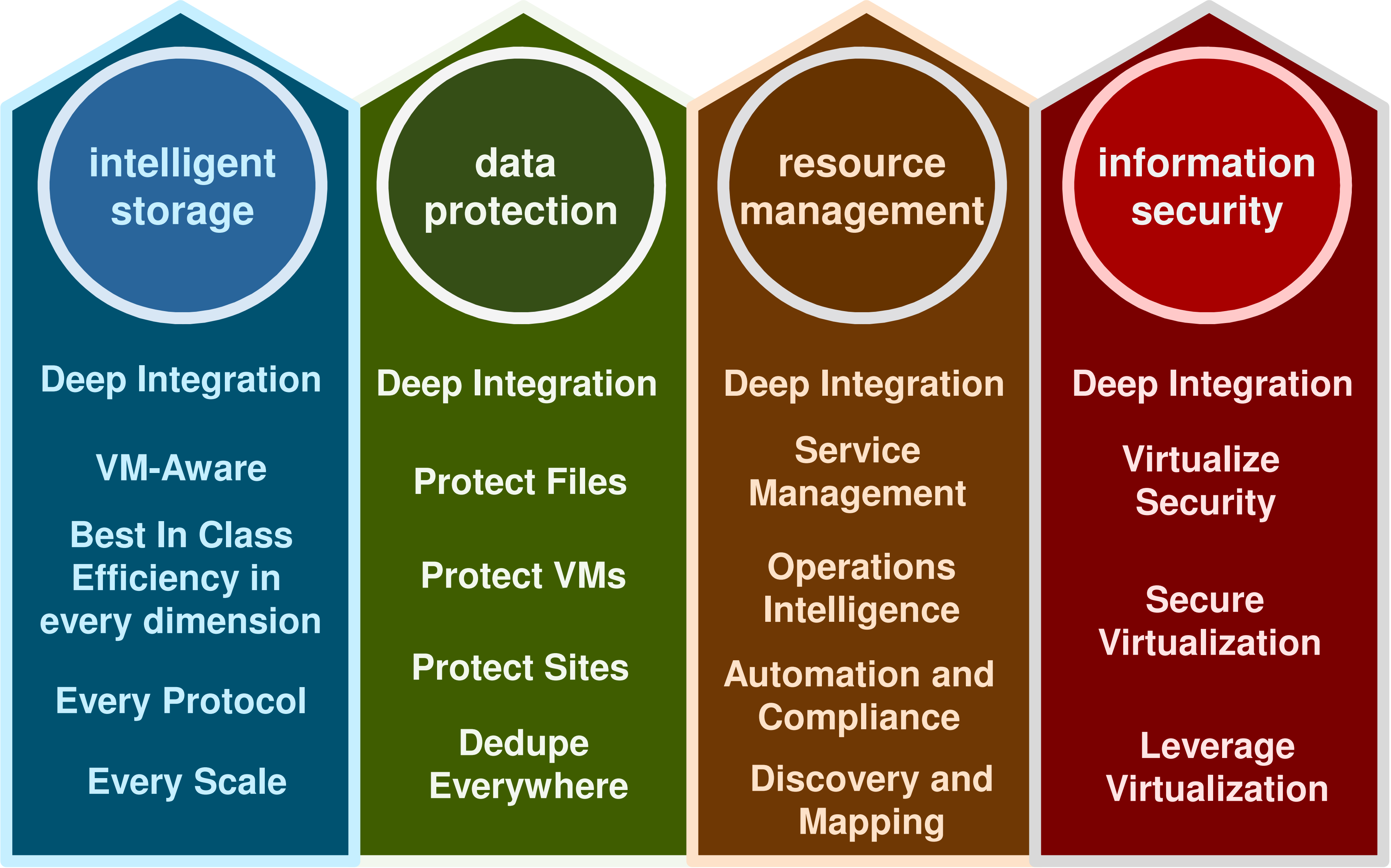
Driving Efficiency. Control. Choice.

Storage Protocol Choices & Storage Best Practices for VMware ESX

Travers Nicholas nicholas_travers@emc.com

VMware Technology Consultant, EMC

EMC – Aligning, Innovating, Integrating



intelligent storage

Deep Integration

VM-Aware

Best In Class Efficiency in every dimension

Every Protocol

Every Scale

data protection

Deep Integration

Protect Files

Protect VMs

Protect Sites

Dedupe Everywhere

resource management

Deep Integration

Service Management

Operations Intelligence

Automation and Compliance

Discovery and Mapping

information security

Deep Integration

Virtualize Security

Secure Virtualization

Leverage Virtualization

EMC – Aligning, Innovating, Integrating

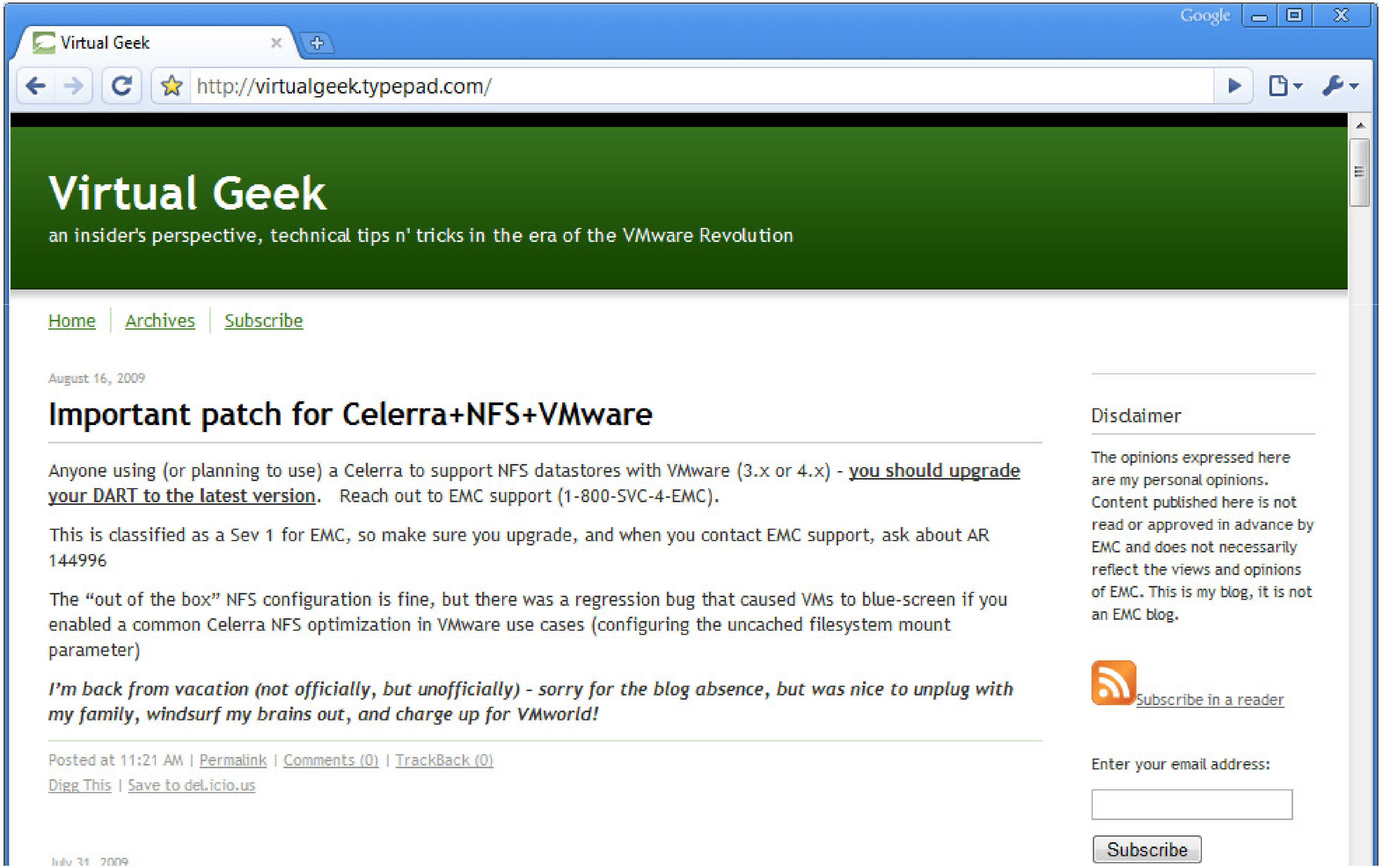
Store (EMC)

Protect (EMC)

Manage (EMC Ionix)

Secure (RSA)

Virtual Geek - Chad Sakac’s Blog



Virtual Geek

an insider's perspective, technical tips n' tricks in the era of the VMware Revolution

Home | Archives | Subscribe

August 16, 2009

Important patch for Celerra+NFS+VMware

Anyone using (or planning to use) a Celerra to support NFS datastores with VMware (3.x or 4.x) - **you should upgrade your DART to the latest version**. Reach out to EMC support (1-800-SVC-4-EMC).

This is classified as a Sev 1 for EMC, so make sure you upgrade, and when you contact EMC support, ask about AR 144996

The “out of the box” NFS configuration is fine, but there was a regression bug that caused VMs to blue-screen if you enabled a common Celerra NFS optimization in VMware use cases (configuring the uncached filesystem mount parameter)

I’m back from vacation (not officially, but unofficially) - sorry for the blog absence, but was nice to unplug with my family, windsurf my brains out, and charge up for VMworld!

Posted at 11:21 AM | Permalink | Comments (0) | TrackBack (0)
Digg This | Save to del.icio.us

July 31, 2009

Disclaimer

The opinions expressed here are my personal opinions. Content published here is not read or approved in advance by EMC and does not necessarily reflect the views and opinions of EMC. This is my blog, it is not an EMC blog.

Subscribe in a reader

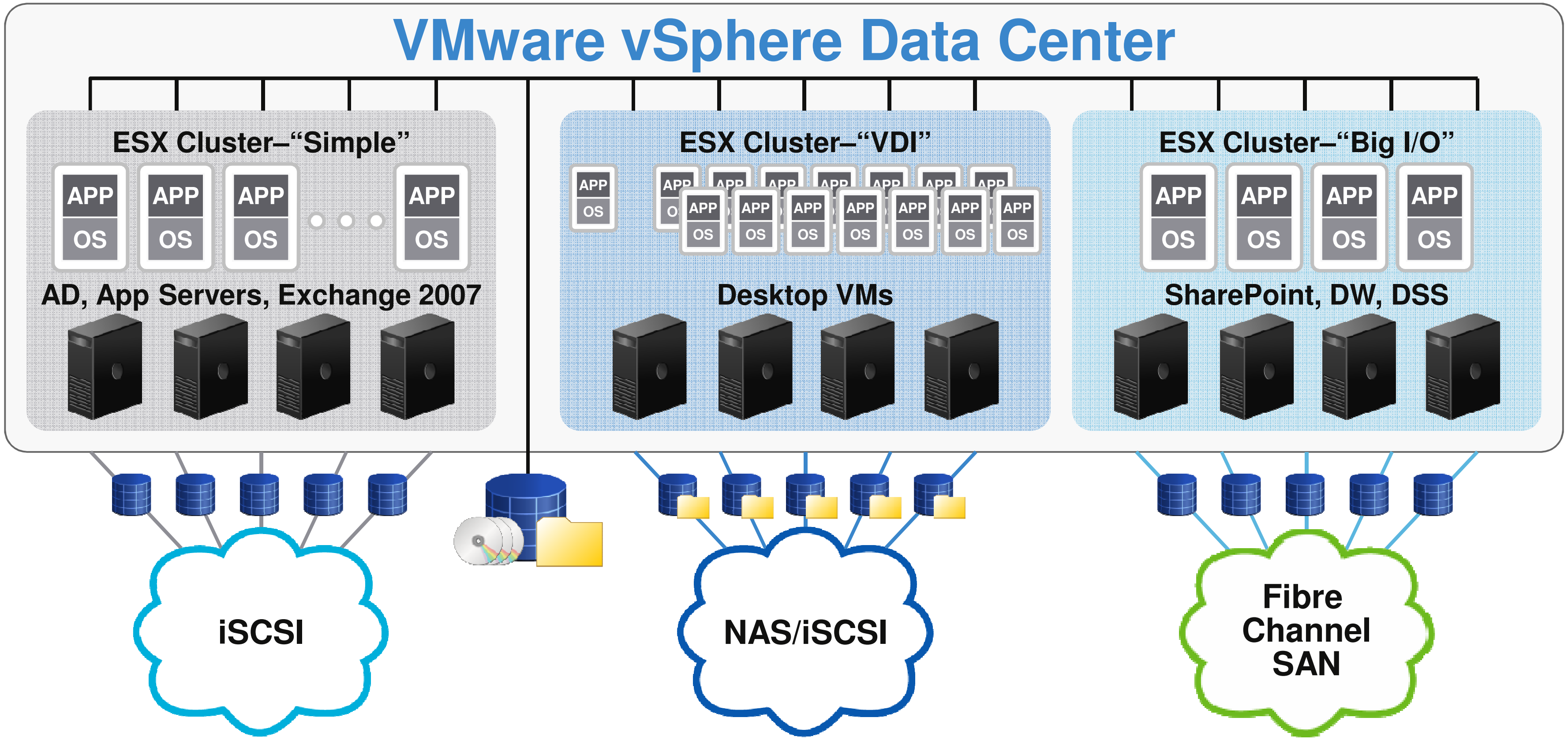
Enter your email address:

Subscribe

Storage Considerations for VMware ESX

- Standard LUN Size?
- How many VMs per LUN?
- NFS, RDM or VMFS?
- iSCSI, or FC?
- How do I scale iSCSI and NFS?
- Queue Depths?
- I need a 3TB LUN... How?
- Why do I need multi-pathing?
- I need SRM support... How?

VMware Leverages Multi-protocol Storage



Storage Protocol Selection

Feature	Fibre Channel SAN	iSCSI	NFS
ESX boot	Yes	Hardware initiator	No
Virtual machine boot	Yes	Yes	Yes
Raw device mapping	Yes	Yes	N/A
LUN extension	Yes	Yes	Yes
Replication Mgr Support	Yes	Yes	No (’09-Q3)
vCenter Site Recovery Manager	Yes	Yes	No (‘09-Q3)
Virtual machine as initiator	No	Yes	No
Security	N/A	CHAP	UNIX_Auth
Storage granularity	LUN	LUN	VM, files in VM

The choice of connectivity with VMware ESX is largely driven by application requirements and preference

Measuring Storage Performance

1. IOs per second (or commands/sec, or IOPS)
2. Response time (latency generally measured in latency)
3. Throughput (bandwidth, megabytes/sec)

Bet the Business IP Networking for Storage

- Separate storage and network traffic on different ports
- Use Cat6 cabling rather than Cat5/5e
- Enable Flow-Control (should be set to receive on switches and transmit on iSCSI targets)
- Enable spanning tree protocol with either RSTP or portfast enabled
- Filter / restrict bridge protocol data units on storage network ports
- Configure jumbo frames end-to-end (support added in 3.5U3+)
- Ensure Ethernet switches have the proper amount of port buffers and other internals to support iSCSI and NFS traffic optimally

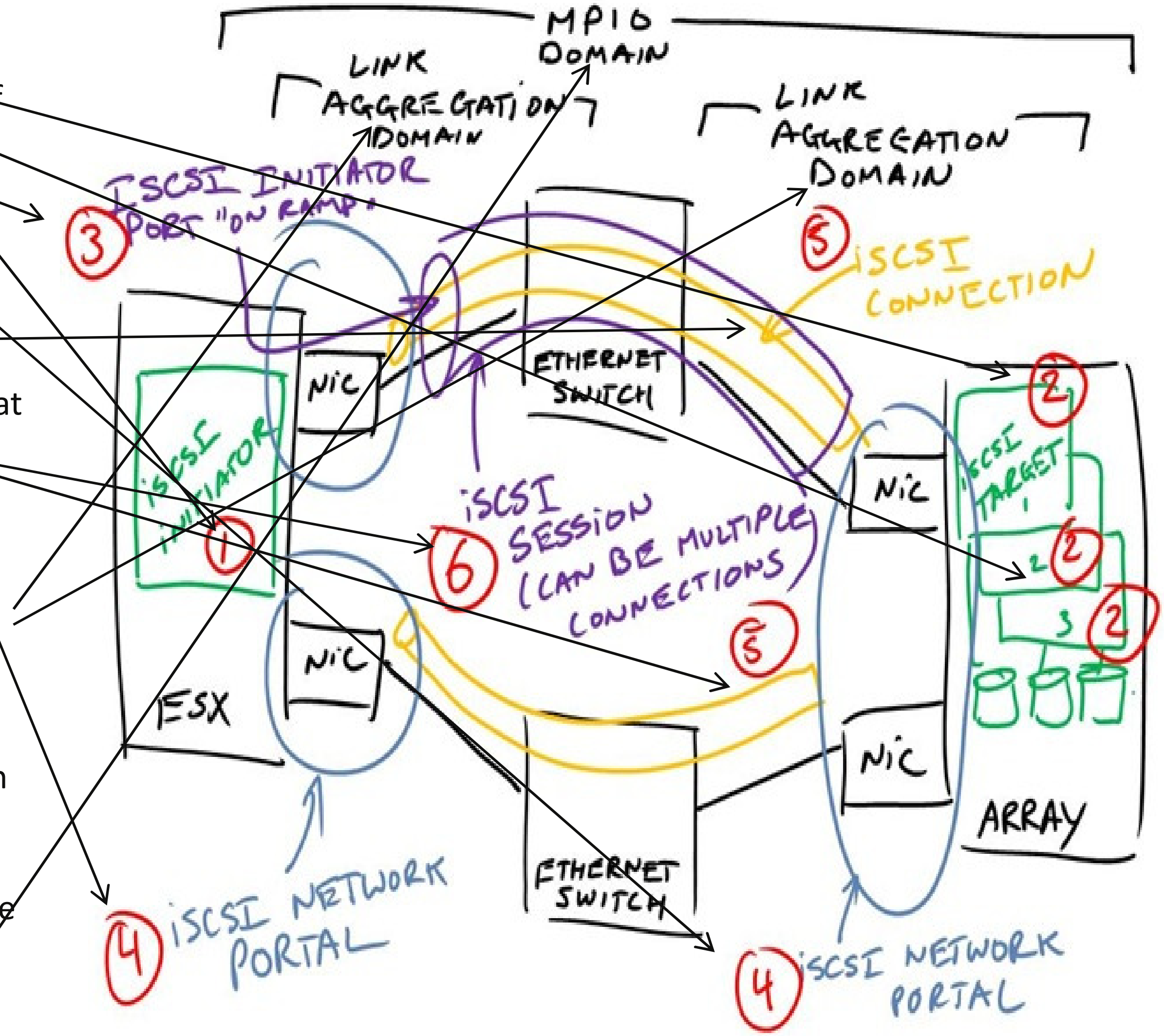
iSCSI Storage Fundamentals (ESX 3.5)

1. **Initiator** = an iSCSI client
2. **Target** = an iSCSI server
3. **Initiator port** = the "ramp" for data of an iSCSI session
4. **Network portal** = an IP address(es) used by initiator or target
5. **Connection** = carries control info, SCSI commands, and data being read or written
6. **Session** = one or more connections that form an initiator-target session

Link Aggregation = aggregates physical links to transmit a given connection

Multiple Connections per Session (MC/S) = multiple connections within a single session to an iSCSI target

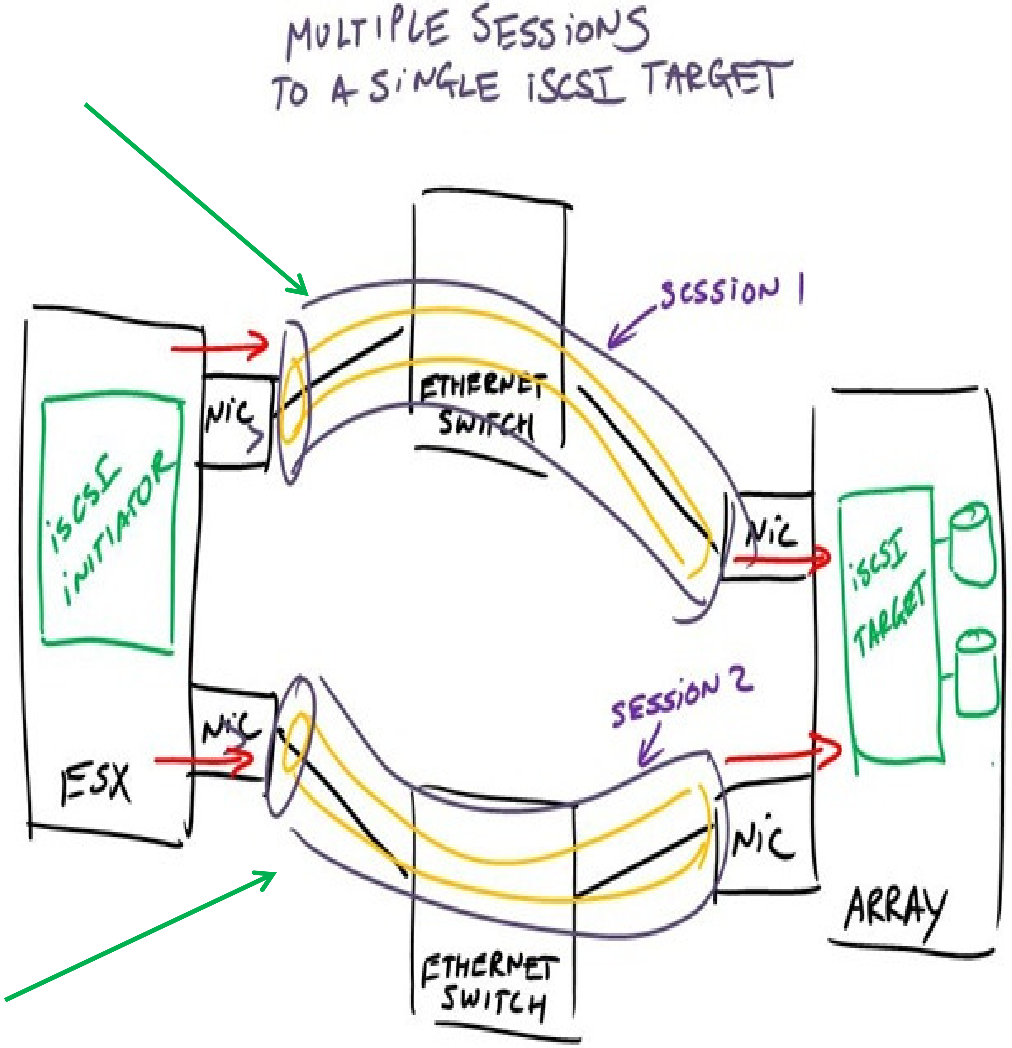
MPIO = multiple sessions, each with one or more connections to any iSCSI target



iSCSI in VMware ESX 4.0

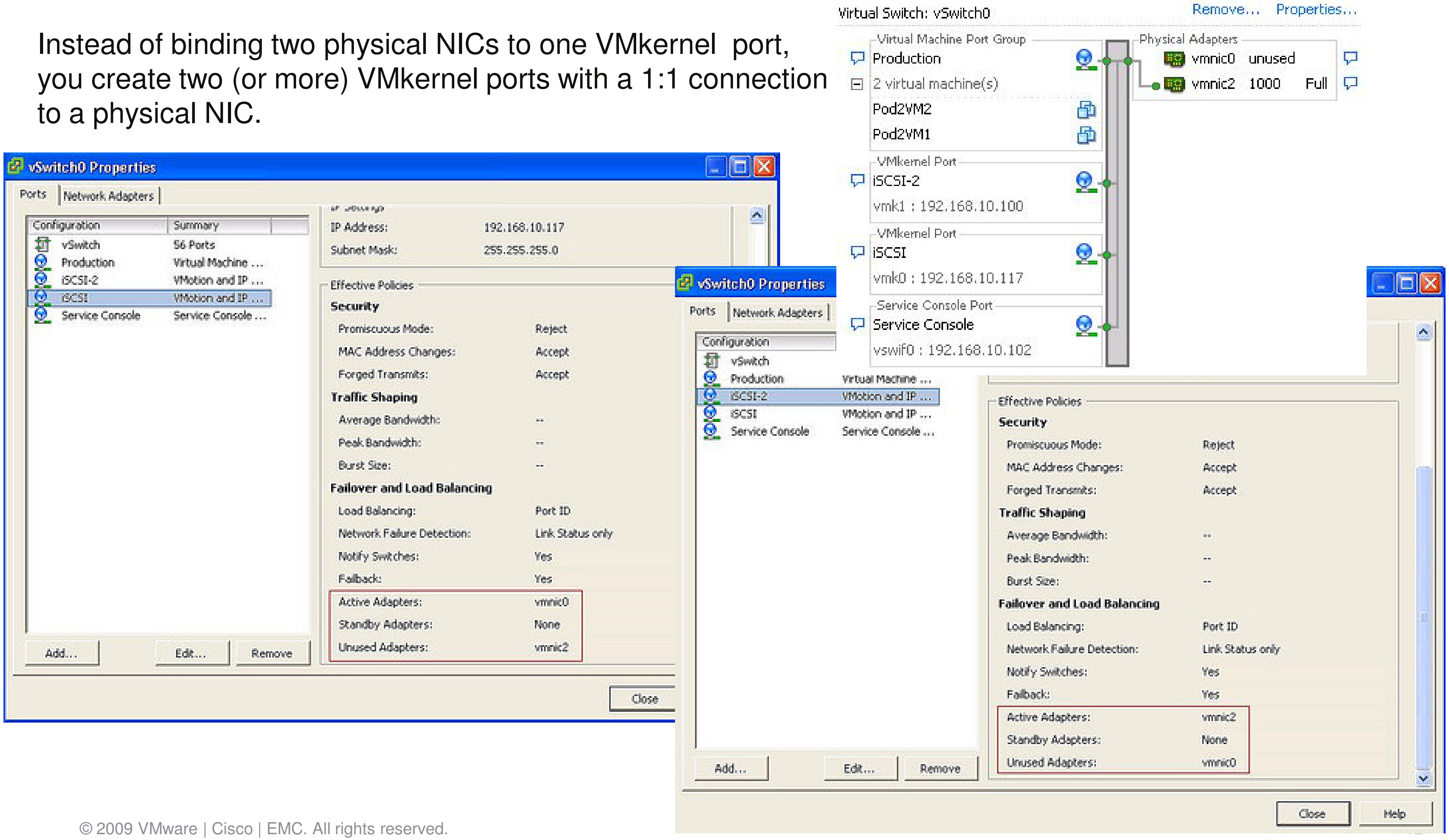
The ESX 4.x software iSCSI initiator can only establish one connection per session established to each target, but can now establish multiple sessions per iSCSI target.

Multiple sessions means multiple “on-ramps” for MPIO. ESX 4.x also brings core multi-pathing improvements in the vStorage initiative: NMP round robin, ALUA support, and EMC PowerPath.



iSCSI Multi-pathing in ESX 4.0

Instead of binding two physical NICs to one VMkernel port, you create two (or more) VMkernel ports with a 1:1 connection to a physical NIC.



Virtual Switch: vSwitch0 — Remove... Properties...

Virtual Machine Port Group: Production — 2 virtual machine(s): Pod2VM2, Pod2VM1

Physical Adapters:

vmnic0	unused	
vmnic2	1000	Full

VMkernel Port: iSCSI-2 — vmk1 : 192.168.10.100

VMkernel Port: iSCSI — vmk0 : 192.168.10.117

Service Console Port: Service Console — vswif0 : 192.168.10.102

vSwitch0 Properties (iSCSI selected)

Configuration	Summary
vSwitch	56 Ports
Production	Virtual Machine ...
iSCSI-2	VMotion and IP ...
iSCSI	VMotion and IP ...
Service Console	Service Console ...

IP Address: 192.168.10.117
Subnet Mask: 255.255.255.0

Effective Policies

Security
- Promiscuous Mode: Reject
- MAC Address Changes: Accept
- Forged Transmits: Accept

Traffic Shaping
- Average Bandwidth: --
- Peak Bandwidth: --
- Burst Size: --

Failover and Load Balancing
- Load Balancing: Port ID
- Network Failure Detection: Link Status only
- Notify Switches: Yes
- Failback: Yes
- Active Adapters: vmnic0
- Standby Adapters: None
- Unused Adapters: vmnic2

vSwitch0 Properties (iSCSI-2 selected)

Configuration	Summary
vSwitch	
Production	Virtual Machine ...
iSCSI-2	VMotion and IP ...
iSCSI	VMotion and IP ...
Service Console	Service Console ...

Effective Policies

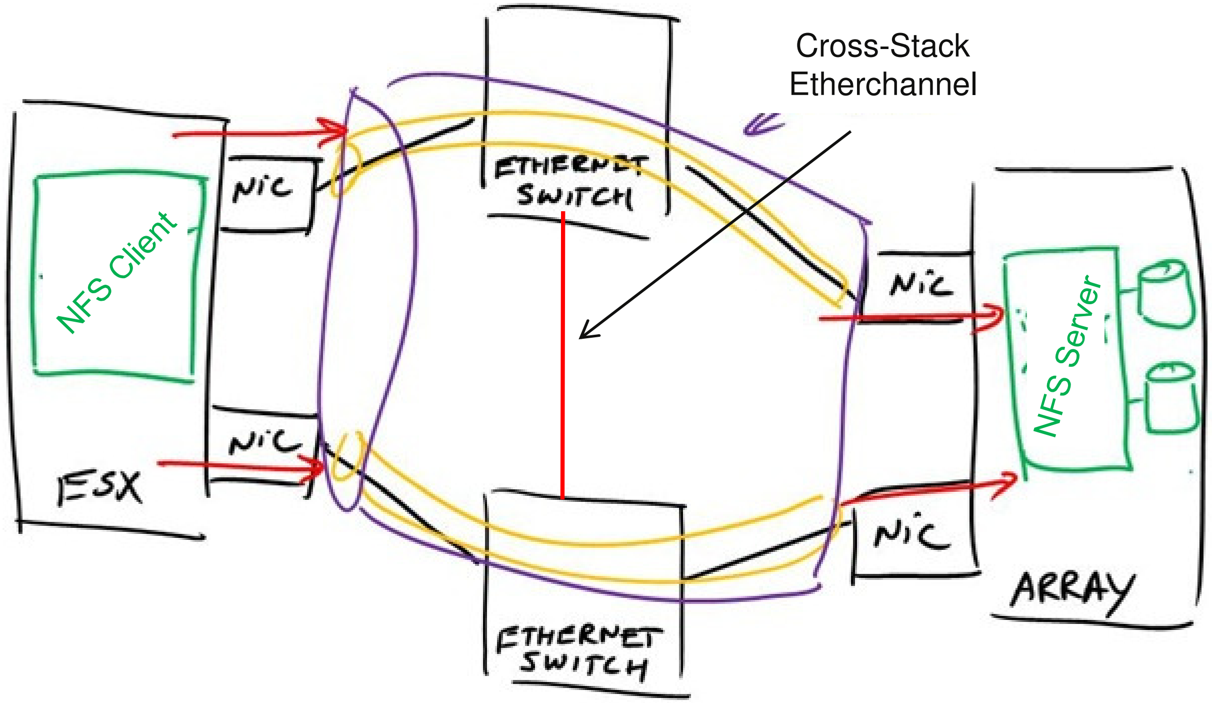
Security
- Promiscuous Mode: Reject
- MAC Address Changes: Accept
- Forged Transmits: Accept

Traffic Shaping
- Average Bandwidth: --
- Peak Bandwidth: --
- Burst Size: --

Failover and Load Balancing
- Load Balancing: Port ID
- Network Failure Detection: Link Status only
- Notify Switches: Yes
- Failback: Yes
- Active Adapters: vmnic2
- Standby Adapters: None
- Unused Adapters: vmnic0

NFS Fundamentals (ESX 3.5 & ESX 4.0)

1. Maximum of two (2) TCP sessions per datastore using an NFS mount
   a. One (1) control session (1% of bandwidth)
   b. One (1) data transmission session (99% of bandwidth)
2. To ensure High Availability (HA) and adequate bandwidth, two (2) options are available:
   a. IP Subnetting
   b. Link Aggregation (Etherchannel) as shown below
3. Scale out the number of datastores and properly distribute virtual machines to get around the maximum number of TCP sessions per datastore (or move to 10GbE)



NFS Configuration Best Practice

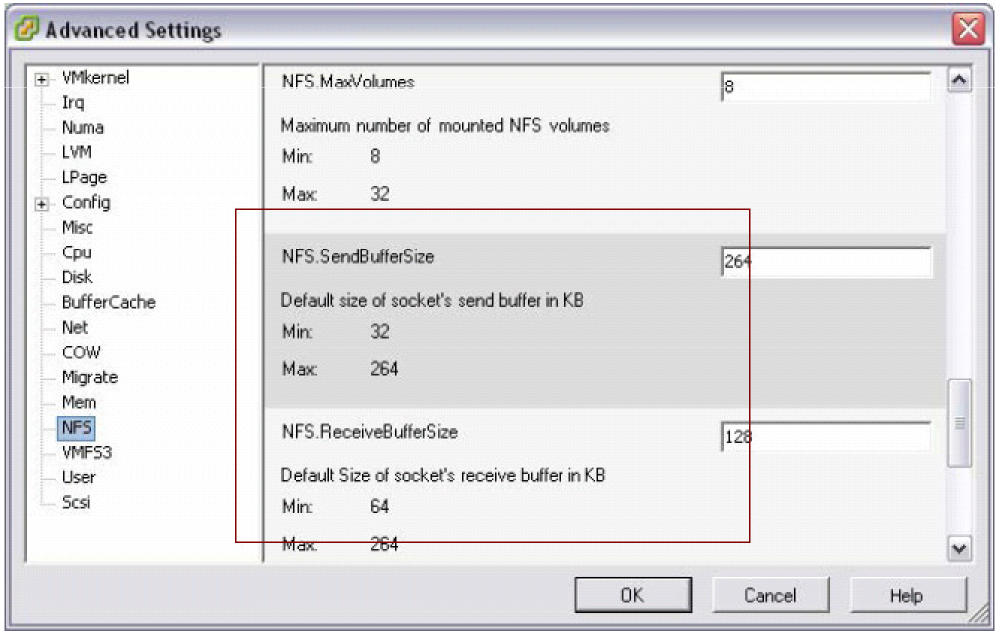
Celerra:

- Enable the **uncached write mechanism** for all file systems (30% improvement):
- Disable the **prefetch read mechanism** for file systems consisting of VMs with small random accesses patterns:

ESX:

- Set **NFS.MaxVolumes** to 32 (64 in ESX 4.0)
- Set **SendBufferSize** to 64
- Set **ReceiveBufferSize** to 64
- Set **Net.TcpipHeapSize** 30 MB
- Set **NFS.HeartbeatFrequency** to 12
- Set **NFS.HeartbeatTimeout** to 5
- Set **NFS.HeartbeatMaxFailures** to 10

See VMware ESX Server Optimization with EMC® Celerra® Performance Study - Technical Note P/N 300-006-724 for additional information



NFS.MaxVolumes 8

Maximum number of mounted NFS volumes

Min: 8

Max: 32

NFS.SendBufferSize 264

Default size of socket's send buffer in KB

Min: 32

Max: 264

NFS.ReceiveBufferSize 128

Default Size of socket's receive buffer in KB

Min: 64

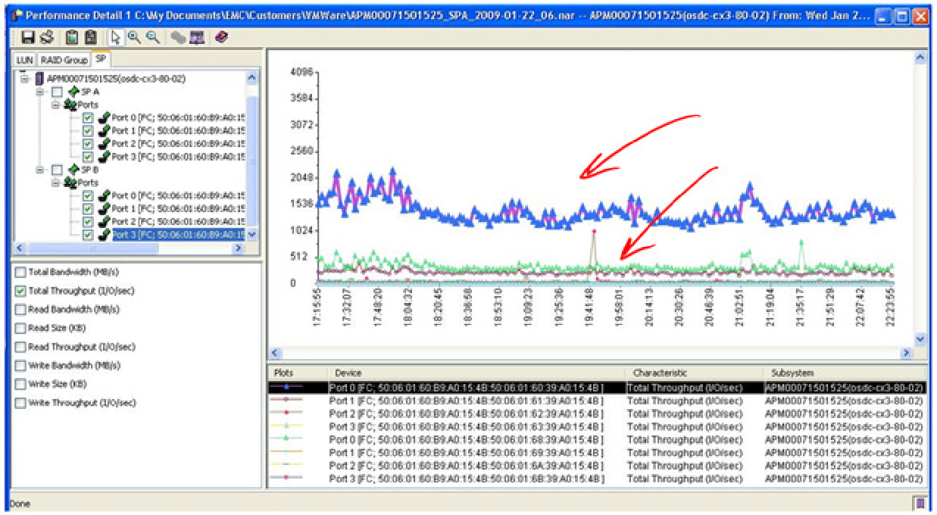
Max: 264

Active/Passive Arrays & MRU

ISSUE: When statically load-balanced configurations are used, every time the ESX host boots, the LUNs default their active path to the first enumerated path. This means that after a while, and you've run some VUM remediations, you'll find one of your array ports abnormally busy (e.g., SPA0). Further, if not correctly zoned, all LUNs will be on a single SP.

RESOLUTION: Ensure that each HBA is zoned correctly to both SPs. Configure the multi-pathing based on the VML and not the vmhba identifier (which can be done at CLI using esxcfg-mpath). The VML to LUN relationship is shown in the /vmfs/devices/disks directory, just do a ls -l.

http://communities.vmware.com/message/598649



Performance Detail 1 C:\My Documents\EMC\Customers\VMWare\APM00071501525_SPA_2009-01-22_06.nar -- APM00071501525(osdc-cx3-80-02) From: Wed Jan 2...

LUN | RAID Group | SP

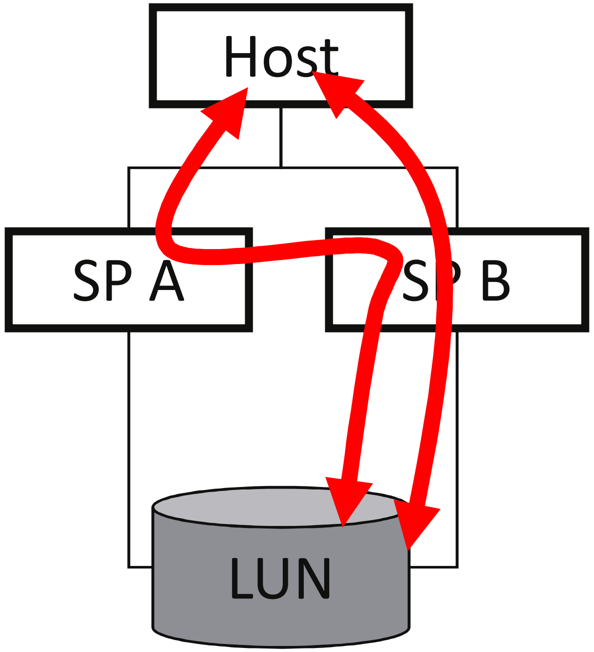
- APM00071501525(osdc-cx3-80-02)
  - □ SP A
    - Ports
      - ☑ Port 0 [FC; 50:06:01:60:B9:A0:15
      - ☑ Port 1 [FC; 50:06:01:60:B9:A0:15
      - ☑ Port 2 [FC; 50:06:01:60:B9:A0:15
      - ☑ Port 3 [FC; 50:06:01:60:B9:A0:15
  - □ SP B
    - Ports
      - ☑ Port 0 [FC; 50:06:01:60:B9:A0:15
      - ☑ Port 1 [FC; 50:06:01:60:B9:A0:15
      - ☑ Port 2 [FC; 50:06:01:60:B9:A0:15
      - ☑ Port 3 [FC; 50:06:01:60:B9:A0:15

□ Total Bandwidth (MB/s)
☑ Total Throughput (I/O/sec)
□ Read Bandwidth (MB/s)
□ Read Size (KB)
□ Read Throughput (I/O/sec)
□ Write Bandwidth (MB/s)
□ Write Size (KB)
□ Write Throughput (I/O/sec)

Plots	Device	Characteristic	Subsystem
	Port 0 [FC; 50:06:01:60:B9:A0:15:4B:50:06:01:60:39:A0:15:4B]	Total Throughput (I/O/sec)	APM00071501525(osdc-cx3-80-02)
	Port 1 [FC; 50:06:01:60:B9:A0:15:4B:50:06:01:61:39:A0:15:4B]	Total Throughput (I/O/sec)	APM00071501525(osdc-cx3-80-02)
	Port 2 [FC; 50:06:01:60:B9:A0:15:4B:50:06:01:62:39:A0:15:4B]	Total Throughput (I/O/sec)	APM00071501525(osdc-cx3-80-02)
	Port 3 [FC; 50:06:01:60:B9:A0:15:4B:50:06:01:63:39:A0:15:4B]	Total Throughput (I/O/sec)	APM00071501525(osdc-cx3-80-02)
	Port 0 [FC; 50:06:01:60:B9:A0:15:4B:50:06:01:68:39:A0:15:4B]	Total Throughput (I/O/sec)	APM00071501525(osdc-cx3-80-02)
	Port 1 [FC; 50:06:01:60:B9:A0:15:4B:50:06:01:69:39:A0:15:4B]	Total Throughput (I/O/sec)	APM00071501525(osdc-cx3-80-02)
	Port 2 [FC; 50:06:01:60:B9:A0:15:4B:50:06:01:6A:39:A0:15:4B]	Total Throughput (I/O/sec)	APM00071501525(osdc-cx3-80-02)
	Port 3 [FC; 50:06:01:60:B9:A0:15:4B:50:06:01:6B:39:A0:15:4B]	Total Throughput (I/O/sec)	APM00071501525(osdc-cx3-80-02)

Done

Asymmetric Logical Unit Access (ALUA)



- I/O is accepted on all ports
- All I/O for a LUN is serviced by its owning storage processor
- I/O received on the non-owning storage processor is forwarded to the owning storage processor for servicing (CLARiiON CMI)
- I/O to the owning storage processor is higher performance than I/O to the non-owning storage processor
- ALUA is not supported in ESX 3.5, but is in supported by ESX 4.x
- PowerPath makes ALUA work on all CLARiiON CX-series arrays, and also balances ALUA to not hammer the weaker paths

Common Production Design Questions

- How do I design for big workloads?
- How many IOPS per datastore?

Consider this data from a real world customer:

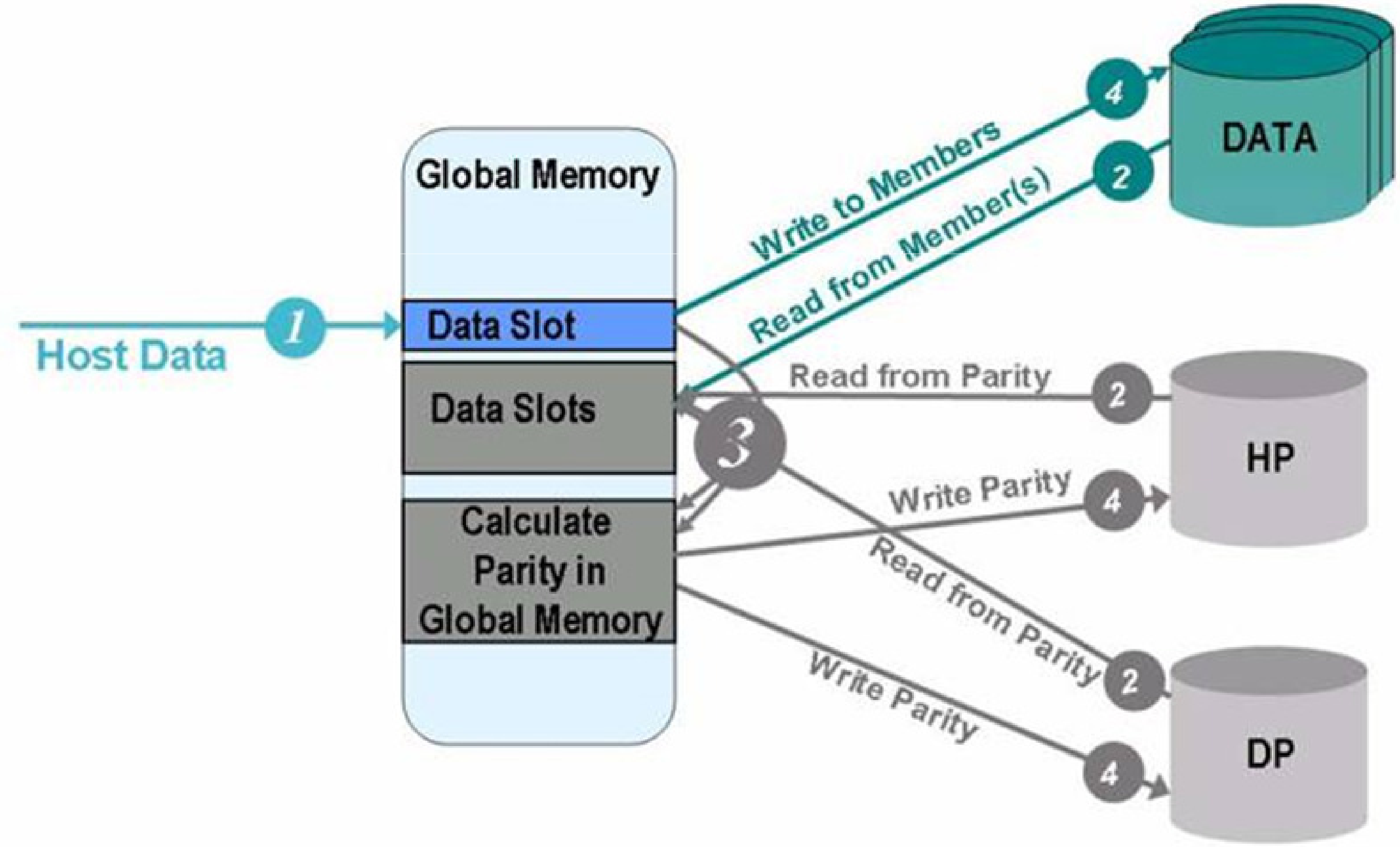
- Boot LUN only
- 175 VMs = 1326 IOPS
- 7.6 IOs per VM
- 35% Read Hit (successfully pre-fetched from to cache)
- 15% Read Miss (random read, straight to disk)
- 50% Write Hit (writes almost always go to cache first)

Designing for 100 Virtual Machines

- 7.6 IOs per VM = 7,600 IOs for 1,000 virtual machines
  - 35% Read Hit = 0 back-end reads
  - 15% Random Read (miss) = 1,140 back end reads
  - 50% Random Write = 3,800 front end writes (writes to cache)

Converting Front-end Writes to Back-end IO

- RAID 1 = 2 IOPS, RAID 5 = 4 IOPS, RAID 6 = 6 IOPS



Calculating Totals

- 3,800 front-end writes equates to:

RAID	Multiplier	Total IOs
1	2	7,600
5	4	15,200
6	6	22,800

- Add back end reads (1,140)

RAID	Writes	Total IOs
1	7,600	8,740
5	15,200	16,340
6	22,800	23,940

Different Disks for Different IO Profiles

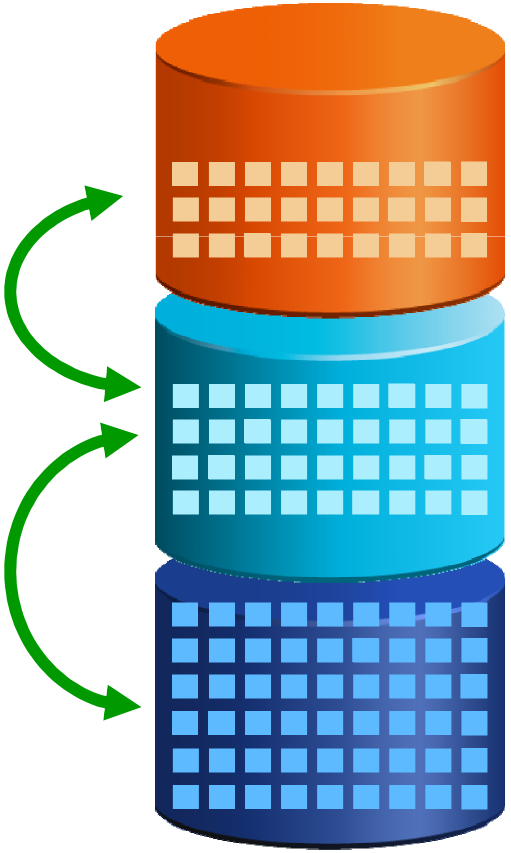
- EFD = 2500+, 15K = 180, 10K = 120, 7.2K = 95
- RAID 5 using 15K Drives means:
  - 16,340 IOPS / 180 = 91 spindles per 1,000 VMs
- RAID 6 using 15K Drives means:
  - 23,940 IOPS / 180 = 133 spindles per 1,000 VMs
- RAID 1 using 15K Drives means:
  - 8,740 IOPS / 180 = 48 spindles per 1,000 VMs

Fully Automated Storage Tiering

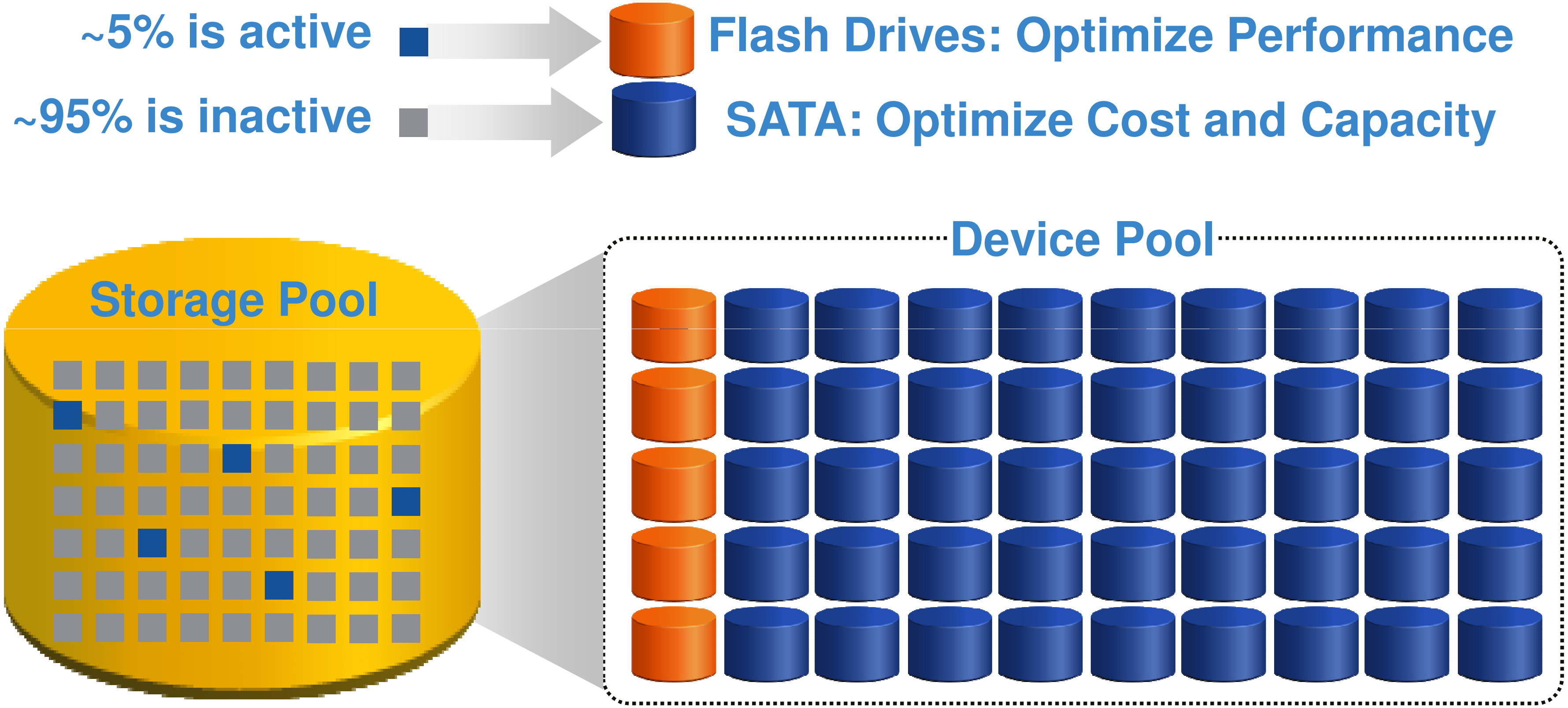
Automates movement and placement of data based on changing needs

- Monitors volumes (LUNs) contained within the storage group
- Identifies candidate LUNs for promotion/demotion moves or swaps
- Configurable via policies
  - Analysis periods
  - Move/swap periods
  - Automatic or Admin approval mode

FAST



Coming Next: FAST v2 - Sub LUN/File Tiering



~5% is active

Flash Drives: Optimize Performance

~95% is inactive

SATA: Optimize Cost and Capacity

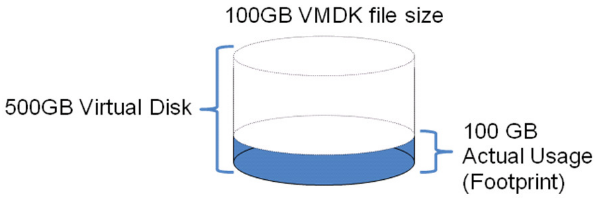
Storage Pool

Device Pool

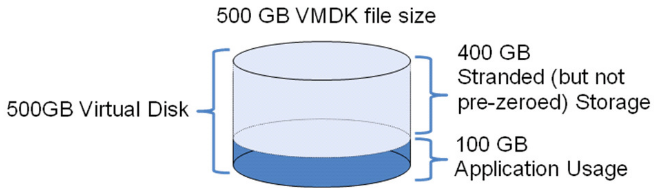
Allows Parts of VMs to Move from Flash Disk to Fibre Channel to SATA with No Performance Impact

Virtual Disk Formats

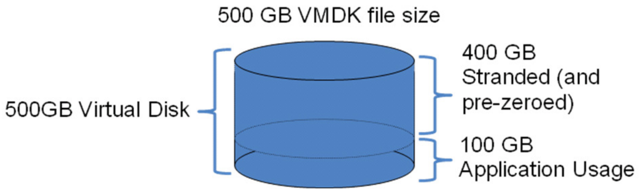
Thin - As I/O occurs in the guest, the vmkernel zeroes out the space needed right before the guest I/O is committed, and grows the VMDK file in 1MB chunks.



Thick (a.k.a., zeroedthick): the size of the VDMK file on the datastore matches the VMDK's total capacity. As I/O occurs in the guest, the vmkernel zeroes out the space needed right before the guest I/O is committed.



Eagerzeroedthick: the size of the VDMK file on the datastore matches the VMDK's total capacity. Within the VMDK file, it is "pre-zeroed" at the time of creation. As I/O occurs in the guest, the vmkernel does not need to zero the blocks prior to the I/O occurring.



Raw Device Mappings (RDMs)

1. Microsoft Windows Clusters (MSCS or WSFC)
2. Storage device must be presented directly to the virtual machine (e.g. CLARiiON Control LUN or Symmetrix gatekeeper)
3. Exchange application-integration (VSS) with storage array snapshot capabilities
4. P to V to P capabilities for databases

Tomorrow: Offload

vStorage APIs for Array Integration (VAAI)

Write Same/Zero

What: 10x less IO for common tasks

How: Eliminating redundant and repetitive write commands – just tell the array to repeat via SCSI commands

Fast/Full Copy

What: 10x faster VM deployment, clone, snapshot, and Storage VMotion

How: leveraging array ability to mass copy, snapshot, and move blocks via SCSI commands

Hardware Offloaded Locking

What: 10x more VMs per datastore

How: stop locking LUNs and start only locking blocks.

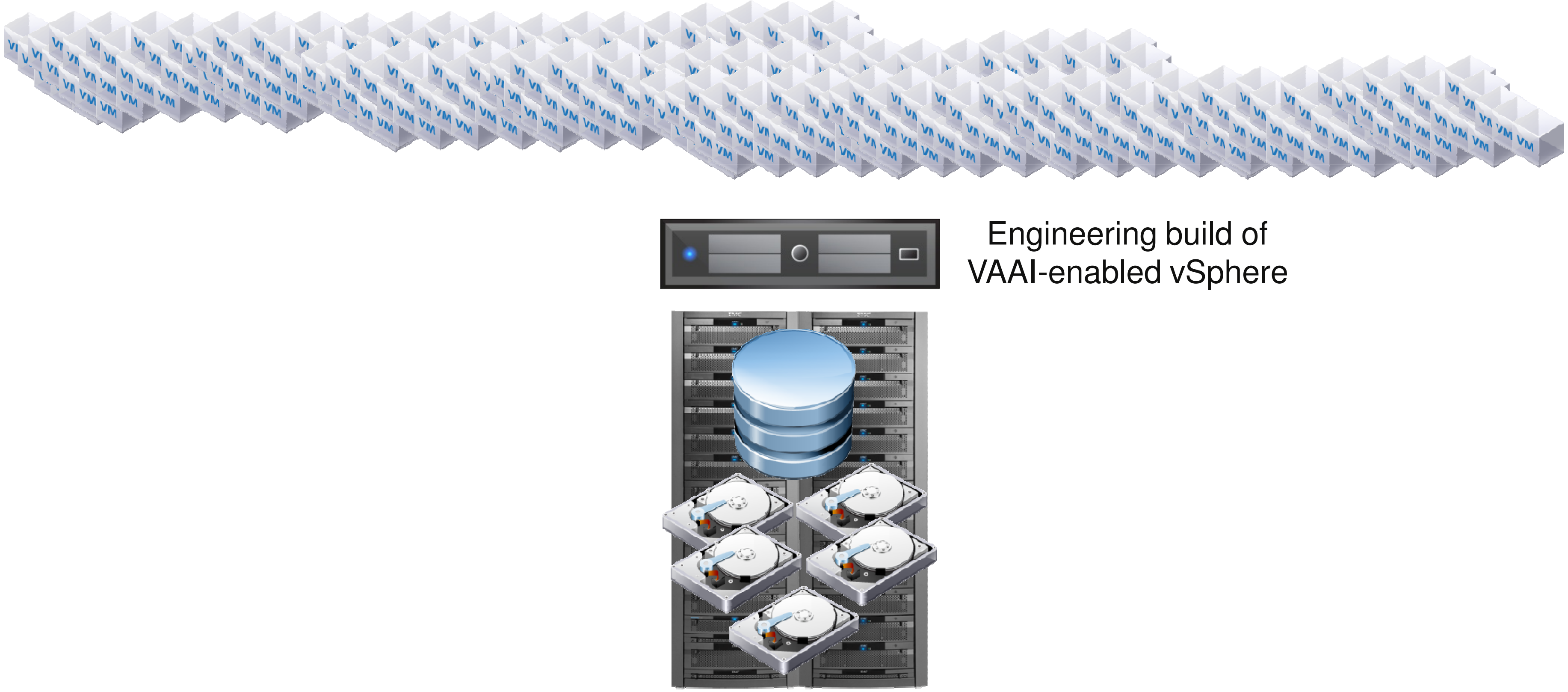
Thin Provisioning Stun

What: Never have an out-of-space disaster

How: reporting array TP state to ESX

Early VAAI Findings

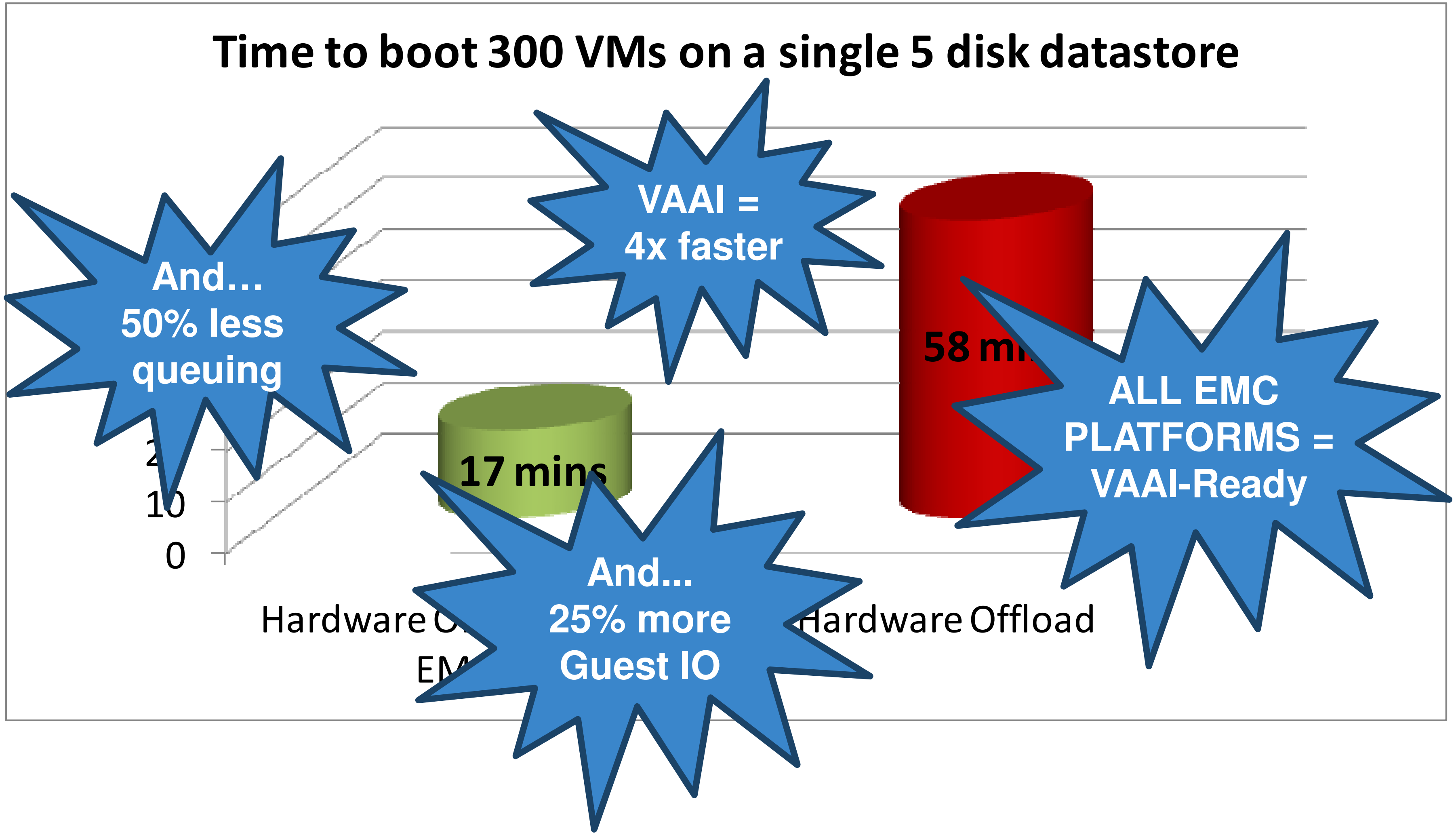
Q: What if you do something ***REALLY out there*** – like using a single VMFS datastore backed by ***only 5 spindles..***
and simultaneously booting 300 VMs on it?



Engineering build of VAAI-enabled vSphere

CX4 w/ engineering build of VAAI-integrated FLARE

Effect of VAAI – Hardware Locking Offload



“Day After Tomorrow”: Invisibility

4 Areas of Key Long Term Collaboration

vStorage + FAST v2 integration

What: "DRS and DPM for Storage"

How: FAST on it's own will auto-tier day 1, working on VM-level policy integrated with vApp policy – dynamically moving VMs between tiers to ***optimize performance, cost, and power.***

Core Storage Stack changes for Storage Virtual Appliances

What: Virtual Storage Appliances – many functions, and large scale

How: pNFS, VMDirectPath IO, changes in the paravirtualized SCSI and vmkernel storage stack

Long-Distance & Cloud VMotion

What: VMotion, between remote datacenters, at the VM-level, and into and out of external clouds

How: Active/active storage virtualization techniques coupled with the ability to access storage as if it were local, ***whether it's not or it's in transit***

vDisk & VM Array Awareness

What: VM-object (and virtual disk) level awareness for block, file, and add object-storage models to vSphere-generation

How: block-list, file-API, and object-storage interfaces in vmkernel, changes in platforms gearing to VM-level function

Storage Protocol Choices & Storage Best Practices for VMware ESX

Travers Nicholas nicholas_travers@emc.com

VMware Technology Consultant, EMC