

ORACLE®



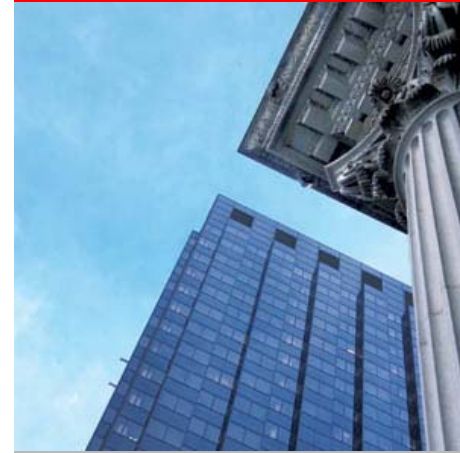
ORACLE®

RAC on Linux Forum

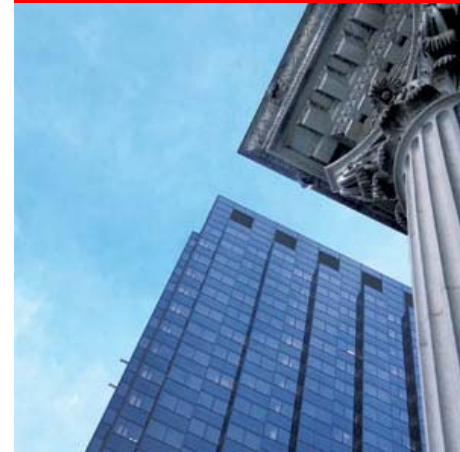
Alejandro Vargas
Oracle Israel
Principal Support Consultant

Agenda

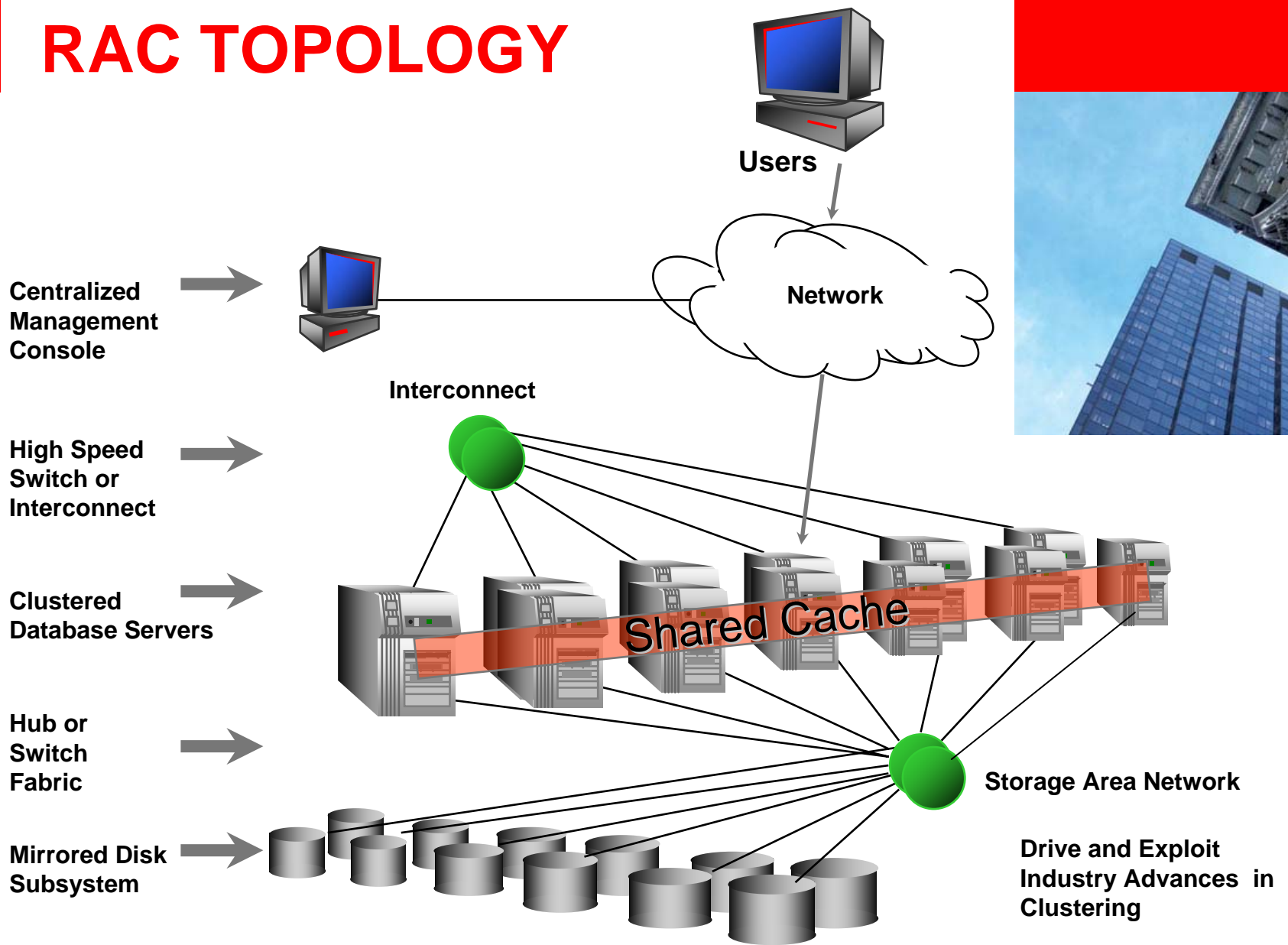
- RAC Topology
- OS Configuration for RAC
- Storage
- Network
- Automatic Storage Management (ASM)
- Database
- Validated RAC on Linux Configurations
- Other important topics (quick review)



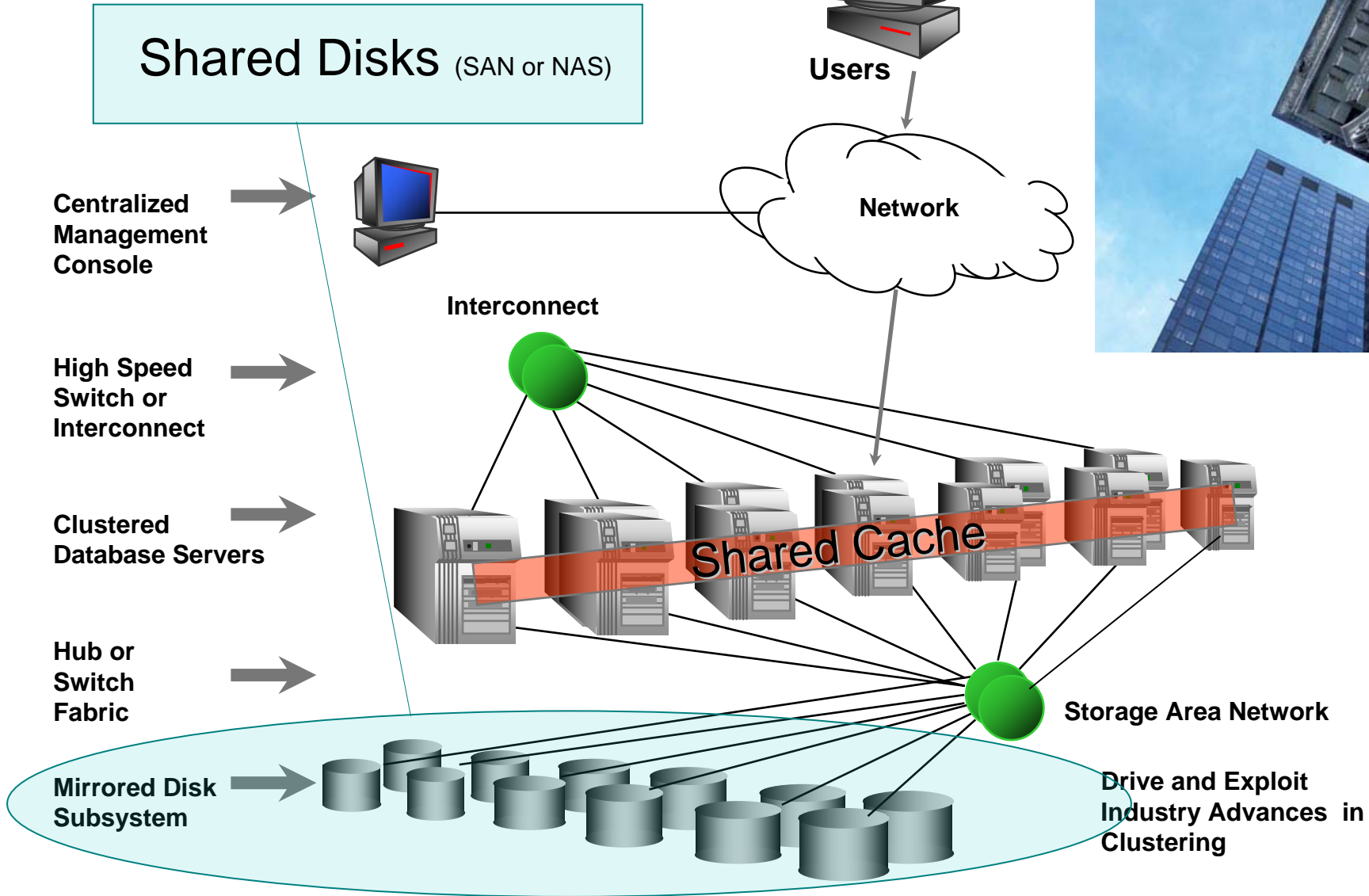
RAC TOPOLOGY



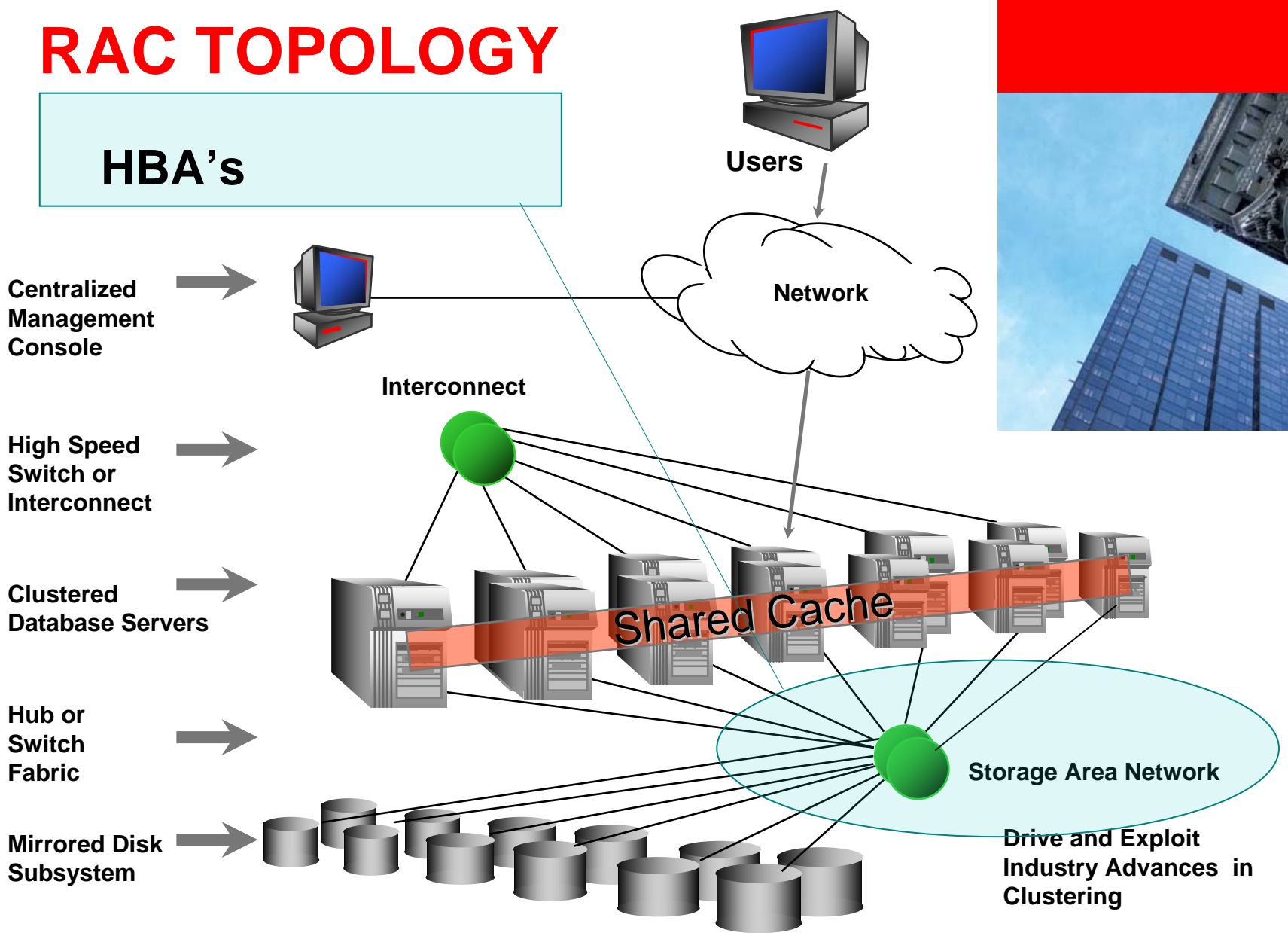
RAC TOPOLOGY



RAC TOPOLOGY

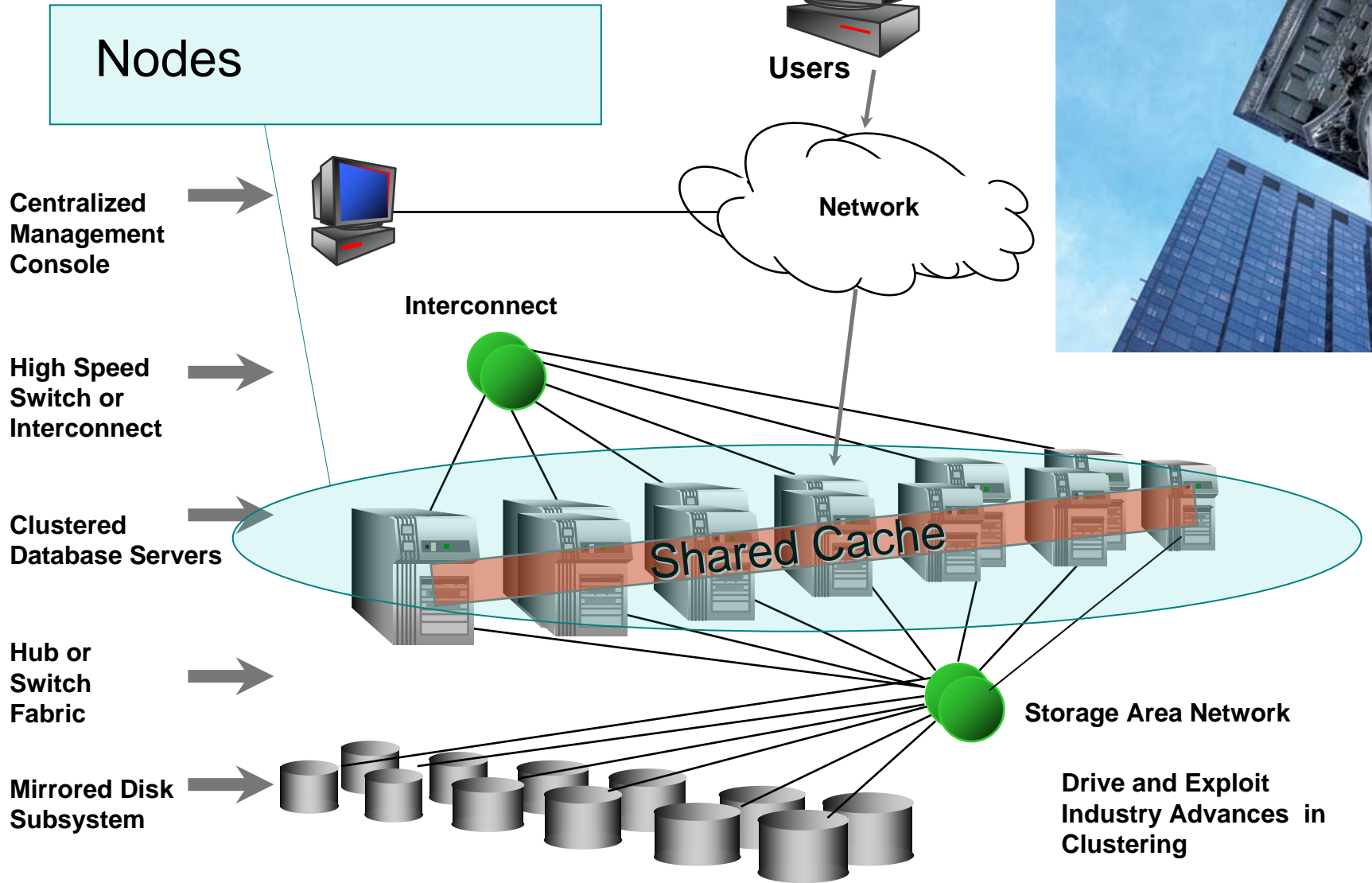


RAC TOPOLOGY

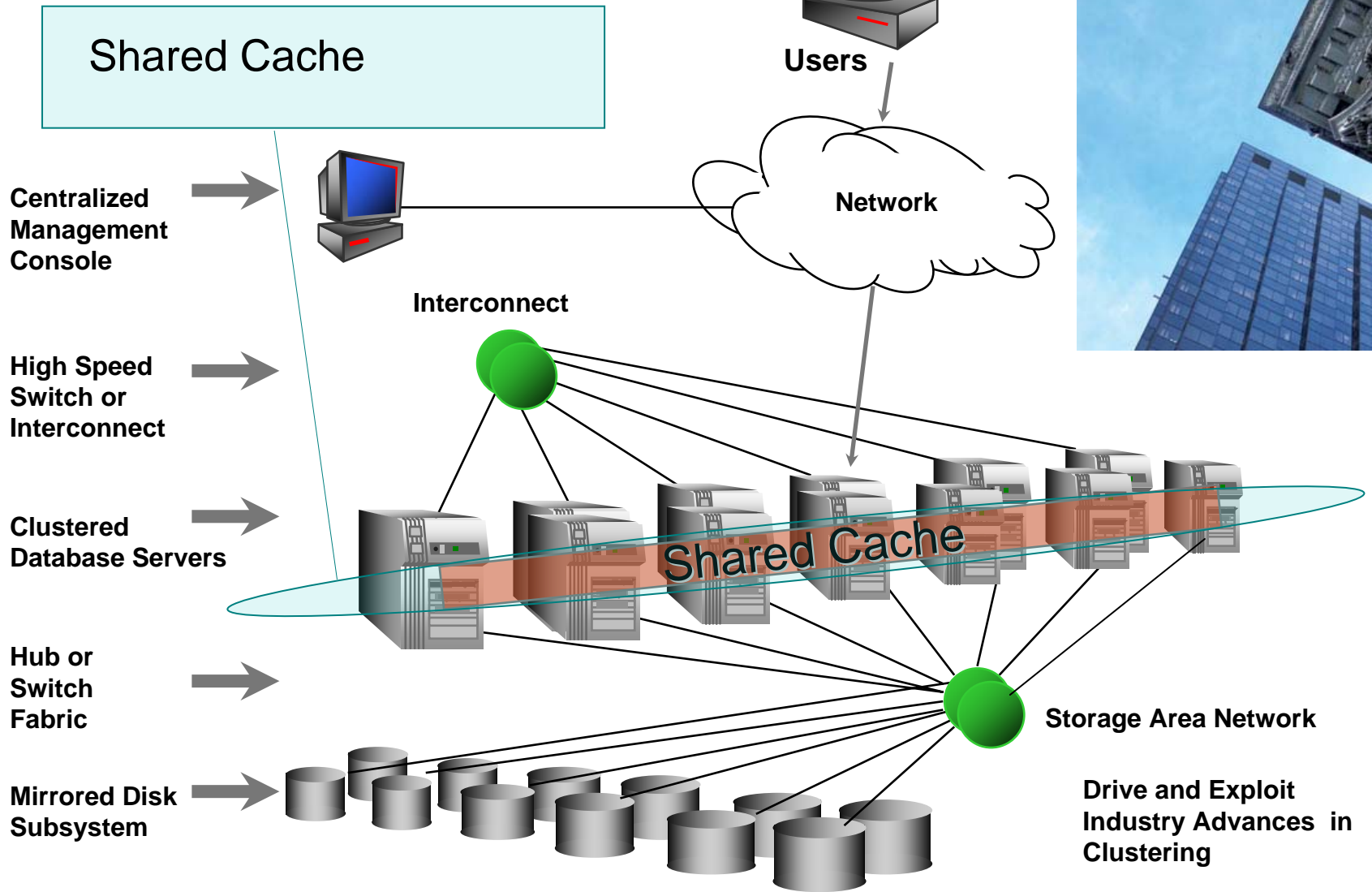


Drive and Exploit
Industry Advances in
Clustering

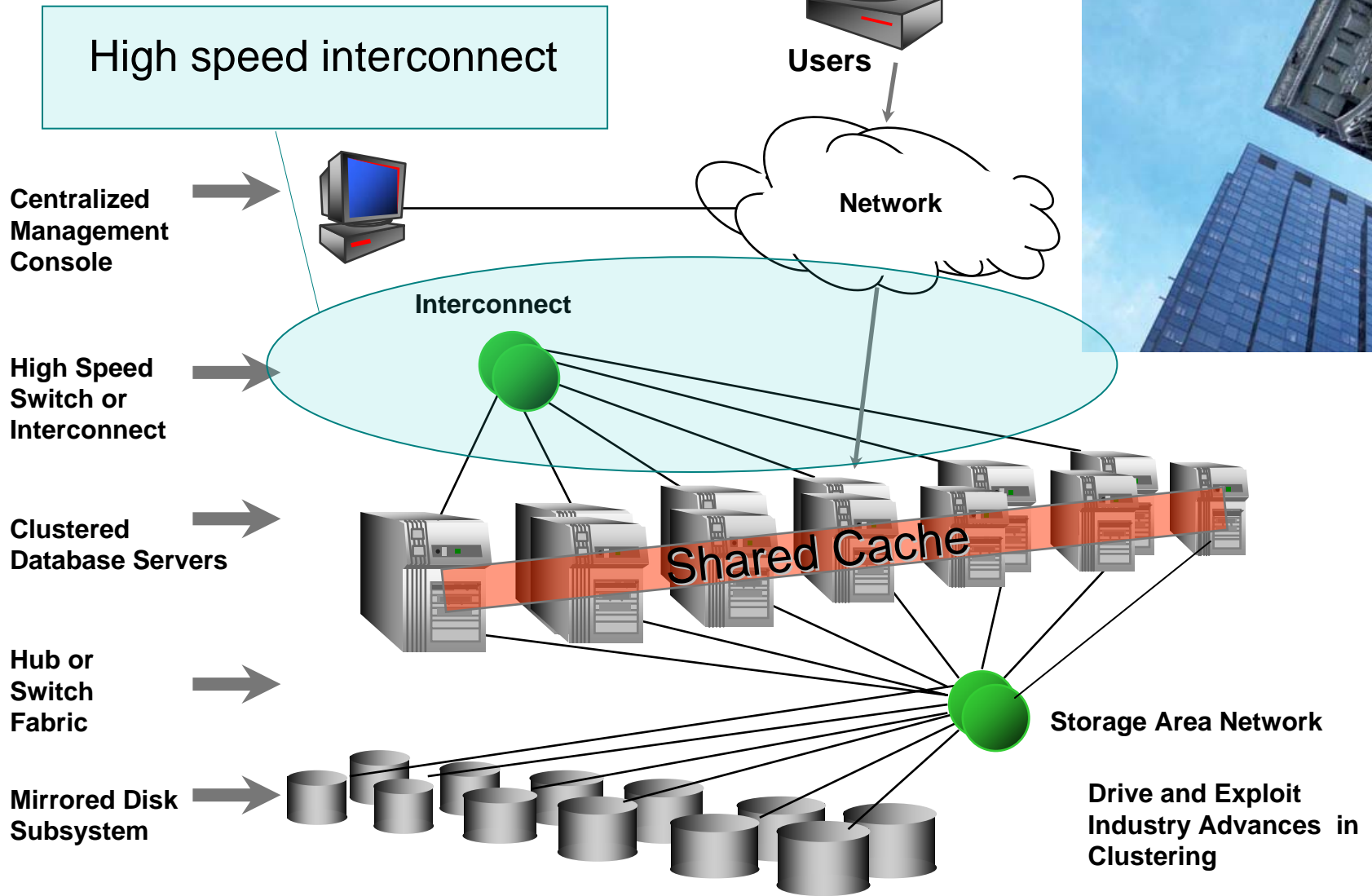
RAC TOPOLOGY



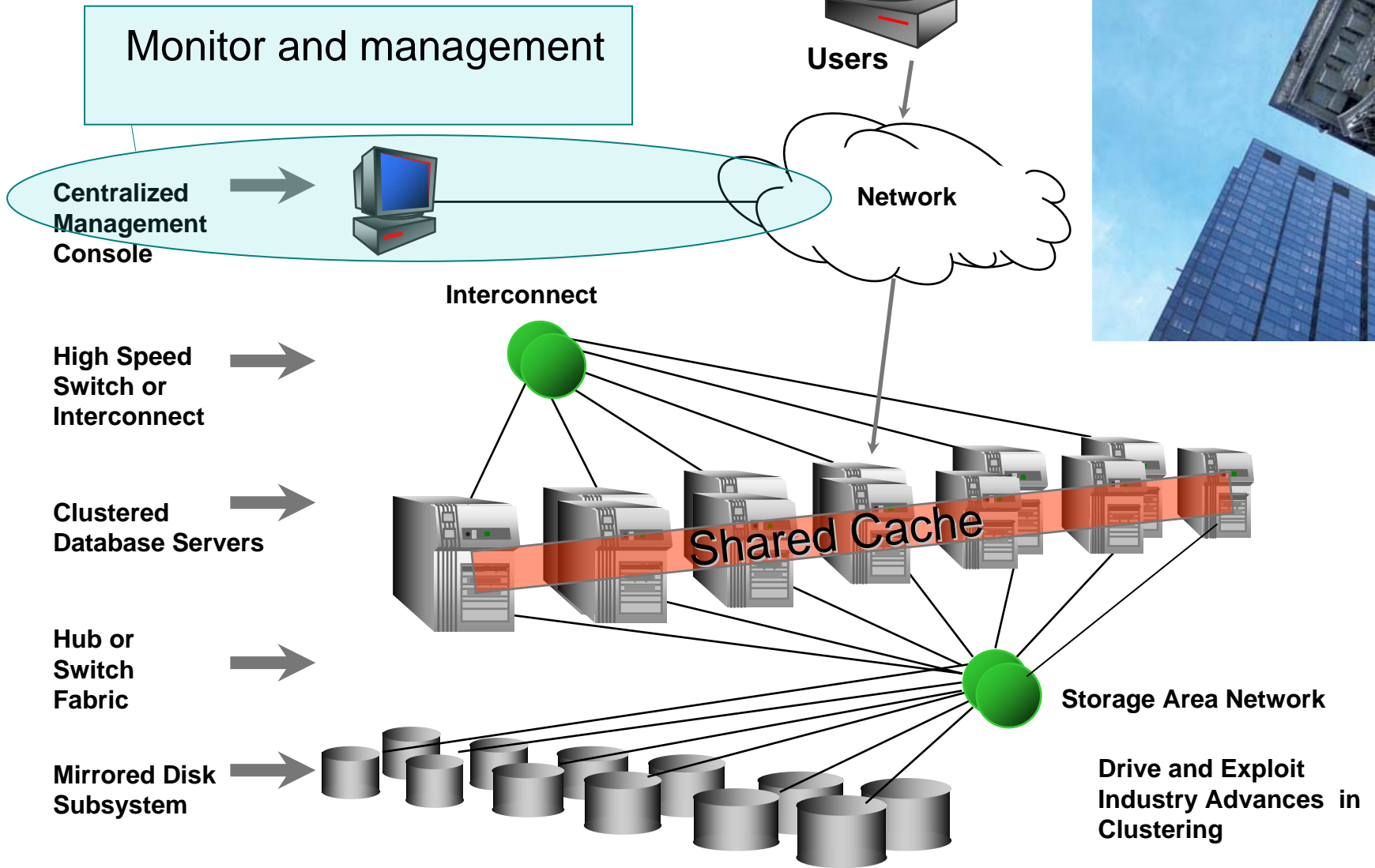
RAC TOPOLOGY



RAC TOPOLOGY



RAC TOPOLOGY



RAC TOPOLOGY

Load Balance and TAF

Centralized Management Console

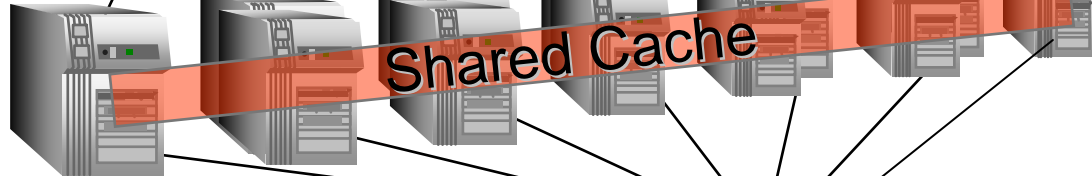


High Speed Switch or Interconnect

Interconnect



Clustered Database Servers



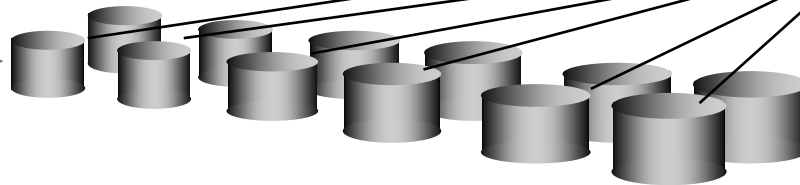
Shared Cache

Hub or Switch Fabric

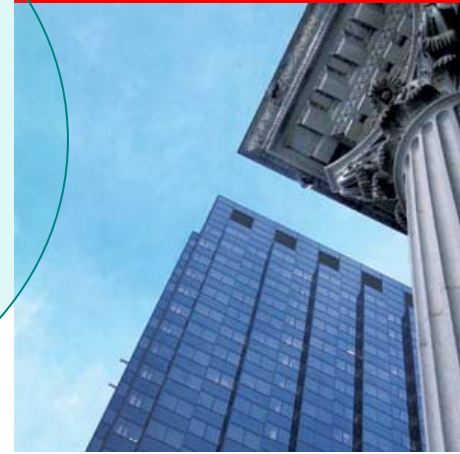
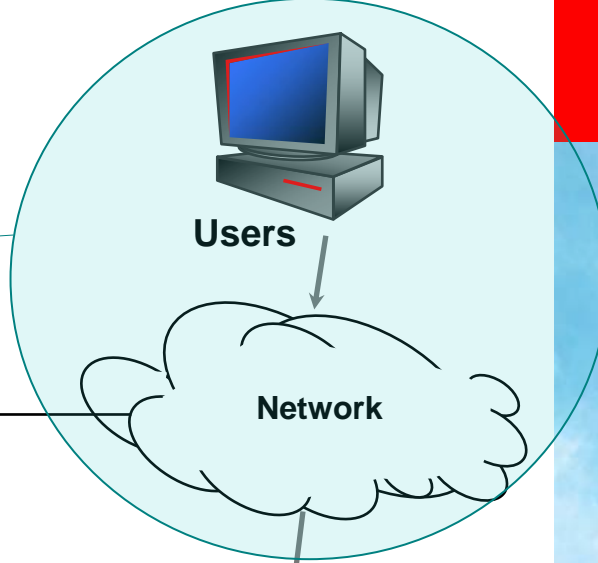


Storage Area Network

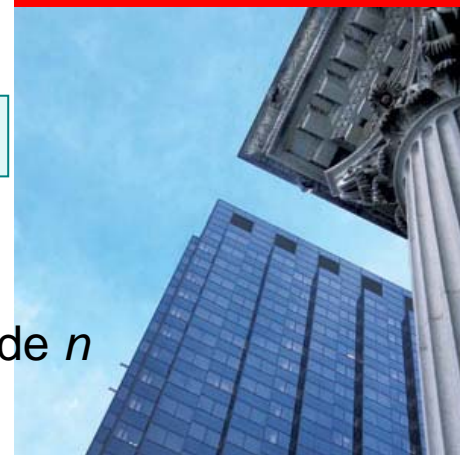
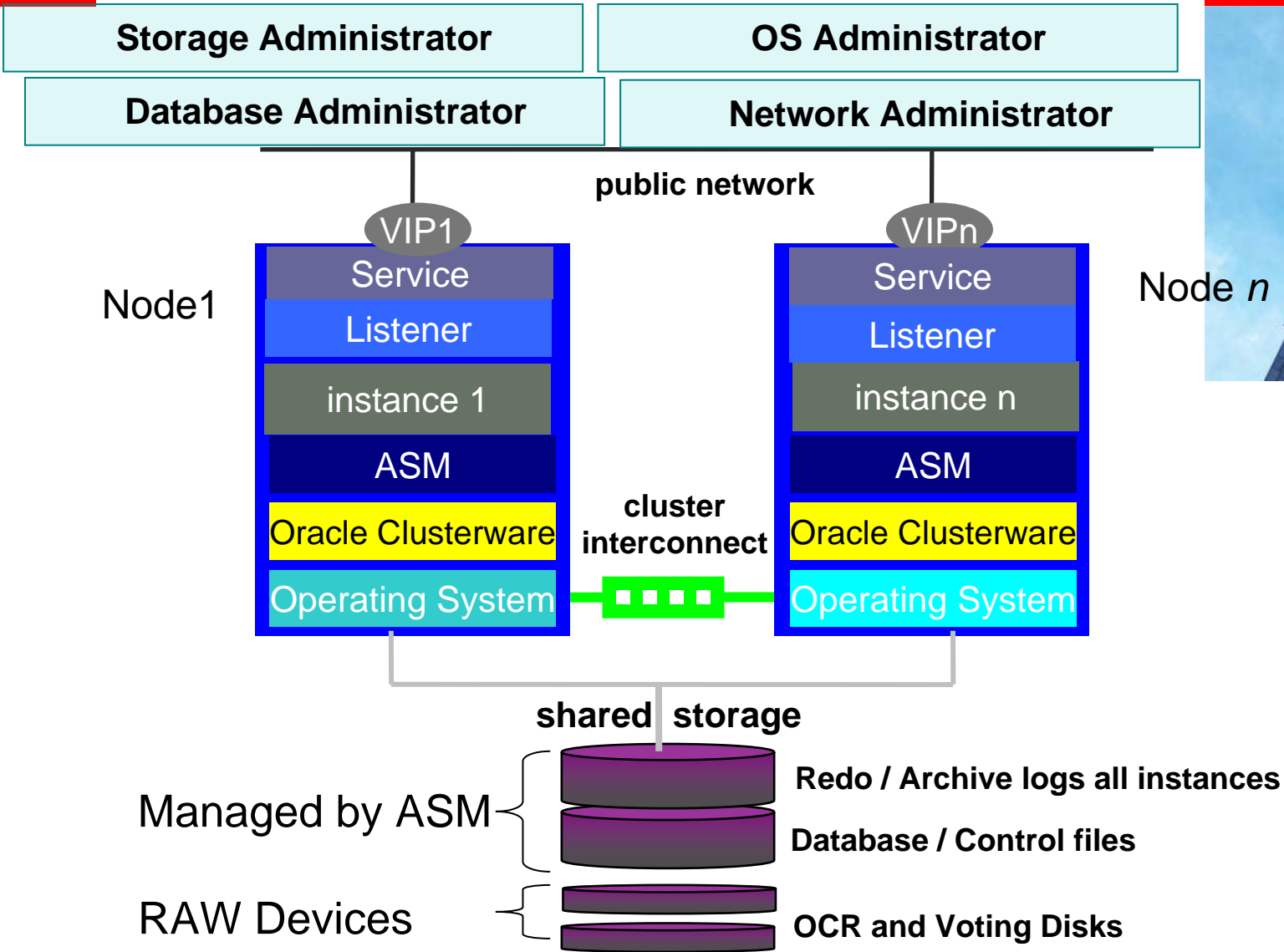
Mirrored Disk Subsystem



Drive and Exploit Industry Advances in Clustering



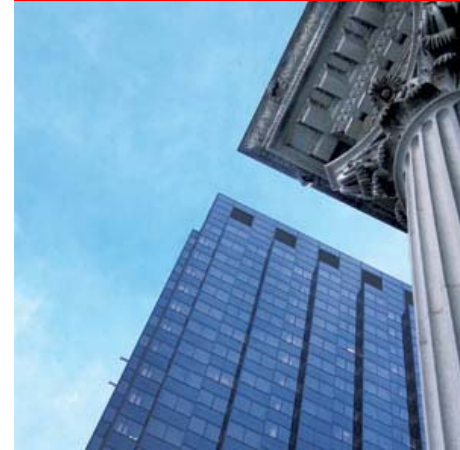
The RAC Team



OS Configuration

Before Starting Check List

- Topology Diagram
- Set Targets for:
 - Availability
 - Scalability
 - Throughput
- Gather the team and explain the configuration in detail
- With your Team, size the configurable components
- Organize step by step check lists of every task
- Use the check lists on every implementation stage



OS Configuration

OS Pre Install Configuration Items

Step by Step Install guides on Oracle Network:

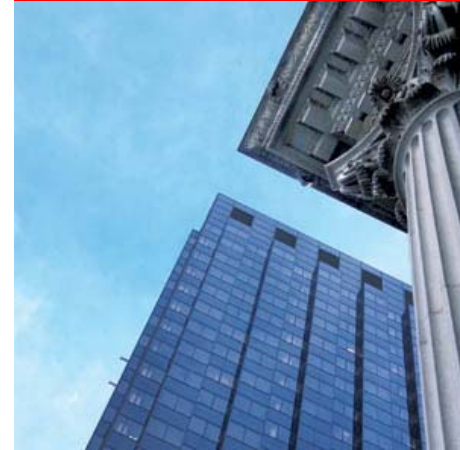
- <http://www.oracle.com/technology/tech/linux/install/index.html>
- http://www.oracle.com/technology/pub/articles/smiley_rac10g_install.html
- <http://static7.userland.com/oracle/gems/alejandroVargas/StepbyStepRAConLinux3.pdf>



OS Configuration

Cluster Verification Utility (CVU)

- Learn and master CVU
- Will certify your environment to the standards
- Will assure stability and productivity



OS Configuration

Deployment of cluvfy

- Install only on local node. Tool deploys itself on remote nodes during execution, as required.

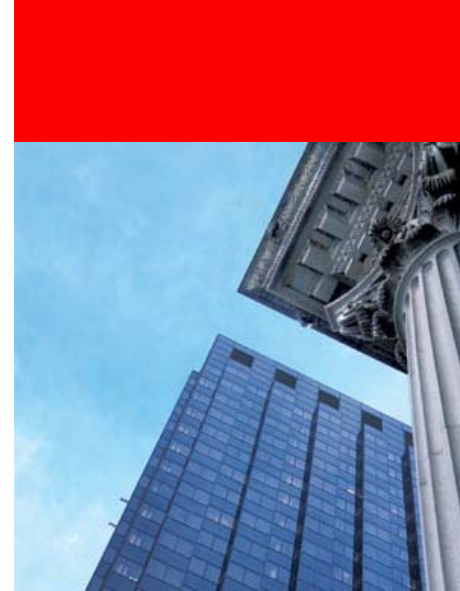
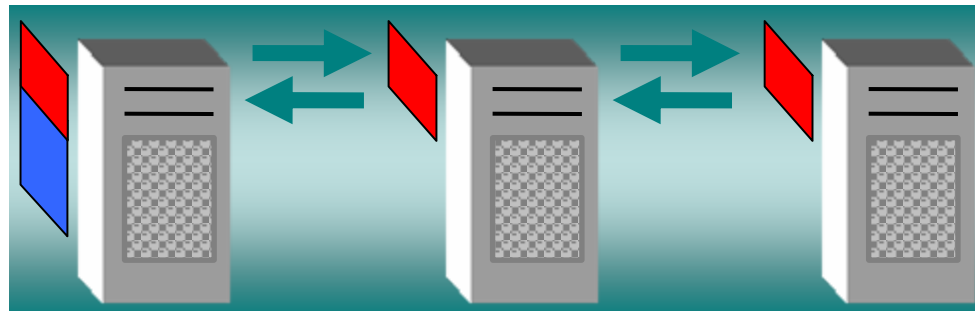
- User installs on local node

- Tool copies the required bits to the remote nodes

- Issues verification command for multiple nodes

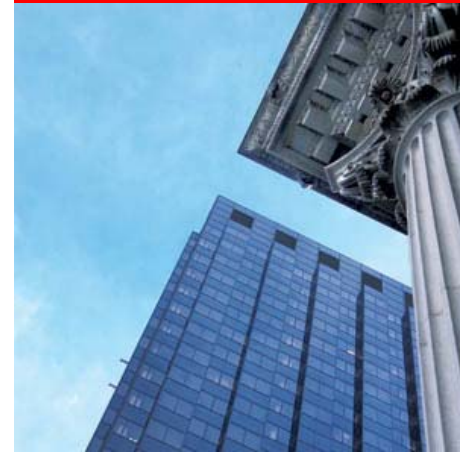
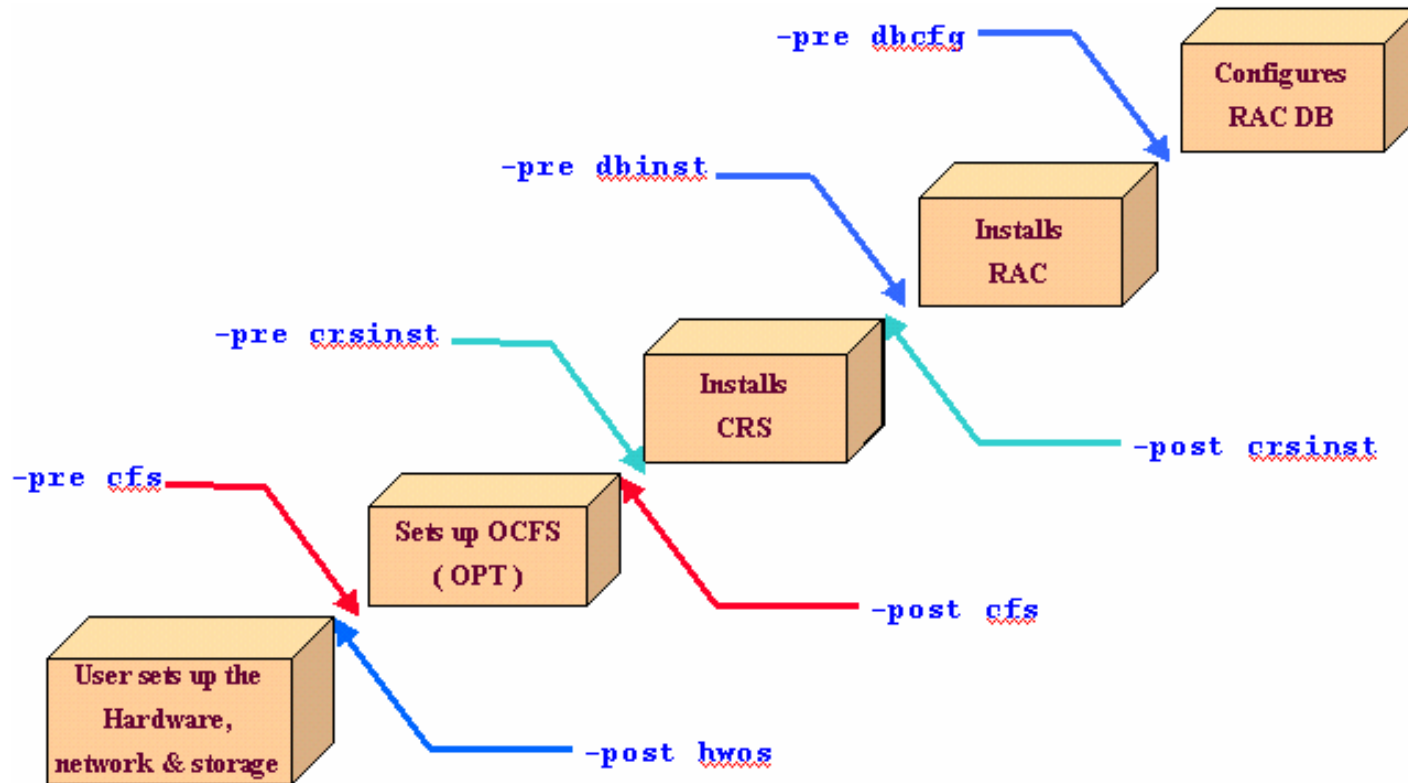
- Executes verification tasks on all nodes and generates report

CVU



OS Configuration

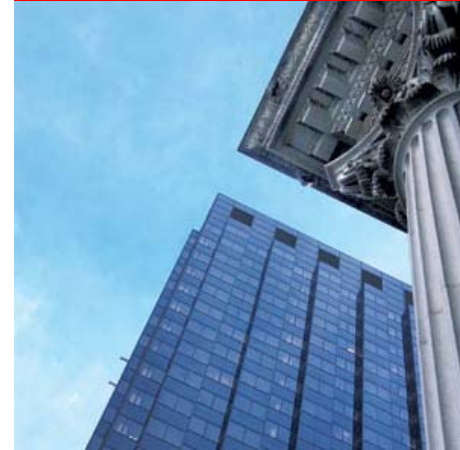
Cluvfy Stage List - Graphical



OS Configuration

CVU locations

- OTN: <http://otn.oracle.com/RAC>
- Oracle DVD
 - [clusterware/cluvfy/runcluvfy.sh](#)
 - [clusterware/rpm/cvuqdisk-1.0.1-1.rpm](#)
- CRS Home
 - [<crs_home>/bin/cluvfy](#)
 - [<crs_home>/cv/rpm/cvuqdisk-1.0.1-1.rpm](#)
- Oracle Home
 - [\\$ORACLE_HOME/bin/cluvfy](#)

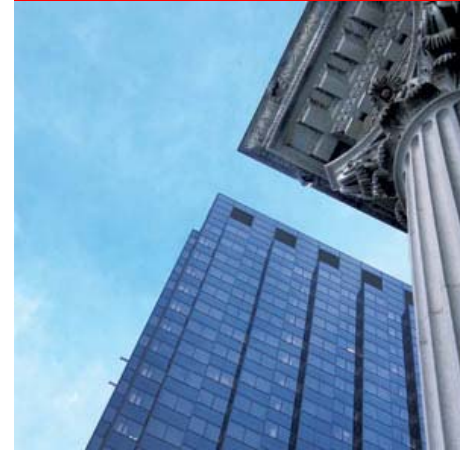


OS Configuration

OS Pre Install Configuration Items

- Validate compatibility matrix for RAC
- Configure Public IP
- Configure Private IP (172.16.* or 192.168.* or 10.10.*)
- Configure Virtual IP on DNS or /etc/hosts only
- Check ping from all nodes to all nodes from both public and private IP's
- Configure the Hangcheck Timer on all nodes
- Configure SSH
- Configure User Equivalence
- Install and Configure ASMLib
- Configure Raw Devices or OCFS2 for OCR/Voting/ASM spfile

NIC's MUST have the same Names on all nodes

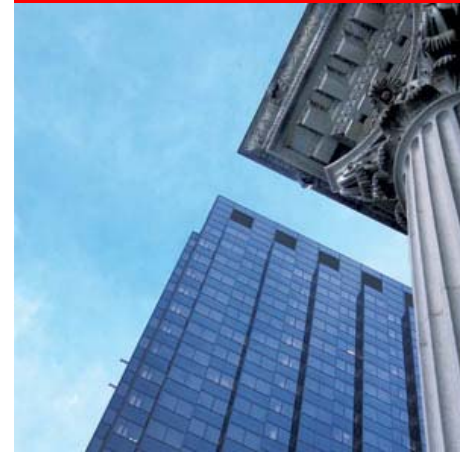
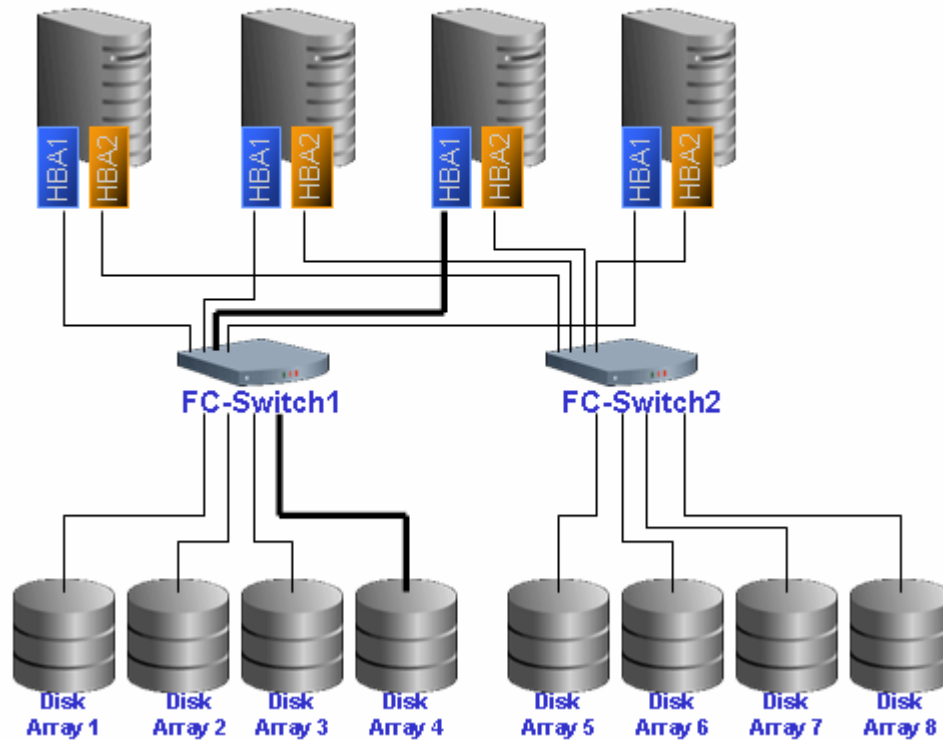


STORAGE



Storage

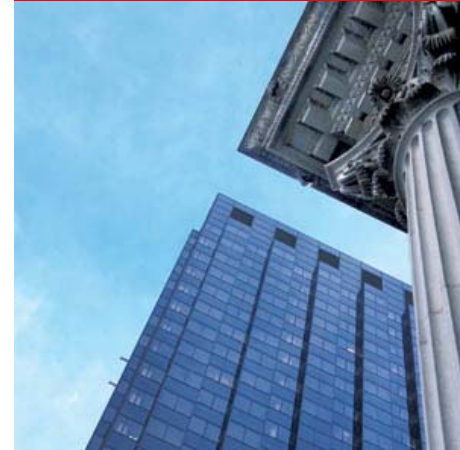
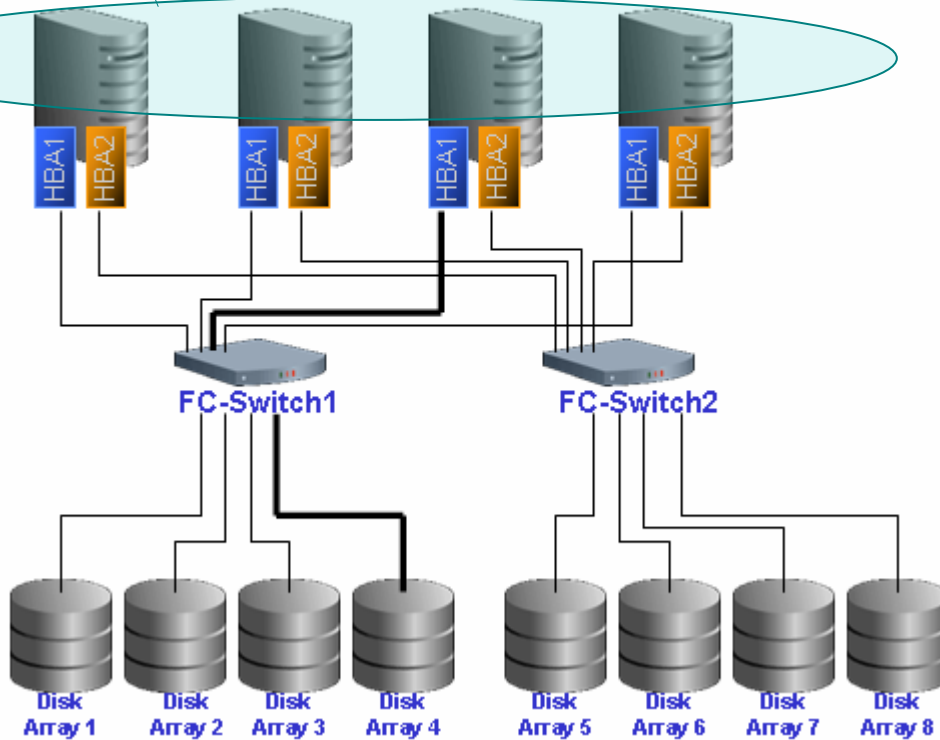
Balanced System



Storage

CPU: Quantity and speed

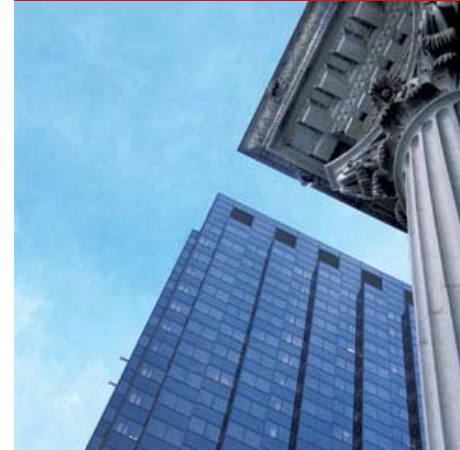
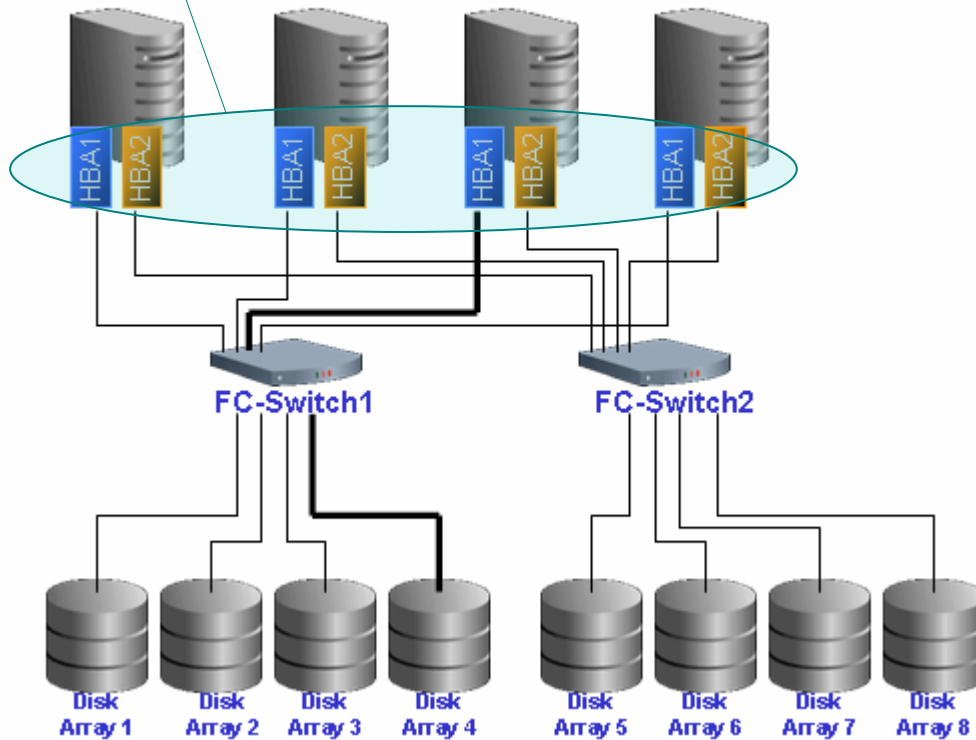
Balanced System



Storage

HBA (Host Bus Adapter):
Quantity and speed

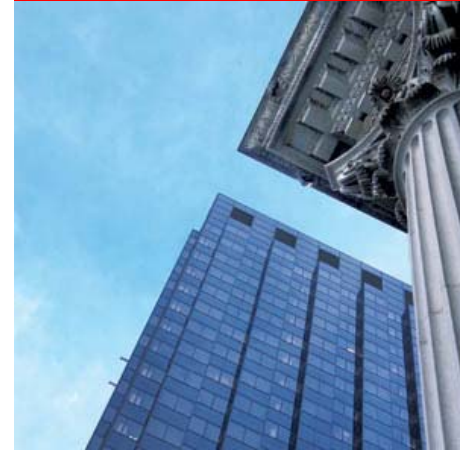
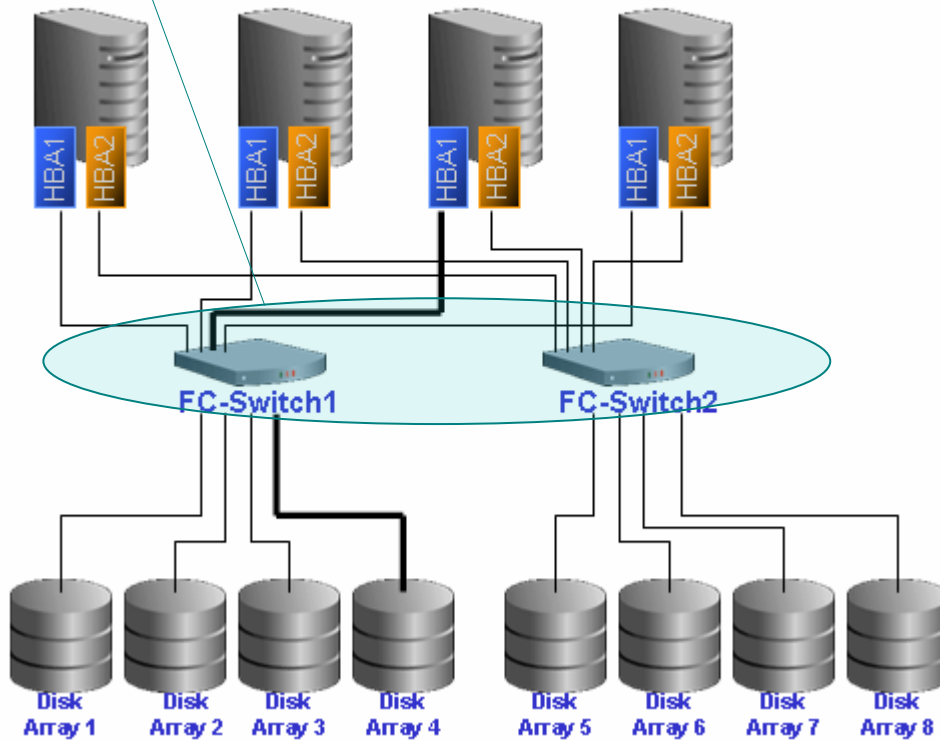
Balanced System



Storage

Switch speed, Controller:
Quantity and speed

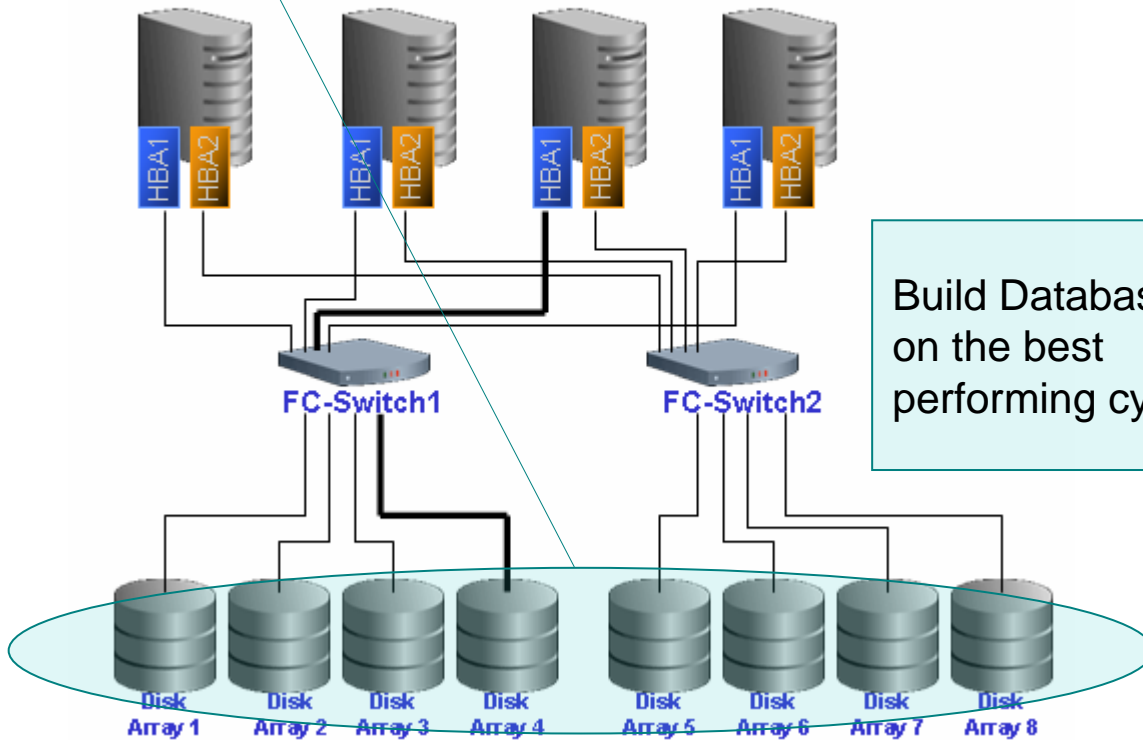
Balanced System



Storage

Disks:
Quantity and speed

Balanced System

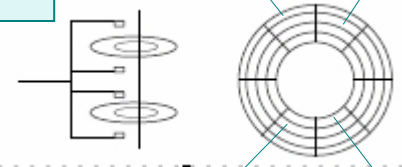


Build Database LUN's
on the best
performing cylinders



Online Redo – Undo – Temp – DB Files

Highest
20% Fastest
40%



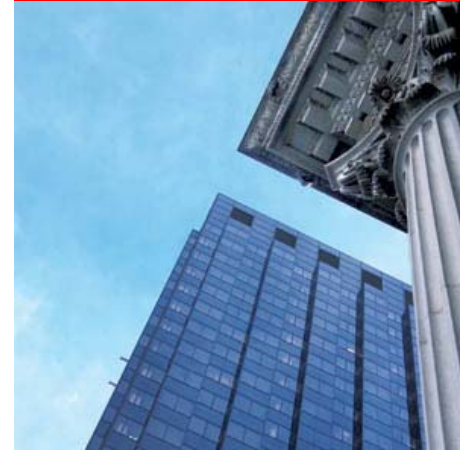
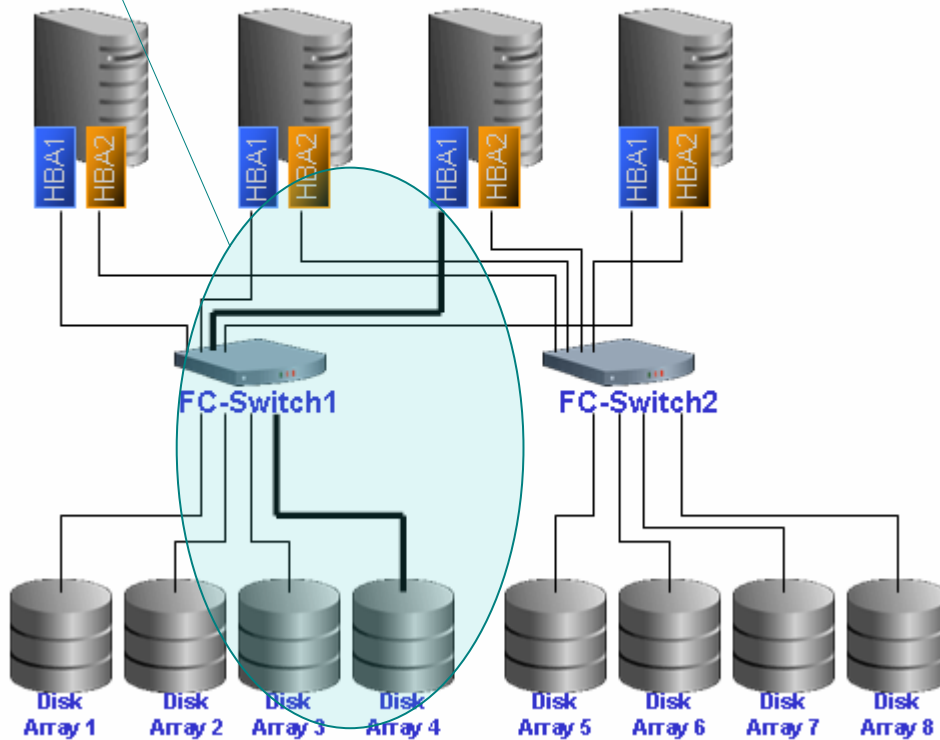
Slow
20% Slowest
20%

Archived logs – Dump destinations

Storage

The weakest link” defines the throughput

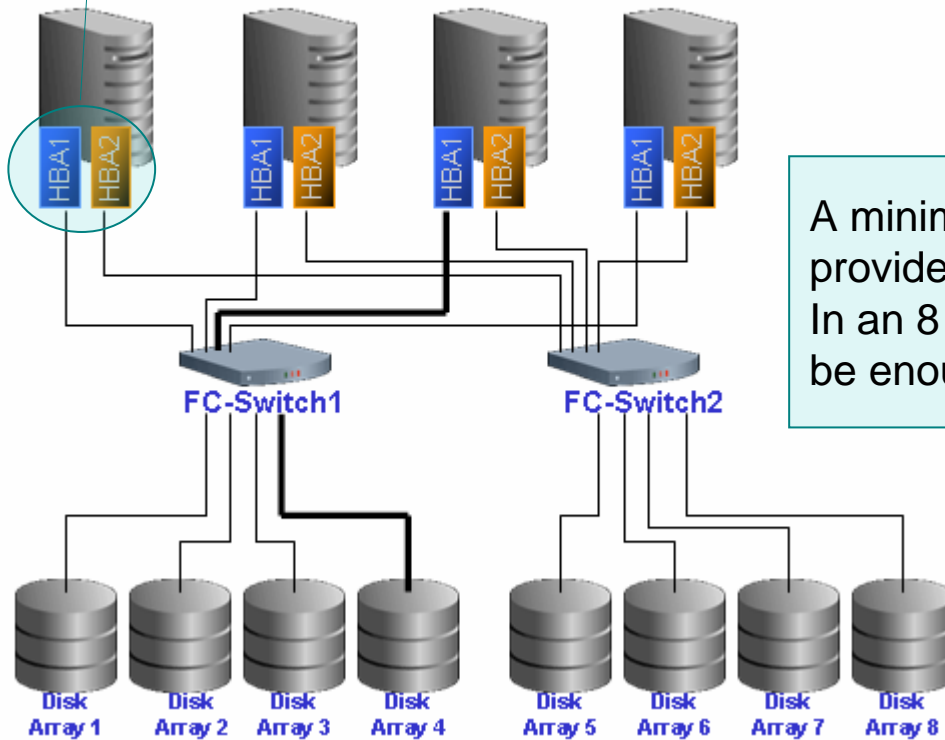
Balanced System



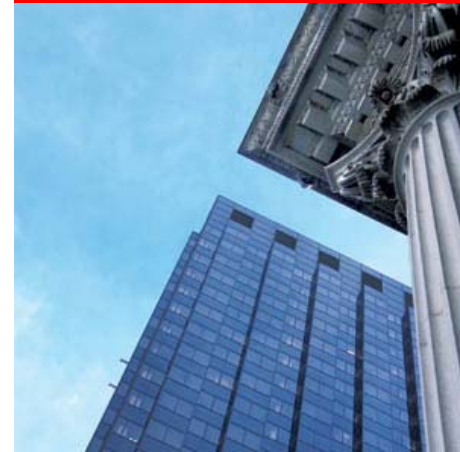
Storage

Most HBA's are 2 or 4 Gbit/s
Aim for 1Gbit/s/core

Balanced System



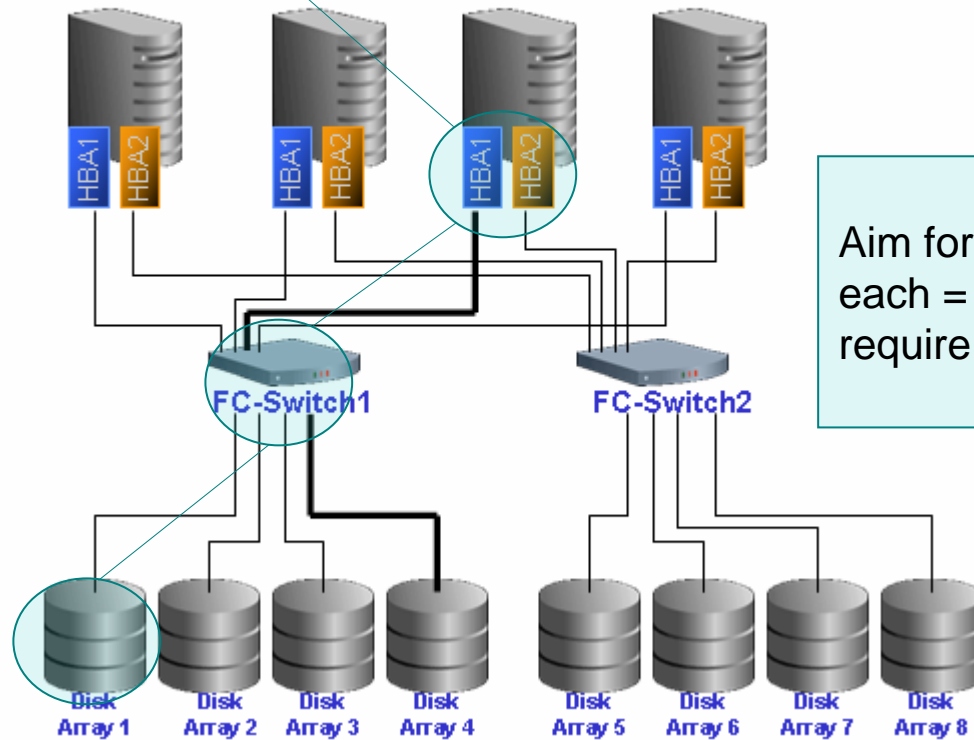
A minimum of 2HBA's per node is required to provide redundancy.
In an 8 core Node 2 HBA's 4Gbit/s each can be enough.



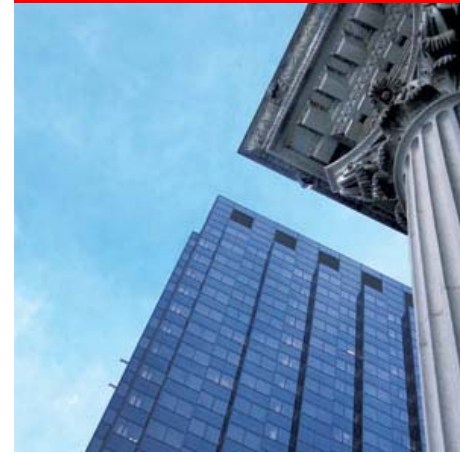
Storage

It is important that bandwidth is sustained all the way from disk to host

Balanced System



Aim for at least 5 spindles per core (20MB/s each = 100MBs/core) On OLTP you may require even more spindles.



Storage

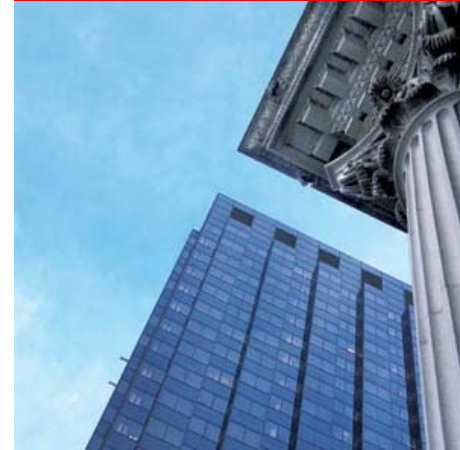
Multi Path Configuration

Device driver combines multiple paths to the same device

Two HBAs become one virtual HBA (Host Bus Adapter)

Failover, Bandwidth aggregation, path rediscovery

Third-party (HP, EMC, IBM, Sun, HDS, Veritas, Qlogic) multipathing on Linux



Storage

Grid Components Rough Sizing numbers

Component	Throughput Performance	
	theory (Bit/s)	maximal Byte/s
HBA	1/2Gbit/s	100/200 Mbytes/s
16 Port Switch	8 x 2Gbit/s	1600 Mbytes/s
Fibre Channel	2Gbit/s	200 Mbytes/s
Disk Controller	2Gbit/s	200 Mbytes/s
GigE NIC	1Gbit/s	80 Mbytes/s
Infiniband	10Gbit/s	890 Mbytes/s
CPU		200MB/s

Storage

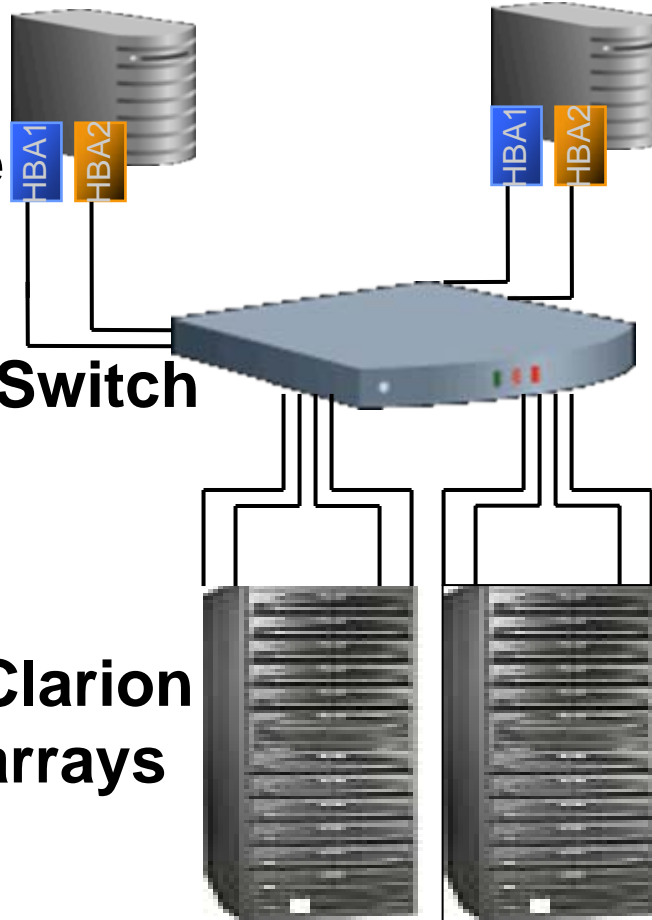
Sample Building Block 2 Node Configuration

2 Dell 6850 2 CPU each

2 HBA's per node

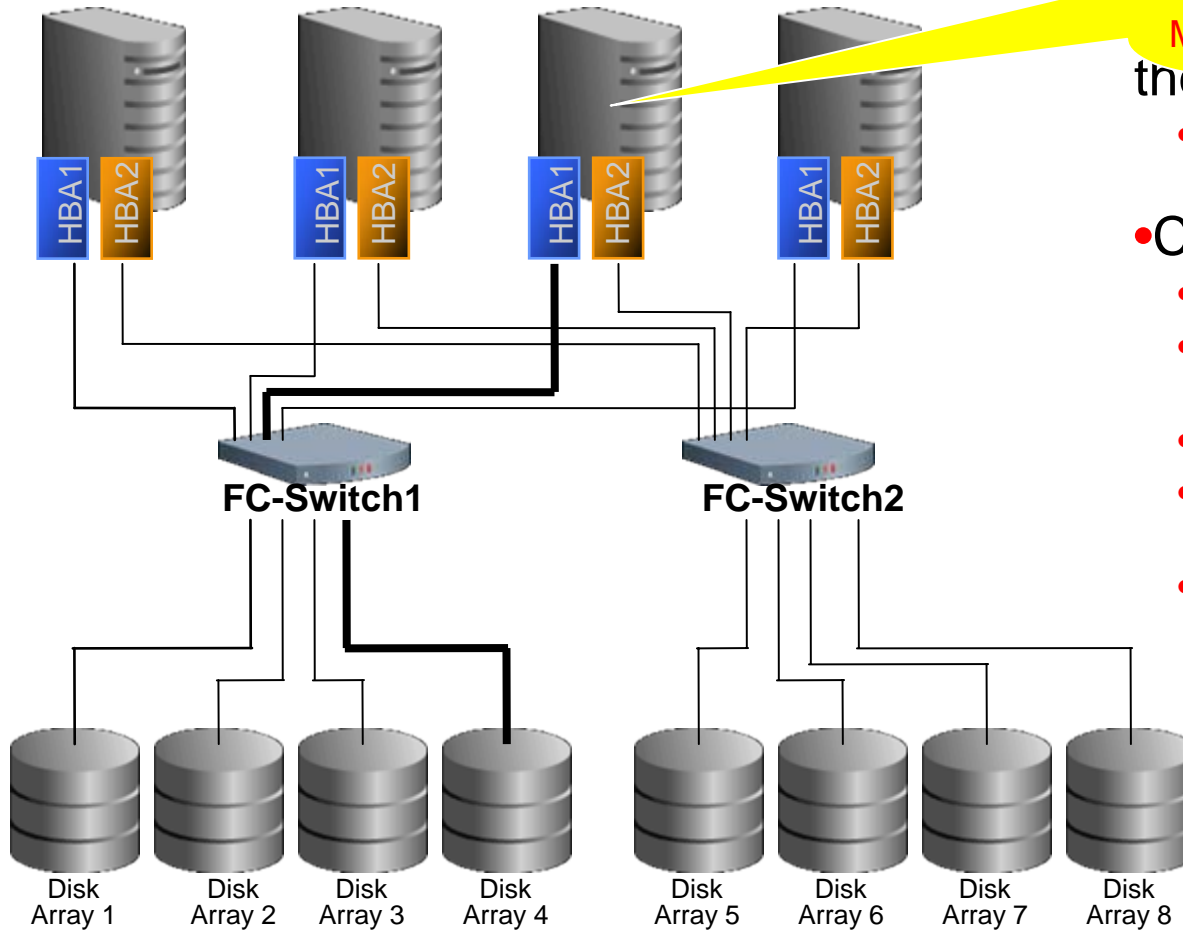
1 FC Switch

2 EMC Clarion
CX500 arrays



Storage

Sizing Example



Each machine has 2 CPUs

All four servers drive about

$$2 * 200\text{MB/s} * 4 = 1600$$

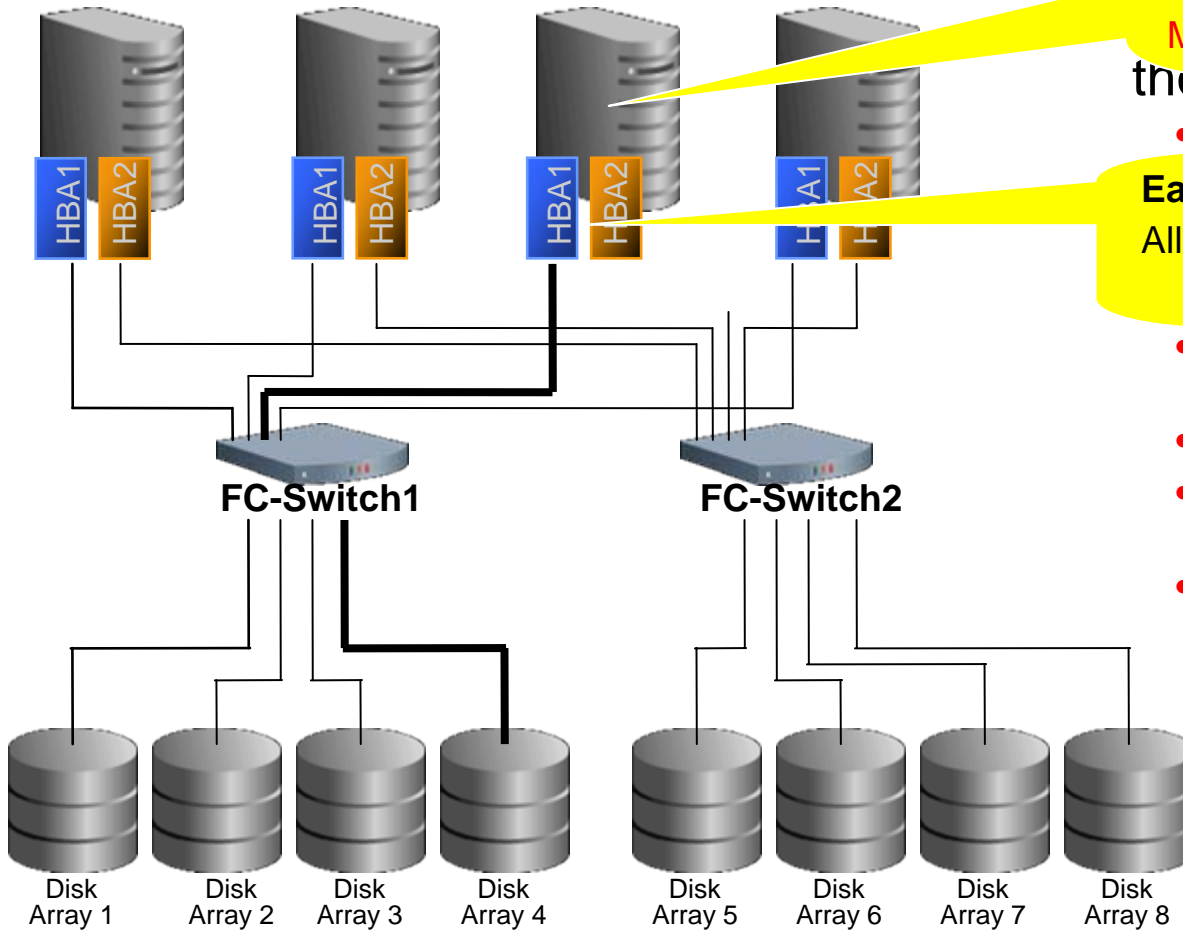
MB/s

the throughput

- Each building block is a balanced unit
- Components to consider:
 - CPU: Quantity and speed
 - HBA (Host Bus Adapter): Quantity and speed
 - Switch speed
 - Controller: Quantity and speed
 - Disk: Quantity and speed

Storage

Sizing Example



Each machine has 2 CPUs

All four servers drive about

$$2 * 200\text{MB/s} * 4 = 1600$$

MB/s

the throughput

- Each building block is a

Each machine has 2 Gb HBAs

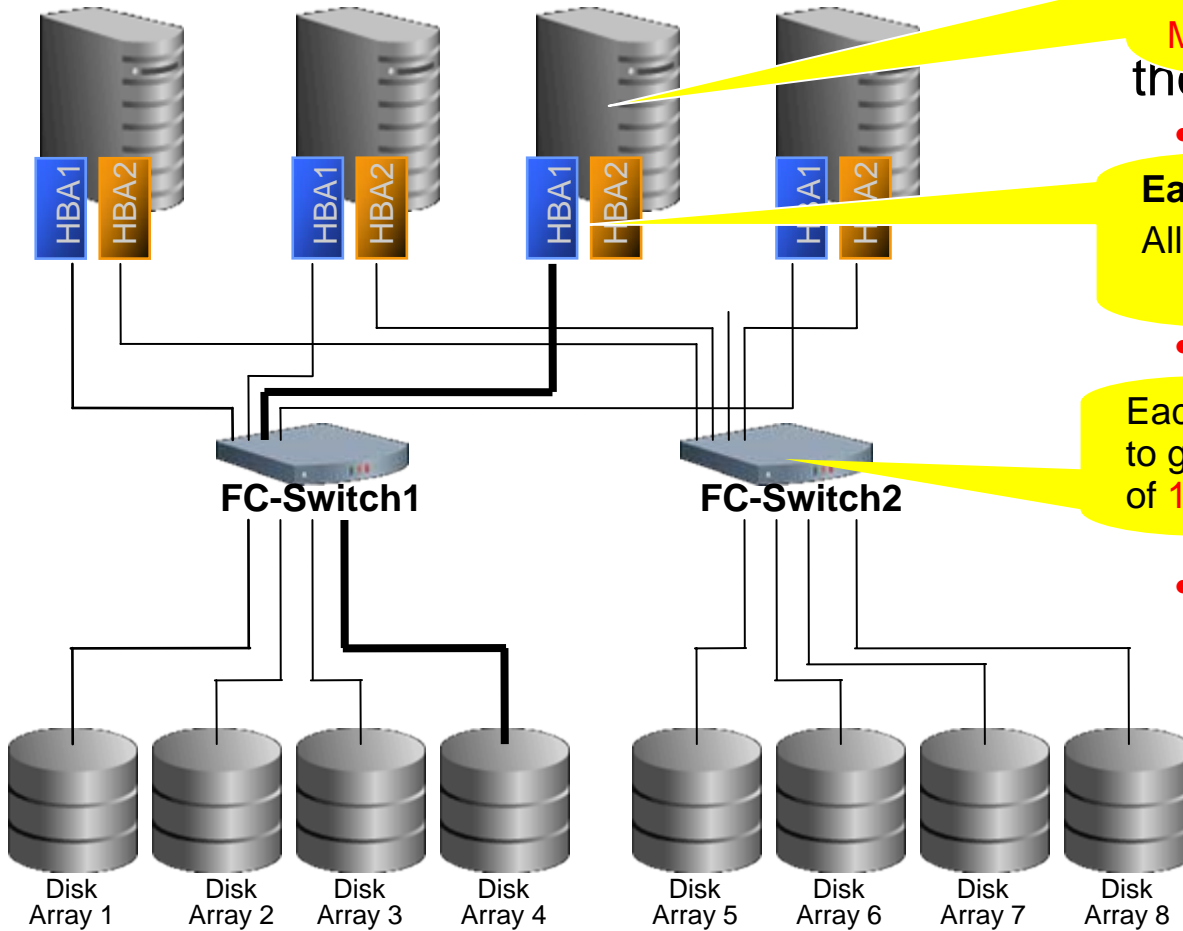
All 8 HBAs can sustain

$$8 * 200\text{MB/s} = 1600 \text{ MB/s}$$

- HBA (Host Bus Adapter): Quantity and speed
- Switch speed
- Controller: Quantity and speed
- Disk: Quantity and speed

Storage

Sizing Example



Each machine has 2 CPUs

All four servers drive about

$$2 * 200\text{MB/s} * 4 = 1600$$

MB/s

the throughput

- Each building block is a

Each machine has 2 Gb HBAs

All 8 HBAs can sustain

$$8 * 200\text{MB/s} = 1600 \text{ MB/s}$$

- HBA (Host Bus Adapter):

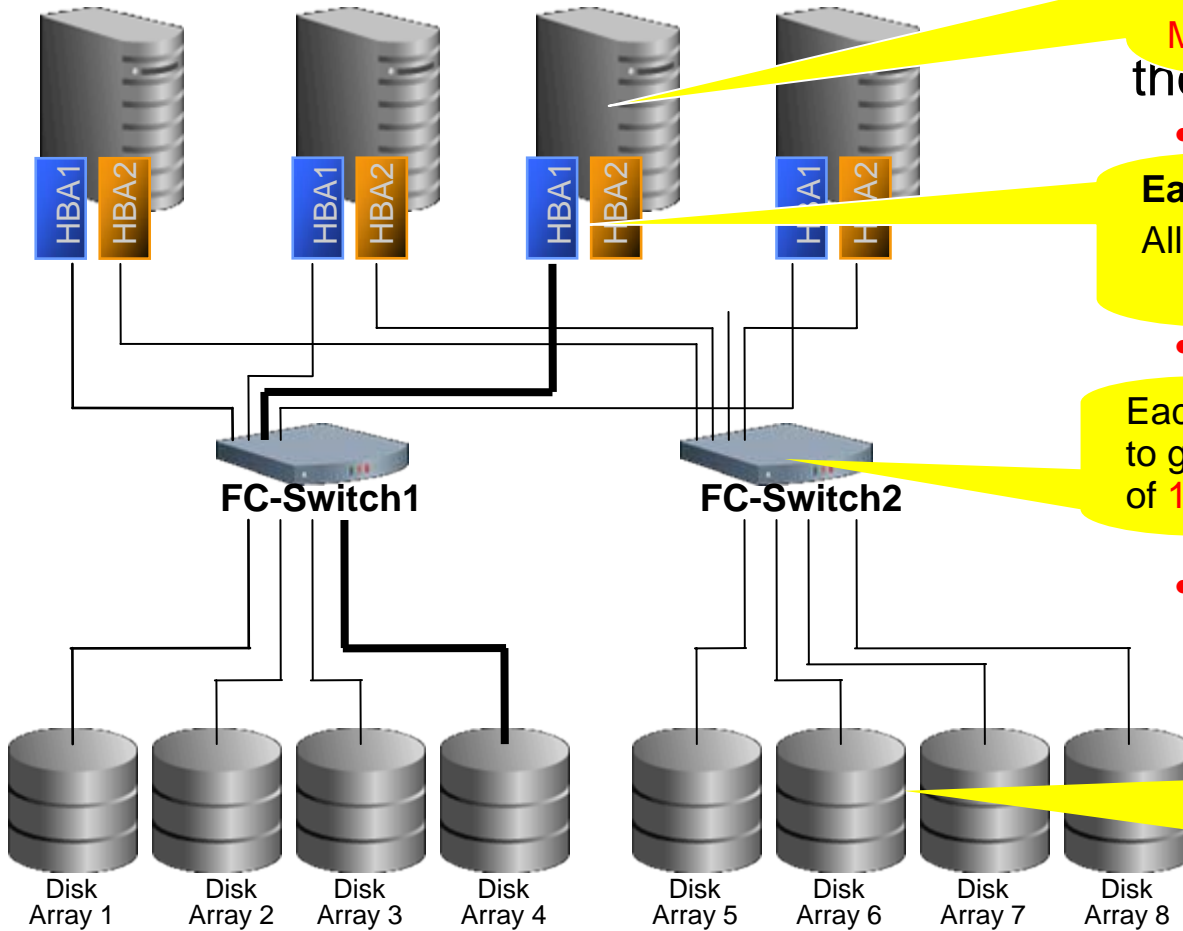
Each switch needs to support 800MB/s to guarantee a total system throughput of 1600 MB/s

speed

- Disk: Quantity and speed

Storage

Sizing Example



Each machine has 2 CPUs

All four servers drive about

$$2 * 200\text{MB/s} * 4 = 1600$$

MB/s

the throughput

- Each building block is a

Each machine has 2 Gb HBAs

All 8 HBAs can sustain

$$8 * 200\text{MB/s} = 1600 \text{ MB/s}$$

- HBA (Host Bus Adapter):

Each switch needs to support 800MB/s to guarantee a total system throughput of 1600 MB/s

speed

- Disk: Quantity and speed

Each disk array has one 2Gbit controller

All 8 disk arrays can sustain

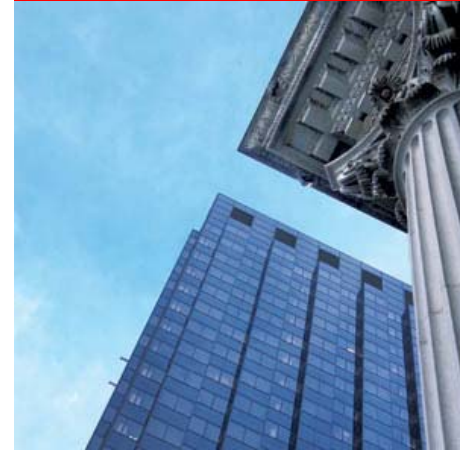
$$8 * 200\text{MB/s} = 1600 \text{ MB/s}$$

Storage

Storage considerations for DW

“I only need 10 disks to store my 1 TB database now that we have 130 GB disk drives!”

- Design for throughput not for capacity
- Keep it simple
- Apply the S.A.M.E methodology
 - Stripe And Mirror Everything
 - At the hardware level
 - Using ASM

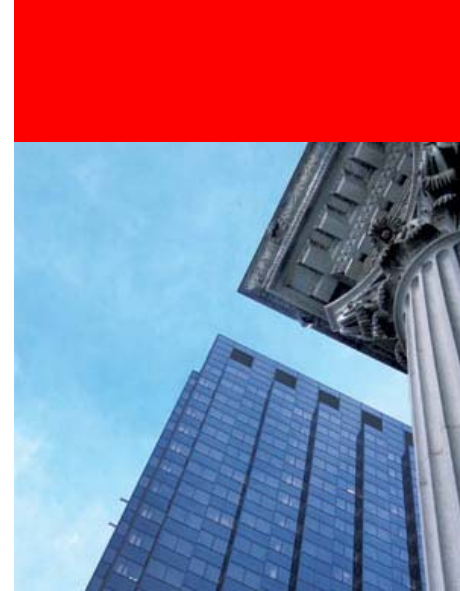


Storage

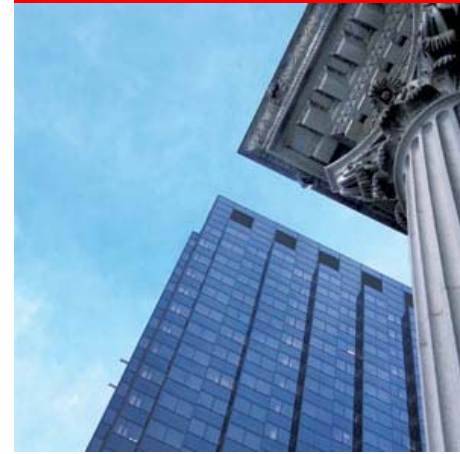
Storage Throughput Probe

- Before RAC implementation probe your storage
- Create the shared raw devices you will use on ASM
- Download, Install and Configure ORION

<http://www.oracle.com/technology/software/tech/orion/index.html>



NETWORK



Network

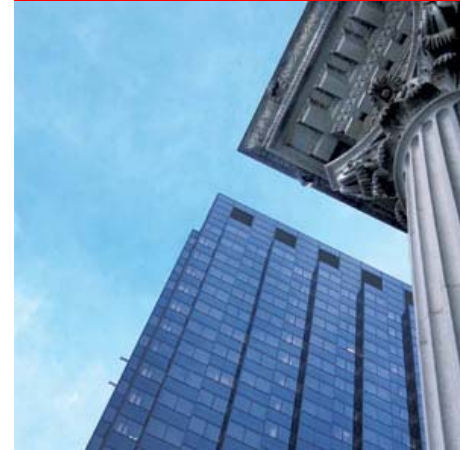
Interconnect Configuration

- Use UDP over Gigabit Ethernet
- OS Bonding/teaming to “virtualize” interconnect
 - Bond two or more GigE cards for performance reasons
 - Failover & Load-balancing & Improved bandwidth
- Set UDP send/receive buffers high enough
 - Platform dependant – typically 256K is adequate
 - `net.core.rmem_max`, `net.core.wmem_max`, `net.core.rmem_default`,
`net.core.wmem_default`
- Use a Switch
 - crossover cable **not** supported



ASM

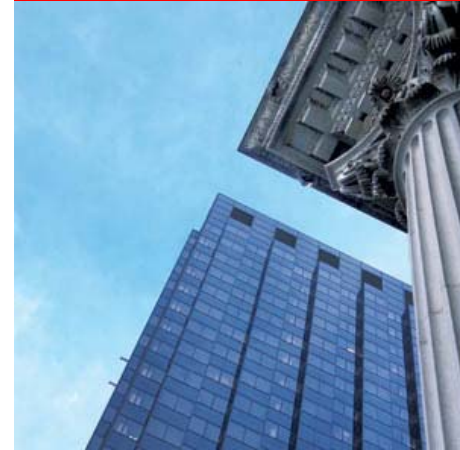
AUTOMATIC STORAGE MANAGEMENT



AUTOMATIC STORAGE MANAGEMENT

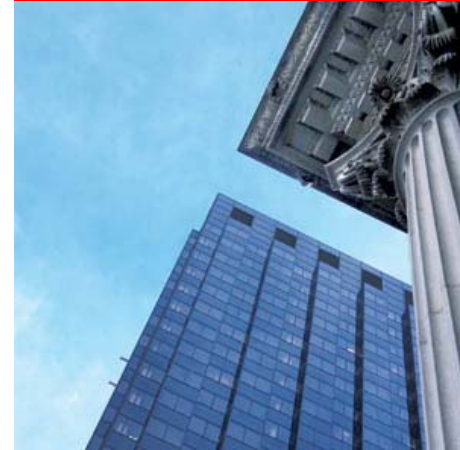
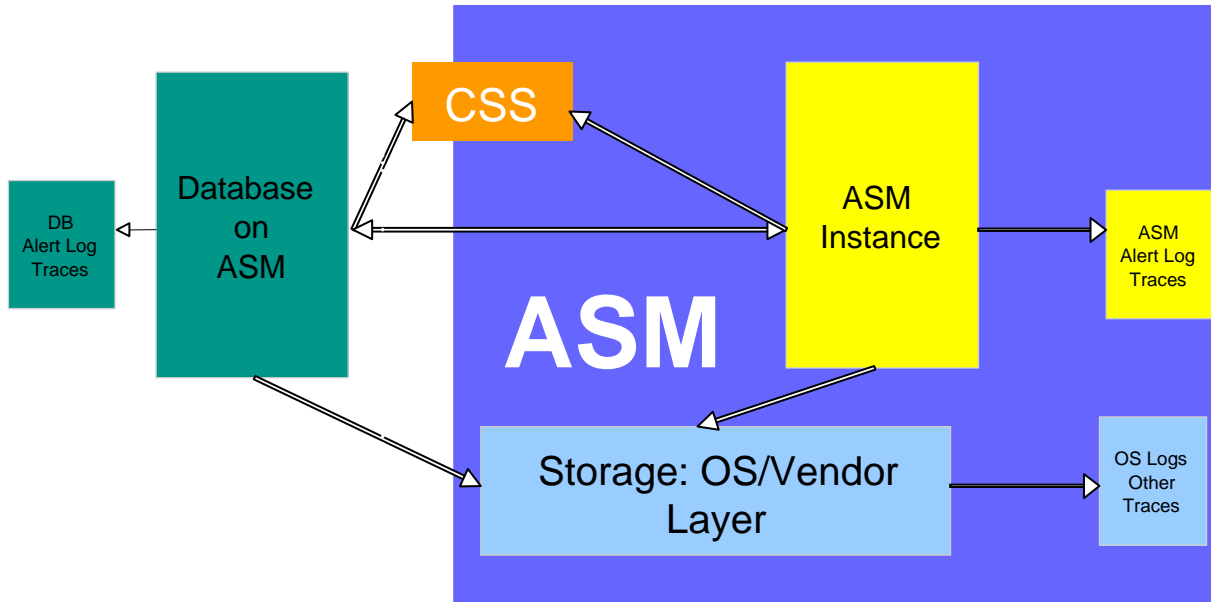
What is ASM?

- A Database File System Which provides
- Cluster file system and
- Volume manager capabilities
- That are integrated into the Oracle database 10g kernel
- At no additional cost



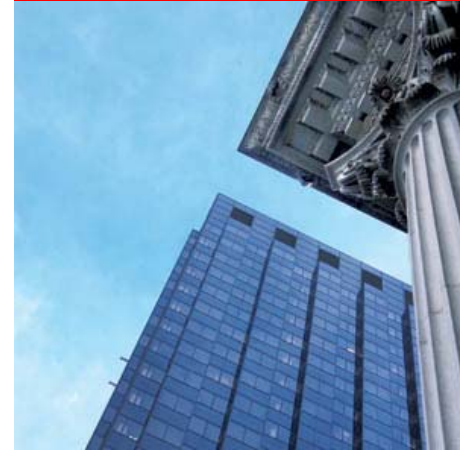
AUTOMATIC STORAGE MANAGEMENT

ASM & DATABASE RELATION

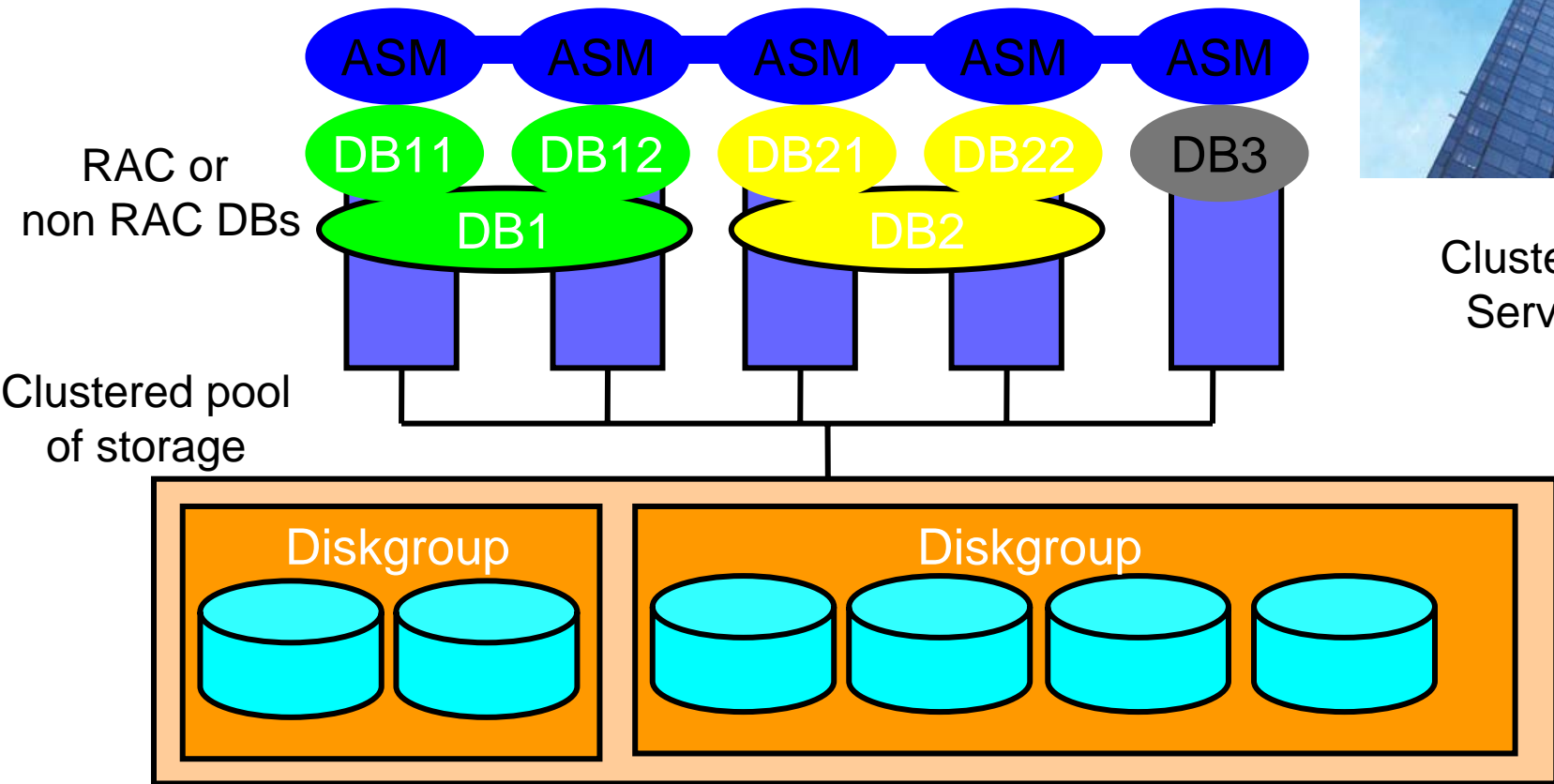


AUTOMATIC STORAGE MANAGEMENT

ASM cluster architecture

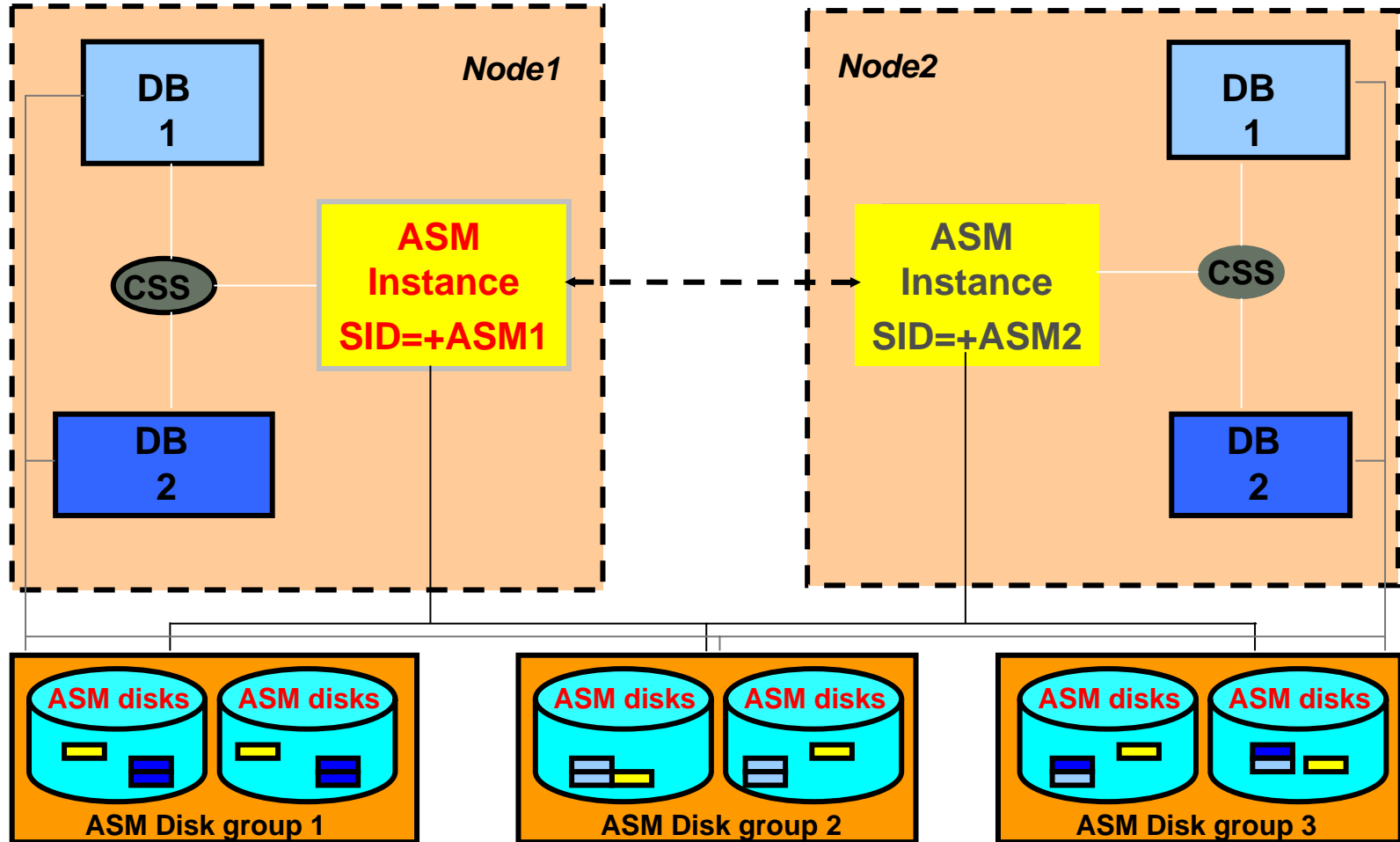


Clustered Servers



AUTOMATIC STORAGE MANAGEMENT

ASM and 2 node RAC



AUTOMATIC STORAGE MANAGEMENT

ASM IO Parameters

Write:

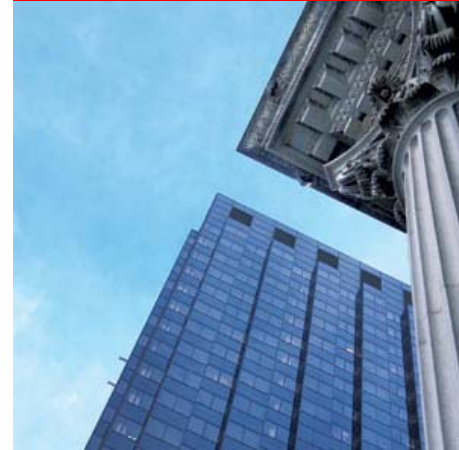
$DB_FILE_DIRECT_IO_COUNT \times \text{Oracle block size} = 1\text{MB}$

In performing tablespace creation, 1MB I/O block size can be 31% faster than 64KB

Read:

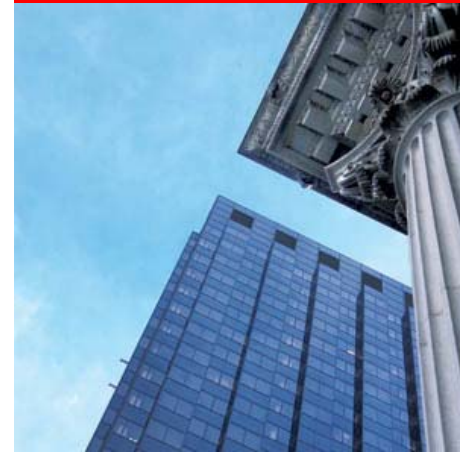
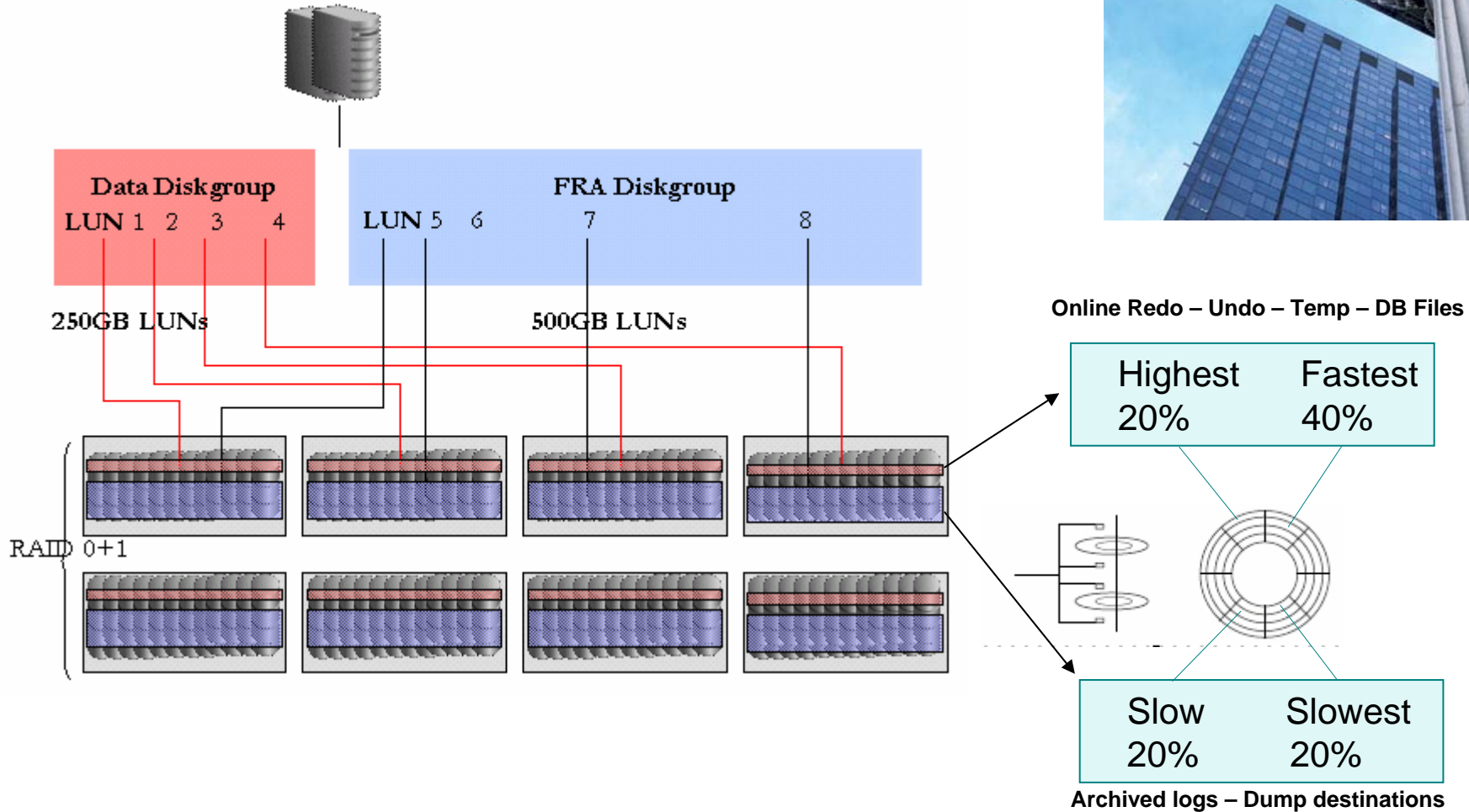
$DB_FILE_MULTIBLOCK_READ_COUNT \times \text{Oracle block size} = 1\text{MB}$

Full table scan was 62% faster with 1MB than 64KB block size



AUTOMATIC STORAGE MANAGEMENT

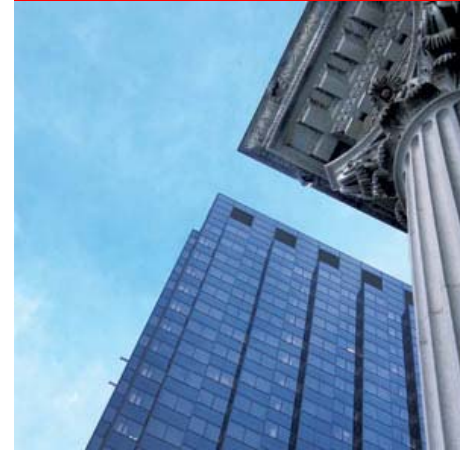
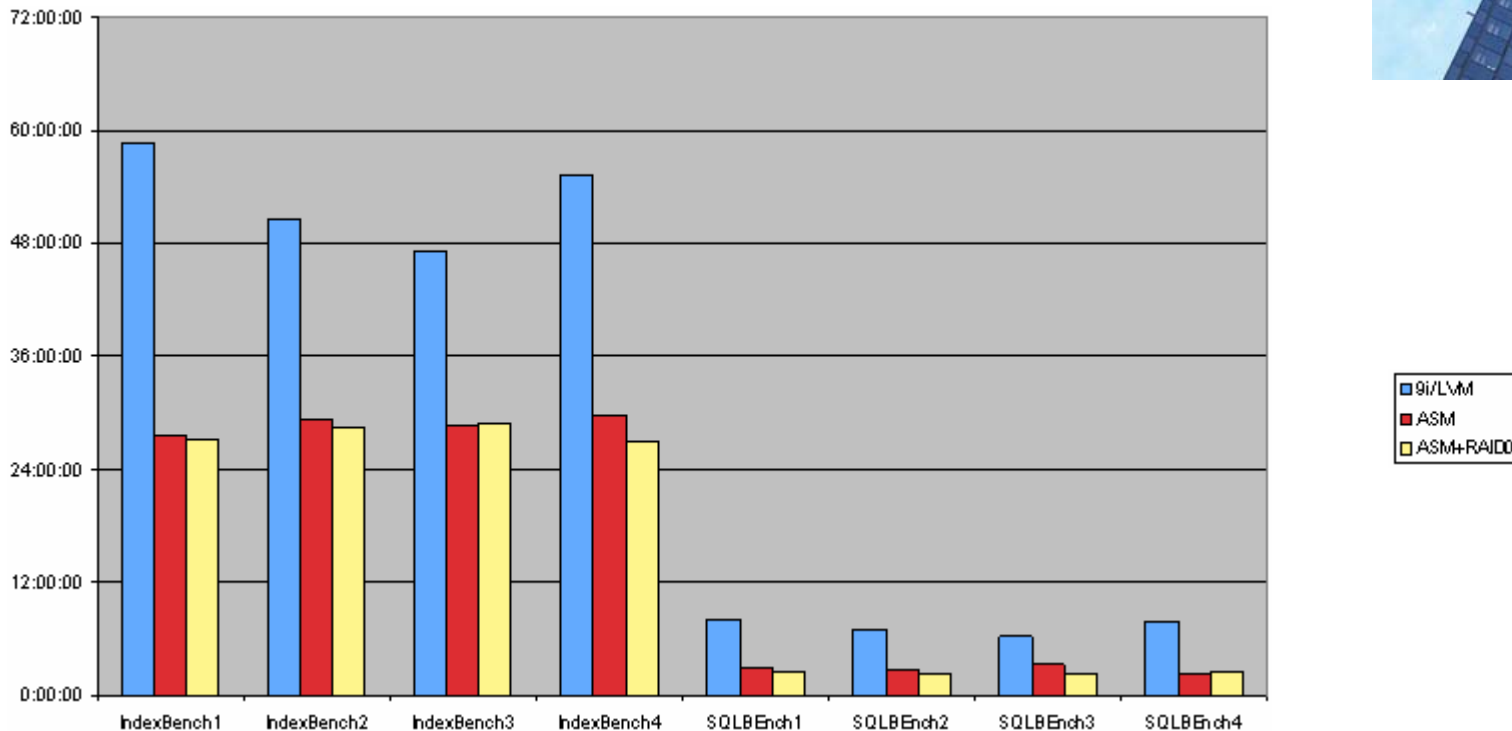
ASM Striping with RAID 0+1



AUTOMATIC STORAGE MANAGEMENT

University of Vanderbilt Customer Benchmark

100-200% improvement in performance with Oracle Database 10g and ASM compared to their 9i and legacy LVM environment.

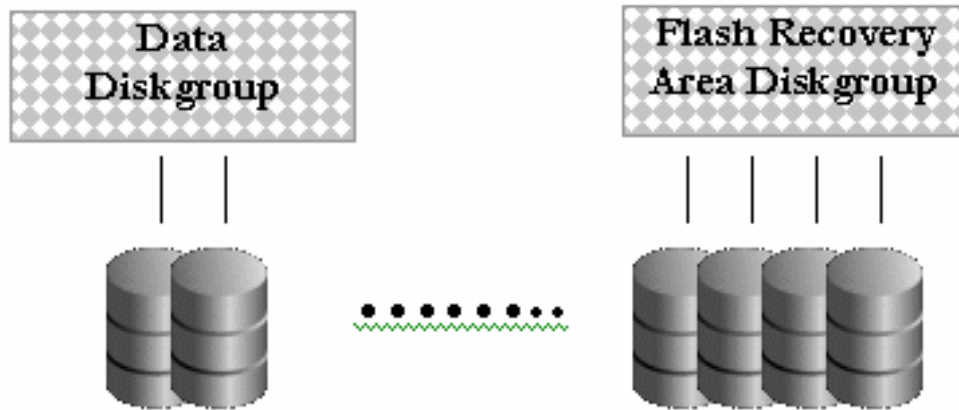


AUTOMATIC STORAGE MANAGEMENT

Recommended ASM Diskgroup Configurations

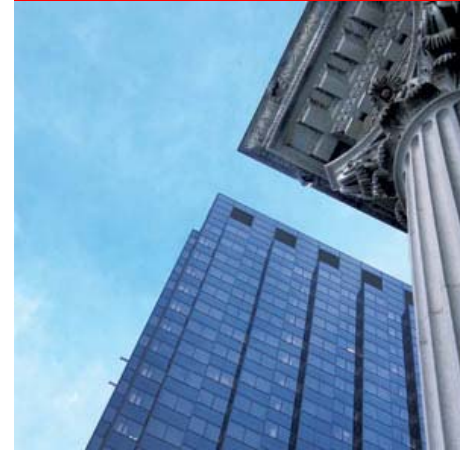
Data : data, log and control files.

Flash Recovery Area : backup, temp, archive log, dumpsets



A typical ASM diskgroup configuration

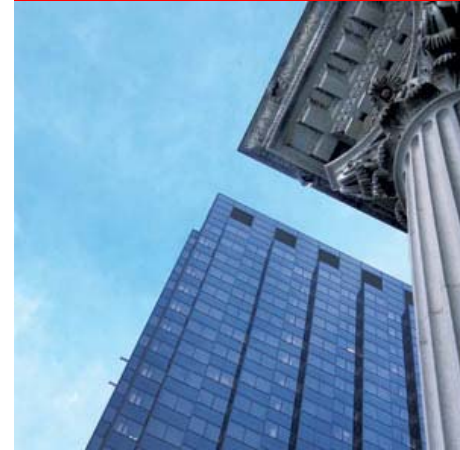
Best Practice: maximize number of spindles per diskgroup.



AUTOMATIC STORAGE MANAGEMENT

ASM BEST PRACTICES REVIEW

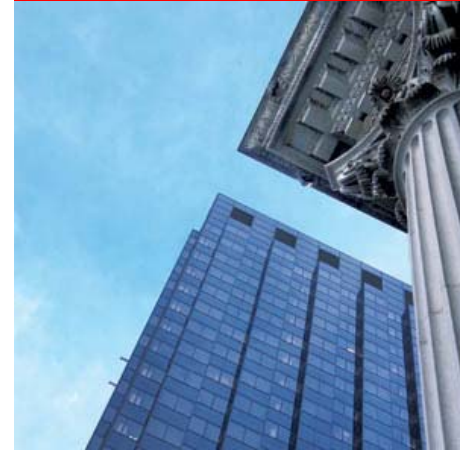
- Configure 2 diskgroups, Data and Flash Recovery Area
- ASM LUNs must have same performance and availability
- ASM LUNs must be the same capacity for each diskgroup
- Use storage array hardware RAID 1 mirroring protection when possible
- ASM mirroring (redundancy) in the absence of a hardware RAID
- Maximize the number of spindles (disks) in your diskgroup
- LUNs should use outside half of disk drives for higher performance



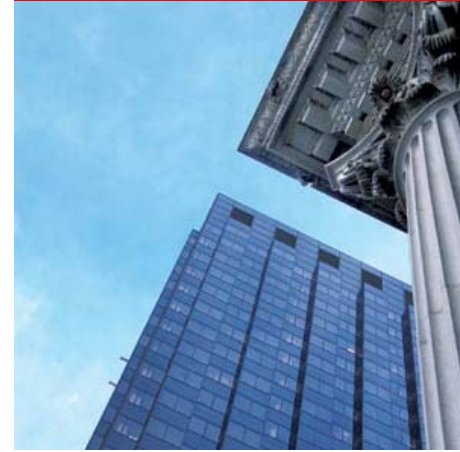
AUTOMATIC STORAGE MANAGEMENT

ASM BEST PRACTICES REVIEW

- Hardware RAID stripe size to be ideally 1MB.
OR choose the max size up to 1MB (128/256/512 etc)
- Use small disks with highest RMP
- Use large LUNs to reduce LUN management overhead
- Set ASM diskgroups on disks or arrays that are not shared with other applications
- Do not use a Logical Volume Manager (LVM)
- Use the Oracle ASMLIB feature



RAC DATABASE

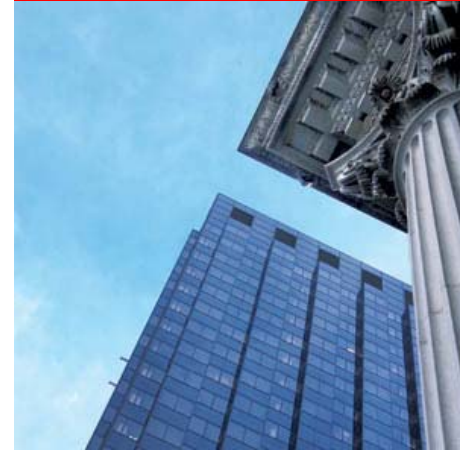


RAC DATABASE

RAC DATABASE REVIEW

RAC provides:

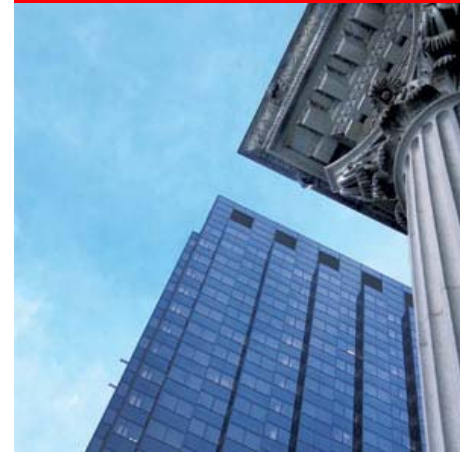
- Scalability
- High availability
- Database Consolidation
- Services for optimal management of multiple applications
- To a well tuned single instance database



VALIDATED RAC ON LINUX CONFIGURATIONS



Oracle Validated Configurations



RDBMS 10.2.0.3 Single Instance and RAC

There are many, Oracle end to end validated RAC on Linux configurations:

Oracle Enterprise Linux 4 Update 5 and RH Enterprise Linux 4 Update 5
AS/ES x86-64 using ASM + SuperMicro S5015 + Compellent Storage Center

Oracle Enterprise Linux 4 Update 4 and RH Enterprise Linux 4 Update 4
AS/ES x86-64 using ASM + Unisys ES7000/One Enterprise Server + EMC CX-600

Oracle Enterprise Linux 4 Update 4 and RH Enterprise Linux 4 Update 4 AS/ES x86-64
using ASM + Dell PowerEdge 6950 + Dell PowerVault MD3000

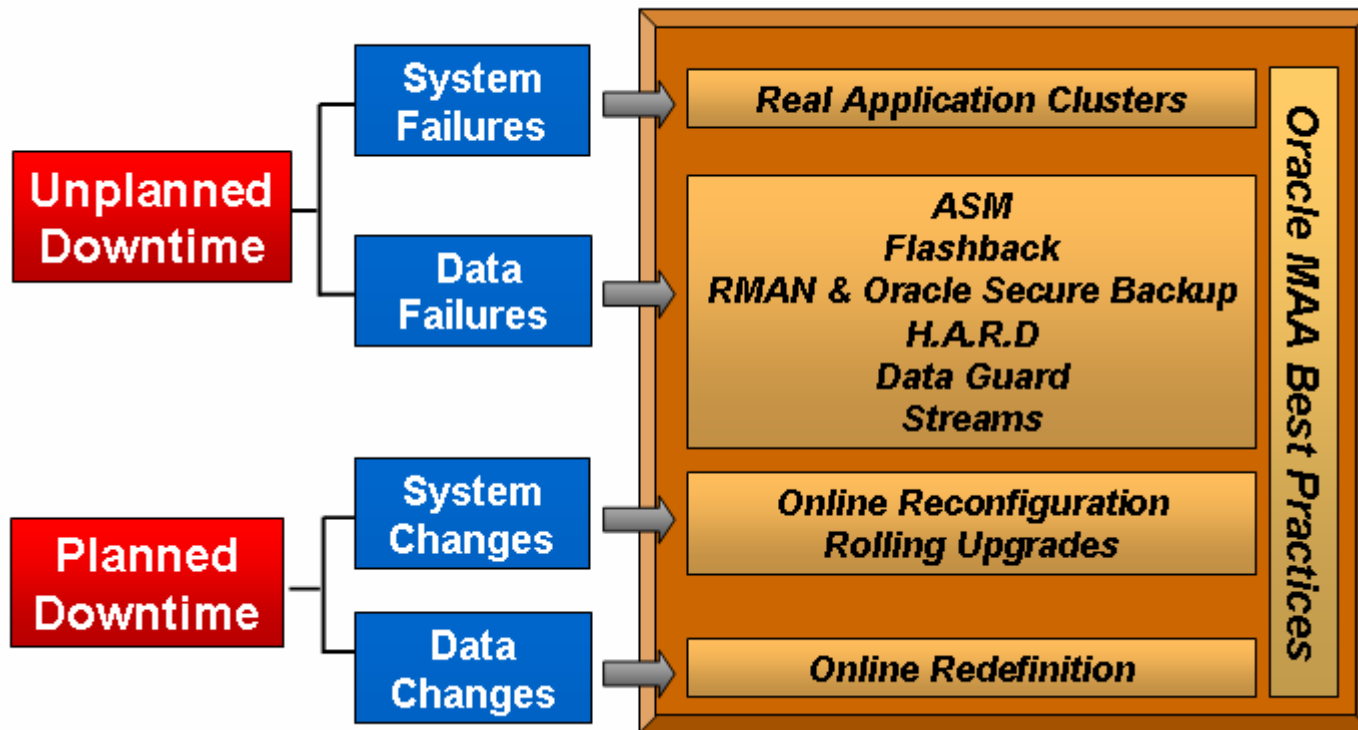
+ 23 validated configurations for 10.2.0.2 Single Instance and RAC

OTHER IMPORTANT TOPICS



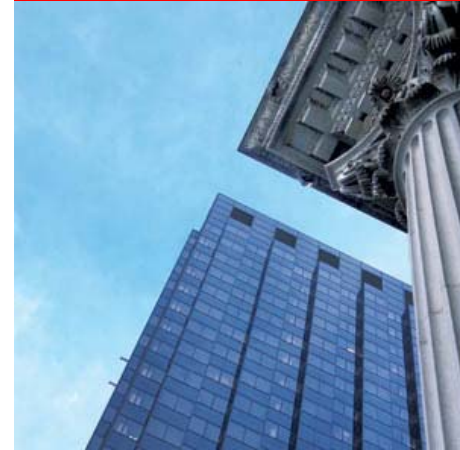
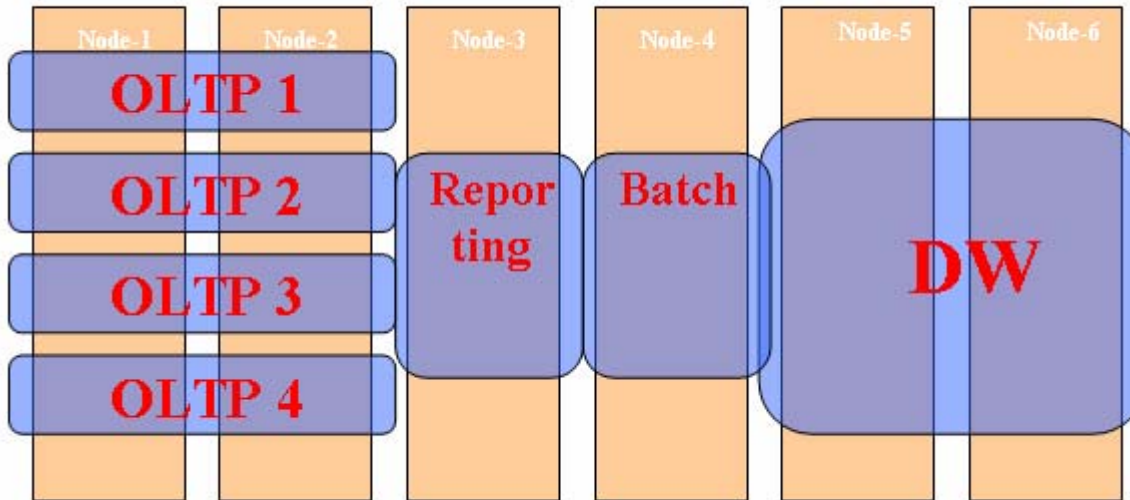
MAA with RAC

MAXIMUM AVAILABILITY ARCHITECTURE



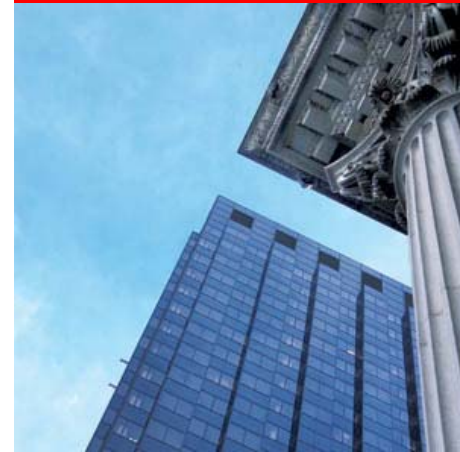
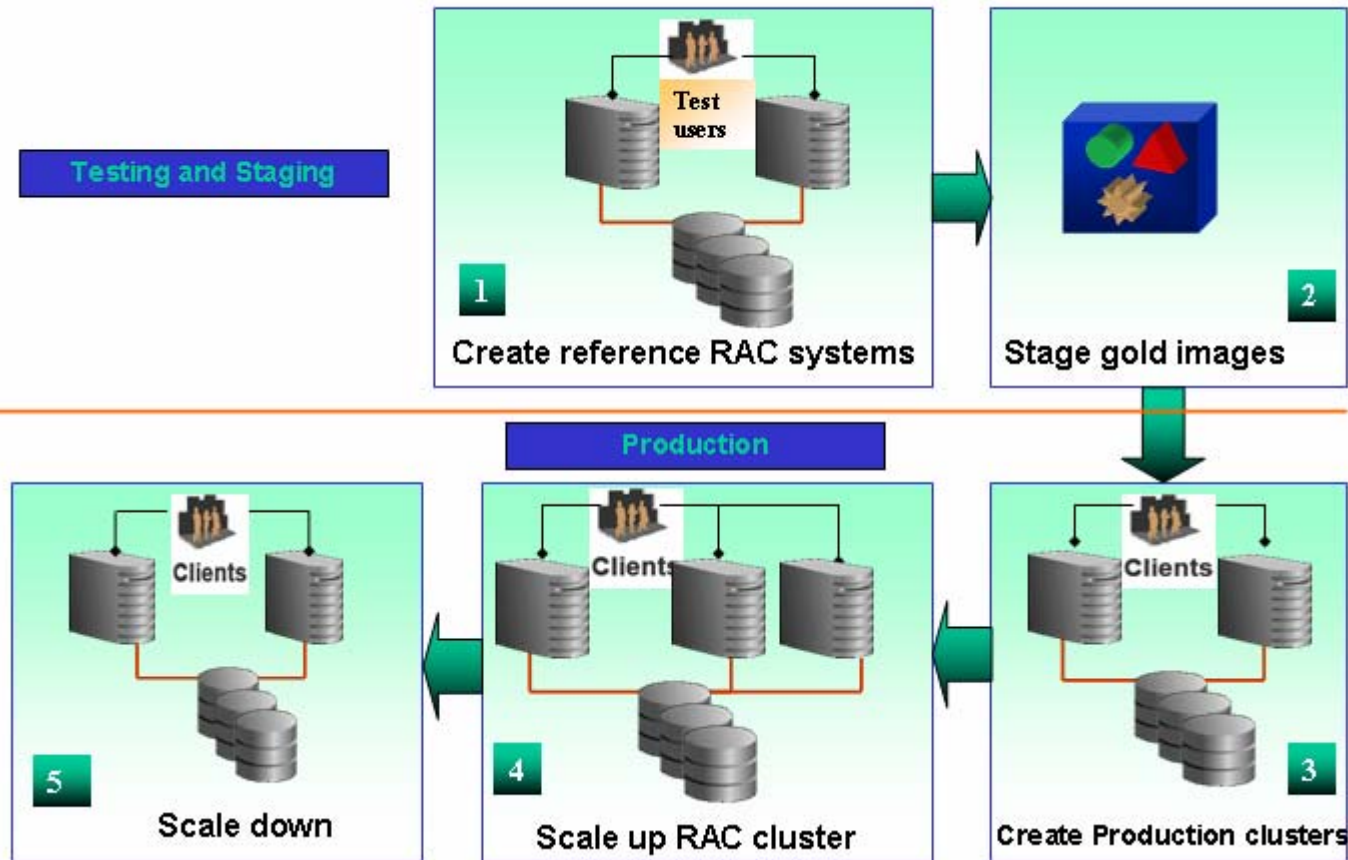
RAC Database Consolidation

DATABASE CONSOLIDATION WITH RAC



RAC Deployment Cycle

RAC DEPLOYMENT LIFE CYCLE



Thanks for Coming!

