VMworld2005
virtualizenow

PAC 267-A
ESX Server Storage I: Tips and Tricks

Mostafa Khalil
Bob Slovick

This presentation may contain VMware confidential information.

Copyright © 2005 VMware, Inc. All rights reserved. All other marks and names mentioned herein may be trademarks of their respective companies.

What We Will Discuss..

- Introduction to ESX Server in a SAN environment
- Design methodologies
- Path management
- Storage performance optimization
- Basic troubleshooting and trouble avoidance
- Boot from SAN

Storage Vendor Partners

- Dell
- EMC
- Fujitsu
- Fujitsu Siemens
- HP
- IBM
- Network Appliance
- NEC
- 3PAR
- StorageTek
- Sun
- Xiotech

First Steps: Virtual Machine Storage Characterization

- How critical is the virtual machine?
- What are its performance requirements?
- What are its availability requirements?
- What are its Point-in-Time (PiT) restoration requirements?
- What are its backup requirements?
- What are its replication requirements?

In short, what “tier” storage does that virtual machine belong on?

What Are Some of the Tiers?

- High Tier: High performance, high availability, often built in snapshots, to facilitate backups and Point-in-Time (PiT) restorations, replication, full controller redundancy, fibre drives. High cost spindles
- Mid Tier: Mid-range performance, lower availability, MAYBE snapshots, some controller redundancy, SCSI drives. Medium cost spindles
- Lower Tier: Low performance, little internal storage redundancy, low end SCSI drives or SATA. Low cost spindles

Tiered Storage…

Warnings…

- Technology changes and push higher tier features to a lower tier: Better, faster, cheaper…
- A virtual machine may change tiers throughout its “lifecycle”, due to changes in criticality or changes in technology
- Criticality is relative, and may change for a variety of reasons, including changes in the organization, operational processes, regulatory requirements, disaster planning, etc.

Tiered Storage…

**How many 9’s
are you willing to pay for?**

- **The truth:** Not all applications need to be on the highest performance, most available storage… At least not throughout their entire lifecycle…
- What if you need some of the functionality of the higher tier, such as snapshots, but don’t want to pay for it? Do it in software… Example…redo logs…

What are Redo Logs?

- They track changes to a virtual machines file system, and allow you to commit the changes, or fall back to a prior Point-in-Time
- Can facilitate backups, and change management
- Done at an ESX Server level, regardless of the tier storage the virtual machine is on

Redo Logs: Considerations…

- They can grow up to the size of the original disk
- The more changes to the disk, the slower the performance
- The redo logs grow in 16 MB increments
- There is a performance penalty on the disk subsystem and ESX Server
- They can be invoked “on the fly”, or with the virtual machine shut down

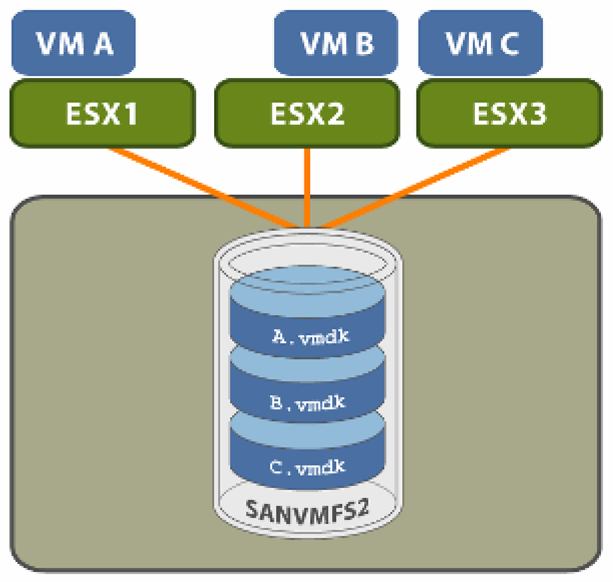
Redo Logs: Considerations…

- If they are invoked with the virtual machine up and running, if the application is not quiesced, meaning “to quiet”, the backup may only be “crash consistent”, instead of being “transactionally consistent”

Note that crash consistent backups may be acceptable for some DR applications.

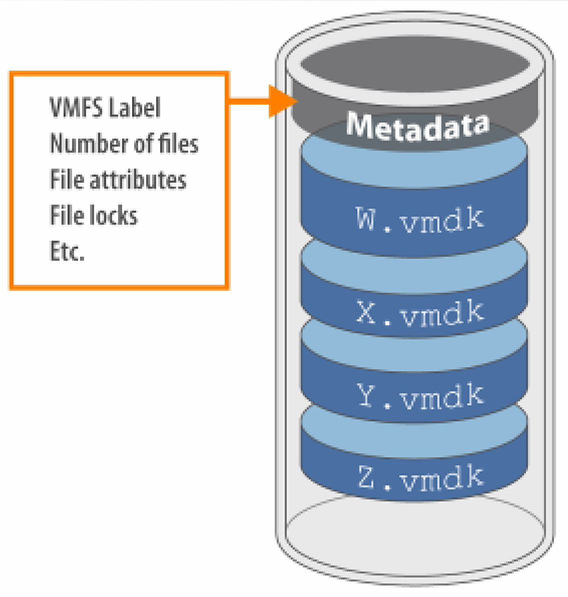
Sharing a VMFS Across ESX Servers

- No communication exists between multiple ESX Servers accessing the same VMFS volume
- To coordinate access to virtual disk files ESX Server uses file-level locks
- To coordinate access to VMFS internal file system information ESX Server uses SCSI reservations on entire LUN



Structure of a VMFS

- A VMFS holds files and has its own metadata
- Metadata gets updated through:
  - Creating a file
  - Changing a file’s attributes
  - Powering on a virtual machine
  - Powering off a virtual machine
  - Growing a file
- When metadata is updated, the VMkernel places a non-persistent SCSI reservation on the entire VMFS volume
  - Lock held on volume for the duration of the operation
  - Other VMkernels are prevented from doing metadata updates


VMFS Label
Number of files
File attributes
File locks
Etc.

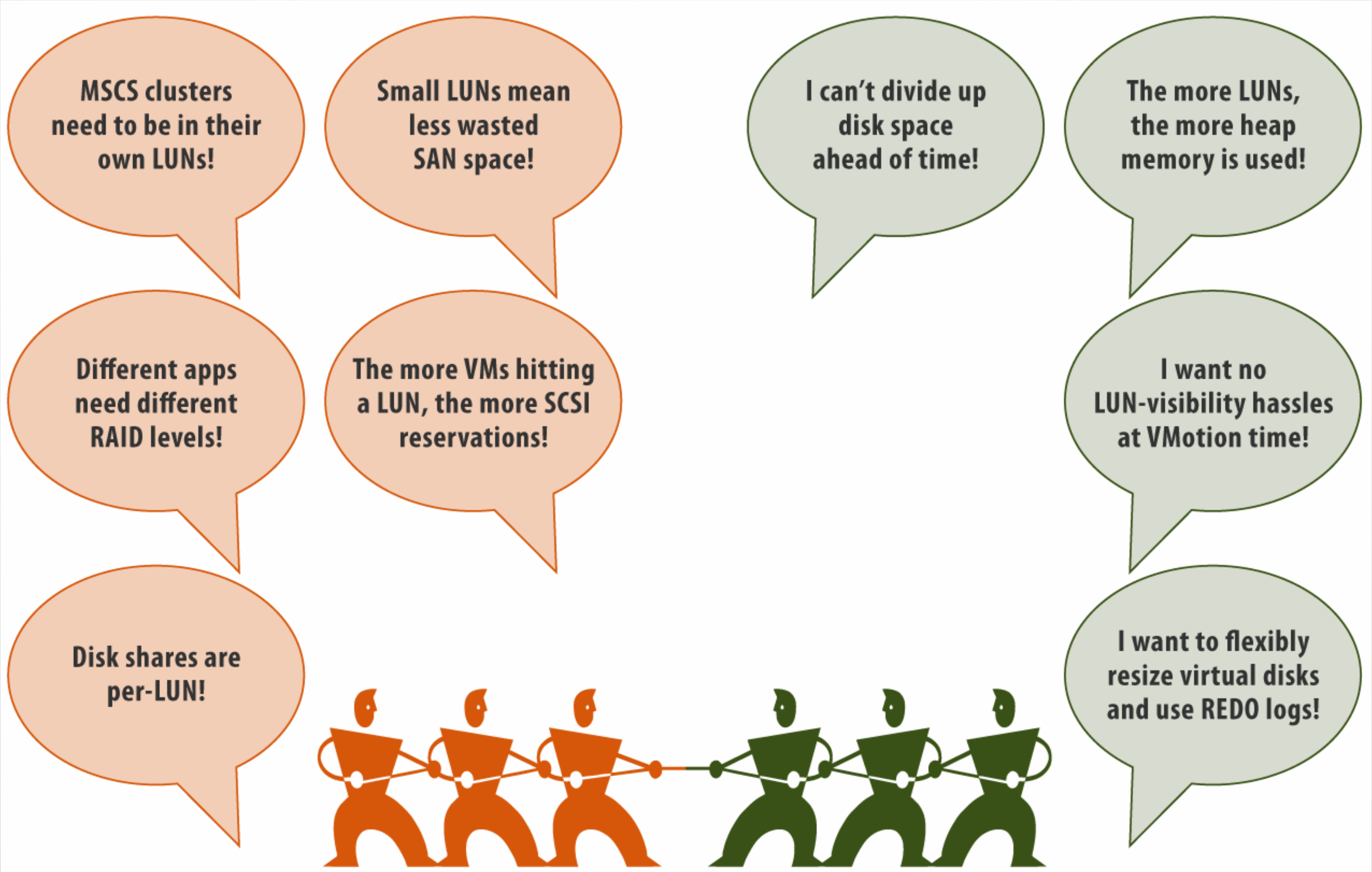
Related Considerations

- VM storage access should be planned for peak periods, but different applications may have different peak access periods
- Virtual machines sharing a common VMFS make it more difficult to characterize peak access periods, or optimize performance
- The more VMs sharing a VMFS, the greater the potential of performance degradation due to I/O contention

Some Tips to Consider…

- Each LUN should have the right RAID level and storage characteristic for applications in virtual machines that will use it
- One LUN should contain only one single VMFS volume
- If multiple virtual machines accessing same LUN, use disk shares to prioritize virtual machines

How Many VM Disks on a VMFS?



MSCS clusters need to be in their own LUNs!

Small LUNs mean less wasted SAN space!

Different apps need different RAID levels!

The more VMs hitting a LUN, the more SCSI reservations!

Disk shares are per-LUN!

I can't divide up disk space ahead of time!

The more LUNs, the more heap memory is used!

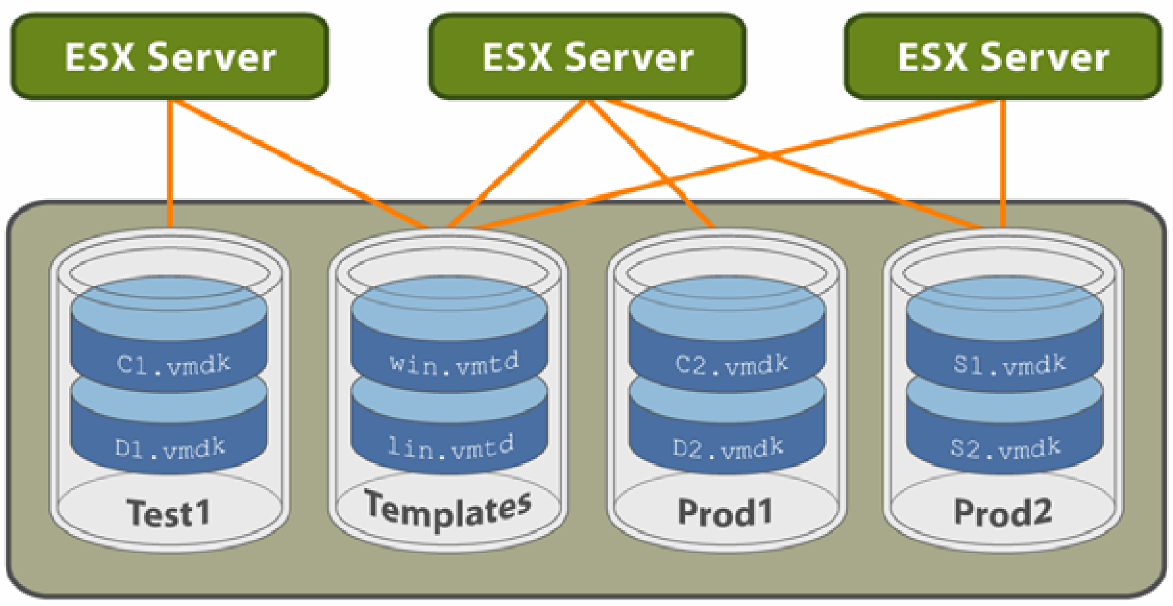
I want no LUN-visibility hassles at VMotion time!

I want to flexibly resize virtual disks and use REDO logs!

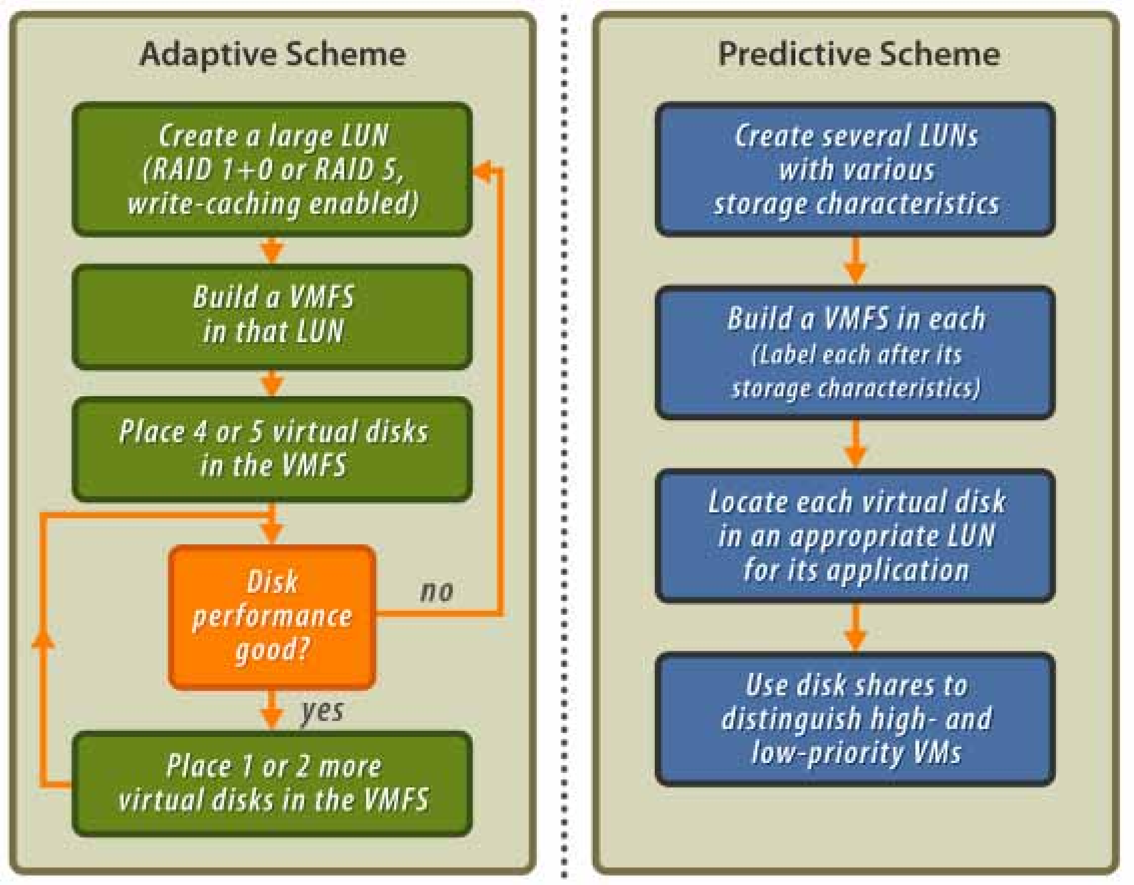
← More, smaller LUNs   Fewer, bigger LUNs →

Multiple VMFS

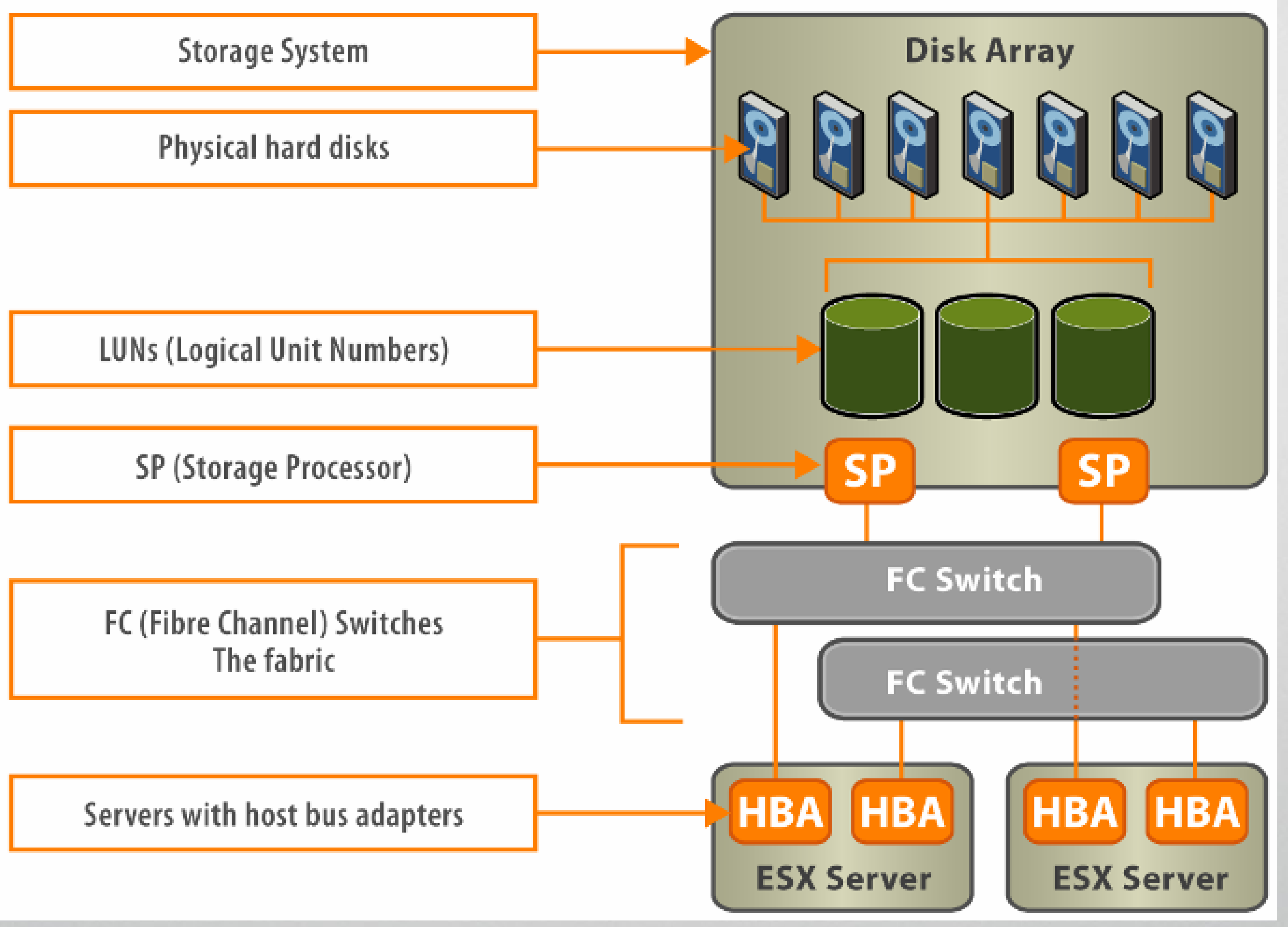
- Separate test virtual machines from production virtual machines. Test virtual machines often run in redo mode, production virtual machines normally run in persistent mode
- Store templates for deploying virtual machines using VirtualCenter, and optimize it for sequential read operations (read-ahead)



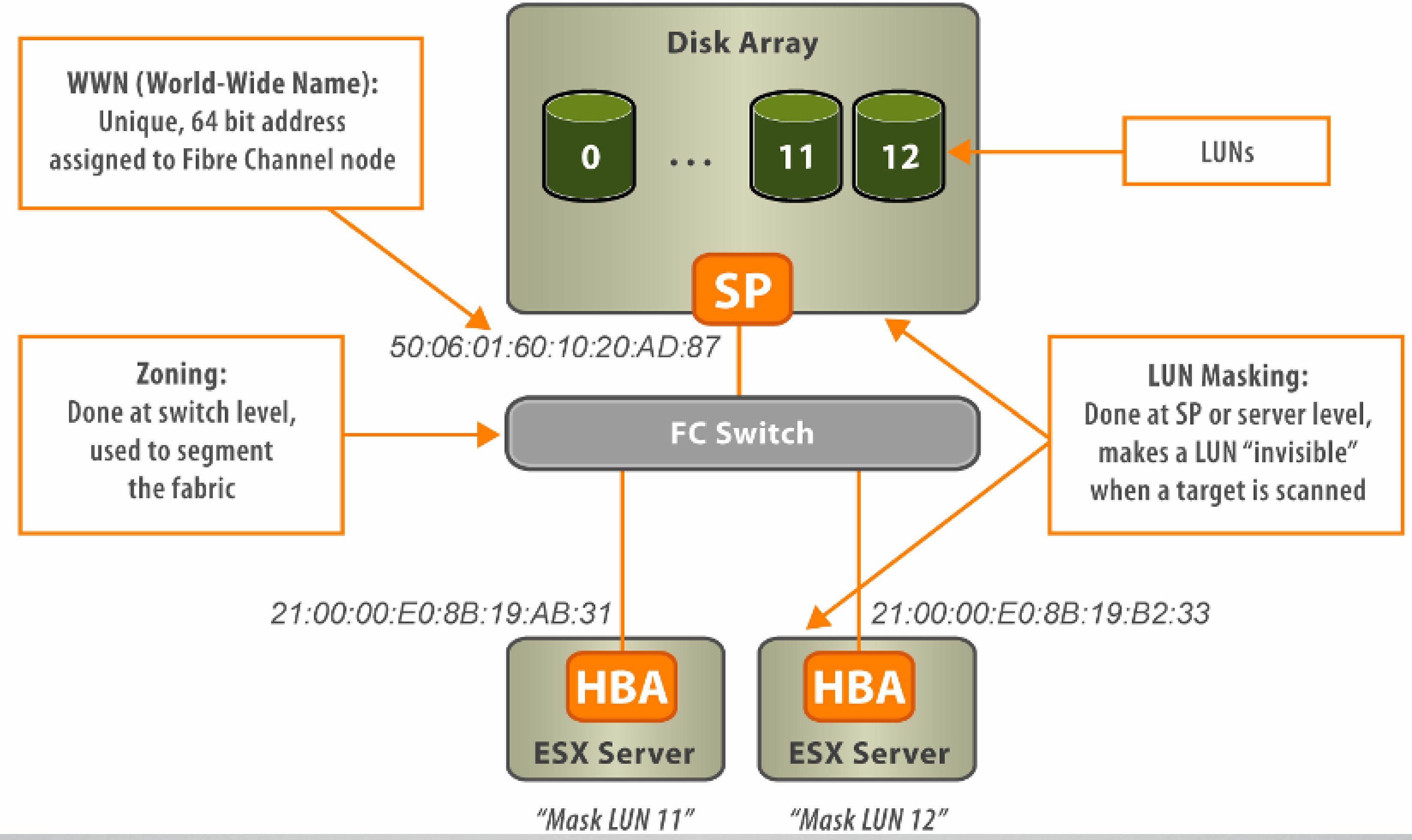
What Can You Do When Data to Characterize Storage for a Virtual Machine Is Not Available?


Adaptive Scheme
Create a large LUN (RAID 1+0 or RAID 5, write-caching enabled)
Build a VMFS in that LUN
Place 4 or 5 virtual disks in the VMFS
Disk performance good?
no
yes
Place 1 or 2 more virtual disks in the VMFS
Predictive Scheme
Create several LUNs with various storage characteristics
Build a VMFS in each (Label each after its storage characteristics)
Locate each virtual disk in an appropriate LUN for its application
Use disk shares to distinguish high- and low-priority VMs

A Basic SAN


Storage System
Physical hard disks
LUNs (Logical Unit Numbers)
SP (Storage Processor)
FC (Fibre Channel) Switches
The fabric
Servers with host bus adapters
Disk Array
SP
SP
FC Switch
FC Switch
HBA
HBA
HBA
HBA
ESX Server
ESX Server

Addressing and Access Control in a SAN



WWN (World-Wide Name):
Unique, 64 bit address
assigned to Fibre Channel node

LUNs

50:06:01:60:10:20:AD:87

Zoning:
Done at switch level,
used to segment
the fabric

LUN Masking:
Done at SP or server level,
makes a LUN "invisible"
when a target is scanned

21:00:00:E0:8B:19:AB:31

21:00:00:E0:8B:19:B2:33

"Mask LUN 11"

"Mask LUN 12"

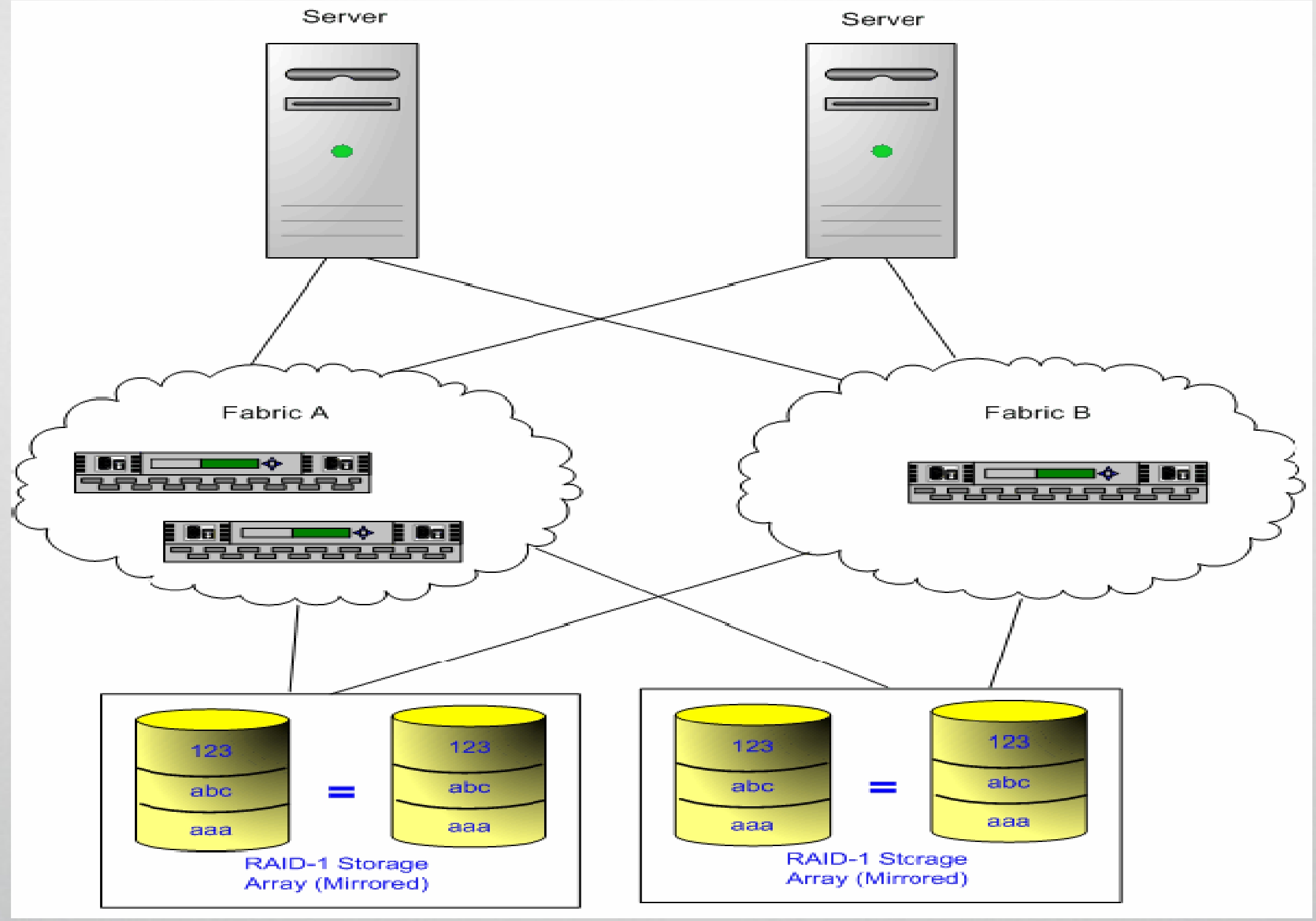
How Can Zoning Help?

- Prevents non-ESX Servers from seeing a particular storage system, and possibly destroying ESX Server VMFS data
- Reduces the number of targets and LUNs presented to an ESX Server
- Controls and isolates paths within a fabric
- Separates test from production environments
- Remember that all ESX Servers in a VirtualCenter farm must be in the same zone

How Can Access Control help?

- Reduces the number of LUNs presented to an ESX Server
- Prevents non-ESX Servers from seeing ESX Server LUNs, and possibly destroying VMFS volumes
- Remember that all ESX Servers in a VirtualCenter farm must be enabled to see all LUNs for VMotion to work

A Highly-Available SAN



Path Management

- Review the most recent SAN Configuration and Compatibility Guides
- Review the SAN implementation guide which most of the major storage vendors have developed for ESX Server
- Pay close attention to FC-HBA and storage controller model/firmware versions, configuration settings and proper failover parameters in ESX Server (MRU or Preferred Path)

Failover Policies

Policy/Controller	Active/Active	Active/Passive
MRU	Administrator action required to fail back after path failure	Administrator action required to fail back after path failure
Fixed	VMkernel resumes using preferred path when connectivity is restored	VMkernel’s attempt to resume preferred path may thrash or fail, because another SP now owns the LUN

Failover Policies

Examples of disk-array types:

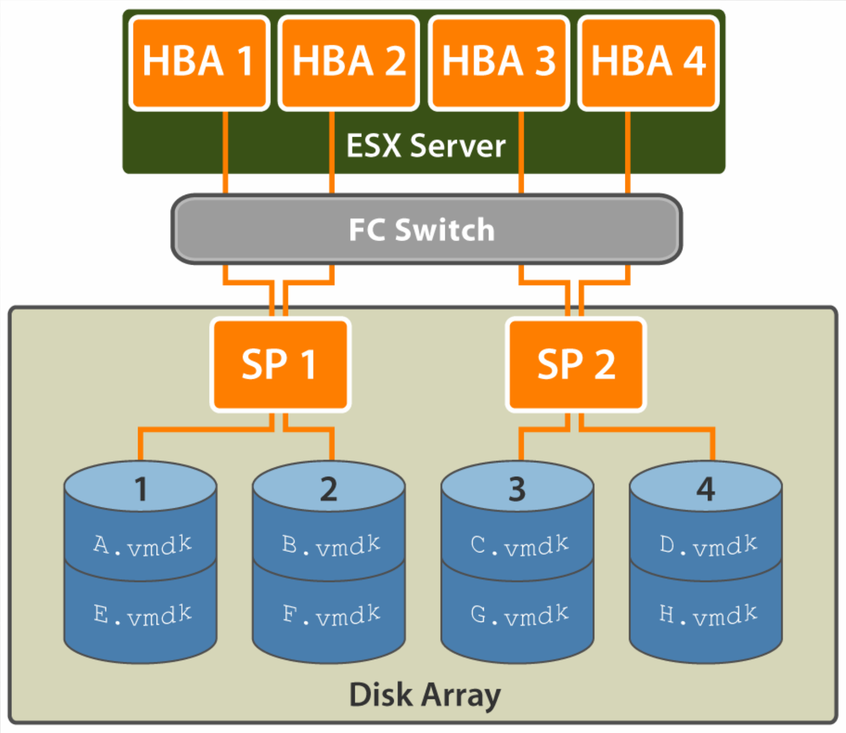
- Active/Active: EMC Symmetrix, IBM ESS (“Shark”), Hitachi 9900
- Active/Passive: HP MSA 1000, EMC CLARiiON, IBM FAStT
  - HP EVA is technically Active/Active, but ESX Server uses it as Active/Passive

Path Management

- Balance out loads among available paths
- Be aware that if a path fails, the surviving paths will be carrying ALL the traffic
- Path switchovers may take a minute or more, as the fabric may “reconverge” with a new topology to try to restore service. This is the preferable method to restore service. If the host path failover time is too short, it may work against fabric reconvergence

Example: Manual Load Balancing

- Active/Active SPs
- 4 ESX Servers in production
- 4 FC HBAs in each server
- Director Class SW
- 4 Fabric Zones
- Define preferred paths
  - LUN 1: vmhba1:1:1
  - LUN 2: vmhba2:1:2
  - LUN 3: vmhba3:2:3
  - LUN 4: vmhba4:2:4
  - Path Policy: fixed



Avoiding and Resolving Problems

Document EVERYTHING!

Include:

- Zoning, access control, storage, switch, server and FC-HBA configuration, and software/firmware versions, and storage cable plant

Take your topology maps, make several copies, and play the game:

- If this element fails, what happens to my SAN?
- Cross off different links, switches, HBAs and other elements to insure you didn’t miss a critical failure point in your design…

Avoiding and Resolving Problems

- When installing ESX Server on a production system, disconnect Fibre Channel HBAs if doing a local install

  Danger: Installer lets you wipe any accessible disks, including SAN LUNs others may be using

- If possible, VMkernel’s resources should be on a path not exposed to SAN-administrator error
  - VMkernel’s core dump partition
  - VMkernel swap file (VMFS partition)
- Dedicate HBAs to VMkernel; do not share with Service Console
  - Eliminates risk of I/O contention between the two
  - Preserves ability to dynamically scan for new SAN LUNs

Avoiding and Resolving Problems

- Insure that the FC-HBAs are installed in the correct slots in the server, based on slot/bus speed and to balance PCI bus load among the available busses in the server
- Become familiar with the various monitor points in your storage network, at all visibility points, including ESX Server, FC switch statistics, and storage performance statistics

What about the Guest OS?

Disks on a SAN may lose connectivity, due to various “fabric events”. This is not a VM specific issue, and can affect any SAN based OS . To change Windows Server 2000/2003, to increase the timeout value to make it more tolerant to the effect, change the following registry key:

- **HKEY_LOCAL_MACHINE\System\CurrentControlSet\Services\Disk\TimeOutValue**
- **REG_DWORD to 60 seconds or more (0x0000003c hexadecimal)**
- **Knowledge Base Article KB 1014**

What is Boot from SAN?

- ESX Server supports installation into, and boot from, SAN disk arrays
  - Only certain Fibre Channel adapters and disk arrays are supported
  - Read the latest documentation before you begin

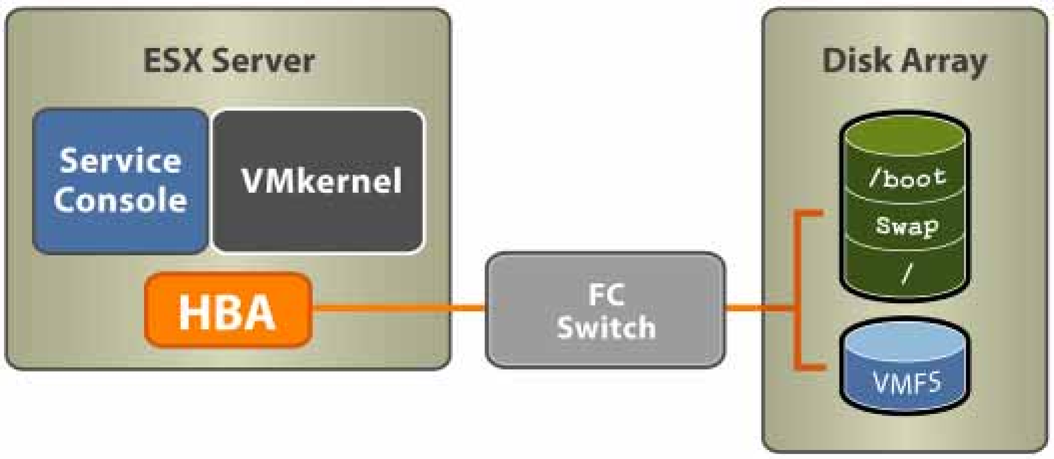
When to Use Boot From SAN

- In hardware configurations, such as on some blade systems
- When maintenance of local Service Console storage is undesirable
- If easy cloning of Service Consoles is desired

When NOT to use Boot from SAN

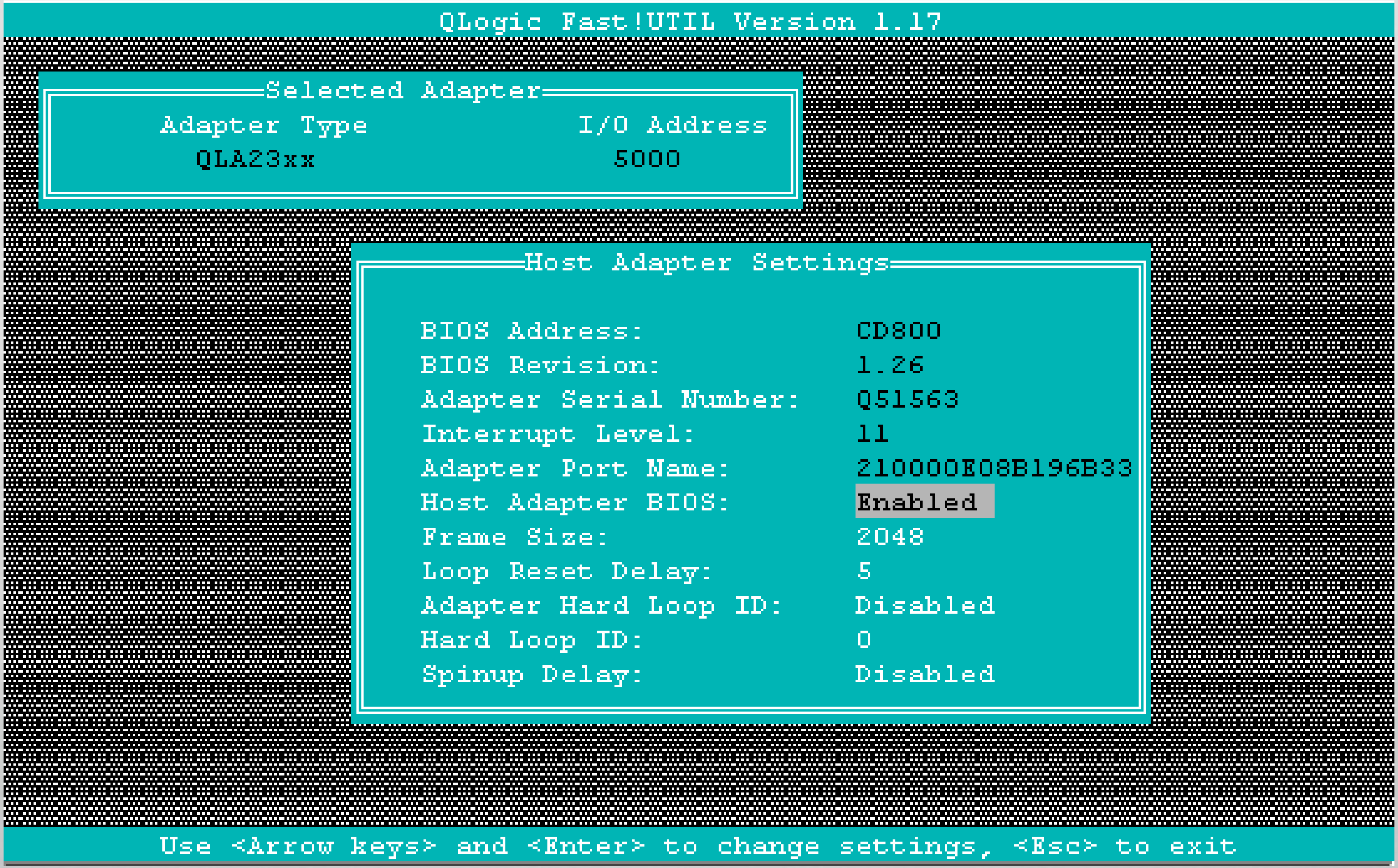
- When there is a risk of I/O contention between Service Console and VMkernel
- When the use of cluster-across-boxes is required
- When the use of VMFS raw disk mappings is required
- When additional dependencies between Service Console and VMkernel are undesirable

How Boot From SAN Works…



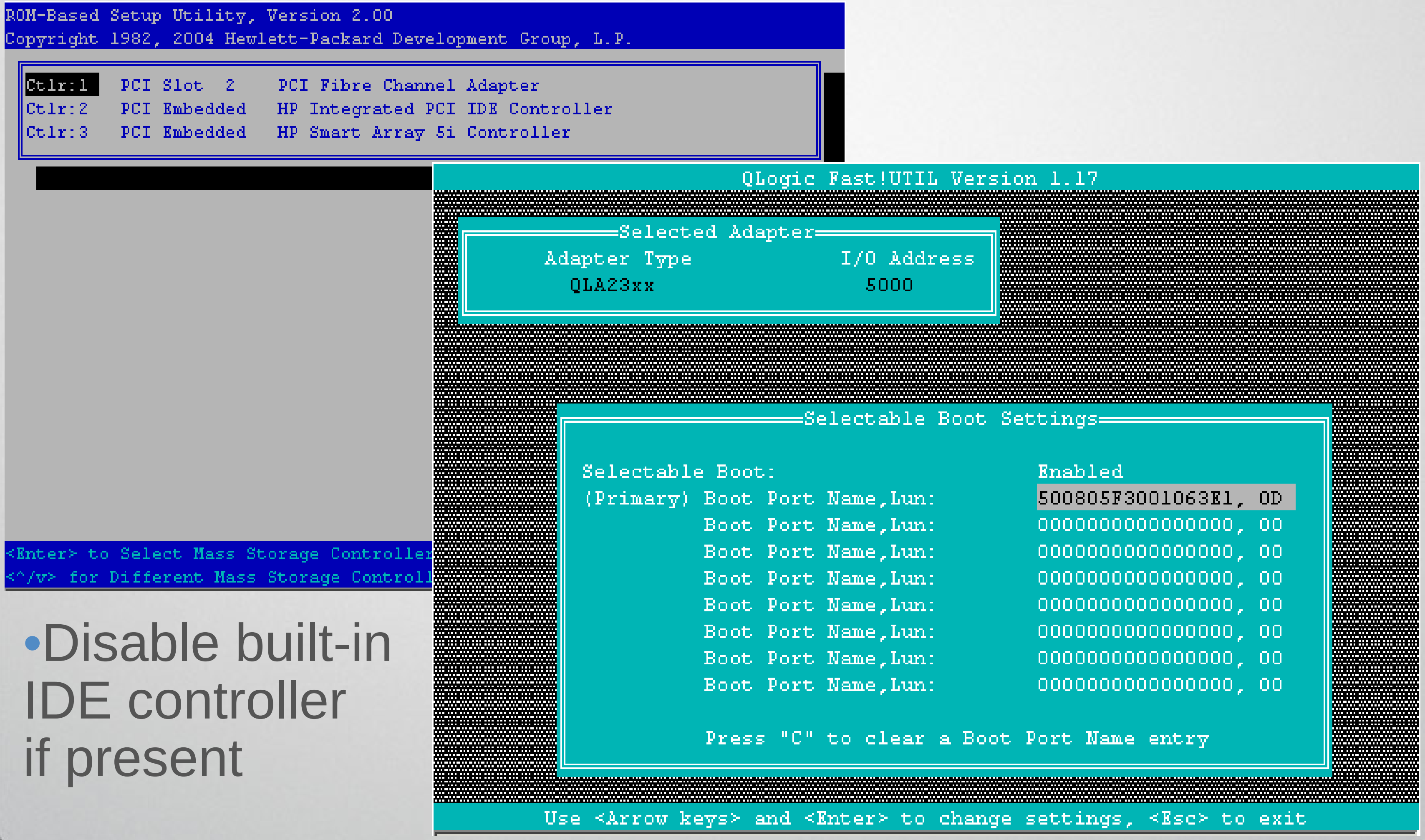
- Server BIOS must designate Fibre Channel card as boot controller
- Fibre Channel card must be configured to initiate a “primitive” connection to the target boot LUN
- Fibre Channel card must be shared between Service Console and VMkernel

Boot from SAN- FC-HBA Setup



```
                    QLogic Fast!UTIL Version 1.17

 ═══════Selected Adapter═════════
      Adapter Type            I/O Address
        QLA23xx                  5000

                  ═══════Host Adapter Settings═══════

                    BIOS Address:               CD800
                    BIOS Revision:              1.26
                    Adapter Serial Number:      Q51563
                    Interrupt Level:            11
                    Adapter Port Name:          210000E08B196B33
                    Host Adapter BIOS:          Enabled
                    Frame Size:                 2048
                    Loop Reset Delay:           5
                    Adapter Hard Loop ID:       Disabled
                    Hard Loop ID:               0
                    Spinup Delay:               Disabled

        Use <Arrow keys> and <Enter> to change settings, <Esc> to exit
```

Set FC-HBA and LUN as Boot Device



```
ROM-Based Setup Utility, Version 2.00
Copyright 1982, 2004 Hewlett-Packard Development Group, L.P.

Ctlr:1    PCI Slot  2    PCI Fibre Channel Adapter
Ctlr:2    PCI Embedded   HP Integrated PCI IDE Controller
Ctlr:3    PCI Embedded   HP Smart Array 5i Controller

<Enter> to Select Mass Storage Controller
<^/v> for Different Mass Storage Controll
```

```
                    QLogic Fast!UTIL Version 1.17

          Selected Adapter
   Adapter Type              I/O Address
     QLA23xx                    5000

                    Selectable Boot Settings

   Selectable Boot:                          Enabled
   (Primary) Boot Port Name,Lun:             500805F3001063E1, 0D
             Boot Port Name,Lun:             0000000000000000, 00
             Boot Port Name,Lun:             0000000000000000, 00
             Boot Port Name,Lun:             0000000000000000, 00
             Boot Port Name,Lun:             0000000000000000, 00
             Boot Port Name,Lun:             0000000000000000, 00
             Boot Port Name,Lun:             0000000000000000, 00
             Boot Port Name,Lun:             0000000000000000, 00

             Press "C" to clear a Boot Port Name entry

        Use <Arrow keys> and <Enter> to change settings, <Esc> to exit
```

- Disable built-in IDE controller if present

Resources…

- Storage vendor’s web site
- ESX Server SAN Configuration Guide
- ESX Server SAN Compatibility Guide
- ESX Server Backup Compatibility Guide
- ESX Server Administrators Guide
- ESX Server Raw Device Mapping paper
- ESX Server Performance white papers

This presentation covers the current versions of our products. Details about future releases of our products are available in select sessions at VMworld, including:

- **PAC879:** “The Next Phase of Virtual Infrastructure: Introducing ESX Server 3.0 and VirtualCenter 2.0”
- **PAC177:** “Distributed Availability Services Architecture”
- **PAC484:** “Consolidated Backup with ESX Server: In-Depth Review”
- **PAC485:** “Managing Data Center Resources Using the VirtualCenter Distributed Resource Scheduler”

Thank You!

Backup Slides

Command Line Path Management

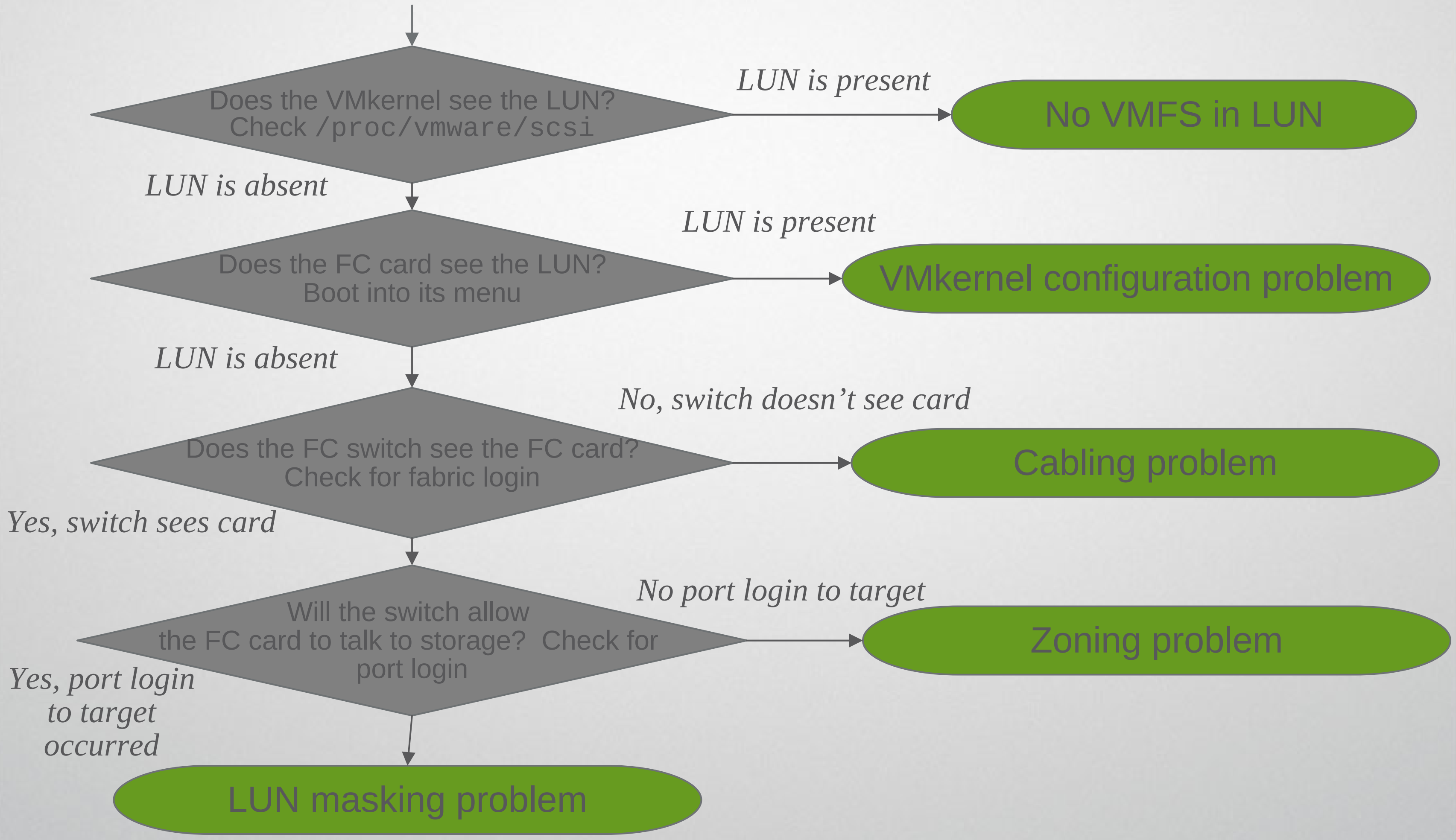
- `wwpn.pl` : show HBA or target (storage processor) on active path
  - `wwpn.pl` : View adapter (HBA) only, on the active path
  - `wwpn.pl –v` (or `wwpn.pl –d`): View adapter, storage processors and related ports on the active path
  - `wwpn.pl –t` : View storage processors and related ports on the active path.
- `vmkmultipath` : multipath maintenance
  - `vmkmultipath –q` : View the current multipathing configuration
  - `vmkmultipath –s –p` *policy* : Set multipathing policy
  - `vmkmultipath –s –r` *path* : Set active path
  - `vmkmultipath –S` : Save configuration

Path Management

- Multipathing allows continuous availability of a SAN LUN in the event of a hardware failure
  - Administrator may set preferred paths for each LUN
- ESX Server supports failover with any supported HBAs
  - Failover occurs automatically, with a HBA configurable delay
- Do not attempt to combine ESX Server failover with other multipathing solutions, as other software or hardware multipathing will conflict with the VMkernel
- Use zones to enforce access from your ESX Server to your disk array
- Choose the right failover policy for your disk array

VMworld 2005 virtualize now

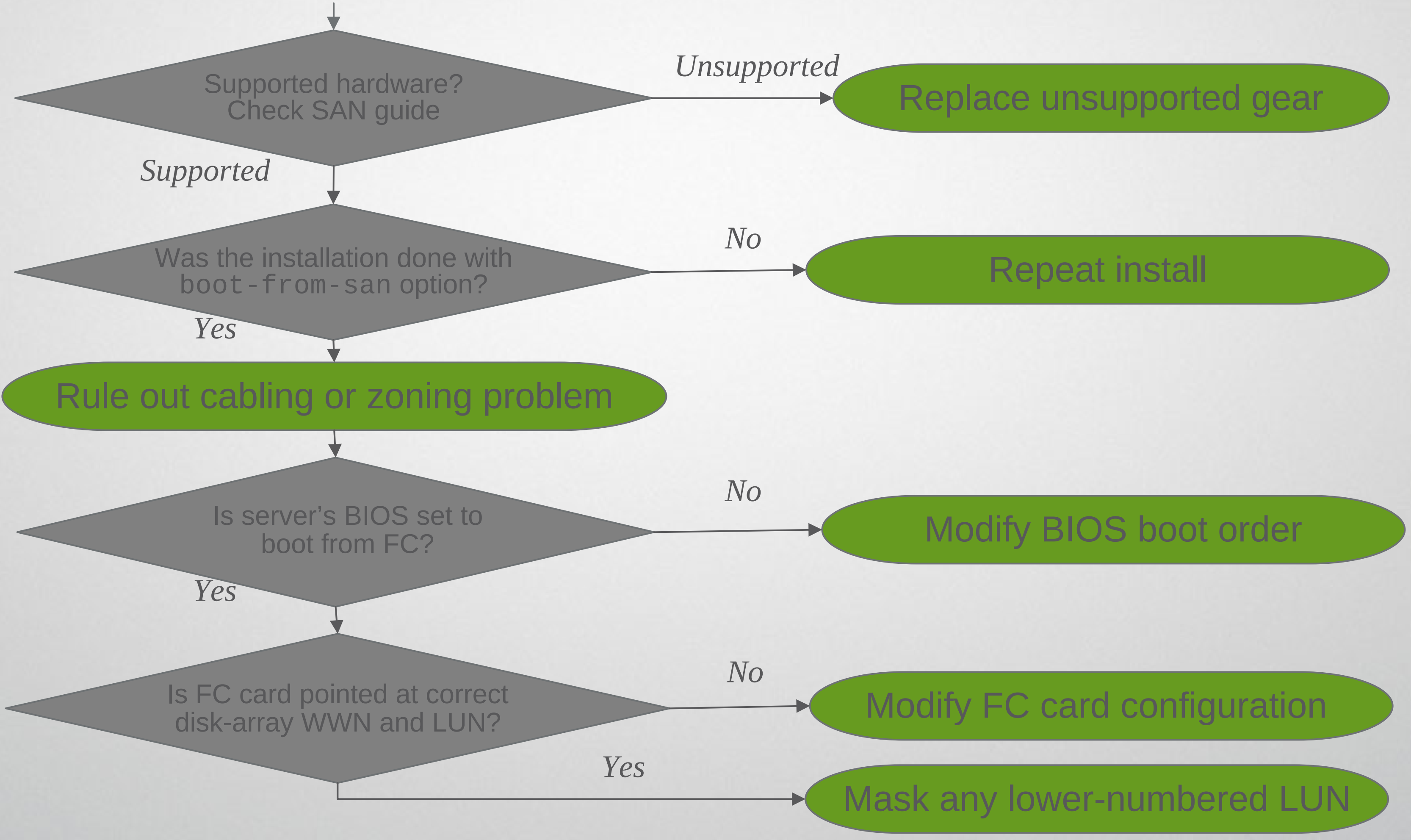
SAN-Based VMFS not Visible



Does the VMkernel see the LUN? Check /proc/vmware/scsi
LUN is present
No VMFS in LUN
LUN is absent
Does the FC card see the LUN? Boot into its menu
LUN is present
VMkernel configuration problem
LUN is absent
Does the FC switch see the FC card? Check for fabric login
No, switch doesn't see card
Cabling problem
Yes, switch sees card
Will the switch allow the FC card to talk to storage? Check for port login
No port login to target
Zoning problem
Yes, port login to target occurred
LUN masking problem

VMworld2005 virtualizenow

Troubleshooting Boot from SAN



Supported hardware? Check SAN guide
Unsupported
Replace unsupported gear
Supported
Was the installation done with boot-from-san option?
No
Repeat install
Yes
Rule out cabling or zoning problem
Is server's BIOS set to boot from FC?
No
Modify BIOS boot order
Yes
Is FC card pointed at correct disk-array WWN and LUN?
No
Modify FC card configuration
Yes
Mask any lower-numbered LUN