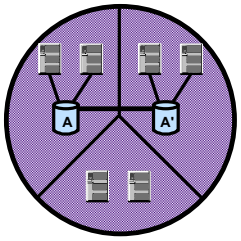
Disaster-Tolerant, Highly Available Cluster Architectures

Bob Sauers

Hewlett Packard Company
HPWorld '99 Tutorial 027
August 1999



Welcome

Bob Sauers
High Availability Solutions Architect

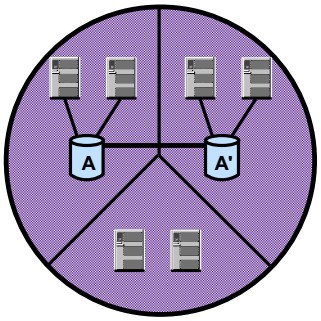
Business Critical Computing Business Unit (BCCBU)
Availability & Consolidation Solutions Lab (ACSL)
HA Advanced Technology Center

Good URL for technical whitepapers on HA & DR:
http://docs.hp.com

Agenda

- Disaster Tolerance Concepts
- Disaster-Tolerant Cluster Architectures
  - Campus Clusters with Physical Data Replication
  - MetroCluster with Physical Data Replication
  - ContinentalClusters with Physical or Logical Data Replication
- Network Examples
  - Campus Clusters and MetroCluster
  - ContinentalClusters
- Questions

Disaster Tolerance Concepts



High Availability

- Redundancy involves multiple pieces of hardware that will take over immediately or within a short time, in case of component failure

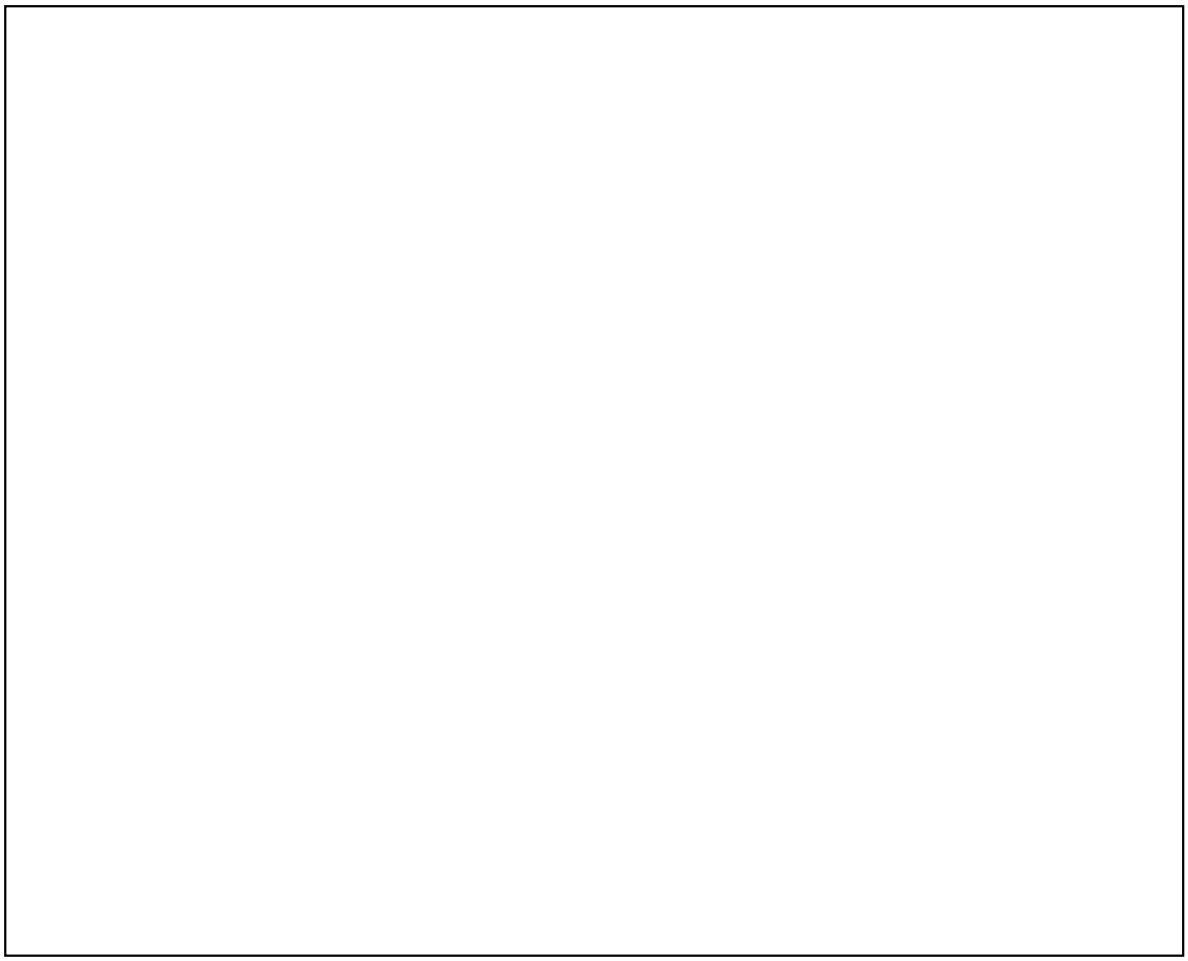
- Increased investment in hardware & software
- Sometimes results in no disruption of the service
  - **networks**
  - **protected disks (RAID)**
- Other times, a short outage occurs while switching to the redundant hardware
  - **systems**
  - **data centers**

Weighing the Cost of High Availability

- Cost of additional hardware versus cost of downtime to the organization
  - loss of revenue
  - loss of productivity
  - loss of customers
  - loss of reputation
- Example
  - Belt: $20 to $30
  - Braces: $30 to $40
  - Cost of downtime: Embarrassment and absence

Ensuring Availability

- Combination of:
  - People
  - Processes
  - Services
  - Technology

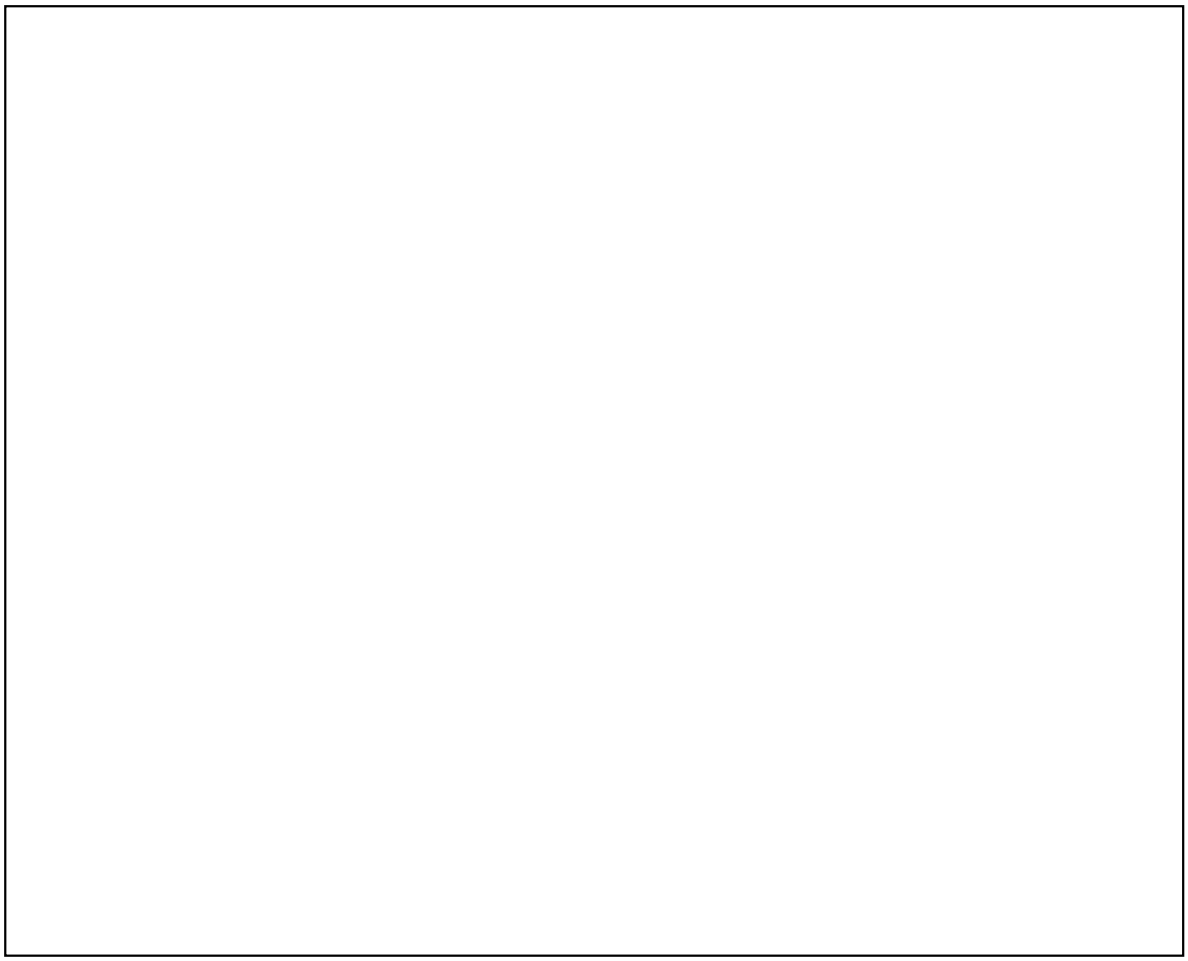


Traditional Disaster Recovery

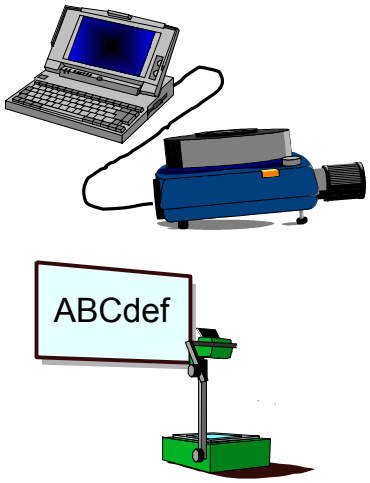
- Services
  - Risk Assessment
  - DR Planning
  - DR Event Practice
  - Shared Systems, PCs & External Networks at a remote disaster recovery site
  - Trailers with Dedicated Systems, PCs, PBXs brought to your location
  - Some vendors now provide capability for remote data replication

Disaster-Tolerant Architectures

- Hardware owned by the Organization
  - Geographically dispersed data centers
  - Additional systems, disk storage & networks
  - Software to detect failures & take action
  - Data Replication Methods
- Services
  - Risk Assessment
  - DR Planning
  - DR Event Practice



Disaster Tolerance



- Redundancy involves multiple pieces of hardware in another location that will take over within a reasonably short time in case of certain disasters
- An outage occurs while switching to the redundant hardware or location

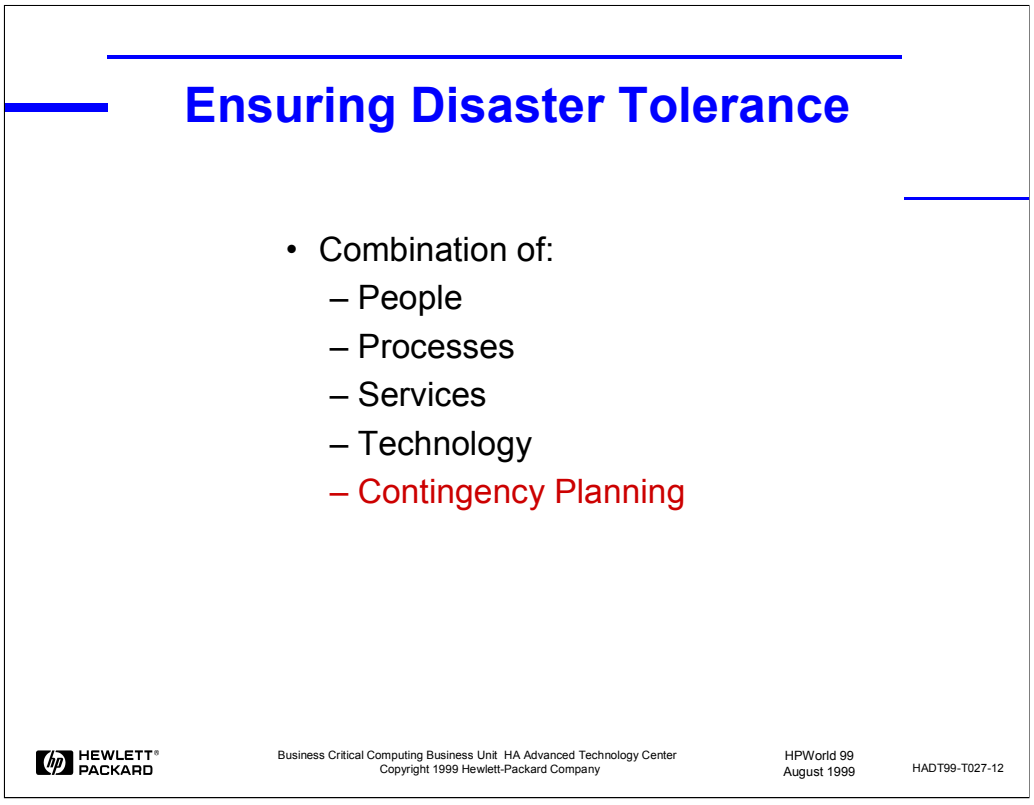
Disasters include:

- natural disasters such as
  - earthquakes
  - floods
  - hurricanes and tornadoes
- other disasters such as
  - fires
  - sabotage
  - explosions
  - large-scale power outages of long duration
  - human error
  - terrorism

Weighing the Cost of Disaster Tolerance

- Cost of additional hardware, staff and facilities versus cost of downtime to the organization
  - loss of revenue
  - loss of productivity
  - loss of customers
  - loss of reputation
  - being driven out of business

- Example
  - Transparencies: $150 in materials, few hundred in labor, extra weight in carry-on
  - Cost of downtime: Embarrassment, reputation suffers, dissatisfied conference attendees, not invited back

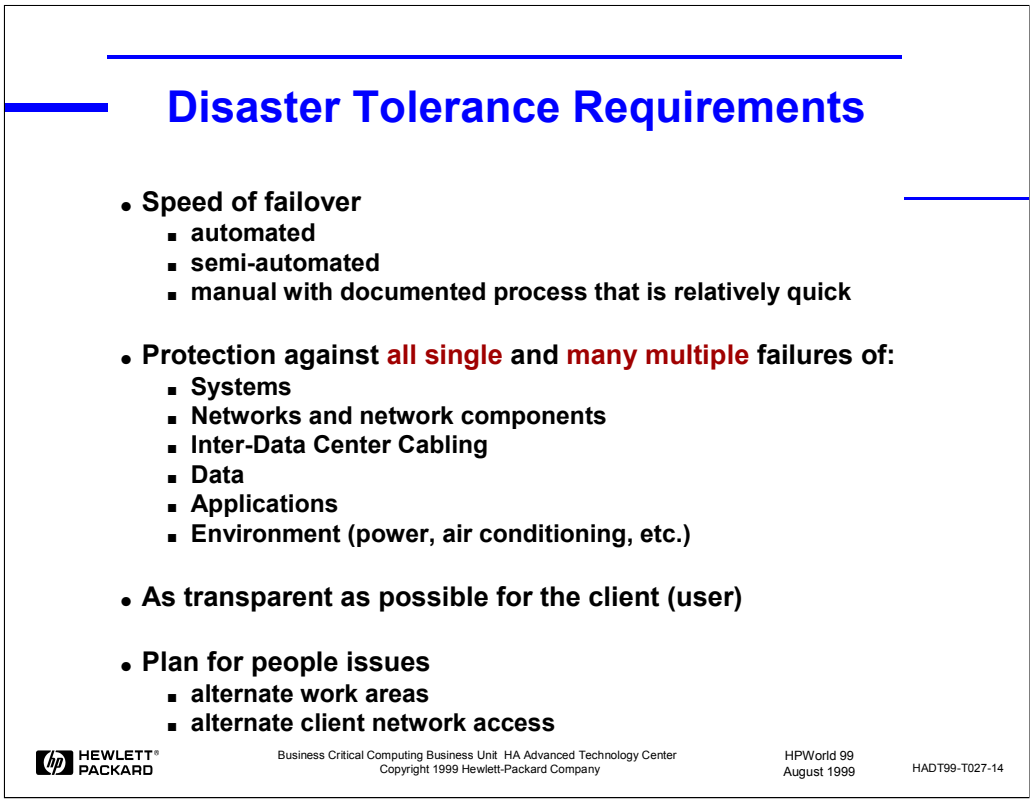


Contingency planning includes such things as:

- evaluation of risk of experiencing certain disasters
- evaluation of probability of occurrence of certain disasters
- evaluation of the cost of downtime
- test and practice planning

Disaster-Tolerant Architectures

- **Cluster architecture that provides the ability to automatically or semi-automatically recover *quickly* *after certain disasters***
- **Based on *redundant***
  - **components (LANs, network devices, Disk Host Adapters)**
  - **data protection (mirroring, replication, etc.)**
  - **systems (cluster techniques)**
- **Assumes some level of geographic dispersed hardware**
  - **multiple data centers**
  - **multiple buildings in a campus**
  - **multiple buildings in a city**
  - **multiple cities**



Disaster-tolerance protects against all Single Points of Failure (SPOFs) and many Multiple Points of Failure (MPOFs). It may NOT protect against rolling disasters. Examples of protecting against SPOFs:

- backup node takes over when a primary node fails
- backup network is used when a primary network fails
- mirror copy of data is used when the primary copy fails

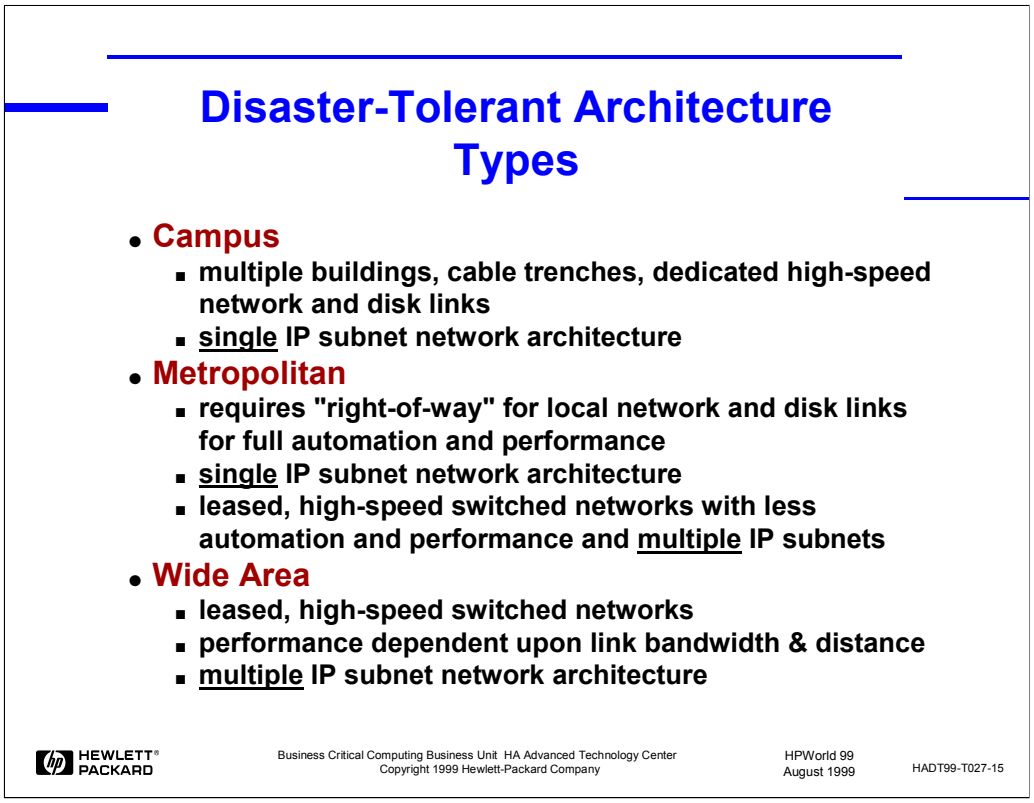
Examples of protecting against MPOFs:

- backup site takes over when entire primary site fails

Examples of MPOFs that may not be disaster-tolerant:

- the failure of ALL networks among ALL data centers
- the loss of power in more than one data center
- the loss of all copies of the on-line data

Providing a disaster-tolerant hardware environment without planning for the "people" aspects is a waste of money. Alternate work areas and client networks must be part of the architecture.



Campus architectures can use Ethernet, FDDI or Fibre Channel networks and are limited more by the maximum disk link than the network

Fibre Channel disk links are currently limited to 10 km using long wave hubs

Metropolitan architectures can use FDDI networks and hard-wired SRDF or Continuous Access XP disk links.This combination is currently limited to 100 km

Metropolitan architectures can also use leased, switched networks. However, these networks pose technical problems with single cluster architectures. Today, these networks can only be used with multiple cluster architectures.

Wide Area architectures use leased, switched networks. This architecture is currently not supported by MC/ServiceGuard due to the requirement that heartbeat networks use a single IP subnet.

The major issue with this architecture is the bandwidth of the link and the data replication requirements.

Disaster-Tolerant Architecture Rules

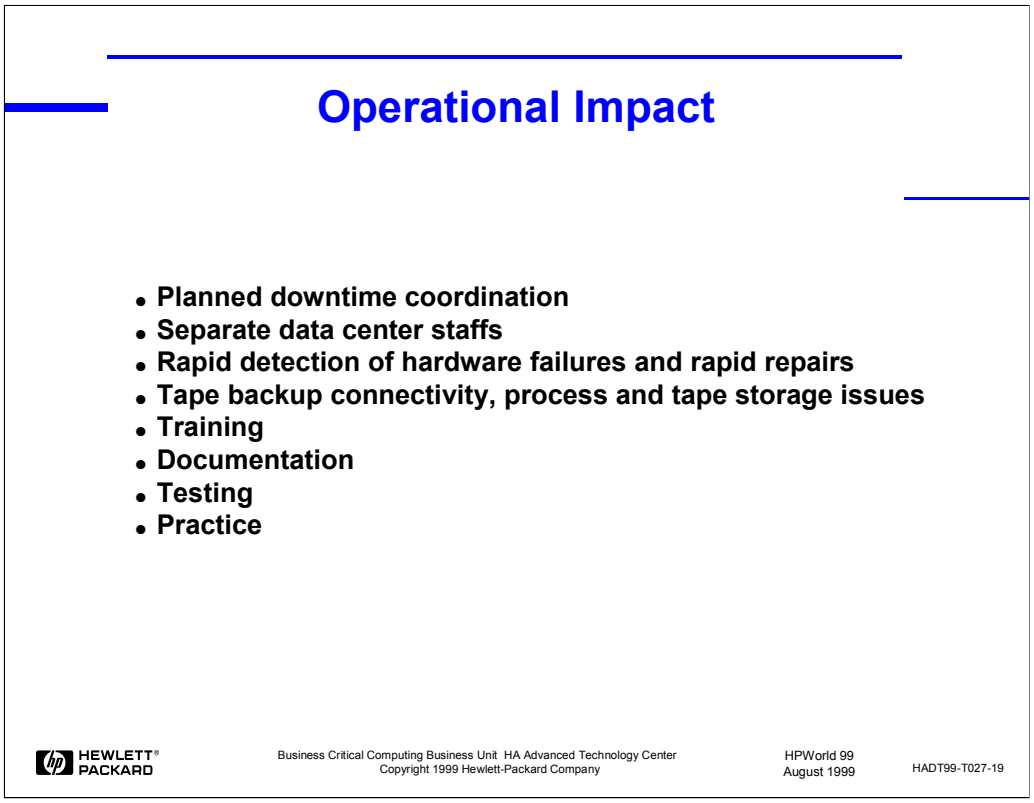
- **Local Site High Availability**
  - no Single Points of Failure (SPOFs)
  - failover among systems at local site is preferred
- **Remote Site High Availability**
  - may be lower level of HA than at Local Site
  - failover among systems at remote site is possible
  - may be running mission-critical applications, also
- **Disaster Tolerance additional requirements**
  - data replication between sites
  - reliable network links between sites
  - redundant links routed differently to prevent the "backhoe" problem

Disaster-Tolerant Design Guidelines

- **Local failover is better for most failures**
  - Faster in some cases
  - May involve less recovery
  - Fewer chances of problems
  - May be more transparent to the clients
- **Disaster Recovery failure should be used only in case of entire site failures**
  - Data may be unprotected while failed over to DR site

Tradeoffs for Disaster Tolerance

- **Ability to recover from certain disasters results in**
  - longer failover time even for local failovers
  - having to manage a cluster across multiple data centers
  - having to design more fault-resilience into networks
  - potential for slower performance due to data replication
- **Greater availability for mission critical applications**
  - greater cost
- **Preserving the ability to recover after a rolling disaster**
  - manual intervention is required for certain failures



There are impacts to the operation of the systems for a cluster, such as having to use cluster commands to bring the applications up and down.

Additional impacts occur in the campus cluster environment:

- taking systems off-line for planned maintenance must be more carefully thought out since it may leave the cluster vulnerable to another failure

- data centers often have separate operations staff -- they will now be required to communicate with each other

- hardware failures must be detected rapidly and repairs made quickly so that the cluster is not vulnerable to additional failures

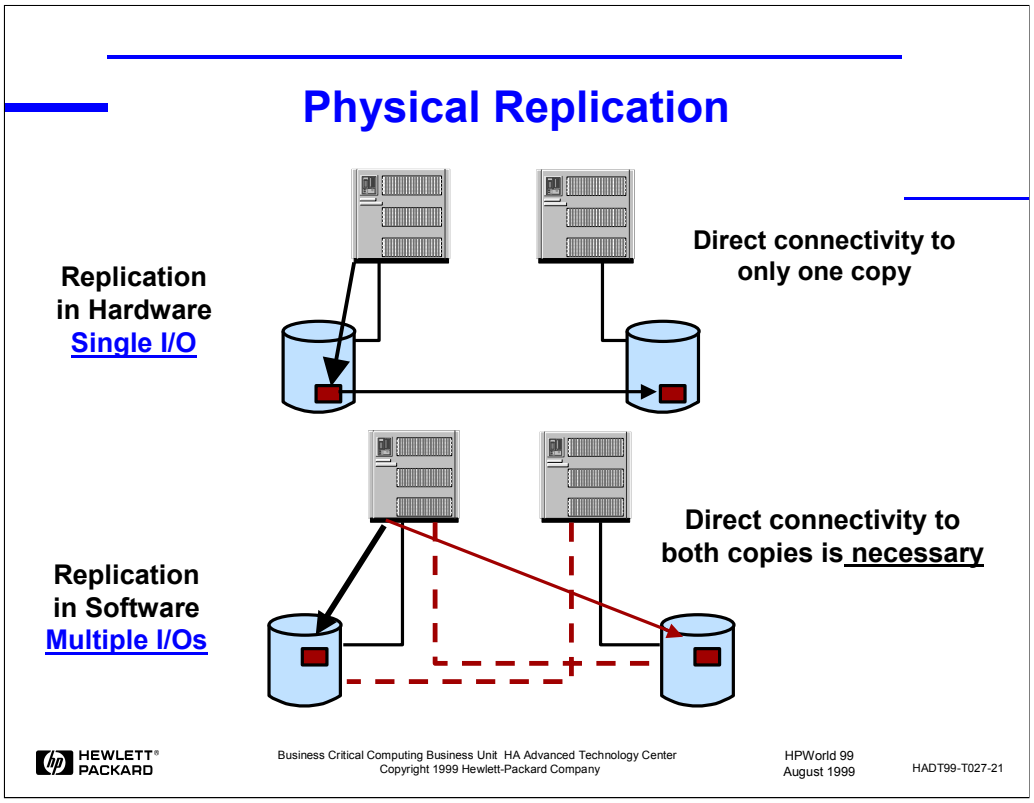
- training and documentation are more complex since the cluster is split across multiple data centers

- testing is more complex and requires personnel in each of the data centers

- disaster practice is important to any disaster recovery environment

Data Replication

- **Definition**
  Scheme by which data is copied from one site to another site for disaster tolerance
- **Physical Replication**
  - **Hardware**
  - **Software**
- **Logical Replication**
  - **File System**
  - **Database**
- **Issues**
  - **Data Consistency**
  - **Data Currency**
  - **Data Recoverability**
  - **Data Loss**



Examples of physical replication in hardware is EMC Symmetrix Remote Data Facility (SRDF) and HP XP256 Continuous Access.

An example of physical replication in software is MirrorDisk/UX.

Advantages and Disadvantages of Physical Replication in Hardware

+consumes no additional CPU overhead

+hardware deals with resynchronization if link or disk fails

+resynchronization is independent of CPU failure
(if disks stay up, CPU failure does not initiate resynchronization)

+write mode is configurable and applies unless the disk or link is down

+built-in ability to copy from replica to the primary copy

+little or no time lag in getting data to the replica

–human errors and database corruption are replicated

–distance is limited by array-to-array capabilities

–requires additional hardware

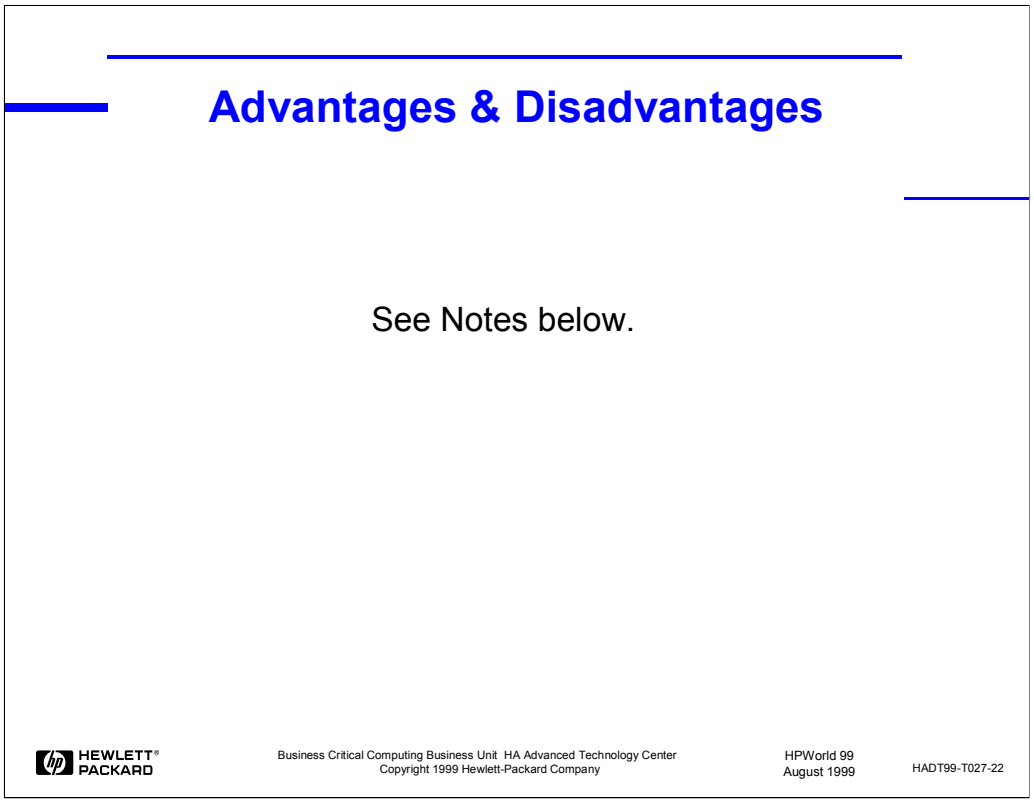
–requires specialized hardware

–may affect performance of I/Os to/from the CPU

–no benefit for reads

–cannot easily monitor the replication or resynchronization

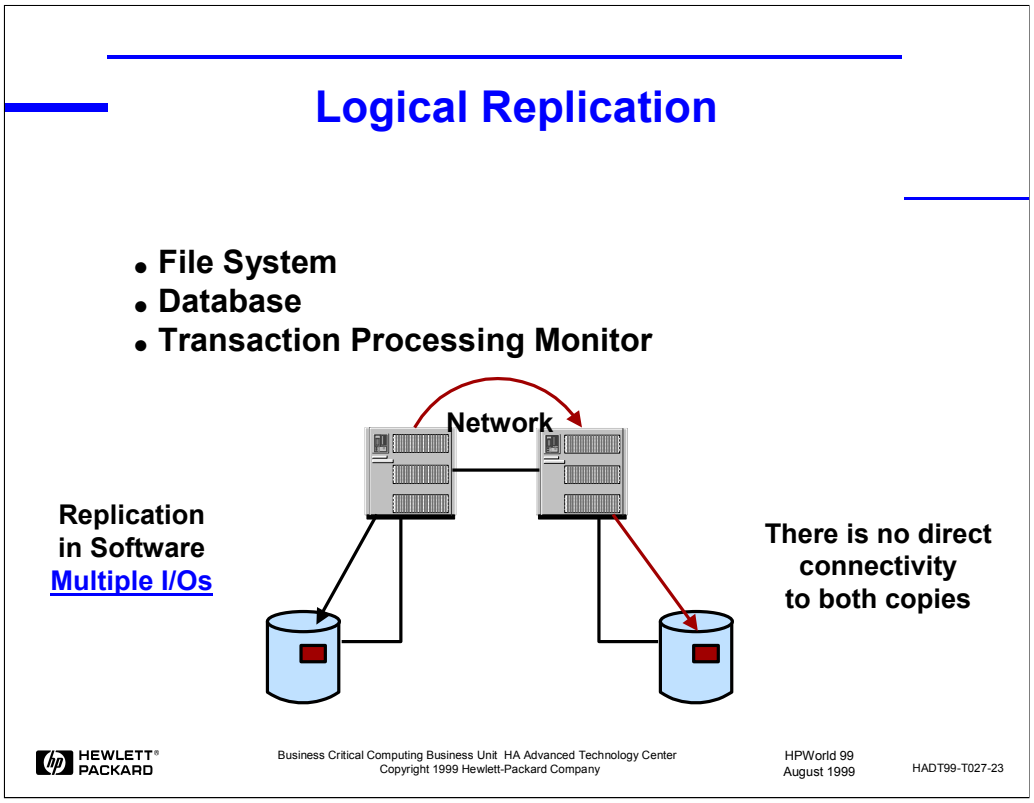
–resynchronization does not currently preserve order of original I/Os



Advantages & Disadvantages of Physical Replication in Software

+independent of disk technology
+may improve read performance (multiple read devices)
+writes are synchronous unless the link or disk is down
+data copies are peers (no master/slave)
+little or no time lag in getting data to the replica

–human errors and database corruption are replicated
–consumes additional CPU overhead for mirroring
–CPU must deal with resynchronization
–CPU failure causes resynchronization even if not needed
–typically degrades write performance
–doubles the I/Os from the CPU
–distance is limited to physical disk link capabilities
–requires additional hardware
–resynchronization does not preserve order of original I/Os



Quest Software offers a product that does file system replication (SharePlex/UX). see http://www.quests.com

Database replication products include:

-Oracle: Advanced Replication and Standby Database products

-Quest Software offers a replication product for Oracle databases (SharePlex/Replication).

-Informix: standard replication product

-Sybase: Replication Server

- **Logical replication of database**
  - **+distance is limited only by the network**
  - **+requires no additional hardware**
  - **+multiple copy corruption is unlikely since transactions are replicated**
  - **+roll forward and rollback capabilities**
  - **–consumes additional CPU overhead**
  - **–consumes network bandwidth**
  - **–may be a significant time lag with getting transactions to replica**
  - **–no automated way to copy data back to the primary if the replica becomes the active copy (due to primary copy failure)**

Logical replication of file system (same as with database plus:)

- **–data in OS buffers may be lost upon CPU failure**
- **–no roll forward/rollback capabilities**

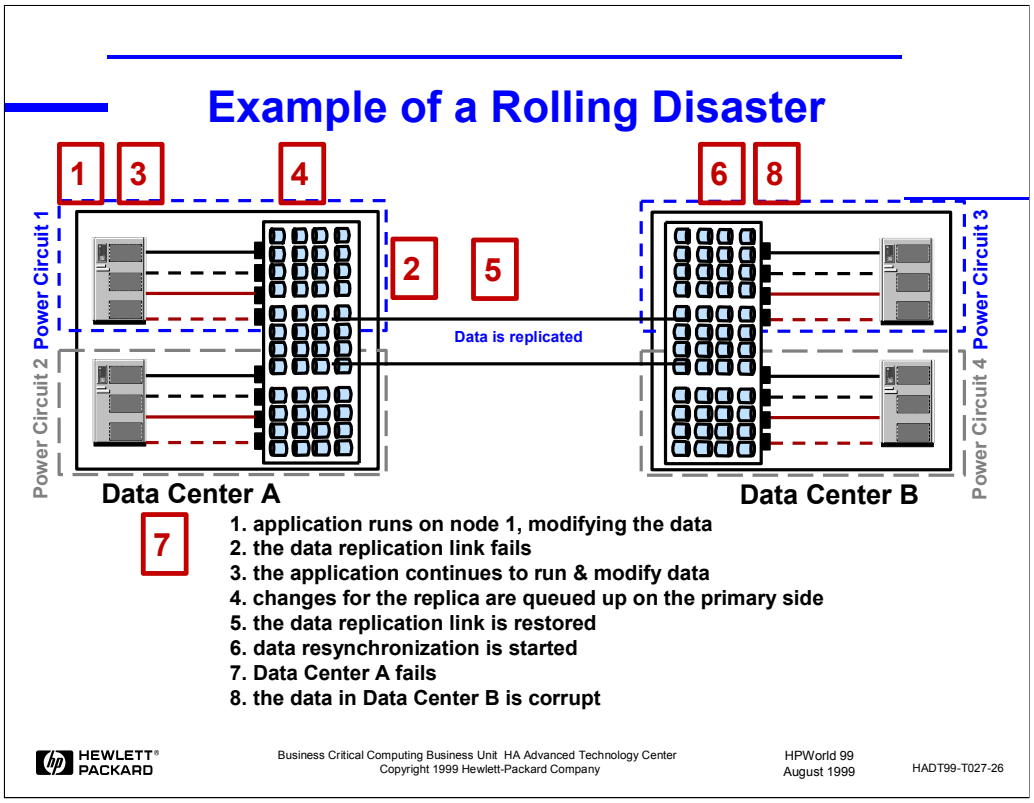
Ideal Situation

- replicate physically for speed and currency
- replicate logically for consistency and to recover from human error
- use the replica only when all other physical copies are corrupt

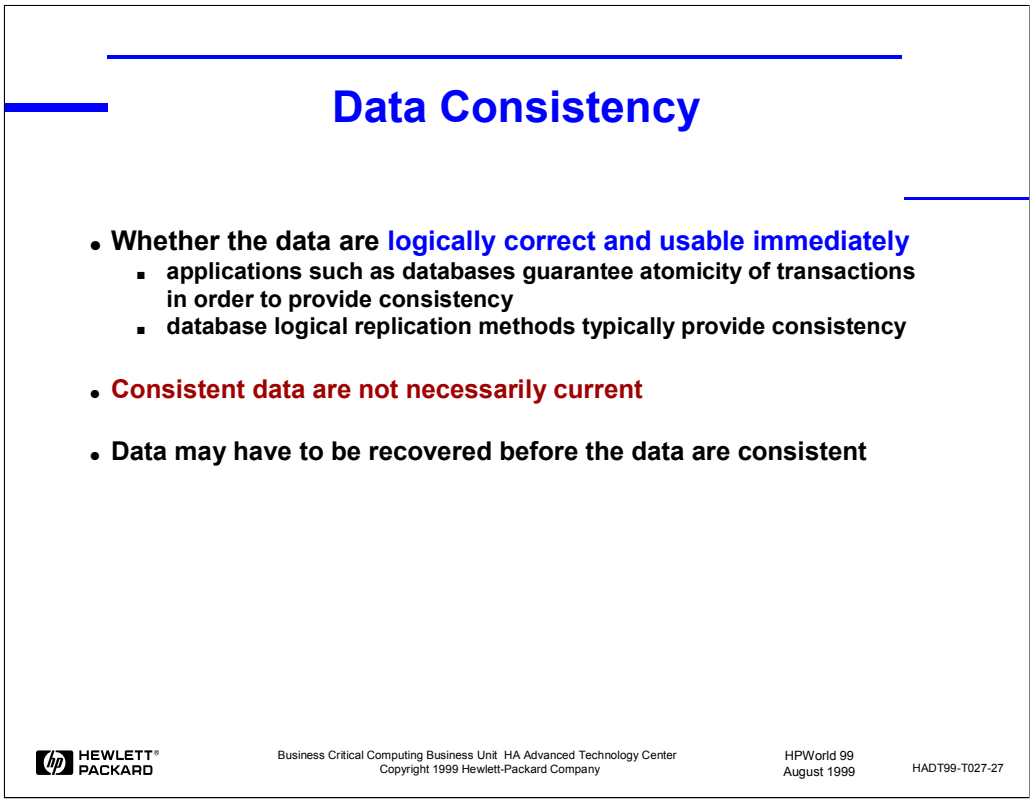
Rolling Disasters

- Rolling disasters occur when
  - a second failure occurs before the recovery from the first failure
  - using physical data replication
- Rolling disasters may result in
  - inconsistent data (corrupt)
  - non-current data
- Recovery from a rolling disaster
  - is typically manual
  - may require data reload from tape
- Rolling disasters are more common in the wide-area case due to the higher probability of network link outages, even of short duration

Cluster architectures and products must be designed to minimize the probability that a rolling disaster might occur.



A rolling disaster such as this may occur when using MirrorDisk/UX, Symmetrix Remote Data Facility or XP256 Continuous Access to replicate the data



Physical replication does not preserve the order of the writes during resynchronization after a link failure or remote disk failure. If the primary site fails during a resynchronization, the remote data will be left in an inconsistent state. Therefore, it is advisable to use one of two methods to guarantee consistency but not necessarily currency:

- a point-in-time split-off copy at the remote site

- logical database replication

For a file system, the data are consistent only if

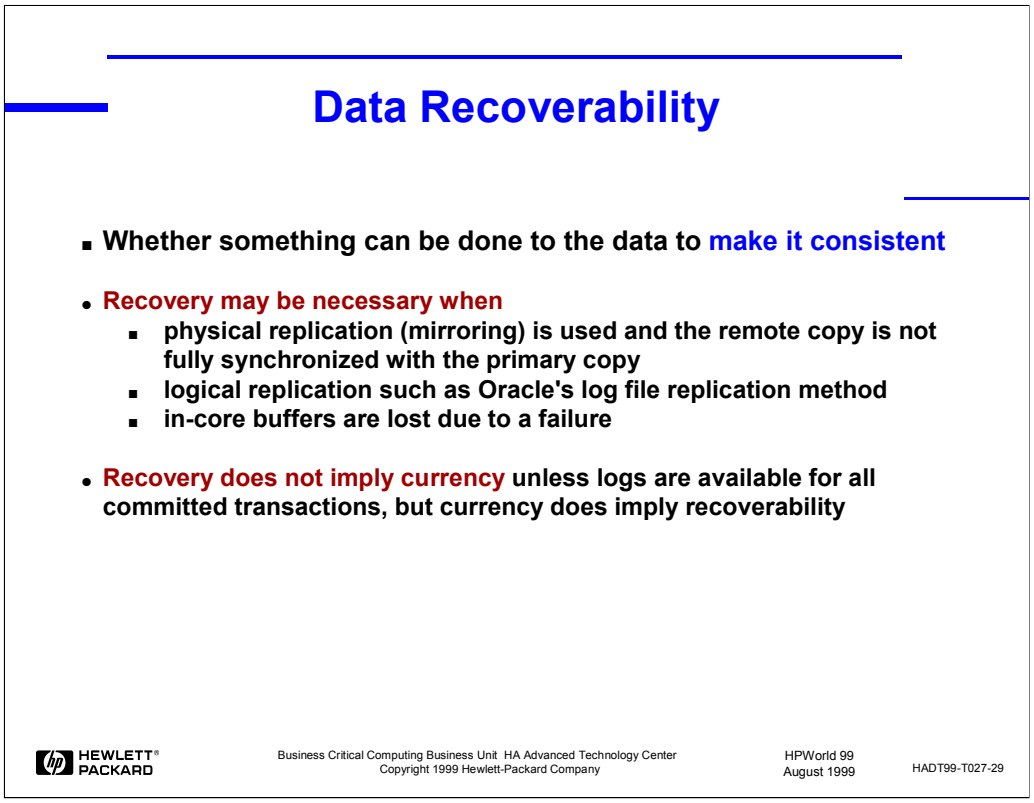
- **the O_SYNC flag is used to force synchronous writes**
- **the file system has been cleanly unmounted or the buffers fully flushed**

- **For a database, the data are consistent if only committed transactions have been applied (all uncommitted transactions rolled back)**

Data Currency

- Whether the remote database can be recovered to include all committed transactions that were applied to the local database
- Synchronous physical replication guarantees currency of the remote copy only while the link remains operational
  - if the link fails, data changes are queued up, resulting in non-current data on the replica
- Most database logical replication methods allow the remote copy to lag behind the primary copy (local site) because they operate asynchronously for performance -- the replica is, therefore, not current

- Split-off copies are *not current if transactions continue to the main copy*
- Uncommitted transactions (in-flight transactions) that are rolled back are not a currency issue
- Any transactions (data) in a computers memory that is not yet written to disk are lost upon system failure



In-core memory buffers may be lost due to

- memory fault
- power loss
- system crash

- **For a database, the data may be recoverable, for example, by either or both**
  - **reload of the database from a point-in-time copy (tape or disk)**
  - **roll forward of committed and roll back of uncommitted transactions**
- **For a file system, the metadata are recoverable for**
  - **an HFS with fs_async=0**
  - **a JFS, due to the intent log**
- **For a file system, the data are not recoverable if not already consistent**

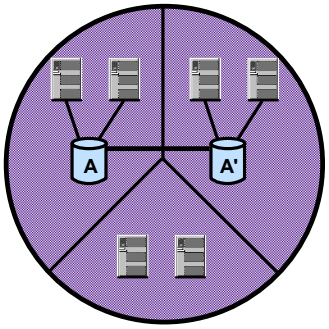
Data Loss

- **Data loss will occur in a cluster (it's usually a matter of how much)**
  - **if recovery fails**
  - **if a system or disk failure occurs in the middle of a data reorganization**
  - **due to human error**
  - **during a rolling disaster**
  - **due to logical software bugs**
  - **if non-synchronous replication is used and a failure occurs**
  - **if anything else that would cause data loss on a single system**
- **Certain critical applications must ensure that no transaction is ever lost. Special techniques must be used to prevent data loss:**
  - **Transaction Processing Monitors (TPMs)**
  - **transaction logging at another site with replay capabilities**

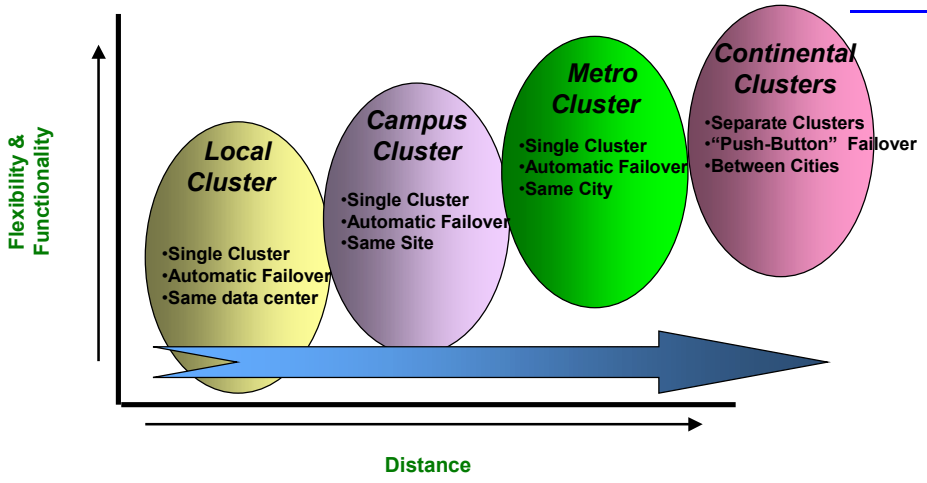
Even the airlines are willing to suffer a certain amount of data loss with their reservation systems, such as a lost seat assignment.

Some applications such as in the finance area require no loss of data.

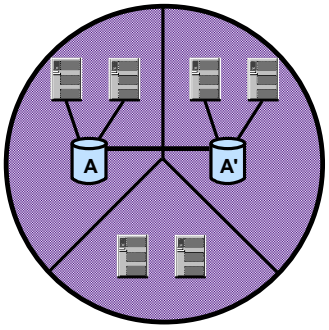
Disaster-Tolerant Cluster Architectures



Range of Architectures



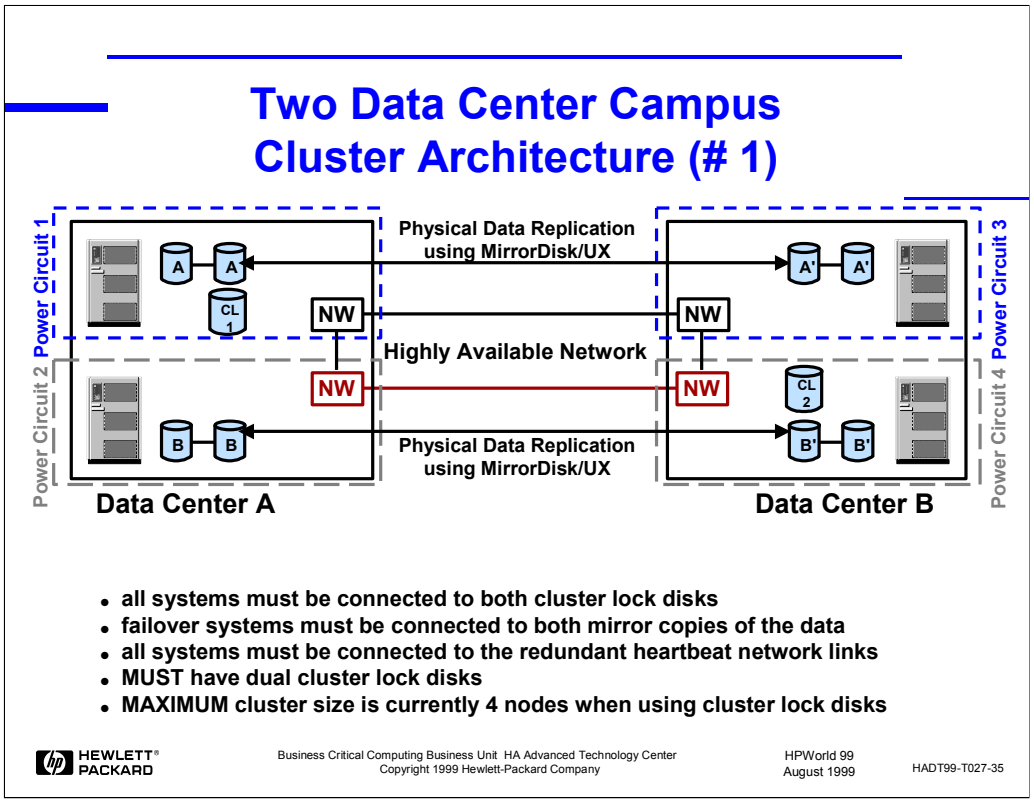
Campus Clusters



Campus Cluster Rules

- **Single campus cluster with automated failover**
  - **Maximum cluster size is 4 nodes**
  - **Dual cluster lock disks to maintain quorum in case an entire data center fails**
  - **Maximum distance between data centers is 10 km (FibreChannel)**
- **Network**
  - **Redundant network connections routed differently**
  - **Redundant network components powered separately**
  - **Must have at least two networks for cluster heartbeat**
- **Data**
  - **Physical data replication using MirrorDisk/UX software**
  - **Redundant data connections routed differently**
  - **Redundant data components (e.g., Fibre Channel Hubs) powered separately**

The redundant network and data replication links now become the most critical resource. It is very important to architect these links correctly.

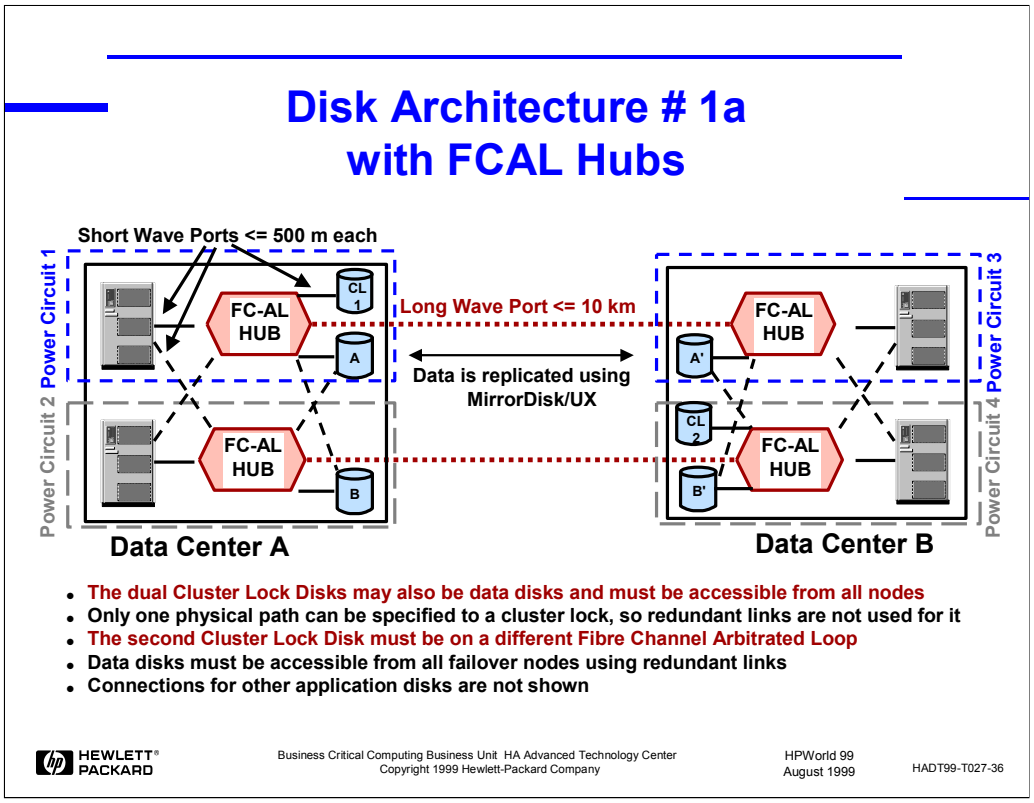


Implementations:

- **EMC Symmetrix with Fibre Channel Arbitrated Loop (FCAL)**
  - **FCAL Point-to-Point**
  - **FCAL with Hubs (Max 2 hubs & 4 hosts per loop)**
- **HP SureStore E Disk Array XP256 with Fibre Channel Arbitrated Loop (FCAL)**
- **HADA Model 30 FC**
  - **FCAL with Hubs (Max 2 hubs & 4 hosts per loop)**

Model 30 used as Cluster Lock must have Auto-Trespass disabled

- **AutoRAID**
  - **FibreChannel/SCSI Mux, no Hubs (2-node cluster only)**
  - **FibreChannel/SCSI Mux, Hubs (Max 2 hubs & 4 hosts per loop)**



Advantages and Disadvantages of Configuration # 1:

+lowest cost

+only two data centers are needed

+no Arbitrator(s) is/are needed

+all systems are connected to both copies of the data (good if the primary disk fails, but the primary systems stay up)

+resynchronization may occur from either side

+bi-directional replication is possible

–slight chance of split brain with dual cluster locks

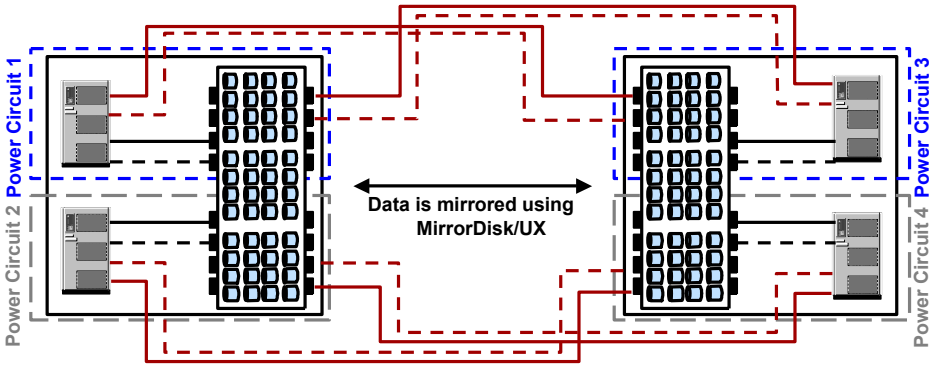
–maximum 10 km between data centers

–increased CPU overhead (for mirroring)

–Cost issues

- **–Fibre Channel disk links are required for local and remote connectivity**
- **–all systems MUST be connected to both copies of the data**
- **–Dual Cluster Lock Disks are required**

Disk Architecture # 1b with FCAL Point-to-Point



- Disks may be EMC Symmetrix or HP XP256
- Maximum distance from any host to Symmetrix is 500 m
- Each Symmetrix must contain one of the dual Cluster Lock Disks
- Host connections must be made using redundant links

Circumstances for Split Brain

With the 2-Data Center Architecture and Dual Cluster Lock Disks, split brain syndrome will occur if:

- **ALL heartbeat networks fail**

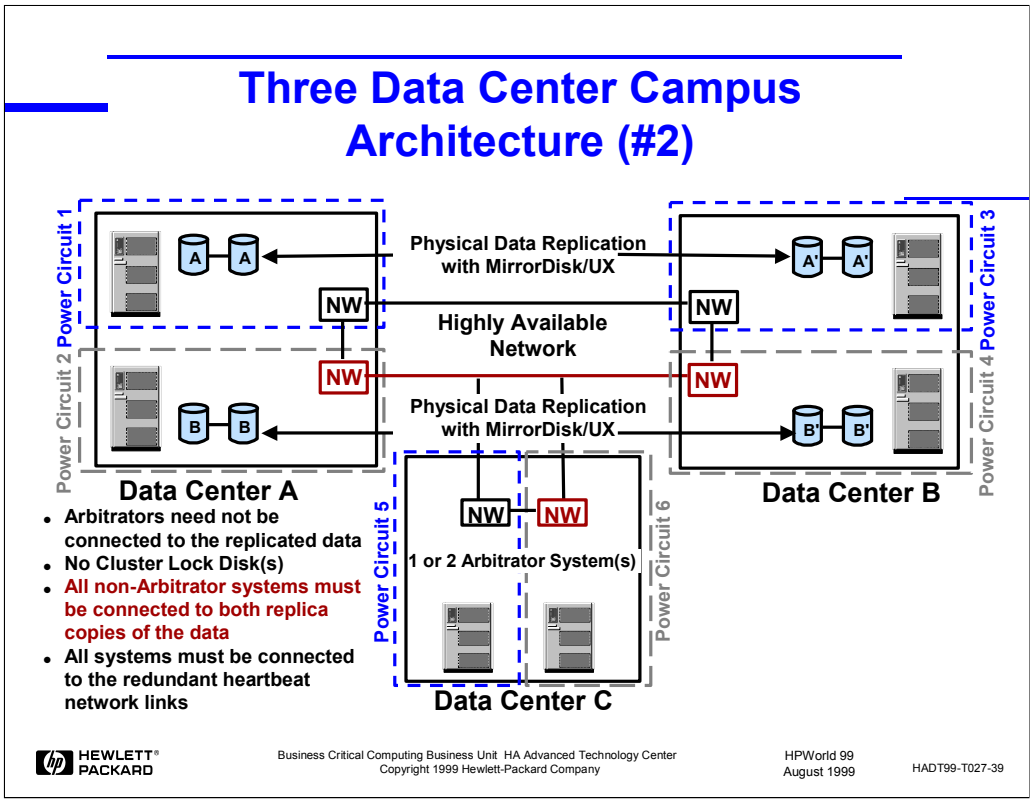
AND

- **the disk link from Data Center A to Cluster Lock # 2 fails**

AND

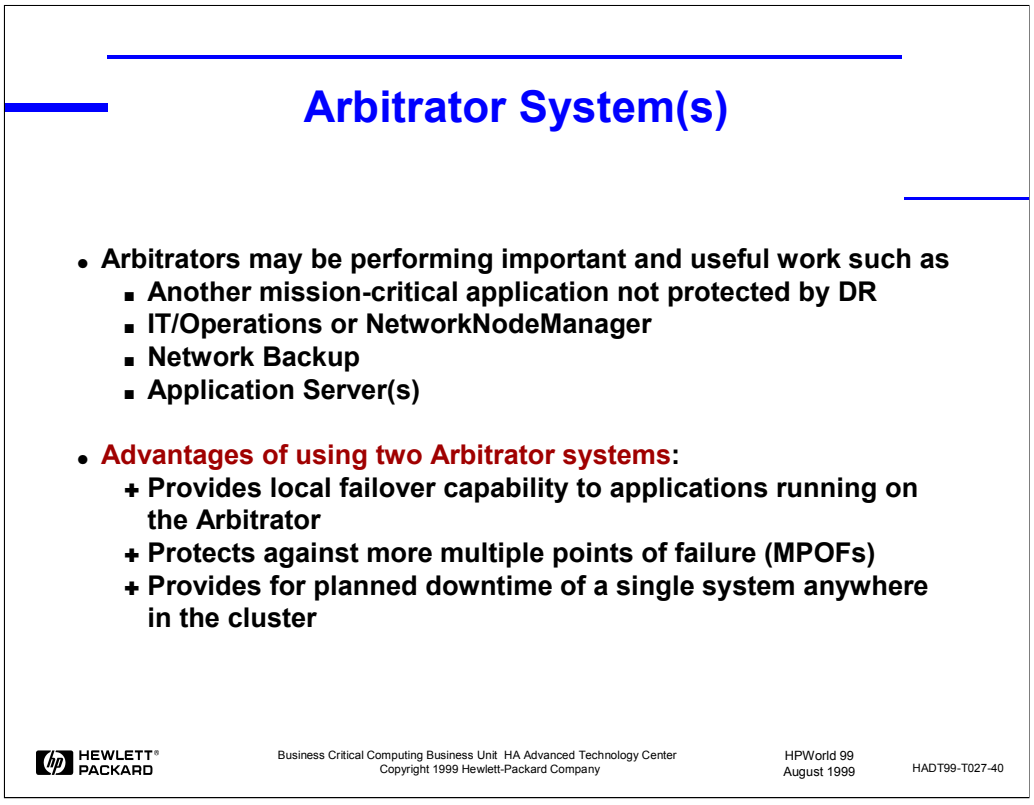
- **the disk link from Data Center B to Cluster Lock # 1 fails**

The result is data corruption!



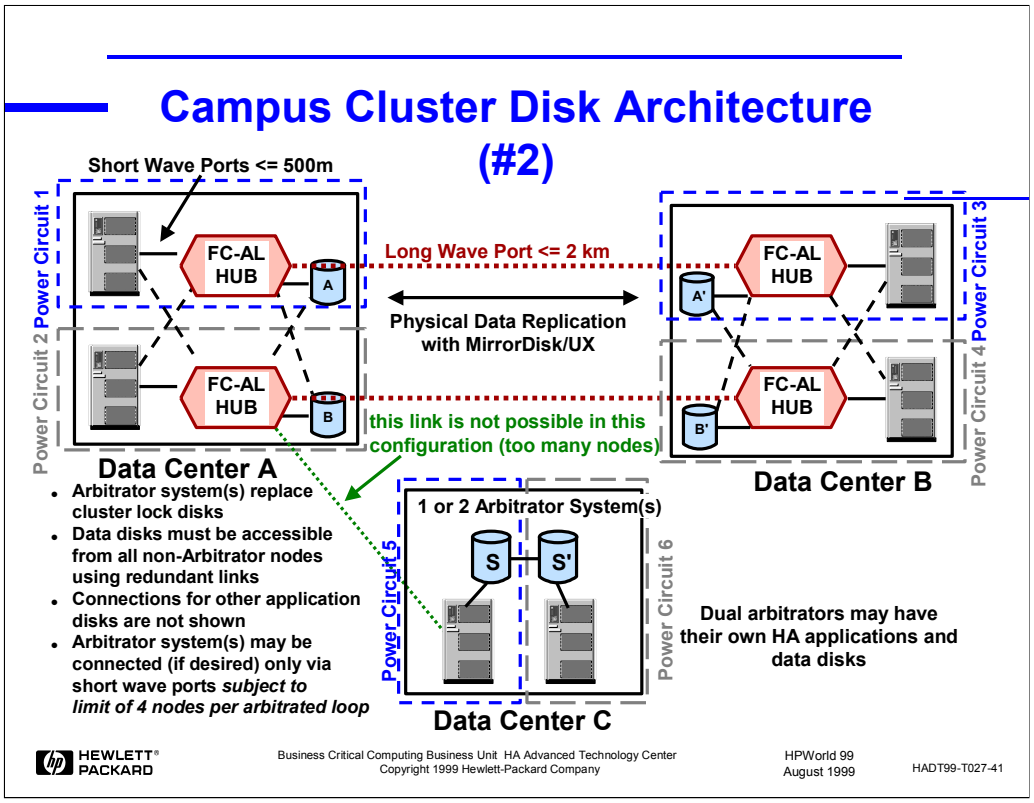
Implementations:

- EMC Symmetrix with Fibre Channel Arbitrated Loop (FCAL)
  - FCAL Point-to-Point
  - FCAL with Hubs (Max 2 hubs & 4 hosts per loop)
- HP SureStore E Disk Array XP256 with Fibre Channel Arbitrated Loop (FCAL)
- High Availability Disk Array Model 30FC
  - FCAL with Hubs (Max 2 hubs & 4 hosts per loop)
- AutoRAID
  - FibreChannel/SCSI Mux, no Hubs (2-node cluster only)
  - FibreChannel/SCSI Mux, Hubs (Max 2 hubs & 4 hosts per loop)



When using only one Arbitrator, special procedures must be followed during times of planned downtime in order to remain protected. Systems must be taken down in pairs, one from each of data centers A and B. If the Arbitrator itself must be taken down, the DR capability is at risk if one of the other systems fails.

Implementing two Arbitrators provides greater flexibility in taking systems down for planned outages as well as protecting against more multiple failures.



Advantages and Disadvantages of Configuration # 2

+no chance of split brain

+no Cluster Lock Disks are used

+all systems are connected to both copies of the data
(if the primary disk fails, no need for remote failover)

+resynchronization may occur from either side

+bi-directional replication is possible

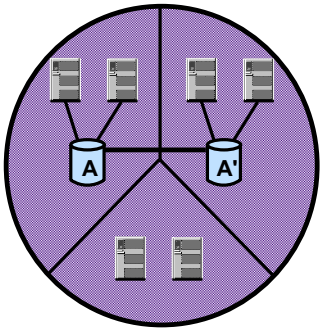
–higher cost (hardware, software & data center)

- –three data centers are needed
- –one or two Arbitrator systems are needed
- –Fibre Channel disk links are required for local and remote connectivity
- –all systems MUST be connected to both copies of the data

–maximum 10 km between data centers

–increased CPU overhead (for mirroring)

MetroCluster



The MetroCluster Products: A Short Look

HP MetroCluster with EMC SRDF (B6264BA)

- A product to automate the failover of MC/ServiceGuard packages among nodes using two Symmetrix disks that are connected by SRDF

HP MetroCluster with Continuous Access XP (B8109BA)

- A product to automate the failover of MC/ServiceGuard packages among nodes using two XP256 disk arrays that are connected by Continuous Access XP (CA)
  - each node is only attached to either the primary or the secondary disk array
  - both local and remote failover is supported
  - both failover and failback are supported

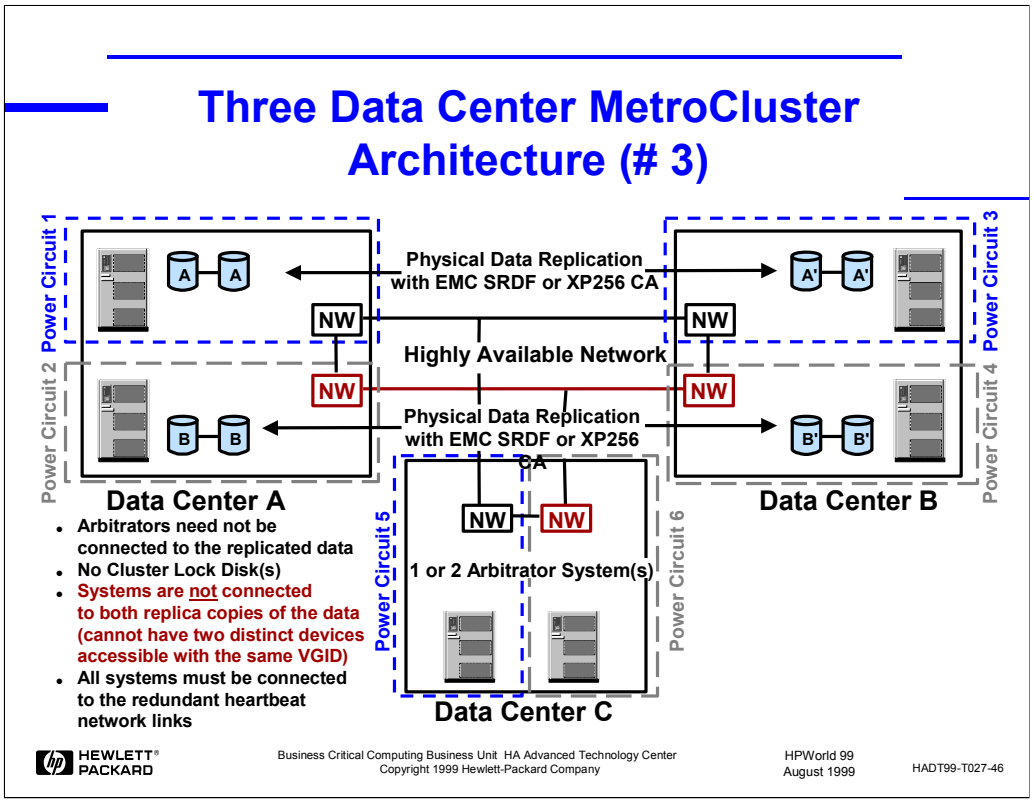
Product Dependencies

- **HP-UX & MC/ServiceGuard**
  - HP-UX 10.20, MC/ServiceGuard 10.10 only
  - HP-UX 11.0, MC/ServiceGuard 11.01 and later
- **Host Adapters**
  - HP-PB: F/W SCSI (requires SCSI Pass-Through patch for HP-UX 10.20 with EMC SRDF only)
  - HSC: F/W SCSI, Fibre Channel
  - PCI: F/W SCSI, Fibre Channel
- **EMC Symmetrix**
  - SRDF Automatic Failover Module (product # SRDF-HP-MC) also called SymCLI, software version T3.0 and later
  - No shared devices (except BCVs) with HP or non-HP hosts that are outside of the MetroCluster
- **HP SureStore E Disk Array XP256**
  - Raid Manager software
  - No shared devices (except BCs) with HP or non-HP hosts that are outside of the MetroCluster

MetroCluster Rules

- **Single campus or metropolitan area cluster with automated failover**
  - **All nodes are members of a *single* MC/ServiceGuard cluster**
  - **Maximum cluster size**
    - **8 nodes with HP-UX 10.10 and later 10.x versions**
    - **16 nodes with HP-UX 11.0 and later 11.x versions**
  - **Same number of nodes in each non-Arbitrator data center to maintain quorum in case an entire data center fails**
  - **Maximum distance among the three data centers is 100 km**
  - **Maximum distance between the disk arrays is 43 or 60 km**
  - **One or two Arbitrator systems for quorum (NO cluster lock disks)**
  - **Exclusive Volume Group activation only**
- **Network**
  - **Redundant network connections routed differently**
  - **Redundant network components powered separately**
  - **Must have at least two networks for cluster heartbeat**
- **Data (Physical data replication in hardware)**
  - **Redundant data connections routed differently**
  - **Redundant data components (e.g., Disk Arrays, ESCON link components) powered separately**

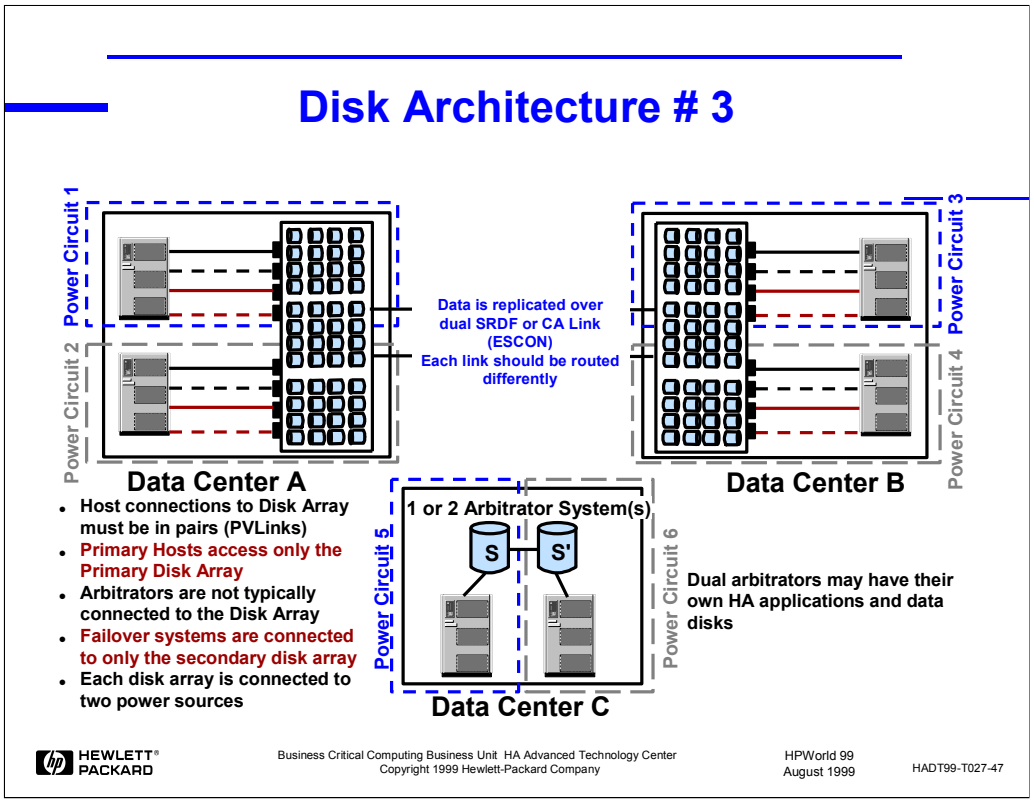
The redundant network and data replication links now become the most critical resource. It is very important to architect these links correctly.



Implementations:

- **EMC Symmetrix with Symmetrix Remote Data Facility (SRDF) for remote data replication (Synchronous Only)**
  Local disk connectivity:
  - **F/W SCSI**
  - **FCAL Point-to-Point**
  - **FCAL with Hubs**

- **HP SureStore E Disk Array XP256 with Continuous Access XP (CA) for remote data replication (Synchronous Only)**
  Local disk connectivity:
  - **FCAL with SCSI Mux**
  - **FCAL Point-to-Point**
  - **FCAL with Hubs**



Advantages and Disadvantages of Configuration # 3

+no chance of split brain

+reduced CPU overhead (replication is in hardware)

+distances up to 60 km (SRDF) or 43 km (CA XP) between disk arrays

+distances up to 100 km among all three data centers (FDDI)

+no Cluster Lock Disks are required

+Fibre Channel or F/W SCSI for local connectivity

+systems are not connected to both copies of the data

+manually invokable feature to copy data back from remote side

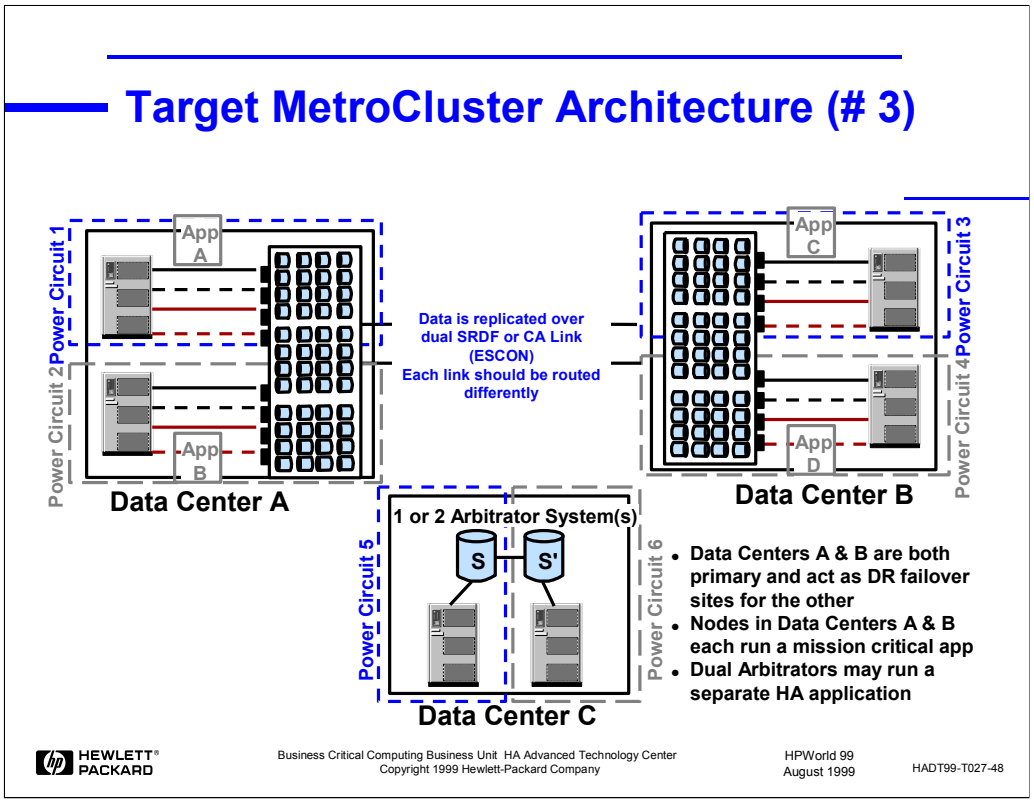
+bi-directional replication is possible

–higher cost

- –three data centers are needed
- –one or two Arbitrator systems are needed
- –SRDF or CA hardware and software

–all systems are connected to only one copy of the data
(primary disk failure requires failover to the remote systems)

–when failed over to the DR site, there is no remote protection for the data



The target architecture involves application packages running on each host, I.e., in both data centers. In this case, the data centers are peers that back each other up in case of disaster.

Of course, systems must be of appropriate capacity if they are to run both their own and the other data center hosts' applications.

Three Data Center Architecture
Numbers of Nodes

Primary Data Center A (with Disk Array)	Primary Data Center B (with Disk Array)	Arbitrator Data Center C (NO Disk Array)
1	1	1
2	2	1
2	2	2*
3	3	1
3	3	2*
4	4	1
4	4	2*
5	5	1
5	5	2*
6	6	1
6	6	2*
7	7	1
7	7	2*

*** Configurations with 2 Arbitrators are preferred**

The same number of nodes must be present in Data Centers A & B

Otherwise, certain failure scenarios will cause the entire cluster to halt

Configurations with two Arbitrators are preferred since they provide a greater degree of availability, especially in cases when a node is down due to a failure or planned maintenance.

Symmetrix Configuration Requirements for Support

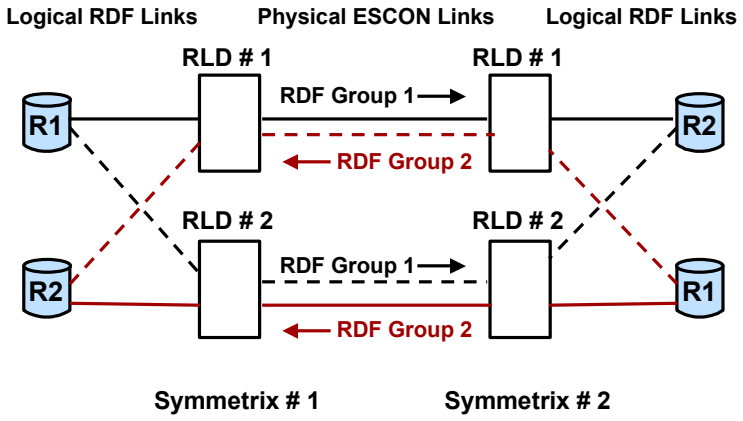
- R1 devices are locally protected (RAID-1 or RAID-S)
- R2 devices are locally protected (RAID-1, RAID-S or BCV)
- Only synchronous mode is enabled during cluster operation (Adaptive Copy mode must be disabled)
- Domino mode is recommended
- No manual changes to the state of Symmetrix devices that belong to the MetroCluster packages during cluster operation
- All Symmetrix devices that belong to a particular MetroCluster package must be in the same state at the same time

Symmetrix SRDF Failback

- After failover to the remote data center, data is unprotected remotely from a DR perspective, until:
  - Primary data center is repaired
    AND
  - Application is failed back

Bi-directional SRDF

- requires 4 physical links for performance reasons
- HA redundancy requires 2 Remote Link Directors (RLDs) (2-port or 4-port)
- RDF groups must be defined to ensure no SPOF



For each RDF group, one Symmetrix is defined as a master and the other as a slave

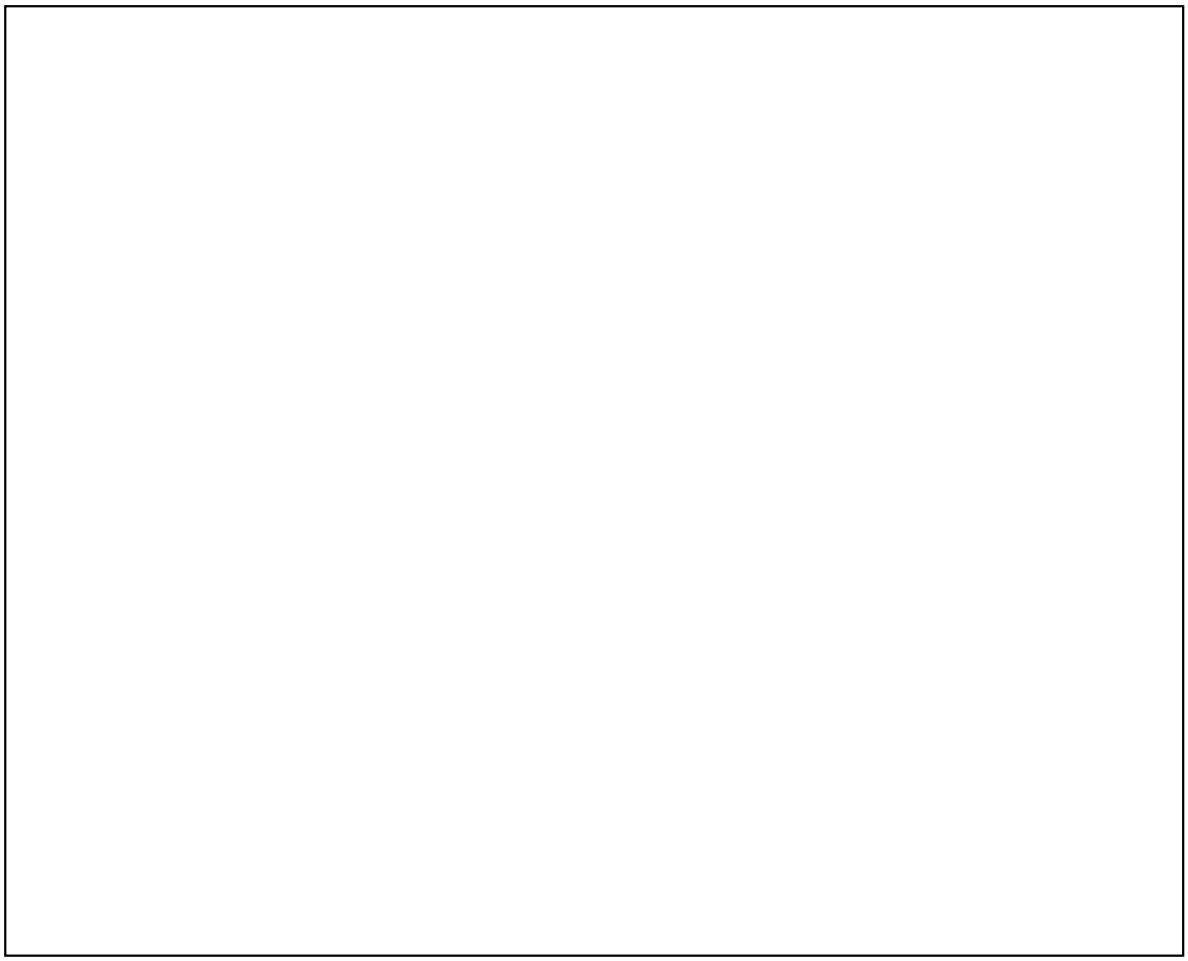
For performance reasons, the R1s on one side should use one RDF group that is assigned to one pair of links while the R1s on the other side use a different RDF group assigned to a different pair of links

XP256 Configuration Requirements for Support

- PVOL devices are locally protected (RAID-1 or RAID-S)
- SVOL devices are locally protected (RAID-1 or RAID-S)
- Only synchronous mode is enabled during cluster operation
- Fence Level Data is recommended
- No manual changes to the state of XP256 devices that belong to the MetroCluster packages during cluster operation
- All XP256 devices that belong to a particular MetroCluster package must be in the same state at the same time

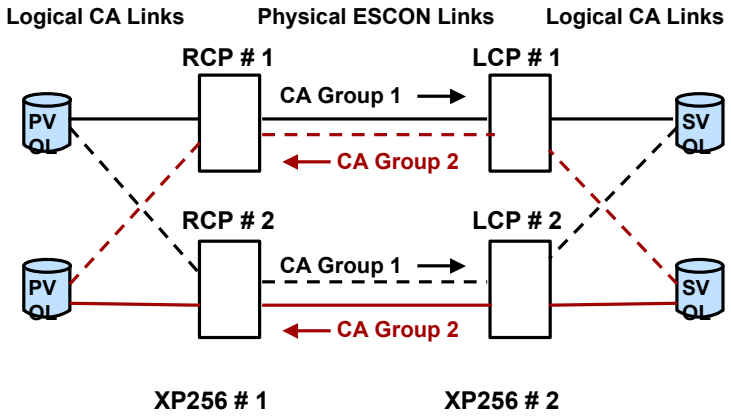
XP256 Continuous Access Failback

- After failover to the remote data center, data is unprotected remotely from a DR perspective, until:
  - Primary data center is repaired
    AND
  - Application is failed back
    OR
  - PVOL/SVOL personalities are swapped

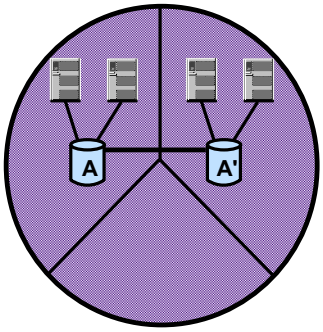


Bi-directional CA

- requires 4 physical links for performance reasons
- HA redundancy requires 2 RCP/LCP pairs
- CA groups must be defined to ensure no SPOF



ContinentalClusters



ContinentalClusters : A Short Look

- **HP ContinentalClusters (B7659BA)**
  - **A product to automate the failover of MC/ServiceGuard packages among TWO separate clusters**
    - **primary and secondary cluster failure notification is configurable**
      - **e-mail (including e-mail to a pager)**
      - **SNMP trap**
      - ***opcmsg***
    - **semi-automatic "push button" initiates automated failover**
    - **choice of various logical or physical data replication methods**
    - **local failover occurs under control of MC/ServiceGuard**
    - **remote failover occurs only when entire primary data center fails**
    - **currently, only uni-directional failover is supported**
    - **includes scripts for EMC SRDF and HP SureStore E Disk Array XP256 (physical replication)**
    - **network failover and failback defined and implemented by user**

Product Dependencies

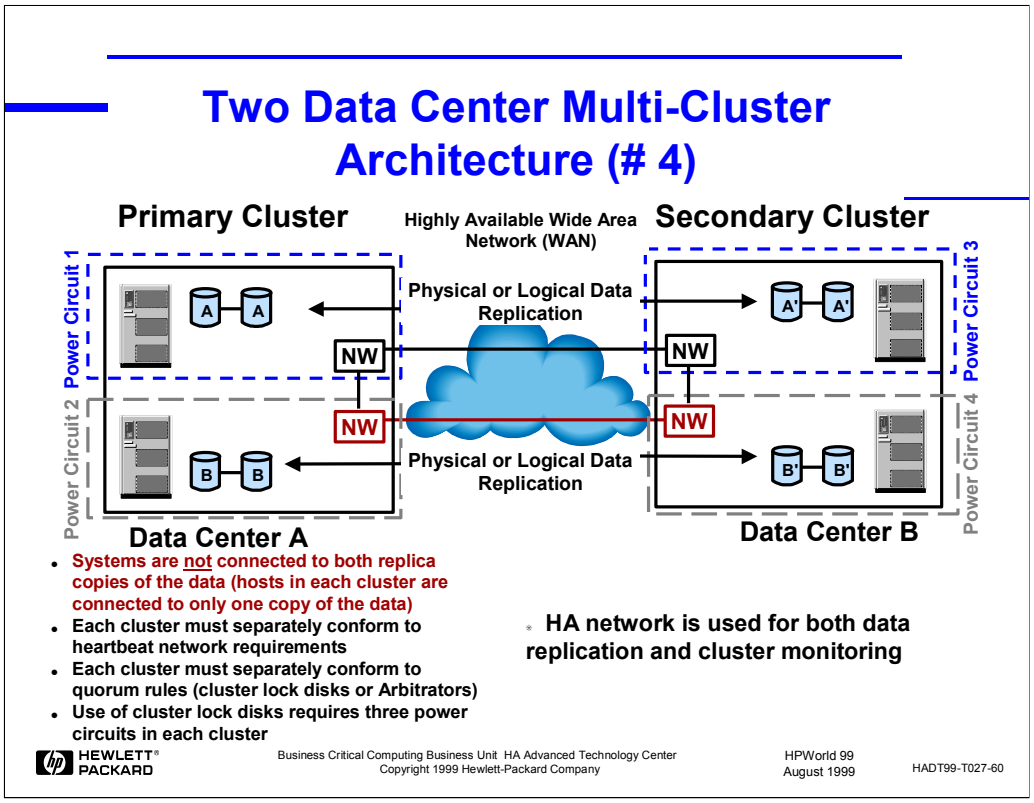
- **HP-UX & MC/ServiceGuard**
  - **HP-UX 11.0, MC/ServiceGuard 11.08 and later**
- **Choice of Logical or Physical Data Replication**
  - **EMC Symmetrix with SRDF (Physical)**
    - **SRDF Automatic Failover Module (product # SRDF-HP-MC) also called SymCLI, software version T3.0 and later**
    - **No shared devices (except BCVs) with HP or non-HP hosts that are outside of the MetroCluster**
  - **HP SureStore E Disk Array XP256 with CA (Physical)**
    - **Raid Manager software**
    - **No shared devices (except BCs) with HP or non-HP hosts that are outside of the MetroCluster**
  - **Oracle Standby Database (Logical)**

ContinentalClusters Rules

- **Dual clusters with semi-automated ("push button") failover**
  - **Maximum cluster size is 16 nodes for each cluster (HP-UX 11.x)**
  - **Each cluster may be composed of a different number of hosts**
  - **Cluster pairs may be MC/ServiceGuard or ServiceGuard OPS Edition**
  - **Each cluster is subject separately to cluster quorum rules**
  - **Maximum distance between clusters is limited by WAN technology for networks and disk replication links (T1, T3/E3, ATM, SONET, etc.)**
- **Network**
  - **Wide Area Network (WAN) must support TCP/IP protocols**
  - **Redundant network connections routed differently**
  - **Redundant network components powered separately**
  - **Recommend at least two networks for inter-cluster monitoring and data replication**
- **Data**
  - **Physical or Logical Data Replication**
  - **Redundant data connections routed differently**
  - **Redundant data components (e.g., Disk Arrays, mirrored disks, ESCON link components) powered separately**

The dual clusters may be either two MC/ServiceGuard clusters or two ServiceGuard OPS Editions (MC/LockManager) clusters.

The redundant network and data replication links now become the most critical resource. It is very important to architect these links correctly.

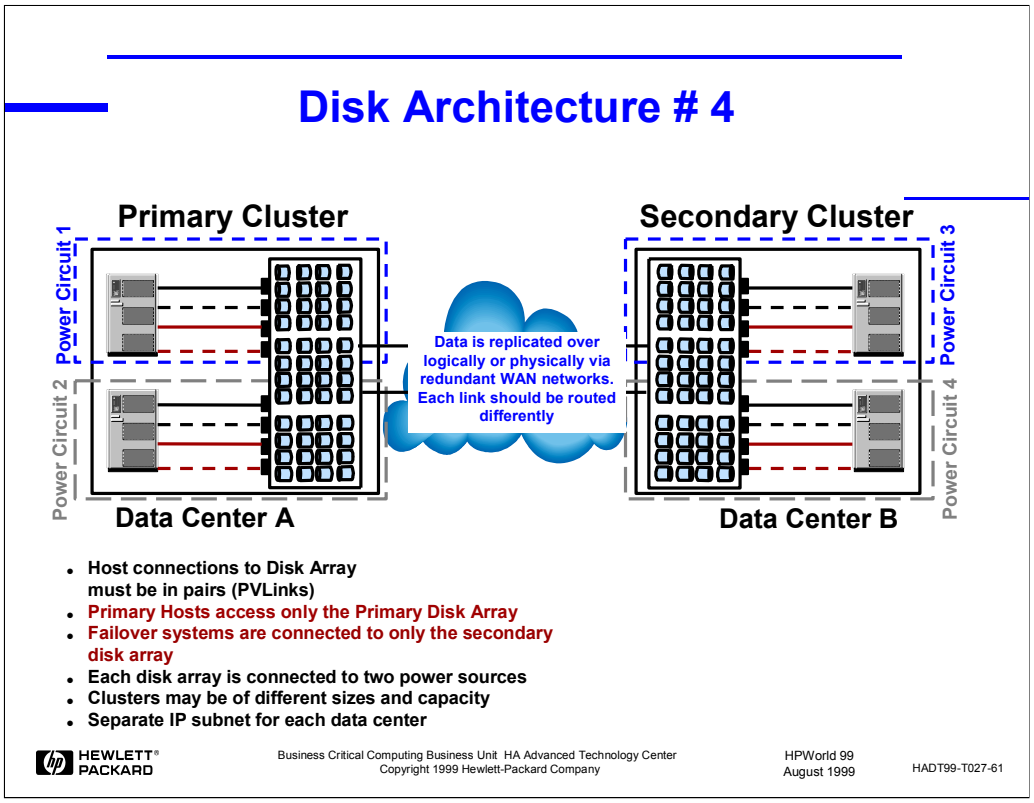


Implementations:

- **EMC Symmetrix with Symmetrix Remote Data Facility (SRDF) for remote data replication (Synchronous Only)**

- **HP SureStore E Disk Array XP256 with Continuous Access XP (CA) for remote data replication (Synchronous Only)**
Local disk connectivity:

- **Oracle Standby Database**



Advantages and Disadvantages of Architecture # 4

+no chance of split brain due to "push button"

+choice of several logical and physical data replication methods

+distances up to limit of network

+systems are not connected to both copies of the data

+manually invokable feature to copy data back from remote side with physical replication in hardware

–higher cost

–special replication hardware or software is needed

–additional CPU overhead for logical replication, if applicable

–client reconnect is more difficult with multiple IP subnets

–no feature to replicate changes back from remote copy with logical replication

–all systems are connected to only one copy of the data
(primary disk failure requires failover to the remote systems)

–bi-directional replication is less feasible (cost, network bandwidth)

–when failed over to the DR site, there is no remote protection for the data

ContinentalClusters : The Process

- **Failover**
  - Failure of primary cluster is detected by the secondary cluster
  - User is notified of the failure of the primary cluster
  - User activates the “push button”
    - Could optionally be automated (e.g., IT/Operations automatic action)
  - The data replication receiver processes (logical replication only) are halted
  - The disk arrays are reconfigured for read/write access (physical replication only)
  - The primary application packages are started on the secondary cluster
- **Failback**
  - Application packages are shutdown
  - Database is backed up and transferred to the repaired or new primary site (logical replication only)
  - Database is copied back to the repaired or new primary site (physical replication only)
  - Application packages are restarted

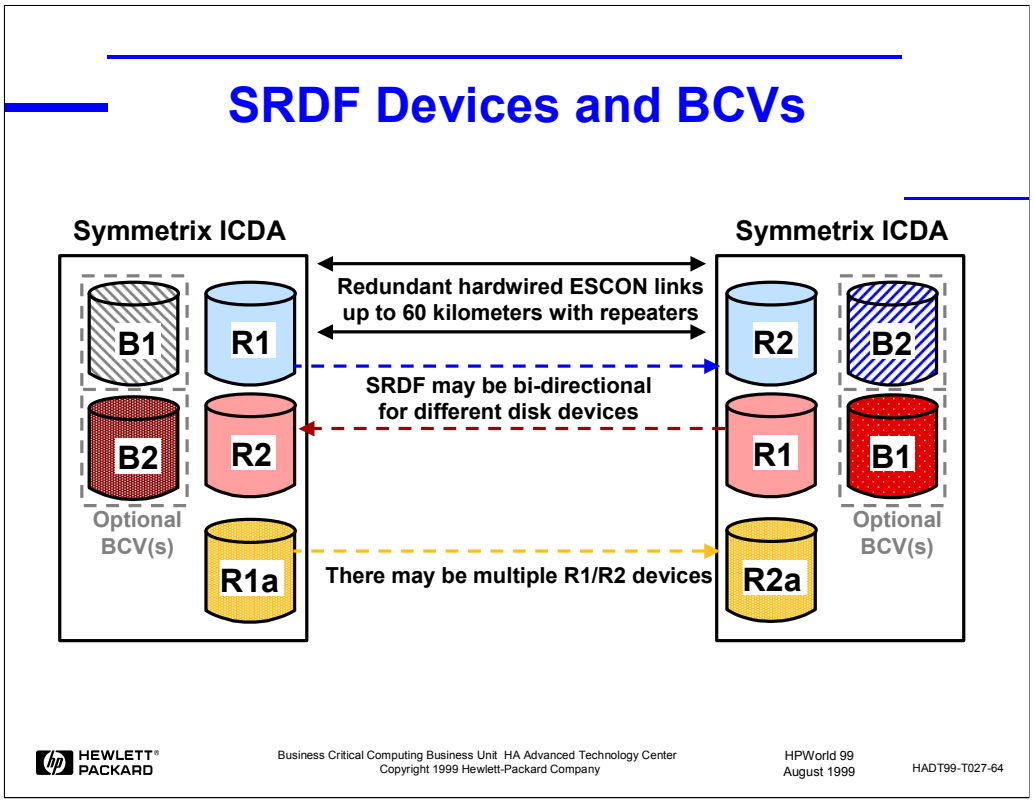
Symmetrix Configuration Requirements for Support

- R1 devices are locally protected (RAID-1 or RAID-S)
- R2 devices are locally protected (RAID-1, RAID-S or BCV)

- Only synchronous mode is enabled during cluster operation (Adaptive Copy mode must be disabled)
  - Secondary copy of the database is almost always current and consistent
  - No committed transactions should be lost upon failover

- Domino mode is recommended
- If Domino mode is not used, BCVs are required to provide point-in-time consistent copies of the data

- No manual changes to the state of Symmetrix devices that belong to the MetroCluster packages during cluster operation
- All Symmetrix devices that belong to a particular MetroCluster package must be in the same state at the same time



•The R1 device is the primary device and is always readable and writable

•The R2 device is the secondary or remote device and is not writable while the link between it and the R1 device is active

•Business Continuation Volumes (BCVs) are additional copies of the data that may be split and re-merged with the main copy of the data

- •requires additional disk devices
- •may be used in either the primary (local) Symmetrix or the secondary (remote) Symmetrix
- •may be split from its primary copy to keep point-in-time consistent copies of the data

SRDF Domino Mode on Symmetrix

- new I/Os are refused if any component associated with an R1/R2 pair or the link between them fails -- causing application to hang or abort
- if the entire R2 data center fails, the application will halt and manual intervention is necessary to restart the application on the R1 side
- in a MetroCluster:
  - very low probability of link failure due to the hardwired ESCON link
- in a ContinentalClusters:
  - high probability of link failure over a WAN (at least temporarily)
- Environments where used:
  - in conjunction with synchronous mode to always guarantee full synchronization (data currency) of data between R1 and R2
  - required for M x N configuration where package data may reside on multiple Symmetrix frames to preserve data consistency
  - also useful in the wide-area environment when a remote BCV exists that can be split from the R2 copy in a consistent state

HEWLETT PACKARD

Domino mode causes new I/Os to be refused if:

- either the R1 or R2 copy of the data becomes unavailable due to hardware failure

- the SRDF link hardware fails

- the entire remote Symmetrix fails

When used in the wide-area environment, a monitor program on local and remote hosts should run periodically to detect when the SRDF link is down. When this state is detected on a remote host (R2), the BCV should be split and domino mode disabled so that the application may continue doing I/Os. When this state is detected on a local host (R1), the SRDF link should be split and the application may then be restarted.

XP256 Configuration Requirements for Support

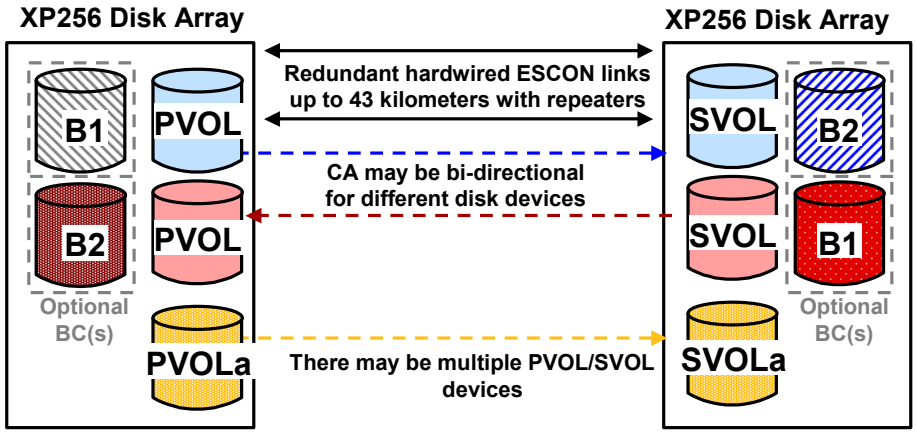
- PVOL devices are locally protected (RAID-1 or RAID-S)
- SVOL devices are locally protected (RAID-1 or RAID-S)

- Only synchronous mode is enabled during cluster operation
  - Secondary copy of the database is almost always current and consistent
  - No committed transactions should be lost upon failover

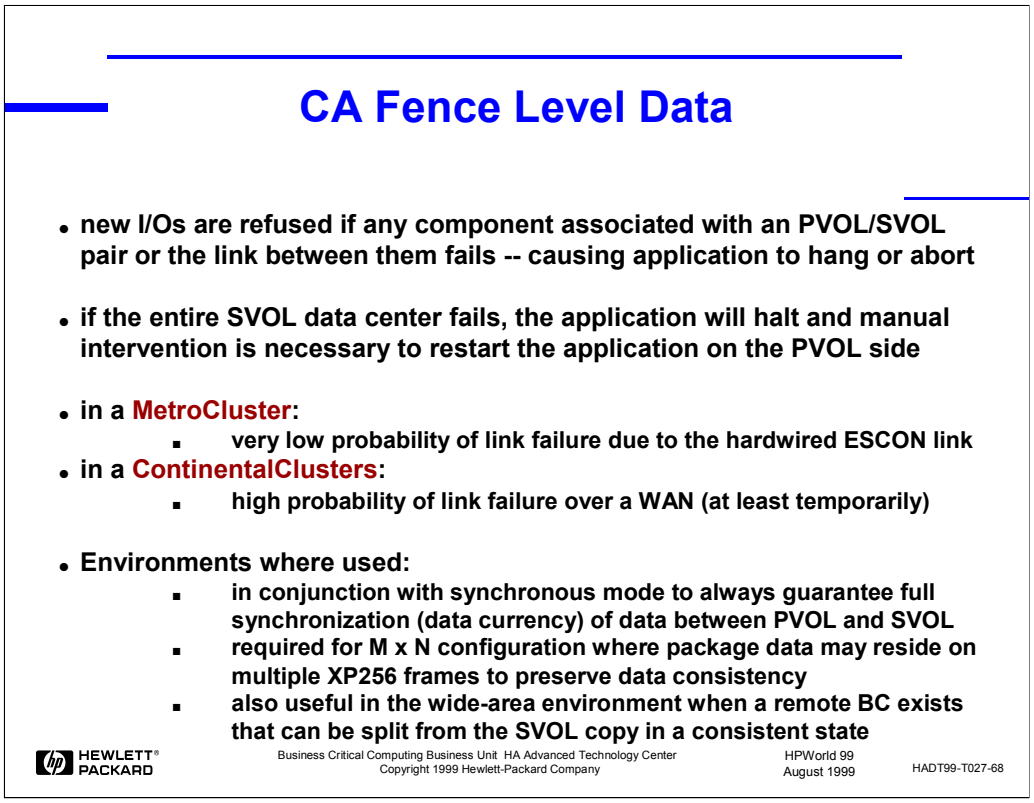
- Fence Level Data is recommended
- If Fence Level Data is not used, BCs are required to provide point-in-time consistent copies of the data

- No manual changes to the state of XP256 devices that belong to the MetroCluster packages during cluster operation
- All XP256 devices that belong to a particular MetroCluster package must be in the same state at the same time

XP256 Continuous Access Devices and BCs



- The PVOL device is the primary device and is always readable and writable
- The SVOL device is the secondary or remote device and is not writable while the link between it and the SVOL device is active
- Business Copies (BCs) are additional copies of the data that may be split and re-merged with the main copy of the data
  - requires additional disk devices
  - may be used in either the primary (local) XP256 or the secondary (remote) XP256
  - may be split from its primary copy to keep point-in-time consistent copies of the data



Fence Level Data causes new I/Os to be refused if:

- either the PVOL or SVOL copy of the data becomes unavailable due to hardware failure

- the CA link hardware fails

- the entire remote XP256 fails

When used in the wide-area environment, a monitor program on local and remote host should run periodically to detect when the CA link is down. When this state is detected on a remote host (SVOL), the BC should be split and fence level data disabled so that the application may continue doing I/Os. When this state is detected on a local host (PVOL), the CA link should be split and the application may then be restarted.

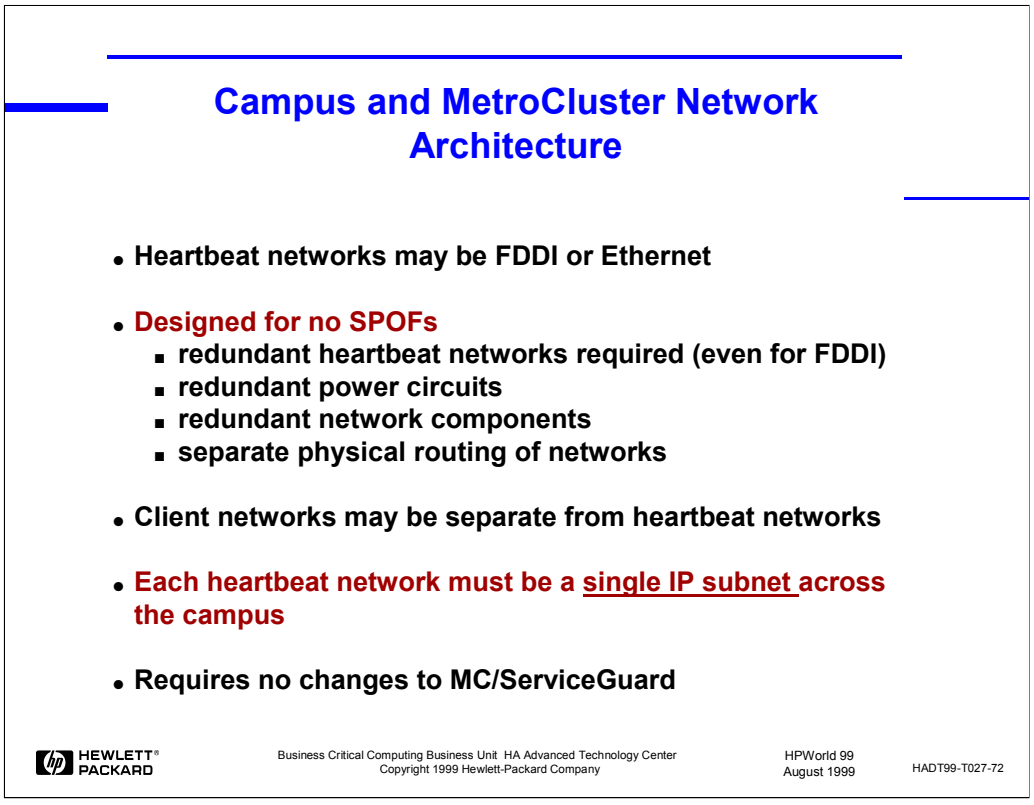
Oracle Standby Database Requirements

- **Physical data protection (RAID) in each data center**
- **Oracle Universal Server or Oracle Parallel Server installed on hosts in each data center**
- **Secondary data center has a copy of the database that is updated asynchronously by the Oracle Standby Database feature**

Oracle Standby Database

- Data is replicated logically and asynchronously using a log file shipping scheme
  - secondary database is almost always non-current, but consistent
  - some amount of data will be lost upon failover
- The transaction logs are transferred via the network and applied to a copy of the database that is running in recovery mode
- Failover involves
  - Waiting for any logs to be applied
  - Changing the database from recovery mode to online mode
- Failback involves
  - Shutting down the application
  - Performing a full backup of the database
  - Transferring the full backup to the primary site
  - Creating the database
  - Loading the database from the full backup

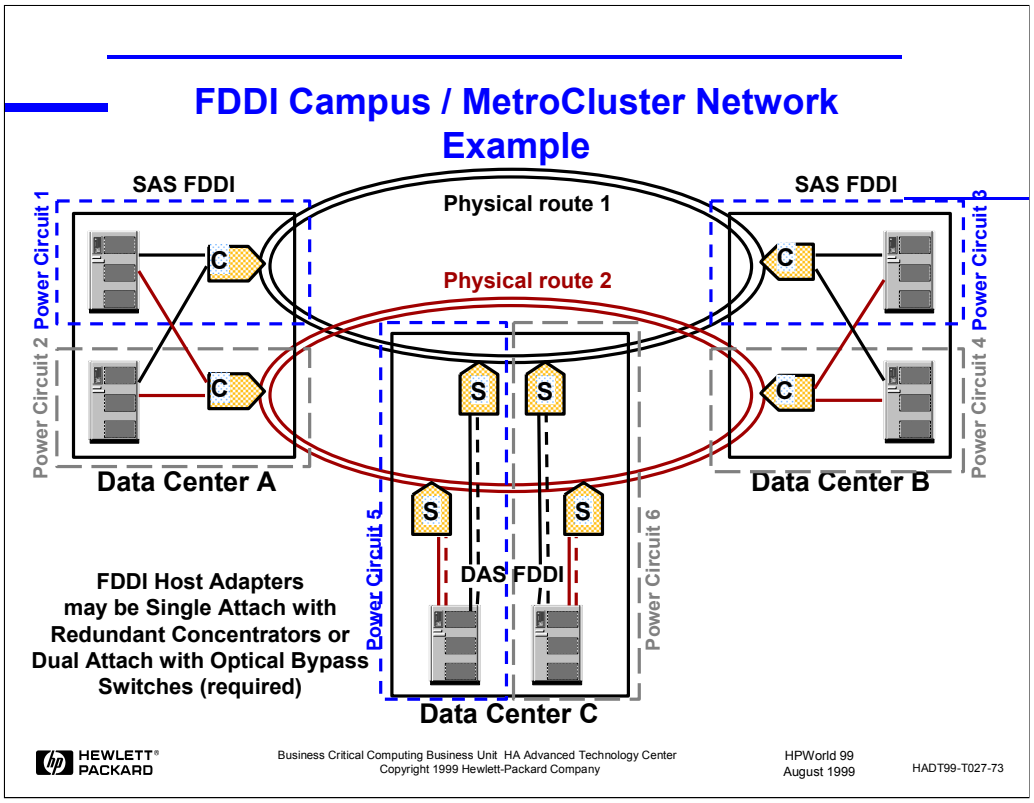
Network Examples



The network between the Data Centers is a critical component of the campus cluster. It is important that redundant network components be powered separately and that redundant cables follow different physical routes.

Each heartbeat network must be configured as a single IP subnet.

Client networks may be configured with redundant routers (discussed later).

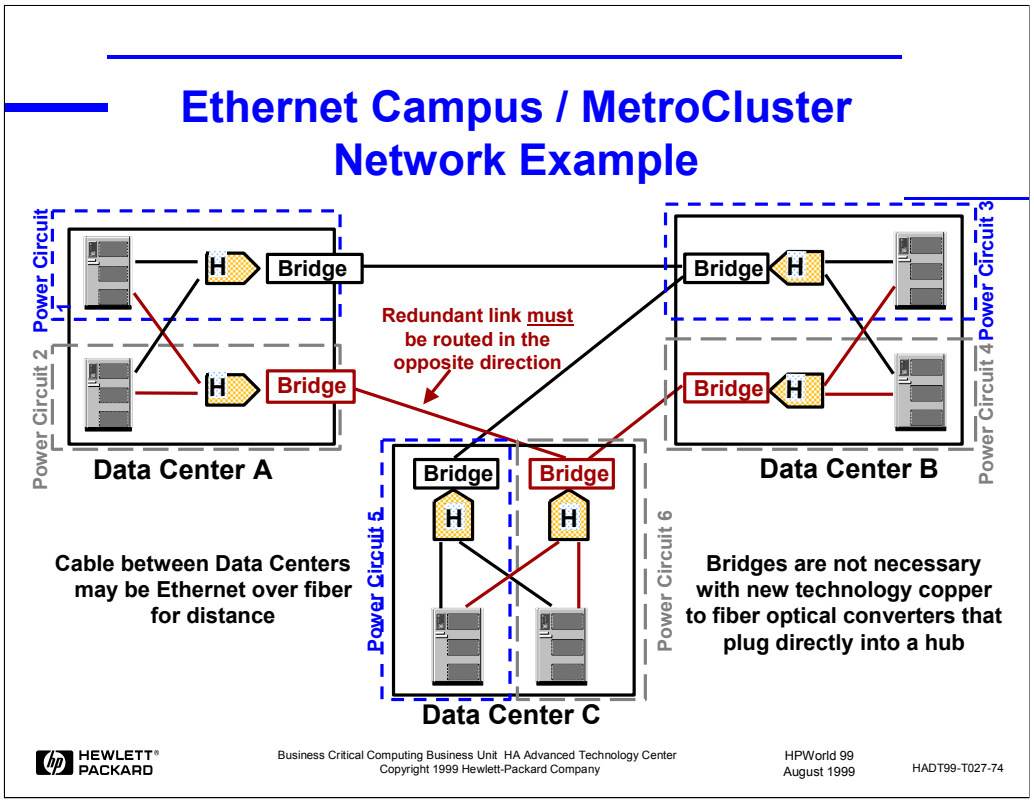


Two configurations are possible using FDDI networks.

One configuration uses two Single Attach Station (SAS) FDDI host adapters in each host. Each adapter is connected to a different FDDI Concentrator. The redundant concentrators are connected to completely different dual FDDI rings. The two rings must be routed in different physical paths.

The other configuration uses two Dual Attach Station (DAS) FDDI host adapters in each host. Each adapter is connected to a different FDDI bypass switch. The redundant switches are connected to completely different dual FDDI rings.

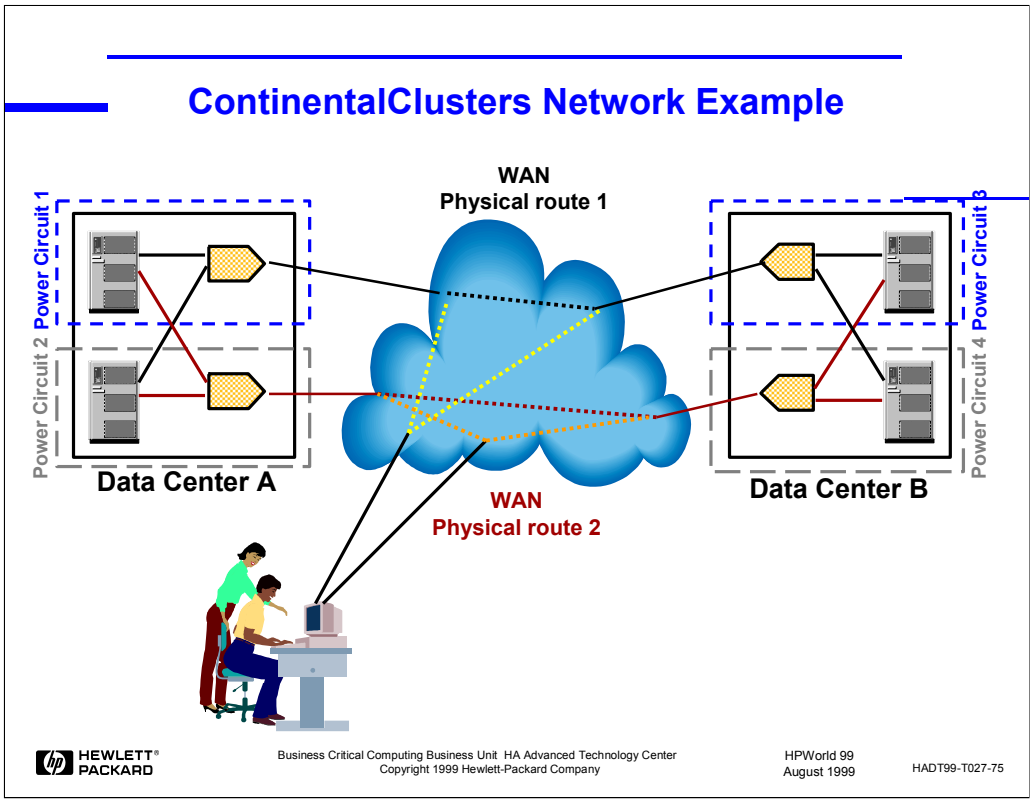
A combination of the two configurations is possible as shown in this slide.



The campus cluster may also be configured using Ethernet links. Hosts are connected to redundant Hubs and Bridges using two 10BaseT or 100BaseT host adapters.

Because Ethernet is a bus architecture rather than a ring, the redundant Ethernets must be routed in opposite directions to prevent Data Center failure from breaking both Ethernet networks.

Bridges or repeaters that convert from copper to fiber optic cable may be used to span longer distances (up to about 10-12 km).



Each data center has its own network and IP subnet. Normal MC/ServiceGuard rules apply for networks.

The network between data centers would typically be a Wide Area Network (WAN). Examples of WAN technologies would be:

- ATM
- T1 / T3 / E3
- SONET
- Satellite
- FDDI

WANs for large organizations might incorporate multiple link technologies.

Client reconnect is typically more difficult in this environment. Routers used by clients must have knowledge of two different IP subnets, each in a different location.

Questions