A Survey of Cluster Technologies

Ken Moreau

Solutions Architect
Hewlett-Packard

HP WORLD 2003
Solutions and Technology Conference & Expo

But first, a word from our sponsor…

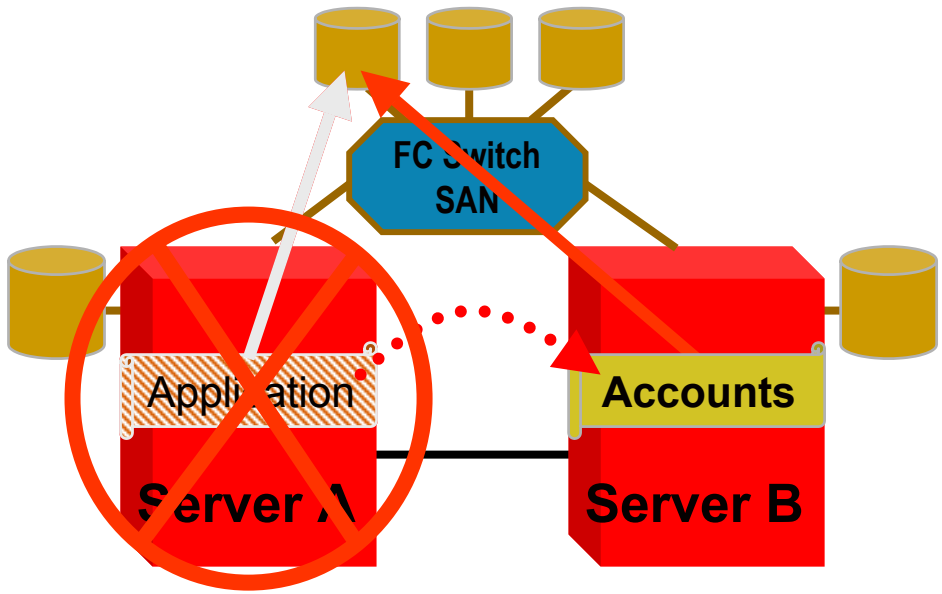
- What this talk will cover
- What this talk won't cover
- Political considerations

Topics

- Single/Multi System View
  - Shared Root
  - System Management
  - Cluster Alias
- Cluster File Systems
  - Network and Cluster File Systems
  - Distributed Lock Manager
- Configurations
  - Interconnect
  - Quorum
- Application Support
  - Special coding for clusters?
  - Failover scripting
- Resilience
  - Data Replication
  - Disaster Tolerance

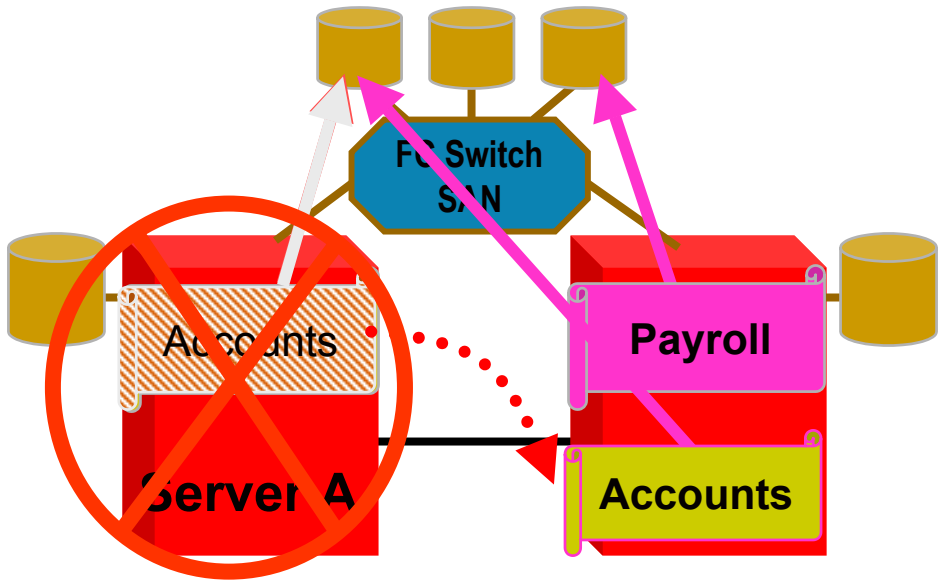
Multi System View Clusters

- Systems are relatively independent entities
- Disks are physically cabled to multiple systems, but are available to only one system at a time
- Therefore, there is no simultaneous data access from multiple systems
- Provides application failover capability only
- The systems look and act differently
- The systems are managed independently
- This is active-standby



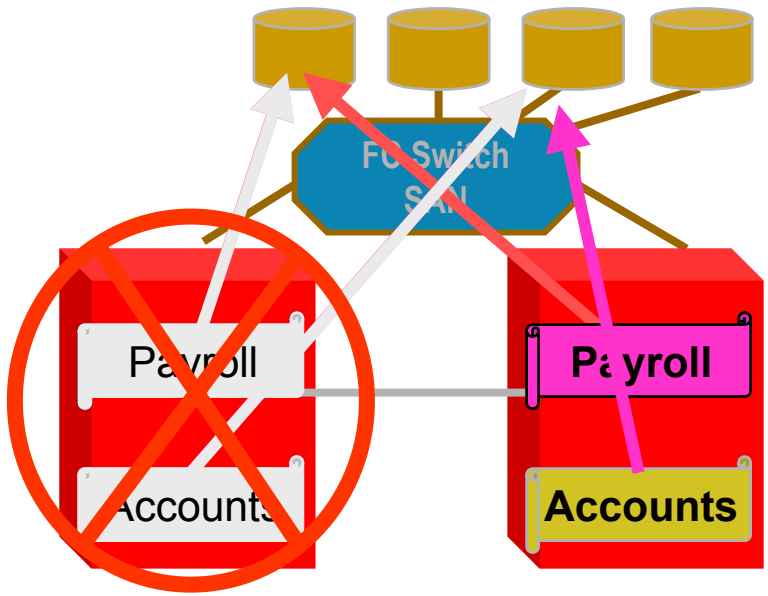
Multi System View Clusters (continued)

- Systems are relatively independent entities
- Disks are physically cabled to multiple systems, but are available to only one system at a time
- There is no simultaneous data access from multiple systems
- You can run different applications, or different instances of the same application, on different systems of the cluster
- Provides application failover capability only
- The systems look and act differently
- The systems are managed independently
- This is active-active



Single System View Clusters

- Systems cooperate very closely
- Disks are physically cabled to all systems, and are available to all systems all the time
- Therefore, simultaneous data access is easy
- Provides both application failover and simultaneous execution
- The systems look and act the same
- The systems are managed as a single entity
- This is active-active

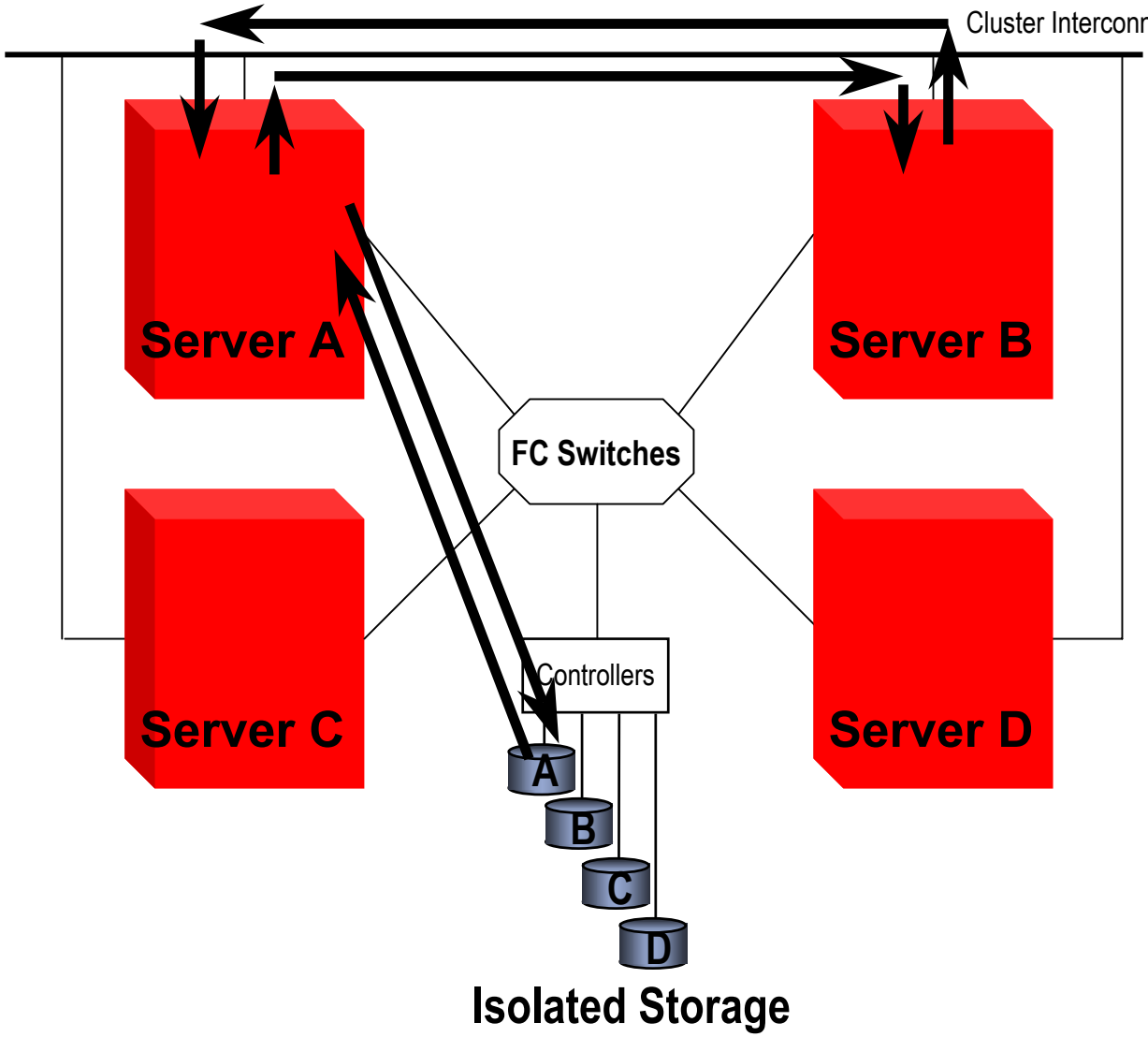


Multi or Single System View

	Multi System View	Single System View	Shared Root
LifeKeeper Linux, Windows	Yes	No[1]	No
Serviceguard	Yes	No[1]	No
NonStop Kernel	Yes	Yes	Each node (1-16 CPUs)
OpenVMS Cluster Software	Yes	Yes	Yes
TruCluster	No	Yes	Yes
Windows 2000 DataCenter	Yes	No[1]	No

[1] Oracle 9*i* RAC supplies SSI for the database and Oracle files

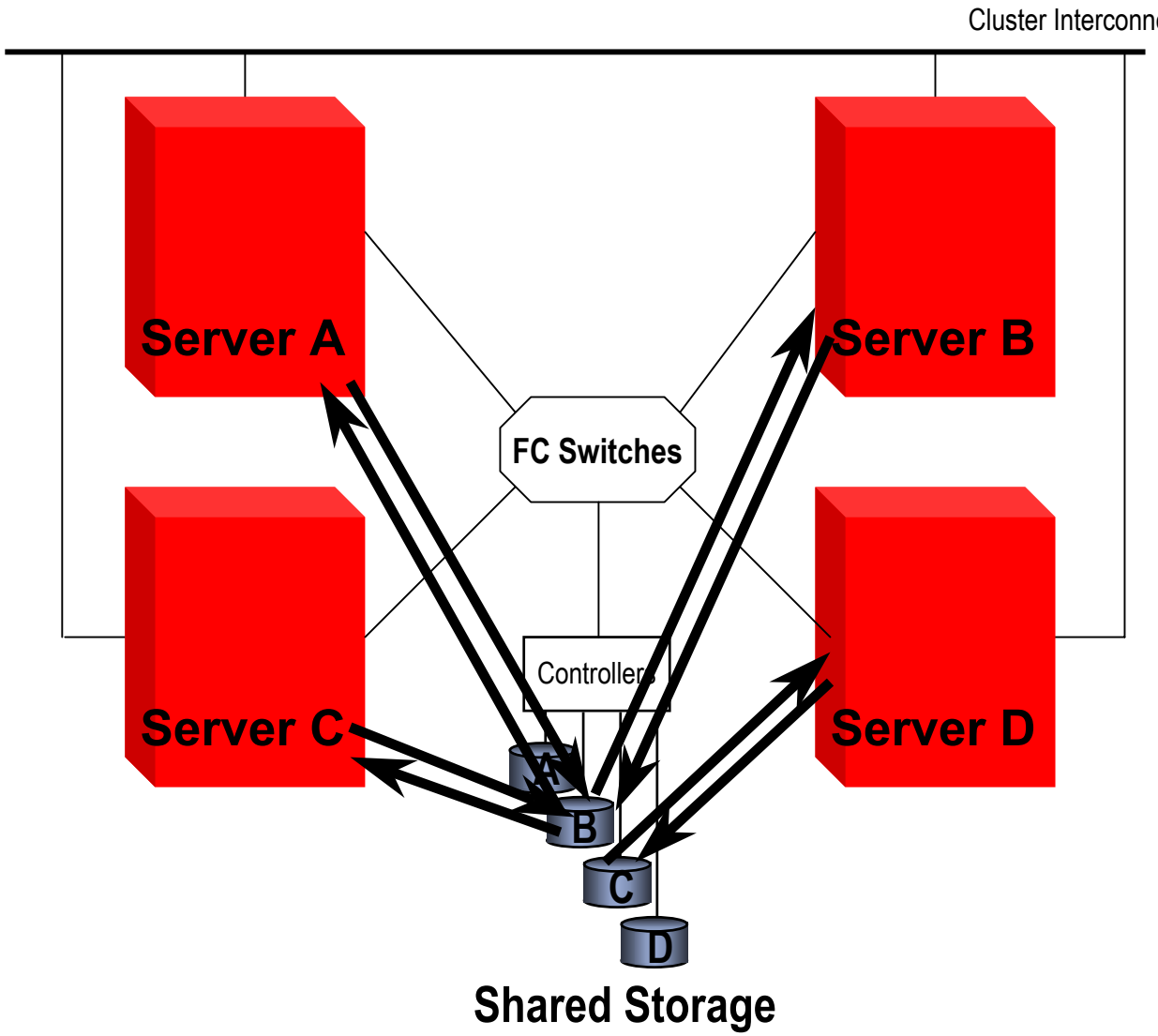
Network File Systems I/O



Client systems talk to Server systems

- Requires 3 I/Os for each disk access
- Examples include NTFS on Windows, MSCP on OpenVMS, NFS on all systems

Direct Access I/O



Direct Access I/O means all nodes in the cluster can talk directly to all disks in the cluster

- Provides full transparency and cache coherency
- Eliminates 2/3's of the I/Os in each access to a disk
- Only tokens and a few locks go on the interconnect

Cluster File Systems

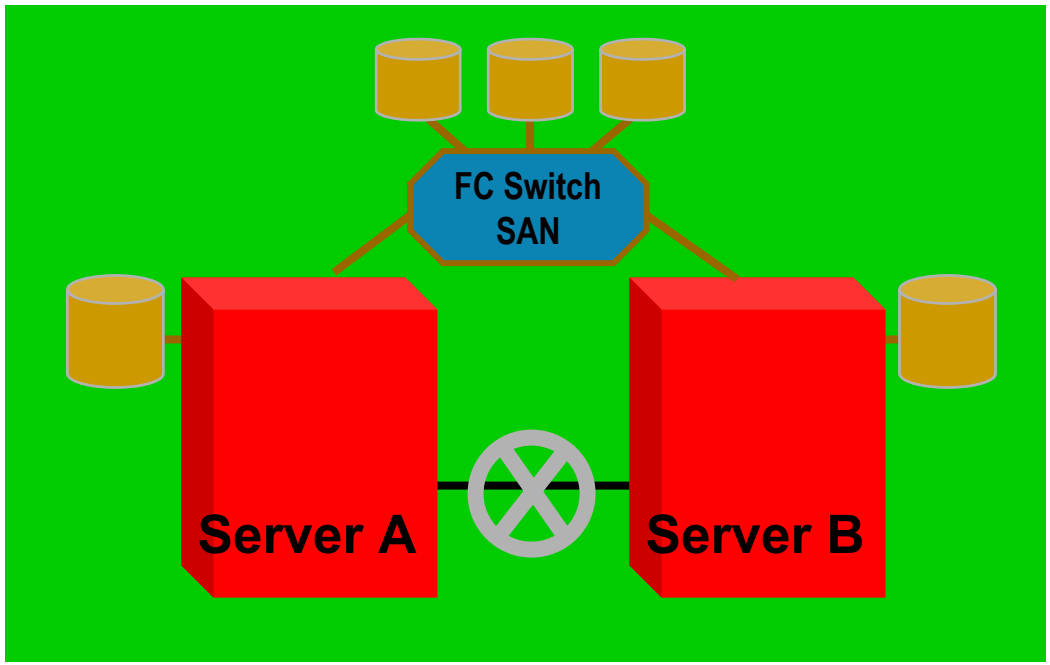
	Network File Systems I/O	Direct Access I/O	Distributed Lock Mgr
LifeKeeper Linux, Windows	NFS	Oracle raw devices, GFS	Supplied by Oracle
Serviceguard	Yes	Oracle raw devices	OPS/RAC Edition
NonStop Kernel	Data Access Manager	Effectively Yes	Not applicable
OpenVMS Cluster Software	Mass Storage Control Protocol	Files-11 on ODS-2 or -5	Yes
TruCluster	Device Request Dispatcher	Cluster File System	Yes
Windows 2000 DataCenter	NTFS	Supplied by Oracle	Supplied by Oracle

Cluster Configurations

	Max # Servers In A Cluster	Cluster Interconnect	Quorum Device
LifeKeeper Linux, Windows	16	Network, Serial	Yes (Optional)
Serviceguard	16	Network, HyperFabric	Yes = 2, optional >2
NonStop Kernel	255	ServerNet, TorusNet	No
OpenVMS Cluster Software	96	CI, Network, MC, Shared Mem	Yes (Optional)
TruCluster	8 generally, 32 w/Alpha SC	100Enet, QSW, Memory Channel	Yes (Optional)
Windows 2000 DataCenter	4	Network	Yes

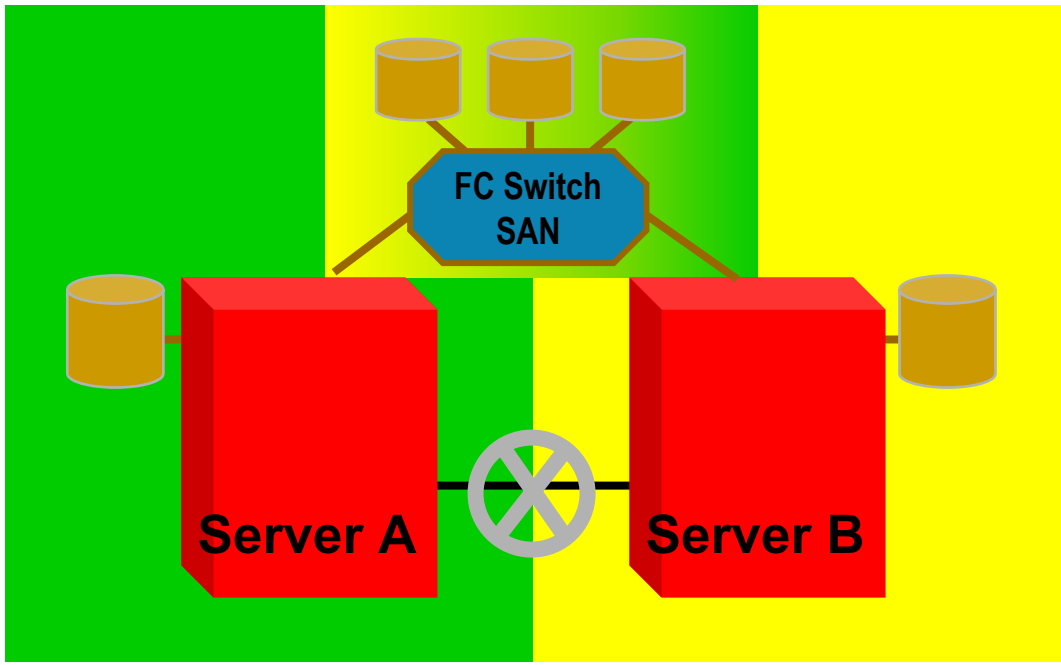
A 2-node cluster without a quorum disk

- All disks mounted cluster wide
  - Required quorum = (expected_votes + 2) / 2 = (2+2)/2 = 2
  - Actual quorum = (*actual_votes* + 2) / 2 = (2+2)/2 = 2



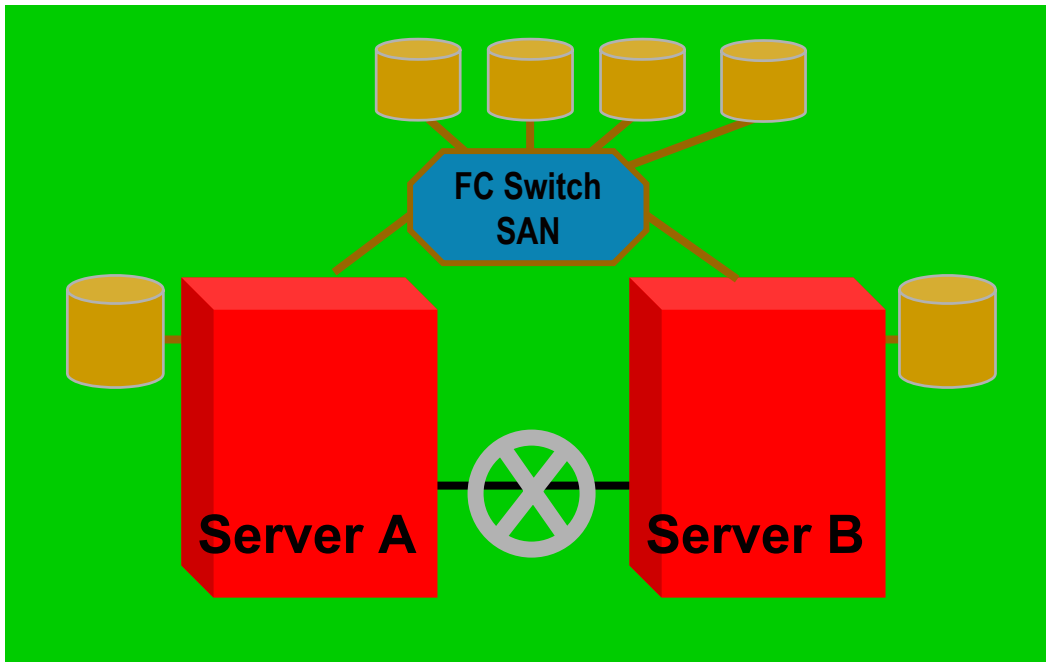
A 2-node cluster without a quorum disk

- Servers A and B each try to form a cluster
  - Actual quorum = (*actual_votes* + 2) / 2 = (1+2)/2 = 1
  - Less than required quorum, so no cluster is formed
- What would happen if this scheme wasn't in place?



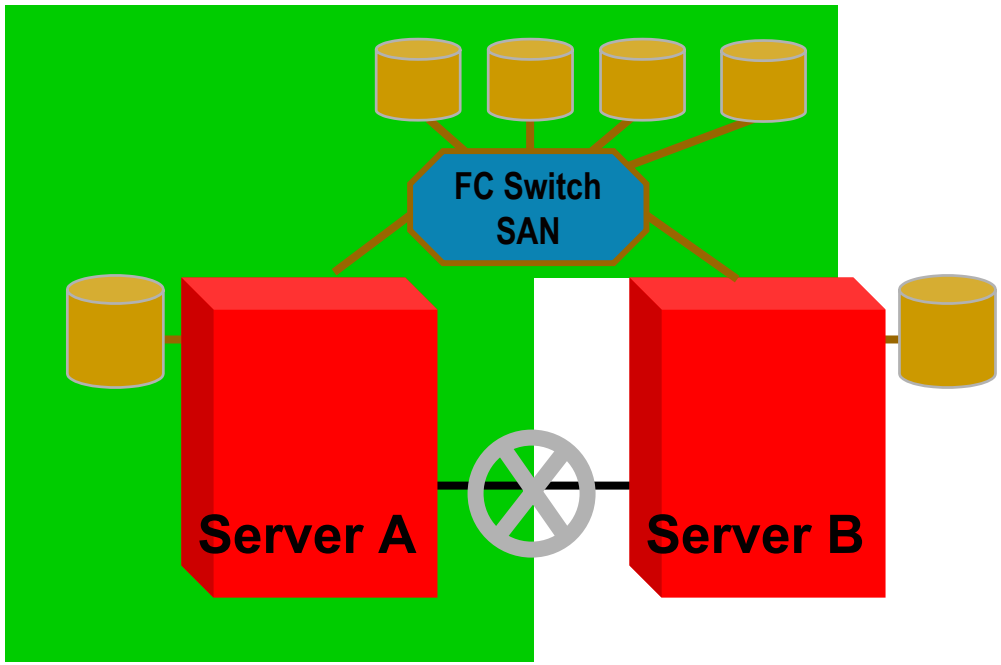
A 2-node cluster with a quorum disk

- All disks mounted cluster wide
  - Required quorum = (expected_votes + 2)/2 = (3+2)/2 = 2
  - Actual quorum = (*actual_votes* + 2)/2 = (3+2)/2 = 2



A 2-node cluster with a quorum disk

- Server A forms a cluster
  - Actual quorum = (*actual_votes* + 2)/2 = (2 + 2)/2 = 2
- Server B does not form a cluster
  - Actual quorum = (*actual_votes* + 2)/2 = (1 + 2)/2 = 1



Application Support

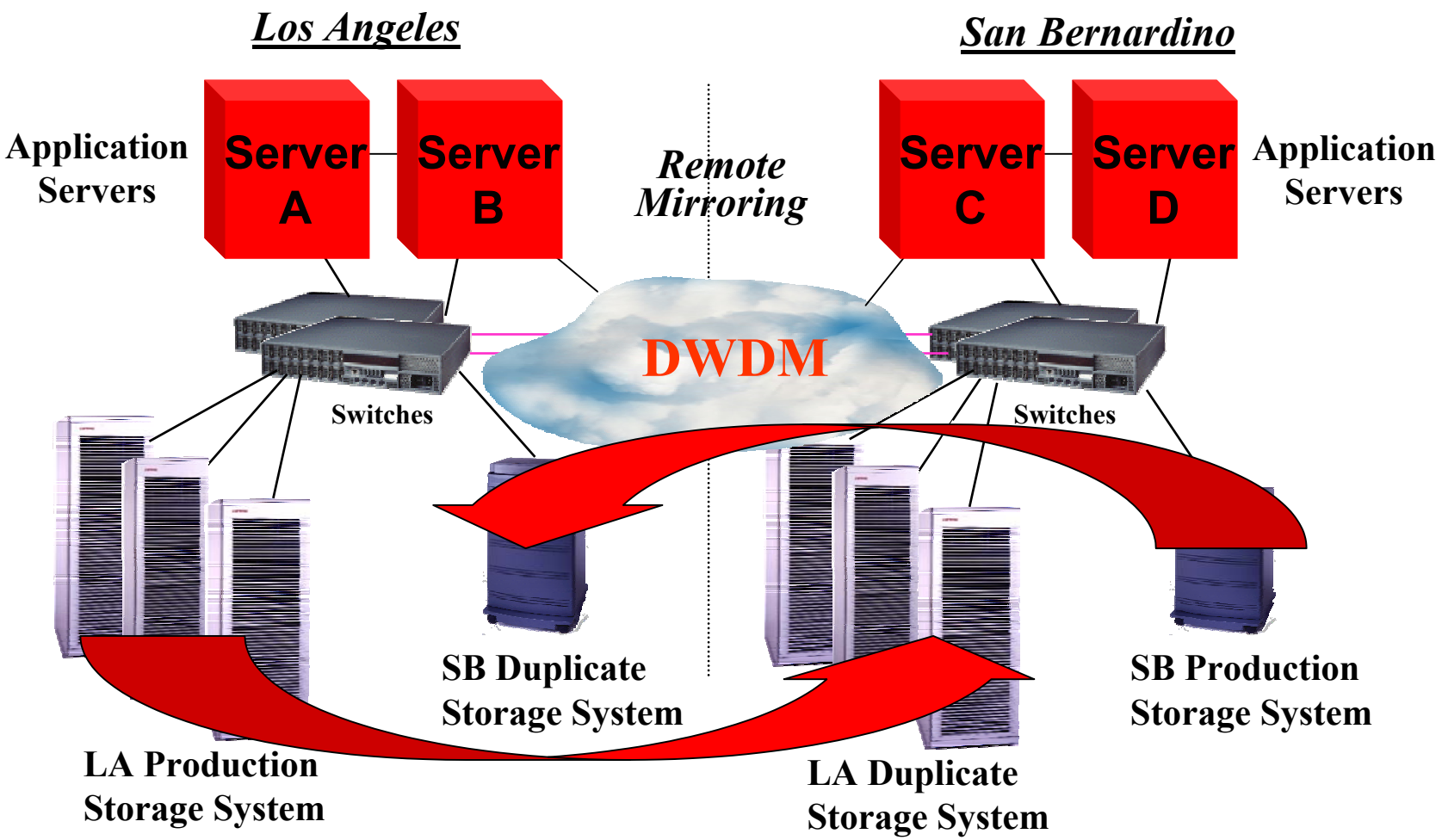
	Single-instance (failover mode)	Multi-instance (cluster-wide)	Recovery Methods
LifeKeeper Linux, Windows	Yes	No[1]	Scripts
Serviceguard	Yes	No[1]	Packages and Scripts
NonStop Kernel	Yes Takeover	Effectively Yes	Paired Processing
OpenVMS Cluster Software	Yes	Yes	Batch /RESTART
TruCluster	Yes	Yes	Cluster Application Availability
Windows 2000 DataCenter	Yes	No[1]	Registration, cluster API

[1] Oracle 9i RAC supplies multi-instance for the database only

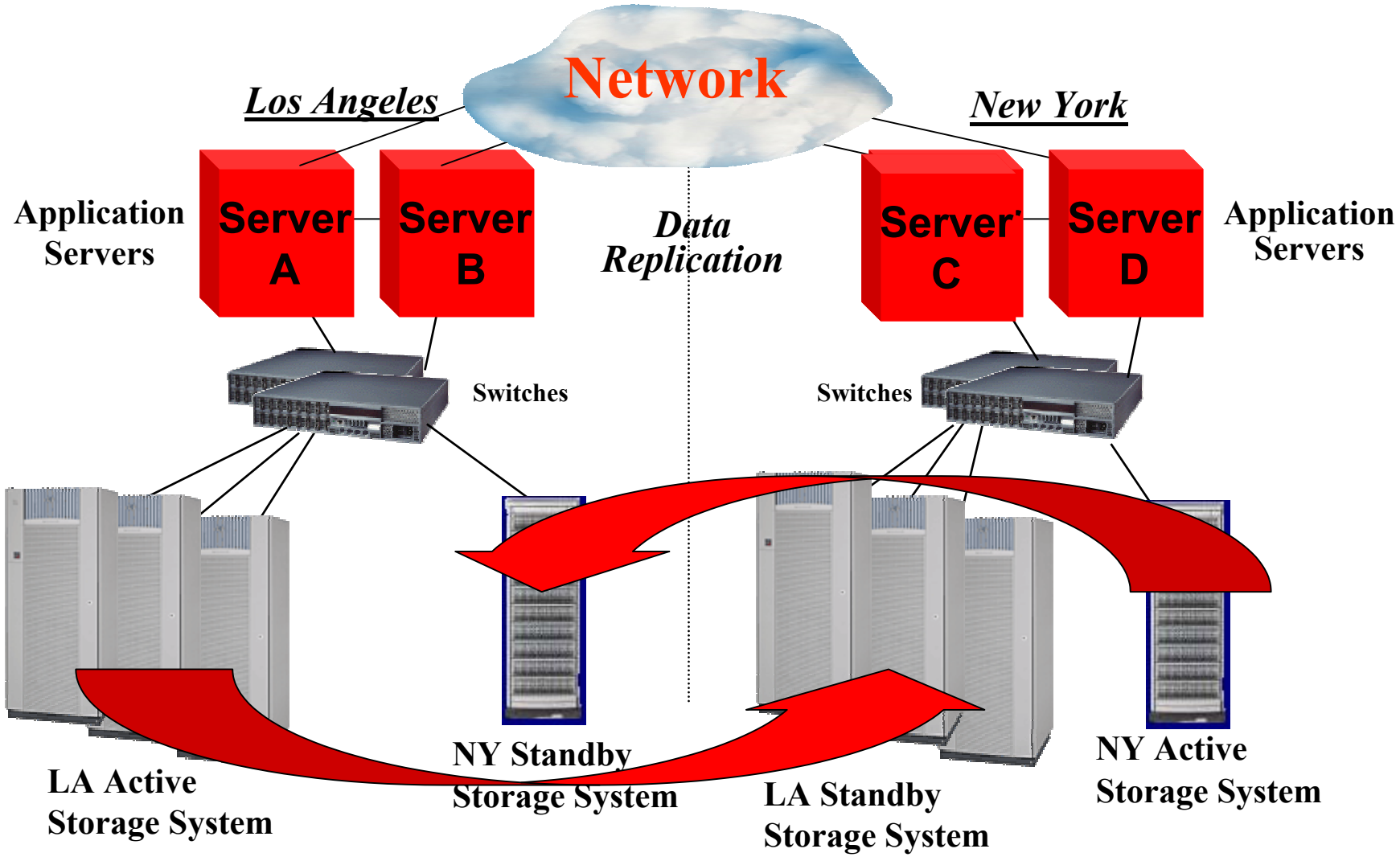
Resilience

	Data High Availability	Dynamic Partitions	Disaster Tolerance
LifeKeeper Linux, Windows	Distributed Replicated Block Device (DRBD)	No	Extended Mirroring
Serviceguard	MirrorDisk/UX, Multi-Path I/O (a)	vPars	Extended Clusters
NonStop Kernel	RAID-1, Multi-Path I/O (p), Process Pairs	No	Remote Database Facility
OpenVMS Cluster Software	HBVS RAID-1, Multi-Path I/O (p)	Galaxy	DTCS, StorageWorks CA
TruCluster	LSM RAID-1, Multi-Path I/O (a)	No	StorageWorks CA
Windows 2000 DataCenter	NTFS RAID-1, SecurePath(p)	No	StorageWorks CA

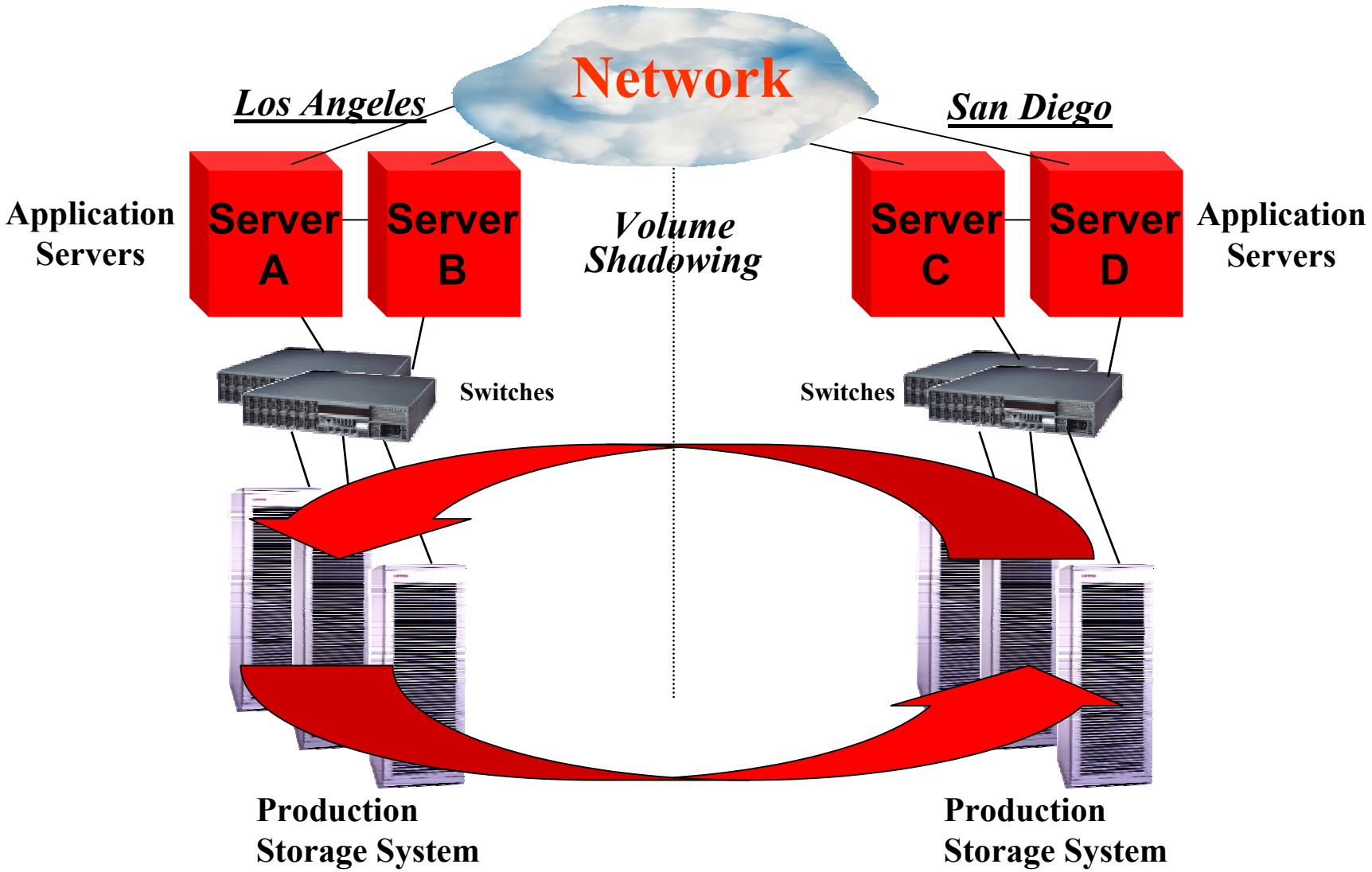
Physical Replication - data synchronization



Logical Replication



Disaster Tolerant OpenVMS Clusters



Summary

- Every system offers a high availability option
  - But the recovery times vary from many minutes to transparent
- Every system can scale outside the box
  - But the sizes vary from 2-node to 255-node clusters
- Every system has the option of disaster tolerance
  - But the technologies vary from one-way data replication between separate clusters, to full active/active cooperation of a single cluster spread over several geographically dispersed datacenters
- Understand the options and choose the right technologies
- Understand what you get and don't get with each technology

Resources

- Linux LifeKeeper
  - http://h18000.www1.hp.com/solutions/enterprise/highavailability/linux/index.html
- Serviceguard
  - http://docs.hp.com/hpux/ha/index.html
- NSK
  - http://h71033.www7.hp.com/page/TIM_Prod.html
- TruCluster
  - http://h30097.www3.hp.com/docs/pub_page/cluster_list.html
- OpenVMS Cluster Software
  - http://h71000.www7.hp.com/openvms/products/clusters/index.html
- Windows 2000
  - http://www.microsoft.com/windows2000/en/datacenter/help
- Books
  - “Clusters for High Availability”, Peter Weygant, ISBN 0-13-089355-2
  - “In Search of Clusters”, Gregory F. Pfister, ISBN 0-13-899709-8

Interex, Encompass and HP bring you a powerful new HP World.