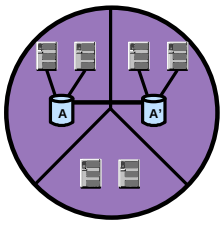
Disaster Tolerant Data Replication & HA Cluster Architectures

Bob Sauers

Hewlett Packard Company
InterWorks '00 Tutorial 023
March 2000



Welcome

Bob Sauers
Chief Architect, Mission Critical and Disaster Tolerant Solutions

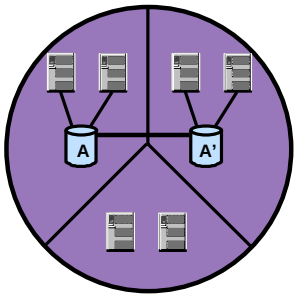
Business Critical Computing Business Unit (BCCBU)
Availability & Consolidation Solutions Lab (ACSL)
HA Advanced Technology Center

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

Agenda

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

Disaster Tolerance Concepts



Levels of Availability

- Basic availability (99.5% to 99.8%)
- High Availability (HA) (99.8% to 99.95% [< 4 hours])
- Continuous Availability (99.999%)
- Disaster Tolerance (99.8% to 99.95%)
  - increases availability
  - provides automated failover during certain disasters
- Disaster Recovery (DR)
  - no increase in availability, *per se*
  - days or weeks to recover

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 Tolerance

- Goes beyond HA, Continuous Availability or Disaster Recovery
- Additional computing hardware geographically distant from the main location
- Depends upon active data replication
- Speed of failover (minutes to hours)
- Automatic or Semi-Automatic failover
- Automation of application recovery
- Protects against all *single* points of failure (SPOFs) and many *multiple* points of failure
- Should be as transparent as possible for the client

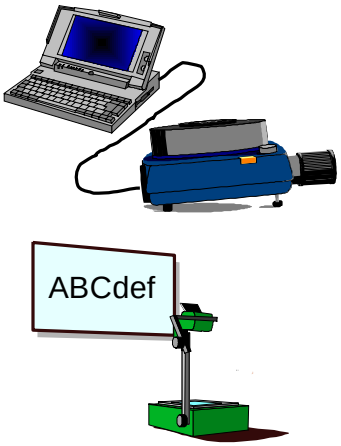
Disaster Tolerance

- Guards against
  - fire
  - flood
  - power outage
  - sabotage
  - certain user error
  - terrorism
  - earthquake

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

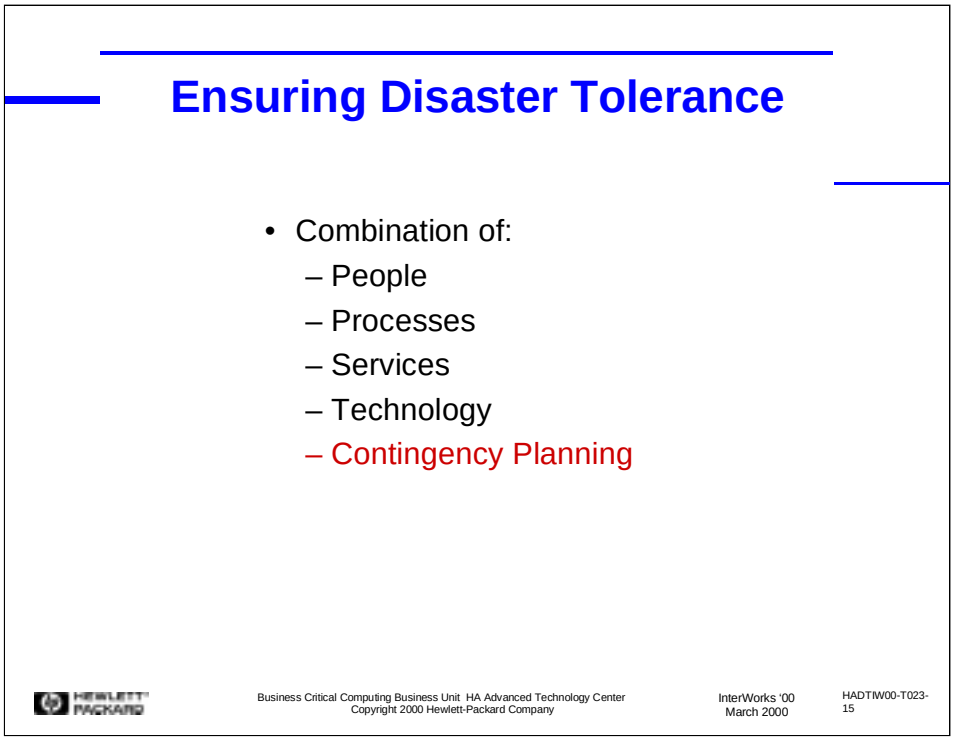
Disasters include:

- 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

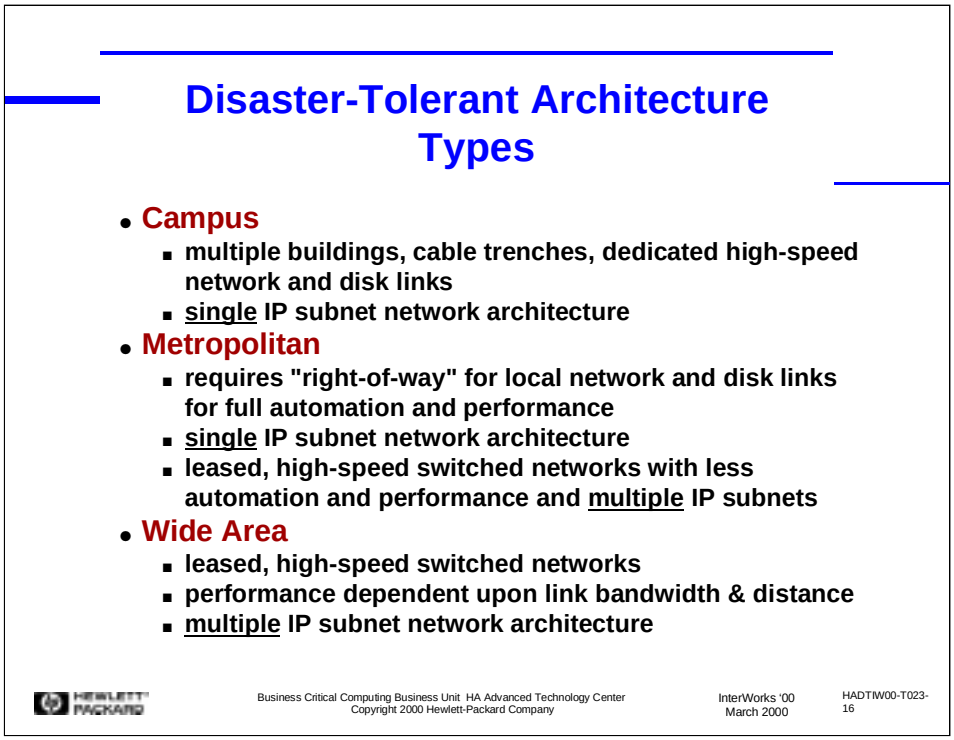
- Cost of additional hardware, staff and facilities versus cost of downtime to the organization
  - loss of revenue
  - loss of productivity
  - loss of customers
  - loss of reputation
  - being driven out of business

- 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



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

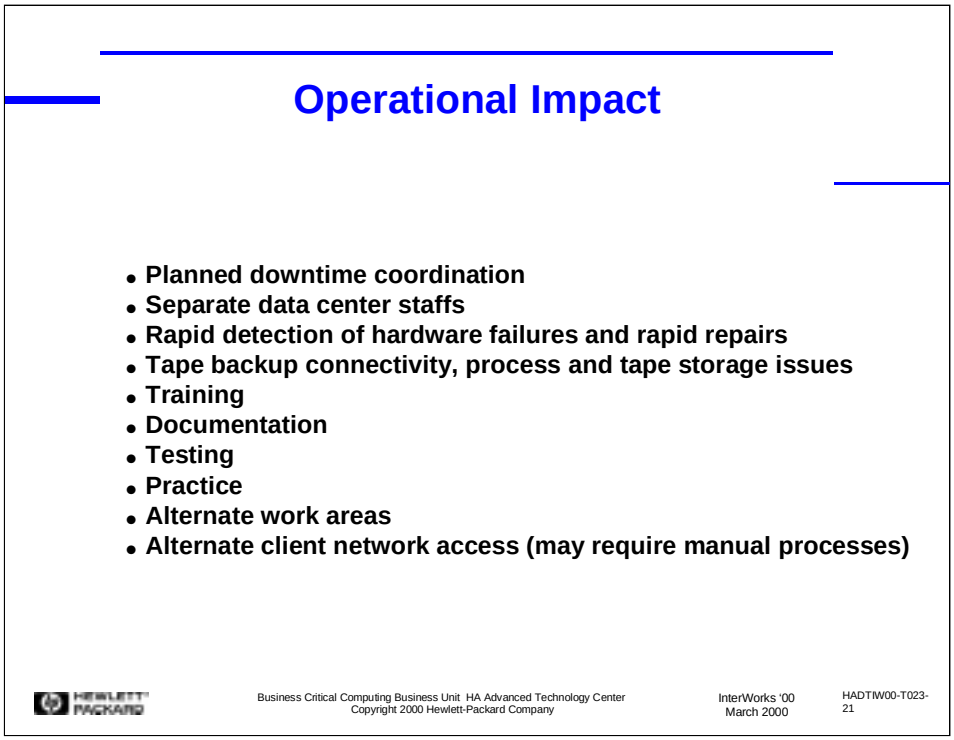
- **Local failover is better for most failures**
  - **Faster in some cases**
  - **May involve less recovery**
  - **Fewer chances of problems**
  - **May be more transparent to the clients**
- **Disaster Recovery failure should be used only in case of entire site failures**
  - **Data may be unprotected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

Choosing the Right Solution

- Criticality of the application
- Business needs
- Budget
- Cost of downtime
- Vulnerability (risk) analysis
- Cost / Benefit analysis

Criticality of the application

- Usually defined by the business unit
- How detrimental loss of application availability is to the business unit
- Service Level Agreement
  - **criticality**
  - **acceptable planned downtime**
  - **acceptable unplanned downtime**
  - **hours of availability**
  - **number of users**
  - **acceptable response time**
  - **etc.**



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

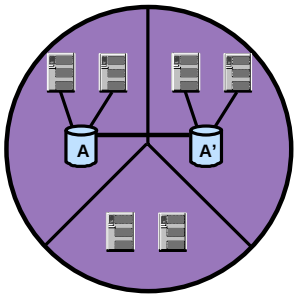
- hardware failures must be detected rapidly and repairs made quickly so that the cluster is not vulnerable to additional failures

- 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

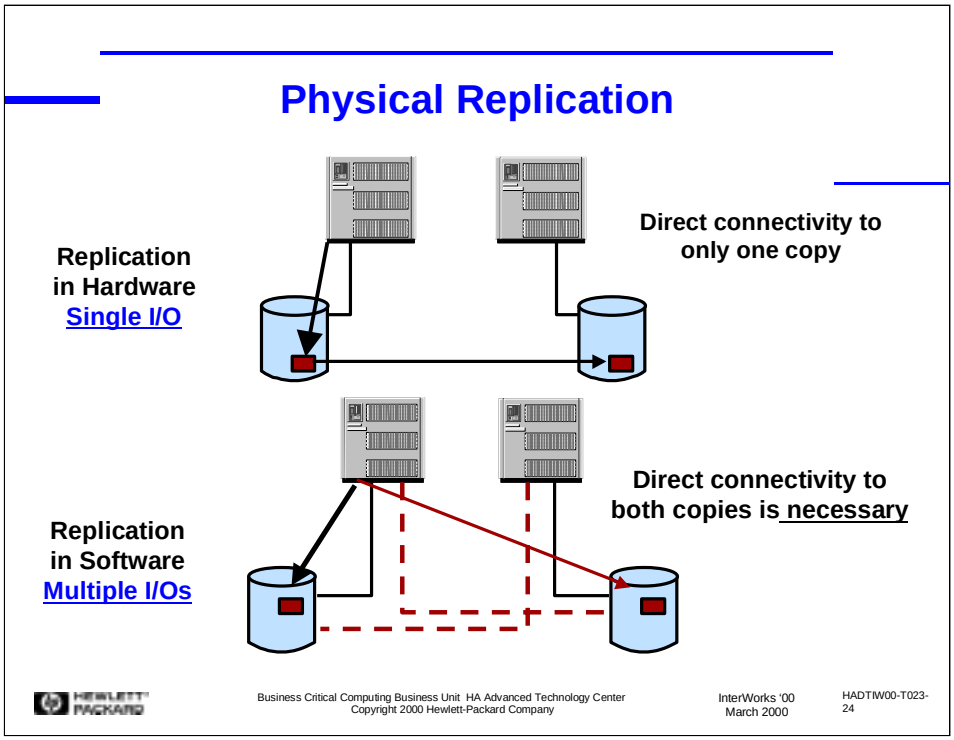
- disaster practice is important to any disaster recovery environment

Data Replication



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**



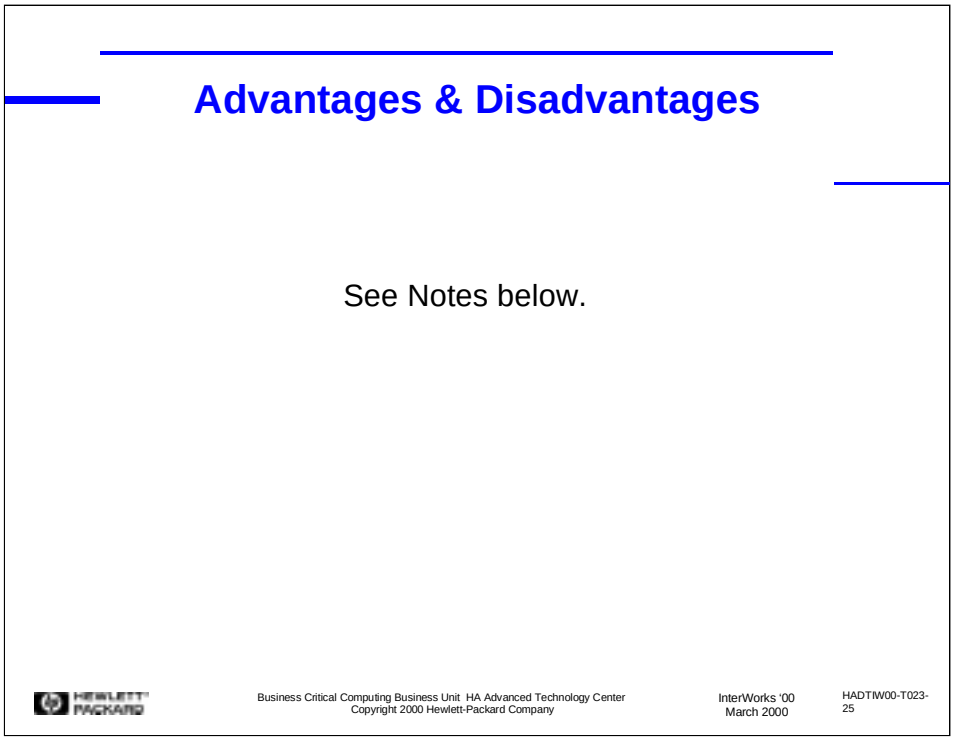
Examples of physical replication in hardware is EMC Symmetrix Remote Data Facility (SRDF) and HP XP256 Continuous Access.

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**



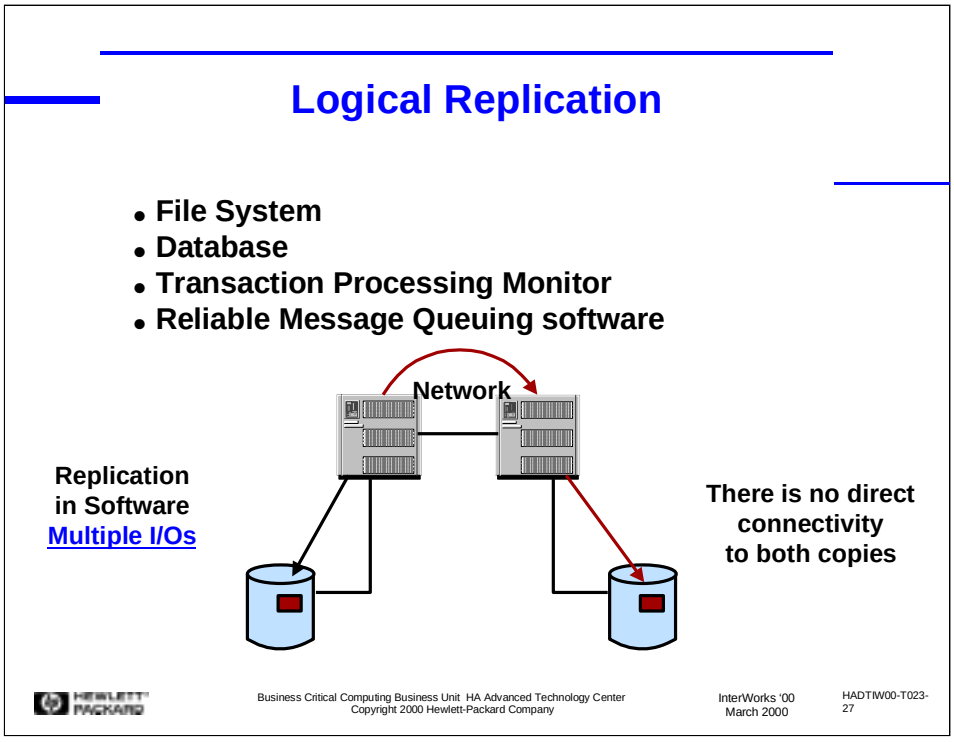
Advantages & Disadvantages of Physical Replication in Software

+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

Physical Replication in Hardware

- Physical replication performed by the disk array
  - **offloads the host from data replication task**
  - **synchronous, asynchronous or non-synchronous**
  - **dedicated fiber links up to 60 km**
  - **converters used to extend to continental & intercontinental distances**
- Synchronous mode keeps copies of data current when link is operational, but may degrade application performance
- Rolling disaster can cause data corruption -- HP has solutions to avoid data corruption
- Non-synchronous modes are deemed unacceptable
- HP SureStore E Disk Array XP256 with Continuous Access XP
- EMC Symmetrix with Symmetrix Remote Data Facility



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**

Logical Replication

- Logical replication in software are usually offered by database vendors
- Copies either the transaction itself or the transaction log
- Consumes additional total CPU since transaction is applied twice
- Preserves the order of committed transactions
- Maintains logically equivalent databases (not necessarily physically equivalent)
- Replica usually lags the primary database, and is not current
- Oracle Standby Database has been integrated with MC/ServiceGuard and ContinentalClusters

Transaction Processing Monitors

- Alternative to logical data replication products
- Requires application modification
- Can increase currency and transparency to user if TPM maintains copy of transaction on another system until it has completed at both sites
- Synchronous or asynchronous methods
- Transaction queuing when system is down

Reliable Message Queuing Software

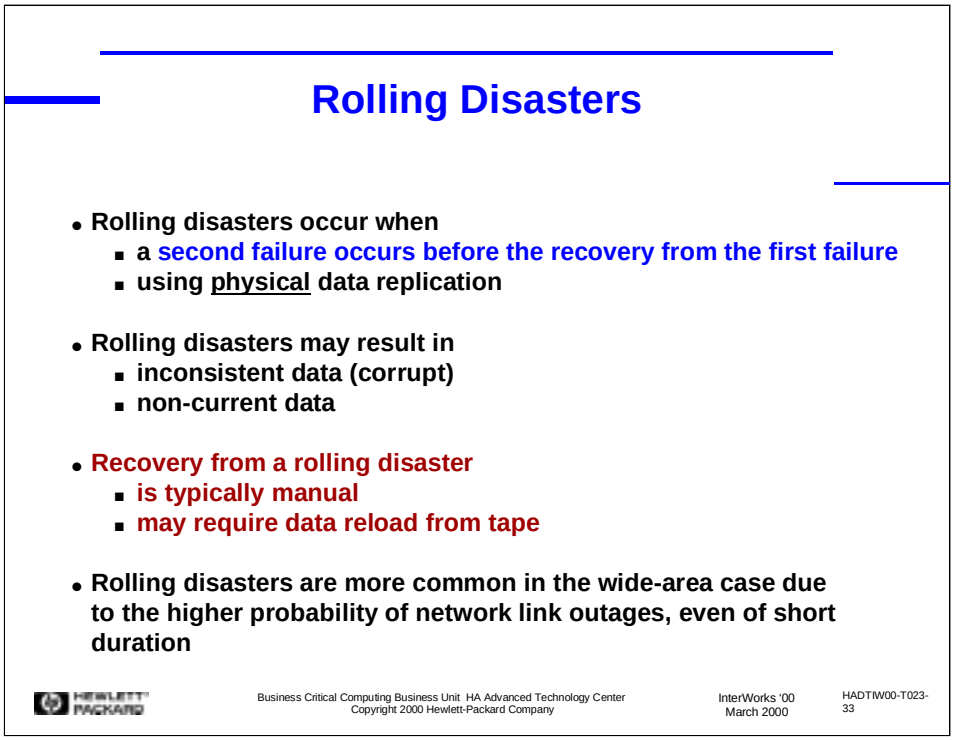
- Alternative to logical data replication products
- Does not usually require application modification
- Can increase currency and transparency to user if MQ software maintains copy of transaction on another system until it has completed at both sites
- Synchronous or asynchronous methods
- Message queues when receiving system is down

Two Phase Commit

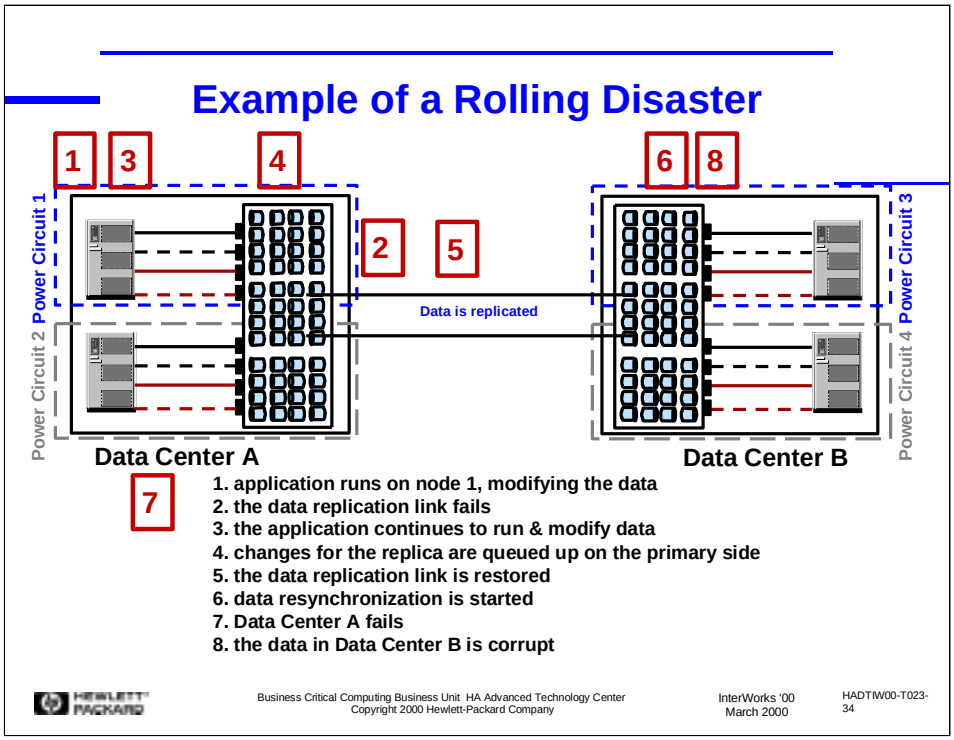
- In open systems environment, applications can be written to use two phase commit semantics
- Ensures all copies of the database are exactly current at the cost of aborting a transaction if one copy is unavailable
- Seldom used due to
  - performance degradation
  - loss of data availability causes transactions to fail

Ideal Situation

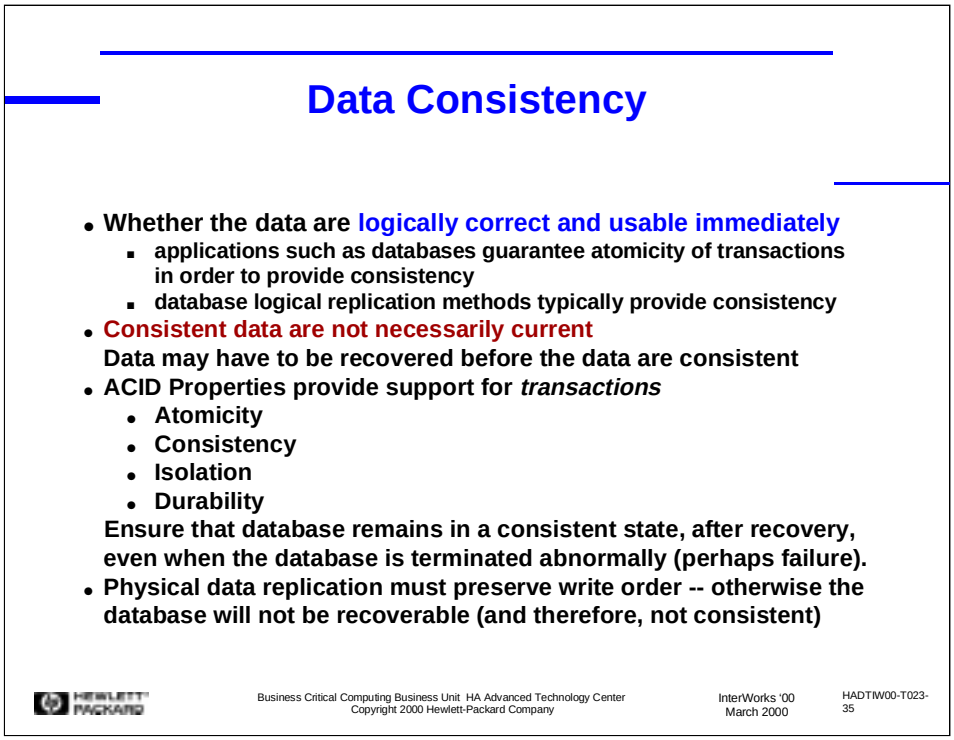
- replicate physically for speed and currency
- replicate logically for consistency and to recover from human error
- use the replica only when all other physical copies are corrupt



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:

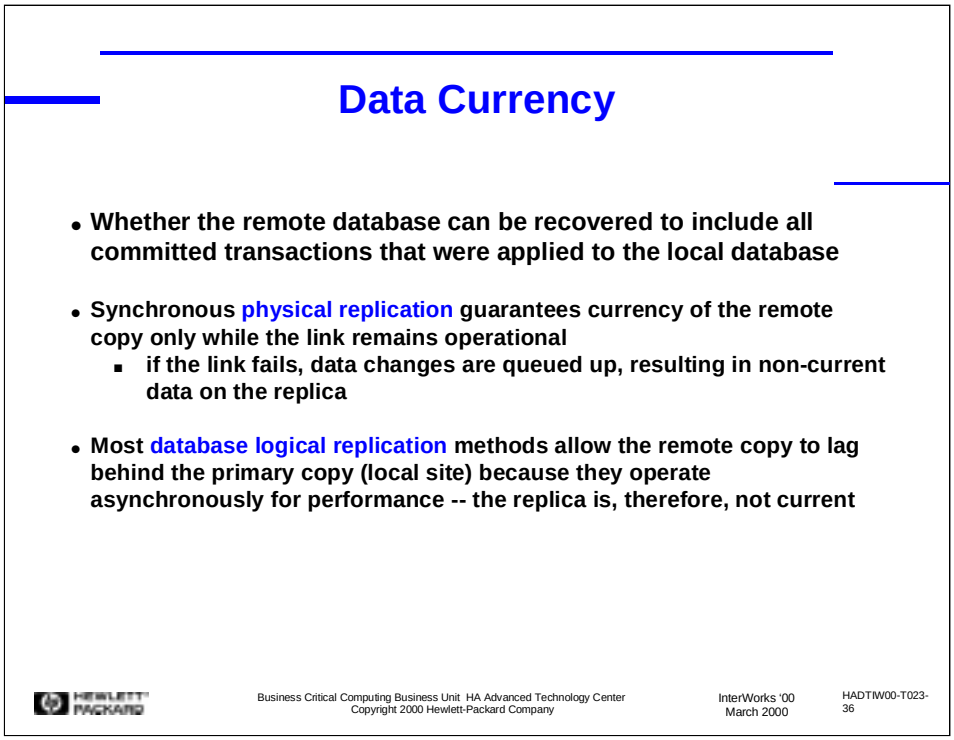
- a point-in-time split-off copy at the remote site

- 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**

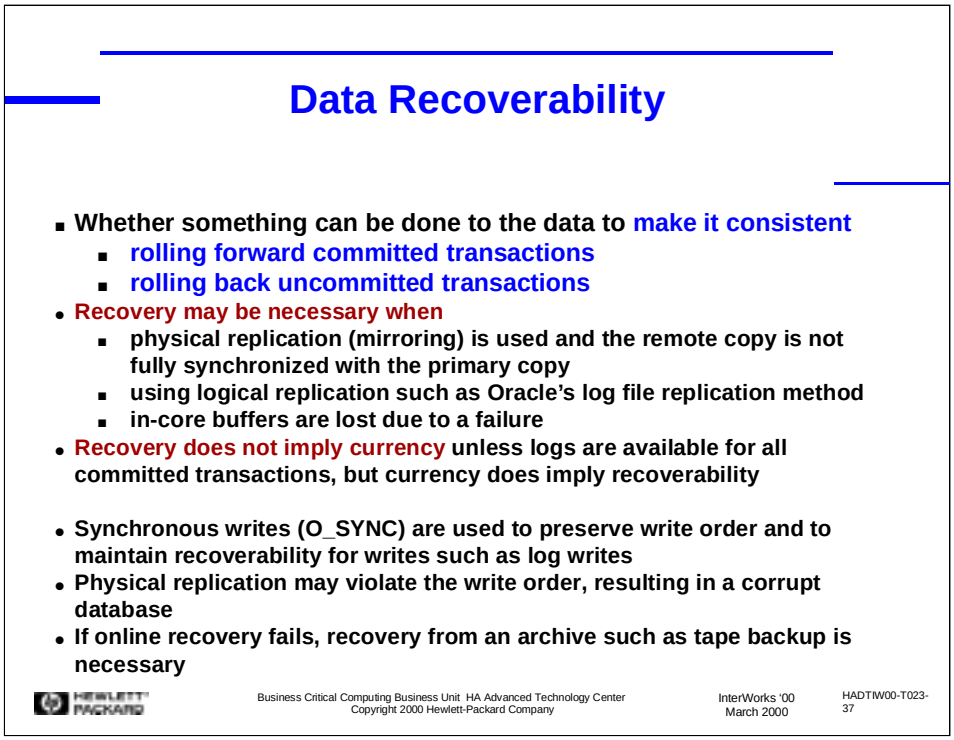
- **For a database, the data are consistent if only committed transactions have been applied (all uncommitted transactions rolled back)**



- 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

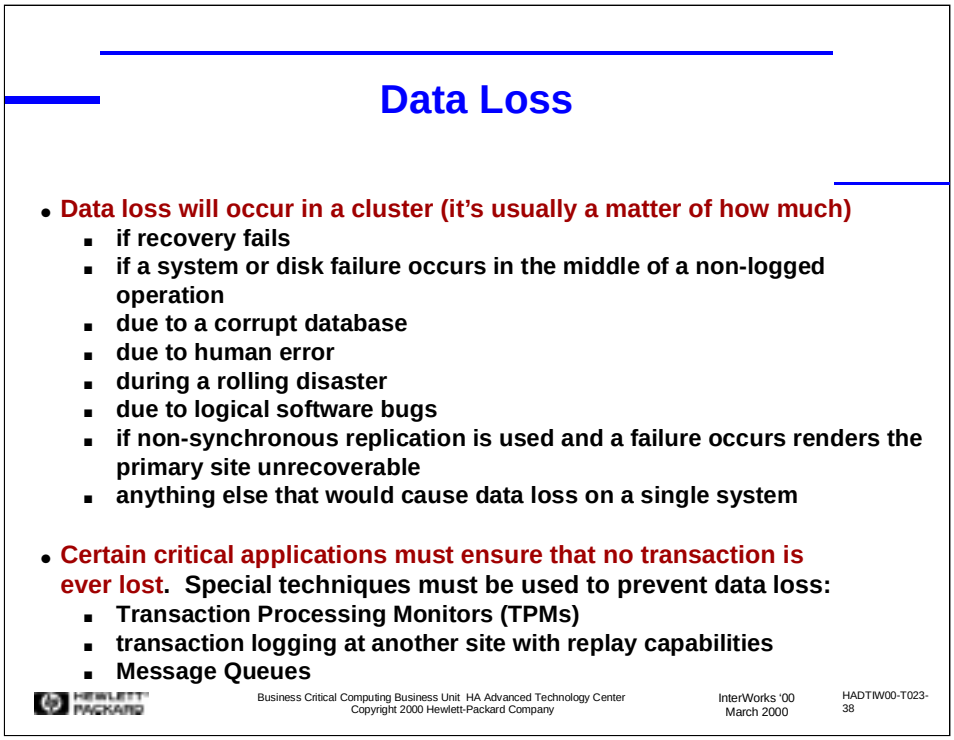
- Any transactions (data) in a computers memory that is not yet written to disk are lost upon system failure



In-core memory buffers may be lost due to

- memory fault
- power loss
- system crash

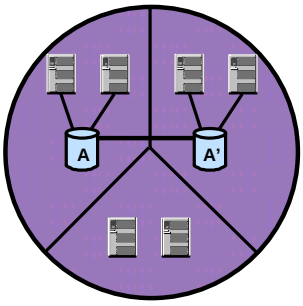
- For a database, the data may be recoverable, for example, by either or both
  - reload of the database from a point-in-time copy (tape or disk)
  - roll forward of committed and roll back of uncommitted transactions
- For a file system, the metadata are recoverable for
  - an HFS with fs_async=0
 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



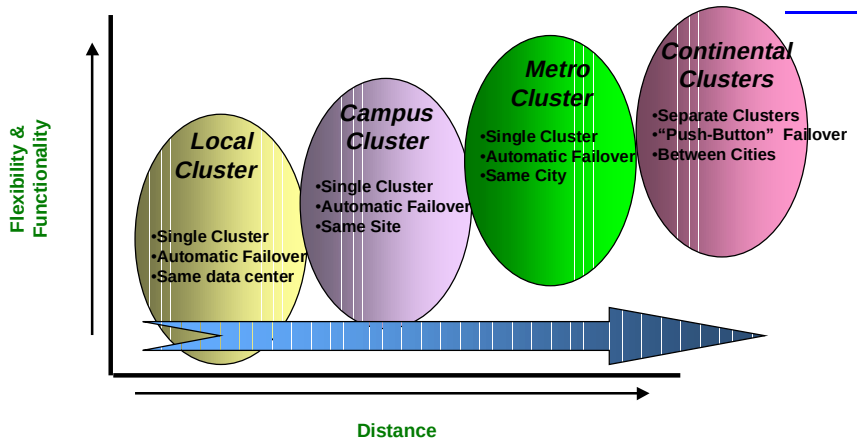
Even the airlines are willing to suffer a certain amount of data loss with their reservation systems, such as a lost seat assignment.

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

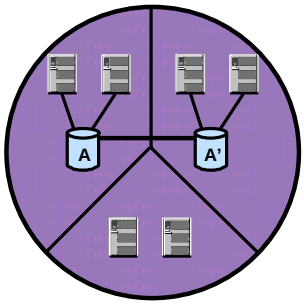
Disaster-Tolerant Cluster Architectures

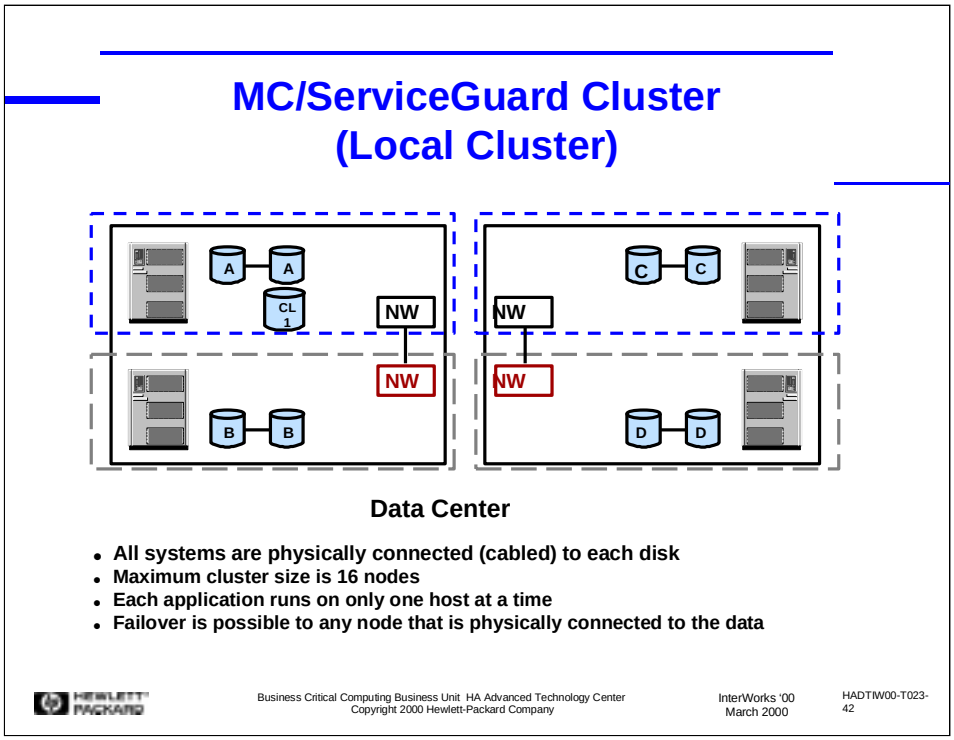


Range of Architectures



Local Cluster MC/ServiceGuard





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)

MC/ServiceGuard

- Protects against failures of
  - hosts
  - networks
  - applications (services)
  - user-defined resources
  - OS resource problems (e.g., shared memory)
- Provides
  - transparent IP address failover
  - rolling upgrade for OS and some applications
  - shared connectivity to single copy of data
  - integration with Oracle 8i Server

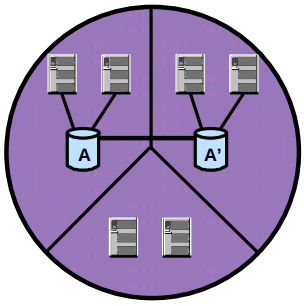
MC/ServiceGuard

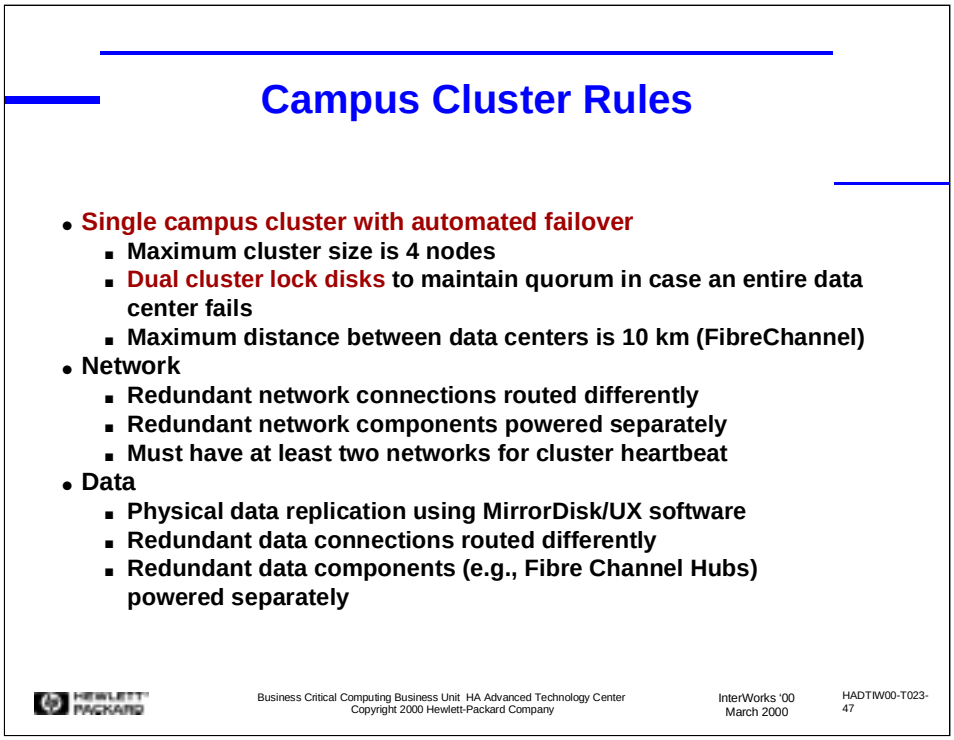
- Does not protect against
  - failures caused by certain types of human error
  - database corruption
  - application bugs
  - natural and human caused disasters

MC/ServiceGuard Comparative Features

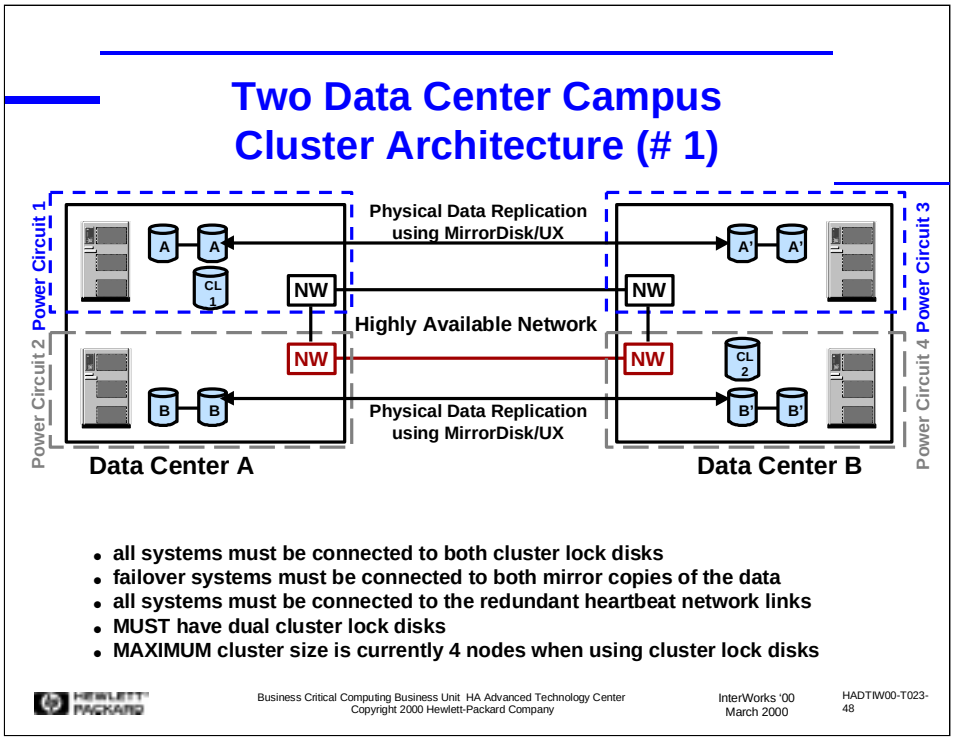
Cluster Topology	**Single Cluster up to 16 nodes**
Geography	**Data Center or Campus**
Network Subnets	**Single IP Subnet**
Network Types	**Dedicated Ethernet, FDDI or Token Ring**
Cluster Lock Disk	**Required for 2 nodes, optional for 3-4 nodes, not used with larger clusters**
Failover Type	**Automatic**
Failover Direction	**Omni-directional**
Data Replication	**NONE**

Campus Clusters





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

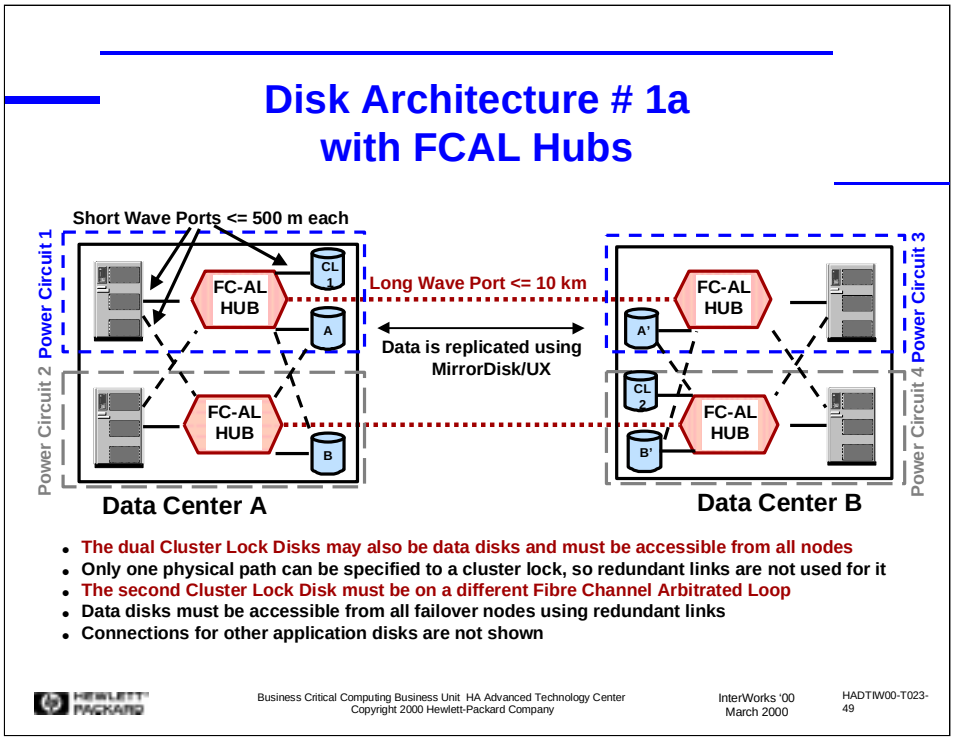


Implementations:

- **EMC Symmetrix with 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**

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**

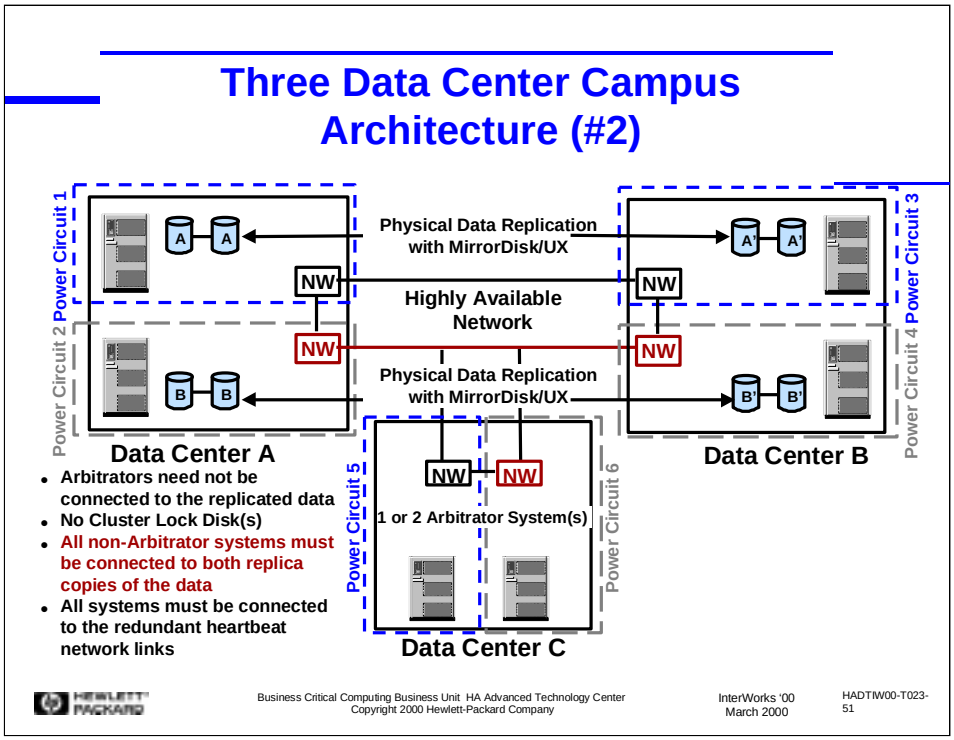
AND

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

AND

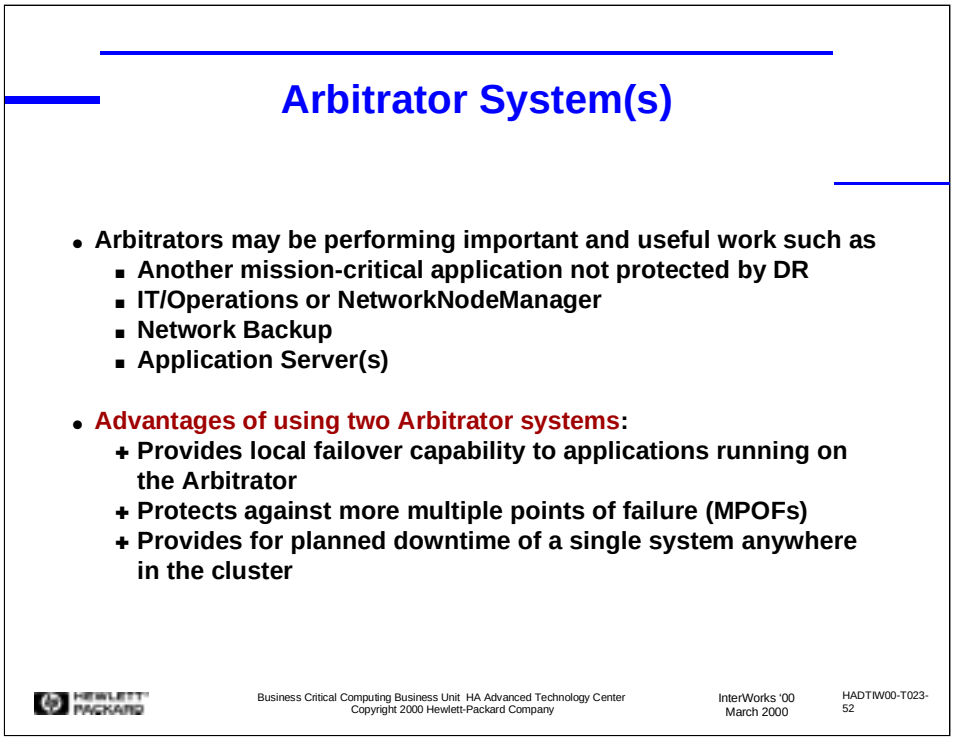
- **the disk link from Data Center B to Cluster Lock # 1 fails**

The result is data corruption!



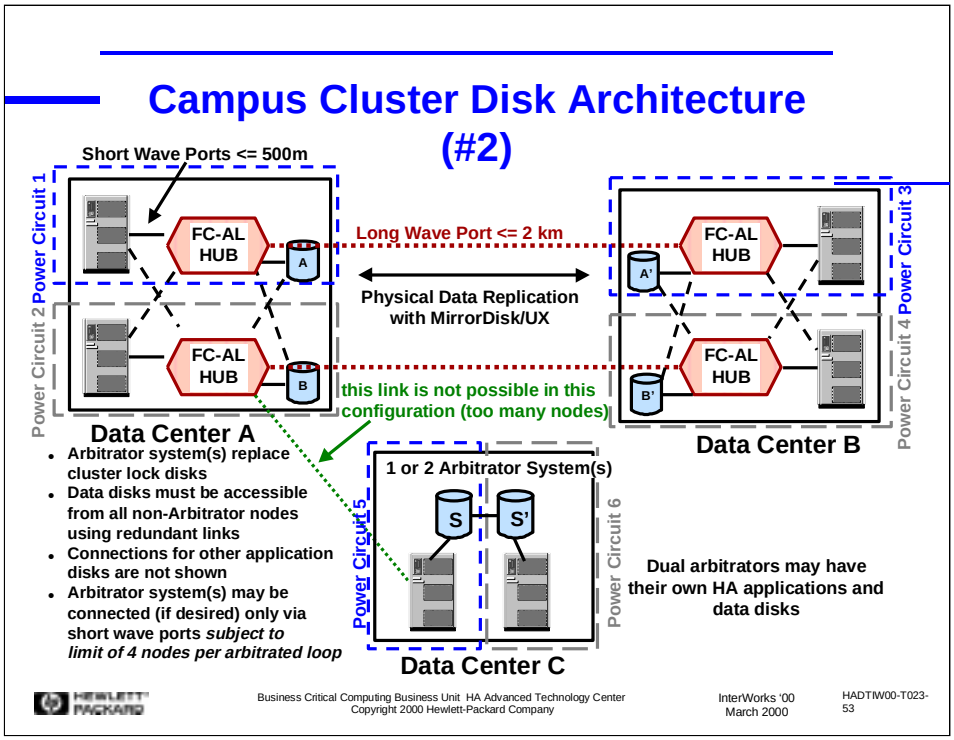
Implementations:

- **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

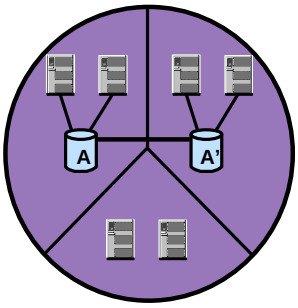
–maximum 10 km between data centers

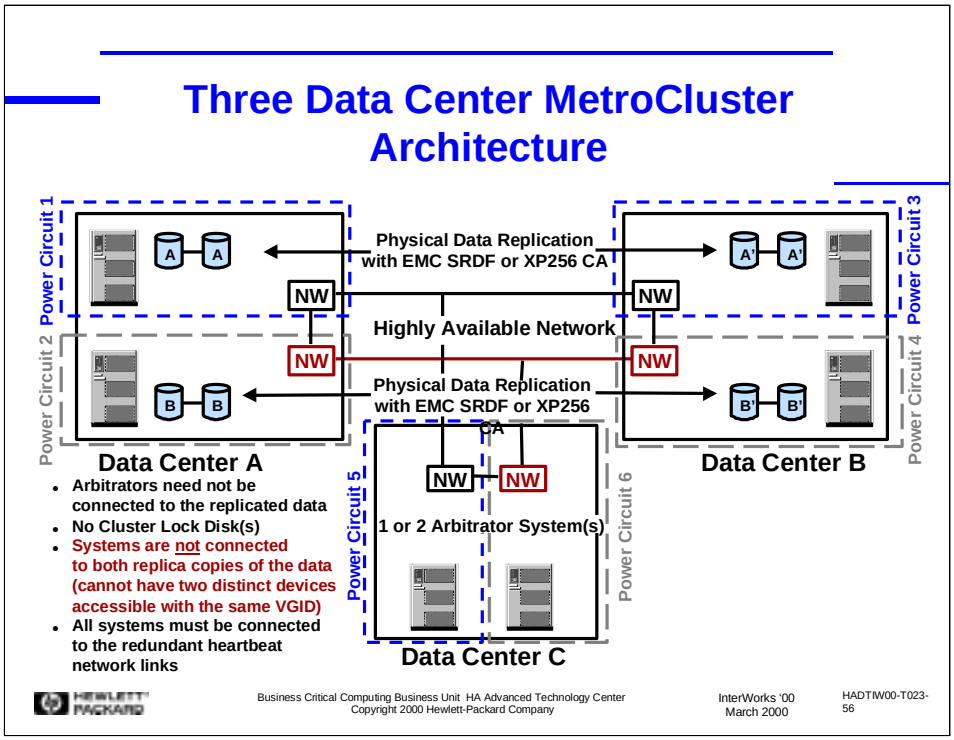
–increased CPU overhead (for mirroring)

Campus Cluster Comparative Features

Cluster Topology	**Single Cluster up to 4 nodes across 2 data centers or up to 16 nodes across 3 data centers**
Geography	**Campus, up to 10 km (Fibre Channel limitations)**
Network Subnets	**Single IP Subnet**
Network Types	**Dedicated Ethernet, FDDI or Token Ring**
Cluster Lock Disk	**Required for 2 nodes, optional for 3-4 nodes, not used with larger clusters**
Failover Type	**Automatic**
Failover Direction	**Bi-directional**
Data Replication	**MirrorDisk/UX**

MetroCluster





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
- Links planned for support in a future release
  - FCAL Point-to-Point
  - FCAL with Hubs

MetroCluster

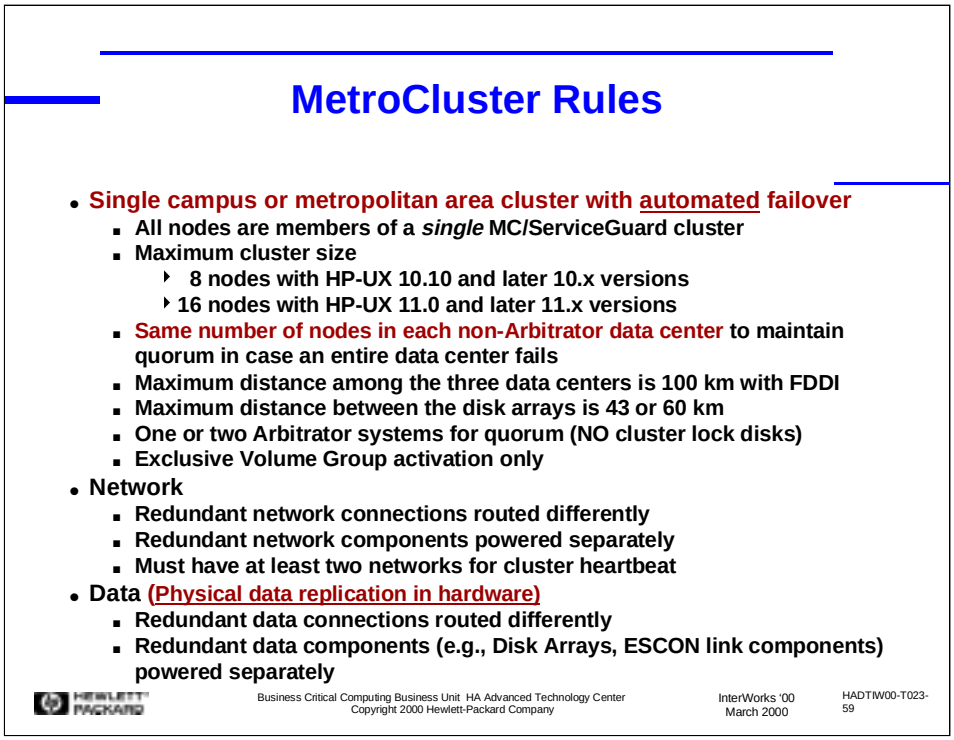
- Extends protection of MC/ServiceGuard cluster to cover certain disasters than affect a data center
- Uses physical data replication in hardware
- Automated local and remote failover
- MetroCluster works with most MC/ServiceGuard applications
- Does not protect against
  - Human error
  - Database corruption
  - Some application bugs
  - Natural disasters that affect an entire metro area

The MetroCluster Products: A Short Look

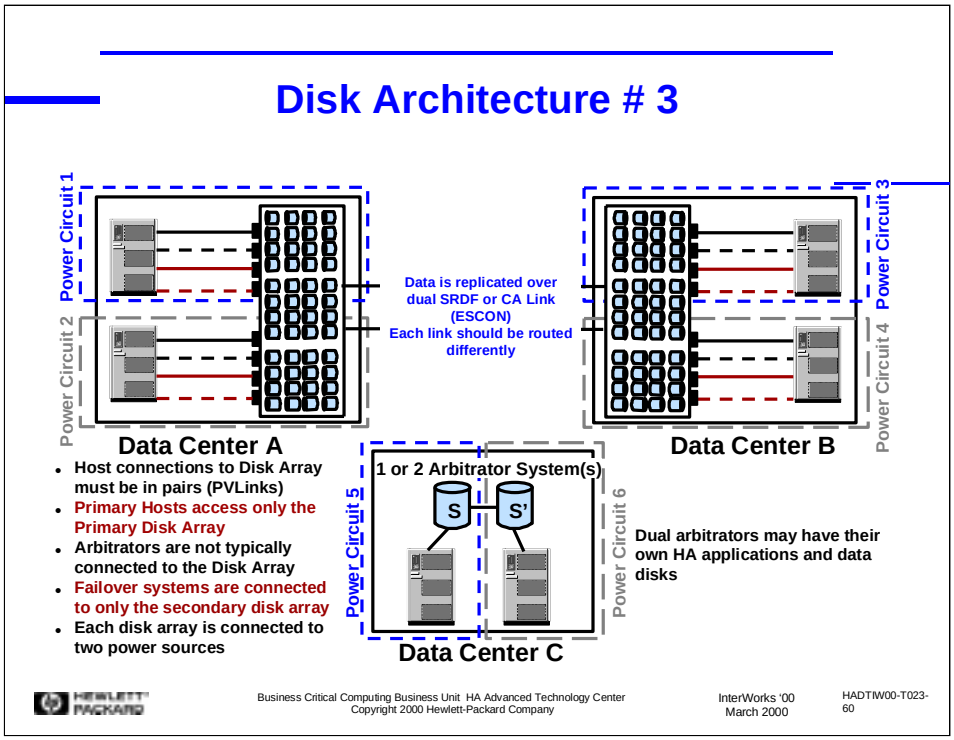
- **HP MetroCluster with EMC SRDF (B6264BA)**
  - **A product to automate the failover of MC/ServiceGuard packages among nodes using two Symmetrix disks that are connected by SRDF**

- **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**



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



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

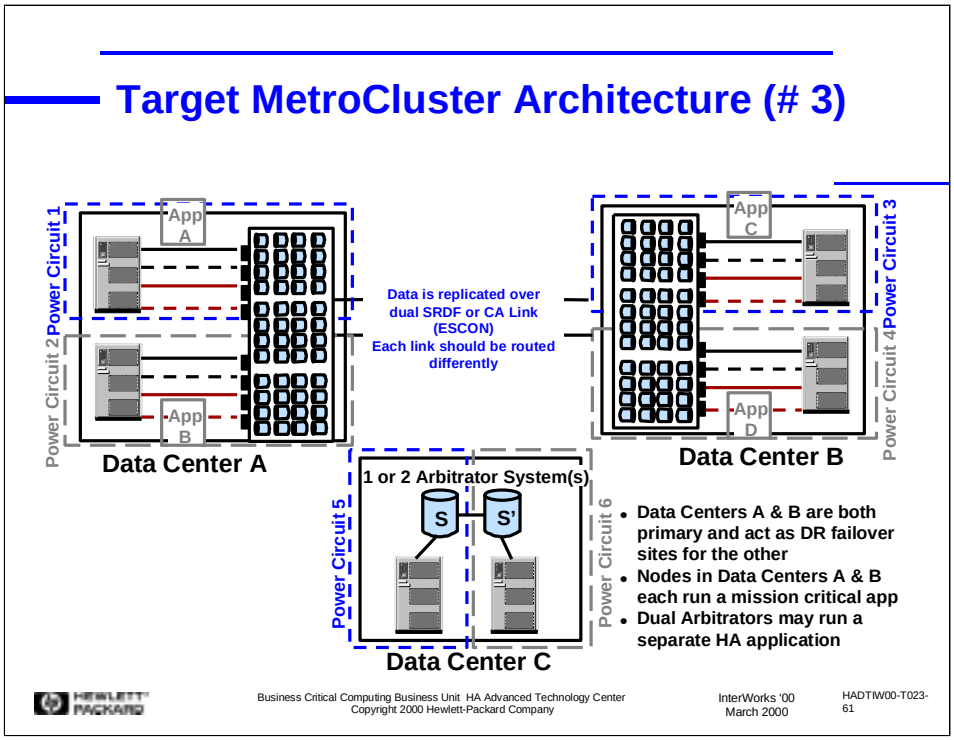
+bi-directional replication is possible

–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**

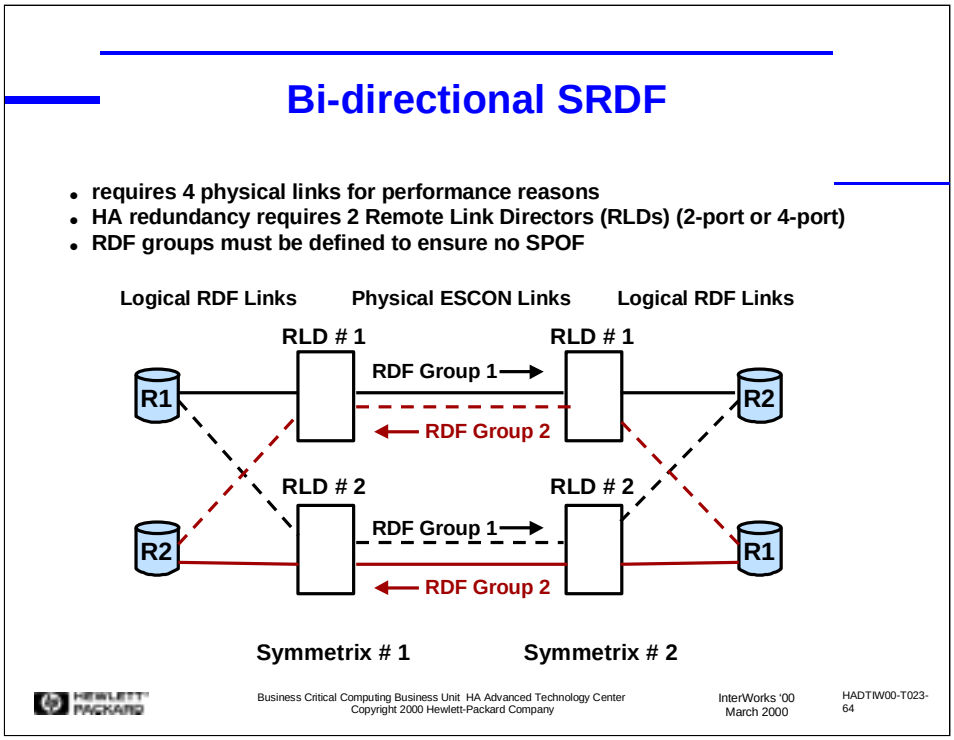
The same number of nodes must be present in Data Centers A & B

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 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



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

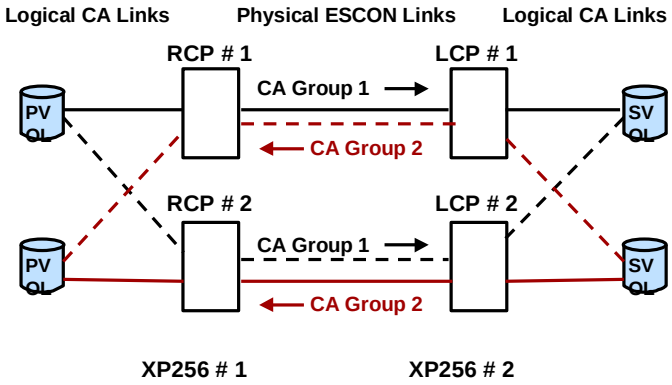
For performance reasons, the R1s on one side should use one RDF group that is assigned to one pair of links while the R1s on the other side use a different RDF group assigned to a different pair of links

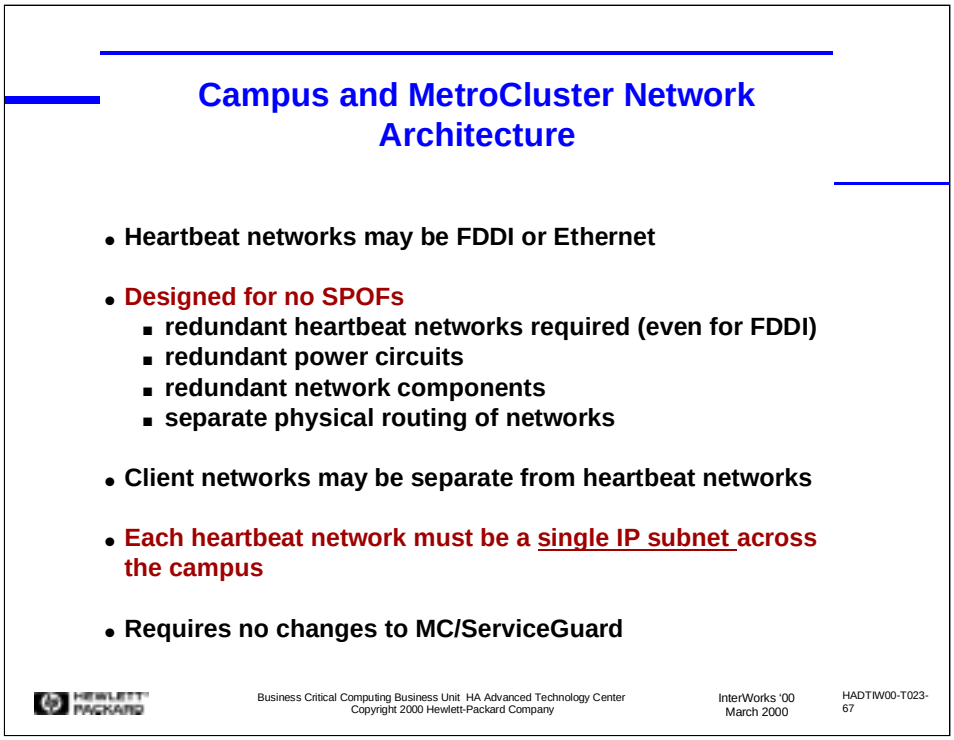
XP256 Continuous Access Failback

- 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

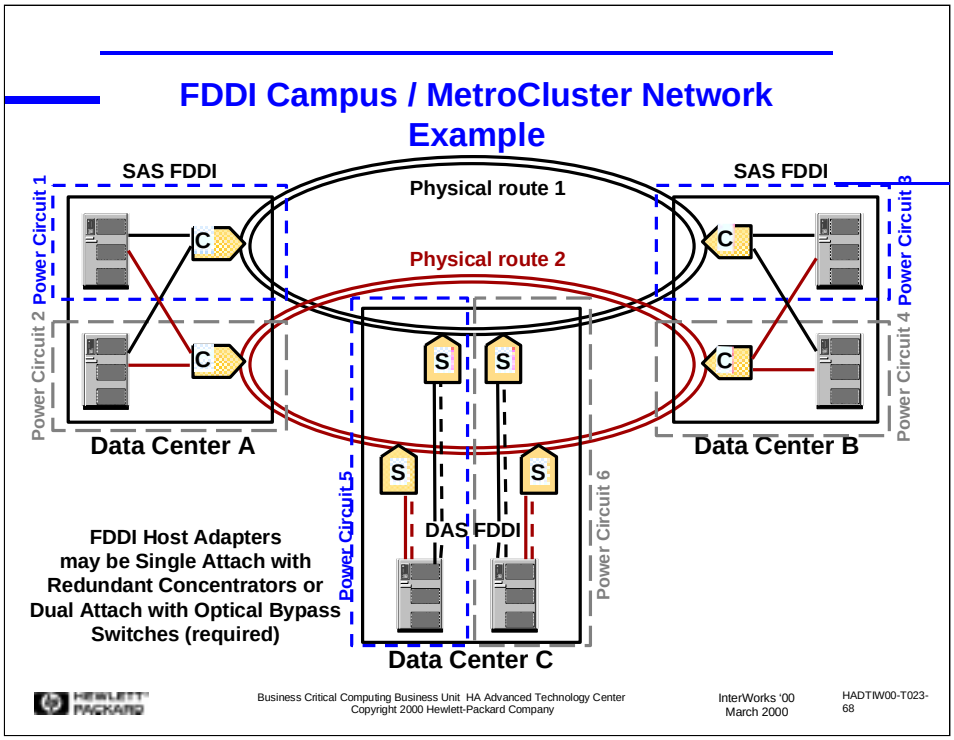




The network between the Data Centers is a critical component of the campus cluster. It is important that redundant network components be powered separately and that redundant cables follow different physical routes.

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

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

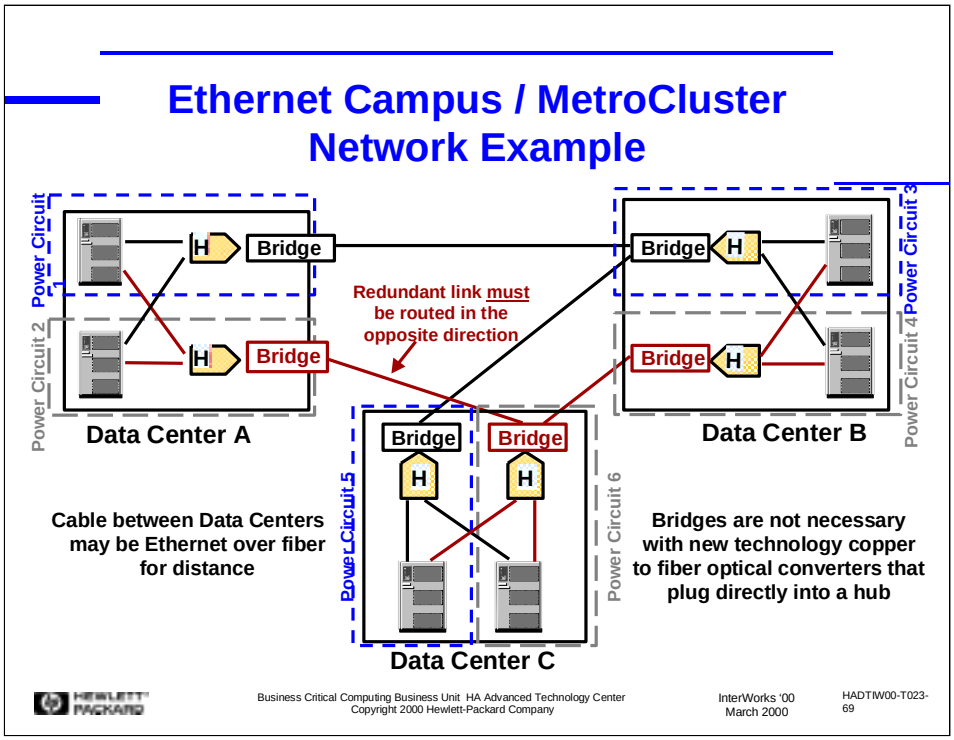


Two configurations are possible using FDDI networks.

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.

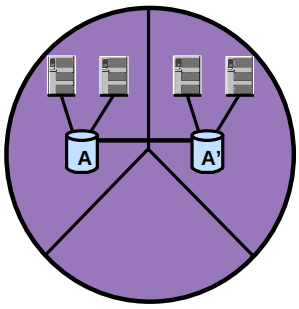
Because Ethernet is a bus architecture rather than a ring, the redundant Ethernets must be routed in opposite directions to prevent Data Center failure from breaking both Ethernet networks.

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

MetroCluster Comparative Features

Cluster Topology	**Single Cluster up to 16 nodes spread across 3 data centers**
Geography	**Campus or Metropolitan area**
Network Subnets	**Single IP Subnet**
Network Types	**Dedicated Ethernet, or FDDI**
Cluster Lock Disk	**Not Used; 1-2 Arbitrators in third data center act as tie breaker**
Failover Type	**Automatic**
Failover Direction	**Bi-directional**
Data Replication	**Physical, in hardware (XP256 CA or EMC SRDF)**

ContinentalClusters

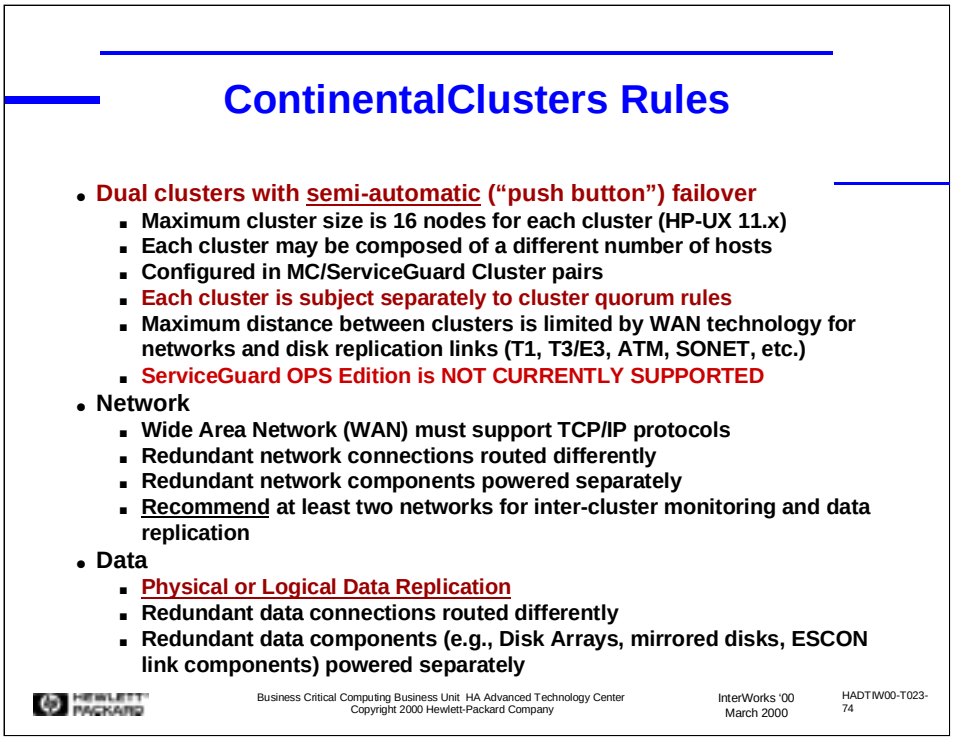


ContinentalClusters : A Short Look

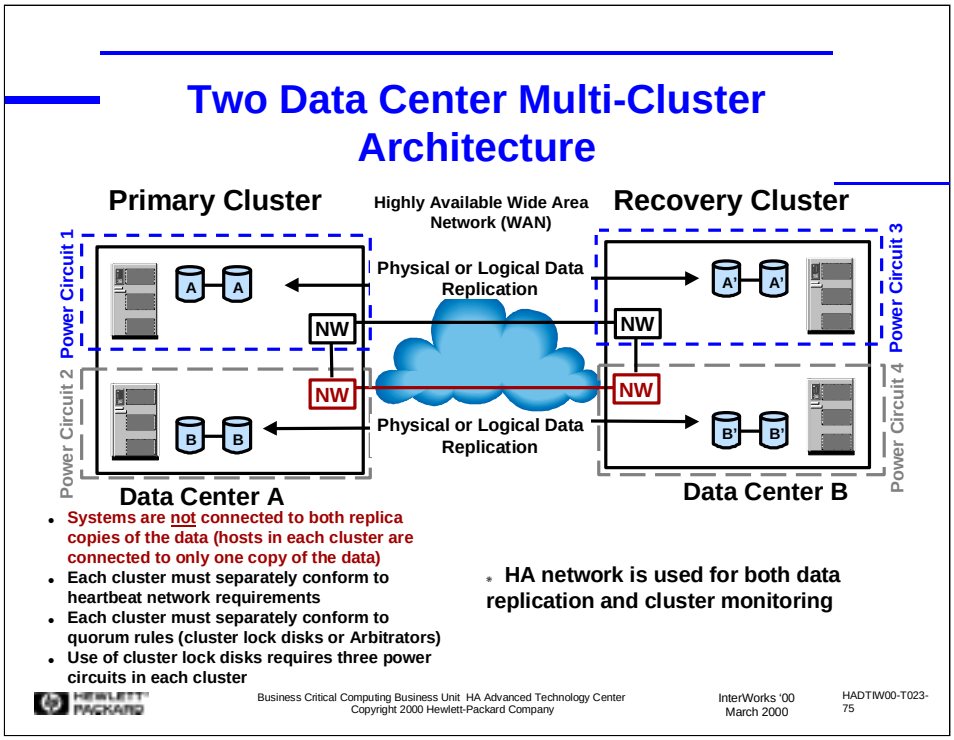
- **HP ContinentalClusters (B7659BA)**
  - **A product to automate the failover of MC/ServiceGuard packages among TWO separate clusters**
    - **primary and secondary cluster failure notification is configurable**
    - **semi-automatic "push button" initiates automated failover**
    - **choice of various logical or physical data replication methods**
    - **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)**



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



Physical and Logical Data Replication Solutions:

- 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

ContinentalClusters

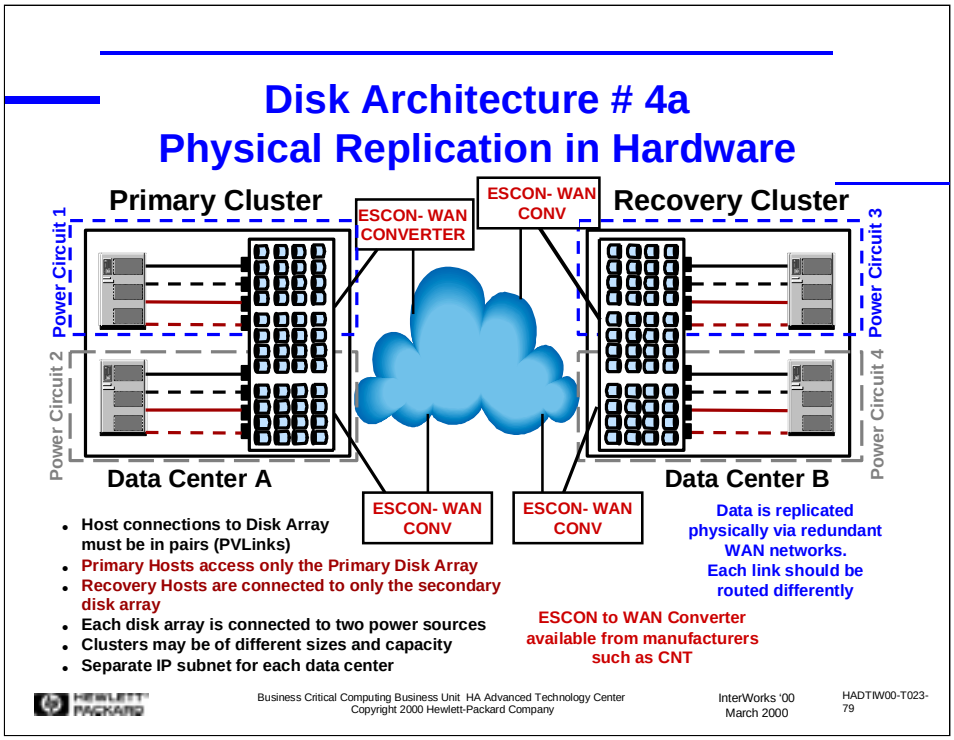
- Automates the failover of applications between two clusters that reside in separate data centers
- Human decision is necessary to initiate the failover
- Cluster problem notification
  - text files
  - system console
  - e-mail
  - SNMP trap
  - opcmsg (IT/Operations)
- Local failover still occurs within the primary data center
- Remote failover is used only when the entire primary cluster fails

ContinentalClusters Failover

- **Semi-automatic failover (based upon notification rules) of ONE pre-configured set of packages from the Primary CLUSTER to the Recovery CLUSTER when**
  - **the entire Primary CLUSTER is DOWN (nodes may be running, but there is no MC/ServiceGuard cluster formed)**
  - **the entire Primary CLUSTER is UNREACHABLE (the network may be down)**
- **Failover to the Recovery CLUSTER does NOT occur when**
  - **individual nodes in the cluster fail**
  - **individual packages are not running**

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



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)

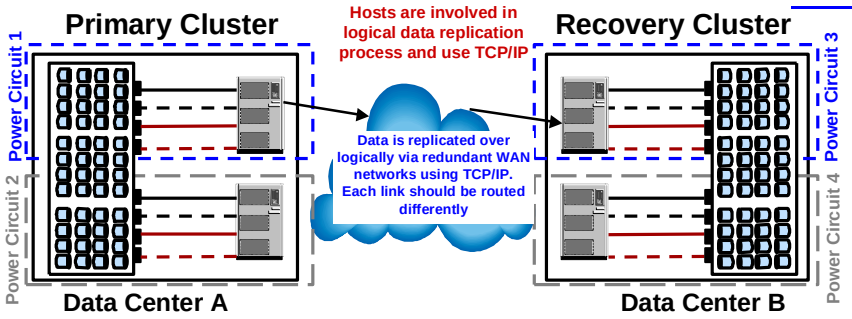
–bi-directional replication is less feasible (cost, network bandwidth)

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

Physical Data Replication

- ContinentalClusters supports the choice of physical data replication methods
- Integrated data replication solutions that have been tested by HP with ContinentalClusters (HP will SUPPORT the Integration SCRIPTS only)
  - HP SureStore E Disk Array XP256 with Continuous Access XP (NOTE: the XP256 is a fully-supported HP product)
  - EMC Symmetrix with Symmetrix Remote Data Facility (SRDF)
- No other physical replication solutions are supported for use with ContinentalClusters

Disk Architecture # 4b
Logical Replication with Oracle



Primary MC/ServiceGuard Cluster with Active Oracle Instance

Recovery MC/ServiceGuard Cluster with Recovery Mode Oracle Instance ONLY

ContinentalClusters with Oracle Standby Database

- Oracle 8i Server (not Oracle 8i Parallel Server currently) installed on hosts in each data center
- Database in each data center is protected with RAID technology
- One or more Oracle 8i apps run on various nodes in the primary cluster
- Secondary data center has a copy of the database that is updated asynchronously by the Oracle Standby Database feature
- Applications are started at the secondary cluster using the replica data in case of disaster at the primary data center

Oracle Standby Database

- Data is replicated logically and asynchronously (via TCP/IP) 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

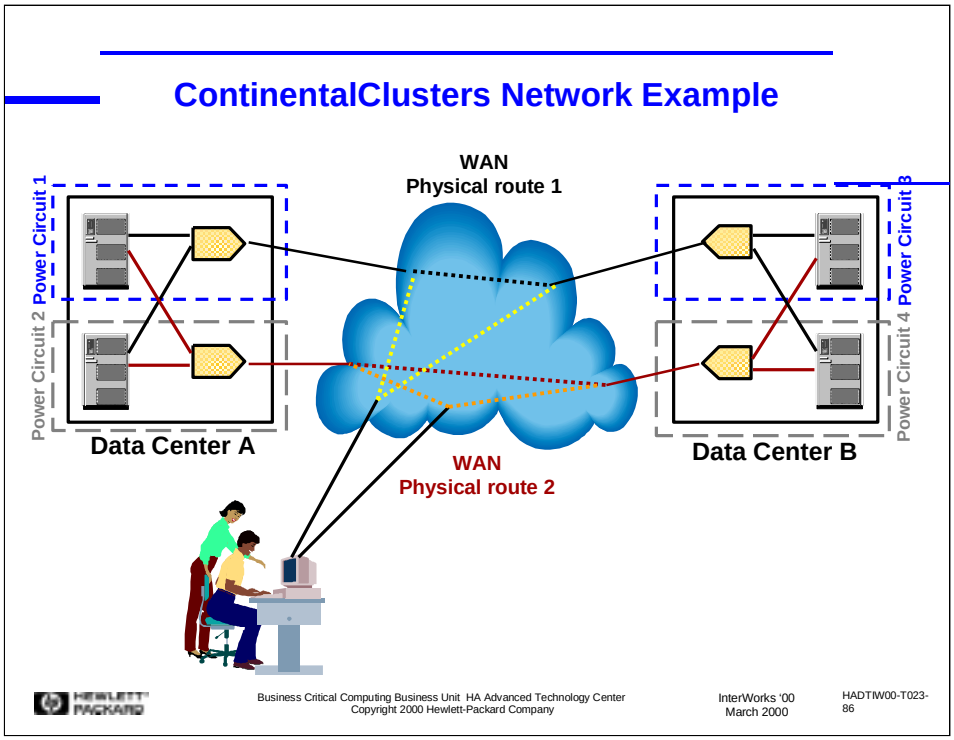
Oracle Standby Database

- **Swapping personalities** involves
  - Shutting down the application
  - Performing a full backup of the database and creating a new standby control file
  - Start the application
  - Transferring the full backup of the database, the new standby control file, and any newly generated archive log files to the old primary site
  - Starting up the new standby database at the old primary site using the new standby control file
  - Applying all the newly generated archive log files that were copied over from the source database at the standby site (that is now primary)
  - Turning on “Managed Recovery Mode” on the standby database at the old primary site (now the new standby site)

- **Data lost upon failure**
  - currently open log file
  - log file currently being transferred
  - any log files queued up to be transferred

Requirements for Oracle 8i Server

- **Requires MC/ServiceGuard clusters**
- **Oracle 8i Server with Logical Replication**
  - **Active Oracle Instance on a host in the Primary Cluster**
  - **Recovery Mode Oracle Instance on a host in the Recovery Cluster**
- **Active Instance**
  - **Oracle processes actively running**
  - **Attached to database**
  - **Users connected**
- **Recovery Mode Instance**
  - **Oracle processes actively running**
  - **Attached to database in RECOVERY MODE**
  - **Users NOT connected**



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:

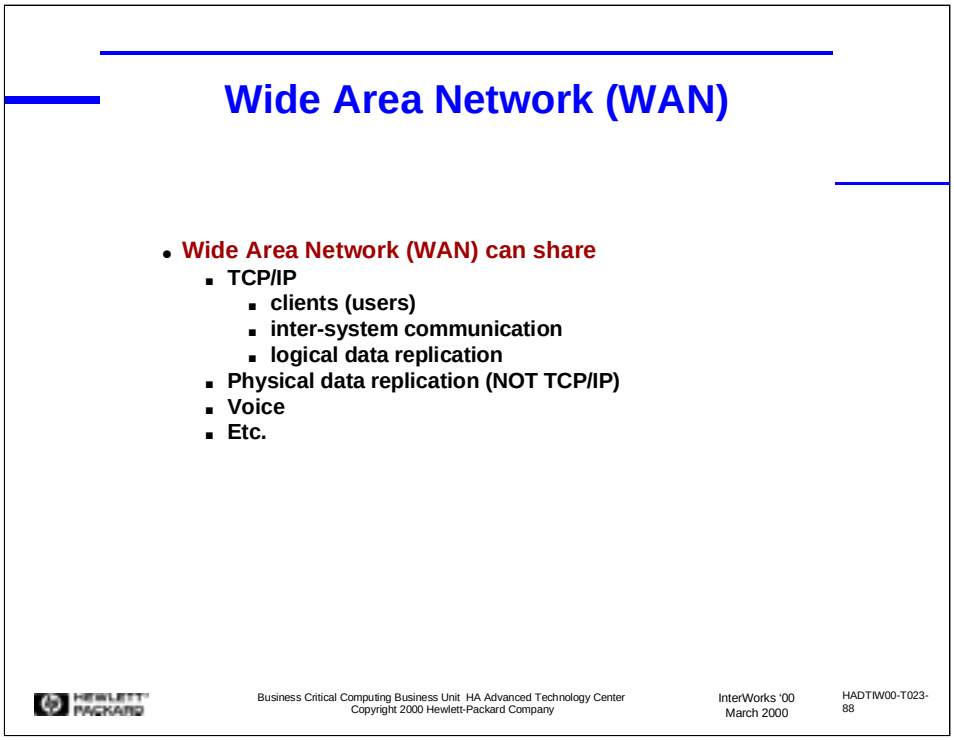
- ATM
- T1 / T3 / E3
- SONET
- Satellite
- FDDI

WANs for large organizations might incorporate multiple link technologies.

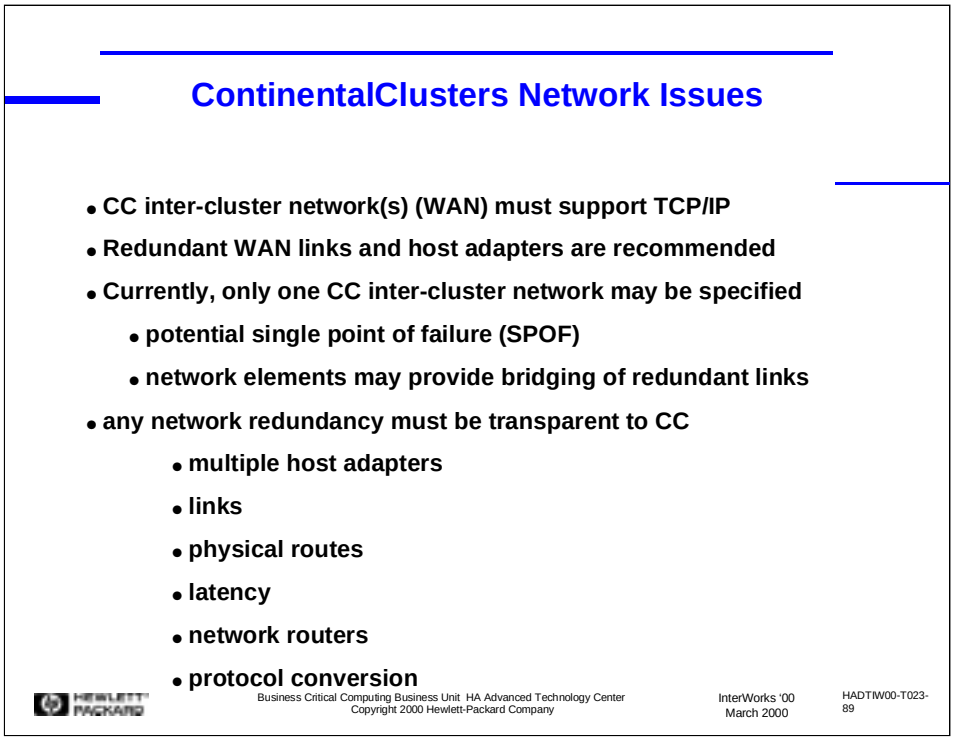
Client reconnect is typically more difficult in this environment. Routers used by clients may need knowledge of two different IP subnets, each in a different location.

Wide Area Networks (WANs) & ContinentalClusters

- ContinentalClusters works with various Wide Area Network technologies as long as the technology supports TCP/IP
  - ATM (Asynchronous Transfer Mode)
  - Frame Relay
  - T1 / T3 / E3
  - SONET
  - Satellite
  - FDDI
  - Internet
  - others



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



• the monitor interval may have to be adjusted to compensate for network latency

• synchronous data replication is very sensitive to network latency and may have a significant negative impact on application performance

ContinentalClusters Network Issues

- ContinentalClusters does not switch the IP address like MC/ServiceGuard does
- Package at Primary Cluster and Package at Recovery Cluster would have different IP addresses due to WAN
- Part of procedure during failover to Recovery Cluster would have to include any reprogramming of network elements
- Access by the clients (users) must be handled externally
  - Network routers
  - DNS servers
  - Client software - try different IP addresses
  - Oracle 8 - multiple IP addresses using TNS names
  - User hostname selection

ContinentalClusters Network Issues

- Data replication network issues
  - Theoretical latency is 4 ns / m and is affected by such things as
    - Protocol overhead
    - Protocol conversions
    - Buffering
    - Network re-routing
  - Multiply network latency by 2 for synchronous replication
  - Distance may affect (depending on data replication mode)
    - Transaction response time
    - Transaction throughput

ContinentalClusters Network Issues

- Data replication network issues
  - Transaction rate (more important with logical replication)
  - Rate of write transactions (more important with physical replication
  - Data replication mode (synchronous, asynchronous, etc.)
  - Response time required or specified as acceptable
  - Amount of data deemed acceptable for the remote system to be non-current
  - Available network technology
  - Cost of sufficient network bandwidth
  - Budget available for one-time and on-going network costs

ContinentalClusters Comparative Features

Cluster Topology	**Two Clusters, each up to 16 nodes**
Geography	**Continental or Inter-continental**
Network Subnets	**Dual IP Subnets**
Network Types	**Dedicated Ethernet or FDDI within each data center, Wide Area Network (WAN) between data centers**
Cluster Lock Disk	**Required for 2 nodes, optional for 3-4 nodes, not used with larger clusters**
Failover Type	**Semi-Automatic**
Failover Direction	**Uni-directional**
Data Replication	**Physical, in hardware (XP256 CA or EMC SRDF) Logical in software (Oracle Standby Database, etc.)**

Questions