TECHNOLOGY BRIEF

May 1997

Compaq Computer Corporation

CONTENTS

Virtual Interface Architecture for SANs

In a joint effort by Compaq Computer Corporation, Intel Corporation, and Microsoft Corporation to expedite the development of clustering solutions, the Virtual Interface Architecture (VI Architecture) Specification was developed by the three founding members. These three leading companies then presented the initiative to an expanded group of industry leaders at a conference in late January. At that time, all of those present were invited to participate in the process either by providing input to the specification or participating in one of the workgroups that will be created to address issues as they arise. Compaq, Intel, and Microsoft are the ideal leaders for such an initiative as the number one server provider, leading architecture developer, and the leading operating system provider respectively. As leaders of this initiative these three companies will be responsible for the following key tasks:

- *manage the specification process as it moves forward*
- *gather, track, assign, and collate issues and comments that are submitted*
- *host events to ensure open contribution to the specification*
- *establish workgroups to address issues as needed*
- *develop and schedule the efforts on the standard*
- *manage logistics*

Compaq's key goals with regard to VI Architecture include: 1) fulfilling customer requirements by providing reliable, cost effective building block products, 2) ensuring the adoption of standards-based clustering technology, and 3) accelerating the expansion of the clustering market. However, none of this can be accomplished without broad industry adoption and support. The input and endorsement of key industry leaders are essential to the success of the initiative.

This document outlines an initiative by industry leaders to develop an interface standard to extend the capabilities within system area networks (SAN) which are used to implement clusters of servers or workstations. The outcome of this initiative is the Virtual Interface Architecture Specification, which defines a new system architecture for more efficient communication within clusters. VI Architecture will have significant impact on server technology by increasing the scalability, reliability, and availability of future clustering solutions.

EXECUTIVE SUMMARY

Within the next year and a half, clustering technology will evolve from the current 2-node application failover solutions, to 2-node (and eventually multi-node) parallel application clusters. As this evolution occurs, reliable messaging between clustered servers will become increasingly critical. Parallel application clusters will create increased messaging traffic between servers, so a very efficient interface is required. Some current protocol packages and hardware interfaces that are used for the high-speed communication within clusters produce high software overhead because they were not designed for a system area network (SAN) environment. In order to relieve this overhead, VI Architecture was developed which will be optimized for use with SANs. SANs are used to implement high-performance "clusters" of servers or workstations that are grouped via a SAN hardware interface to perform as a single, integrated system. VI Architecture acts as a standard reliable messaging interface to significantly increase the efficiency of the communications between these clustered servers over some current solutions, which use non-optimized interfaces.

Please direct comments regarding this communication to the ECG Technology Communications Group at this Internet address: tech_com@bangate.compaq.com

COMPAQ

NOTICE

Virtual Interface Architecture for SANs
First Edition (May 1997)
Document Number 508A/0597

COMMUNICATION WITHIN CLUSTERS

When addressing the issue of a messaging interconnect for the communications between clustered servers or workstations, there have traditionally been only two choices: to use *local area networks* (LAN) *network interace controllers* (NIC) or to build an interconnect specifically designed for messaging. Neither of these two options is ideal.

A LAN is a communications network connecting servers, workstations, and a network operating system (OS). LAN NICs are not optimized for a cluster environment. They are traditionally used for the communications between a server and a client. This assumes that connections are being made over long distances, through the use of any of several different media. As such, the protocols and interfaces focus on flexibility for networking requirements. Therefore, the LAN NICs use a great deal of the system resources just to ensure that messages are transmitted. This is not necessary with system area networks since the servers in question are connected in a dedicated compute environment, and do not need the rich multi-protocol interfaces used in LANs.

The second option, to develop a dedicated SAN interconnect, has the advantage that the protocol and hardware implementations can be optimized for clustering applications. Servernet, originally developed by Tandem Computer Incorporated and migrated to standards-based servers by Compaq and Tandem, is one such dedicated SAN. However, historically there has been a multiplicity of these interfaces, presenting software developers with the task of potentially supporting many different interfaces. Developing a common interface standard that can be applied to many SAN implementations, such as Servernet, can expedite proliferation of software solutions across a wide range of SAN implementations.

Unlike other approaches to scaleable systems, SANs enable the use of existing volume server technology without requiring any hardware modifications. SANs are independent of the internal server bus structure, a feature that is very cost effective and reduces redesign/validation time. This feature also allows SAN technology to develop at its own rate. Other approaches to scaleable systems, such as NUMA architecture, require changes in high volume product that result in increased cost, delayed schedules, and additional validation responsibilities. The importance of bringing new technology to market as fast as possible, at the lowest cost is undisputed. Since SANs are the future for high availability and scalability, the development and adoption of a common interface will expand their use by software developers across implementations, offering many price:performance options.

THE VI ARCHITECTURE CONCEPT

The engineers for VI Architecture designed it to minimize message processing delays and allow more efficient communication within system area networks. VI Architecture is the messaging interface for applications that reside on servers connected in a SAN. Figure 1 depicts the relationship between VI Architecture and SANs.

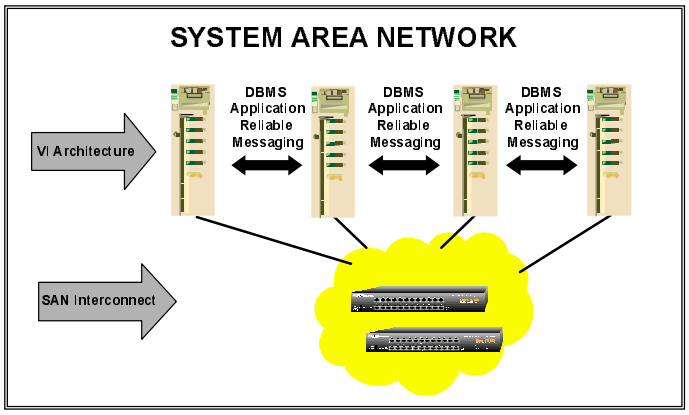

Figure 1: Relationship between VI Architecture and SAN

The concept behind VI Architecture has its roots in the traditional messaging between applications and a computer's NIC. Computer applications operate as if they have unlimited memory. In reality, the operating system gives and takes the actual physical memory away from applications as it is needed to run other applications. Traditionally, if an application wanted to send messages to the NIC using the physical memory, the request had to go through the kernel, which caused processing delays. With VI Architecture, the application can use its virtual memory addresses to speak directly to the SAN NIC without going through the kernel. This is true of the most common types of messages that are passed, such as Send and Receive messages. The less common types of messages will still go through the kernel to access the NIC, but they comprise a very small percentage of the messages that are passed. This will greatly reduce the message latency. Virtual Interface Architecture was named for the fact that it exists in this virtual address space that the application program uses to communicate with the SAN NIC.

Although applications bypass the operating system by sending messages from their virtual addresses, the operating system continues to provide security and messaging setup for the applications. Following are some of the key advantages of using VI Architecture in a SAN environment:

- allows applications to use their own memory to directly send messages to the SAN NIC
- eliminates the buffer copying and kernel overhead of traditional NIC protocol
- preserves security and authentication provided by the operating system

When addressing a solution to the communication limitations of clusters, several criteria were identified to ensure that the key goals of the VI Architecture initiative were met. It was determined that the interface must:

- promote high performance SANs
- ensure internode communications are reliable
- embrace existing technology
- encourage industry participation
- increase marketability of SANs

Performance and Reliability

Because VI Architecture can bypass the operating system for message passing, clusters will be able to offer much higher performance. VI Architecture's messaging system also contributes to the improved performance of SANs because it allows system resources to be used productively rather than for message error checking. The VI Architecture specification allows for a spectrum of reliability, spanning hardware-based reliability for high performance, to software-based reliability, yielding low cost solutions. This is important because the more reliable the hardware is, the less software overhead will be incurred to check for messaging errors. With a redundant interconnect that has no single point of failure, VI Architecture allows SANs to produce extremely reliable results, greatly improving availability. The section entitled "How VI Architecture Works" will provide a more detailed explanation of the higher performance and reliability that VI Architecture allows.

Use of Existing Technology

VI Architecture is both operating system and processor independent. This will allow SANs to utilize existing hardware and software technologies. VI Architecture is also hardware independent and compatible with current interconnects such as Ethernet and ServerNet. New system hardware will not be required to run a SAN with VI Architecture. This ability to use existing technology will help foster the rapid growth of new VI Architecture compliant technology.

Industry Participation

Because VI Architecture is operating system, processor, network, and CPU independent, every hardware and software vendor has the opportunity to participate and compete in the development of new products for use with SANs. This level playing field will also help speed the evolution of SAN technology.

Increase of Marketability of SANs

VI Architecture increases the marketability of SANs in a few ways. As described above, the improved performance, increased reliability, and ability to use existing technology all expand the marketability of SAN systems. VI Architecture is also expected to accelerate the acceptance of server and workstation clusters as a lower cost alternative to the larger, more expensive, proprietary machines. SANs operating with VI Architecture will allow customers to purchase the less expensive, industry standard systems and still attain the desired results.

HOW VI ARCHITECTURE WORKS

Implementations of the VI Architecture consist of three primary components, 1) the SAN media hardware with the VI register interfaces, 2) the VI Primitive Library (VIPL), and 3) VI OS kernel support routines. Figure 2 illustrates the relationship between these three components and how applications communicate with each other in a SAN that uses VI Architecture. The three components act together to allow applications to access the VI Architecture interface. Each VI Architecture register represents a hardware send and receive queue between two connection endpoints. These registers are accessed and manipulated by a thin software layer called the VI Primitive Library. An application calls the VI kernel support routines to establish an endpoint of a connection to another instance of the application on another node in the SAN. Once the endpoints are established, the application will then register a portion of its virtual memory with the kernel to be used for sending and receiving messages or transferring data. After the memory is registered, the application can then send and receive messages directly from user space without

incurring the overhead of kernel system calls. Implementations of VI Architectures that utilize an inherently reliable SAN media interface can avoid the high overhead of software protocols.

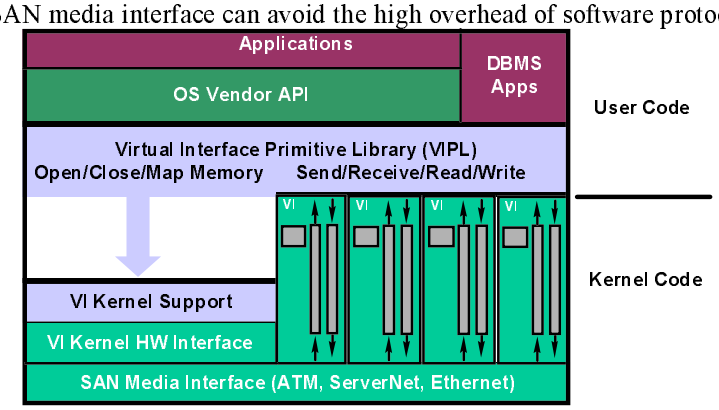

Figure 2: Relationship between SAN media hardware, VI Primitive Library, and VI OS kernel support

For the most part, current software protocols run inside the operating system kernel; and to get to them, a number of contact switches, buffer copies, and interrupts are involved. Because VI Architecture will be operating system independent, it will not have to go through the kernel. Thus, access times will be faster.

VI Architecture will be hardware independent and compatible with current hardware interconnects such as ATM, ServerNet, and Ethernet. This will help expedite the migration from current software protocols to VI Architecture and will speed change by allowing the use of existing technology. Additionally, VI Architecture is CPU independent, enabling protocols to be put into silicon. This allows CPU overhead to be offloaded to adapter silicon. This, combined with the fact that VI Architecture does not waste a lot of system resources for message error checking, generates extremely low processor overhead and high bandwidth.

WHY NOW?

Current cluster offerings do not produce high messaging demands. Future solutions will require a greater amount of communication interface between the clustered servers or workstations to ensure high availability and scalability are maintained. It is critical that we pave the way now for upcoming applications. We already know what some of the inherent messaging limitations of future clusters will be and, with VI Architecture, we have a solution for many of these limitations.

In the next section, the evolution of clustering technology is outlined. This section includes descriptions of current and future cluster offerings and an explanation of the messaging challenges of future clusters.

EVOLUTION OF CLUSTERS

The demands of business-critical applications have increased dramatically in the past few years. Businesses have much lower tolerance for failures and downtime and are growing at a much faster rate than ever before. They require hardware and software support that will be highly available, reliable, and scaleable as their businesses grow. There are two solutions to these increased customer demands and Compaq is developing products that support both methods. One solution is to continue to expand the performance and availability of individual servers. The other technique is clustering. Both of these methods offer performance, availability, and cost choices. Compaq provides both to allow customers to make the right choice for their applications.

Clusters are groups of interconnected servers that work as a single system to provide high-speed, reliable, and scaleable service. Until recently, only very expensive, proprietary systems could deliver the levels of speed, reliability, and scalability required for enterprise computing. With clustering, less expensive, industry-standard systems now have the same capability. Currently, there are two types of clusters: *application failover* and *parallel application*.

Current Application Failover Clusters

In an application failover cluster, two or more servers, or nodes, are connected together. Currently Compaq has the capability to connect two servers in an application failover cluster, called a 2-node cluster. In this configuration, an application runs on the first server (server A). If there is a hardware failure on server A, the application shuts down and starts almost instantaneously on the second server (server B). Data that is stored on external storage devices, and connected to the cluster via SCSI cables, also switches to server B. This failover is accomplished automatically without system administrator intervention, ensuring high availability of data with minimal disruptions.

Compaq currently offers two application failover solutions, *Standby Recovery Server* and the *On-Line Recovery Server*. With the Standby Recovery Server option, one server acts as a standby to the other. If the main server fails, the second (or recovery) server automatically takes over the functions of the main server. No system administrator intervention is required for this switchover. When not taking over for the main server, the recovery server stands idle.

In the case of the On-Line Recovery Server option, two servers are connected in a cluster. However, in this configuration the servers process independent workloads during normal operation. They may even be running different applications. No server stands idle as happens with the Standby Recovery option. In addition, the On-Line Recovery Server option can failover in either direction. Therefore, if either of the servers fails in this configuration, the other must be able to accommodate the workload of the failed server in addition to its own workload.

Note: Please see "Competitive Analysis: Clustering Solutions" white paper (document no. 245A/0996) and "Compaq On-Line Recovery Server" white paper (document no. 286A/1196) for more information on Standby Recovery Server and On-Line Recovery Server configurations.

Future Application Failover Clusters

While Compaq's current application failover solutions use SCSI cables to connect the nodes to the external storage devices, the solution that Compaq will announce in the second half of 1997 uses Fibre Channel. This upcoming application failover offering also has a redundant messaging interface and will be aligned with Microsoft Wolfpack. The use of Fibre Channel technology, redundant messaging interface, and Microsoft Wolfpack will allow for a richer clustering environment with more application support and more reliable messaging. Figure 3 shows Compaq's current and upcoming application failover configurations.

Note: Please see "Strategic Direction for Compaq Fibre Channel-Attached Storage" white paper (document no. 227A0397) for more information on Fibre Channel technology.

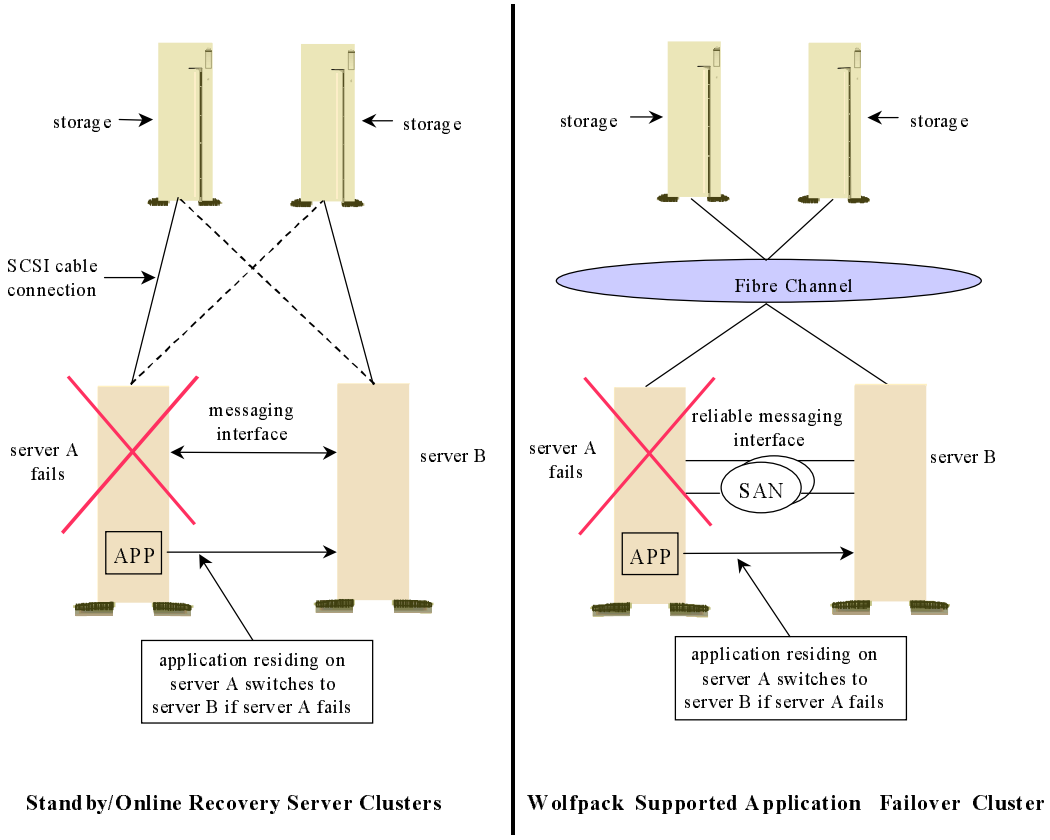

Figure 3: Current and Upcoming Application Failover Clusters

Parallel Application Clusters

A parallel application cluster is also a grouping of two or more servers. In this configuration a copy of each application resides on each server, allowing additional scalability over application failover clusters. Client requests are divided between the servers, which exchange information about which instance of the application will work on which data for each user. Therefore, if server A experiences a hardware failure, the jobs it is working on can be instantly redistributed across the remaining server(s). And, as with application failover clusters, the external storage device that server A was using switches its resources to server B (Figure 4). In this instance the clients never see an application crash. To them the application is continuously available.

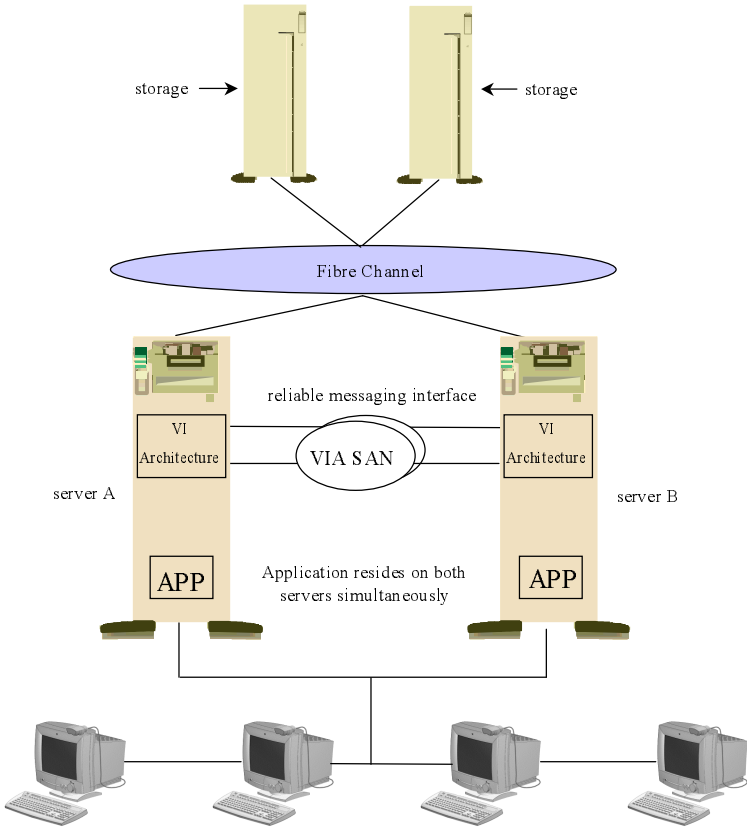

Figure 4: Parallel application cluster with VI Architecture SAN

Currently, there are very few applications that can run in parallel environments. They are faced with the challenges of using LAN NICs for messaging. Compaq's parallel application offerings, which will be announced mid-1998, will use VI Architecture to address these challenges.

Parallel Application Database Example

For balance of workloads between the two servers to work properly, both servers must constantly monitor both the hardware status and the intermediate points of the application for each server. Figure 5 shows how a database application in a parallel application cluster is monitored on both servers for workload balancing.

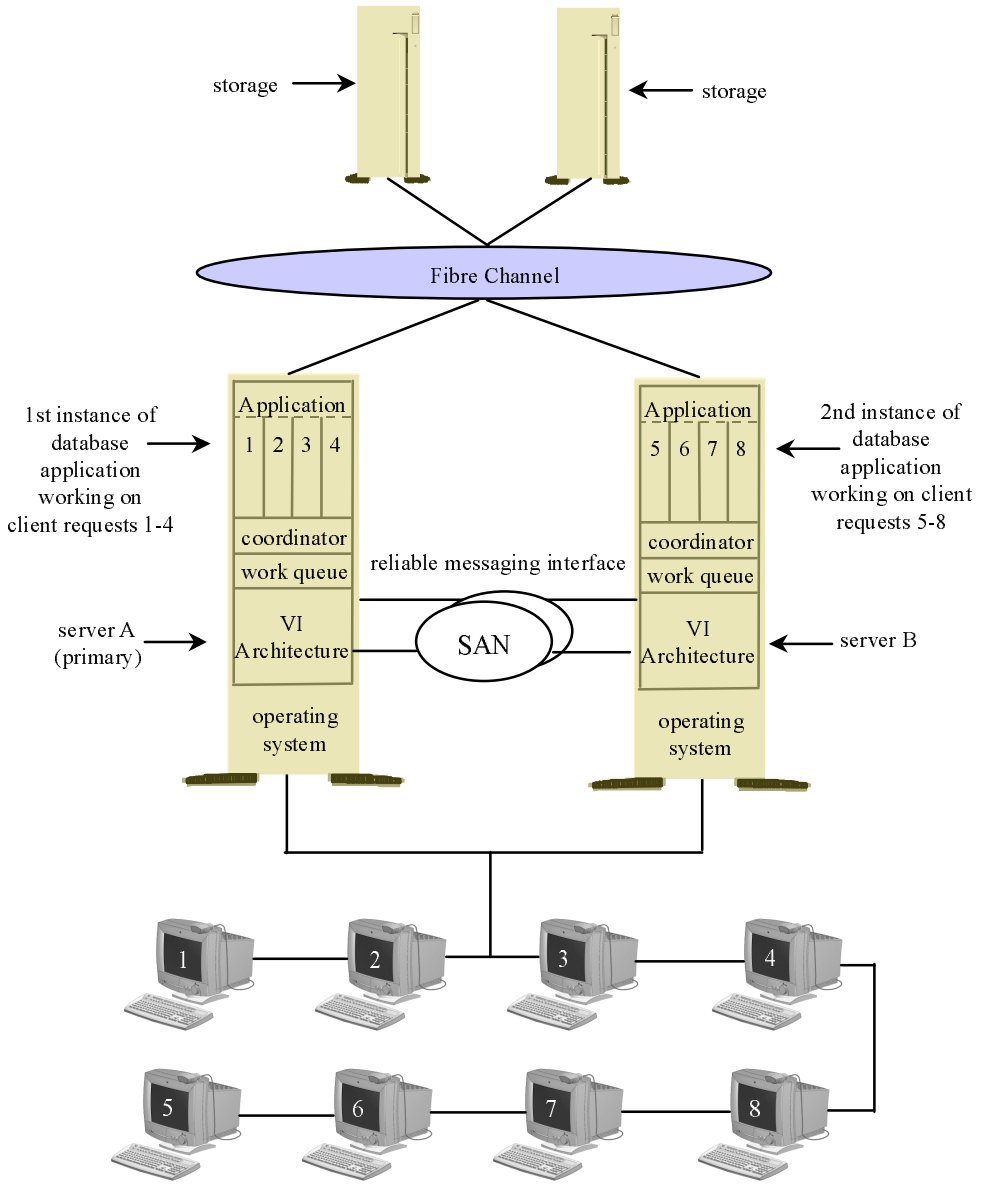

Figure 5: Database application in parallel application cluster. Message passing between coordinators enables load balancing.

In this situation, the database application exists on both servers. Eight end users, or clients, are accessing the database through the operating system on the primary server (server A). Work requests from the end users are put into the queue. A "coordinator"[1] on server A talks to the coordinator on server B and they agree which instance of the application will handle each request.

[1] Possible types of coordinators include transaction monitors, distributed lock managers and data partitioners.

In addition to distributing the work requests, the coordinator also has to keep track of which instance of the application is working on which request at a given time.

Obviously there is a greater amount of communication going on between the servers in a parallel application cluster than in failover clusters. Inefficiencies of current protocols, such as LAN technology, create a lot of overhead. The need for a more efficient message handling interface for clusters is unquestionable. The LAN, storage, and CPU subsystems have traditionally been the major components of a computer. Now, the messaging subsystem is a major component that must be designed to perform optimally in a cluster environment. So the question is: *What type of messaging interface can ensure the most efficient use of the system resources, allowing clusters to perform optimally?* With the Virtual Interface Architecture Specification, Compaq, Intel, and Microsoft are leading the effort to answer this question.

Cluster Construction

Multiple servers and VI Architecture are obviously not the only requirements for a cluster. Several components are required, among them:

- two or more servers or workstations
- enabled operating system
- external data storage devices
- messaging interface
- one or more applications (including databases)

Some of these components will need to be updated either by a hardware redesign or software emulation for optimization in a SAN environment. These updates are required because of new tasks these components have to perform:

- The operating system must provide as much transparency as possible to hide which node is performing which tasks. This transparency is required so if one node shuts down, it can be replaced without disrupting the end users. The operating system must also coordinate the failover event and resources.
- Storage must be accessible by more than one node to ensure high availability. If only one node is able to access the storage, the number of problems a node can address is limited.
- The application must be modified to exchange information through the coordinator.

The most difficult of these five components to configure for SAN is the messaging interface between the clustered servers. However, Compaq, Intel, and Microsoft believe in and support VI Architecture as the answer to this challenge.

PRODUCT LINE ALIGNMENT

The VI Architecture initiative is synergistic with current clustering strategies. Compaq's 1997 clustering offerings are aligned with Microsoft Wolfpack, which targets application failover cluster solutions that have light to moderate messaging demands. VI Architecture is focused primarily, but not exclusively, on the upcoming parallel application cluster technology. The VI Architecture efforts will enable efficient and reliable messaging in these clusters. Compaq anticipates that these future parallel application clusters will be aligned with subsequent Wolfpack versions. It is expected that VI Architecture software emulation products and modified SAN NICs will be available in 1998. It is also expected that multi-node failover and parallel application clusters will be introduced in 1998.

SUMMARY

As customers need greater reliability, availability, and scalability, SANs are the way of the future. Clustering is moving towards multi-node parallel application clusters. It is therefore imperative that the SAN messaging interface standard is completed, well received across the industry, broadly adopted, and well supported. VI Architecture provides an opportunity to exploit high bandwidth SANs as well as existing standard networks such as LAN and WAN. However, although neither LANs nor WANs are the appropriate solution for a SAN interconnect, their use and value will remain unchanged as SANs are not intended to replace them in their current capacities.

VI Architecture will also potentially remove a barrier to the use of future interconnect technologies that have not yet been developed. VI Architecture provides a way to do all of this at minimal cost, with the least amount of effort for the customer, and in the quickest time frame possible. It will also allow Compaq to fulfill customers' requirements for scalability, increased reliability, high bandwidth, low latency, and low cost cluster solutions while spurring innovation and ensuring the continued growth of the cluster market. Compaq is committed to supporting the VI Architecture effort now and will offer compatible products when they can provide tangible benefits to our customers.