WHITE PAPER

January 1998

Prepared By
High Availability
Products Group

Compaq Computer
Corporation

CONTENTS

COMPAQ

Increasing Availability of Cluster Communications in a Windows NT Cluster

Whether you currently own a Compaq ProLiant Cluster, or are considering purchasing one, your interest in the product most likely stems from a need to increase the availability of data and applications.

A ProLiant Cluster integrates software and hardware technologies. These technologies enable a set of loosely coupled servers and storage to present a single image to network clients, and to operate as a single system. To create a system of loosely coupled servers, the cluster must have a mechanism with which the cluster nodes (servers) can communicate with each other. Additionally, to present a single image to network clients, the cluster must employ a resilient communication mechanism between the cluster system and the network clients.

In this paper, each of these communication mechanisms is identified and defined, and a description of their significance to the cluster is included. In addition, the downtime issues associated with a failure of communication are detailed, along with a description of several configurations that minimize the occurrence and impact of a communication breakdown.

It is assumed that the reader is familiar with Compaq ProLiant Clusters, Microsoft Cluster Server software, and Microsoft Windows NT Server 4.0 Enterprise Edition. In addition, the reader must have a solid fundamental understanding of clustering concepts, particularly cluster communication over an interconnect and a client local area network (LAN).

NOTICE

The information in this publication is subject to change without notice.

COMPAQ COMPUTER CORPORATION SHALL NOT BE LIABLE FOR TECHNICAL OR EDITORIAL ERRORS OR OMISSIONS CONTAINED HEREIN, NOR FOR INCIDENTAL OR CONSEQUENTIAL DAMAGES RESULTING FROM THE FURNISHING, PERFORMANCE, OR USE OF THIS MATERIAL.

This publication does not constitute an endorsement of the product or products that were tested. The configuration or configurations tested or described may or may not be the only available solution. This test is not a determination of product quality or correctness, nor does it ensure compliance with any federal, state or local requirements. Compaq does not warrant products other than its own strictly as stated in Compaq product warranties.

Increasing Availability of Cluster Communications in a Window's NT Cluster

First Edition (January 1998)
ECG051/1297

TYPES OF CLUSTER COMMUNICATION

Two types of interconnects have an immediate and dramatic effect on the availability of cluster communication. These are intra-cluster communication, and cluster-to-LAN communication.

Intra-cluster communication consists of information passed from one cluster node to another. The communication is performed over an interconnect. This consists of, at a minimum, two network controllers (one in each cluster node) and a cable to connect the controllers.

Intra-cluster communication uses the interconnect data path to:

- Communicate individual resource and overall cluster status
- Send and receive heartbeat signals
- Update registry information

Cluster-to-LAN communication consists of requests and responses to and from cluster nodes and network clients. This type of communication also exists in a non-clustered environment. As can happen with a stand-alone server, failure of a key network component results in downtime for network clients, when they can no longer access their server. Availability is of primary importance because you are, or will be, operating in a clustered environment. Ensuring network clients have access to their clustered applications and data depends on the availability of the cluster-to-LAN communication path.

SIGNIFICANCE OF CLUSTER COMMUNICATION PATHS

Since these communication mechanisms operate in a cluster environment, before discussing their significance, it is necessary to understand the terminology used to describe failures in a cluster. The Compaq ProLiant Cluster is highly available, rather than continuously available, so it is important to understand what parts of the system are vulnerable to faults. When any hardware or software cluster component whose failure prevents, for an unacceptable period of time, intervention to reestablish access with the cluster resources, that component is identified as a single point of failure (SPOF). Due to the serious nature of single points of failure, a cluster should be designed to eliminate as many of them as possible.

Not all failures that interrupt cluster operations are single points of failure. As long as the cluster can recover from the failure, the data, applications[1], and network clients will return to normal operations as soon as the recovery process is complete. However, the period of time during the recovery process is considered unplanned or unscheduled downtime. While recovery is taking place, some network clients will not have access to their cluster groups. Though not as catastrophic as a single point of failure, measures should also be taken to prevent these types of disruptions, and thereby reduce downtime.

As it pertains to cluster communications, there are two issues that adversely affect the operation of the cluster. The first issue momentarily disrupts operation by causing a failover event. The second issue is a single point of failure and disrupts operation until manual intervention by an administrator resolves the problem.

The first issue involves downtime associated with failover and failback events. When Cluster Server detects an error that adversely affects the operation of a cluster **group**[2], it fails the cluster group from the one node to the other node. When Cluster Server detects an error that affects the operation of an entire cluster **node**[3], it fails all cluster groups running on that node to the other node. One such error occurs when communication between the cluster nodes is disrupted. If only one interconnect exists in a cluster configuration, each time the interconnect is permanently disrupted Cluster Server will bring all cluster groups off-line on one of the nodes and fail them over to the other node. The process of failing over cluster groups takes time. The groups must be taken off-line on their primary node, the resources of each group (applications, drive volumes, IP addresses) must be transferred over to the other node, and the transferred data must be validated on the surviving node[4]. While all of these operations occur, network clients are unable to access their cluster groups. Creating a redundant intra-cluster communication path easily and inexpensively minimizes the amount of downtime incurred due to this failure.

The second issue involves network clients' ability to access their clustered applications. Cluster Server operates its failover and failback events at a cluster group level. A cluster group usually consists of an application, service, or file share, along with any dependent resources, such as drive volumes and IP addresses. Cluster Server will likely be configured such that some cluster groups operate on cluster Node 1, and some on cluster Node 2. Each node is physically connected to the client LAN via a network controller, network cable, and a network hub.

A disruptive event will occur if the physical connection from an individual cluster node (ex. Node 1) to the client LAN is disrupted while the interconnect is still operational. For network clients whose cluster groups reside on Node 1, this event will prevent the clients from accessing their

[1] The application can only recover from a failure if it can function correctly in a clustered environment, and can be failed over.
[2] A cluster group usually consists of an application, service, or file share, along with any dependent resources, such as drive volumes and IP addresses.
[3] A node is a computer server operating as an entity in cluster configuration.
[4] This is only true if the client network is not enabled as the secondary cluster network.

cluster groups (applications). Automatic failover of the cluster groups will not occur since Cluster Server, via the interconnect, believes both cluster nodes are operating normally. Until an administrator realizes the problem, discovers the root cause is a network error, and manually fails over all the cluster groups from Node 1 to Node 2, the clients cannot make use of Node 1's clustered applications. Creating a redundant cluster-to-LAN communication path easily and inexpensively minimizes the probability of this failure event.

COMMUNICATION POINTS OF FAILURE

Several components make up the physical network of the cluster-to-LAN and intra-cluster communication paths. The failure of any one of these components renders the entire path inoperable, and results in the failure scenarios previously described. Unless redundancy has been designed into the communication paths, a component failure will cause either a failover event or a complete disruption of access to certain cluster groups.

Note: A fifth hardware item, an entire network hub, is also a single point of failure. However, the hub is viewed as a piece of the larger network, whose availability is a concern whether operating in a clustered environment or in a stand-alone server environment. In the "Examples" section, failure of a network hub is noted as a single point of failure when appropriate. Discussing how to resolve the effects of such a failure is beyond the scope of this document.

Understanding how each of these components plays a role in the interconnect and cluster-to-LAN data paths will help you comprehend the solutions discussed later in this paper. The following four hardware items are the primary points of failure.

- A port on a multiported network controller (client or interconnect)
- An entire network controller (client or interconnect)
- A network cable
- A port on a hub

The diagram below depicts each of these failure points.

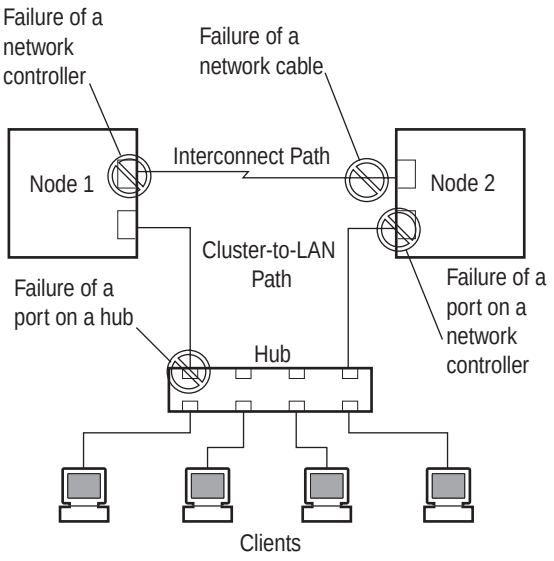

Figure 1: Communication Points of Failure

BUILDING BLOCKS OF COMMUNICATION PATH REDUNDANCY

Now that you are able to identify the primary causes of failure in the communication paths, the next step is to understand what technologies are available to combat these points of failure. As you will see in the next section, an integration of hardware and software technologies supplies the surest way to create redundancy. This increases both the resiliency of cluster communications, and the overall availability of clustered applications and data.

Compaq Advanced Network Fault Detection and Correction Feature

The Compaq Advanced Network Fault Detection and Correction feature consists of combining Compaq software with Compaq 10/100 Fast Ethernet Network Interface Controllers (NICs). With this combination, two NICs, or two ports in a single NIC, can be configured to be primary and backup paths for network communication; creating a redundant pair of network controllers. This feature is enabled with the Compaq Advanced Network Control Utility, which can be found on the Compaq Support Software Diskette for Windows NT (NT SSD). Following is a sample screen from the utility.

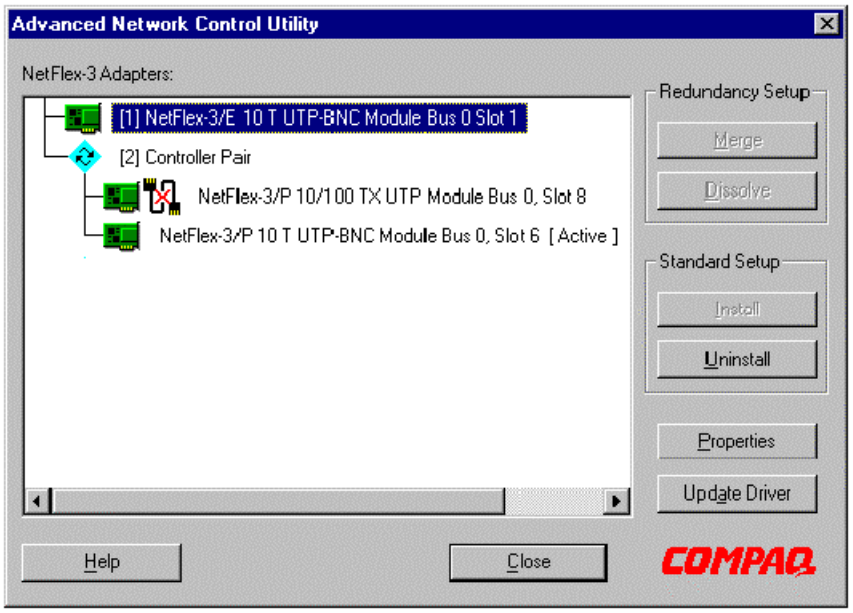

Figure 2: Illustration of an Active, Stand-Alone Controller in Slot 1 and a Redundant Pair in Slots 6 and 8

A quick summary of this feature can be found in Appendix A of this paper. For a complete description of the product, refer to the white paper entitled *Compaq Advanced Network Error Correction Support in a Microsoft Windows NT Server Environment.*

Redundant Network Interface Controller (NIC)

A redundant NIC can be defined as two physical network controllers configured such that one controller is in an active mode and the other is in a standby mode. When the active NIC fails, the redundant NIC then becomes active. The benefit of a redundant NIC is that the failover of the active NIC occurs instantaneously, and network traffic continues as if the originally active NIC were still processing data.

Network interface controller redundancy in a Windows NT environment is accomplished with software furnished by Compaq. This feature is available only with Compaq 10/100 Fast Ethernet products, and is enabled with the Compaq Advanced Network Control Utility.

Dual-Ported Network Interface Controller (NIC)

The majority of network interface controllers have a single-port with which a single network cable connects. Ordinarily, if two distinct network communication paths are needed from a single server, two NICs are placed in the server, and two expansion bus slots are used. A dual-ported NIC, however, has two ports, each of which supports its own network connection. Only one expansion bus slot is used. Compaq offers a dual-ported PCI NIC called the Netelligent Dual 10/100 TX PCI UTP Controller.

In the previous section, it was noted that redundant NICs could be configured with two separate network controllers. An exciting feature of the Compaq Advanced Network Control Utility is that it can be used to configure two ports of a dual-ported NIC to be redundant. In this configuration, one of the ports on a Netelligent Dual 10/100 TX PCI UTP Controller is configured as a hot backup for the other. The primary port will operate normally, sending and receiving data. Meanwhile, the second port remains in a standby state until the primary port encounters a failure. When the primary port encounters a failure, the standby port will take over. Therefore, no interruption of data flow is encountered.

Furthermore, the Advanced Network Fault Detection and Correction feature can be employed with any two NICs, regardless of whether the NICs are single-ported, dual-ported, or a combination. For example, assume a dual-ported NIC and a single-ported NIC reside in Node 1. A redundant NIC configuration can be made using one of the ports on the dual-ported NIC as the active port, and the port on the single-ported NIC as the standby port.

Interconnect Paths

Building redundancy into your cluster communication paths requires knowledge of interconnect paths. Two types of interconnect paths exist. A *private interconnect* (also known as a dedicated interconnect) is used exclusively for intra-cluster communication. A *public interconnect* (also known as a shared interconnect) not only takes care of communication between the cluster nodes, it handles cluster-to-LAN communication.

Use of a private interconnect precludes heavy cluster-to-LAN network traffic from diminishing the flow of important intra-cluster communication. Additionally, a private interconnect is easy to set up, maintain, and monitor.

There are two methods of physically creating a private interconnect. The first directly connects, using a crossover cable, the network controllers in each cluster node. A network hub is not required since the crossover cable plugs directly into each controller and enables comprehensible communication to occur between the NICs. The crossover cable appears to be a standard Ethernet cable, but it is not. The internal wiring of the cable differs from a standard Ethernet cable. You should consider labeling the crossover cable to distinguish it from a standard network cable. One crossover cable is included with each ProLiant Cluster kit.

The second physical interconnect utilizes a network hub, or even a series of hubs, repeaters, and switches. If the cluster nodes will be more than several meters apart, you will need to use this method. As long as both interconnect controllers reside on the same IP network, intra-cluster communication will occur over any combination of networking devices.

A public interconnect is not recommended as the primary path for intra-cluster communication, because cluster-to-LAN traffic can be heavy at times, and may interfere with node-to-node traffic.

Still, it is recommended that a public interconnect be configured for intra-cluster communication (as a redundant path) while a dedicated interconnect is created to serve as the primary path.

PCI Hot Plug

Another important building block in communication path redundancy is PCI Hot Plug technology. PCI Hot Plug is a new industry standard technology, which offers customers greater availability by eliminating both planned and unplanned downtime. This technology allows customers to remove and replace PCI expansion boards without having to power down the server or suspend any processes. The initial software release implementation of PCI Hot Plug support allows users to replace a failed board with an identical 'like for like' board.

Note: PCI Hot Plug is not supported on all Compaq servers; be sure to check the product specifications for your ProLiant server to see if PCI Hot Plug is available.

Compaq has implemented PCI Hot Plug so that each PCI bus slot can be controlled individually. This provides greater flexibility and availability, as service in adjacent slots is not effected if power is removed from one or more of the slots. In addition, greater availability is achieved through redundant failover capabilities available for certain Compaq PCI expansion boards. This function provides auto failover capabilities that allow a standby board to assume the workload, even while the failed board is being replaced.

PCI Hot Plug, when used in conjunction with redundant network interface controllers, brings even greater availability to your cluster communications. For example, assume you have two Compaq 10/100 Fast Ethernet controllers placed in slots that support PCI Hot Plug. The two controllers have been configured with the Compaq Advanced Network Control Utility to operate as redundant controllers. The primary controller encounters a failure, and network operation switches over to the standby (redundant) controller. At this point the primary controller can be physically removed from the cluster node without interrupting any operation of the cluster node. A new controller, the same model as the removed one, is placed in the open slot. Once the installation is complete, the newly installed controller becomes the standby for the currently active network controller. You are now back to a redundant configuration without having to power down the cluster node or disrupt client connections.

Microsoft Cluster Server allows the administrator to configure any certified network controller for intra-cluster communication, cluster-to-LAN communication, or both. Employing redundancy for the interconnect requires that at least two network controllers be configured, via Cluster Server, for intra-cluster communication.

The following picture is of the main screen Cluster Administrator. Notice how the *New Cluster Network* item under the *Networks* folder is highlighted. The user has right clicked on the selection to bring up the small menu to the right of the selection.

Figure 3: Screen Shot of Cluster Administrator's Network Controller Configuration Screen

Clicking on *Properties*, brings up the following picture of Cluster Administrator's Properties screen for network controllers. This is where controllers are configured for cluster-to-LAN use (depicted below as *Use only for client access*), for intra-cluster communication (depicted below as *Use only for internal cluster communications*), or for both (depicted below as *Use for all communications*).

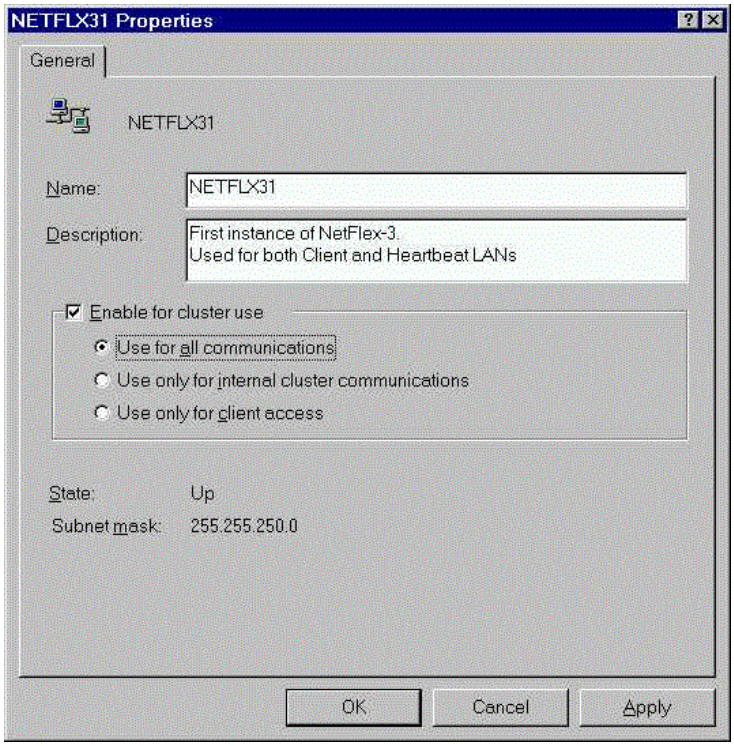

Figure 4: Cluster Administrator Properties Screen for Network Controllers

INTEGRATING THE BUILDING BLOCKS: COMPAQ RECOMMENDED STRATEGY

Now that each of the building blocks is understood, it is time to create solutions by integrating them. Recall the two issues described in the "Significance of Cluster Communication Paths" section. They are addressed again, only now each is described with a Compaq recommended strategy. In the next section, there are several example configurations that display the pros and cons of various integrations of the building blocks.

Compaq's minimum recommendation employs a single dedicated interconnect along with a single cluster-to-LAN network path. The dedicated interconnect is used only for intra-cluster communication. The cluster-to-LAN network path should be configured to handle client and intra-cluster communication.

Minimizing the number of failovers due to interconnect disruption is addressed first. At a minimum, Compaq recommends employing a single dedicated interconnect along with a single cluster-to-LAN network path. The dedicated interconnect is used only for intra-cluster communication. The cluster-to-LAN network path should be configured to handle client and intra-cluster communication. For a configuration that meets this recommendation, see Example 4.

It is important to note that when redundant interconnect paths are configured, each interconnect path must reside on its own unique IP network.

The second issue described in the "Significance of the Cluster Communication Paths" section, resulted in network clients being unable to access cluster groups. Access is denied because a physical interruption in the network path, from the clients to the cluster node containing their cluster groups, exists. Removing this single point of failure is accomplished by employing a redundant cluster-to-LAN network path. To accomplish this, the Compaq Advanced Network Fault Detection and Correction feature must be employed with either a dual-ported NIC, or with two single-ported NICs. Additionally, the interconnect issue stated above is again resolved by configuring the cluster-to-LAN network paths to handle client and intra-cluster communication.

Several configurations exist that will resolve both issues. The incremental cost of implementing one of these configurations is minor when compared to the savings accrued by reducing potential downtime. Although the configurations will be more complex, since they require multiple NICs and the use of the Compaq Advanced Network Fault Detection and Correction feature, in the long run the benefits outweigh the cost and configuration time. See Examples 5 and 6 for samples of integrated building blocks that posses excellent availability.

Example Configurations

Below, several solutions employing varying degrees of availability are discussed. These examples are intended to depict the various types of physical configurations that can be used for cluster communications. The examples begin with the least redundant configuration and gradually increase availability characteristics until the most redundant configurations are shown.

The primary goal of these examples is to educate the reader on how assorted network components and software configurations can impact the availability of cluster communication. A rating is given to each example to achieve a secondary goal, which is to let the reader know how well the configuration meets Compaq's Recommendation Strategy. The ratings are:

- Not Recommended
- Minimum Recommendation
- Suggested Recommendation
- Most Robust Recommendation

Example 1 – Not Recommended

This configuration consists of a single NIC. The NIC handles both types of communication; intra-cluster and cluster-to-LAN. No hardware or software redundancy exists for either the interconnect or the client-to-LAN data paths.

A failure of the network controller, a port in the network controller, or a network cable connected from the hub to either cluster node, will result in a cluster failover event. This temporarily interrupts client access to data files and applications. In most instances, clients will have access once the surviving server brings its cluster resources on-line and the clients reconnect to the cluster. Although clients will eventually regain access, minimizing the number of failover events will increase the overall availability of clustered data and applications. In the vast majority of cases, the benefit of purchasing and configuring a redundant NIC significantly outweighs the cost of the controller.

The purpose of creating a cluster is to increase the availability of your data files and applications. Using this configuration directly contradicts the effort by allowing easily removed single points of failure to exist in your cluster configuration.

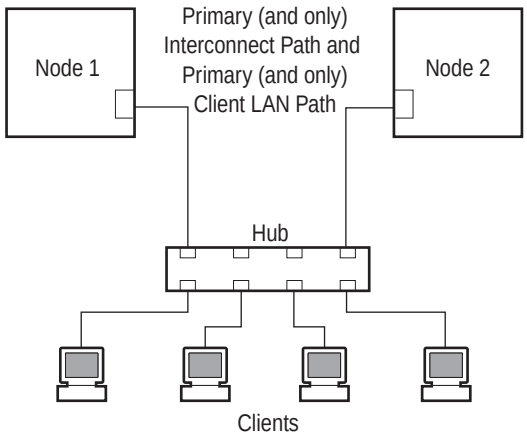

Figure 5: One Single-Port NIC Configuration

Redundancy Characteristics: Intra-Cluster Communication

1. None

Redundancy Characteristics: Cluster-to-LAN Communication

1. None

Point(s) of Failure: Intra-Cluster Communication

1. Failure of a network controller port used for the primary interconnect path is a single point of failure.
2. Failure of a network controller used for the primary interconnect path is a single point of failure.
3. Failure of the network cable(s) used for the primary interconnect path is a single point of failure.
4. Failure of the port(s) on the hub used for the primary interconnect path is a single point of failure.
5. Failure of the entire hub is a single point of failure.

Single Point(s) of Failure: Cluster-to-LAN Communication

1. Failure of a network controller port used for the primary cluster-to-LAN path is a single point of failure.
2. Failure of a network controller used for the primary cluster-to-LAN path is a single point of failure.
3. Failure of the network cable(s) used for the primary cluster-to-LAN path is a single point of failure.
4. Failure of the port(s) on the hub used for the primary cluster-to-LAN path is a single point of failure.
5. Failure of the entire hub is a single point of failure.

Example 2 – Not Recommended

This configuration makes use of the Compaq Advanced Fault Detection and Correction feature. Of the two ports available on a dual-ported NIC, one is configured to be the primary port and the other to be the backup port. The primary port actively sends and receives data. If the port should encounter a failure, or if the network segment to which it is attached should encounter a failure, the Advanced Network Fault Detection and Correction software will switch communication to occur over the backup port.

While this configuration offers greater availability than the single-port NIC described above, it has a noticeable flaw. Difficulties arise when either of the dual-ported NICs fails. If the failure of the NIC is not detected by Cluster Server, all applications running on the server with a failed NIC are unreachable by network clients. Furthermore, the applications are not failed over. If the failure of the NIC is detected by Cluster Server, then Cluster Server will failover all the clustered applications to the surviving server. This causes a brief interruption while the failover occurs and could result in slower overall performance since one server is now processing all clustered applications.

Another downside to this configuration is that both intra-cluster and cluster-to-LAN communication occur over the same network path. Although the intra-cluster communication is not intensive, it is an additional amount of traffic on the cluster-to-LAN network path. Furthermore, if the cluster-to-LAN network path is saturated, it could delay the transmission of intra-cluster information. The most extreme effect of this delay would be a failover event, caused by what the cluster would deem a loss of intra-cluster communication.

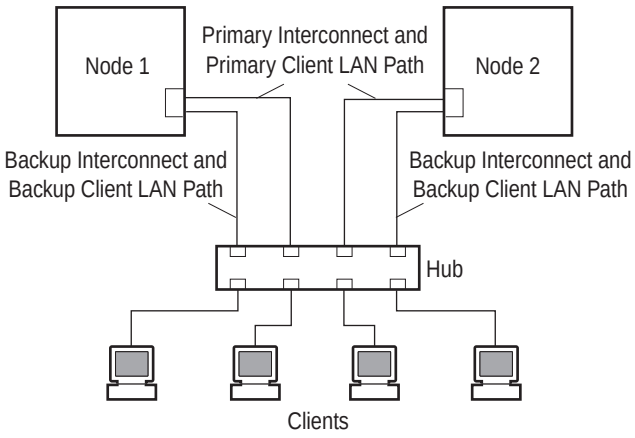

Figure 6: One Dual-Port NIC Configuration

Redundancy Characteristics: Intra-Cluster Communication

1. Remains available if a network controller port used for the primary interconnect path fails.
2. Remains available if the network cable(s) used for the primary interconnect path fails.
3. Remains available if the port(s) on the hub used for the primary interconnect path fails.

Redundancy Characteristics: Cluster-to-LAN Communication

1. Remains available if a network controller port used for the primary cluster-to-LAN path fails.
2. Remains available if the network cable(s) used for the primary cluster-to-LAN path fails.
3. Remains available if the port(s) on the hub used for the primary cluster-to-LAN path fails.

Point(s) of Failure: Intra-Cluster Communication

1. Failure of a network controller used for the primary interconnect path is a single point of failure.
2. Failure of the entire hub is a single point of failure.

Single Point(s) of Failure – Cluster-to-LAN Communication

1. Failure of a network controller used for the primary cluster-to-LAN path is a single point of failure.
2. Failure of the entire hub is a single point of failure.

Example 3 – Not Recommended

The configuration, from a hardware perspective, is very similar to the one described above. It uses a single dual-ported NIC but, in contrast to the above configuration, one of the ports is used as a dedicated interconnect, over which intra-cluster communication is passed. The second port is used as the primary, and only, cluster-to-LAN path. The second port is also configured, via Cluster Server, to serve as a backup interconnect.

From a software perspective, the Compaq Advanced Network Fault Detection and Correction feature is not enabled. Both ports are active simultaneously, and neither is a backup for the other.

When compared against the previous, similar configuration, the interconnect receives greater availability whereas the cluster-to-LAN communication path loses availability. The interconnect is now only affected by one single point of failure; the failure of an entire network controller. The cluster-to-LAN path is not as redundant as it was in the above configuration. The failure of an entire network controller, of the utilized network controller port, and of the utilized hub port will result in the failover of clustered applications to the surviving cluster node.

If you choose to disregard Compaq's recommendation and employ a configuration with a single dual-ported NIC, the availability is better when configured as shown in Example 2.

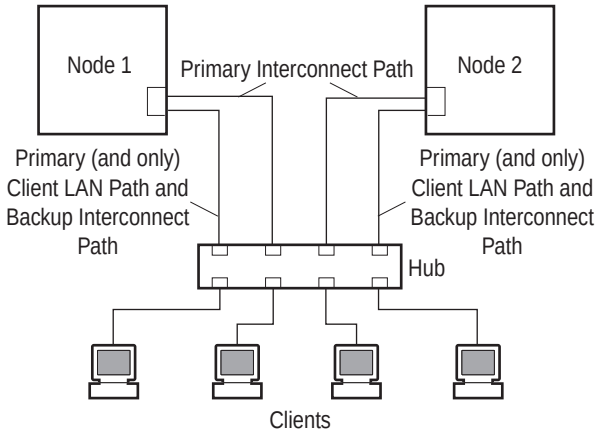

Figure 7: One Dual-Port NIC Configuration

Redundancy Characteristics: Intra-Cluster Communication

1. Remains available if a network controller port used for the primary interconnect path fails.
2. Remains available if the network cable(s) used for the primary interconnect path fails.

Redundancy Characteristics: Cluster-to-LAN Communication

1. None

Point(s) of Failure: Intra-Cluster Communication

1. Failure of a network controller used for the primary interconnect path is a single point of failure.

Single Point(s) of Failure: Cluster-to-LAN Communication

1. Failure of a network controller port used for the primary cluster-to-LAN path is a single point of failure.
2. Failure of a network controller used for the primary cluster-to-LAN path is a single point of failure.
3. Failure of the network cable(s) used for the primary cluster-to-LAN path is a single point of failure.
4. Failure of the port(s) on the hub used for the primary cluster-to-LAN path is a single point of failure.
5. Failure of the entire hub is a single point of failure.

Example 4 – Minimum Recommendation

This configuration consists of two single-port NICs. The first NIC is dedicated for intra-cluster communication. The second NIC is the primary, and only, path for cluster-to-LAN communications. The second NIC is also configured via Cluster Server to be the backup interconnect path. The Compaq Advanced Network Fault Detection and Correction feature is not enabled.

In this configuration the interconnect receives excellent availability. The cluster-to-LAN communication path has several single points of failure, and the interconnect is fully redundant. The failure of any component of the dedicated interconnect results in the switching of intra-cluster communications to the cluster-to-LAN network path.

The failure of an entire network controller, of the utilized network controller port, or of the utilized hub port, can interrupt the cluster-to-LAN communication. Any of these failures will result in the failover of clustered applications to the surviving cluster node.

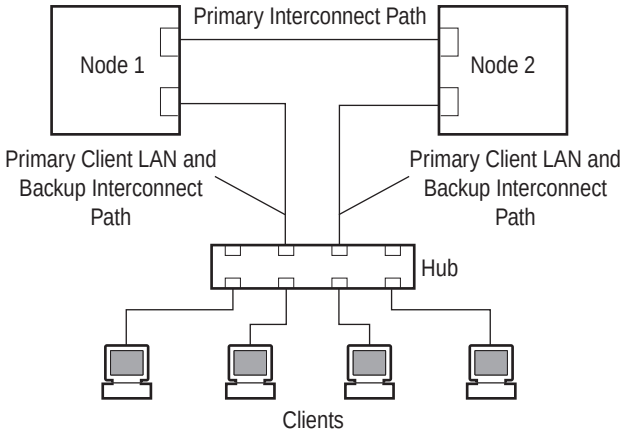

Figure 8: Two Single-Port NICs Configuration

Redundancy Characteristics: Intra-Cluster Communication

1. Remains available if a network controller port used for the primary interconnect path fails.
2. Remains available if a network controller used for the primary interconnect path fails.
3. Remains available if the network cable(s) used for the primary interconnect path fails.

Redundancy Characteristics: Cluster-to-LAN Communication

1. None.

Point(s) of Failure: Intra-Cluster Communication

1. None.

Single Point(s) of Failure: Cluster-to-LAN Communication

1. Failure of a network controller port used for the primary cluster-to-LAN path is a single point of failure.
2. Failure of a network controller used for the primary cluster-to-LAN path is a single point of failure.
3. Failure of the network cable(s) used for the primary cluster-to-LAN path is a single point of failure.
4. Failure of the port(s) on the hub used for the primary cluster-to-LAN path is a single point of failure.
5. Failure of the entire hub is a single point of failure.

Example 5 – Suggested Recommendation

This configuration, from a hardware perspective, is similar to the one described above. However, in implementation it differs significantly. As you will see, it removes the solitary cluster-to-LAN single point of failure without adversely affecting intra-cluster communications.

The dual-port NIC is essentially split. The first port is a dedicated interconnect. The second port is the backup redundant client cluster-to-LAN network path. The single-port NIC is configured as the primary path for cluster-to-LAN communication. The Compaq Advanced Network Control Utility

is used to configure the second port on the dual-port NIC as the backup port of a redundant pair. The single-port on the other NIC is configured to be the primary port of the redundant pair.

The interconnect retains its fully redundant status when Cluster Server is configured to use the other network ports as backups for the interconnect. Failure of the primary interconnect path results in intra-cluster communications occurring over the single NIC. Should the entire dual-port NIC fail, the cluster nodes will still have a working communication path over the single-port NIC.

The cluster-to-LAN communication becomes even more redundant than the configuration described just above. Above, the solitary single point of failure was the failure of the dual-port NIC. With this configuration, even a failure of the dual-port NIC results only in the transfer of the intra-cluster communication to the single-port NIC. Other than the network hub encountering a failure, any failure of any cluster network component will be resolved by the redundancy of this configuration.

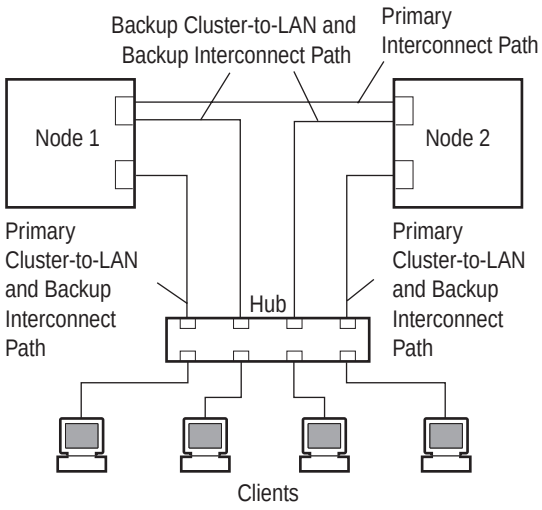

Figure 9: One Single-Port, One Dual-Port Configuration

Redundancy Characteristics: Intra-Cluster Communication

1. Remains available if a network controller port used for the primary interconnect path fails.
2. Remains available if a network controller used for the primary interconnect path fails.
3. Remains available if the network cable(s) used for the primary interconnect path fails.

Redundancy Characteristics: Cluster-to-LAN Communication

1. Remains available if a network controller port used for the primary cluster-to-LAN path fails.
2. Remains available if a network controller used for the primary cluster-to-LAN path fails.
3. Remains available if the network cable(s) used for the primary cluster-to-LAN path fails.
4. Remains available if the port(s) on the hub used for the primary cluster-to-LAN path fails.

Point(s) of Failure: Intra-Cluster Communication

1. None.

Single Point(s) of Failure: Cluster-to-LAN Communication

1. Failure of the entire hub is a single point of failure.

Example 6 – Most Robust Recommendation

This configuration consists of two NICs, each dual-ported. As in Example 4, the use of the dual-ported NICs is split. The top port of the first NIC is used as a dedicated interconnect, and rather than utilizing a crossover cable, it is connected through a network hub. The bottom port of the first NIC is used as the backup path for cluster-to-LAN communication. The second NIC (the lower one in the drawing) is configured so that its top port is the backup dedicated interconnect and its bottom port is the primary cluster-to-LAN path.

The Compaq Advanced Network Control Utility is used to configure a single-port on each NIC as a primary/standby pair. The top port of the first NIC is paired with the top port of the second NIC. The port on the first NIC is configured as the primary port, while the port on the second NIC is configured as the standby. Similarly, the bottom port of the second NIC is configured to be the primary port in its pair, and the bottom port of the first NIC is set up to be the standby.

This configuration should use a network hub for the interconnects. If a hub is not used and a dedicated (crossover) cable is used to connect each controller, the failure of a single-port on the controller in Node 1, would render that interconnect path inoperable. However, if the port on the controller in Node 2 had not encountered a failure, it would continue to send information to the failed port in Node 1. Eventually, Node 2's NIC would determine that the primary path is no longer working and would switch interconnect communication to the second port, which defeats the purpose of using redundantly configured NIC ports.

Failure of the primary interconnect path results in intra-cluster communications occurring over the redundant dedicated interconnect path. When Cluster Server is configured to send intra-cluster communication over the LAN NIC as well, the cluster nodes will still have a working communication path even if one of the dual-ported NICs fails.

Other than the network hub used for client communication encountering a failure, any failure in the cluster-to-LAN path should be resolved by the redundancy of this configuration.

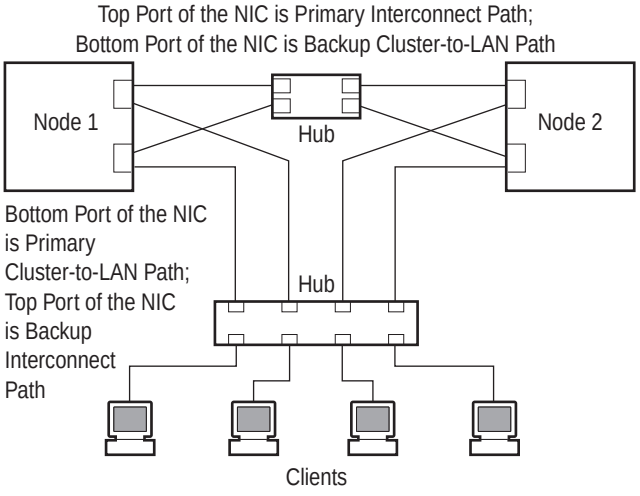

Figure 10: Two Dual-Port NICs

Redundancy Characteristics: Intra-Cluster Communication

1. Remains available if a network controller port used for the primary interconnect path fails.
2. Remains available if a network controller used for the primary interconnect path fails.
3. Remains available if the network cable(s) used for the primary interconnect path fails.
4. Remains available if the port(s) on the hub used for the primary interconnect path fails.
5. Remains available if the hub used for the primary interconnect path fails.

Redundancy Characteristics: Cluster-to-LAN Communication

1. Remains available if a network controller used for the primary cluster-to-LAN path fails.
2. Remains available if a network controller port used for the primary cluster-to-LAN path fails.
3. Remains available if the network cable(s) used for the primary cluster-to-LAN path fails.
4. Remains available if the port(s) on the hub used for the primary cluster-to-LAN path fails.

Point(s) of Failure: Intra-Cluster Communication

1. None.

Single Point(s) of Failure: Cluster-to-LAN Communication

1. Failure of the entire hub is a single point of failure.

OTHER INFORMATION

Integrated NIC

All Compaq ProLiant Servers ship with a network interface controller. In some servers, the NIC is integrated onto the motherboard of the computer. In other cases, the NIC is a separate controller placed in one of the expansion bus slots.

When configuring your cluster's network communications, it is recommended that, whenever feasible, you utilize the NIC shipped with the server. The reason for this recommendation is simply that you use one less expansion bus slot than if you placed another NIC in the server.

If you will be using a configuration that requires the Compaq Advanced Fault Detection and Correction feature, check to see if your integrated NIC supports it. If it does you can use the integrated NIC as part of a redundant pair. If it does not, you should consider using the NIC as part of a dedicated interconnect or dedicated cluster-to-LAN network path.

Ethernet Crossover Cable versus Ethernet Hub

When configuring a dedicated Ethernet interconnect, two cabling methodologies exist. In the first, an Ethernet crossover cable can be used to directly connect one Ethernet controller in cluster Node 1 to the corresponding Ethernet controller in cluster Node 2. The crossover cable is included with the ProLiant Cluster.

The alternative is to use an Ethernet hub. If an Ethernet hub is used, it is recommended that the hub be dedicated to the interconnect. That is, only two cables are plugged into the hub, one cable from the interconnect controller in cluster Node 1 and the other from the interconnect controller in cluster Node 2. No benefits have been found in using a hub, and it is therefore a matter of personal preference as to whether one is employed or not.

ServerNet

The Tandem ServerNet controller is a bi-directional, high-bandwidth, low-latency, redundant path interconnect. It is designed as a dedicated interconnect. Its use is for cluster configurations that require the most rigorous and fastest intra-cluster communication.

The ServerNet controller is a PCI card that uses a single PCI slot and is dual-ported. The controllers are connected directly to each other via special cables that ship with the ServerNet card.

Note: The ServerNet controller cannot be used as part of the Compaq Advanced Network Fault Detection and Correction feature.

Throughout this document you will see references to and figures of dedicated interconnects, which are generally shown as Ethernet interconnects. ServerNet is an alternative to dedicated Ethernet interconnects. The following figures show a comparison of how a cluster communications configuration would appear with a dedicated interconnect and then with a ServerNet controller.

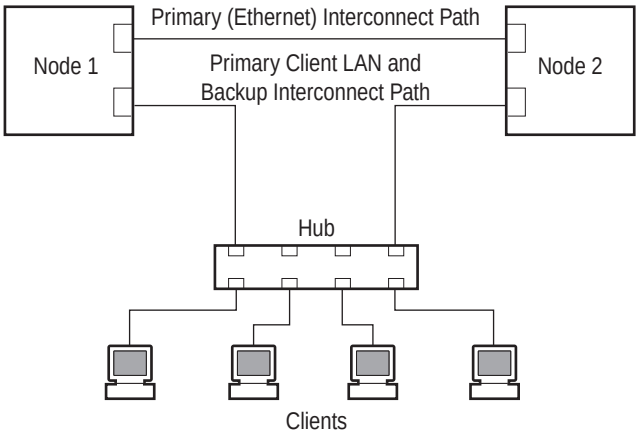

Figure 11: Dedicated Ethernet Interconnect

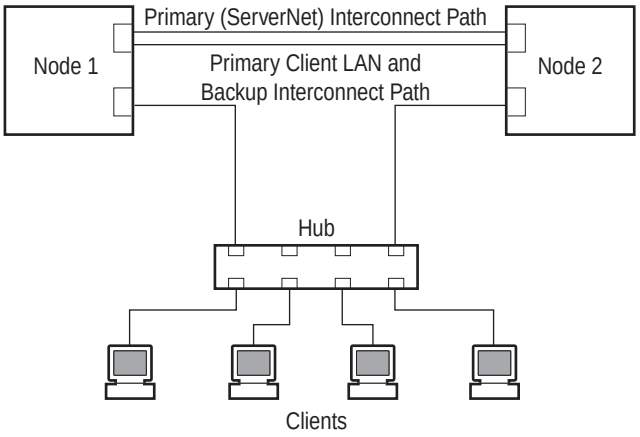

Figure 12: Dedicated ServerNet Interconnect

Configuring an Interconnect for Cluster Nodes Separated by Significant Distance

Depending on the model of ProLiant Cluster, you may have the ability to separate the cluster nodes by distances of up to 1000 meters. To do so, both the cluster storage system and the cluster interconnect must support a separation of this distance.

Microsoft Cluster Server requires both cluster nodes to exist on the same IP network.

Many methods can be employed to configure an Ethernet interconnect: using hubs, routers, switches, and repeaters. Any of these devices can be used with Microsoft Cluster Server as long as both cluster nodes exist on the same IP network. This is a Cluster Server requirement.

Note that ServerNet supports distances between nodes of up to 30 meters. If you require a greater distance between nodes, you should use Ethernet for your interconnect.

Additional Suggested Reading

- Compaq Advanced Network Error Correction Support in a Microsoft Windows NT Server Environment (first edition November 1996. Doc #114A/1196)
- Compaq Advanced Network Error Correction Support using PCI Hot Plug with Windows NT (first edition August 1997. Doc # ECG057.0897)
- Deploying PCI Hot Plug on Compaq Servers in a Microsoft Windows NT Environment (first edition July 1997. Doc # 064A/0797)

APPENDIX A

The Compaq Advanced Network Error Detection and Correction Feature

Overview

Compaq has implemented a two-part, software-based solution to ensure network availability on Compaq server networks. The first part of the solution is provided by the Compaq NetFlex-3 device driver, which includes built-in network fault protection features, such as recovery from hardware failure events, network status events, and errors.

The second part of the solution is provided by the Compaq Advanced Network Control Utility. This utility allows for the designation of a secondary controller that acts as a standby, and assumes network responsibilities if the primary controller fails. This secondary controller is idle until a failure occurs with the primary controller.

Figure 12 illustrates how the NetFlex-3 driver and the Advanced Network Control Utility provide network fault protection in Windows NT.

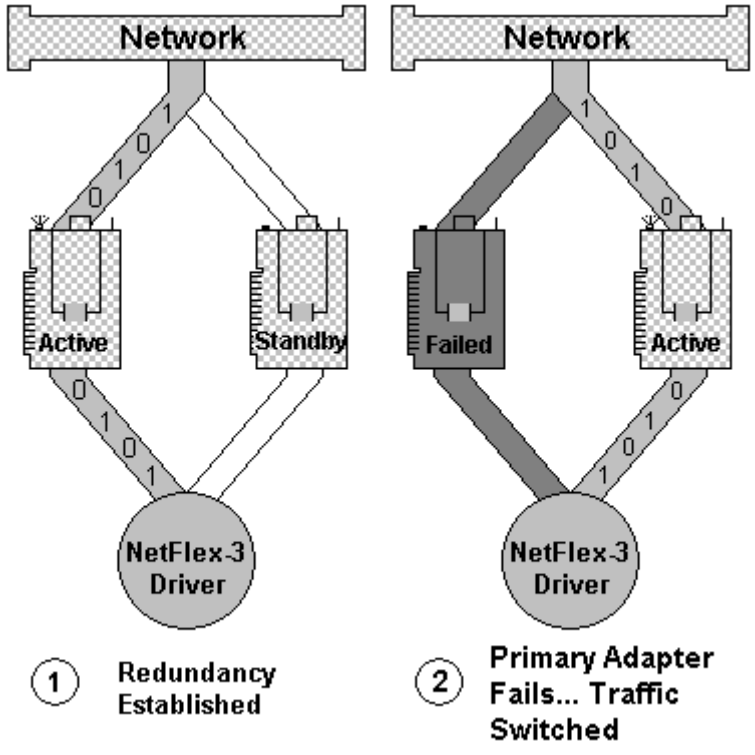

Figure 12: Controller Pair Illustration: "Active Primary/Idle Standby" and "Idle Primary/Active Standby" After a NetFlex-3 Switchover Occurs

When you configure a controller pair on a Compaq server, you are providing an added level of network fault tolerance protection. A single NetFlex-3 driver instance controls both the primary and secondary controllers when the Advance Network Control Utility is installed. When the driver detects failures on the primary controller, and the secondary controller has a valid link, network traffic is switched to the secondary controller. The default operating mode is *Switch on Fail*. The switch occurs without any interruption to the operating system or network services.

The Windows NT operating system views the controller pair as a single controller. In a controller pair, if a primary network controller fails, the Media Access Control (MAC) address of the primary controller transfers to the secondary controller. At that time, the driver switches the MAC address from the primary controller to the secondary controller. The MAC address is a unique address assigned to a particular network controller. The MAC address identifies that network controller on the network. The secondary controller automatically switches to the active controller and takes control of the traffic on the network. The entire process occurs without the intervention of network protocols or operating system applications.

Controller Pairing

Creating a controller pair with the Advanced Network Control Utility is easy, and you can immediately view the results in the utility. A controller pair consists of two Compaq NetFlex-3 or Netelligent Controllers. The paired controller configuration provides connectivity over the network so that the network maintains operation even if a controller connection is lost or if one of the controllers fails.

Although a controller pair configuration is installed and monitored with one instance of the NetFlex-3 driver, the driver is always aware that there are two physical controllers. One controller is active and the other non-active controller is a backup (standby). The NetFlex-3 driver controls both controllers and can automatically switch traffic if one controller fails. The operating system views the controller pair as a single controller and the switch over from one physical controller to the other is completely transparent to the operating system and any application programs.

Each controller pair consists of a primary and a secondary physical controller. Only one of these controllers can be active at a time. When the system starts up, the driver tries to direct network traffic over to the primary controller, so the primary controller is usually the *active* controller. Should the primary controller fail or lose connection, the secondary or *standby* controller can become active depending on the operating mode of the controller pair.

For detailed information about the tremendous capabilities of this feature refer to the Compaq White Paper entitled *Compaq Advanced Network Error Correction Support in a Microsoft Windows NT Server Environment*