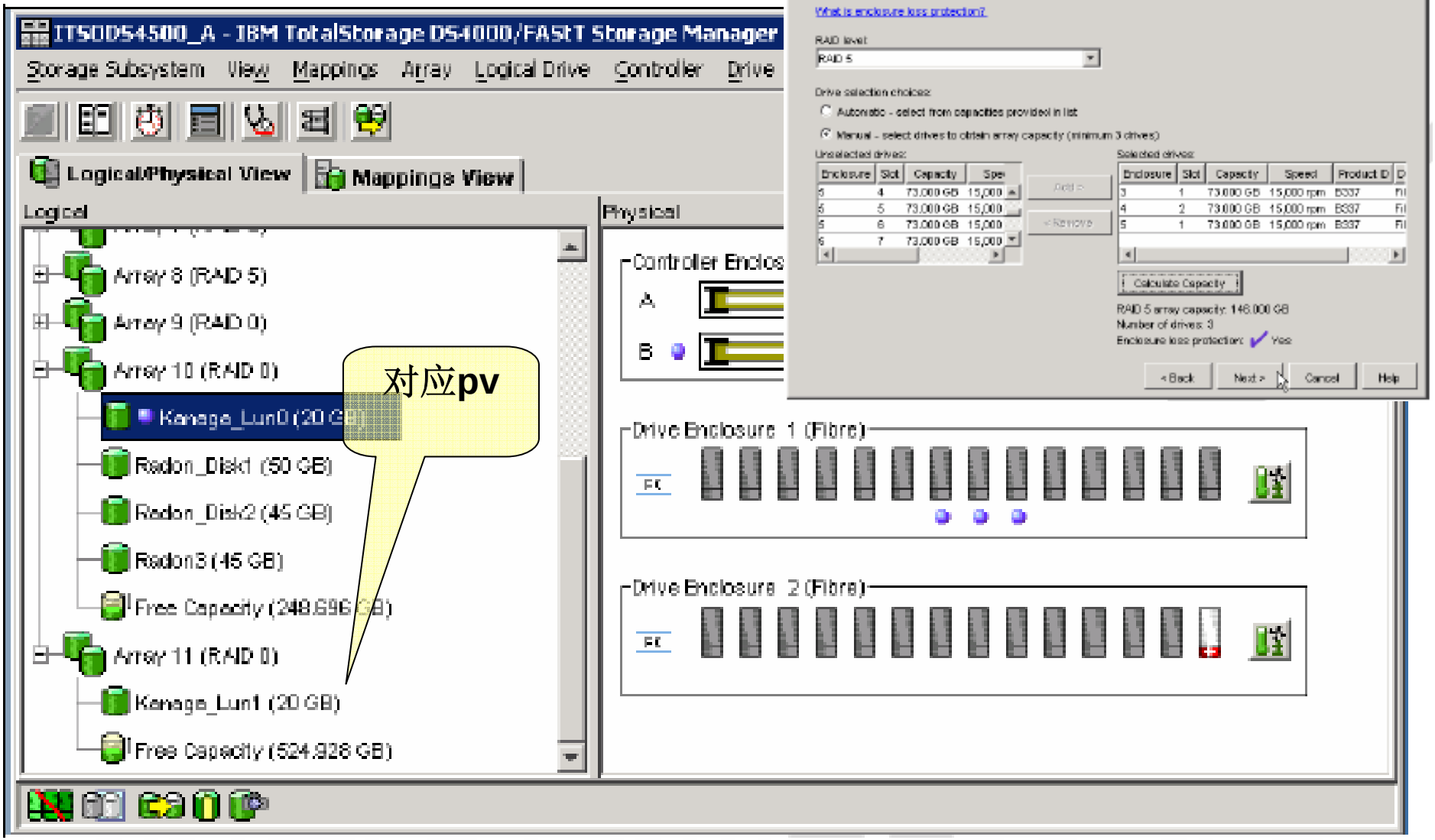


AIX连接DS4000-硬件连接-连接限制4

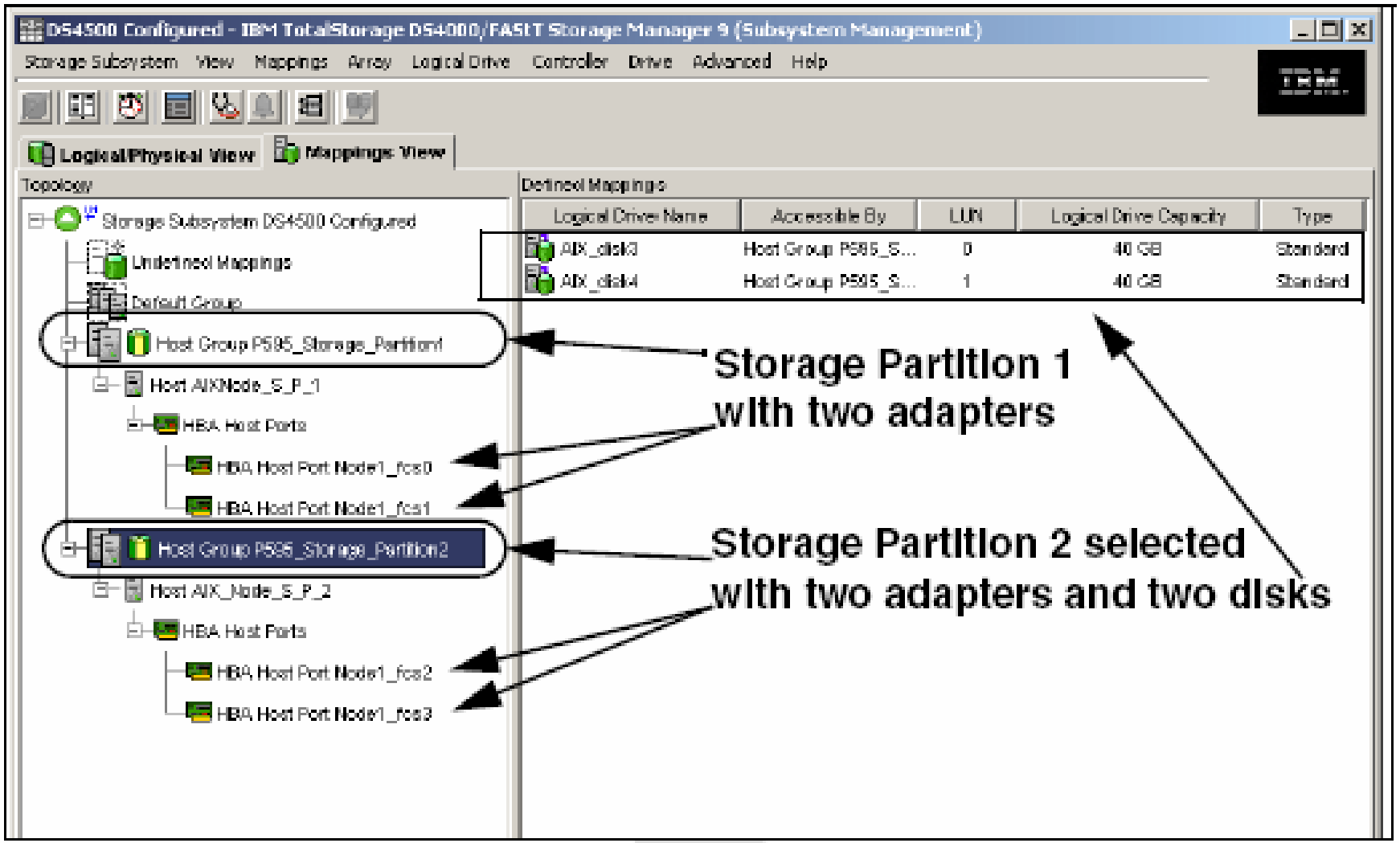
- **Other storage devices, such as tape devices or other disk storage, must be connected through separate HBAs and SAN zones.**
- 在**LAN-Free**备份模式下，必须使用这种配置！



DS4000-划分盘


DS4800 - Specify Array (Create Logical Drive)
To create an array, you must specify the redundancy protection (RAID level) and its overall capacity (number of drives). You can either select the capacity from a list of automatic choices or manually select the drives. If you manually select the drives, you must use the Calculate Capacity button to determine the overall capacity.
What RAID level is best for my application?
What is enclosure loss protection?
RAID level:
RAID 5
Drive selection choices:
Automatic - select from capacities provided in list
Manual - select drives to obtain array capacity (minimum 3 drives)
Unselected drives:
Selected drives:
Calculate Capacity
RAID 5 array capacity: 146.000 GB
Number of drives: 3
Enclosure loss protection: Yes
< Back
Next >
Cancel
Help
IT5DD54500_A - IBM TotalStorage DS4000/FAStT Storage Manager
Storage Subsystem
View
Mappings
Array
Logical Drive
Controller
Drive
Logical/Physical View
Mappings View
Logical
Physical
Array 8 (RAID 5)
Array 9 (RAID 0)
Array 10 (RAID 0)
Kanaga_Lun0 (20 GB)
Radon_Disk1 (50 GB)
Radon_Disk2 (45 GB)
Radon3 (45 GB)
Free Capacity (248.696 GB)
Array 11 (RAID 0)
Kanaga_Lun1 (20 GB)
Free Capacity (524.928 GB)
对应pv
Controller Enclos
A
B
Drive Enclosure 1 (Fibre)
Drive Enclosure 2 (Fibre)

DS4000- 配置AIX主机&映射



（三）AIX连接DS4000-内容

- 驱动程序
- 逻辑设备名
- **AIX**中与磁盘有关的常用命令

AIX连接DS4000- 驱动程序

Install the RDAC driver on AIX

You need the following filesets for the AIX device driver:

- devices.fcp.disk.array.rte - RDAC runtime software
- devices.fcp.disk.array.diag - RDAC diagnostic software
- devices.common.IBM.fc.rte - Common FC Software

You also need one of the following drivers depending on your HBA:

- devices.pci.df1000f7.com - Feature code for 6227 and 6228 adapters require this driver.
- devices.pci.df1000f7.rte - Feature code 6227 adapter requires this driver.
- devices.pci.df1000f9.rte - Feature code 6228 adapter requires this driver.
- devices.pci.df1080f9.rte - Feature code 6239 adapter requires this driver.
- devices.pci.df1000fa.rte - Feature code 5716 adapter requires this driver.

AIX连接DS4000- 驱动程序

- **dar**
  - The FAStT storage server device
  - The number of dars should be equal to the number of FAStT boxes attached to the system
  - Some FAStT server options could be set by changing dar device attributes
- **dac**
  - The controllers on the FAStT servers
  - Each FAStT server should have 2 controllers
  - Some controller options could be set by changing dac device attributes
- **hdisk**
  - FAStT LUNs mapped to AIX
  - The number of hdisks should be equal to the number of mapped LUNs
  - Some LUN options could be set by changing hdisk device attributes
- **fcs**
  - HBA logical device in AIX
  - The number of fcs should be equal to the number HBAs installed in pServer

AIX中与磁盘有关的常用命令

- **cfgmgr** 重新扫描磁盘
- **lsdev –Cc disk** 列出盘
- **lscfg –vl fcs0** 列出**fcs0**的属性，关注**WWN(Network)**
- **Remove fibre channel devices from system:**
  - **rmdev –dl dar0 –R**
  - **rmdev –dl dac0 –R** **rmdev –dl dac1 –R**
  - **rmdev –dl fcs0 –R** **rmdev –dl fcs1 –R**
- **lsattr -El dar0** 查看**dar**属性
- **lsattr -El dac0** 查看**dac**属性
- **lsattr -El hdisk2** 查看**hdisk2**属性
- **lsdev -Cc array** 查看**location codes of dacX**
- **lsdev -Cc driver** 查看**location codes of fscsiX**
- **lsslot -c pci** 查看**pci** 插槽中的卡对应的逻辑设备
- **lsdev –Ccadapter** 列出卡，关注**fcs0**、**fcs1**。。。
  - fcs0 Available 1n-08 FC Adapter
- lsdev -C|grep 1n-08 # will show all sub-devices under certain fcsX
- p615_2:/>lsdev -C|grep 1n-08
  - **dac0 Available 1n-08-01 3542 (200) Disk Array Controller**
  - **dac1 Available 1n-08-01 3542 (200) Disk Array Controller**
  - **fcnet0 Defined 1n-08-02 Fibre Channel Network Protocol Device**
  - **fcs0 Available 1n-08 FC Adapter**
  - **fscsi0 Available 1n-08-01 FC SCSI I/O Controller Protocol Device**
  - **rmt0 Available 1n-08-01 IBM 3580 Ultrium Tape Drive (FCP)**
  - **smc0 Available 1n-08-01 IBM 3582 Library Medium Changer (FCP)**

AIX连接DS4000- 驱动程序—fget_config看盘的映射

- # fget_config -v -A
- ---dar0---

User array name = 'DMWS FT1'

- dac0 ACTIVE dac2 ACTIVE
- dac0-hdisk4 Pr1A01DWH01vg_Share
- dac2-hdisk5 Pr1B03DWH03vg_Share
- dac0-hdisk6 Pr1A05DWH05vg_Share
- dac2-hdisk7 Pr1B07DWH07vg_Share
- dac2-hdisk12 Pr1A20Oracle_dmws92_1
- dac2-hdisk13 Pr1B21Oracle_dmws92_2
- dac2-hdisk14 Pr1B25Unix_dmws92
- dac0-hdisk20 Pr1A00Unix_dmws92
- dac0-hdisk22 PM1Chordiant

有1个DS4000设备
名字为DMWS FT1

DS4000中的LUN名字

该LUN在哪个控制器上

（四）性能优化-内容

- 磁盘
- 扩展柜
- **RAID**
- 盘阵
- **SAN**
- 主机
- 操作系统
- 应用

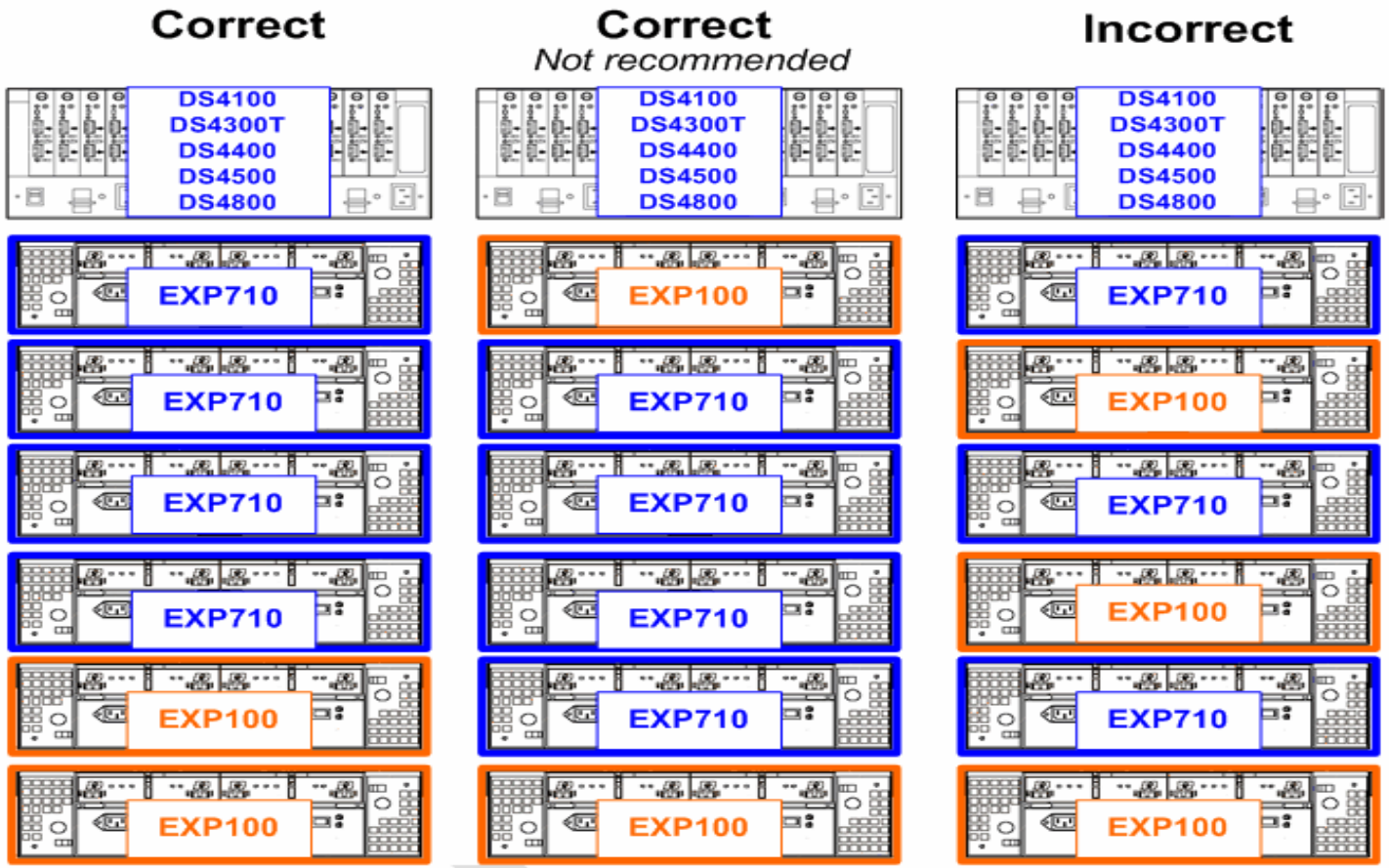
磁盘

按照对性能的要求选择磁盘
——价格与性能的权衡

	Fibre Channel	SATA	SATA difference
Spin Speed	10K and 15K	7.2K	
Command Queuing	Yes 16 Max	No 1 Max	
Single Disk IO Rate[a] (# of 512 bytes IOPS)	280 & 340	88	.31 & .25
Read Bandwidth (MB/s)	69 & 76	60	.86 & .78
Write Bandwidth (MB/s)	68 & 71	30	.44

扩展柜

正确连接扩展柜
——避免大材小用



RAID

按照对性能的要求选择RAID
——空间与性能的权衡

RAID	Description	APP	Advantage	Disadvantage
0	Stripes data across multiple drives.	IOPS Mbps	Performance due to parallel operation of the access.	No redundancy. One drive fails, data is lost.
1	Disk's data is mirrored to another drive.	IOPS	Performance as multiple requests can be fulfilled simultaneously.	Storage costs are doubled.
10	Data is striped across multiple drives and mirrored to same number of disks.	IOPS	Performance as multiple requests can be fulfilled simultaneously. Most reliable RAID level on the DS4000	Storage costs are doubled.
3	Drives operated independently with data blocks distributed among all drives. Parity is written to a dedicated drive.	Mbps	High performance for large, sequentially accessed files (image, video, graphical).	Degraded performance with 8-9 I/O threads, random IOPS, smaller more numerous IOPS.
5	Drives operated independently with data and parity blocks distributed across all drives in the group.	IOPS Mbps	Good for reads, small IOPS, many concurrent IOPS and random I/Os.	Writes are particularly demanding.

RAID

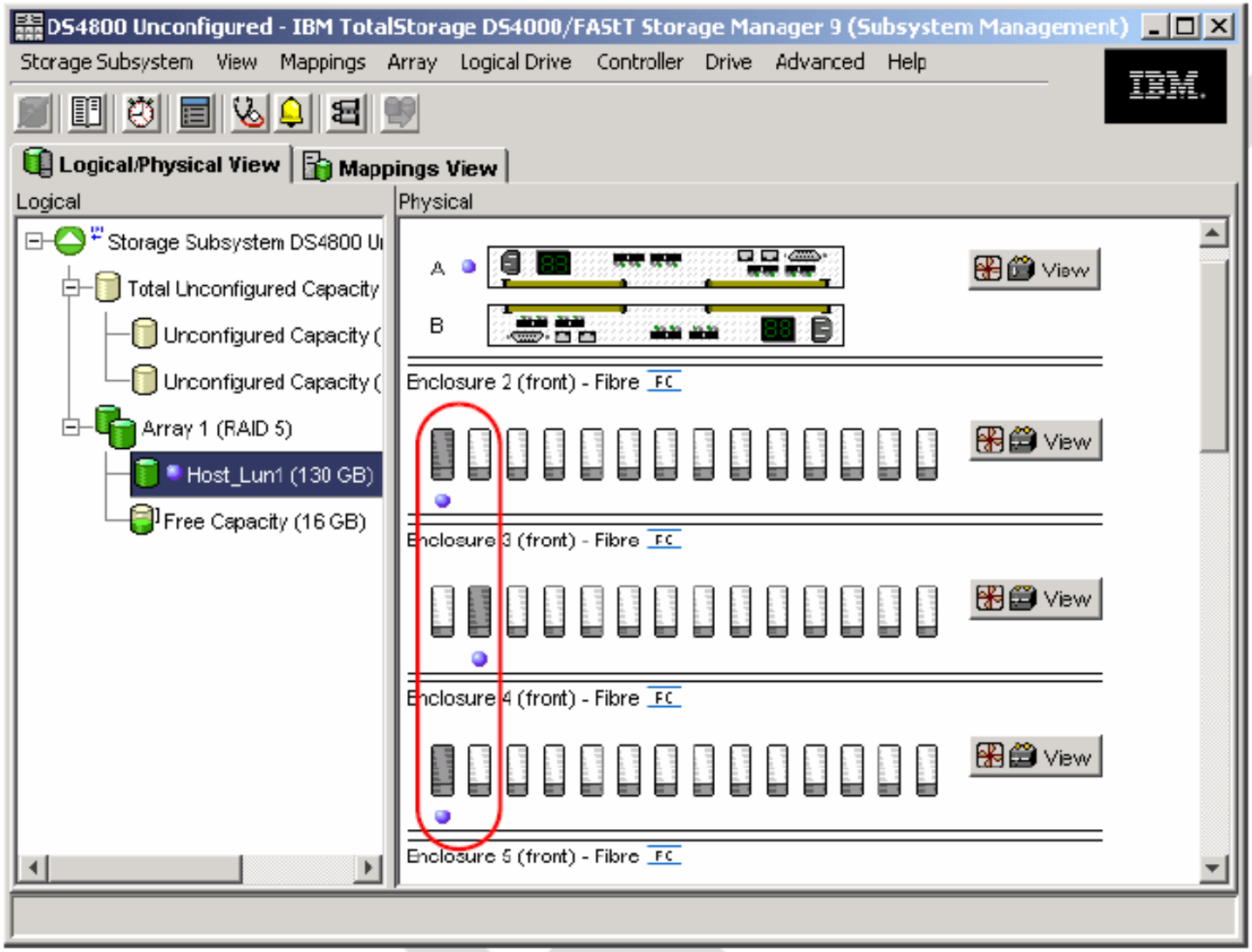
按照对性能的要求选择RAID
——空间与性能的权衡

Table 2-4 RAID level and performance

RAID levels	Data capacity[a]	Sequential I/O performance[b]		Random I/O performance[b]	
		Read	Write	Read	Write
Single disk	n	6	6	4	4
RAID-0	n	10	10	10	10
RAID-1	n/2	7	5	6	3
RAID-5	n-1	7	7[c]	7	4
RAID-10	n/2	10	9	7	6

a. In the data capacity, n refers to the number of equally sized disks in the array.
b. 10 = best, 1 = worst. We should only compare values within each column. Comparisons between columns are not valid for this table.
c. With the write back setting enabled.

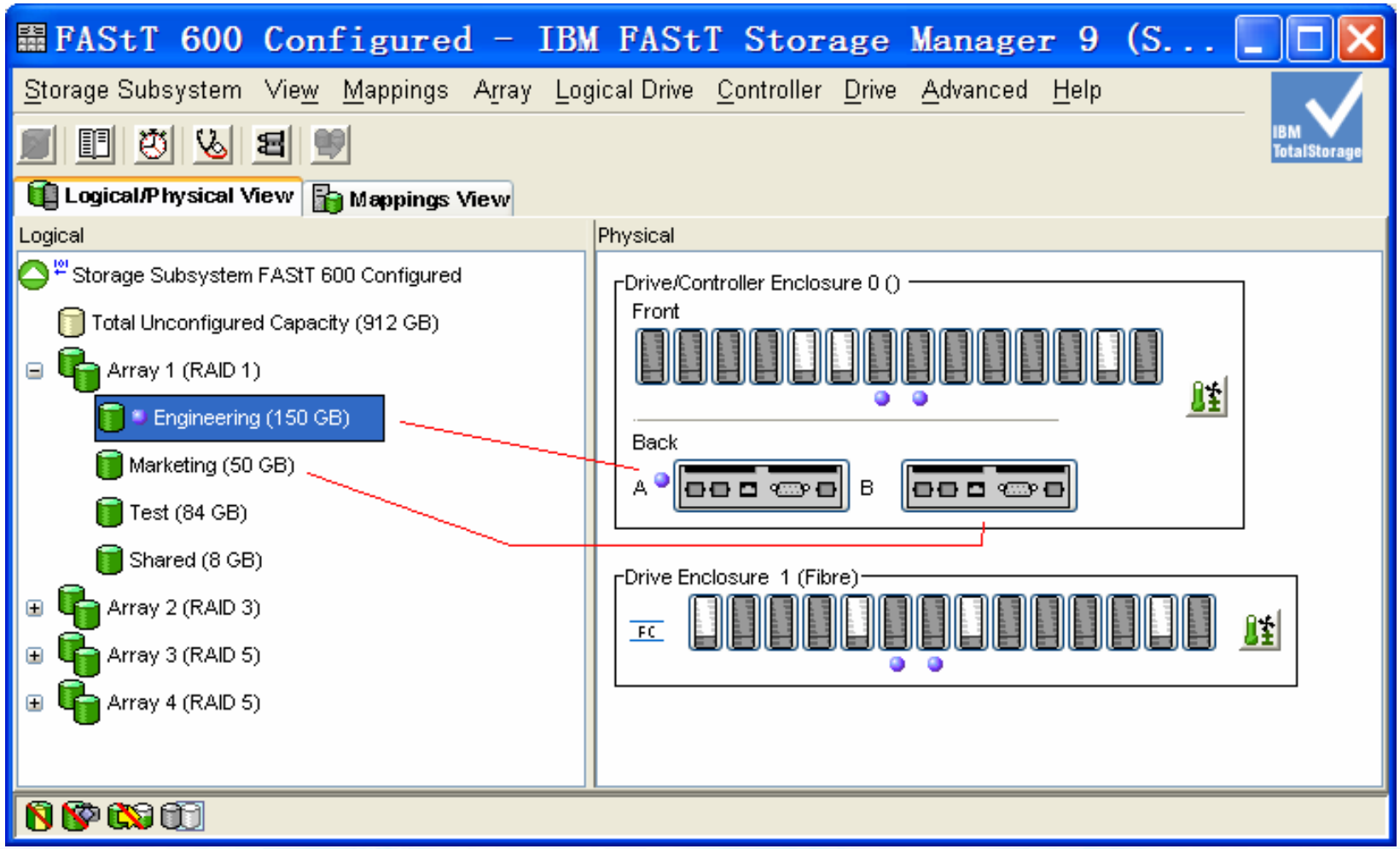
盘阵-优化1-磁盘路径负载均衡



同一个RAID中盘，会同时访问。

盘阵-优化2-控制器间负载均衡

让多控制器同时工作



盘阵-优化3-缓存块大小

•Best Practice:Set the cache blocksize to 4K for the DS4000 system normally for transaction intense environments.

•Best Practice: Set the cache blocksize to 16K for the DS4000 system normally for throughput intense environments

•Tip: Throughput operations though impacted by smaller cache blocksize can still perform reasonable if all other efforts have been accounted for. Transaction based operations are normally the higher concern, and therefore should be the focus for setting the server wide values if applicable.

盘阵-优化4-Cache flush control settings

•With the cache flush settings you can determine what level of write cache usage can be reached before the server will start to flush the data to disks, and at what level the flushing will stop.

•Best Practice: Start with “Start/Stop flush settings of 50/50, and adjust from there. Always keep them equal to each other.

盘阵-优化5-Logical drive segments

•The segment size is the maximum amount of data that is written or read from a disk per operation before the next disk in the array is used.

•Best Practice: With the DS4000 Storage Server it is recommended that the segment size be 64KB to 128KB for most high transaction workloads.

•Best Practice: In the throughput environment you desire the stripe size to be equal to, or a even multiple of the host IO size.

盘阵-优化6-cache read-ahead multiplier

•This parameter is used to increase the number of segments that are read into cache to increase the amount of data that is readily available to present to the host for sequential IO requests. To avoid excess read IOs in the random small transaction intense environments you should disable the cache read-ahead multiplier for the logical drive by setting it to “0”.

•Best Practice: For high throughput with sequential IO, enable cache read-ahead multiplier. For high transactions with random IO, disable it.

盘阵-影响性能的因素

- **Media Scan**
  - Best Practice: Setting media scan to 30 days has been found to be a good general all around value to aid in keeping media clear and server background process load at an acceptable level.
- **Defragmenting an array**
  - **Important:** Once this procedure is started, it cannot be stopped; and no configurationchanges can be performed on the array while it is running.
- **Copyback**
  - **Copyback refers to the process of copying data from a hot-spare drive (used as a standby in case of possible drive failure) to a replacement drive. When you physically replace the failed drive, a copyback operation automatically occurs from the hot-spare drive to the replacement drive.**
- **Initialization**
  - This is the deletion of all data on a drive, logical drive, or array. In previous versions of the storage management software, this was called format.

盘阵-影响性能的因素

- **Dynamic Segment Sizing (DSS)**
  - Dynamic Segment Sizing (DSS) describes a modification operation where the segment size for a select logical drive is changed to increase or decrease the number of data blocks that the segment size contains. A segment is the amount of data that the controller writes on a single drive in a logical drive before writing data on the next drive.
- **Dynamic Reconstruction Rate (DRR)**
  - Dynamic Reconstruction Rate (DRR) is a modification operation where data and parity within an array are used to regenerate the data to a replacement drive or a hot spare drive. Only data on a RAID-1, -3, or -5 logical drive can be reconstructed.
- **Dynamic RAID Level Migration (DRM)**
  - Dynamic RAID Level Migration (DRM) describes a modification operation used to change the RAID level on a selected array. The RAID level selected determines the level of performance and parity of an array.

盘阵-影响性能的因素

- **Dynamic Capacity Expansion (DCE)**
  - **Dynamic Capacity Expansion (DCE) describes a modification operation used to increase the available free capacity on an array. The increase in capacity is achieved by selecting unassigned drives to be added to the array.**

- **Dynamic logical drive Expansion (DVE)**
  - **Dynamic logical drive Expansion (DVE) is a modification operation used to increase the capacity of a standard logical drive or a FlashCopy repository logical drive. The increase in capacity is achieved by using the free capacity available on the array of the standard or FlashCopy repository logical drive.**

SAN

- 划zone隔离，VSAN隔离，交换机隔离
- 用硬件zone
- 用“2点” 划zone

主机-HBA卡负载均衡

让多HBA卡同时工作

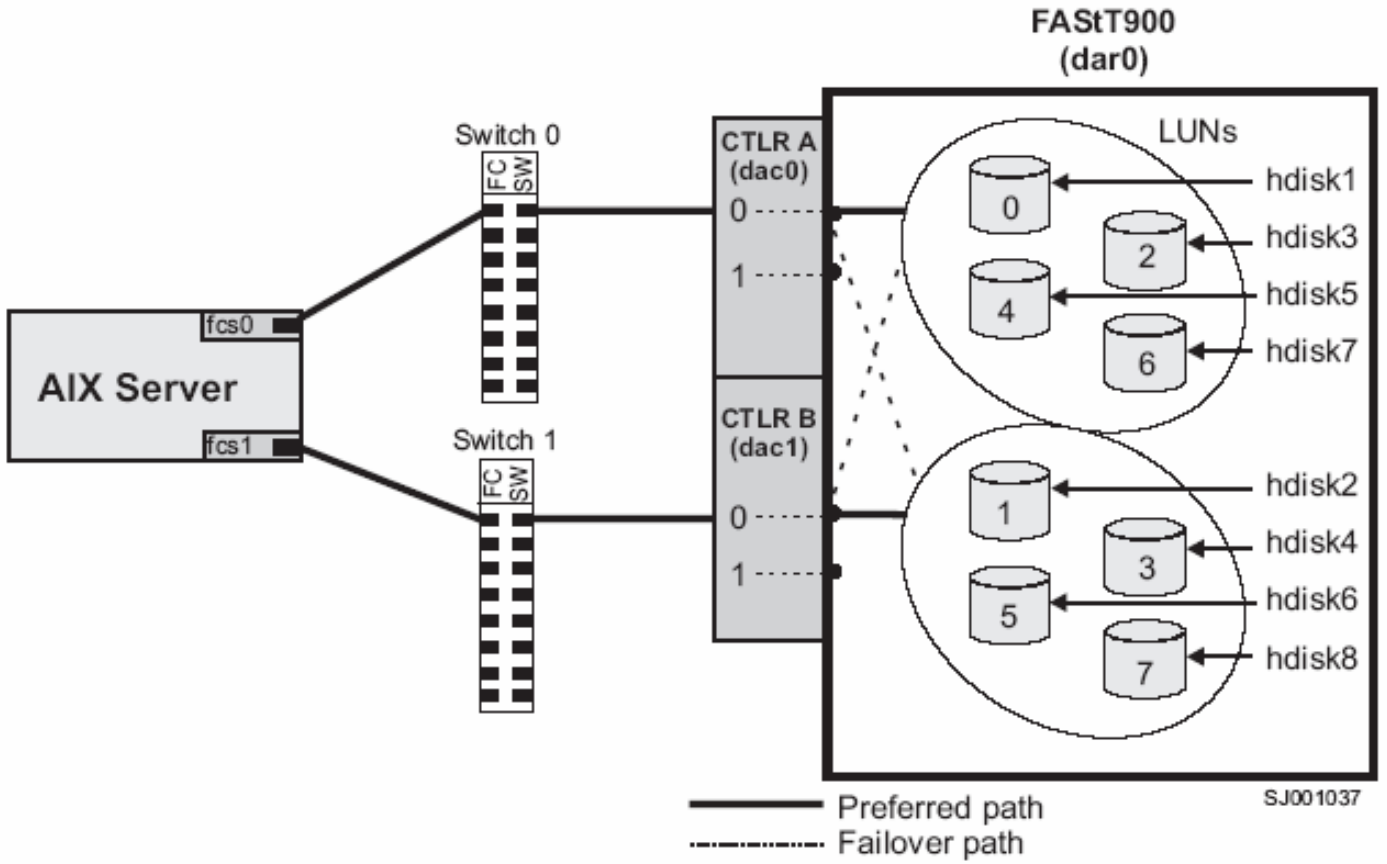


Figure 7. FAStT/AIX system configuration

主机-PCI负载均衡

分散到不同的**PCI** 总线上

操作系统

- 利用各种**OS**的**LVM**
  - 将一个LV分散到不同的HBA卡、光路、控制器。
  - RANGE of physical volumes [maximum]
- 调整**IO**块大小
- ***HBA Queue depth***
  - **The queue depth is the maximum number of commands that can be queued on the system at the same time. The DS4000 controller firmware version 05.30.xx.xx or earlier, the queue depth is 512; For the DS4000 controller firmware versions 06.1x.xx.xx or 05.4x.xx.xx, the queue depth is 2048. This is represents 1024 per controller. The formula for the correct queue depth on the host HBA for this level of firmware code is:**

    2048 / (number of hosts * LUNs per host)
  - **For example, a system with four hosts, each with 32 LUNs, would have a maximum queue depth of 16: 2048 / (4 * 32) = 16.**

应用

- 使用裸卷，要调整I/O块大小
- 区分I/O的随机性质
  - 数据库应用是随机I/O
  - 媒体、备份/恢复、归档应用是连续I/O
- 区分I/O的读写性质
  - 媒体是读占多
  - 备份/恢复、归档应用是写占多
  - 数据库根据表的性质优化表空间、容器

调优

- 没有“绝对值”
- 有时需要权衡利弊
- “性能与其说是调整，不如说是规划”

谢谢！