hp storage

HP World
September 2002

technical white paper

Extended Distance SAN with MC/ServiceGuard Opens New Disaster Recovery Opportunities

executive summary

In the past 18 months a number of new technologies have become available that provide for the support of Fibre Channel SAN Inter Switch Links (ISLs) at distances up to 100km. These include Dense Wave-length Division Multiplexing (DWDM), long-haul GBICs, and Extended Fabric capabilities on the Fibre Channel switches used in the SAN solutions HP sells and supports.

HP customers have come to recognize these new technologies and capabilities and want to leverage them in support of Extended MC/ServiceGuard cluster configurations. Prior to the advent of extended SAN capabilities, for customers requiring disaster tolerant clusters of more than 10km, HP only offered solutions based on the MetroCluster product. MetroCluster leverages storage array functionality that provides data replication for distances up to 100km (Continuous Access XP). Although MetroCluster is a very good solution for many customers, many others desire an alternative to the MetroCluster solution, one that uses standard Fibre Channel components and relatively inexpensive modular storage.

summary

In the past, an Extended MC/ServiceGuard cluster was referred to as a "CampusCluster". The old campus clusters relied on the capability of the Fibre Channel technology, using switches and/or hubs, to provide interconnection of up to 10km for storage connectivity to the servers in the cluster. Current changes in Fibre Channel technology now allow for distances up to 100km, far from these old "campus" distances. Since these new clusters now extend far beyond "campus" distances we are referring to them as "extended " clusters.

This paper is intended to be a tool for customer to use in understanding the requirements for designing and deploying Extended MC/ServiceGuard clusters. Many decisions have to be made when designing an extended cluster. These decisions can have a tremendous impact on the availability of the solution, consistency of the data, and overall cost of the solution. This paper will discuss the design choices available for Extended MC/ServiceGuard clusters and their ramifications.

for more information

"Clusters For High Availability", second edition, by Peter S. Weygant.

[March 2002 version 1.0]

overview

This paper is intended to provide explanation of some of the underlying concepts of the MC/ServiceGuard product, concepts concerning inter site link technologies, and then will demonstrate how to properly design a disaster tolerant Extended MC/ServiceGuard cluster using these concepts. This paper is not intended to teach the reader how to configure MC/ServiceGuard, how to set up and configure HP-UX LVM Mirroring (MirrorDisk/UX), or the basics of SAN design. Users of this white paper are expected to be very well versed in MC/ServiceGuard cluster design and configuration, SAN design and configuration, and have a very strong background in the HP-UX LVM and MirrorDisk/UX products.

current situation

An Extended MC/ServiceGuard cluster is a normal single MC/ServiceGuard cluster that spans two (or in some cases three) data centers. It has cluster nodes and storage devices in two major data centers and uses host based mirroring (MirrorDisk/UX) to replicate (mirror) the MC/ServiceGuard packages applications' data between these two data centers. A proper design allows an application to failover from one data center to the other in case of a site disaster.

hp solution overview

The primary factor leading a customer to deploy an extended cluster solution is protection from the failure of an entire data center due to a limited geographic disaster. Fire, flooding, and site power failure are just a few examples of limited geographic disasters that can result in the failure of an entire data center. Both Extended MC/ServiceGuard clusters and MetroClusters can provide protection for disasters of limited scale, as the maximum distance allowed between data centers using either of these products is now 100km (about 60 miles). For customers interested in disaster protection from widespread geographic disasters such as hurricanes or earthquakes, products like HP's "ContinentalClusters" or other disaster recovery products should be investigated. This paper focuses on Extended MC/ServiceGuard clusters.

On the surface, an Extended MC/ServiceGuard cluster looks exactly like an ordinary (single data center) MC/ServiceGuard cluster. There are no special configuration files or parameters that differentiate an ordinary MC/ServiceGuard cluster from an Extended MC/ServiceGuard cluster. In fact, by simply looking at the server configuration and cluster configuration files you cannot tell if a cluster is an ordinary MC/ServiceGuard or Extended MC/ServiceGuard cluster.

The primary difference between a single data center MC/ServiceGuard cluster and an Extended MC/ServiceGuard cluster is the physical location of the components (servers and storage devices) that make up the cluster, and the purpose behind separating the cluster components. In a standard MC/ServiceGuard cluster, generally, all the hardware components reside in a single data center (Figure 1), and in most cases HP MirrorDisk/UX is only used to protect application data in the cases where the storage of choice is Just a Bunch of Disks or JBODs[1].

[1] There may be cases where a particularly cautious customer uses host based mirroring to mirror data that is also protected on RAID storage arrays but these are the exceptions and not the rule.

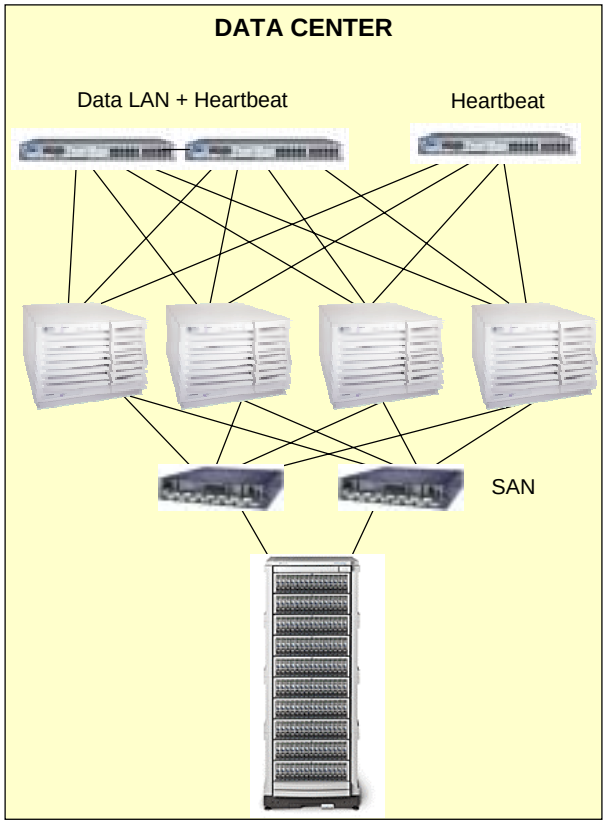

Figure 1: Cluster in a single data center

In an Extended MC/ServiceGuard cluster, the hardware components reside in two geographically separated data centers separated from one another by a few hundred meters up to 100 kilometers (Figure 2). HP MirrorDisk/UX is always used to replicate application packages' data between the two data centers in an Extended MC/ServiceGuard cluster, even if the data is stored on RAID disk arrays. This physical separation of the cluster hardware components results in additional cluster and solution design considerations over and above those required for a simple MC/ServiceGuard cluster residing in a single data center. In short, an Extended MC/ServiceGuard cluster is a MC/ServiceGuard cluster that relies on strong architectural rules, design rules, and requirements to provide the ability to survive a site disaster. Additionally, it always uses host based mirroring (MirrorDisk/UX) to replicate (mirror) the application packages' data between two separate storage devices that reside in physically separated data centers.

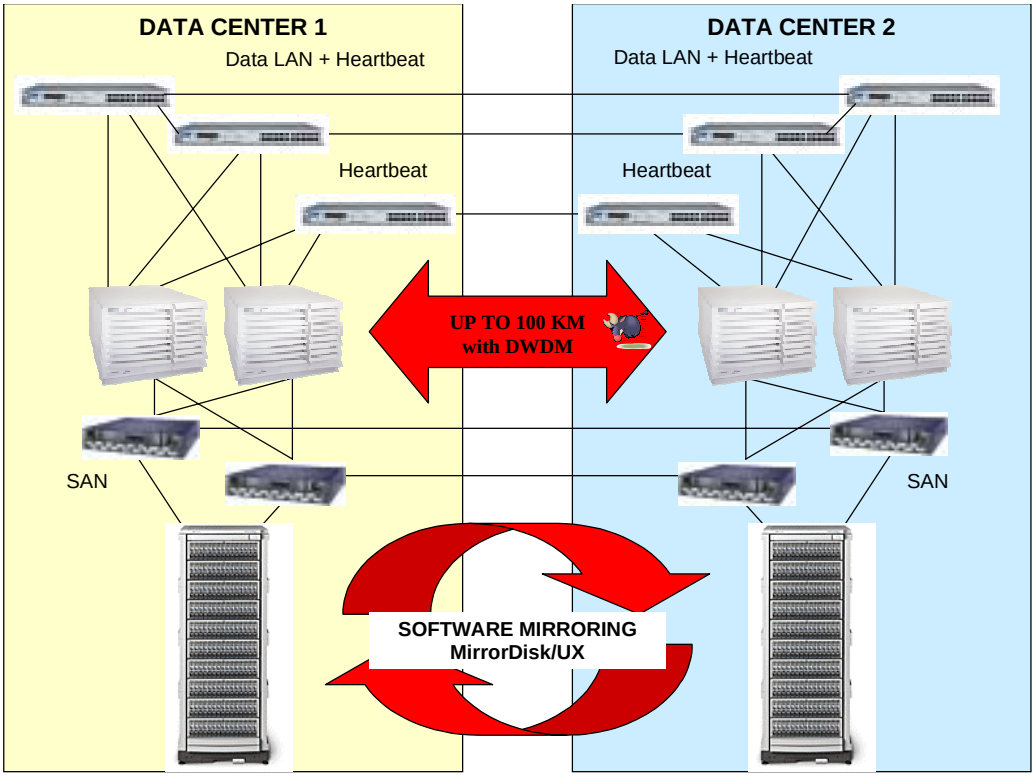

Figure 2: Cluster spanning two data centers

No special licensing, no special package control scripts, no special storage, and no special host or cluster configuration is required for an Extended MC/ServiceGuard cluster. Extended clusters use standard products sold by HP in their design. There are however specific design rules, configuration rules, and architecture rules necessary to build a proper Extended MC/ServiceGuard cluster. These additional rules add requirements and limitations to an extended cluster that may or may not be imposed on a standard MC/ServiceGuard cluster. These differences include but are not limited to:

- the use of at least two physically separate sites to contain the cluster components (a three site solution is also available)
- data replication using MirrorDisk/UX to mirror application package data between sites
- special considerations for ensuring that cluster quorum can be achieved when cluster reformation occurs
- special considerations for the physical routing of the cables between sites containing the cluster components
- integration of tools outside MC/ServiceGuard into the overall solution to help detect failures in the solution and help guarantee data consistency

When designing an Extended MC/ServiceGuard cluster, all the rules and requirements that apply to a normal MC/ServiceGuard cluster must be adhered to. An Extended MC/ServiceGuard cluster that does not follow all standard cluster design and configuration rules and requirements will not be supported by the Hewlett-Packard Company.

problems, challenges, and constraints

The following sections get into the details of MC/ServiceGuard cluster quorum and inter-site link technologies before actually tackling the issue of Extended MC/ServiceGuard cluster design. A thorough understanding of cluster quorum and the inter-site link technologies available is necessary if the best choices in design tradeoffs between solution cost and solution availability are going to be made.

understanding cluster quorum

As stated previously, the primary purpose for a customer to deploy an Extended MC/ServiceGuard cluster is protection from the failure of an entire data center. In order to design a cluster that does this properly requires an understanding of how the MC/ServiceGuard cluster reformation protocol works and how "cluster quorum" is achieved after a failure so that a new cluster may be formed. Understanding cluster quorum and how the associated algorithms work will help:

- ensure proper cluster design
- articulate different design tradeoffs and their ramifications on the ability of the cluster to respond to different failure scenarios
- articulate how the cluster will behave if a failure occurs during maintenance periods when cluster nodes have been manually halted
- explain what split-brain syndrome is, why it is dangerous, what can be done to eliminate the potential for it
- guide you to make good choices on the overall design of the Extended MC/ServiceGuard cluster

cluster quorum requirements

Since an Extended MC/ServiceGuard cluster is just a plain old vanilla MC/ServiceGuard cluster, the standard MC/ServiceGuard cluster formation protocols apply. For quorum to be achieved during a cluster reformation, one of the two following requirements, on the number of nodes forming the cluster, must be met or a new cluster will not be created.

MC/ServiceGuard Cluster Quorum Requirements

- Strictly, more than 50% of the active members from the previous cluster membership must be present (all cluster members are required when a cluster is initially started unless a manual override is specified)
- 50% of the members of the previous cluster must be present and a cluster lock disk must be claimed

If neither of these conditions can be met, i.e. cluster quorum cannot be achieved, the cluster protocols will ensure the nodes attempting to form the cluster issue a Transfer of Control (TOC) (A TOC looks just like a PANIC in Unix).

The cluster protocols are defined to operate in this manner to guarantee data protection. Since all the nodes in a running cluster are connected to the same set of disk devices, if a failure occurs that causes multiple clusters to attempt to form, the protocols will ensure that only one new cluster successfully forms. If two separate clusters are allowed to form and they both write to the common disk devices, data corruption will occur. The cluster formation protocols and cluster quorum requirements are designed to prevent this from occurring. If a cluster tries to form and quorum cannot be attained, the nodes attempting to form the non-quorum attaining cluster will TOC. This ensures that if another cluster (which has attained quorum) forms, only it will have write permission to the common disk devices. A properly designed cluster will never get into a state in which a new cluster cannot form after a failure (assuming only a single failure caused the cluster reformation[2]).

cluster lock disk(s)

A cluster lock disk is needed whenever a cluster containing exactly ½ of the nodes from a prior cluster wants to form a new cluster. It can also act as a tiebreaker when two separate clusters attempt to form after a failure and each of the clusters contains exactly ½ of the nodes that were active cluster members before the failure. A simple

[2] There are many multi-failure scenarios that MC/ServiceGuard can recover from. Loss of an entire data center and its associated cluster members is considered a single failure.

example of a situation where this would occur is a two-node cluster that suffers a networking segmentation failure that prevents the two nodes from exchanging cluster heartbeats or cluster reformation protocol information. The lack of heartbeat messages between the nodes will result in cluster reformation occurring on each node. Both nodes will try to form a one-node cluster consisting of just itself. When the cluster reformation protocol gets to the point where it checks to see how many nodes are in the new cluster it would detect that ½ of the nodes from the prior cluster are present (each node is forming a one node cluster and the prior cluster contained two nodes). At this point, the new clusters would attempt to claim the cluster lock disk. The protocol that controls claiming of the cluster lock disk is designed to guarantee that one and only one cluster will gain access to the cluster lock disk and the other cluster will not get access. The node that acquires the cluster lock disk will have achieved quorum and will become the new active cluster. The node that did not acquire the cluster lock disk will not have achieved quorum and will TOC leaving just one active cluster connected to the common disk storage.

Note: Cluster lock disk(s) are only supported for MC/ServiceGuard clusters containing four nodes or fewer. They are not supported for clusters containing more than four nodes.

dual cluster lock disks

A MC/ServiceGuard cluster can be designed with one or two cluster lock disks. As has already been discussed, a cluster lock disk is used as a "tie-breaker" when a potential cluster containing ½ the nodes of the previous cluster attempts to form. This means assurances must be taken during the design of a cluster, so that a single cluster lock disk does not become part of a SPOF (Single Point Of Failure) that keeps a new cluster from forming (i.e. we must ensure that a single failure cannot cause ½ of the nodes and the cluster lock disk to fail simultaneously). In the case that a single cluster lock disk is part of a SPOF then the cluster must be designed with dual (two) cluster lock disks. In a cluster with dual cluster lock disks, one of the disks is referred to as the primary cluster lock disk and the other as the secondary cluster lock disk. Here is an explanation of how the dual cluster lock disk protocol works:

> A set of nodes trying to form a new cluster will attempt to claim the primary cluster lock disk followed by the secondary cluster lock disk. If the set of nodes attempts to claim the primary cluster lock disk and the primary cluster lock disk has already been claimed by another cluster, the set of nodes will stop trying to form a cluster and will TOC (without trying to claim the secondary cluster lock disk). However, if the primary cluster lock disk is not accessible (i.e. the cluster cannot communicate with the primary cluster lock disk), the set of nodes will wait for a given timeout value (contained in the cluster reformation protocol) before trying to claim the secondary cluster lock disk. If it now acquires the secondary cluster lock disk, it will form a cluster. If after waiting the timeout, the secondary cluster lock disk has already been claimed by another cluster, the set of nodes will not form a cluster and will TOC. The following flowchart diagrams dual cluster lock disk behavior:

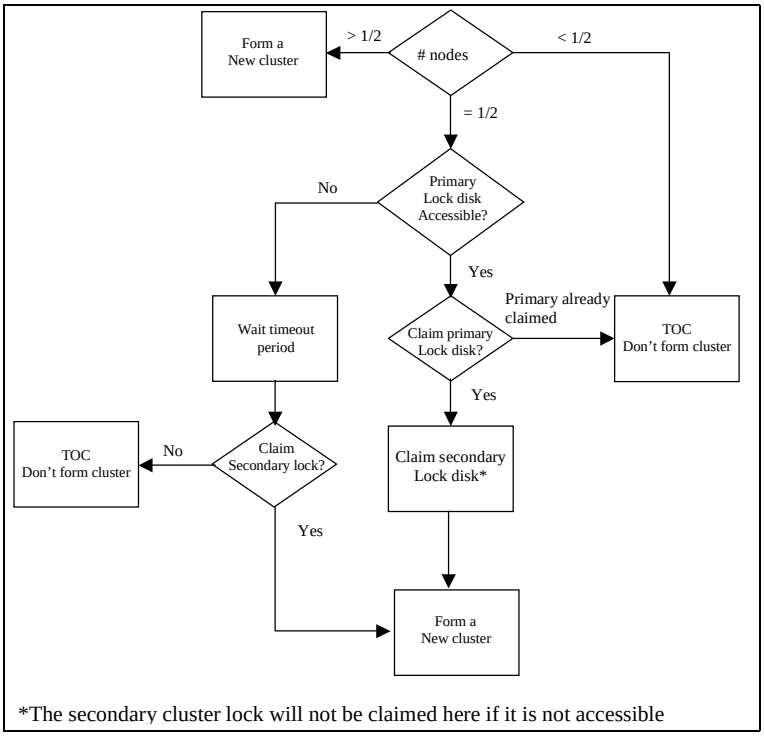

Figure 3: Dual cluster lock disk algorithm flowchart

Cluster lock disk SPOF Example: this example demonstrates how a single cluster lock disk can potentially be a SPOF for the cluster. This example is intended only to demonstrate the need for dual cluster lock disks. This is a somewhat simplistic example and assumes all the cluster components have a single power input (today, almost all servers and storage provide support for redundant power connections, so all cluster components would be connected to redundant power circuits).

One cluster design example where a single cluster lock disk is a SPOF for the entire cluster is that in which half of the nodes in the cluster and the single cluster lock disk can all become disabled simultaneously. This can occur in a design where only two power circuits are available to power the cluster nodes and the cluster lock disk (Figure 4).

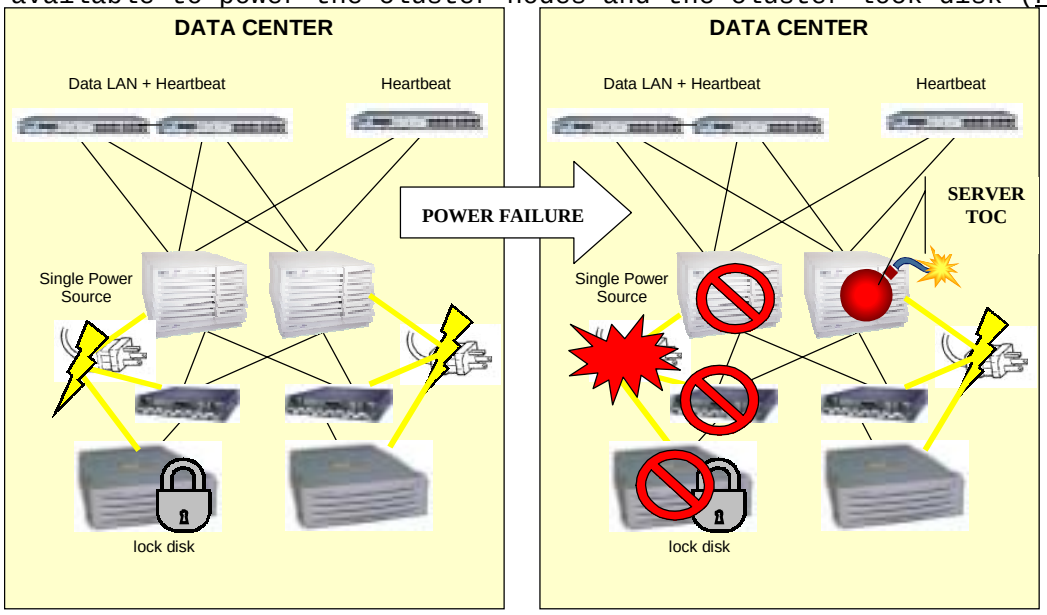

Figure 4: Cluster lock disk as a SPOF

If the cluster is designed with a single cluster lock disk, the cluster lock disk must share a power circuit with half of the cluster nodes. If this critical power circuit fails, the remaining nodes will not be able to achieve quorum when they form a new cluster since they will not be able to claim the now unavailable cluster lock disk. The result is

the entire cluster fails. To protect from this, the cluster would have to be configured with dual cluster lock disks, one on each power circuit used by the cluster. In this case, after the failure of one power circuit a new cluster is able to form (Figure 5).

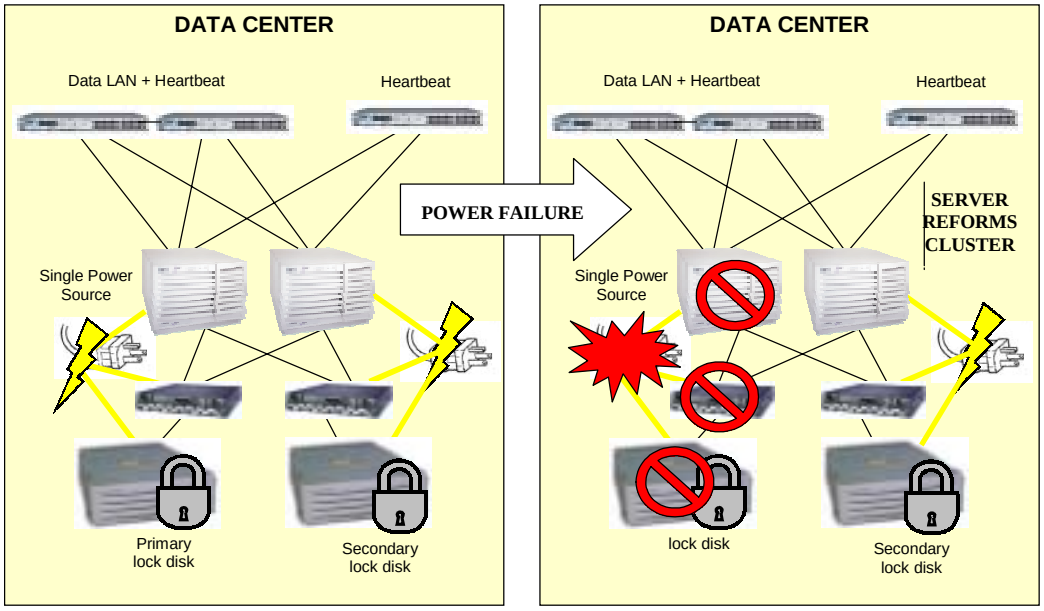

Figure 5: Dual Cluster lock disk

split brain syndrome

"Split Brain Syndrome" describes a situation where, after a failure on a cluster configured with dual cluster lock disks (actually multiple failures if the cluster is properly designed), two separate clusters are able to achieve cluster quorum and form. This can occur if multiple network failures result in a partitioning of the networks carrying cluster heartbeat and multiple failures of the SAN links connecting the cluster servers to the cluster lock disk devices occur (these failures do not have to occur at the same time but the SAN failure must occur at or before the network failure). What happens in a split brain failure scenario is that ½ of the cluster nodes end up on one side of a network partition and the other ½ of the nodes end up on the other side. The SAN connections must also be partitioned where one cluster lock disk (from a dual cluster lock disk configuration) ends up on each side of the SAN partition. When this occurs, two separate clusters form, each containing ½ of the nodes from the prior cluster plus a cluster lock disk device (Figure 6). Split-brain is an undesirable state for the cluster to attain and is only possible in a configuration containing dual cluster lock disks.

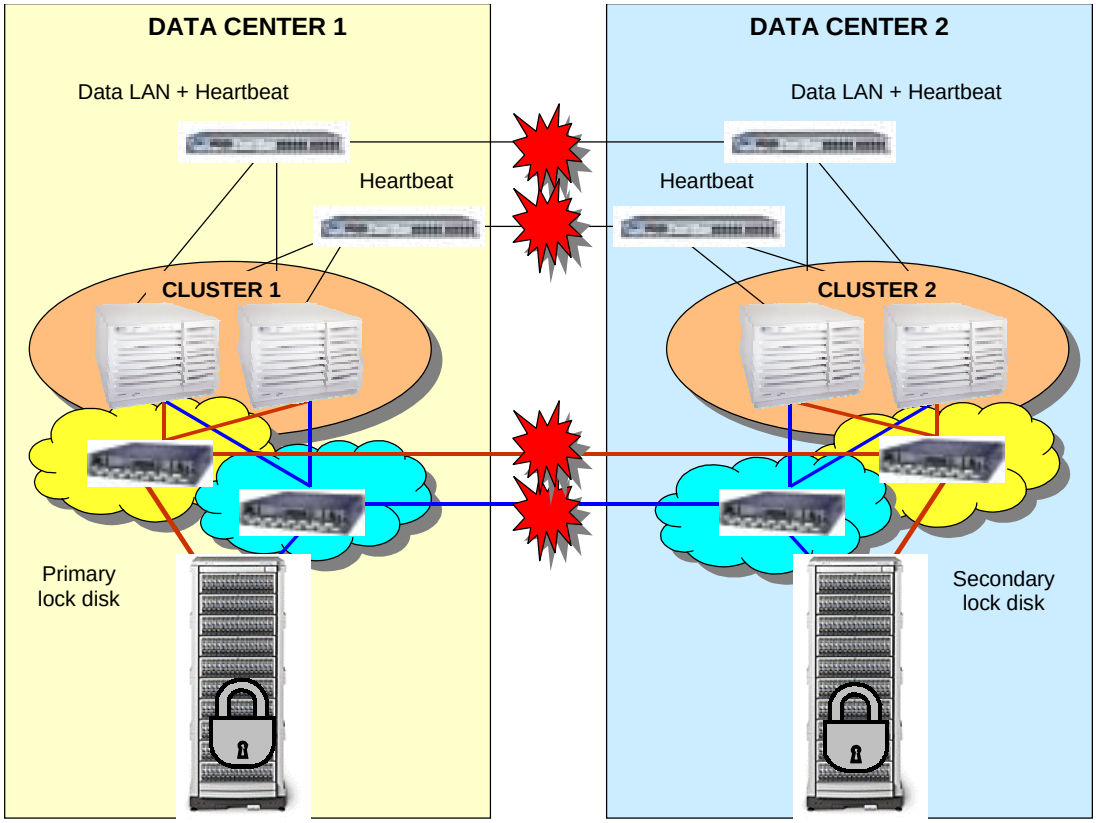

Figure 6: Split-brain example
(The data subnet is not HA in this diagram)

designing to ensure cluster quorum

In the case of an Extended MC/ServiceGuard cluster, designing the cluster in a manner that ensures the cluster can achieve quorum after a failure is a critical issue. In addition to the standard cluster protection against a single node failure, an extended cluster must protect against a site disaster in which up to ½ of the nodes in the cluster may simultaneously fail. Based on the definition of cluster quorum, in a two site cluster, if we employ a simplistic design that equally splits the resources between two data centers, taking no precautions to guarantee the ability to achieve quorum, the loss of a single site will always result in a total cluster failure as it will be impossible to reform a new cluster in the surviving site. If we design the cluster with dual cluster lock disks, one in each center, a new cluster will be capable of forming after a site failure. If we design a solution using three data centers and ensure no single data center contains more than ½ of the nodes we remove the requirement for dual cluster lock disks.

In essence, there are two primary configurations that all extended cluster solutions must be based upon to guarantee the viability of the design. These two solutions are:

- Two site design with dual cluster lock disks
- Three site design without cluster lock disks

These two design options will be discussed in detail in the section "extended cluster design".

inter-site link technologies and requirements

This section discusses, in detail, the physical and functional requirements for both the IP networks and the Storage Area Network, SAN, links used by an Extended MC/ServiceGuard cluster. There are both distance and functionality restrictions on the interconnects that must be adhered to when designing a solution.

requirements for physical interconnection between sites

Making good choices for the physical interconnect between sites for the networking and the SAN is very important. Connections (both IP networking and SAN) must not only be redundant, they must be physically routed in such a manner that a single event cannot cause all IP and all SAN paths to fail simultaneously. This physical separation of the links is actually an aspect of cluster design that can be applied to standard MC/ServiceGuard clusters, Extended MC/ServiceGuard clusters, MetroClusters, and ContinentalClusters.

Protecting from physical failures starts inside the data centers that house the cluster nodes and extends to the physical routes the cables take between the sites containing the data centers. This means that inside the data center, redundant networking and SAN components should be physically separated from one another to protect from physical destruction due to incidents like a limited fire, falling objects, or the improper operation of fire sprinklers in the building. Remember, most failures resulting from a physical incident (fire for example) don't affect an entire building. They most often affect a small portion of a building (most fires don't burn down an entire building, but are contained to a single room or a closet and even more often to a single piece of equipment). Cabling within the building should ensure that redundant cables for the cluster are not co-located. Care should be taken to try and ensure that redundant cables for the solution do not share a common exit point from the building(s). Once outside the building, at least two separate physical conduits, taking separate paths, must be utilized between the sites and they must be sufficiently far apart from one another that both interconnects cannot be broken as a result of a single event. An example of such an event would be a backhoe hitting the conduit containing the physical cables between the sites. See Figure 7 for an example showing physically separate cluster infrastructure components.

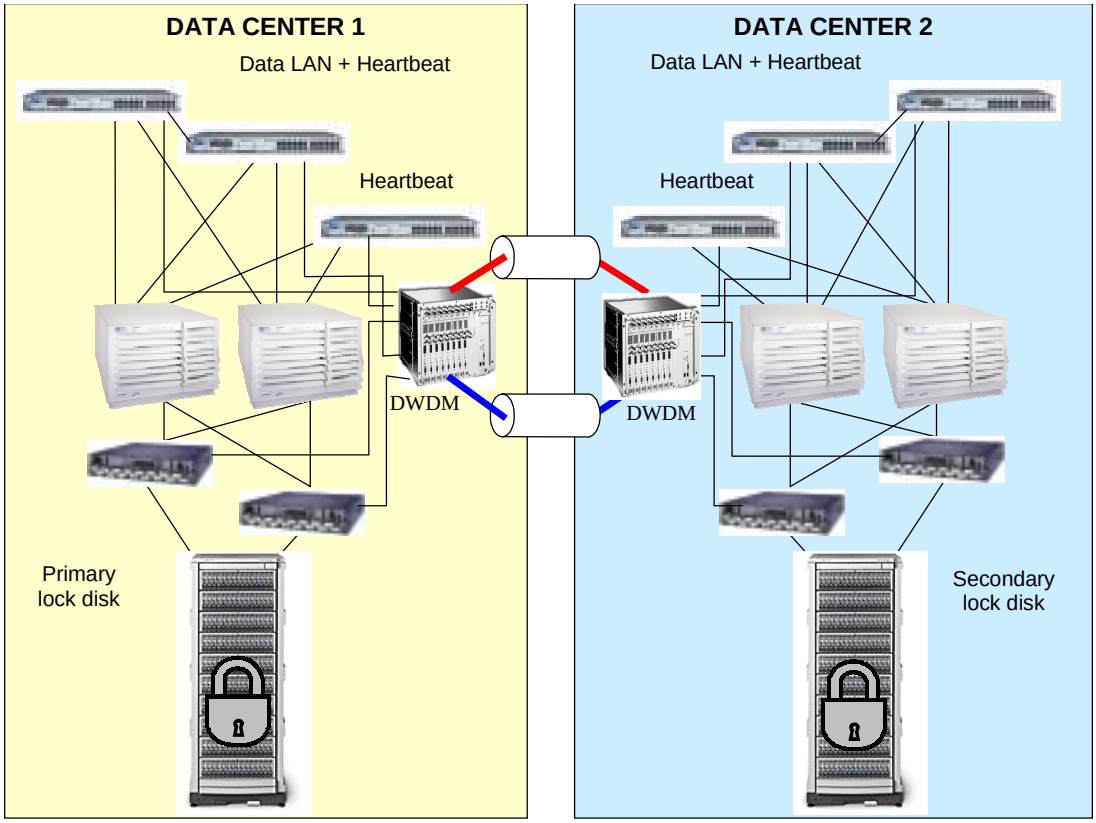

Figure 7: Separate physical intra-site infrastructure with DWDM (fault tolerant)

storage area networking

The SAN is the key component of an Extended MC/ServiceGuard cluster. It is not only necessary to provide a means for the replication of the data generated by the applications running on the cluster, it also plays a part in cluster reformation if the solution utilizes cluster lock disks. Care must be taken to ensure it is designed in a manner that not only provides availability but also meets the performance requirements of the solution during normal operations, after a SAN component failure, and during (re)synchronization operations.

Prior to the writing of this white paper, extended campus clusters were limited to a maximum distance of 10km. This was the maximum distance supported by a Fibre Channel SAN using standard hardware components (hubs, switches and long-wave GBICs). Now, the advent of new technologies allows SAN design and implementation to be extended up to 100km.

short-wave, long-wave and long-haul ISLs[3]

Fibre Channel is capable of running across different type of mediums, copper cables and optical fibers. HP supports only optical fiber based Fibre Channel as SAN interconnect links in its HP-UX cluster solutions. Also, only fiber optics allow long distances connections. If the SAN connections between the sites will be utilizing fiber optic cables dedicated to the SAN (no plan to have them be shared to support both the SAN and the IP networking) then simply connecting hubs[4] or switches together directly, using short-wave, long-wave, or long-haul GBICs is acceptable.

The distance that can be reached depend on multiple factors including:

- the characteristics of the optical fiber
- the quality of the optical fiber
- the wavelength of the laser

[3] ISL stands for Inter-Switch Link. This paper will use ISL to refer to the SAN optical fiber connections between the sites that containing the data centers.
[4] There are serious restrictions on HUB performance. See the section on SAN performance for details.

The different wavelengths available are commonly referred to as "short-wave" or "long-wave" and are associated with the laser module (GBIC) devices. In this context, short-wave connections use light in the 780 and 850 nm range and 62.5/125 or 50/125 µm multi-mode optical fibers reaching distances of up to 500 meters. Long-wave[5] connections use light in the 1330 nm range and 9/125 µm mono-mode optical fibers and can reach distances up to 10km. Recently, a new type of GBIC module has been introduced. The laser for these new modules operate in the upper limit of the long-wave range, around 1550 nm, and are commonly called " long-haul" GBICs. The fiber required by these laser modules is the same as that necessary for long-wave support, 9/125 µm mono-mode optical fiber. The distance that can be reached with these new long-haul modules essentially depends on the characteristics of the chosen hardware but can be up to 80km with the Finisar FTR-1519.

The following table summarizes the distances provided by the different connection types and supported for Extended MC/ServiceGuard clusters.

Optical fiber specification	62.5/125	50/125	9/125
Short-wave	175 m	500 m	-
Long-wave	-	-	10 km
Long-haul	-	-	80 km

Table 1: distance, wavelength and optical fiber specification

DWDM

DWDM stands for Dense Wavelength Division Multiplexing. It is an opto-electronic technology whose concept is simple: simultaneously transmit separate optical signals through the same fiber at different wavelengths or colors of light. A DWDM device takes as input an optical signal at a particular wavelength, it translates it to another wavelength, and multiplexes it with other optical signals (whose wavelengths have been translated too) over a single fiber. The signals are restored at the opposite end of the DWDM connection, de-multiplexed and translated back to their original wavelength. DWDM has no knowledge of any protocols. Each input signal uses one channel of the DWDM. It's a kind of repeater with wavelength modification and, potentially, a multiplexing of multiple wavelength signals. The maximum distance allowed between a DWDM device pair depends on the particular DWDM vendor product used, but can reach distances as high as the 100-120 km range.

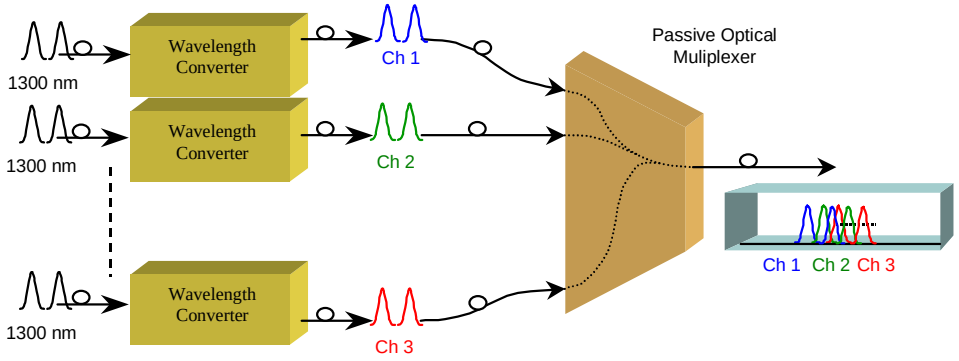

Figure 8: DWDM illustrated

HP does not require any particular vendor's DWDM equipment be used. The customer is responsible for the selection and maintenance of any DWDM equipment. However, HP can provide, upon request, a list of **tested** but **not certified** DWDM vendor's equipment.

DWDM is an alternative for solutions that require more than two ISLs between the sites or for solutions where the customer wants to be able to carry both optical networking and Fibre Channel traffic on the same pair of optical fibers. Since DWDM has no protocol awareness of any kind, multiple protocols can be run simultaneously over a single optical fiber via DWDM: Ethernet 100BaseFX, FDDI, Gig Ethernet, Fibre Channel and ESCON (essentially any optical based protocol).

A DWDM solution can support multiple separate clusters and multiple non-cluster related optical connections simultaneously over a redundant pair of optical fibers. The redundant

[5] The wavelength associated with long-wave in the Fibre Channel redbook is between 1300 and 1550 nm.

fibers between the DWM equipment must follow the standard physical routing requirements for an extended cluster in that they must follow different physical routes between the sites they connect.

choosing the right inter-site link technology

There are different ISL options available when designing A SAN. The choice of ISL in the design of the SAN is particularly important as it not only affects the initial cost of the solution, it can have future ramifications on the scalability of the solution if the customer's SAN bandwidth requirements increase. A number of factors must be evaluated before making an ISL choice. These factors include but are not limited to:

- **Distance between sites:** For distances up to 80km direct connections between switches in the SAN can provide the distance support necessary or DWDM can be used. For ISLs over 80km, DWDM must be used to provide the ISL links in the solution.
- **SAN Bandwidth requirements:** Proper sizing of the SAN during its design will determine how many ISLs are required between the sites to meet the necessary bandwidth requirements during both normal processing and (re)synchronization operations. When using short-wave, long-wave, or long-haul GBICs, a separate optical fiber is required for each ISL. Multiple optical fibers between sites can become very expensive. For solution requiring more than two ISLs, use of DWDM becomes an option.
- **Number of optical fibers between sites:** If the customer wants to leverage the same pair of optical fibers for the both IP networking and the SAN between the sites, then DWDM will be required. Using DWDM limits the number of optical fibers needed between sites to two.
- **Telecom provider requirements (i.e. dark fiber requirements):** We are encountering more and more telecom providers who only offer dark fiber over DWDM. Check with the local provider to determine if DWDM is an option or requirement for the solution.
- **Cost:** An analysis must be performed to determine if dedicated ISL fibers and networking infrastructure are more cost effective than DWDM or not.
- **Customer preferences**

internetworking

The network capability requirements for a two site extended campus cluster are exactly the same as those required for a standard MC/ServiceGuard cluster and a three site Extended MC/ServiceGuard cluster's networking requirements are exactly the same as a MetroCluster. The network must consist of a single IP subnet (to support relocatable IP addresses) and support DLPI communications end-to-end (to support MC/ServiceGuard cluster formation protocols). The latency must not interfere with the cluster heartbeat messages or the cluster reformation protocols. The overall end-to-end network latency for the cluster must be at or below 200ms.

Note: The overall end-to-end network latency, for the cluster IP network, must be at or below 200ms.

The network must be designed and deployed in a manner such that no SPOF in the network exists that can result in the entire cluster to failing because cluster heartbeats cannot be communicated between cluster nodes or cluster reformation protocol messages cannot be exchanged between enough cluster nodes to maintain cluster quorum.

Two recommended networking options for extended clusters are:

- FDDI
- Optical Ethernet (Gigabit or 100-FX)

FDDI and Optical Ethernet will operate and are supported over DWDM for an extended clusters with data centers at distances of up to 100km apart from one another.

extended MC/ServiceGuard cluster design: bringing the pieces together

This section delves into the overall design of an Extended MC/ServiceGuard cluster including the SAN and Networking infrastructure. It discusses both two-site and three-site designs, articulating the tradeoffs of the two choices.

extended MC/ServiceGuard cluster design philosophy

In all cases, the first and foremost requirement of an HA or DR design is that it protect from data corruption. A solution that keeps the applications running but allows data to become corrupt is useless (applications cannot run properly on corrupt data). An HA or DR design's second requirement is that it provide for application availability. Data protection must always precede availability.

When designing an extended cluster, there are many acceptable architectures, options, and tradeoffs that go into the design. Many of these design decisions and tradeoffs occur as a result of the customer's requirements. Sometimes these tradeoffs consist of choosing between two undesirable failure scenarios that cannot both be avoided, either because it is not technically feasible to design a solution that protects from both or because the customer doesn't want to spend the money necessary to protect from both. Whatever the reason for the choices made, it is imperative that the customer's availability requirements are well understood and that any tradeoffs made are agreed to. It's not uncommon for the customer's availability requirements to conflict with his design requirements. In those cases where there is a conflict steps must be taken to educate the customer about the conflict, explain why it exists, and articulate the choices available to minimize the conflict while still providing for data protection and availability. It is very important that the customer clearly understand the ramifications of all design choices made on the data integrity and availability of the solution.

single points of failure (SPOF) in a design

In the disaster tolerant solution realm two different types of SPOF should be considered:

- Technical SPOF
- Design SPOF

Disaster tolerant clusters can actually be designed to survive multiple failures (everything from several hardware components in the cluster to an entire site failure). It is recommended that, when designing and deploying extended cluster, all measures are taken to design a solution that can recover from as many multiple failures as possible.

technical SPOFs

A technical SPOF is a situation where the solution is not properly designed to protect from the failure of a critical hardware component. An example would be an extended cluster designed around two non-fault tolerant Fibre Channel switches for disk connectivity. The switches themselves have limited or no high availability features built into them. If one of the switches were to fail cluster nodes will loose all communications capability to the storage devices served by that switch. This may result in cluster package failures and a lack of data replication. Designs containing technical SPOFs are not acceptable and will not be supported by the Hewlett-Packard Company.

design SPOFs

Design SPOFs are different from technical SPOFs in that a conscious decision was made to accept the exposure resulting from not protecting a single critical piece of hardware that in and of itself is highly available. An example would be a cluster designed around two fault tolerant FC64 Fibre Channel switches. Since these switches are near fault tolerant (no single component failure within the switch will result in a complete failure of the switch), a single pair of these switches is acceptable and is supported for an Extended MC/ServiceGuard cluster solution. However, the fact that there are only two of them in the solution makes each of them a potential SPOF. If for some reasons one of the switches is physically damaged (for example, it is damaged or destroyed by fire), the ability for the cluster to replicate data between data centers would be compromised and the DR protection provided by the cluster would be rendered ineffective.

It is extremely important that the customer understand the ramifications of all design SPOFs before they accept them. The ramifications of these decisions can be very far reaching when a failure that the customer thought highly unlikely does in fact occur.

Attention: Any Extended MC/ServiceGuard cluster designs that contain technical SPOFs will not be supported by the Hewlett-Packard Company. Clusters containing design SPOFs will have to be evaluated on a case-by-case basis and should be thoroughly documented in a TCO

Complete solution design and planning engagement from the TCO organization. See your HP sales rep for details.

key areas of design consideration

There are four key areas of concern when designing an Extended MC/ServiceGuard cluster.

Key Design Areas

- The number of sites the cluster will occupy (two or three).
- The Ethernet network communication infrastructure that supports the cluster. (includes the networks for cluster communication and client attachment to the cluster members)
- The Storage Area Network (SAN) communication infrastructure that connects the cluster servers to the mass storage (disk) devices.
- Ensuring cluster quorum can be achieved after a site failure.

All of these areas must be addressed with data protection (consistency and currency) and cluster quorum capability in mind. The remainder of this section will go into detail on supported configurations, limitations, and how to properly design for each of the areas listed above for both "two data center" and "three data center" designs.

two site cluster design with dual cluster lock disks

As the name implies, this configuration defines a solution in which the cluster components are geographically separated into two independent data centers. Dual cluster lock disks are required in a two data center solution in order to guarantee recovery from an entire data center failure.

Warning: Any time dual cluster lock disks are configured into a solution the door has been opened to split-brain syndrome. Since a two data center design uses dual cluster lock disks it is susceptible to split-brain syndrome. Special care should be taken in the overall design of the cluster to limit the possibility of split-brain occurring. In a properly designed, configured, and deployed two-site solution it would be very difficult for split-brain to occur (as a minimum, two links must fail) but it is still possible.

two site design requirements and restrictions

There are some special requirements and restrictions placed upon two data center Extended MC/ServiceGuard cluster solutions that must be adhered to. Since dual cluster lock disk devices are required, a two-site cluster must contain either two nodes or four nodes. Three node clusters are not allowed, as there must be an equal number of nodes in each data center. This is a direct consequence of the cluster quorum mechanism. Also, clusters with greater than four nodes are not allowed because MC/ServiceGuard only supports cluster lock disks for clusters containing four nodes or less.

Two Site Design Limitations and Requirements

- Two separate data centers are used
- Symmetric clusters only – equal number of nodes at each site
- Two node and four node cluster only – no three node clusters
- Lock disks only supported for clusters with up to four nodes.
- Dual cluster lock disks required – one in each data center, each on a separate bus
- Special care must be taken to minimize the potential for split-brain syndrome to occur – physical connections between the data centers (cables) must be redundant and must be routed via at least two separate physical paths

Refer to Figure 9 for an example diagram of a supported two site extended cluster.

why a restriction of either two nodes or four nodes?

Why must a two data center cluster contain an even number of nodes (two or four) and at the same time not contain more than four nodes? Why must the nodes be allocated evenly between the two data centers? The answers to these two questions are fairly straightforward. They relate to cluster quorum and a MC/ServiceGuard cluster lock disk configuration limitation.

Remember, the purpose for deploying an extended cluster is to protect from the failure of an entire data center. Having an even number of nodes in each data center and dual cluster lock disks relates directly to the operation of the MC/ServiceGuard cluster reformation protocols and cluster quorum. After the failure of a node or nodes in a MC/ServiceGuard cluster, the cluster reformation protocol requires that at least ½ of the nodes that were active in the cluster prior to the failure must be present during reformation for a new cluster to form – in the case of ½ of the nodes forming a cluster, a cluster lock disk must also be available.

Cluster Reformation Protocol Requirements for Quorum

- At least ½ of the nodes that were active in the cluster prior to the failure must be present during reformation for a new cluster to form
- In the case of ½ of the nodes forming a cluster, a cluster lock disk must also be available

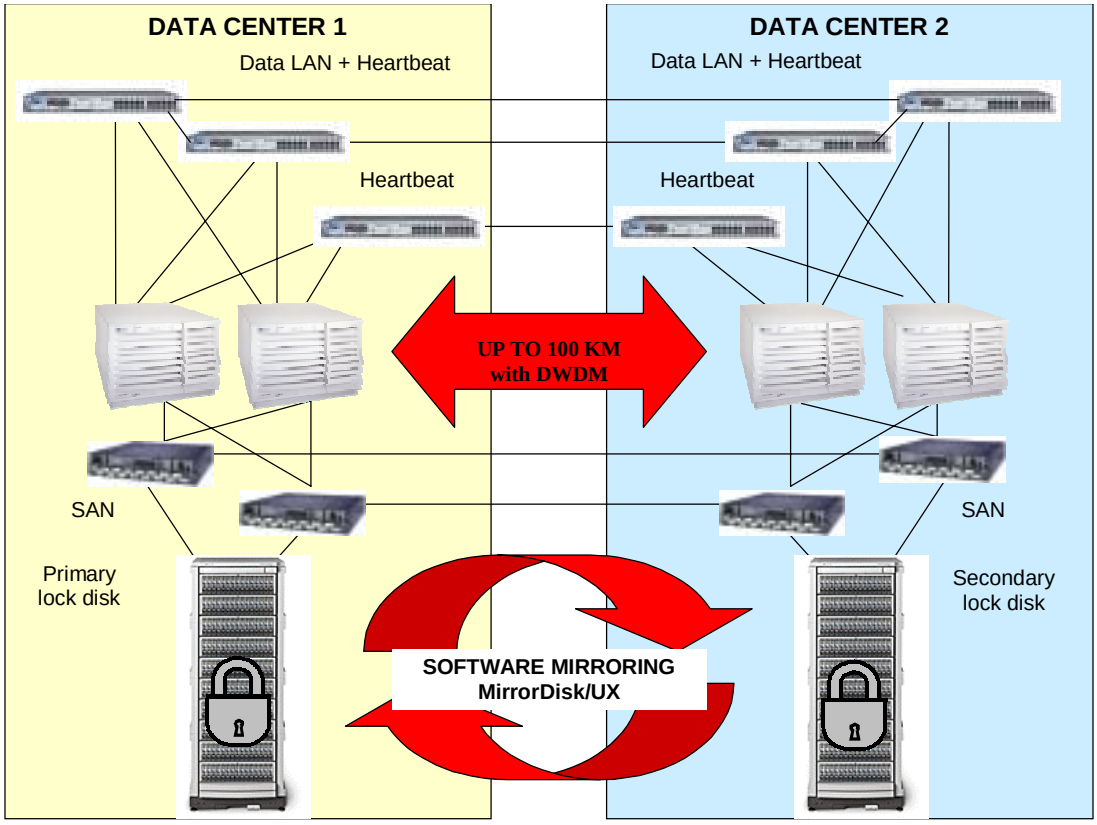

Figure 9: Two-site cluster design with 4 nodes

If we had an asymmetrically designed cluster where one data center contained more nodes than the other, if the data center with the greater number of nodes in it fails, the nodes in the surviving data center will not be able to achieve cluster quorum as defined by the MC/ServiceGuard cluster protocols. As a result, they would all TOC as specified by the cluster reformation protocol, resulting in a failure of the entire cluster (this brings up an interesting issue related to cluster maintenance procedures and ensuring there is always a symmetrical number of nodes in the cluster. This is the reason for the requirement of an even number of nodes distributed symmetrically between the two data centers.

The basis for the requirement that dual cluster lock disks (one in each data center) be present in the cluster is that, in addition to requiring at least ½ of the nodes from the prior cluster, the cluster reformation protocols may require the acquisition of a cluster lock disk device for a new cluster to form. By configuring a cluster lock disk in each data center we ensure that a cluster lock disk will be available if an entire data center fails and a new cluster needs to be formed.

Split-brain syndrome is when two separate clusters form. It can only occur in a cluster configured with dual cluster lock disks. In a properly designed and deployed cluster, split-brain syndrome can only occur when a muti-point failure occurs that allows exactly one half of the cluster nodes to communicate with each other and have access to one of the two cluster lock disks, and at the same time, the other ½ of the cluster nodes have communications among themselves and access to the other cluster lock disk. Since zero or one cluster lock disks is not an option with a two data center extended cluster, the only way to avoid the potential of split brain is to architect a three data center cluster which does not use cluster lock disks.

Warning: Any time a cluster is configured with dual cluster lock disks there exists the potential for split brain syndrome to occur. The only way to **guarantee** split brain will never occur is to never configure a cluster with dual cluster lock disks.

three site cluster design

In this design paradigm the nodes in the cluster reside in three independent data centers. Clusters of three nodes and clusters with greater than four nodes are supported as are asymmetric clusters (clusters that do not contain an equal number of nodes in all three

data centers) of five nodes or more. The attractiveness of a three data center cluster design is that it does not need, and in fact cannot use, cluster lock disks to achieve cluster quorum after a data center failure. Since it doesn't use cluster lock disks, a three site extended cluster is not susceptible to split brain syndrome. A properly designed three site Extended MC/ServiceGuard cluster solution will always be able to achieve cluster quorum after a site failure because >50% of the nodes will always be available to form a new cluster.

three site design requirements and limitations

There are requirements and limitation for three-site clusters that do not exist for two-site clusters. Special care must be taken when designing a three-site cluster to ensure cluster quorum can be achieved after an entire site failure. The nodes in a three-site cluster do not have to be distributed symmetrically, they may be distributed in an asymmetric manner but there are requirements on node layout to ensure quorum can be achieved after a site failure. There must be an equal number of nodes in both of the major data centers. Also, in a three-site cluster not all cluster nodes are connected to all of the storage. Nodes connected to the mass storage devices will be referred to a primary node and will reside in the two major data centers. Nodes not connected to the mass storage will be referred to as arbitrator or secondary nodes and reside in the third data center. Disaster recovery occurs between the two major data centers since these data centers contain the mass storage devices between which application data is being replicated (Figure 10).

Three Site Design Limitations and Requirements

- Asymmetric clusters supported (i.e. not all three data centers have an equal number of nodes)
- The major data centers must contain an equal number of nodes[6]
- Node layout must guarantee quorum after a site failure
- No data center may contain >= ½ of the cluster nodes
- Clusters may contain three nodes or greater than four nodes
- Four node clusters are not supported
- Only nodes in the major data centers are connected to the storage devices
- Cluster lock disks not needed and not supported
- The two networks connecting the three sites must be routed in different directions so a one site failure does not become a SPOF for both networks
- Up to 16 nodes in a cluster

[6] An exception to this is a cluster where one major data center and the data center containing the arbitrator nodes each contain two nodes and the second major data center contains one node. See table 2.

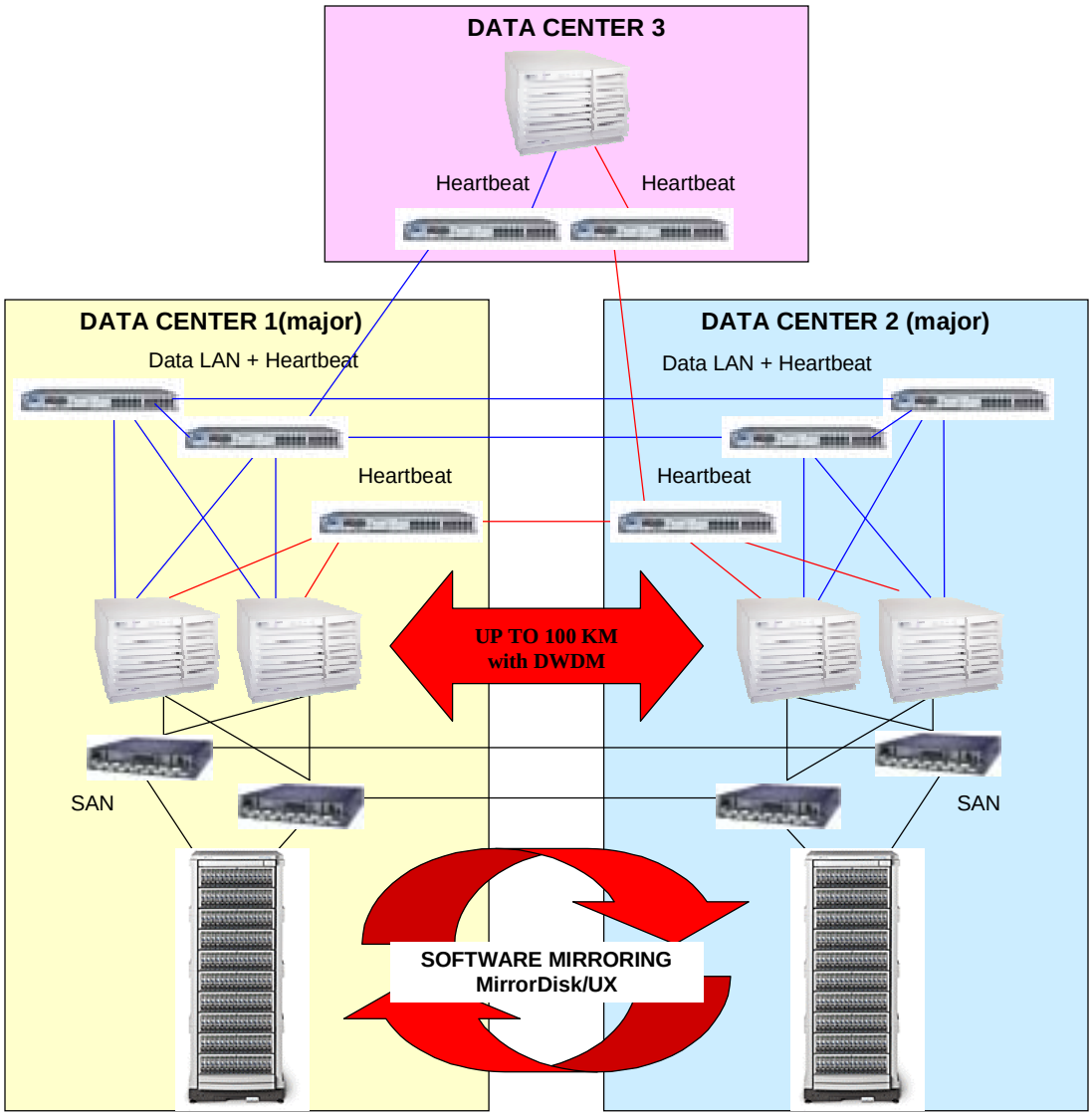

Figure 10: Three-site cluster design

cluster lock restrictions and limitations

For MC/ServiceGuard to support cluster lock disks in a cluster all the nodes in the cluster must have access to all cluster lock disk devices. In a three-site cluster, only the nodes in the two major data centers are connected to the storage, the nodes in the third data center are not. Therefore, cluster lock disks are not supported in a three site extended cluster solution.

Since cluster lock disks are not supported, how is quorum achieved if an entire site fails? In a properly designed three-site cluster, cluster lock disks are not necessary. In case of a site failure, quorum will be achieved by ensuring greater that ½ of the nodes from the prior cluster are still available in the two surviving sites. Refer to the next section for details.

Node layout restrictions and limitations

The nodes in a three site extended cluster may be distributed between the sites asymmetrically. This must be done in a manner that ensures no one site is a SPOF for the entire cluster. This means no one site may contain one half or more of the total nodes in the cluster. If we allow one site to contain one half or more of the cluster nodes and that site suffered a total failure, the surviving nodes will not be able to achieve quorum and the entire cluster will fail. As a result of this design requirement, four node clusters are not allowed in a three-site design. With a four-node cluster there is no way to distribute the nodes across three sites and not have one site contain ½ of the nodes. Three node clusters are supported and clusters with greater than four nodes are supported, but not four-node cluster.

Demystifying the “ Arbitrator” server(s)

In a three-site cluster only the two major data centers actually contain cluster nodes that run MC/ServiceGuard application packages. The third site contains a server(s) which is part of the cluster, but its primary purpose is to be part of the cluster reformation process so that quorum can be achieved without the need for cluster lock disks. In the past (in MetroCluster) the server(s) in the third data center have been referred to as "arbitrator" server(s). Although they are referred to by this special name, there is nothing special about their configuration or role in the MC/ServiceGuard cluster. They are only included in the cluster to ensure that if one of the major data centers fail, enough cluster nodes will still be active (> ½) to ensure cluster quorum can be achieved without the need for a cluster lock disk.

It is possible to configure two arbitrator servers at the third site and have these servers share some common local disk storage. You can then configure a MC/ServiceGuard package(s) that fails over between these two nodes, but which always remains local to this third data center.

hints on choosing the number of nodes for the cluster

It is fairly easy and straightforward to architect a three-site cluster in which a site failure is not a SPOF. However, special consideration must be given to the design if we do not want a SPOF to exist after an active cluster member has left the cluster. In other words, if one node leaves the cluster due to a failure or maintenance, the cluster may be susceptible to a total failure resulting from a site failure. It is not only possible to design the cluster in a manner that will close this hole, it is recommended (though not required). The easiest method for accomplishing this is to ensure that all three sites contain a minimum of two nodes each. This will ensure that there are always enough active nodes to achieve quorum even if the active cluster is running short one member. Keep in mind that this only applies after a node leaves the initial cluster that contains all cluster members as active members.

Let's look at a table showing the number of nodes necessary at each site to ensure the cluster's ability to recover from a site failure following a single node failure (or a node removed from the cluster administratively)

Note: The total number of nodes contained in sites 1, 2 and 3 must be equal to or less than the number of nodes supported by the version of MC/ServiceGuard you are using.

As can be seen from the table below, designs containing >= 2 nodes in both major data centers and two arbitrator nodes in the third data center will provide protection from a site failure even after a single node has left the cluster dues to a failure or for administrative operations (i.e. OS upgrade, hardware upgrade, cluster software upgrade). For this reason we recommend two arbitrator nodes for all three-site cluster containing two or more nodes in the major data centers.

Major Data Center 1	Major Data Center 2	Data Center 3 (arbitrator site)	Guarantees protection from site failure after node failure
1	1	1	No
1	1	2	Not Supported
2	2	1	No
2	1[7]	2	No
2	2	2	Yes
3	3	1	No
3	3	2	Yes
4	4	1	No
4	4	2	Yes
5	5	1	No
5	5	2	Yes
6	6	1	No
6	6	2	Yes
7	7	1	No
7	7	2	Yes

Table 2: Per-site node quantity impact

extended MC/ServiceGuard cluster SAN design

There are no "special" SAN design criteria required for an Extended MC/ServiceGuard cluster. The standard requirements for availability, path redundancy, and consideration for bandwidth requirements found in normal SAN design are sufficient when designing an Extended MC/ServiceGuard cluster SAN.

ISL support for extended MC/ServiceGuard clusters

For extended clusters, HP supports SAN designs using short wave GBICs, long-wave GBICs, Finisar long-wave GBICs, and DWDM for switch ISL links. A single hop between two switches, of up to 100km in length is allowed in the SAN for an extended Service Guard cluster. This means the sites containing the disk storage devices can be up to 100km apart from one another. The following table outlines the ISL lengths supported for different optical fiber specifications:

Optical fiber specification	62.5/125	50/125	9/125
Short-wave	175 m	500 m	-
Long-wave	-	-	10 km
Long-haul	-	-	80 km
DWDM			100 km

Table 3: distance, technology and optical fiber specification

SAN design rules and requirements

When designing the SAN portion of an Extended MC/ServiceGuard cluster solution on HP-UX servers, all SAN design rules outlined By the Hewlett-Packard Company must be followed or the SAN will not be supported by the Hewlett-Packard Company.

dual SAN clouds

As stated earlier, Extended MC/ServiceGuard clusters rely on HP MirrorDisk/UX to replicate the data between the data centers containing the disk storage devices. Proper design of the Fibre Channel infrastructure and configuration of the volume groups and logical

[7] This is the ACSL MetroCluster 2-1-2 configuration. See the ACSL Support web page for details

volumes used by the application packages is necessary to ensure proper protection of the data belonging to the applications. It is recommended that at least two separate SAN clouds be used to connect the servers to the disk devices in the solution. The two halves of a mirrored logical volume should be connected to each server in the cluster via separate SAN clouds. As a minimum, one SAN cloud can connect the cluster servers to the disk storage devices in the local data center (one half of the mirror) and the second SAN would connect the servers to the disk storage devices in the remote data center (the other half of the mirror). See Figure 11.

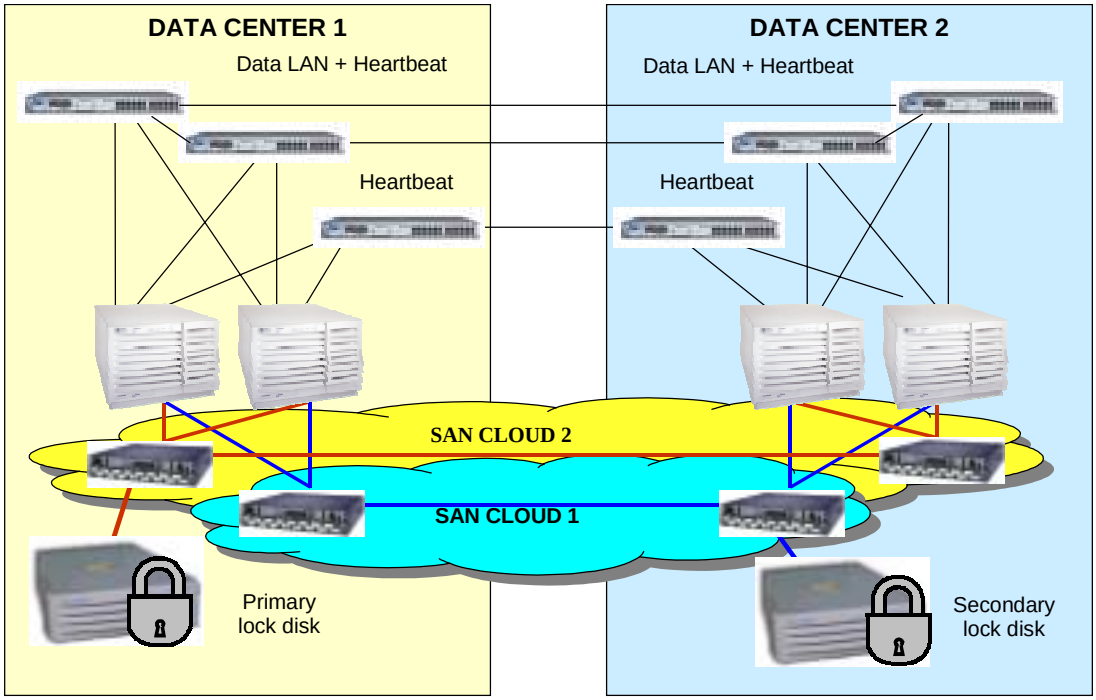

Figure 11: Dual SAN clouds without PV-Link support

It is not a requirement, but it is highly recommended that the cluster servers and storage devices be connected to the SAN in such a manner that PV-Links (alternate paths) can be configured (Figure 12). Whether PV-Links can be configured or not depends on whether the storage devices in the solution provide redundant connection capabilities or not (currently all HP Fibre Channel based disk storage devices provide connection redundancy that can be used to support PV-Links). For an extended cluster, configuring PV-Links will help ensure MirrorDisk/UX can continue to replicate the data between sites even if one of the SAN paths becomes unusable (Figure 13). Furthermore, not configuring PV-Links exposes applications to data loss in the event of a SAN link failure followed by a site disaster.

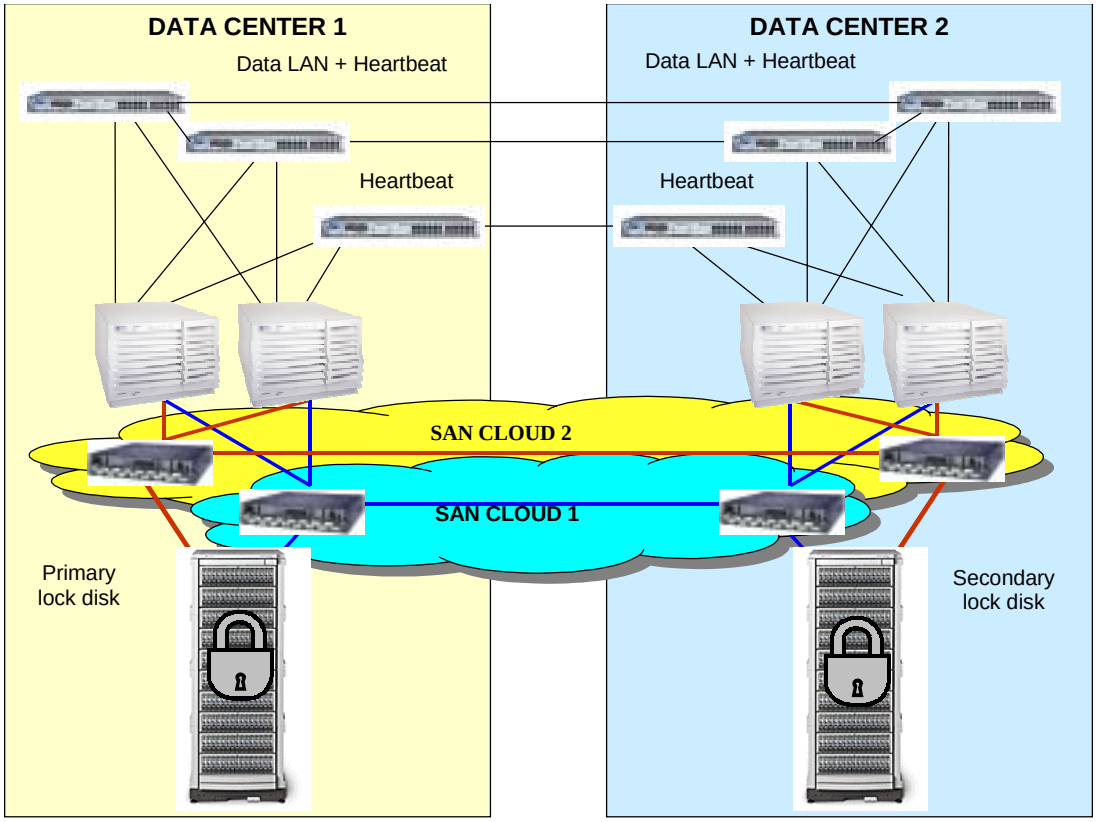

Figure 12: Dual SAN clouds with PV-links - Ideal design

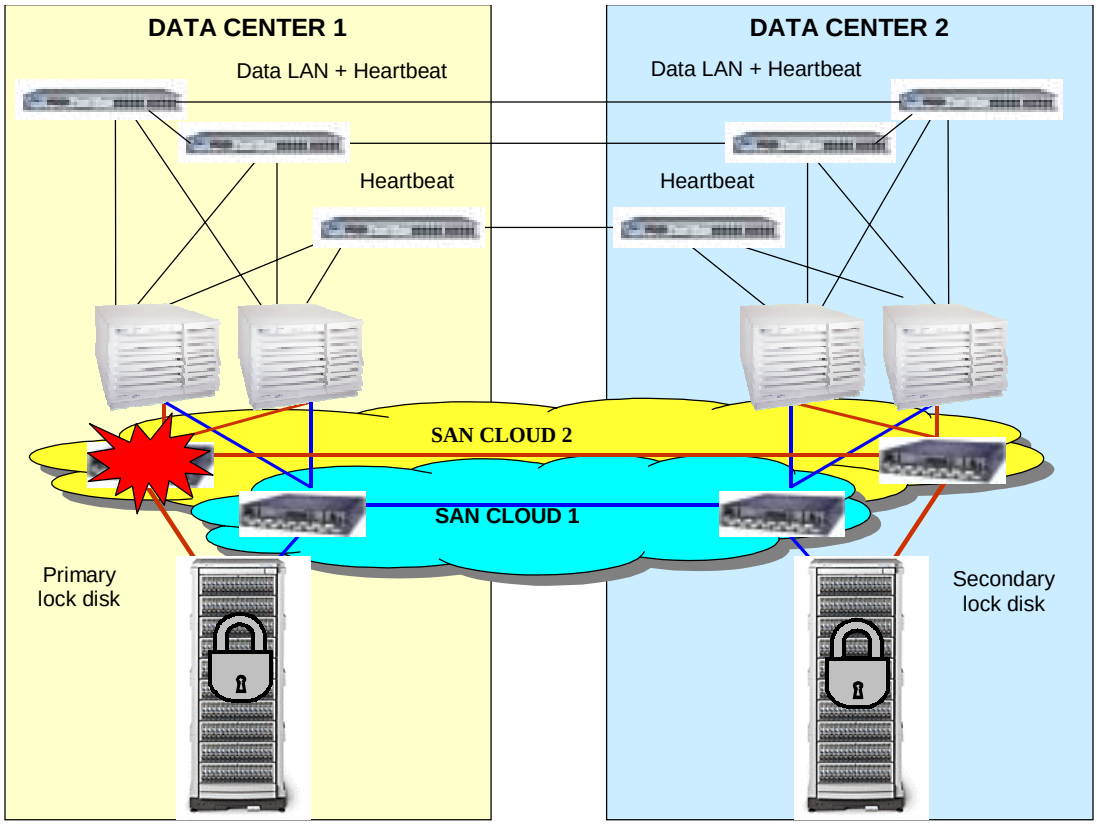

Figure 13: Uninterrupted data replication with dual SAN cloud with PV-Links after a switch or link failure

extended cluster IP network design

In an extended cluster, all of the standard requirements for MC/ServiceGuard networking must be met:

MC/ServiceGuard Networking Requirements

- Routed networks cannot be used as heartbeat networks
- Heartbeat networks must support DLPI protocol
- Maximum 200ms network latency heartbeat networks
- Solution must be designed so no single failure can result in all heartbeat subnets failing simultaneously

As with all other aspects of an extended cluster, the IP networks used in a solution should not only be designed for high availability, they must be designed so that the failure of an entire data center does not result in all subnets failing.

All of the standard network interface types supported by MC/ServiceGuard are supported for extended clusters. Keep in mind that any distance limitation on the network link chosen will define the maximum distance supported for the extended cluster solution. The following optical based networking links are likely to be used in an extended cluster as there are not as many distance limitation constraints on these interfaces as there are on copper based LAN links (all of these can be extended with DWDM):

- 100Base-FX
- 1000BaseT (gigabit Ethernet)
- FDDI[8] (see Figure 14 for an FDDI over DWDM example)

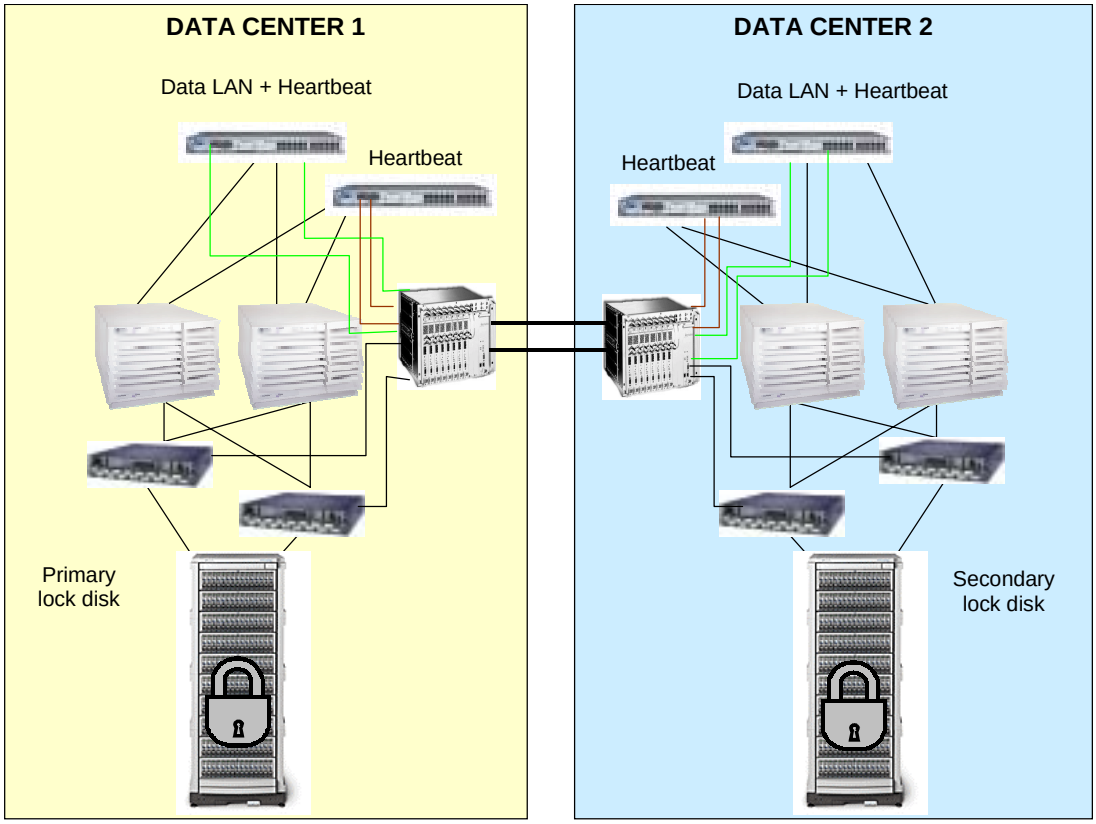

Figure 14: FDDI dual ring, redundant networks over DWDM
(The data subnet is not HA in this diagram)

[8] Each FDDI requires two slots in the DWDM converter to support its dual-ring topology. Two FDDI subnets would require four slots in each DWDM converter

dual data center network design

For a dual data center solutions the networking requirements are exactly the same as any standard MC/ServiceGuard cluster with one major difference, in an extended cluster, the physical paths the two subnets follow between data centers must be different (a standard requirement for all extended cluster inter site links). It is highly recommended that the subnet used by clients for connecting to cluster servers be made HA. All the examples in this paper show a cluster with two networks. Notice that the user network is designed to be HA and the dedicated heartbeat network is not. See Figure 12.

Three data center network design

The network design requirements for three data center extended clusters are the same as the requirements for a MetroCluster. Two separate networks are required to connect the data centers together (both must carry heartbeat). The user subnet should be made HA if possible. In a three data center design, the networks must be deployed in such a manner that one of the data centers is not a SPOF for both subnets. Figure 10 shows a properly designed and deployed network for a three data center design. In this diagram, both networks are "bus" type networks. One of the subnets has endpoint in Data center 1 and Data center 3 and the other subnet has endpoint in Data center 2 and Data center 3. If any one data center fails, the two surviving data centers will still have a communication path between them. Notice that the arbitrator server does not have HA connections to either subnet.

glossary

Continuous Access XP
: A software product available on Hewlett-Packard XP disk arrays. It provides hardware mirroring of data between two separate arrays over either and ESCON connection or a Fiber Channel connection at distances of up to 100km

DWDM (Dense Wavelength Division Multiplexing)
: An opto-electronic technology whose concept is simple: simultaneously transmit separate optical signals through the same fiber at different wavelengths or colors of light

GBIC (GigaBit Interface Converter)
: The component in a Fiber Channel device that contains the optical laser. They come in three flavors, short-wave (up to 500m), long-wave (up to 10km), and long-haul (up to 80km)

ISL (Inter Switch Link)
: A fiber optic connection between e-ports on two Fibre Channel switches.

LVM
: HP-UX Logical Volume Manager

MC/ServiceGuard
: Hewlett-Packards industry leading High Availability cluster software product

Metro Cluster
: A DR cluster product which combines MC/ServiceGuard and Continuous Access XP (or SRDF) to provide disaster tolerant clusters at up to 100km over Fiber Channel (50 km with EMC SRDF)

MirrorDisk/UX
: A licensable product that operated with HP-UX LVM to provide a host based data mirroring solution.

SRDF
: A software product available on EMC ICDAs. It provides hardware mirroring of data between two separate arrays at distances of up to 50km