COMPAQ
White Paper

September 1998
ECG097/0998

Prepared by Industry Standard
Server Division

Compaq Computer Corporation

Contents

ServerNet - A High Bandwith, Low Latency Cluster Interconnection

This paper provides an overview of ServerNet technology, what benefits can be realized from ServerNet in clustered server SANs, and how Compaq is implementing SAN concepts using ServerNet as the primary cluster interconnect for its enterprise vision of the future.

Abstract: Clustering is mentioned almost everywhere in current computer industry news, but it is more than just linking computers. Enabling enterprise-ready cluster solutions requires reliable, scalable interconnect architecture and robust, parallel software.

The computer interconnect technology is the nerve system of a cluster, tying together all its various components - servers, storage, networks, and other system resources. There are many possible cluster interconnect technologies—Ethernet, asynchronous transfer mode (ATM), and other high-speed services. One of the emerging technologies is ServerNet, representing the industry's first System Area Network (SAN) solution, which offers fault-tolerant high bandwidth and extremely low latency for distributed message passing and clustered server communications.

ServerNet, a new open systems approach for high-speed interconnection of both shared-disk and shared-nothing clusters, considers all the major components of the system as independent elements so that any cluster component—server, disk, or I/O device—can directly communicate with any other component without processor intervention.

The benefits of ServerNet inter-processing communications include high speed, low latency communications with the added benefit of truly non-stop interfaces through the multi-plane ServerNet Switch. These hardware features, in conjunction with the developing Thin Winsock-2 and VI protocol architectures for distributed application and parallel processing, make ServerNet an ideal candidate for inter-processor and cluster communications.

As ServerNet evolves from the ProLiant Cluster Series F and continues to scale into larger clustered server configurations, such as the VI - SAN technology demonstration using IBM Universal Database Version 5.1 and Compaq's industry leading cluster TPC-C benchmark, this technology can continue to establish itself as the industry standard cluster interconnect.

Notice

The information in this publication is subject to change without notice and is provided “AS IS” WITHOUT WARRANTY OF ANY KIND. THE ENTIRE RISK ARISING OUT OF THE USE OF THIS INFORMATION REMAINS WITH RECIPIENT. IN NO EVENT SHALL COMPAQ BE LIABLE FOR ANY DIRECT, CONSEQUENTIAL, INCIDENTAL, SPECIAL, PUNITIVE OR OTHER DAMAGES WHATSOEVER (INCLUDING WITHOUT LIMITATION, DAMAGES FOR LOSS OF BUSINESS PROFITS, BUSINESS INTERRUPTION OR LOSS OF BUSINESS INFORMATION), EVEN IF COMPAQ HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

The limited warranties for Compaq products are exclusively set forth in the documentation accompanying such products. Nothing herein should be construed as constituting a further or additional warranty.

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 or product quality or correctness, nor does it ensure compliance with any federal state or local requirements.

Product names mentioned herein may be trademarks and/or registered trademarks of their respective companies.

Compaq, Contura, Deskpro, Fastart, Compaq Insight Manager, LTE, PageMarq, Systempro, Systempro/LT, ProLiant, TwinTray, ROMPaq, LicensePaq, QVision, SLT, ProLinea, SmartStart, NetFlex, DirectPlus, QuickFind, RemotePaq, BackPaq, TechPaq, SpeedPaq, QuickBack, PaqFax, Presario, SilentCool, CompaqCare (design), Aero, SmartStation, MiniStation, and PaqRap, registered United States Patent and Trademark Office.

Netelligent, Armada, Cruiser, Concerto, QuickChoice, ProSignia, Systempro/XL, Net1, LTE Elite, Vocalyst, PageMate, SoftPaq, FirstPaq, SolutionPaq, EasyPoint, EZ Help, MaxLight, MultiLock, QuickBlank, QuickLock, UltraView, Innovate logo, Wonder Tools logo in black/white and color, and Compaq PC Card Solution logo are trademarks and/or service marks of Compaq Computer Corporation.

Microsoft, Windows, Windows NT, Windows NT Server and Workstation, Microsoft SQL Server for Windows NT are trademarks and/or registered trademarks of Microsoft Corporation.

NetWare and Novell are registered trademarks and intraNetWare, NDS, and Novell Directory Services are trademarks of Novell, Inc.

Pentium is a registered trademark of Intel Corporation.

Information from this paper was compiled from information found on websites and includes news releases, product summary sheets, product guides, and technical papers.

ServerNet - A High Bandwith, Low Latency Cluster Interconnection
White Paper prepared by Industry Standard Server Division

First Edition (September 1998)
Document Number ECG097/0998

Clusters – an Emerging Paradigm in the Intel NT Server Marketplace

Microsoft's Windows NT Server has established itself as an easy-to-use, easy-to-program operating system running on open industry PC hardware. But for the Windows NT Server platform to move into the area of essential business computing, information technology professionals need to see systems that exhibit true enterprise-level availability and scalability.

Clustering has its roots in fault-tolerant, massively parallel computing. Today, the computer industry understands that greatly enhanced scalability and reliability can be realized through the clustering of multiple small-scale, thus low-cost, Symmetric Multiprocessor (SMP) nodes and, in some cases, single processor nodes.

Intelligent middleware, distributed databases, and high-speed interconnect technologies are now falling into place to support clustered environments. Microsoft has embraced the cluster paradigm for Windows NT Server and is providing the requisite operating system-level functionality while supporting the development of standard clustering services for the Windows NT Server platform.

Microsoft Cluster Server (MSCS) technology is a key enabler of clustering in the Windows NT Server environment. However, it is only the software component of the solution. Hardware models and interconnect schemes such as ServerNet, along with software message passing architectures such as Virtual Interface Architecture (VI) will ultimately define cluster performance, scalability, and reliability.

ServerNet: The Interconnect Vehicle for Enterprise Clusters

Traditional local and wide area networking technologies, which may make suitable client connections, are often too slow to provide effective inter-server communications. These technologies do not provide sufficient performance due to their heritage of involved communication protocols and secure message-passing architectures. These technologies can burden the cluster's servers with significant protocol overhead when trying to provide high bandwidth and throughput.

These operating characteristics necessitated a new approach to meet the needs of both shared-disk and shared-nothing clusters. This new concept, the industry's first, is referred as a System Area Network (SAN), and encompasses a high speed, low latency transfer mechanism. ServerNet is Compaq's current choice as a SAN vehicle.

ServerNet Initial Concept

The concept of a high-speed, non-blocking inter-processor link was first developed by Tandem, a division of Compaq Computer Corporation, as a proprietary solution to meet a fault-tolerant clustering requirement. This concept was to satisfy requirements of no transaction or data losses due to a system error. ServerNet was designed from the ground up to provide the low latency, reliability, and scalability that are essential for high-end, commercial applications. The initial implementation was not only for inter-processor communications keeping clustered applications and server nodes in step, but also as a fault-tolerant link to data storage and networking.

System fault tolerance is achieved by having dual memories and dual paths to memory on a single node and paired nodes, so if one node or one processor in the node fails, it is taken offline but the transaction completes successfully on the paired node.

This implementation was further enhanced in the Tandem Himalaya S7000 using the Tandem Non Stop Kernal (NSK) operating system. In this implementation, ServerNet was used strictly as a message-passing interface sending checkpoint messages between "process pairs" or two node clusters within a multi-node cluster called Himalaya. Part of this matched pair processing was the implementation of a transaction monitor to further ensure transaction completion. ServerNet was also used here as an I/O vehicle to attach and ensure consistent and continuous data flows into the cluster.

ServerNet as Industry Standard

Today ServerNet is being externalized for the open systems server market and is comprised of two hardware devices: the ServerNet PCI Adapter and the six port ServerNet Switch.

Tandem has been promoting its technology to industry hardware and software providers as an industry standard device. At present over 40 OEM partners are working with the ServerNet PCI form factor adapter and switch to implement clustered server configurations.

ServerNet has been offered in the CS150 and Integrity XC products offered from Tandem which are based on the Microsoft NT Server and Santa Cruz Operation (SCO) UnixWare operating systems. Compaq has also referenced the use of ServerNet in their ProLiant Cluster Series F model clusters as a high-speed point-to-point interconnect. In addition, Compaq has utilized ServerNet in multi-node cluster solutions for its Workstation Division’s NT Workstation clusters in technical computing and parallel commercial applications.

The leading database vendors, Oracle and IBM, have committed to the VI Architecture and are developing distributed message interfaces to VI as well. This will allow Server Net and the industry standard VI Architecture to provide scalable servers with transparent parallel transaction and decision support services that have not been available in past systems.

Compaq is supporting the further development of ServerNet to support VI Architecture for the open systems industry and will continue to assist Tandem in their development of ServerNet.

Customer Benefits Using ServerNet Technology

Customers using ServerNet are assured of an exceptionally scalable, low latency interconnect providing an ultra-reliable interconnect for storage, processing, and communications. Customer benefits include the following:

- **Improved cost per transaction**. By utilizing ServerNet as an interprocessor or server interconnect at speeds that are equal or exceed current network technologies, bandwith, transactions, and data can be spread across a wide range of server systems. This cluster will reduce the overall cost of application systems as well as provide a more fault-tolerant processing complex for customers’ business critical applications.
- **High scalability**. ServerNet users can scale resources when needed to meet expected application demands across the cluster. If more resources are required, adding additional processors or memory to existing systems is less expensive than adding an entire new system to meet capacity needs.

Success Story

The San Jose Mercury News has a 4 processor S7000. Since installation on January 1, 1998, the S7000 has had one scheduled downtime with NO unscheduled downtimes. Now that the S7000 system is online, the response time is never more than one second, and the DP staff can now add compiles and other jobs during the day, which they could not do before. "We can't slow it down," said Tim Benjamin, Systems Support Manager. Benjamin added that he plans on adding the Contra Costa Times (a system newspaper) with 75 more Compaq workstations online by the end of the year.

- **High availability**. ServerNet allows systems to work as a true client/server hierarchy so applications servers can be administered independently of data servers yet with near I/O bus performance across the ServerNet interface. This, in conjunction with the VI Architecture and the future integration of Microsoft's Winsock 2 distributed command processing, make ServerNet the choice for fault-tolerant, high bandwith cluster communications.
- **Integration and application compatibility.** Users can be assured that ServerNet is well-tested and integrated into Compaq's server and cluster solutions. Compaq has offered ServerNet as interface option in its ProLiant Cluster Series F clusters after years of joint engineering and development with Tandem.

  ServerNet technology also includes software development for proprietary communications between Tandem clustered servers but also for more general forms of network communications over traditional communications protocols. This service development has been in two areas: TCP/IP and NDIS protocol stacks for NT Server and NT Workstation compatibility, and legacy application compatibility for applications such as Oracle Parallel Server and Informix XPS.

  The goal for this development is to migrate applications, which use these common protocol tacks easily for high-speed, fault-tolerant data and command processing. In this mode, applications will gain inter-processor communication performance increases due to the speed of the ServerNet Adapter and Switch hardware with its WormHole Routing techniques and multi-plane processing. WormHole Routing and multi plane processing (discussed in more detail further in this paper) are features of ServerNet switch technology. At this point, the applications can take advantage of the ServerNet hardware and software without change. In the future, the application developer may wish to migrate to one of the higher speed protocol stacks such as the LTI (Light Transport Interface) or Winsock 2, which would provide an even faster interconnect for clustered servers.

ServerNet and Virtual Interface Architecture (VI)

ServerNet technology is in its initial implementation from Compaq Computer Corporation. It is being extended to support additional distributed and parallel processing communications with the introduction of the Virtual Interface Architecture (VI) Distributed Message Passing architecture across the Compaq ProLiant Server and Workstation lines. ServerNet adapters are being updated to include VI protocol specifications on the adapter for hardware level communications processing as well as software stack processing, as is the case today.

VI is an emerging industry standard protocol for Distributed Message Passing and command processing. This development has been co-authored by Compaq, Intel, and Microsoft and published as a formal specification on December 16, 1997. Tandem has committed to implement the VI libraries and emulation services for ServerNet based on the Specification 1.0 of the VI architecture. Compaq and Tandem, using IBM Software Group's UDB V5.0 Parallel Database for NT, have demonstrated the use of VI and ServerNet functionality and performance at multiple shows and conferences. At present, over 140 technical organizations and corporations are in the review stages of the VI architecture and specifications.

Advantages of Using ServerNet to Meet the Interconnect Technology Requirements of Clusters

ServerNet considers all the major components of the system as independent elements so that any cluster component – server, disk, or I/O device – can directly communicate with any other component without processor intervention. Using ServerNet provides many technical advantages including:

High Bandwidth

The first cluster interconnect requirement is for high bandwidth. The aggregate data throughput capacity of a SAN (between nodes and shared storage or between the PCI buses of each connected node) must exceed the demands of a local PCI bus–connected controller such as a SMART-2 RAID controller or a Gigabit Ethernet NIC. Presently ServerNet allows 50MB data transfer between each of its two channels providing up to 100MB of data transfer between two or more SAN connected servers or workstations. Compaq has customers who are implementing multiple ServerNet adapters in servers for added bandwidth as well as fault tolerant alternative-pathing in the event of a link failure.

Low transmission overhead is also important for clustered systems and interconnects. Speed cannot come at the expense of increased CPU overhead, which constrains the cluster’s scalability. Scalability needs to be supported along multiple dimensions: additional processing power, increased I/O bandwidth, added interconnection bandwidth, and assured availability.

In support of enhanced reliability, the interconnect technology needs to exhibit fault tolerance, protecting against component failures as well as detecting and ensuring rapid recovery from corrupted data transfers.

Low Latency

The second cluster interconnect requirement is for any-to-any addressing. Today, this is a static connection via a user-defined routing table but in the future, with larger server clusters (up to 128 servers), any-to-any capability through self-discovery by ServerNet will make cluster growth seamless. Like traditional LAN and WAN implementations, ServerNet enables cluster resources to be shared. ServerNet recognizes all major components of the system as independent elements. This means that any cluster component—server, disk, or I/O device—can directly communicate with any other component without processor intervention. This support is still under development for Compaq systems but has been in Tandem systems since the early ‘90s. VI and I2O (Intelligent I/O) systems concepts as well as the forthcoming Intel I/O system architectures make the SAN concept an industry standard.

Latency is system overhead that occurs when data is moved between nodes in a cluster. ServerNet’s any-to-any switching fabric will enable virtually limitless cluster growth. As nodes are added to handle increased data volume, more complex transactions or queries, or larger numbers of concurrent users, the switching fabric can be scaled to accommodate the additions without adding new overhead. This is accomplished by incorporating all the logic necessary for the interconnection directly in the components. The fabric, or any part of it, can also be replicated for fault tolerance.

ServerNet technology uses six-way crossbar ASIC-based switches (providing a total of 300GB/second throughput theoretical limit for a single switch) to move traffic in an any-to-any fashion between system components: processor to processor, processor to disk, disk to disk, etc.

Processor intervention is not required in the switching process, since the switching fabric handles intelligent routing. As a result, ServerNet virtually eliminates the effects of CPU drag. Compared to other LAN-based networks such as Ethernet, ServerNet links have demonstrated the ability to move five times the data with about one tenth the CPU overhead, making ServerNet fifty times more efficient than Ethernet. This is due to both the speed of the network service as well as the switching technology, which sets ServerNet apart from traditional, interconnect technologies.

Error checking is also performed by the ServerNet fabric, which provides delivery of all data packets, in sequence, without loss or corruption. In addition, the fabric can be duplicated to provide fault tolerance.

Scalable Interconnect Architecture for Scalable Bandwidth

ServerNet technology enables scalable I/O bandwidth by providing the ability to add multiple data paths that can be cascaded to support as large a bandwidth as necessary. When a server is expanded, more data paths are added, and aggregate bandwidth of the ServerNet interconnect increases. ServerNet scales I/O bandwidth by supporting multiple simultaneous I/O transfers.

To accomplish this scalability, ServerNet technology embeds a reliable network transport layer into a single very large scale integration (VLSI) integrated circuit (IC) hardware device to connect a processor or I/O device to a scalable interconnect fabric. This fabric is composed of many very high-speed point-to-point data paths. Each high-speed path uses a hardware protocol to guarantee delivery of data between devices. The data paths allow system elements (processors, storage, I/O) to be joined into a cluster of servers and, potentially, storage.

Data paths from system elements are connected together within the cluster by means of six-port routers, which are single VLSI devices that use switching technology to direct requests to the correct data path. Using these routers, the system elements are assembled into as large a server as desired.

Wormhole Routing for Lowest-Latency Switching

ServerNet technology uses a technique known as wormhole routing to reduce network latency. With this technique, a packet does not need to be completely received before being sent to its next destination (as with store and forward). Wormhole routing works by allowing the router to decode the header of the packet as it is received, then locate the port on which the packet will exit by using the header's destination address and the internal routing table. This one-time operation allows the packet to be directed through the crossbar switch, all while the router is still receiving the remainder of the packet. As a result, the head is routed and retransmitted well before the tail has been received. Thus, the latency incurred by the router is much less than that of store-and-forward technology.

ServerNet Fault Tolerance and Data Integrity

For fault tolerance, dual ServerNet interconnects provide full dual-path connectivity to all system elements. If any component fails, data is transferred via the other interconnect. In addition, both interconnects are concurrently active, providing data-transfer capability in parallel while maintaining fault tolerance. Continuous availability is maintained by the dual independent ServerNet planes.

Data integrity is maintained with multiple techniques: command link integrity isolates single-bit errors, cyclic redundancy checks that maintain both data and control integrity of transfers crossing the interconnect fabric, and hardware protocol acknowledgments ensure that end-to-end reliable data transfer has occurred.

Mesh Concepts for Added Scalable Performance

With traditional shared-bus architectures, adding processors or I/O devices to the buses does not increase a server's aggregate bandwidth, so the same bandwidth must be divided among more devices. In contrast, adding processors or I/O devices with ServerNet technology actually expands a server's aggregate bandwidth, since every addition brings with it another 50-megabyte-per-second data-transfer path. The technology minimizes software latency through a "push/pull" ability to extract or deliver data autonomously to a node. Interconnect data transfers can themselves contain the addresses of information in other node(s) to "push" (write) data to or "pull" (read) data. A node can then request subsequent transfers from another node without requiring software interaction, as the ServerNet device performs the operation without disturbing any processors.

Another key advantage is that processor performance is no longer hindered by massive I/O transfers, increasing system processor utilization and overall system throughput. Thus, server configuration is no longer dominated by the need to handle I/O bandwidth; rather, it can be based on how best to manage processor resources independently of I/O.

Conclusion

With clustering, computing’s most elusive goal stands within the reach of companies of every size and type. For clusters to deliver fully on their promise of reliable, scalable, affordable and open computing, a robust interconnect technology is needed to map applications to the underlying cluster architecture. ServerNet, an emerging technology, offers fault-tolerant high bandwith and extremely low latency for distributed message passing and clustered server communications. As ServerNet continues to scale into larger clustered server configurations, and with the advent of VI as an open-end standard distributed processing protocol, ServerNet can continue to establish itself as the industry-standard cluster interconnect.

For More Information

To access additional Compaq white papers, please visit http://www.compaq.com/support/techpubs/whitepapers

Related Compaq white papers include:

System Area Networks – A New Approach to Clustering

Virtual Interface Architecture – The New Open Standard for Distributed Messaging Within a Cluster

Compaq E2000 Architecture: Meeting All Levels of Enterprise Computing Requirements with Standards-Based Solutions

Feel free to visit the Virtual Interface Architecture Web site at
http://www.viarch.org