HP WORLD 2004
Solutions and Technology Conference & Expo

Session 3511: It's 2:00 AM: What Is Your Exchange SAN Storage Doing?

Gary Ketchum, MASE, MCSE, MCDBA
NSS CFT, Storage Consultant
Hewlett-Packard

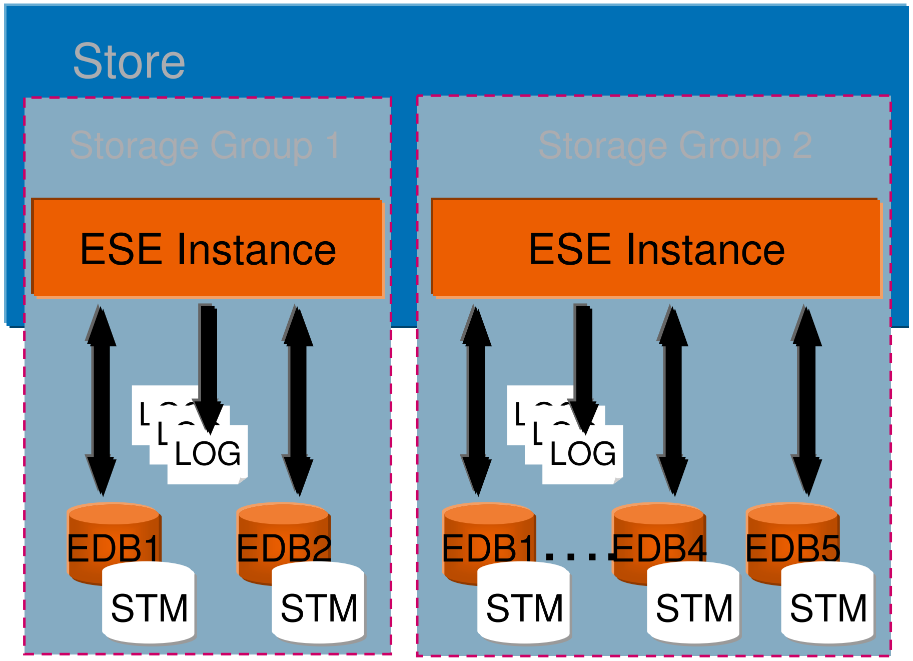
Agenda

- Here's what you may be asking & what you'll learn…

1. Exchange Storage Fundamentals
   - Workload requirements
   - User profiles
2. Design the EVA for Exchange
   - Disk Groups
   - Case Study
   - How to Validate the Storage Subsystem
3. Windows Server 2003 Features
   - VSS
   - StorPort, MPIO

1. Exchange Storage Fundamentals

- Exchange is a transactional database with the following types of files
  - Transaction Logs
    - 1 Set per Storage Group (SG)
    - 4 SG per Server (Enterprise)
  - Database – EDB + STM pair
    - Up to 5 per SG (20 per server)
    - Private or Public
  - Other: SMTP mailroot, tracking, temp conversion, etc.



Knowledge of your planned deployments requirements is essential for designing your EVA

- What are the I/O requirements?
- What applications or user groups will access the subsystem (shared storage)?
- If an application is data transfer-intensive, what is the required transfer rate?
- If it is I/O request-intensive, what is the required response time?
- What is the read/write ratio for a typical request?
- Is the data being stored mission-critical?

I/O Requirement

- Determine your storage IOPS requirement.
  Establish a baseline and user profile and determine your required IOPS.

- Your mileage may vary but a typical MAPI profile is heavier on reads than writes. For example Microsoft internally has a profile of 75% read and 25% write.

- Choice of RAID affects IOPS of LUN (Disk Volume)
  - IOPS approximates to READS + (WRITES * RAID factor).
    RAID factor is 4x for RAID5 and 2x for RAID1
  - VRaid1 performance is great for Exchange database, logs
  - VRaid5 has higher latency and supports less I/O, good for D2D backups

- Database is primarily 4K random I/O
  - Transaction logs are mostly 4K sequential*
  - Streaming database is usually ignored in sizing MAPI environments

Determine the number of disks required to support peak user IOPS and capacity

- Exchange database requires sufficient physical disks to support the Exchange **bursty** random I/O patterns

- Design to support peak random I/O before capacity
  - 15K disks are capable of 150 IOPS per disk (Vraid1)
  - 10K disks are capable of 100 -120 IOPS per disk
  - Conservative use of 100 IOPS per disk for design allows room for peak periods
  - Storage Groups/databases may requires 4K or higher IOPS per server
  - Transaction logs require 100 - 150 IOPS per storage group

Determine User Profile (how many IOPS required)

- The conversion of user activity to IOPS is critical
  - The EVA performs with minimal latency for Exchange random IOPS when designed correctly (12K IOPS is a sweet spot).
  - Convert user actions to I/O

- Use Windows System Monitor to collect disk counters or use the Fudge factor as a starting point

- Fudge factor = Corporate MAPI users range from light to heavy users. That can be a range of .2 to .6 IOPS per user.

- Larger mailbox size has been shown to increase the amount of I/O required. So if you are planning 200 MB mailboxes than increase IOPS per active user.
    - If you had heavy .6 users with 100MB mailbox, then increase to 1 IOPS for 200MB mailbox

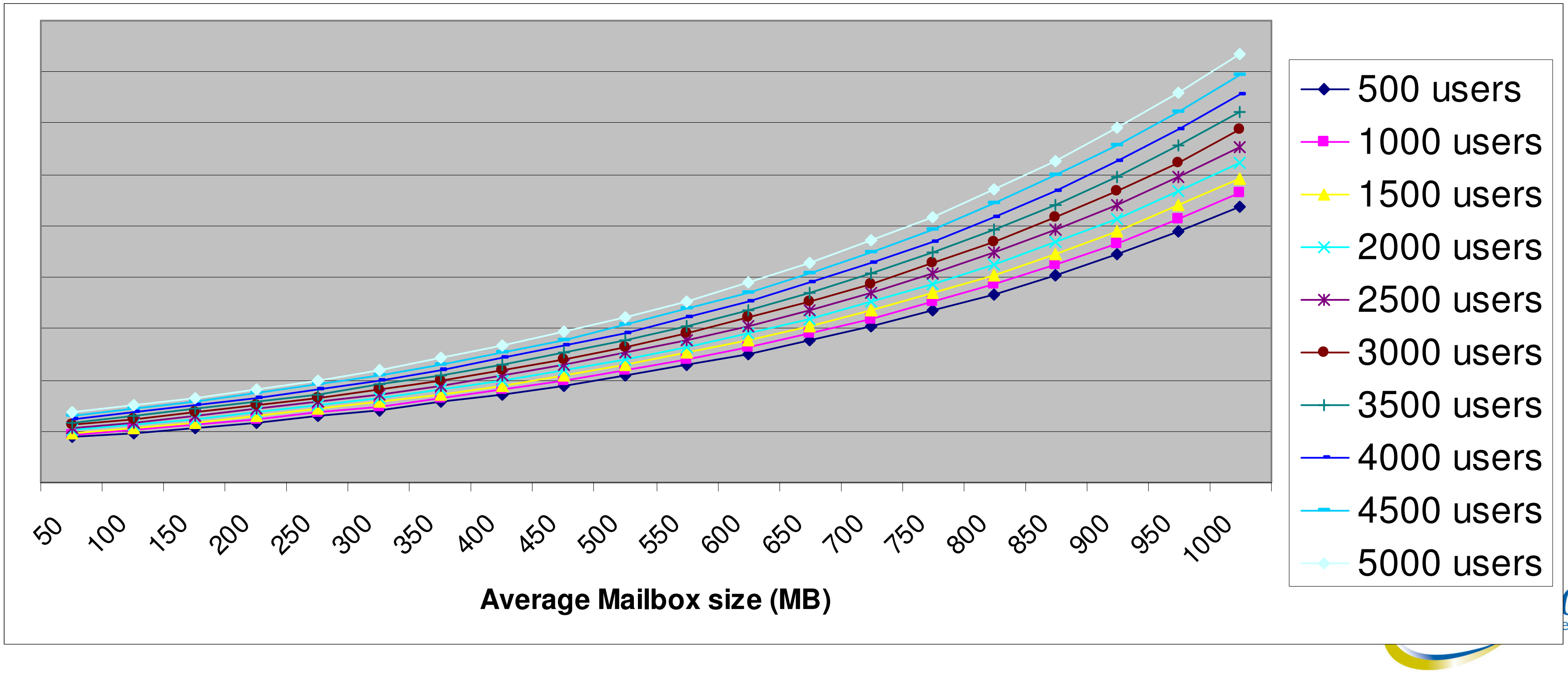
- Total I/O = IOPS/ number of active users. The Total I/O required is matched against the expected available I/O from a disk group, and supportable IOPS from the controller pairs.

Database IOPS/u Rates

Exchange 2003 IOPS/u rates (MAPI user, 200MB mailbox)

- ISP / “light” user: <.3 IOPS/u
- Medium corporate: .5 IOPS/u
- Heavy corporate: >.75 IOPS/u
- Microsoft: 1.2 – 3 IOPS/u

Rates do not stay constant as load, mailbox sizes increase



Sizing Exchange: Software

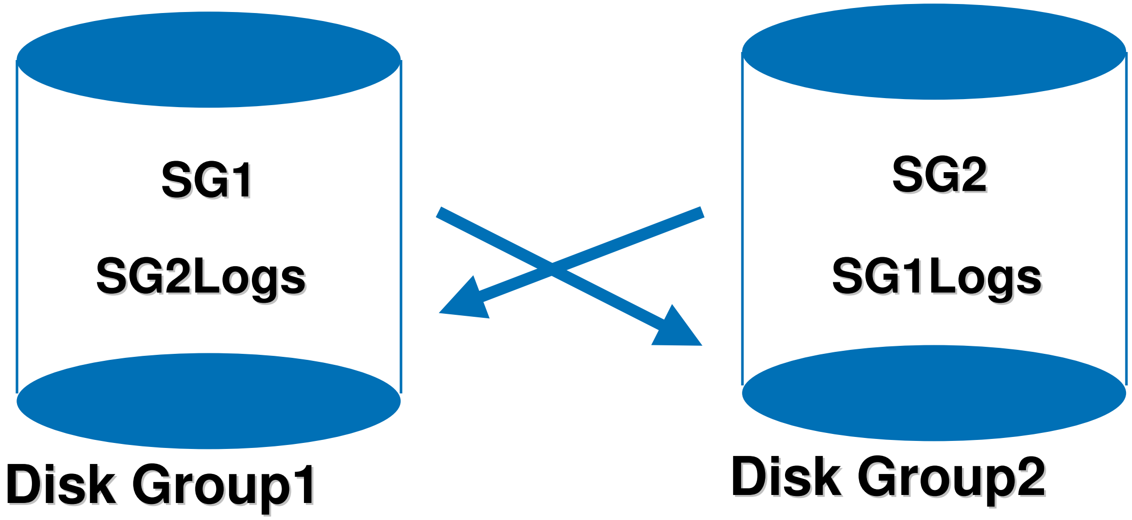
- Software component impacting CPU
  - Anti-Virus
  - Exchange components
    - Content indexing
    - DDLs
    - Cached mode
- Mobile device support
  - E.G. 1 Blackberry user = 2.21 MAPI users for CPU and network, but not storage

Designing the EVA for Exchange

2. Designing the EVA for Microsoft Exchange Server 2003 (Mailbox Server)

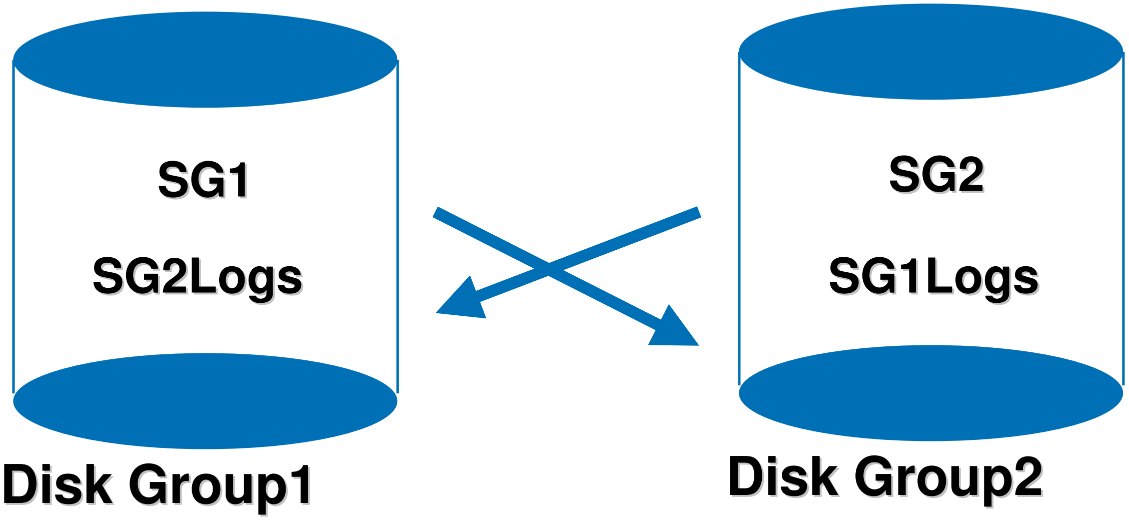
- The configuration of Exchange storage groups and their underlying storage subsystem is the cornerstone of successful Microsoft Exchange Server 2003 deployment.
- Designing today’s Exchange Server 2003 SAN on EVA's to support high performance, high availability and server consolidation.

This old rule still applies – separate transaction logs from databases



- Cross reference disk group design
  - Disk group 1 contains Storage Group 1 and Storage Group 2's logs
  - Disk group 2 contains Storage Group 2 and Storage Group 1's logs
- Note : Good Performance but Lower availability

Cross Reference Database to Logs

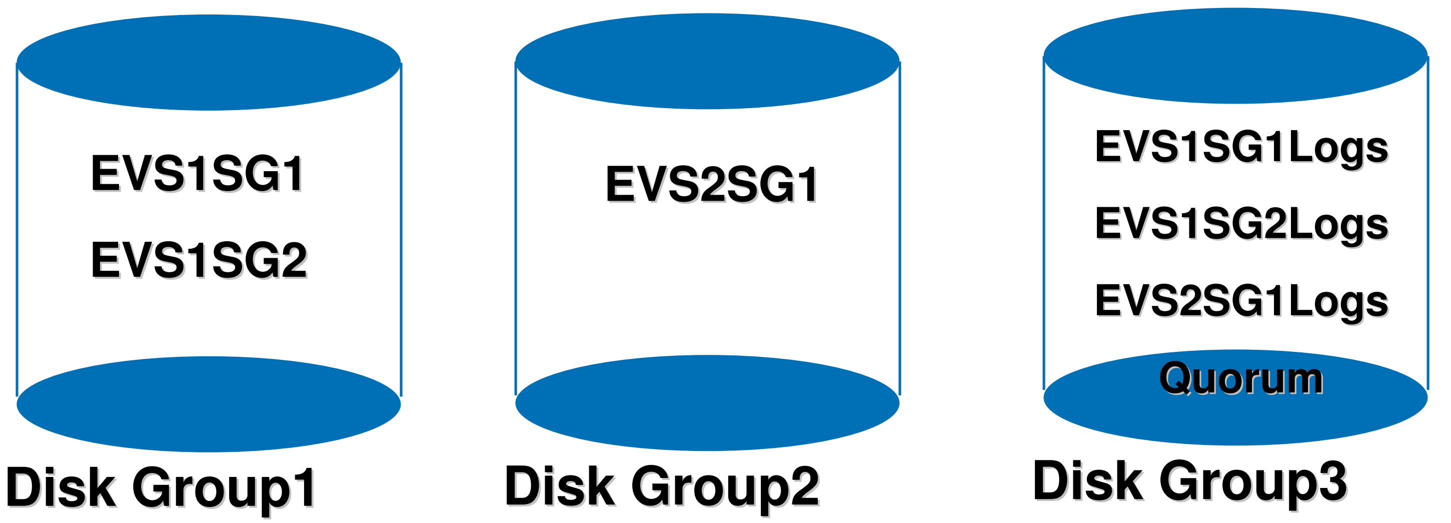


- Cross reference disk group design (less desirable)
  - Exchange database I/O wears physical disks out quicker

- Cluster failover requires devices move cleanly between nodes
  - Multiple nodes in the same disk group were impacted with early VCS drops for cluster failover times

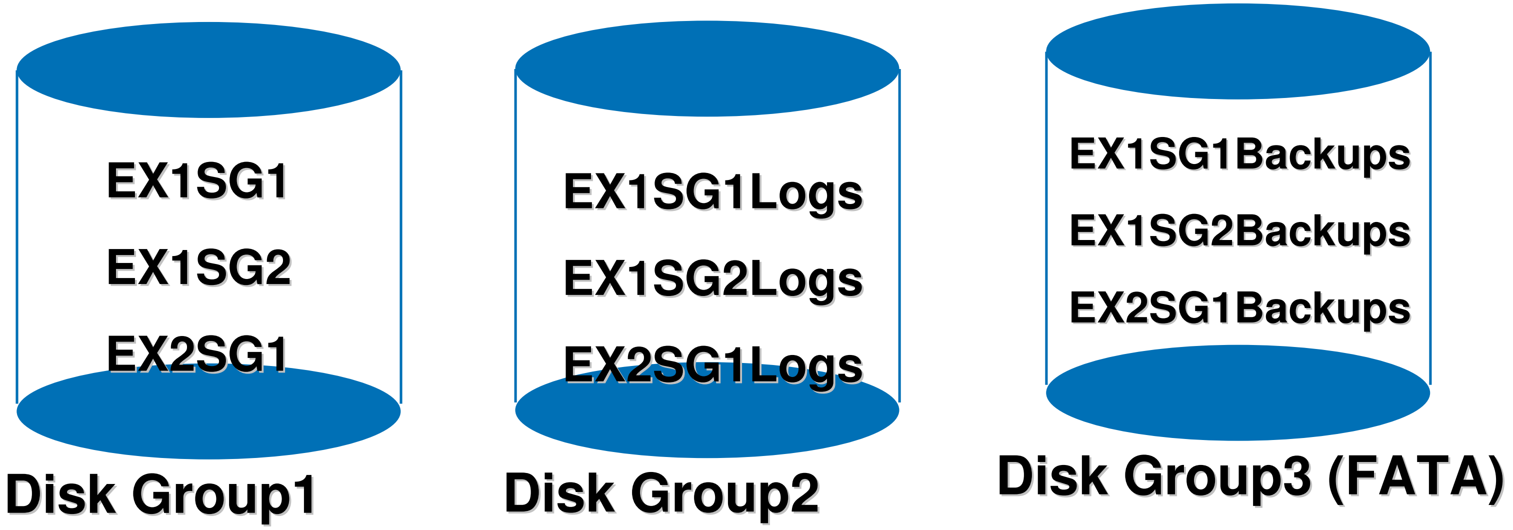
Failure of a single Disk Group – fails both Exchange Servers

Separate Disk Group for all Log devices



- Highly available Disk Group 3 for Transaction Logs
- Disk Group failure, I/O boundary for database, logs
- If failure to DG1 or DG2 lose one Exchange Server
- Do not mix backup I/O streams in with Database (D2D)
- Note: Good Performance, isolate Exchange physical disks from other I/O streams.

Separate Disk Group for all Log devices



- Highly available Disk Group 2 for Transaction Logs
- Disk Group failure, I/O boundary for database, logs
- If failure to DG1 lose all Exchange Server
- Do not mix backup I/O streams in with Database (D2D)
  - Disk Group3 FATA Raid5 backup LUNs
- Note: Good Performance, isolate Exchange physical disks from other I/O streams.

Case Study: Microsoft Exchange 2003 Server Consolidation on EVA

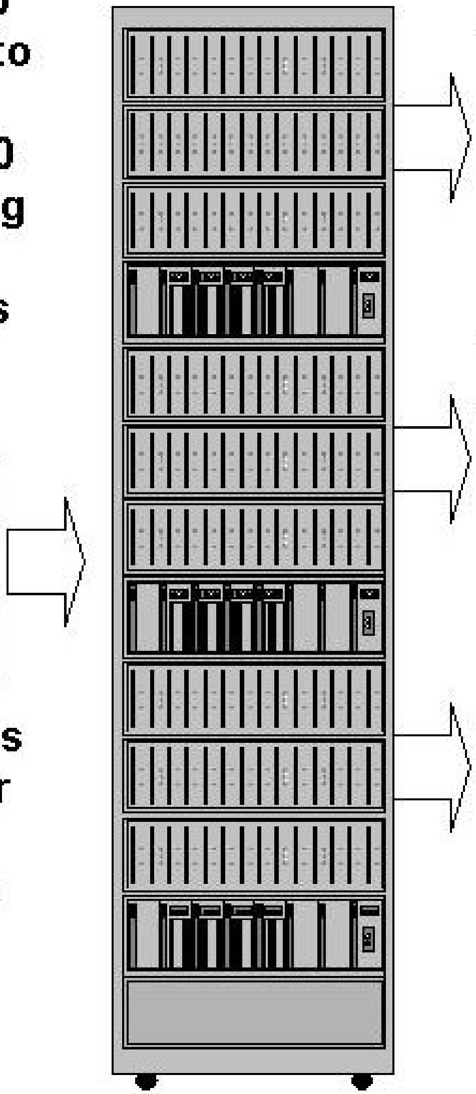
Case Study: OTG Exchange Deployment (Pre consolidation)

- Pre Consolidation Stats
  - 115 mailbox servers in 75 worldwide locations
  - Regional Servers
    - DAS connected
    - 500 to 1500 100MB mailboxes (location specific)
    - Network based backups
  - Redmond
    - Predominantly SAN connected servers
    - 3000 100MB mailboxes
      - 42 drive EMA8000 per Exchange Storage Group
      - 1000 mailboxes per Exchange Storage Group
      - 15 mailstores per server
    - Two stage backup
      - Disk to Disk (NTBackup 10MB/sec per Exchange Storage Group)
      - Disk to Tape (Used GigE fiber for throughput)

Redmond Exchange 2000 legacy SAN

10 PL8500 attached to 10 EMA12000 supporting ~30,000 mailboxes

ProLiant
8500
8 CPU
4 GB
RAM



3000 users per server

3 Storage Groups per EMA 12000

Storage Group 1

LOG	DATA	BACKUP
(6) x 18- GB	(24) x 18- GB	(12) x 18- GB
RAID 0+1	RAID 0+1	RAID 5
~50- GB LUN	~200- GB LUN	~200- GB LUN

Storage Group 2

LOG	DATA	BACKUP
(6) x 18- GB	(24) x 18- GB	(12) x 18- GB
RAID 0+1	RAID 0+1	RAID 5
~50- GB LUN	~200- GB LUN	~200- GB LUN

Storage Group 3

LOG	DATA	BACKUP
(6) x 18- GB	(24) x 18- GB	(12) x 18- GB
RAID 0+1	RAID 0+1	RAID 5
~50- GB LUN	~200- GB LUN	~200- GB LUN

Scalability Goal

- Plan to provide a clustered solution using VSS integrated backup capability for the entire Windows product group
- Increase scale from 3K to 5K mailboxes using an Active (A) Passive (P) cluster
- Double mailbox limits from 100MB to 200MB
- Storage - EVA 5000
  - 168 72GB 10K RPM disks
- Servers – Proliant DL580-G2
  - Originally Quad 1.6GHz (upgraded to 1.9Ghz) 4GB Ram

Next Approach

- Design clusters for global deployment
  - High scale for Redmond
  - Reduced scale for Regions (minimize risk)

- Determining the number of mailboxes to deploy per SAN was hardest decision
  - Validated internally that the EVA in an optimized configuration would sustain 12K transfers with acceptable read and write latency
    - ~ 16ms for read
    - ~ 4ms for write
  - Peak mailbox I/O requirements were trended (Microsoft profile)
    - 100MB – 0.6 to .08 transfers/sec – legacy limit
    - 200MB – 1.0 to 1.2 transfers/sec – new corporate limit
  - 8K mailboxes required 9800 transfers peak
  - Maintain 20% buffer for unexpected I/O – Bulk mailbox moves etc.

Cluster Designs

- Redmond
  - 7 node clusters in an A/A/A/A/P/p/p design
    - (p) refers to Alternate Passive node
    - Main role in life to handle second stage backup to tape
    - One allocated per SAN per cluster
    - 4 Active Exchange instances
    - 4K mailboxes per Exchange instance
    - 200 mailbox per mail store
    - 20 databases per Exchange Instance
    - 2 EVA's per cluster (24TB Raw storage)
- Regional
  - 5 node clusters in an A/A/A/P/p design
    - 3 Active Exchange instances
    - 2.7K mailboxes per instance
    - 135 mailbox per mail store
    - 20 databases per Exchange Instance
    - 1 EVA per cluster (12TB Raw storage)

Redmond Multi-Node Cluster Design - 16,000 Mailboxes
MACS - OTG

PCI Slot Configuration for 580-G2's
Slot 1 Emulex LP952 Tape Library (future)
Slot 2 NC7770 Private Network
Slot 3 Emulex LP952 - Fabric A
Slot 4 NC7770 - Primary Public Network
Slot 5 Exulex LP952 - Fabirc B
Slot 6 NC7770 - Secondary Public Network

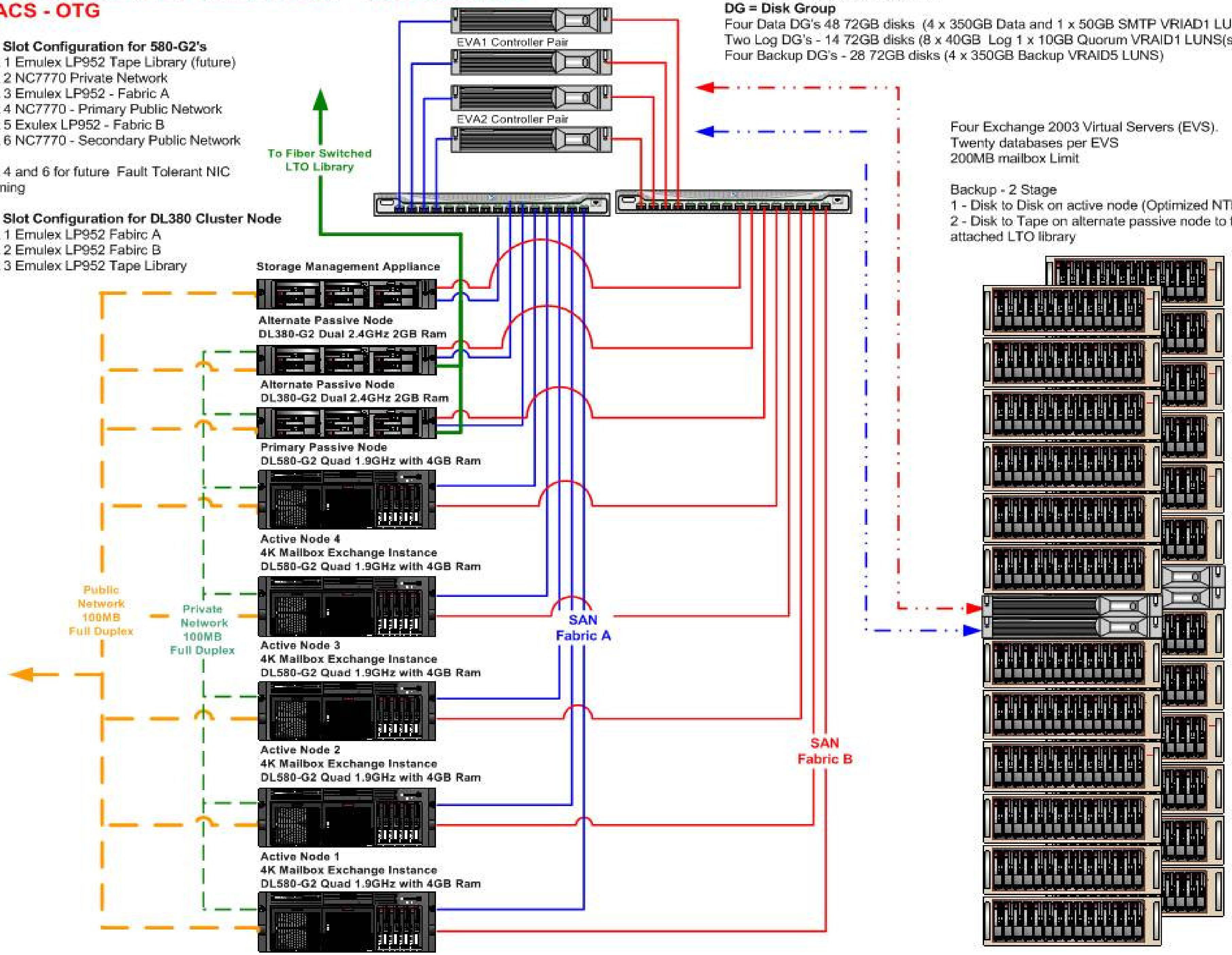
Slot 4 and 6 for future Fault Tolerant NIC teaming

PCI Slot Configuration for DL380 Cluster Node
Slot 1 Emulex LP952 Fabirc A
Slot 2 Emulex LP952 Fabirc B
Slot 3 Emulex LP952 Tape Library

Two HP Enterprise Virtual Array
DG = Disk Group
Four Data DG's 48 72GB disks (4 x 350GB Data and 1 x 50GB SMTP VRIAD1 LUN(s))
Two Log DG's - 14 72GB disks (8 x 40GB Log 1 x 10GB Quorum VRAID1 LUNS(s))
Four Backup DG's - 28 72GB disks (4 x 350GB Backup VRAID5 LUNS)

Four Exchange 2003 Virtual Servers (EVS).
Twenty databases per EVS
200MB mailbox Limit

Backup - 2 Stage
1 - Disk to Disk on active node (Optimized NTBackup)
2 - Disk to Tape on alternate passive node to fiber attached LTO library


EVA1 Controller Pair
EVA2 Controller Pair
To Fiber Switched LTO Library
Storage Management Appliance
Alternate Passive Node
DL380-G2 Dual 2.4GHz 2GB Ram
Alternate Passive Node
DL380-G2 Dual 2.4GHz 2GB Ram
Primary Passive Node
DL580-G2 Quad 1.9GHz with 4GB Ram
Active Node 4
4K Mailbox Exchange Instance
DL580-G2 Quad 1.9GHz with 4GB Ram
Active Node 3
4K Mailbox Exchange Instance
DL580-G2 Quad 1.9GHz with 4GB Ram
Active Node 2
4K Mailbox Exchange Instance
DL580-G2 Quad 1.9GHz with 4GB Ram
Active Node 1
4K Mailbox Exchange Instance
DL580-G2 Quad 1.9GHz with 4GB Ram
Public Network 100MB Full Duplex
Private Network 100MB Full Duplex
SAN Fabric A
SAN Fabric B

Fully understand decision points for configuring EVA storage for Exchange

- Number of disk groups
  - Design the minimal number of disk groups that satisfies the failure domain and I/O domain requirements
    - For Exchange a single Disk Group is not recommended
  - Separate Exchange Storage Group Log files from its database with a Disk Group boundary
  - Understand and monitor I/O that impacts disk groups
  - Be careful with Clusters (disk group I/O on earlier VCS impacted failover)
  - Design to avoid the cost of stranded storage and reconfiguration
  - Read EVA Best Practices - Cost, Performance, and Availability http://h18006.www1.hp.com/storage/arraywhitepapers.html

Follow Best Practices for designing Disk Groups

- Vraid1 is recommended for database and transaction logs
  - Even number of disks
  - All disks should be arranged in a vertical fashion, i.e.,distribute the disks among the shelves such that the same bay in each shelf has a disk

- Vraid 5 or mixed Vraid disk group for backups
  - minimum of 8 shelves in a configuration for VRaid 5
  - All disks should be arranged in a vertical fashion, i.e, distribute the disks among the shelves such that the same bay in each shelf has a disk
  - The total number of disks in a disk group should be an integer multiple of eight.
  - Higher latency and less IOPS – use diskpar utility

16K Mailbox Spindle Distribution (cluster)

DG = Disk Group
EVS = Exchange Virtual Server
MP = Mount Point

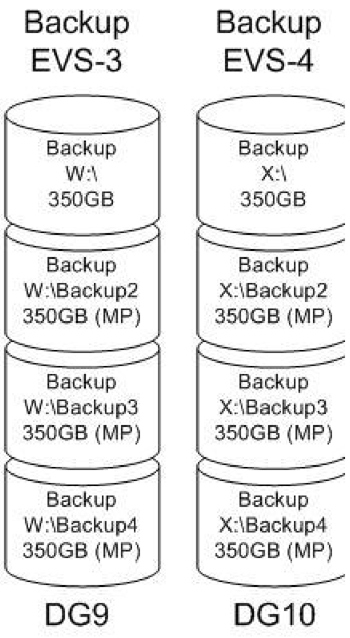
Utilizing Two 12TB Enterprise Virtual Arrays

48 72GB 10K disks: DG1,DG2 EVA1 - DG6,DG7 EVA2 - VRAID1
16 72GB 10K disks: DG3 EVA1 - DG7 EVA2 - VRAID1
28 72GB 10K disks: DG4,DG5 EVA1 - DG9,DG10 EVA2 - VRAID5

LUN Controller Distribution - optimized backup performance
SG1 and SG2 Data, Log and Backup on HSV110 Controller 1
SG3 and SG4 Data, Log and Backup on HSV110 Controller 2

Note:
Clone integration will require Backup LUN elimination and replacement with VRAID1 LUNS.

VRAID5 LUNS are presented to hosts with mirrored write back cache disabled to overcome bottleneck backup process.



Backup EVS-3

Backup W:\ 350GB

Backup W:\Backup2 350GB (MP)

Backup W:\Backup3 350GB (MP)

Backup W:\Backup4 350GB (MP)

DG9

Backup EVS-4

Backup X:\ 350GB

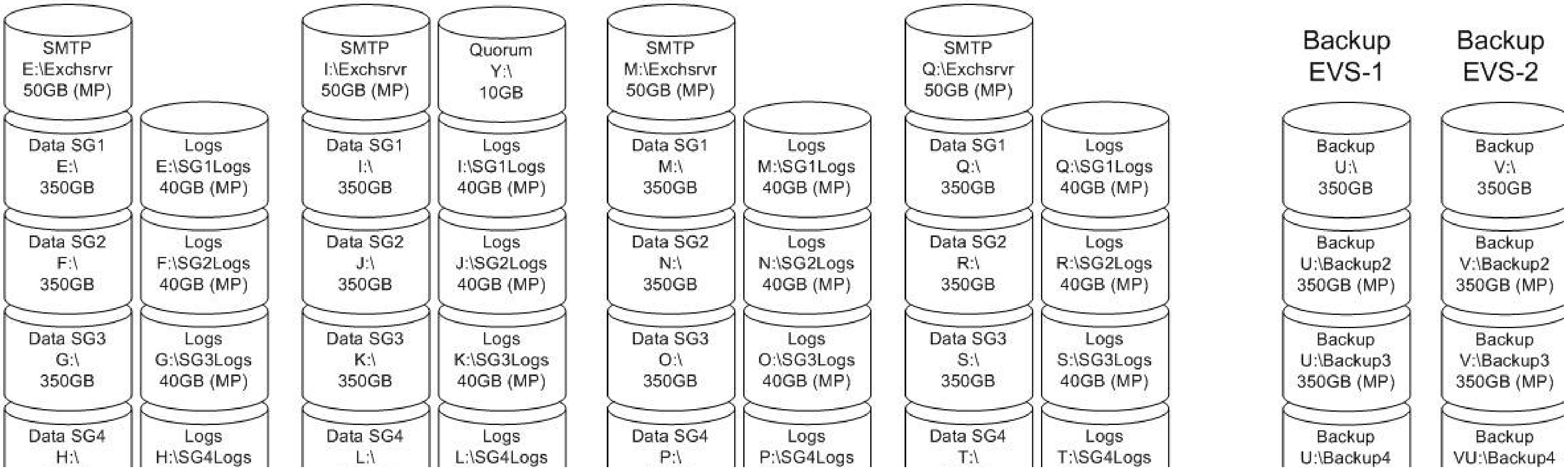
Backup X:\Backup2 350GB (MP)

Backup X:\Backup3 350GB (MP)

Backup X:\Backup4 350GB (MP)

DG10

EVS-1 EVS-2 EVS-3 EVS-4



EVS-1

SMTP E:\Exchsrvr 50GB (MP)

Data SG1 E:\ 350GB

Data SG2 F:\ 350GB

Data SG3 G:\ 350GB

Data SG4 H:\

Logs E:\SG1Logs 40GB (MP)

Logs F:\SG2Logs 40GB (MP)

Logs G:\SG3Logs 40GB (MP)

Logs H:\SG4Logs

EVS-2

SMTP I:\Exchsrvr 50GB (MP)

Data SG1 I:\ 350GB

Data SG2 J:\ 350GB

Data SG3 K:\ 350GB

Data SG4 L:\

Quorum Y:\ 10GB

Logs I:\SG1Logs 40GB (MP)

Logs J:\SG2Logs 40GB (MP)

Logs K:\SG3Logs 40GB (MP)

Logs L:\SG4Logs

EVS-3

SMTP M:\Exchsrvr 50GB (MP)

Data SG1 M:\ 350GB

Data SG2 N:\ 350GB

Data SG3 O:\ 350GB

Data SG4 P:\

Logs M:\SG1Logs 40GB (MP)

Logs N:\SG2Logs 40GB (MP)

Logs O:\SG3Logs 40GB (MP)

Logs P:\SG4Logs

EVS-4

SMTP Q:\Exchsrvr 50GB (MP)

Data SG1 Q:\ 350GB

Data SG2 R:\ 350GB

Data SG3 S:\ 350GB

Data SG4 T:\

Logs Q:\SG1Logs 40GB (MP)

Logs R:\SG2Logs 40GB (MP)

Logs S:\SG3Logs 40GB (MP)

Logs T:\SG4Logs

Backup EVS-1

Backup U:\ 350GB

Backup U:\Backup2 350GB (MP)

Backup U:\Backup3 350GB (MP)

Backup U:\Backup4

Backup EVS-2

Backup V:\ 350GB

Backup V:\Backup2 350GB (MP)

Backup V:\Backup3 350GB (MP)

Backup VU:\Backup4

Reality Check – Does Mark, Dave, or the customer like it?

- The storage administrator design goals typically focus on how much capacity is required, what applications are supported, IOPS required, and how fast can data be recovered.
- The EVA is a high performance storage system that is easy to configure for Exchange Server database. Only a few options are required to design an optimal storage system. The options are based on overall system COST, PERFORMANCE, and AVAILABLILITY
- Careful with marketing poof-or-menace numbers
- http://h18006.www1.hp.com/storage/arraywhitepapers.html

Problems Resolved

- Many pain points detected during validation
- Poor cluster failover with some disk group designs
  - LUNS in a disk group shared with LUNS under load would fail a controlled move within cluster
  - Serious implications for multi node designs
- Impact on sharing sequential and random I/O can be significant within the same disk group
  - Results in excessive latency for Exchange I/O if sequential requirements are shared within the same disk group as Exchange mail stores
- Poor read write latencies beyond 2K transfers
  - QueueDepth on the KGPSA driver parameter
- Throttle down during bus resets
  - Resolved with A16 of KGPSA

HBA Tuning

- Optimal host performance on a Windows host with EVA-only storage requires a registry change to support performance requirements.
- Queue target From 1 to 0
  - Defines how the QueueDepth parameter is interpreted, on a per-LUN or per-target (subsystem) basis
- Queue depth From 25 to 128
  - 128 is max outstanding requests Windows will send to a LUN if Queuetarget = 0
- Number of requests From 50 to 150 (decimal)
- Shallow QueueDepth will limit host transfers
  - OTG was bottlenecked at 2K transfers with 40ms read latency
  - Increased QueueDepth and now sustain 5-6K transfers with 16ms max read latency

D2D backup performance to Vraid5 was no good before Diskpar

- **Customer Advisory OI040301_CW02**

  Windows applications that use EVA "Vraid" Virtual Disks for data stores may experience a write performance issue if partitions are not properly aligned.

  Performance data shows the impact to some applications may be significant.

  Data shows a greater impact is experienced on Vraid5 units than on Vraid1 units. Applications with multiple sequential write streams are more impacted than random write loads. The Windows DISKPAR utility will allow creating a disk partition that is properly aligned.

Diskpar.exe

- Available on the W2K Resource Kit
- Must be run <u>before</u> format of partition
- Not to be confused with DISKPART (yet)

```
You can use Disk Manager to delete all existing partitions
Are you sure drive 12 is a raw device without any partition? (Y/N) Y
---- Drive 12 Geometry Infomation ----
Cylinders = 1305
TracksPerCylinder = 255
SectorsPerTrack = 63
BytesPerSector = 512
DiskSize = 10733990400 (Bytes) = 10236 (MB)
We are going to set the new disk partition.
All data on this drive will be lost. continue (Y/N)? y
Please specify starting offset (in sectors): 64
Please specify partition length (in MB) (Max = 10236): 10236     <---- enter partition
Done setting partition.
---- New Partition information ----
StatringOffset = 32768
PartitionLength = 10733223936
HiddenSectors = 64
PartitionNumber = 1
PartitionType = 7
You now should use Disk Manager to format this partition
```

Host Perfmon counters Database

- **PhysicalDisk\Average Disk sec/Read**
- Indicates the average time (in seconds) to read data from the disk.
  - The average value should be below 20 ms.
  - Spikes (maximum values) should not be higher than 100 ms.
- **PhysicalDisk\Average Disk sec/Write**
- Indicates the average time (in seconds) to write data to the disk
  - The average value should be below 20 ms.
  - Spikes (maximum values) should not be higher than 100 ms.
- **Physical Disk\Disk Transfers / sec**
  - **Host IOPS (measure against controller ports)**
  - **Note – OTG requires < 10ms writes**

Backup @ OTG :http://www.microsoft.com/technet/itsolutions/msit/operations/msgbrtcs.mspx

- Two Stage Backup
  - Disk to Disk using NTBackup (Runs on active node)
    - Registry optimization doubles OOB throughput (20MB/sec)
    - QFE version to provide un-buffered I/O to reduce contention due soon. Will ship as QFE
    - Lot of optimizations
      - Secure Path LUN distribution is critical
      - 6 to 7GB per min is max throughput on the EVA
      - Sector aligned targets
      - Disabled mirror write backup cache on backup targets
  - Disk to Tape using Backup Exec (Runs on Alternate Node)
    - Alternate Passive nodes connected to tape fabric
    - 4 LTO-Type1 per pair of clusters
    - 1.6GB/min per tape device sustained
- Design works around poor connectivity problems with tape fabric. Backup Exec requires hard restart to re-establish connectivity to tape devices on disconnect.
- See links for paper on optimizations

MSIT/OTG Backup

- Total throughput sustained at 4.8GB/mim per Exchange virtual server with four concurrent backups
- Allows MSIT messaging to run eight concurrent backups across two Exchange virtual servers per SAN
- Monitor total disk write bytes/sec
- The backup target disks are configured using Vraid5
- Backup targets are all sector aligned to eliminate the offset caused on a primary partition as a result of the MBR – utilize diskpar to partition LUNS
- Mirrored write back cache is disabled on all backup targets

MSIT/OTG Backup

- Don’t share disk spindles to support backup targets and Exchange databases.
  - IT messaging strongly recommends following this as a best practice to ensure that sequential content streaming to tape does not impact the random access requirements for Exchange.
  - This allows streaming the content to tape as part of the second stage process at any time during the day without impacting users supported on the clusters.
- Throughput to tape sustained at a rate of 1.6GB/min per stream with up to four concurrent streams to four LTO1 tape devices
- Best throughput was achieved by balancing SecurePath
  - SG1 and SG2 Data, Log and Backup disks per virtual server on Controller 1
  - SG3 and SG4 Data, Log and Backup disks per virtual server on Controller 2
- Restore rates can be achieved in the range of 2GB/min.

Links for Case Study

- OTG deployments and best practices can be found on http://www.microsoft.com/technet/itsolutions/msit/operations/msgbrtcs.mspx
- http://www.microsoft.com/technet/itsolutions/msit/operations/mesoptwp.mspx
- Microsoft TechNet http://www.microsoft.com/technet/itshowcase
- http://www.microsoft.com/technet/treeview/default.asp?url=/technet/itsolutions/MSIT/Deploy/Ex03ATWP.asp
- Microsoft Case Study Resources http://www.microsoft.com/resources/casestudies
- D2D two-stage backup optimizations
- http://download.microsoft.com/download/4/3/1/43104b4b-dd07-44d0-90c9-d1cda210f3cd/ExchangeBackupNote.doc
- Storage array systems white papers
- http://h18006.www1.hp.com/storage/arraywhitepapers.html

How To Validate the Storage Subsystem

Sizer Tools are a Starting Point

- Performance Estimator Tool for XP and EVA
  - Hp confidential

NSS Sizer

http://h30144.www3.hp.com/

Exchange workload – creates BOM as starting point

ActiveAnswers Exchange Storage Calculator

http://h71019.www7.hp.com/ActiveAnswers/Render/1,1027,2400-6-100-225-1,00.htm

Validate the Storage Subsystem

- Microsoft tools
- http://www.microsoft.com/exchange/downloads/2003.asp
- **Jetstress (English only)**
  Verify the performance and stability of your disk subsystem by simulating disk I/O load on a test server running Exchange Server before putting your server into a production environment.
  Load Simulator 2003 (LoadSim) (English only)
  Simulate the performance load of MAPI clients with this benchmarking tool, which enables you to test how a server running Exchange Server 2003 responds to e-mail message loads.

Stress Testing Tools (public download)

- LoadSim (standard for Exchange Benchmarks)
  - Requires client machines with 500-750 loadsim users each
  - Creates AD accounts & DLs
  - Initializes mailboxes
  - MAPI Profiles
    - Light, MMB3, Heavy (standard profiles map to I/O)
    - Set of actions simulate ‘user’ and load on the storage array
- JetStress (excellent documentation with download)
  - NOTE: Jetstress document suggests 20ms latency as acceptable. OTG considers 20ms excessive and a poor user experience
  - intended only to simulate Exchange disk I/O activity
  - Does not require Exchange to be installed
  - New UI version soon
- IOMETER
  - Use to initialize first write penalty (60GB an hour init)
  - After initialized, can be used to quickly test storage array

EVA 5000/15K 168 disks, Cross Reference -Jetstress (E:\jetstress.exe –l F:\ -n 28311552 –b 64 >c:\temp\makedb1.txt)

Test 2: 168 73GB 15k rpm drives

Indicator	Cluster 1	Cluster 4	Total
Secure Path Balanced?	No	No	No
Jetstress Threads	2x12	2x12	48
Disk Transfers/sec (Average) M:	**3073**	**2839**	
Disk Transfers/sec(Average) O:	**2866**	**3041**	
Total DB IOPS Achieved	**5939**	**5880**	**11819**
Average Disk Sec/Transfer (Average) M:	**0.004**	**0.004**	
Average Disk Sec/Transfer (Maximum) M:	**0.005**	**0.005**	
Average Disk Sec/Transfer (Average) O:	**0.004**	**0.004**	
Maximum Disk Sec/Transfer (Maximum) O:	**0.005**	**0.005**	
Average Disk Sec/Read (Average) M:	**0.006**	**0.006**	
Average Disk Sec/Read (Maximum) M:	**0.007**	**0.006**	
Average Disk Sec/Read (Average) O:	**0.006**	**0.006**	
Average Disk Sec/Read (Maximum) O:	**0.007**	**0.006**	
Average Disk Sec/Write (Average) M:	**0.001**	**0.001**	
Average Disk Sec/Write (Maximum) M:	**0.001**	**0.001**	
Average Disk Sec/Write (Average) O:	**0.001**	**0.001**	
Average Disk Sec/Write (Maximum) O:	**0.001**	**0.002**	

EVA 5000/15K 168 disks, Cross Reference -Jetstress (E:\jetstress.exe –l F:\ -n 28311552 –b 64 >c:\temp\makedb1.txt)

Indicator	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
Secure Path Balanced?	No	No	No	No	0
Jetstress Threads	2x12	2x12	2x12	2x12	
Disk Transfers/sec (Average) M:	**1838**	**1797**	**1999**	**1861**	
Disk Transfers/sec(Average) O:	**1994**	**2157**	**2562**	**1962**	
Total DB IOPS Achieved	**3832**	**3954**	**4561**	**3823**	16170
Average Disk Sec/Transfer (Average) M:	**0.007**	**0.007**	**0.006**	**0.006**	
Average Disk Sec/Transfer (Maximum) M:	**0.008**	**0.008**	**0.011**	**0.007**	
Average Disk Sec/Transfer (Average) O:	**0.006**	**0.006**	**0.006**	**0.006**	
Average Disk Sec/Transfer (Maximum) O:	**0.007**	**0.007**	**0.011**	**0.007**	
Average Disk Sec/Read (Average) M:	**0.01**	**0.01**	**0.009**	**0.009**	
Average Disk Sec/Read (Maximum) M:	**0.01**	**0.011**	**0.01**	**0.011**	
Average Disk Sec/Read (Average) O:	**0.009**	**0.009**	**0.009**	**0.009**	
Average Disk Sec/Read (Maximum) O:	**0.01**	**0.01**	**0.009**	**0.01**	
Average Disk Sec/Write (Average) M:	**0.002**	**0.002**	**0.002**	**0.002**	
Average Disk Sec/Write (Maximum) M:	**0.003**	**0.004**	**0.011**	**0.003**	
Average Disk Sec/Write (Average) O:	**0.002**	**0.001**	**0.002**	**0.002**	
Average Disk Sec/Write (Maximum) O:	**0.005**	**0.002**	**0.011**	**0.003**	

EVA 5000/15K – 4 Storage Groups, VRaid 1

jetstress_SG1\Jetstress –L F:\jetstress_log1 –n 0 -B 64 –T 8 -I 50 –R 0 –D 50 -z -a -q 86400

Disk Transfer Rates					
	EVS1	EVS2	EVS3	Total	Average
Total Database IO	4694.7	3925.8	4310.4	12930.9	4310.3
Storage Group IO/user	1.269	1.061	1.165		1.165
Total Transaction Log IO	1673.1	1448.8	1588.2	4710.1	1570.0
Transaction Log IO/user	0.452	0.392	0.429	17641.0	0.424

EVA 5000/15K – 4 Storage Groups. VRaid 1

jetstress_SG1\Jetstress –L F:\jetstress_log1 –n 0 -B 64 –T 8 -I 50 –R 0 –D 50 -z -a -q 86400

Disk Latency & Queue Length				
	Threshold	EVS1	EVS2	EVS3
Storage Group 1 (J:)				
Average Disk sec/Read	0.02	Pass	Pass	Pass
Average Disk sec/Write	0.02	Pass	Pass	Pass
Current Disk Queue Length	10	Pass	Pass	Pass
Storage Group 2 (K:)				
Average Disk sec/Read	0.02	Pass	Pass	Pass
Average Disk sec/Write	0.02	Pass	Pass	Pass
Current Disk Queue Length	10	Pass	Pass	Pass
Storage Group 2 (L:)				
Average Disk sec/Read	0.02	Pass	Pass	Pass
Average Disk sec/Write	0.02	Pass	Pass	Pass
Current Disk Queue Length	10	Pass	Pass	Pass
TL 1 (F:)				
Average Disk sec/Read	0.02	Pass	Pass	Pass
Average Disk sec/Write	0.02	Pass	Pass	Pass
Current Disk Queue Length	1	Pass	Pass	Pass
TL 2 (G:)				
Average Disk sec/Read	0.02	Pass	Pass	Pass
Average Disk sec/Write	0.02	Pass	Pass	Pass
Current Disk Queue Length	1	Pass	Pass	Pass
TL 3 (H:)				
Average Disk sec/Read	0.02	Pass	Pass	Pass
Average Disk sec/Write	0.02	Pass	Pass	Pass
Current Disk Queue Length	1	Pass	Pass	Pass

Stress Testing Tools (public download)

- LoadSim (standard for Exchange Benchmarks)
- http://www.microsoft.com/downloads/details.aspx?familyid=92eb2edc-3433-47ca-a5f8-0483c7ddea85&displaylang=en

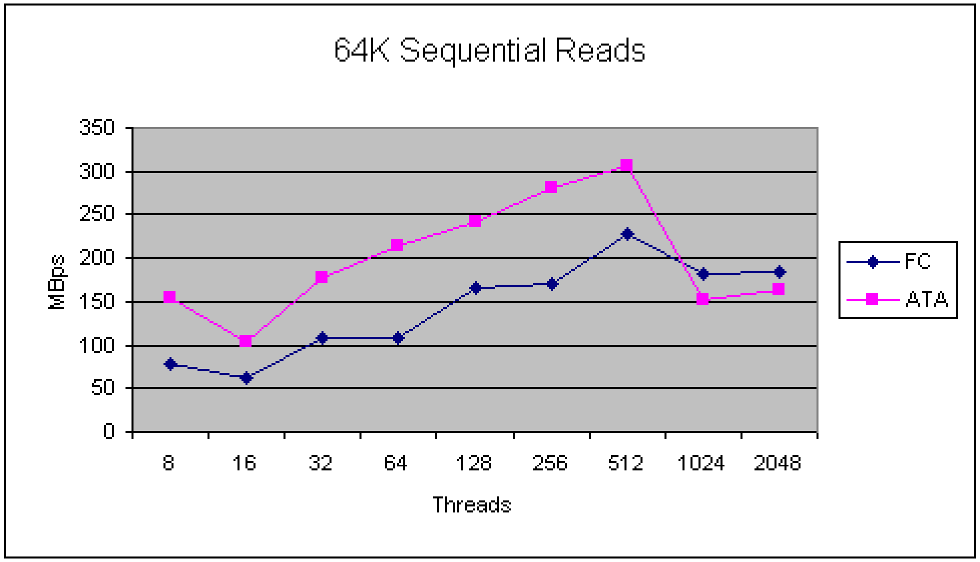
- Loadsim Benchmarks (not real world customer configurations)
- http://www.microsoft.com/exchange/evaluation/performance/mmb3.asp

- JetStress (excellent documentation with download)
  - http://www.microsoft.com/downloads/details.aspx?FamilyId=94B9810B-670E-433A-B5EF-B47054595E9C&displaylang=en

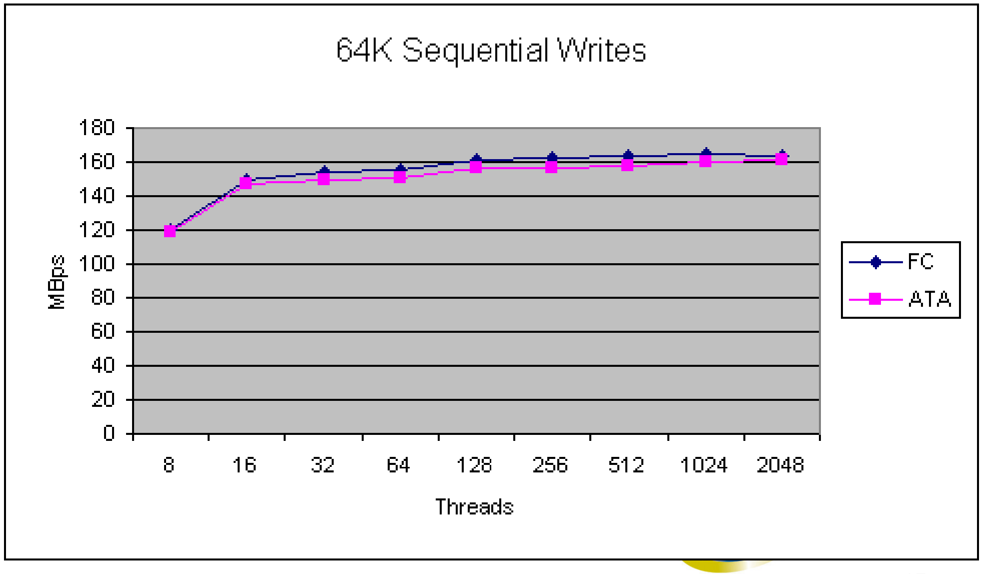
- IOMETER
  - http://www.iometer.org/

IOMeter FATA backup stream

64K Streaming Read (Sequential 64K Read I/O)



64K Streaming Write (Sequential 64K Write I/O)



SAN Storage Features Windows Server 2003

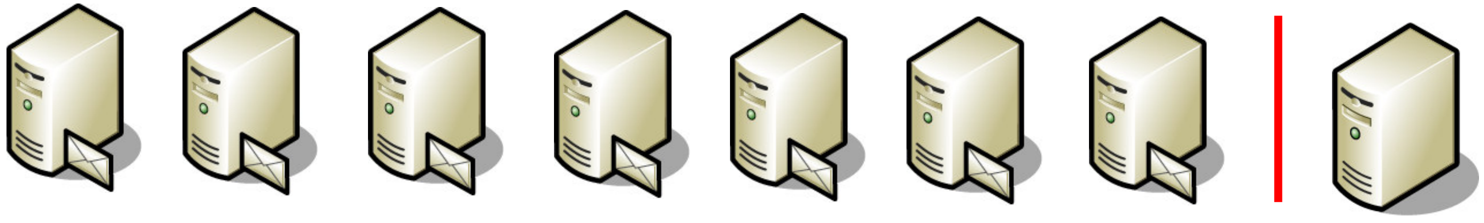
Features of Windows Server 2003 (Storage Centric)

Volume mount points supported on Windows 2003 (318458); useful for 4/8-node clusters

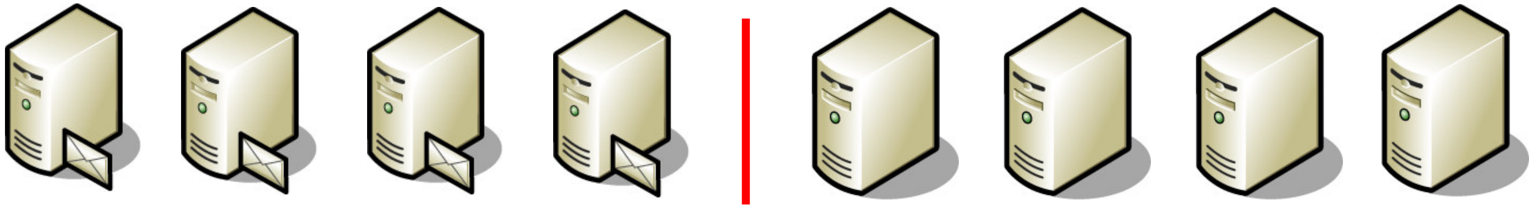
- SAN Boot Support
  - http://h18000.www1.hp.com/products/storageworks/san/documentation.html#sanbg
  - http://support.microsoft.com/default.aspx?scid=kb;en-us;305547
- VDS, Diskpart, Raidisk
- VSS
- Storport
- MPIO
- Exchange SP1

Exchange 2003 Clustering

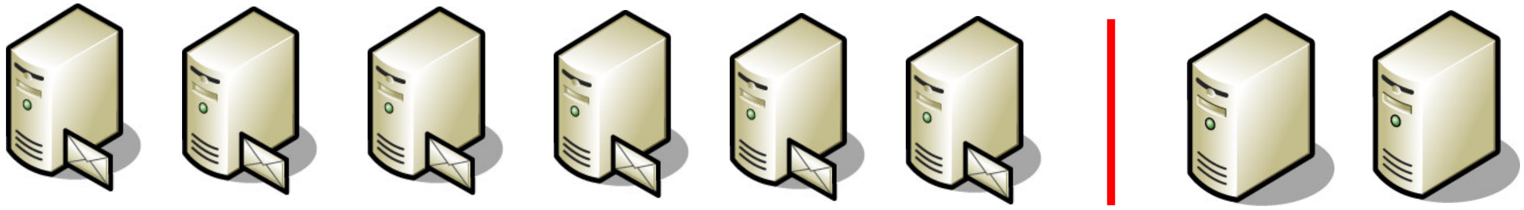
7+1



4+4



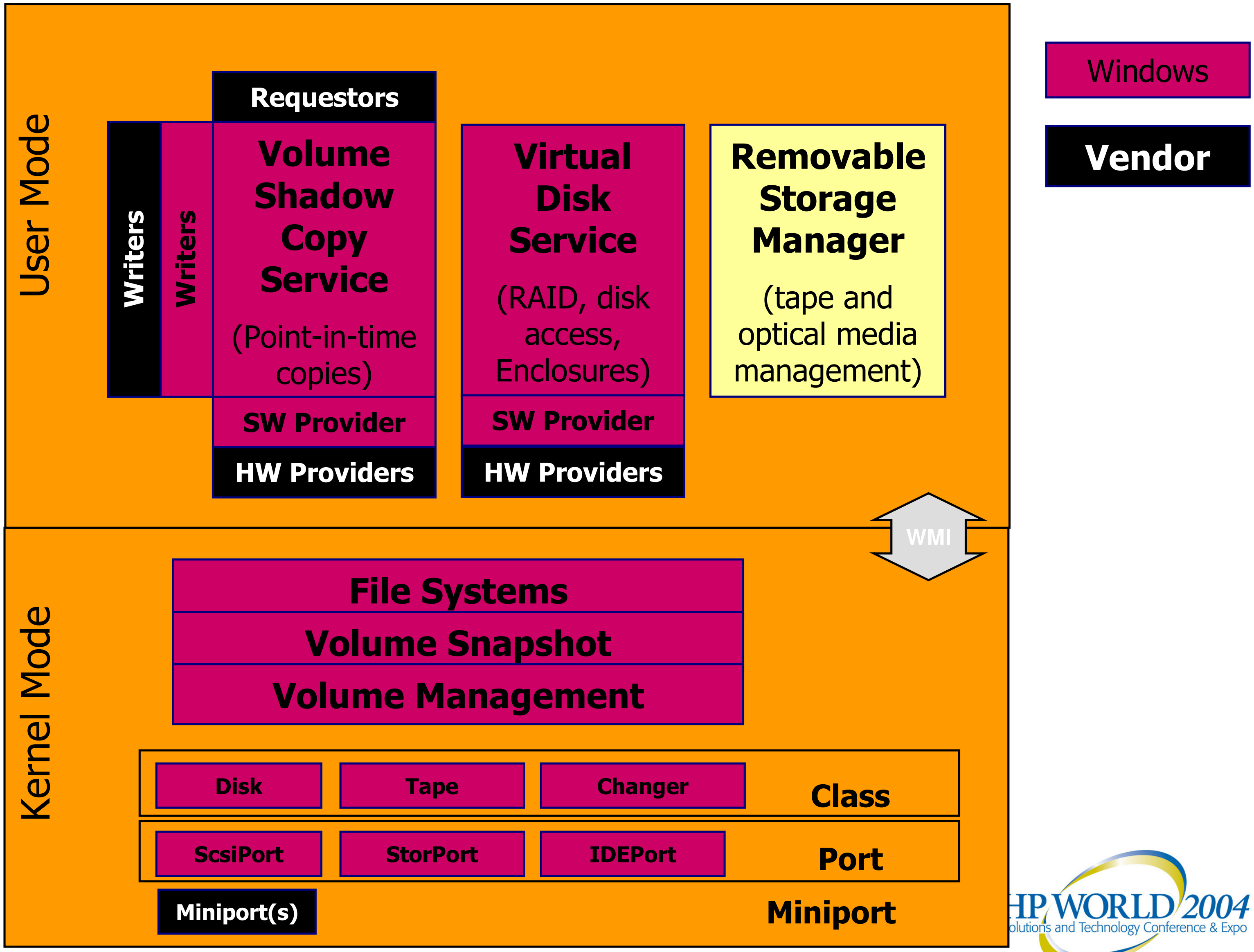
6+2



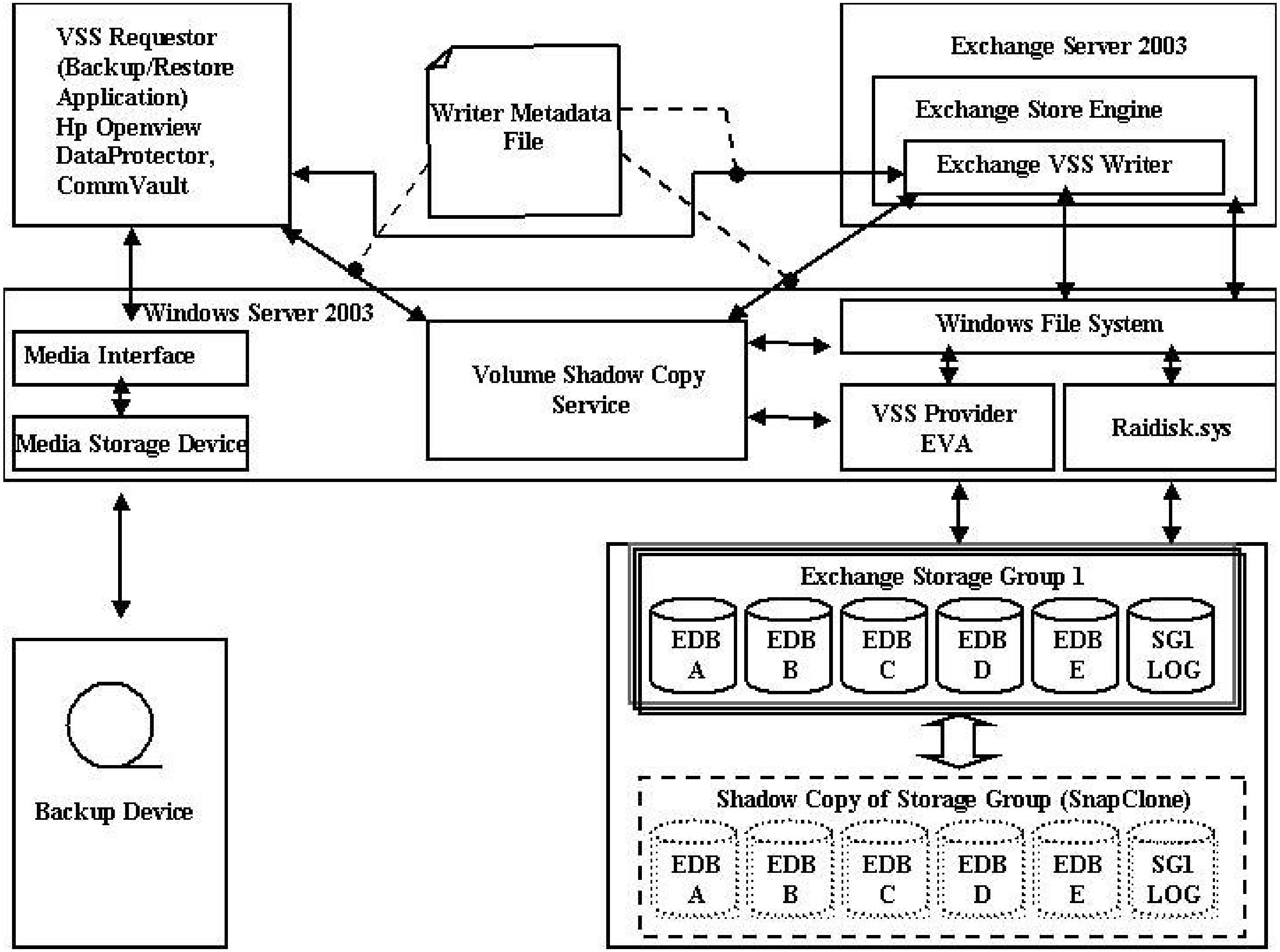
VSS/VDS Components

- Volume Shadow Copy Service Coordinator (VSS)
- Requestors – Backup Apps, Instant Recovery Solutions
- Writers – Represents Apps (ie. SQL, Exchange, AD, etc.)
  - Coordinates with backup apps
  - Differentiates VSS from competitors
- Virtual Disk Service (VDS)
  - VDS is a core service new to Windows Server 2003 and Windows Storage Server 2003. The VDS infrastructure is designed to provide storage administrators with a single user interface for managing multi-vendor storage at the block level.
- VSS/VDS Providers
  - Hardware
  - In-box

Windows 2003 Storage Stack



VSS high-level overview of components for Exchange backup/restore



Volume Shadow Copy Service (VSS) Design Considerations

- In order to use VSS it needs to be considered early in the storage design phase
  - Number and size of Exchange Database LUNs
    - VSS SnapClone operations for Exchange are at the Storage Group or Enterprise level
    - An Exchange SnapShot Set of a StorageGroup includes all database volumes, transaction log file, and checkpoint file
  - VSS SnapClone is successful with 3 or less LUNs in a SnapShot set
    - Best Practice is to use one LUN for the Storage Group (SG)
    - Keep databases 60GB or smaller to keep the SG less than 400GB
    - VSS SnapClone to a separate Disk Group that could support production I/O for that Storage Group
    - Sweetspot tested configurations
- VSS & VDS Hardware Providers:www.hp.com/support/vdsvss

VSS Requestors

- Requestors can add value to the backup solution
  - FRS 2003 instant recovery
    - Recovery includes log replay
  - Many backup vendors are focused on backups
    - Recovery may not be faster than traditional backup
  - Does requestor support Copies, Full, Incremental backups?

  VSS Requestor must check snapshot consistency by running “eseutil /i /k” against the database and log files.

  Due to the snap nature of VSS backups, JET does not get the opportunity to touch all the pages to do the necessary consistency checks. Therefore it is VSS Requestor's responsibility to ensure snapshot consistency.

VSS is the Holy Grail (but still immature)

- Transportable VSS snapshots require Windows 2003 Enterprise Edition
- Microsoft QFE's required
  - QFE83112 - VDS Stale Cache
  - QFE67560 KB833167 - VSS in a cluster configuration
  - Must be installed correctly

Requestor must run integrity check

VSS becomes your primary backup if truncating logs

Design to enable VSS, but validate it fits your backup window

KB822896- Exchange Server 2003 Data Back Up and Volume Shadow Copy Services

Multipath (MPIO) *Microsoft supported architecture*

- MPIO DSM's**:** http://h18006.www1.hp.com/products/sanworks/multipathoptions/index.html
  - **Microsoft Windows MPIO DSMs (device specific modules)**
  - **Download Software & Drivers**
  - » HP MPIO Basic Failover v1.0 for XP Arrays (May, 2004)
    » HP MPIO Basic Failover v1.0 for EVA Arrays (June, 2004)
    » HP MPIO Basic Failover v1.0 for MSA Arrays (June, 2004)
  - **Description**
  - These Device Specific Modules can be downloaded directly from the links provided above.
  - Windows Versions Supported:
  - Microsoft Windows Server 2003 (32 bit EE and 64 bit EE only)
  - HP Arrays Supported:
  - HP MPIO Basic Failover v1.0 for EVA Arrays
    - EVA5000 and EVA3000 with VCS v3.010, v3.014, or or later
  - HP MPIO Basic Failover v1.0 for MSA Arrays
    - MSA1000 v4.36 or later
  - HP MPIO Basic Failover v1.0 for XP Arrays
    - XP 48/512, 128/1024

StorPort

- StorPORT – targeted support in June 2004

Microsoft claims that Storport is 30 percent to 50 percent more efficient than SCSIport and can handle more I/Os per second at a lower CPU utilization

- SCSIport driver has architectural limitations
  - Max 254 I/O requests per SCSI adapter
  - sequential (or half-duplex) I/O functions
  - excessive load at high IRQ levels
  - high buffer-processing overhead
  - I/O queue management limitations

Exchange 2003 SP1

- 1018 fixer
  - 40% of -1018s stem from single-bit errors in 1 page
  - Two page checksums allow recovery from single-bit errors
  - Requires database format change; some back-compat issues
- VSS
  - Now supports for incremental and differential backups
  - Bug fixes & perf improvements
- Recovery Storage Groups
  - Automatically merge mailbox data back when recovering

Summary

Agenda - Summary

- Here's what we covered

1. Exchange Storage Fundamentals
   - Workload requirements
   - User profiles
2. Design the EVA for Exchange
   - Disk Groups
   - Case Study
   - How to Validate the Storage Subsystem
3. Windows Server 2003 Features
   - VSS
   - StorPort, MPIO
   - Exchange SP1

HP WORLD 2004

Solutions and Technology Conference & Expo

Co-produced by:

interex

shared knowledge • shared power

encompass

AN HP USER GROUP

RECOMMENDED TRAINING VENUE FOR THE

HP Certified Professional