New Availability Features of Local Area VAXcluster Systems

By Lee Leahy

Abstract

VMS version 5.4-3 increases the availability of local area VAXcluster (LAVc) configurations by allowing the use of multiple local area network (LAN) adapters in the VAXcluster system. Availability is increased by enabling fail-over between LAN adapters, reducing channel failure detection time, and providing better network troubleshooting. Combined, these changes significantly increase the availability of LAN-based VAXcluster configurations by allowing the VAXcluster system to tolerate and work around network failures.

This paper describes the availability features added to local area VAXcluster (LAVc) support in VMS version 5.4-3. These features support multiple local area network (LAN) adapters, reduce the time required to detect network path (channel) failures, and provide additional support for network troubleshooting. (Table 1 presents definitions for terms used throughout the paper.)

Table 1

LAVc Terminology

Channel	A data structure in PEDRIVER that represents a network path (see network path below). Each channel is associated with a single virtual circuit (VC).
Datagram	A message that is requested to be sent by the client of the LAN driver. A datagram does not have guaranteed delivery to the remote system. The datagram may never be sent, or could be lost during transmission and never received.

LAN Adapter	An Ethernet or fiber distributed data interface (FDDI) adapter. Each type of LAN adapter has a unique set of attributes, such as the receive ring size.

Table 1 (Cont.)

LAVc Terminology

LAN Address	The network address used to reference a specific LAN adapter connected to the Ethernet or FDDI. This address is displayed as six hexadecimal bytes separated by dashes, e.g., 08-00-2B-12-34-56.
LAN Segment	An Ethernet segment or FDDI ring. Each type of LAN has a unique set of attributes, e.g., maximum packet size. LAN segments can be connected together with bridges to form a single extended LAN. However, in such a LAN, the LAN segments can have different characteristics (e.g., different packet sizes for an FDDI ring bridged to an Ethernet).
Network Path	The pieces of the physical network traversed when a datagram is sent from one LAN address to another LAN address. The network path is represented by a pair of LAN addresses, one for the local system and one on the remote system. Each network path has a specific set of attributes, which are a combination of the attributes of the local LAN adapter, the remote LAN adapter, and each of the LAN segments and LAN devices on the path between them.
PEDRIVER	The VMS port driver that provides reliable cluster communication utilizing the Ethernet.
Virtual Circuit	A data structure in PEDRIVER that represents the data path between the local system and the remote system. This data path provides guaranteed delivery for the messages sent. PEDRIVER's datagram service, along with an error recovery mechanism, ensures that a message is delivered to the remote system or is returned to the client with an error. A virtual circuit (VC) has one channel for each network path to the remote system.

New Availability Features of Local Area VAXcluster Systems

We begin the paper with an overview of the added LAVc availability features of VMS version 5.4-3. We then present the multiple-adapter support features of the new release, with comparisons to the previous single-adapter implementation. The detection of network delays is discussed, along with how the system selects alternate paths around these delays after detection. Finally, we discuss the analysis of network failures and the physical descriptions needed to achieve the proper level of failure reporting.

Added Availability Features

VMS version 5.4-3 supports LAVc use of up to four LAN adapters for each VAX system. Availability and performance are increased by connecting each LAN adapter to a different LAN segment. Maximum availability is achieved by redundantly bridging these LAN segments together to form a single extended LAN. This configuration maximizes availability and reduces single points of failure by increasing the number of possible network paths between the different systems within

to the remote system. If not acknowledged within 2 seconds, a datagram is retransmitted. Retransmission continues until the connection between the two systems is declared broken. However, applications can be stalled during this error recovery process. Therefore, reducing the time for detecting channel failures and retransmitting datagrams reduces the amount of application delay introduced by network problems.

VMS version 5.4-3 also increases availability by reducing the delays introduced by network congestion. This latest release measures the network delays on a channel basis. The channel with the lowest computed network delay value is used to communicate with the remote system.

LAVc network failure analysis is a new feature in VMS version 5.4-3. This feature provides an analysis of failing channels by isolating the common network components responsible for the channel failures. LAVc network failure analysis increases availability by reducing the downtime caused by failing network components.

the VAXcluster system. Availability has also been increased at the applications level by reducing the time required to detect channel failures.

The LAVc protocol (NISCA) sends sequenced datagrams

To enable this feature, the system or network manager must provide an accurate physical description of the network used for LAVc communications.

New Availability Features of Local Area VAXcluster Systems

Multiple-Adapter Support

This section describes the availability features added with the multiple-adapter LAVc support in VMS version 5.4-3. Some limitations of the single-adapter implementation are presented for comparison.

Single Points of Failure

In single-adapter LAVc satellites, the Ethernet adapter remains as a single point of failure. This failure "point" actually extends through the network components common to all of the network paths in use for cluster communication. The combination of VMS version 5.4-3 with multiple LAN adapters removes the LAN adapter as a single point of failure in the local system. Additionally, the use of multiple LAN adapters connected to an extended LAN creates multiple network paths to remote systems. This configuration results in a higher tolerance for network component failures and higher cluster availability.

Adapter Selection

The single-adapter implementation is configuration-dependent and does not allow the system manager a choice of adapters. The multiple-adapter support in VMS stop the LAVc protocol on the LAN adapters. This support allows the system manager to select which LAN adapters will run the LAVc protocol.

The means of locating the LAN devices in the system has also changed. The system now maintains a list of LAN devices. As each LAN device driver is loaded into the system, an entry is appended to this list. A new support routine steps through this list and returns a pointer to the next LAN device in the system. The single-adapter implementation requires code changes in PEDRIVER to add a new LAN device; the new implementation no longer requires these changes.

Channel Control Handshake

The channel control handshake is a three-way message exchange. The exchange starts when a HELLO message is received from a remote system and the channel is in the closed state, or any time a CCSTART message is received. Upon receiving a HELLO message on a closed channel, the system responds with a CCSTART message.

Upon receiving a CCSTART message, the system closes the channel if the PATH bit was set. In all cases,

version 5.4-3 configures the system for maximum availability by starting the LAVc protocol on all LAN adapters in the system. Support is also provided to start and

if the cluster password is correct, the system responds with a VERF message. Upon receiving the VERF message, the remote system verifies the cluster password. If

the password is correct, the remote system sends an acknowledgment (VACK) message and marks the channel as usable by setting the PATH bit. The local system, upon receiving the VACK message, also marks the channel as usable by setting the PATH bit.

The channel control handshake now verifies the network path used by this channel, instead of verifying the virtual circuit (VC) as in the single-adapter implementation. Additionally, the handshake is used to negotiate some parameters between the local and remote systems on a channel basis (instead of assuming that the parameters are common for all channels connected to the VC).

Packet size and pipe quota are two characteristics that are now arbitrated between the two systems. These parameters are negotiated on a channel-by-channel basis to allow different channels to fully utilize the capabilities of the specific network path.

With the introduction of FDDI, larger packet sizes are now supported. The channel handshake between two nodes negotiates packet size of 4468 bytes or smaller. An increased packet size reduces the number of messages required when large blocks of data are sent. This increase in packet size results in fewer messages, less handshaking, and thus better network efficiency.

The PIPE_QUOTA value is used to limit the number of messages sent to the remote system before waiting for an acknowledgment. PIPE_QUOTA was implemented to help prevent receiver overrun on the remote system. Instead of using a fixed value, the new implementation uses a value specified by the LAN driver. This value factors in the LAN device's receive ring size and is typically larger than the fixed value of eight messages used previously. Increasing the PIPE_QUOTA value allows more data to be sent between the nodes before an acknowledgment message is required, thus increasing the protocol's efficiency and reducing the network traffic.

These new features in VMS version 5.4-3 have reduced the amount of handshaking required to move data and the number of messages required to move large amounts of data. The result

a packet size that is supported by the entire network path. Any path that utilizes an Ethernet must use a packet size of 1498 bytes or smaller. An FDDI-to-FDDI path on the same extended ring must use a

is greater applications availability through fewer network-based delays.

Use of Hello Messages

The single-adapter implementation uses a HELLO message to maintain the VC and not the channels. Also, the handshake to verify connectivity is performed by the VC, which forces all channels to use the same characteristics. In comparison, the multiple-adapter implementation uses HELLO messages to trigger the channel handshake, test the network path and maintain the channel in the open state, and continuously verify the network topology.

To maintain the channel and test the network path, each system multicasts a HELLO message through each of its LAN adapters every 3 seconds. Upon receipt of a HELLO message (if the channel is not open), a channel handshake begins. If the channel is open, the network delay is computed and the channel packet size is verified. When an open channel does not receive a HELLO message within 8 seconds, it declares a listen time-out and the channel is closed.

Additional topology change detection is required because FDDI-to-FDDI communications use large packets. If two systems using FDDI adapters exchange channel control messages, then both can

separated by an Ethernet segment.)

Detection of the dumbbell configuration is performed using the priority field in the frame control byte of the FDDI message header. This field does not exist in Ethernet messages and must be created when forwarding an Ethernet message to an FDDI ring. Ethernet-to-FDDI LAN bridges set this field's value to zero. All LAVc messages transmitted by the FDDI adapters use a non-zero value for the

priority field. When a channel control message is received, the value of this field is checked. If the value is non-zero, then large messages can be used because the message did not traverse an Ethernet segment.

The priority field is also verified every time a HELLO message is received and the channel is open. A topology change is detected when a change in the priority value is received. If the priority value goes from zero to non-

zero, the packet size is renegotiated and a larger packet size may be used. If the priority value goes from non-zero to zero, the channel packet size must be reduced. If this is the only channel with a larger

agree on a large packet size. However, if the network is configured in the dumbbell configuration, then only the small packet size can be used. (The dumbbell configuration consists of two FDDI rings

packet size, then the VC closes and forces the two systems to reassign the message sequence numbers.

Listen Time-out

VMS version 5.4-3 now consistently times out channels in 8 to 9 seconds, whereas the single-adapter implementation detects the failure in 8 to 15 seconds. Reducing this time reduces the delays experienced by applications when a LAVc node is removed from the cluster. The result is an increase in applications availability.

The single-adapter implementation traverses the VC list and scans each of the receive channels (RCH structures embedded in the VC) to check for time-out. Because this scan is CPU-intensive, the algorithm was designed to scan the VC list only once every 8 seconds. Reducing this scan time required the design of a new algorithm that reduces the CPU utilization required to locate the channels that have timed out.

The VMS version 5.4-3 implementation places each open channel into a ring of time-out queues. The time-out routine maintains a pointer into the ring of queues corresponding to the 8-second time-out. Each second, the time-out routine executes, removes any channels pointed to by receiving HELLO messages are inserted into the ring of queues pointed to by the current time pointer, which prevents them from timing out. This implementation reduces CPU utilization during the time-out scan by looking at only the channels that have timed out.

Changes to Virtual Circuit Maintenance

The single-adapter implementation closes the VC and performs a channel control handshake every time a new channel is established. This implementation also forces each channel to use the same characteristics, specifically packet size, thereby reducing the characteristics to the lowest common denominator.

VMS version 5.4-3 does not close the VC each time a new channel is established. The channel handshake affects only the channel and is used to negotiate the channel characteristics, including packet size. The VC remains open as long as a channel with the corresponding packet size is open. This maintenance increases applications availability by allowing channels to fail and

the time-out pointer, and calls the listen time-out routine for the channel. Next, the time-out pointer and the 8-second time-out pointer are updated to point to a new set of queue headers in the ring. Active channels and channels reestablish transparently without disrupting service at the VC and systems communication services (SCS) layers.

One Channel with Matching Characteristics Required. The VC can be opened as soon as the first channel to the remote system is opened. When the VC opens, its packet size is set to the packet size of the channel being used. The VC can remain open as long as at least one channel with a compatible packet size is open. The packet size is compatible if the channel packet size is greater than or equal to the packet size currently in use by the VC.

Transfers restricted to an FDDI ring can use a larger packet size than those that traverse an Ethernet LAN segment. PEDRIVER now supports variable packet sizes up to the size supported for the FDDI ring. Each time the VC switches channels, the new channel characteristics are copied into the VC. The result is that as soon as the VC switches to using the FDDI-to-FDDI channel, it also switches to using the larger packet size.

Receive Message Caching. VMS version 5.4-3 introduces a receive message cache to prevent any performance degradation when messages are received out of order. Because of transmission and network delays, messages are typically received out of order.

Channel Failure Not Displayed. The multiple-adapter implementation does not display any messages when a channel fails. This choice was made to maintain compatibility with the previous implementation. We also wished to reduce the number of console messages and still provide enough data to isolate the problem. However, without some channel failure notification, all but one channel could fail without notice, thus negating all the availability that was introduced by using multiple adapters.

The LAVc network failure analysis allows the system or network manager to select one of the following levels of channel failure notification: no notification, if not enabled; channel failure notification, when barely enabled; or isolation of the failing network component, when fully enabled. When this feature is fully enabled, a failing network component typically generates only a single console message instead of displaying tens or hundreds of channel failure messages.

Channel Selection

of order at approximately the time a channel switch occurs. Also, most of the channel selections are invoked after locating a channel with a lower network delay value, thus increasing the probability

VMS version 5.4-3 bases its selection of a single transmit channel for a remote system first, on the packet size and second, on the network delay value. The channel selection algorithm searches for

an open channel with a compatible packet size so that the VC does not have to be broken. If more than one channel has a compatible packet size, the network delays are compared and the channel with the lowest network delay value is chosen. The selected channel is used until it fails, encounters an error, or a channel with a lower network delay value is found.

Channel selection is performed independently for each remote system. This implementation means that a two-node cluster increases its availability through the use of more

LAN adapters, but does not achieve a performance benefit by increasing the number of LAN adapters above two. Larger clusters, however, can take advantage of the additional LAN adapters and thus achieve better cluster performance. Multiple LAN adapters can also increase the bandwidth available for use by the LAVc protocol. However, the actual performance is very configuration- and application-dependent.

Channel selection is limited to the transmit channel, but all channels are used to receive data. The receive cache helps

selection algorithm, e.g., in PEDRIVER or in any component implementing NISCA.

Multiple-adapter Availability Summary

The multiple-adapter LAVc support added to VMS version 5.4-3 increases the availability of applications and of the overall cluster. Availability is increased by removing the LAN adapter as a single point of failure. Cluster availability is enhanced through continuous testing of the network paths and correction for network topology changes.

This implementation also increases network utilization and cluster performance by taking full advantage of a channel's characteristics. Larger receive ring sizes reduce the protocol handshaking overhead. Moreover, larger packet sizes reduce the number of messages that must be sent for large transfers.

The next section discusses how the PEDRIVER detects network delays and selects

alternate paths.

Network Delay Detection

VMS version 5.4-3

prevent retransmission by the remote system by placing messages received out of order into the receive cache until the previous messages are received. This receive algorithm is compatible with any transmit channel

increases application availability by detecting significant network delays and selecting alternate paths. As the network gets busy, it becomes more difficult for a LAVc node to send cluster messages. These delays in network

communications cause delays in cluster traffic and translate into delays in the applications. Thus, through delay detection and the use of alternate paths, VMS version 5.4-3 reduces the delays for applications and increases overall cluster performance.

Assumptions and Delay Calculations

PEDRIVER computes network delays through a series of assumptions. The primary assumptions are that the transmit and receive delays for a path are equal, and that there are small internal delays associated with the LAN device. Although these assumptions are occasionally invalid, PEDRIVER uses them because there are no round-trip messages available in the NISCA protocol to compute the delay.

As the first step in the delay calculation for each channel between nodes, each node time-stamps the HELLO message just prior to transmission. When the HELLO message is received, the time stamp is subtracted from the local system time. This resulting value equals the sum of the transmit queue delay, the network delay, the receive queue delay, and the difference in the two system times.

The second step of the delay calculation is to compare the delay times between different channels to the same remote system. This comparison is a subtraction of the values computed above for each channel. The computation removes the common factor (the difference in the two system times) and results in the comparison of the two network delays. When multiple channels exist, PEDRIVER attempts to use the channel with the lowest network delay value.

Problems and Benefits Associated with the Assumptions

The assumptions in the network delay calculation do not always hold true. The arbitration delay to transmit a message on the Ethernet, between a pair of systems, is not always equal in both directions. Over the long term, this assumption would be valid if the systems are sending the same number of messages in each direction; however, this is not typically the case. When this assumption does not hold true, i.e., if the transmit delay is longer than the receive delay, then additional delay is introduced when transmitting messages using this channel.

The assumption that

Applying the assumptions reduces this value to the sum of the network delay and the difference in the two system times.

internal delays are small depends upon the network traffic and the transmit traffic generated for an adapter by the other LAN clients. If another LAN client is a heavy user of

a particular LAN adapter, then transmissions from this adapter experience additional queue delays while waiting for the adapter. If the network is busy, messages in the transmit queue have an additional wait.

Finally, the network delay computed is the delay from the remote system to the local system. Since the delay is not always symmetric, it does not always represent the delay in the other direction, i.e., transmitting messages to the remote system. Yet, because the NISCA protocol does not have any round-trip messages, this is the best possible delay value.

Even with these problems in the assumptions, the network delay calculations increase the availability of the cluster by detecting large network delays. With this data, PEDRIVER is usually able to select alternate paths around the network delays when multiple channels exist, providing better cluster performance and availability.

Figure 1 represents an example of network delay detection. If LAN segment A is very busy, then PEDRIVER can detect an additional network delay for channels A1-B1, A1-B2, and A2-B1. PEDRIVER can then select an alternate path, that is, transmit packets only on channel A2-B2. Use of channels A1-B1, A1-B2, and A2-B1 can resume when the network traffic level on LAN segment A is reduced to about the level of LAN segment B, or if channel A2-B2 fails.

LAVc Network Failure Analysis

VMS version 5.4-3 uses multiple LAN adapters to increase availability by working around network delays and failures. Channels fail as network failures occur, reducing the availability provided by these extra channels. However, the VC remains open, allowing cluster helping to locate the cause of the failure. Also, as the cluster configuration gets larger, or the number of LAN adapters increases, channel failure messages increase (depending on what component failed) beyond the point where they are helpful. Yet to maintain cluster availability, the system or network manager

communication as long as a single channel remains open.

To maintain compatibility with previous VMS versions, only VC failures are displayed on the local console. Displaying messages about channel failures would only indicate a problem without needs to be told of the channel failures that are reducing the availability.

The LAVc network failure analysis, introduced with VMS version 5.4-3, is used to analyze the network failures and display the OPCOM messages that call out the failing network component. This support

requires a description of the physical network used for LAVc communications. Depending upon the description supplied, the system or network manager can select the level of failure reporting. This level may range from channel failure reporting to calling out the actual component that failed.

Display of Channel Failures

There is a significant difference between displaying the channel failures and performing LAVc failure analysis. This difference is shown in Figure 2, which represents a multiple-adapter LAVc configuration.

Looking from system VAX A, the following channels exist: A1-A2, A2-A1, A1-B1, A1-B2, A2-B1, A2-B2, A1-C1, A1-C2, A2-C1, A2-C2, A1-D1, A1-D2, A2-D1, and A2-D2. Let us assume that DELNI B fails, causing the following channel failures: A1-C1, A2-C1, A1-D1, and A2-D1. A display of channel failures would show that some interesting event had just occurred but would leave it up to the system or network manager to isolate the actual failure. Also, since other channels are still open to VAX C and VAX D (A1-C2, A2-C2, A1-D2, and A2-D2), these nodes still remain in the cluster. However, the number of channels to these nodes has been halved, reducing cluster availability.

LAVc network failure analysis uses the physical network description to analyze channel failures. The working channel A1-C2 indicates that VAX A, A1, DELNI A, LAN segment A, Ethernet-to-Ethernet LAN bridge, LAN segment B, DELNI D, C2, and VAX C function. The working channel A2-D2 indicates that A2, DELNI C, D2, and VAX D also function. The remaining components are DELNI B, C1, and causing the failure and is the only network component displayed on the console.

In this small cluster configuration, LAVc network failure analysis has reduced the messages displayed, i.e., from four channel failure messages to one component failure message. This simpler display provides timely notification and better isolation of network component failures, allowing the system or network manager to repair the network earlier and restore the full availability of the cluster.

Physical Network Description

LAVc network failure analysis requires a description of the physical network. This description lists the components used by the LAVc and the network paths that correspond to the LAVc channels.

The network component description consists of several pieces of data, including a component type and text description provided by the system or network manager. Some component types will require additional data. There are several types of network components: NODE,

D1. By reviewing the failing channels for common failures, we see that two channels use component C1, two channels use component D1, and all four channels use component DELNI B. Therefore, DELNI B has the highest probability of

ADAPTER, COMPONENT, and CLOUD. Each NODE component requires a unique node name associated with it that matches the SCSNODE SYSGEN parameter. The ADAPTER component has at least one and sometimes two LAN addresses associated with

it. One LAN address is the hardware address and the other, when specified, is the DECnet LAN address. COMPONENTs are used to describe all pieces of the network, both working and nonworking. CLOUDs describe portions of the network that are working only if all paths are working. Any path failure implies that the CLOUD component may not be working.

Component descriptions can range from actual devices and cables to internal CPU bus adapters. When the component is defined, an ID value is returned for use in the network path description. The choice of the components described is left to the system or network manager and allows the manager to select the desired level of network analysis. Each network component has a reference count and a working count. The reference count is incremented when a network path is defined that utilizes the network component. The working count is incremented each time a LAVc channel is opened, and decremented each time an open LAVc channel is closed.

The network path description consists of a directed list of component identifier (ID) values. using this path. The final component ID value is that of the remote node.

Each network path description must contain two node ID values and two adapter ID values. To be useful for analysis, the path description must contain the node ID value for the node running the analysis. Without this node ID value, the path cannot be matched with any of the LAVc channels on that node.

Channel Mapping and Processing

The network path descriptions are matched with the LAVc channels by using the LAN addresses. If possible, only the LAN hardware address is used for the mapping function. This mapping provides the best analysis because it remains constant with respect to any LAN adapter. In clusters running mixed VMS versions, the LAN hardware address is not available for systems running a version prior to VMS version 5.4-3. In prior versions, the DECnet LAN address is used for the mapping function.

Each time a LAVc channel is opened, the network path database is searched to locate a matching network path description. If found, this description

For proper analysis, this list must start with the ID value for the local node. Each successive ID value in the list must be associated with the next network component through which a message would travel when is connected to the channel and a scan of all the components in the path is performed. For each component in the path, the working count is incremented. If the component switches from

not working to working, then a WORKING message is displayed.

When a LAVc channel fails, the corresponding network path is placed on a failure list. The network path is then scanned and the working count for each component is decremented.

Failure Analysis

Related channel failures are collected by delaying 10 seconds following the channel failure. Each channel failure extends the time delay to the full 10 seconds. Once the 10-second delay has elapsed following the last channel failure, the full list of failing network paths is processed.

Computing the failure probabilities begins by reviewing each of the components in the network path. If a component cannot be proven to work, then it is placed on the suspect list and the component's suspect count is incremented. A component is working if the working count is non-zero; a CLOUD component is working if the working count equals the reference count. This step ends with a list of suspect components, each with a suspect count that represents the number of and a primary suspect is selected. The primary suspect is the first component with the highest suspect count in the network path. Secondary suspects are the other components in the network path with the same suspect count value. The primary and secondary suspects are displayed after all the network paths have been reviewed. The other components in the suspect list are removed from the list, and are not displayed because the failure analysis judged them to be unrelated to any of the channel failures.

There are several limitations to the failure analysis. The analysis requires an accurate description of the physical network. The failure analysis is also looking for a common network component failure. Therefore, an incorrect analysis results from either an inaccurate network description, multiple related failures, or too much detail.

The key to a valid network failure analysis is the correct description of the physical network. In Figure 2, if the network path A1-B1 incorrectly listed DELNI B, then the

times this component could have caused the failure. Suspects are selected by comparing the suspect counts for each of the components in a network path. Each network path is reviewed independently

failure analysis would find that DELNI B is working and remove it from the suspect list. The final analysis would list both C1 and D1 as the failing components. Validation of the network description can be performed by network

fault insertion and by reviewing the network failure analysis. If the description is accurate, then the failure analysis should display the expected messages. If an inaccurate network description exists, unexpected messages may be displayed. In such cases, the network description should be reviewed.

Multiple related failures may also cause an incorrect failure analysis. Referring again to Figure 2, assume a correct network description. Instead of a DELNI B failure, assume that both C1 and D1 have failed. The failure analysis reviews the network description and locates the single component DELNI B because it is common to all of the failures. In this case, the failure analysis does correctly locate the area of the network (something connected to DELNI B). However, further review is required to identify that DELNI B itself has not failed, but rather both C1 and D1.

The choice of the network description, the number of components defined, and the path descriptions, is left to the system or network manager. This choice allows the manager to select the level of failure reporting reduce the components to a single failure. Instead, a primary suspect and several secondary suspects are usually displayed. Too much detail also requires more CPU cycles and memory for analysis, and in general is a bad trade-off.

In Figure 2, if the Ethernet adapter C1 fails, and the transceiver cables are listed in the network description, then the failure analysis displays two messages. The primary suspect is listed as the transceiver cable because it is the first component that matches the failure in the path from A to C. The Ethernet adapter C1 is listed as a secondary suspect, because its suspect count matches the suspect count of the primary suspect. In this example, there are no network paths described that use Ethernet adapter C1 without using the transceiver cable connected between C1 and DELNI B. With the network description provided, there is no way to distinguish between these two components. Therefore, both are displayed when either is a primary or secondary suspect.

Benefits

The LAVc network failure analysis, combined with

needed to troubleshoot the network. However, when the physical network description includes too much detail (e.g., transceiver cables), it becomes difficult for the failure analysis to an accurate description of the physical network, enables the system or network manager to maintain the increased availability gained with the use of multiple LAN adapters.

Timely analysis and reporting of network component failures significantly reduces troubleshooting times and increases the overall cluster availability.

Summary

VMS version 5.4-3 increases the availability of Local Area VAXcluster configurations by providing the following features:

- Faster detection of channel failures
- Support for the use of multiple adapters
- Support for the use of additional network paths
- Detection of network congestion
- Analysis of network failures

The goals of these features are to

- Provide higher cluster availability
- Work around network congestion and network component failures while keeping the cluster running
- Detect problems earlier and report them more accurately, with network data that helps isolate the failing network components

In addition to meeting these goals, the features in VMS version 5.4-3 increase the cluster communication bandwidth.

Acknowledgements

I want to thank Kathy Perko and Steve Mayhew for their help with the design of the multiple-adapter version of PEDRIVER. Kathy reviewed the code during the implementation and provided valuable input for both the code and this paper. Thanks to Scott H. Davis, Sandy Snaman, and Dave Thiel for their contributions to the new PEDRIVER design. Thanks also to the LAN Group (Linda Duffell, Dave Gagne, Rod Gamache, Bill Salkewicz, and Dick Stockdale) for the VAX communication interface to the LAN drivers, which simplified the design of the new PEDRIVER. I also wish to acknowledge the LAN Group for their help during the debug phase of this implementation.