Sequent's NUMA-Q™ Architecture

SEQUENT

Overview

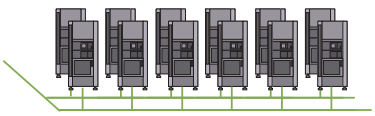
This document overviews the path-breaking technologies introduced by Sequent's® NUMA-Q™ architecture. It describes how the current enterprise-class system architectures are driven by usage models such as on-line transaction processing (OLTP), decision support systems (DSS), and business communications. It also describes Sequent's work to develop a common building block for all enterprise-class system architectures. This includes descriptions of the 4x Intel® Pentium® Pro processor SMP system (quad) building block, a Sequent-developed system interconnect for linking these quads, the architectures to which they can be applied, and the benefits realized from these applications.

Sequent's new NUMA-Q (Non Uniform Memory Access for Quads) architecture yields new levels of performance, availability, and manageability in enterprise-class systems. NUMA-Q is not so much a family of products as it is a quantum leap in symmetric multiprocessing (SMP) and clustered systems architectures, and the realization of highly available and manageable enterprise-class networked server architectures.

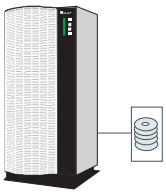
Platform architecture alternatives for the enterprise

Sequent's target market is the commercial data center solving mission-critical problems. Computer systems in these centers have several characteristics in common. They need to be highly available (just minutes of downtime per year), highly reliable, capable of meeting ever increasing performance demands, highly scalable, and finally integrated into a heterogeneous systems management environment. The primary applications, or usage models, found in data center computing today fall into three major categories:

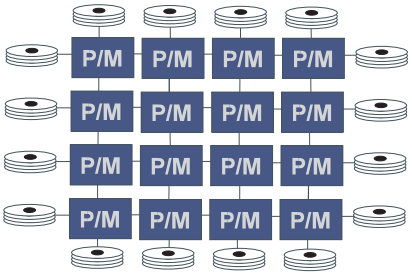
- *OLTP:* On-line transaction processing refers to the day-to-day management of business functions using a relational database.
- *DSS:* Decision support systems refer to the extraction, analysis, and presentation of data from databases to enable decision-making based on operations.
- *Business communications:* Refers to messaging, web servers, document retrieval, and workflow.



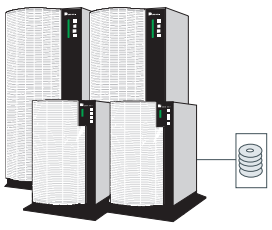
Networked Servers



Large SMP systems



MPP, Shared Nothing systems



Clustered Shared Disk SMP systems

Four fundamental architectures for enterprise-class computing

Architecture	Usage Model	Pros	Cons
Networked servers	Small Digital Libraries	Inexpensive	Management, availability
SMP	DSS, OLTP, Bus. Comms	Easy to program	Limited in size by backplane
Clustered SMP	DSS, OLTP, Bus. Comms	High Availability	Requires more management
MPP	DSS	Can be very large	Data skew problems

OLTP, DSS, and business communications systems designers currently have four architectural options for their computing platforms:

- *Small networked-servers:* Multiple small standalone servers connected over a network.
- *Large SMP nodes:* Many processors and resources running under one operating system.
- *Clustered SMP nodes:* Multiple instances of an application running on separate nodes under separate instances of an operating system, but sharing some storage devices and data.
- *MPP (Massively Parallel Processor) systems:* Many unique instances of an operating system and application, on separate nodes, usually without any shared resource, and communicating by passing messages.

Each of the usage models has different requirements with respect to I/O, memory, processor, and connectivity. Thus, each architecture has characteristics that can be a help or a hindrance depending on the usage model. The choice of architecture is, therefore, largely dependent on the usage model.

Networked servers

The networked servers model suggests that many large computing problems can be solved by a network of small computers or servers. It is true that a collection of networked servers can be successfully and economically applied to some problems such as a small World Wide Web service, or a digital library for presentations and documents in large corporations. However, the networked-servers model is unsuitable for implementing large OLTP, DSS, and business communications applications. It forces an arbitrary distribution of data across servers, leading to the difficult problems of migrating processes or replicating data across a network of servers.

The primary difficulty with implementing a vast network of small servers is in management and availability. Most commodity server companies put more emphasis on cost than reliability and manageability. As a result, networked-server solutions are often driven by a low-cost requirement and suffer availability and manageability problems.

SMP

Large single-node SMP systems have gained popularity, because they are ideally suited to large DSS and OLTP applications. Data managed by an SMP system is centrally located, users share a pool of resources, and SMP systems are easy to manage. Additionally, single SMP nodes make it easy to measure peak performance, and project and plan for future performance needs. Another reason why SMP has become the dominant enterprise architecture is because it provides a smooth migration path for sophisticated uniprocessor applications to high-performance multi-processor systems.

One future drawback for large single-node SMP systems is that the number of processors will be increasingly limited by the size and speed of the backplane and the shared system bus. Physics is the largest contributor to future band-

width limitations. As microprocessor performance continues to dramatically increase, computer system designers are forced to make a bus length/bus speed tradeoff—electrons travel at near light speeds, and no amount of encouragement will speed them up! Large SMP system designs must include shorter backplane/system buses to meet the needs of faster processors and I/O. The smaller backplanes, while faster, support fewer processors simply because of packaging constraints. This architectural limit will constrict the amount of I/O into and out of single-node SMP systems. In the future however, DSS and business communications applications will continue to require increasing amounts of I/O. Another downside to large single-node SMP systems is that there are single points of failure, which can cause application interruptions.

Clustered SMP

The solution to the latter problem is the interconnection of single SMP systems into a cluster of nodes. When implemented to gain availability, clustering provides enough performance on one or more nodes-and access to common resources-to completely replace the unplanned loss of another node. In the event of a single node outage, the other nodes continue to operate and may automatically assume the load of the failed node in a period of minutes. Open-systems relational database management system (RDBMS) companies are evolving their software to support clustered environments to dramatically improve availability beyond what has been practical on traditional single-node SMP systems.

Clusters can also achieve far greater performance and scalability than a single SMP node. This "out of box scaling" is generally due to the increased number of users that can be connected, the increased I/O bandwidth, and the increased amount of processors and memory. It also requires that the application be running on all nodes simultaneously and that nodes communicate before making changes to shared data. The latter point is, in fact, what controls cluster scalability. It is also one of the many factors that limit the application of MPP to business problems. The industry has learned how to make 4-8 SMP nodes communicate effectively, while MPP architectures attempt to pass messages between hundreds of nodes. The immaturity of MPP message-passing software, and the associated overhead, limits MPP's applicability of DSS and OLTP problems. "Out of box" clustered SMP performance is still improving, especially with the advent of software that can take advantage of reflective memory technology such as Sequent's Scalable Data Interconnect (SDI).

The downside of clusters is that they require more thought in management and load balancing. The more nodes, the more complex the problem. The speed and latency of the message passing interconnect is key to improving the scalability of clusters.

MPP

The one overwhelming advantage of MPP architectures is the ability to connect hundreds of processor/memory cells (individual nodes with their own copy of the OS and application). This is also the overwhelming disadvantage. For problems that require an enormous amount of I/O followed by localized computation, and where the intervening results and original data do not have to be shared across the pool of processors, MPP systems can offer satisfactory results. A video server is just such an application. However, for applications that need to scan large data sets in an unpredictable fashion (DSS) or applications that require many updates (and, therefore, locking) like OLTP, the cumbersome messaging of MPP becomes a bottleneck.

This is where SMP's single large memory and processor pool excels. In an SMP system, message-passing between processors is implicit through shared memory and as such is orders of magnitude faster than MPP. The equally short memory access latency of SMP systems makes optimizing performance quite straightforward. The "distributed everything" or "shared nothing" model of most MPP systems is a complex environment that is difficult to program in—thus limiting the range of applications where an MPP system is a suitable choice. MPP systems require software architectures that have not yet been proven in open systems production environments. Currently, MPP vendors are trying to make their processing nodes more powerful by turning them into SMP nodes. The MPP vendors are also beginning to add shared disk capabilities to their systems, made easier because of the advent of optical interconnects. An MPP system made up of many loosely coupled SMP nodes and shared-disk resources is just an SMP cluster. It is becoming apparent that shared-disk clusters are, in fact, the convergence point of MPP and SMP. The old "distributed everything" MPP model will soon be abandoned. This will leave just three effective architectures for implementing mission-critical solutions in open systems: networked-servers, SMP, and clusters.

Sequent's new NUMA-Q architecture

A pervasive trend in the past ten years has been the widespread use of commodity technologies as building blocks for large systems. In the past, merchant microprocessors were the primary building block and were applied all the way from the desktop to the superserver. In the future, however, the microprocessor will be replaced by the 4x Pentium Pro processor quad. The same 4x Pentium Pro processor (quads) used to implement a network of servers can be used to build very large SMP systems.

To overcome the architectural limitations of virtually all the current approaches to building enterprise-class computing systems today, Sequent launched a massive design project in 1992 with the following ambitious goals:

- Creating a new set of CPU and memory-interconnect building blocks for enterprise-class systems beyond the year 2000.
- Meeting the needs of the OLTP, DSS and business communications markets in the same time period.
- Leveraging higher-level integrated components from Intel and other component suppliers.
- Building systems from these building blocks that meet avail-ability and manageability needs.
- Allowing OLTP, DSS, and business communication system designers to choose freely between networked-servers, SMP, and clustered SMP architectures.

In the process, two breakthrough technologies have been invented—the NUMA-Q architecture and the IQ-Link™ interconnect-that will revolutionize system construction in the next ten years. In simplest terms—a system based on the NUMA-Q architecture includes multiple 4x Pentium Pro quad SMP systems tied together with Sequent's new IQ-Link interconnect technology to form a single large computing complex.

The building block: the 4x Pentium Pro processor quad

The Sequent NUMA-Q architecture leverages the new Intel 4x Pentium Pro processor SMP baseboard as a commodity building block for large systems. A four-processor Pentium Pro SMP system may at first glance appear to be a natural progression for 4 processor Pentium systems available in the marketplace today. However, there is one important difference: 4x Pentium Pro processor systems use newer Intel bus logic that allows for third-party control

of the processor bus. Third-party control is the "hook" required to permit the joining of multiple quads together to form a larger system.

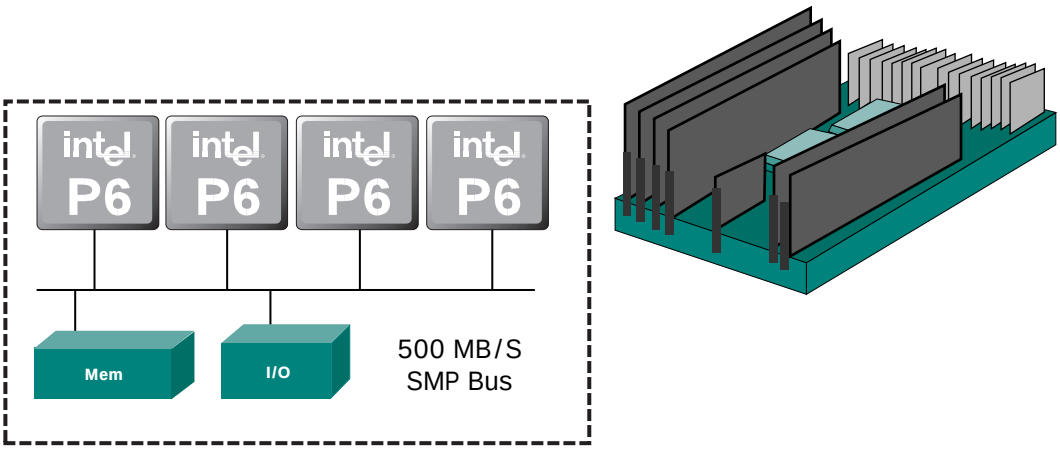
Sequent partnered with Intel to design the 4x Pentium Pro SMP baseboard. However, we made some changes to the 4x quad to make it suitable for high-end, mission-critical, data center applications. These changes include:

- On-Line Replacement/On-Line Insertion of redundant power supplies
- Additional EMI shielding to meet large system requirements
- Greater fault isolation capability
- A 4x Pentium Pro daughter card
- Reliability improvements
- Removal of some unneeded PC logic
- Design of a management and diagnostic processor (MDC)
- Design of our own high-quality memory controller

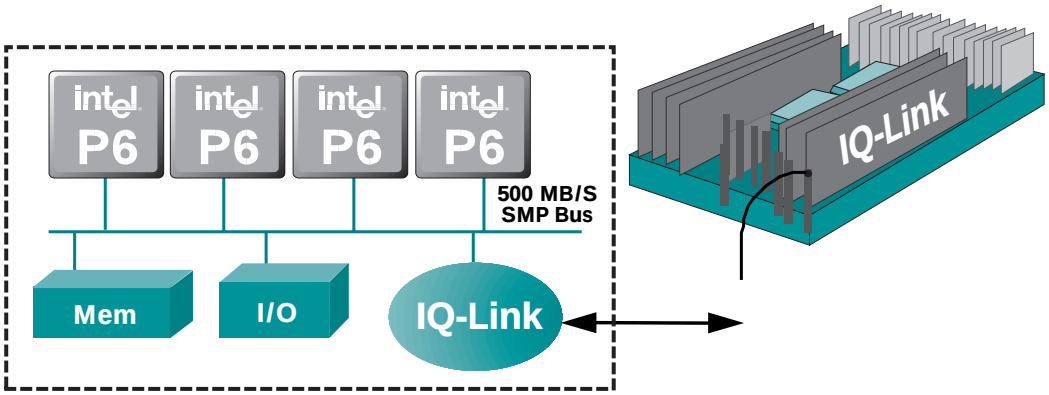
The finished quad comes in its own rackmount box, with two PCI buses accommodating up to seven PCI boards per quad. Each quad includes four processors and between 512 megabytes (MB) to 4 gigabytes (GB) of memory. This is the new building block for enterprise class systems, commercially hardened for mission critical applications by Sequent.

In the quad CPUs, memory and I/O are uniquely arranged. Sequent has essentially pulled memory apart and put pieces of it near each processor. I/O is also closer to each processor, yielding several advantages. The primary advantage is that for all memory and I/O accesses that can be satisfied inside the quad, there is no need to go out and use bandwidth on the interconnect between quads. In today's single-node SMP implementations, all memory and I/O access travel over a single shared bus. In the NUMA-Q architecture, many of these accesses are handled at the quad level. When a memory access does go out on the IQ-Link interconnect, it happens as fast in the new architecture as it does in Sequent's current Symmetry® 5000 systems. Software applications do not have to change to accommodate the architecture.

The value of having memory and I/O in the quad near the processors is that the 500 MB/sec bus that links these together can operate independently of all the other quads until it makes a request that must be fulfilled outside the quad. The effective bus bandwidth of the system is now the summation of all the quads' 500 MB/sec buses, or 32 GB/sec for a 252-processor system. With two PCI buses in each quad, each rated at 133 MB/sec, half of this 32 GB/sec can be used for I/O.



The quad



The quad with an IQ-Link interconnect

The IQ-Link: 1 GB/second SCI link

Engineers at Sequent reasoned that if the industry is at a point where the networked-servers, SMP, and clustered architectures can utilize the same building block to forge a solution, then perhaps one can also craft a common interconnect to suit all these architectures. Why not create a systematic way to manage the quads and the software running on them from a central location, regardless of the architecture? And when you design the quad, interconnect, and software programs, focus on achieving the highest availability possible, because it is key to all three architectures.

The interconnects for processor chips are well understood; they involve caches and buses in SMP systems. Interconnecting systems pose a much greater and less well understood challenge. Sequent has met the challenge of creating an interconnect for the new building blocks of the computer industry, an interconnect that goes beyond just building large SMP systems. Sequent has created an interconnect that can be used to build 252-processor SMP systems, support very large clusters of large SMP systems, or even network hundreds of servers together with unparalleled bandwidth and low latency, all without a backplane.

The Sequent-designed connection technology for linking multiple quads is called IQ-Link. These links can be made memory-coherent (as in creating a single large SMP system from multiple quads), or the links can be used strictly for fast, low-latency message-passing (as in the case of networked-server or clusters architecture). These attributes are also those needed to tie clusters of large SMP servers together to maximize performance. In order for a group of quads to run as a single SMP system (a node), the interconnect creates a single large contiguous coherent view of memory out of the distributed pieces of physical memory found in each quad. Because IQ-Link can provide this unified view of memory, with ranges of the address space parceled out to each quad, one instance of the OS and the applications simultaneously runs on all interconnected quads. The result is a very large, single-node SMP system.

When IQ-Link is used to create a large SMP system, it has the ability to monitor the Pentium Pro quad processor bus and knows when it should respond to requests for specific memory locations (those outside of the range of the portion of memory contained on this quad). IQ-Link examines its own large

cache (L3) for this data, and if it cannot be found there, it puts a request out to the portions of memory on other quads. All of this activity is transparent to the database and application software.

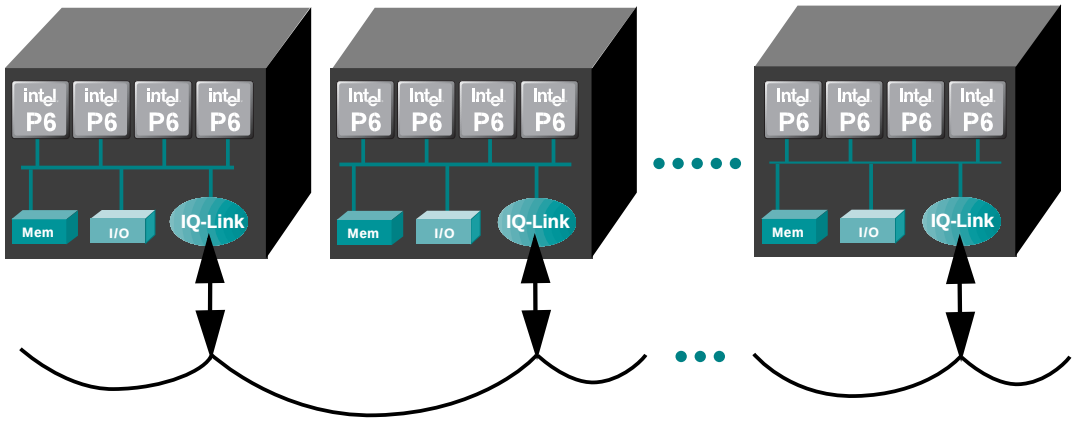
Some memory accesses are resolved quickly, when the data is found in the memory on the same quad that made the request. Other accesses are resolved slower, when the data has to be fetched from a different quad by IQ-Link. This type of architecture is often referred to as CC-NUMA: Cache-Coherent Non Uniform Memory Access. The key to implementing CC-NUMA successfully is to make the fetches to memory in other quads so fast and so rare that the software can effectively ignore it. Poorly implemented interconnects will result in long latency, poor scaling of nodes, and disappointing performance.

The 1 GB/sec IQ-Link is not a backplane. It is in fact a daisy chain connection between quads. It is this point-to-point nature of the IQ-Link that overcomes the length/speed design tradeoff discussed earlier. Indeed, each segment of the IQ-Link is limited in length to achieve high bandwidth, but taken altogether, multiple segments can connect a large number of processors while delivering an aggregate bandwidth of more than 1 GB/sec.

Leveraging key technologies

In NUMA-Q, Sequent is leveraging key developing technologies. SCI, the Scalable Coherent Interconnect, is the basis for the IQ-Link interconnect. Sequent has partnered with Vitesse Corporation to develop a 1 GB/sec data pump chip based on GaAs. This component is integral to the implementation and design of IQ-Link. Sequent's own ASIC designs on IQ-Link are the key to achieving performance far superior to other SCI implementations.

PCI is evolving as the standard I/O bus, replacing VMEbus, for both desktop systems and servers. Fibre Channel is the I/O standard for disks and tapes and provides the reliability, availability, and serviceability, that enterprise class systems need. Lastly, the Pentium Pro processor is SMP-ready and allows third-party control of its bus. Sequent has leveraged all of these emerging technologies in the design and implementation of the NUMA-Q architecture.



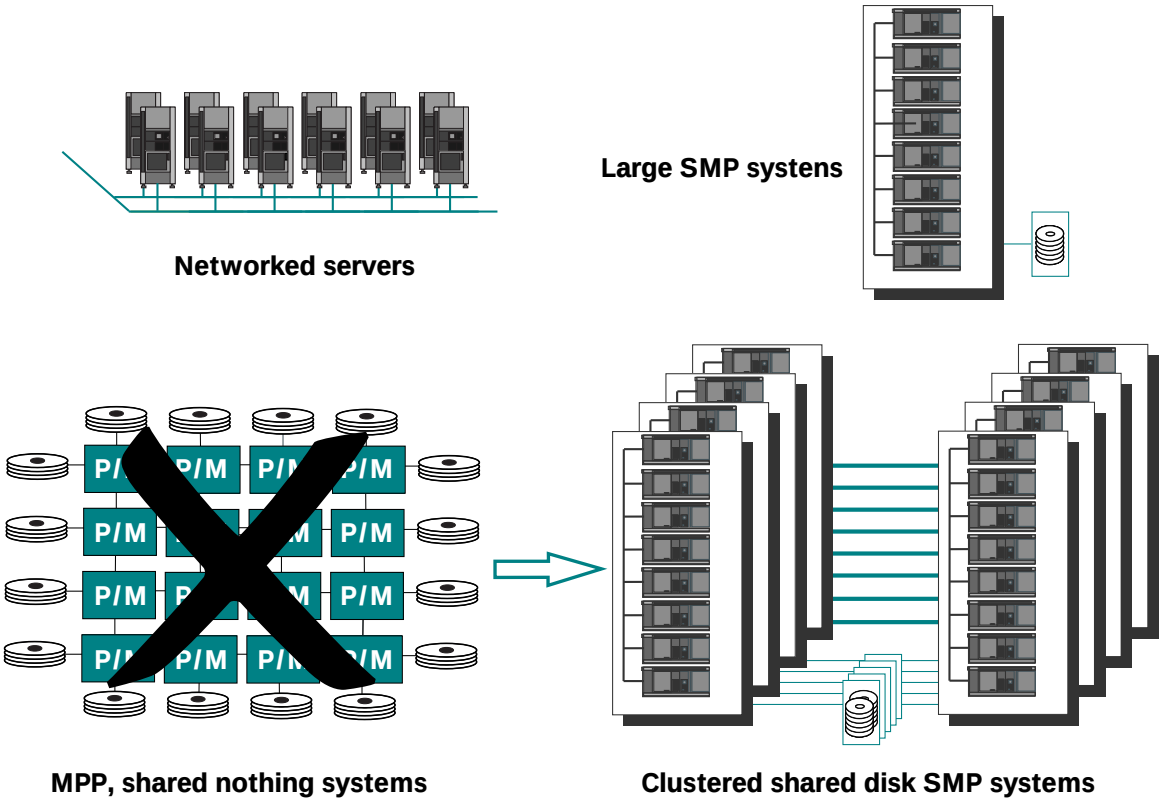
CC-NUMA SMP architecture

Benefits of Sequent's new architecture

Sequent's new NUMA-Q architecture provides an enterprise with a number of important advantages over the next ten years including:

- Flexibility to use one set of components to create a variety of architectures.
- High-end performance leadership.
- Common set of availability and manageability tools for a variety of solutions.
- Higher application availability.
- Higher single-stream and system-wide performance, to support more users and greater throughput.
- Vastly improved system management, reducing operator errors and training, simplifying system administration, and improving application availability.
- Improved serviceability and reliability, reducing customer downtime.
- Increased I/O capacity permitting greater on-line storage and backup.
- Increased memory capacity, permitting more users and faster processing.
- Improved install-ability, reducing the time it takes to install a system and get it operational.
- Headroom for growth of at least 30% per year for the next five years.
- Maximum customer investment protection.
- Binary compatibility with Sequent Symmetry 5000.

The NUMA-Q architecture will allow OLTP, DSS and business communications system designers to build very large mission-critical solutions without the management nightmares of the networked-server implementations, the



The future for the four fundamental architectures of enterprise-class computing

backplane limitation of current SMP systems, and the programming intricacies of the MPP paradigm. In huge information servers comprised of many distinct 4x Pentium Pro quad servers, the new Sequent IQ-Link provides full memory bandwidth communication directly between quads with a fraction of the overhead experienced with networked servers today. In large SMP systems, instead of limited scalability above 30 processors, a single OS instance can manage hundreds of processors. With an effective bus bandwidth of 32 GB/sec, interconnect bandwidth in excess of 1 GB/sec and latencies as low as two microseconds, memory accesses over the IQ-Link to other quads are faster than today's backplane-based SMP systems. This means that applications aren't affected, and in fact never need to know that the backplane has disappeared.

The data center will also see old-style shared-nothing MPP systems gradually evolve into shared-disk clusters of large SMP nodes. For the shared-disk clusters of SMP nodes, the performance of NUMA-Q architecture systems is greater and more predictable than MPPs, management is easier, and availability is higher.

In 1983, Sequent took the lead in leveraging the microprocessor as a building block for creating larger systems and in pioneering the industry move into symmetric multiprocessing. In 1997, Sequent will lead the industry to the next great step for computer architectures: applying one common building block and one common interconnect to all of the architectures needed in enterprise class computing. In successfully leveraging the 4x Pentium Pro processor quad as a building block, and crafting an interconnect that requires no programming changes from the successful SMP model, Sequent is demonstrating the way forward.

Corporate headquarters:
American headquarters:
Sequent Computer Systems, Inc.
15450 SW Koll Parkway
Beaverton, Oregon 97006-6063
(503) 626-5700 or (800) 257-9044
URL: http://www.sequent.com

European headquarters:
Sequent Computer Systems, Ltd.
Sequent House
Unit 3, Weybridge Business Park
Addlestone Road
Weybridge, Surrey KT15 2UF
England
(44) 1932 851111

Asia/Pacific headquarters:
Sequent Computer Systems
(Singapore) Pte Ltd.
80 Robinson Road, #18-03
Singapore 068898
(65) 223-5455

With offices in:
Australia, Austria, Brazil, Czech Republic, France, Germany, Hong Kong, India, Indonesia, Italy, Japan, Korea, Malaysia, Mexico, The Netherlands, New Zealand, Philippines, Poland, Russia, Singapore, Taiwan, Thailand, United Kingdom, and United States.

With distributors in:
Bahrain, Brunei, Croatia, Czech Republic, Egypt, Greece, Hong Kong, Hungary, India, Japan, Korea, Kuwait, Malaysia, Mexico, Oman, People's Republic of China, Philippines, Poland, Russia, Saudi Arabia, Slovenia, South Africa, Sri Lanka, Thailand, Ukraine, United Arab Emirates, and Yugoslavia/Serbia.

PD-1124 6/97

SEQUENT