Scalable Cache Coherent Shared Memory at Cluster Prices

White Paper

NumaConnect™

A high level technical overview of the NumaConnect technology and products

By: Einar Rustad

"Systems based on standard high volume servers interconnected with full-blown distributed cache coherent memory will be extremely interesting for a wide range of compute intensive applications that today can only be efficiently executed on systems costing significantly more", says Professor Petter Bjørstad, Founder of Parallab and Professor and Head of the Department of informatics at the University of Bergen. NW1V2

ABSTRACT

Numascale's NumaConnect™ technology enables computer system vendors to build scalable servers with the functionality of enterprise mainframes at the cost level of clusters. The technology unites all the processors, memory and IO resources in the system in a fully virtualized environment controlled by standard operating systems.

NumaConnect enables significant cost savings in three dimensions; resource utilization, system management and programmer productivity.

According to long time users of both large shared memory systems and clusters in environments with a variety of applications, the former provide a much higher degree of resource utilization due to the flexibility of all system resources.

Contents

1. Background

Numascale's NumaConnect™ technology enables computer system vendors to build scalable servers with the functionality of enterprise mainframes at the cost level of clusters. The technology unites all the processors, memory and IO resources in the system in a fully virtualized environment controlled by standard operating systems.

Systems based on NumaConnect will efficiently support all classes of applications using shared memory or message passing through all popular high level programming models. System size can be scaled to 4k nodes where each node can contain multiple processors. Memory size is limited by the 48-bit physical address range provided by the Opteron processors resulting in a total system main memory of 256 TBytes.

At the heart of NumaConnect is NumaChip; a single chip that combines the cache coherent shared memory control logic with an on-chip 7-way switch. This eliminates the need for a separate, central switch and enables linear capacity and cost scaling.

The continuing trend with multi-core processor chips is enabling more applications to take advantage of parallel processing. NumaChip leverages the multi-core trend by enabling applications to scale seamlessly without the extra programming effort required for cluster computing. All tasks can access all memory and IO resources. This is of great value to users and the ultimate way to virtualization of all system resources.

No other interconnect technology outside the high-end enterprise servers can offer this capability.

All high speed interconnects now use the same kind of physical interfaces resulting in almost the same peak bandwidth. The differentiation is in latency for the critical short transfers, functionality and software compatibility. NumaConnect™ differentiates from all other interconnects through the ability to provide unified access to all resources in a system and utilize caching techniques to obtain very low latency.

Key Facts:

- Scalable, directory based Cache Coherent Shared Memory interconnect for Opteron
- Attaches to coherent HyperTransport (cHT) through HTX connector, pick-up module or mounted directly on mainboard
- Configurable Remote Cache for each node
- Full 48 bit physical address space (256 Tbytes)
- Up to 4k (4096) nodes
- ≈1 microsecond MPI latency (ping-pong/2)
- On-chip, distributed switch fabric for 2 or 3 dimensional torus topologies

1.1 Expanding the capabilities of multi-core processors

Semiconductor technology has reached a level where processor frequency can no longer be increased much due to power consumption with corresponding heat dissipation and thermal handling problems. Historically, processor frequency scaled at approximately the same rate as transistor density and resulted in performance improvements for most all applications with no extra programming efforts. Processor chips are now instead being equipped with multiple processors on a single die. Utilizing the added capacity requires software that is prepared for parallel processing. This is quite obviously simple for individual and separated tasks that can be run independently, but is much more complex for speeding up single tasks.

The complexity for speeding up a single task grows with the logic distance between the resources needed to do the task, i.e. the fewer resources that can be shared, the harder it is. Multi-core processors share the main memory and some of the cache levels, i.e. they are classified as Symmetrical Multi Processors (SMP). Modern processors chips

are also equipped with signals and logic that allow connecting to other processor chips still maintaining the same logic sharing of memory. The practical limit is at two to four processor sockets before the overheads reduce performance scaling instead of increasing it. This is normally restricted to a single motherboard.

Currently, scaling beyond the single/dual SMP motherboards is done through some form of network connection using Ethernet or a higher speed interconnect like InfiniBand. This requires processes running on the different compute nodes to communicate through explicit messages. With this model, programs that need to be scaled beyond a small number of processors have to be written in a more complex way where the data can no longer be shared among all processes, but need to be explicitly decomposed and transferred between the different processors' memories when required.

NumaConnect™ uses a scalable approach to sharing all memory based on distributed directories to store information about shared memory locations. This means that programs can be scaled beyond the limit of a single motherboard without any changes to the programming principle. Any process running on any processor in the system can use any part of the memory regardless if the physical location of the memory is on a different motherboard.

2 NumaConnect Value Proposition

NumaConnect enables significant cost savings in three dimensions; resource utilization, system management and programmer productivity.

According to long time users of both large shared memory systems (SMPs) and clusters in environments with a variety of applications, the former provide a much higher degree of resource utilization due to the flexibility of all system resources. They indicate that large mainframe SMPs can easily be kept at more than 90% utilization and that clusters seldom can reach more than 60-70% in environments running a variety of jobs. Better compute resource utilization also contributes to more efficient use of the necessary infrastructure with power consumption and cooling as the most prominent ones (account for approximately one third of the overall coast) with floor-space as a secondary aspect.

Regarding system management, NumaChip can reduce the number of individual operating system images significantly. In a system with 100Tflops computing power, the number of system images can be reduced from approximately 1 400 to 40, a reduction factor of 35. Even if each of those 40 OS images require somewhat more resources for management than the 1 400 smaller ones, the overall savings are significant.

Parallel processing in a cluster requires explicit message passing programming whereas shared memory systems can utilize compilers and other tools that are developed for multi-core processors. Parallel programming is a complex task and programs written for message passing normally contain 50%-100% more code than programs written for shared memory processing. Since all programs contain errors, the probability of errors in message passing programs is 50%-100% higher than for shared memory programs. A significant amount of software development time is consumed by debugging errors further increasing the time to complete development of an application.

In principle, servers are multi-tasking, multi-user machines that are fully capable of running multiple applications at any given time. Small servers are very cost-efficient measured by a peak price/performance ratio because they are manufactured in very high volumes and use many of the same components as desk-side and desktop computers. The price per CPU socket is now less than USD 2,000. However, these small to medium sized servers are not very scalable. The most widely used configuration has 2 CPU sockets that hold from 4 to 16 CPU cores each. They cannot be upgraded with-

out changing to a different main board that also normally requires a larger power supply and a different chassis. In turn, this means that careful capacity planning is required to optimize cost and if compute requirements increase, it may be necessary to replace the entire server with a bigger and much more expensive one since the price increase is far from linear. For the most expensive servers, the price per CPU socket is the range of USD 50,000 – 60,000.

NumaChip contains all the logic needed to build Scale-Up systems based on volume manufactured server components. This drives the cost per CPU core down to the same level while offering the same capabilities as the mainframe type servers.

Where IT budgets are in focus the price difference is obvious and NumaChip represents a compelling proposition to get mainframe capabilities at the cost level of high-end cluster technology. The expensive mainframes still include some features for dynamic system reconfiguration that NumaChip systems will not offer initially. Such features depend on operating system software and can be also be implemented in NumaChip based systems.

3 Technology

3.1 Multi-core processors and shared memory

Shared memory programming for multiprocessing boosts programmer productivity since it is easier to handle than the alternative message passing paradigms. Shared memory programs are supported by compiler tools and require less code than the alternatives resulting in shorter development time and fewer program bugs. The availability of multi-core processors on all major platforms including desktops and laptops is driving more programs to take advantage of the increased performance potential.

NumaChip offers seamless scaling within the same programming paradigm regardless of system size from a single processor chip to systems with more than 1,000 processor chips.

Other interconnect technologies that do not offer cc-NUMA capabilities require that applications are written for message passing, resulting in larger programs with more bugs and correspondingly longer development time while systems built with NumaChip can run any program efficiently.

3.2 Virtualization

The strong trend of virtualization is driven by the desire of obtaining higher utilization of resources in the datacenter. In short, it means that any application should be able to run on any server in the datacenter so that each server can be better utilized by combining more applications on each server dynamically according to user loads.

Commodity server technology represents severe limitations in reaching this goal. One major limitation is that the memory requirements of any given application need to be satisfied by the physical server that hosts the application at any given time. In turn, this means that if any application in the datacenter shall be dynamically executable on all of the servers at different times, all of the servers must be configured with the amount of memory required by the most demanding application, but only the one running the app will actually use that memory. This is where the mainframes excel since these have a flexible shared memory architecture where any processor can use any portion of the memory at any given time, so they only need to be configured to be able to handle the most demanding application in one instance. NumaChip offers the exact same feature, by providing any application with access to the aggregate amount of memory in the system. In addition, it also offers all applications access to all I/O devices in the system through the standard virtual view provided by the operating system.

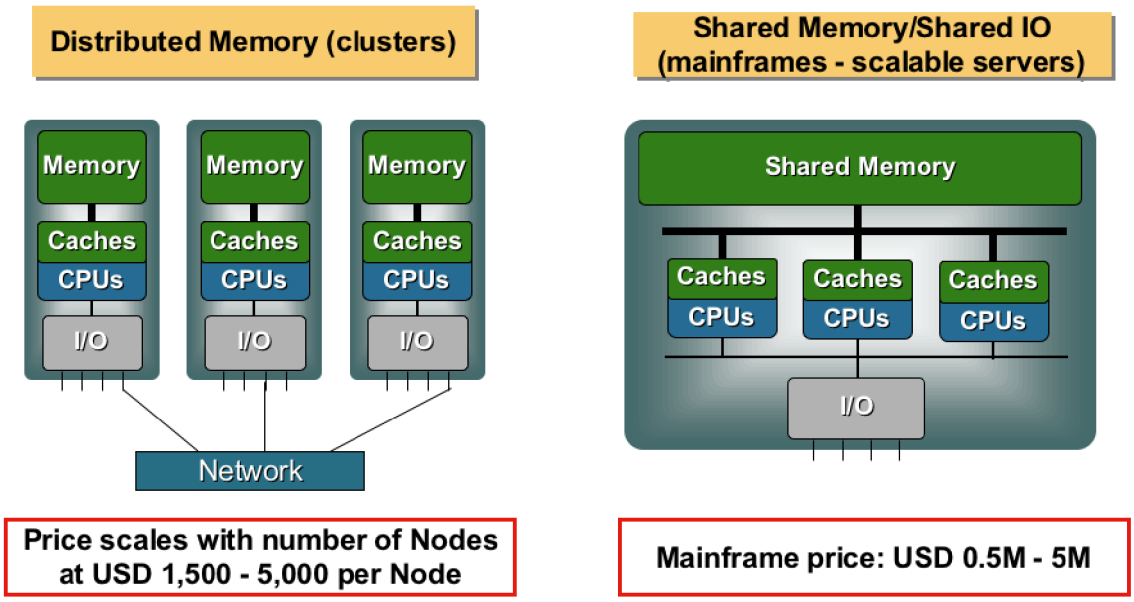

Figure 1, Clustered vs Mainframe Architecture

The two distinctly different architectures of clusters and mainframes are shown in Figure 1. In Clusters processes are loosely coupled through a network like Ethernet or InfiniBand. An application that needs to utilize more processors or I/O than those present in each server must be programmed to do so from the beginning. In the mainframe, any application can use any resource in the system as a virtualized resource and the compiler can generate threads to be executed on any processor.

In a system interconnected with NumaChip, all processors can access all the memory and all the I/O resources in the system in the same way as on a mainframe. NumaChip provides a fully virtualized hardware environment with shared memory and I/O and with the same ability as mainframes to utilize compiler generated parallel processes and threads.

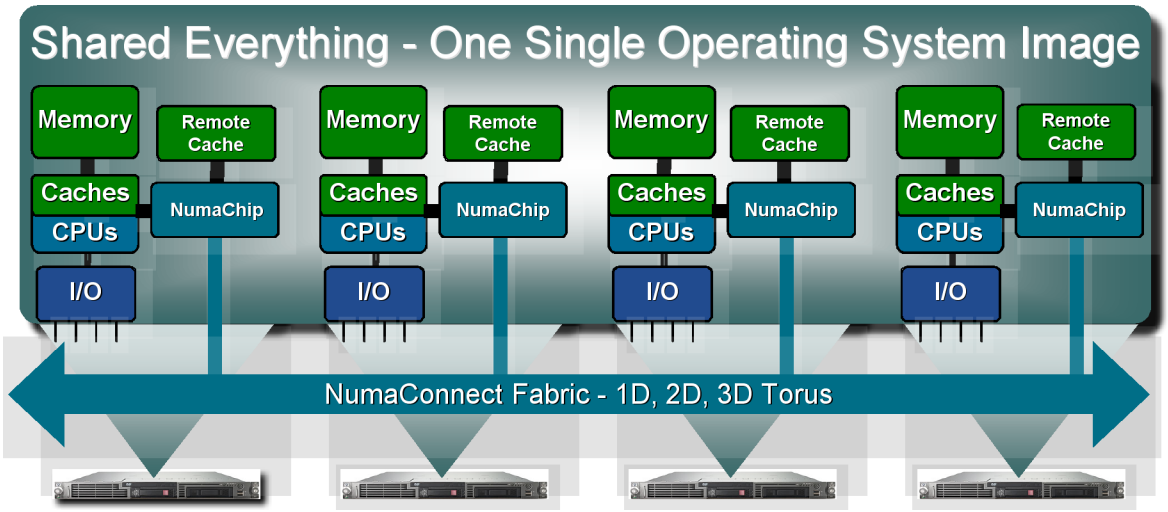

Figure 2, NumaChip System Architecture

3.3 Operating Systems

Systems based on NumaChip can run standard operating systems that handle shared memory multiprocessing. Examples of such operating systems are Linux, Solaris and Windows Server. Numascale provides a bootstrap loader that is invoked after powerup and performs initialization of the system by setting up node address routing tables. Initially Numascale has tested and provides bootstrap for Limux.

When the standard bootstrap loader is launched, the system will appear as a large unified shared memory system.

3.4 Cache Coherent Shared Memory

The big differentiator for NumaConnect compared to other high-speed interconnect technologies is the shared memory and cache coherency mechanisms. These features allow programs to access any memory location and any memory mapped I/O device in a multiprocessor system with high degree of efficiency. It provides scalable systems with a unified programming model that stays the same from the small multicore machines used in laptops and desktops to the largest imaginable single system image machines that may contain thousands of processors.

There are a number of pros for shared memory machines that lead experts to hold the architecture as the holy grail of computing compared to clusters:

- Any processor can access any data location through direct load and store operations - easier programming, less code to write and debug
- Compilers can automatically exploit loop level parallelism – higher efficiency with less human effort
- System administration relates to a unified system as opposed to a large number of separate images in a cluster – less effort to maintain
- Resources can be mapped and used by any processor in the system – optimal use of resources in a virtualized environment
- Process scheduling is synchronized through a single, real-time clock - avoids serialization of scheduling associated with asynchronous operating systems in a cluster and the corresponding loss of efficiency

These features are all available in high cost mainframe systems from IBM, Oracle (Sun), HP and SGI. The only catch is that these systems hold a price tag that is up to 30 times higher per CPU core compared with commodity servers. In the low end, the multiprocessor machines from Intel and AMD have proven multiprocessing to be extremely popular with the commodity price levels: Dual processor socket machines are by far selling in the highest volumes.

3.5 Scalability and Robustness

The initial design aimed at scaling to very large numbers of processors with 64-bit physical address space with 16 bits for node identifier and 48 bits of address within each node. The current implementation for Opteron is limited by the global physical address space of 48 bits, with 12 bits used to address 4 096 physical nodes for a total physical address range of 256 Terabytes.

A directory based cache coherence protocol was developed to handle scaling with significant number of nodes sharing data to avoid overloading the interconnect between nodes with coherency traffic which would seriously reduce real data throughput.

The basic ring topology with distributed switching allows a number of different interconnect configurations that are more scalable than most other interconnect switch fabrics. This also eliminates the need for a centralized switch and includes inherent redundancy for multidimensional topologies.

Functionality is included to manage robustness issues associated with high node counts

and extremely high requirements for data integrity with the ability to provide high availability for systems managing critical data in transaction processing and realtime control. All data that may exist in only one copy are ECC protected with automatic scrubbing after detected single bit errors and automatic background scrubbing to avoid accumulation of single bit errors.

3.6 Integrated, distributed switching

NumaChip contains an on-chip switch to connect to other nodes in a NumaChip based system, eliminating the need to use a centralized switch. The on-chip switch can connect systems in one, two or three dimensions. Small systems can use one, medium sized system two and large systems will use-all three dimensions to provide efficient and scalable connectivity between processors.

The two- and three-dimensional topologies (called Torus) also have the advantage of built-in redundancy as opposed to systems based on centralized switches, where the switch represents a single point of failure.

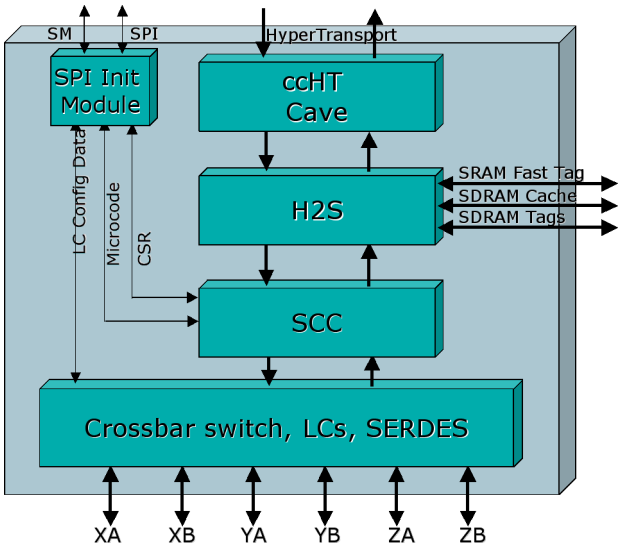

Figure 3, Block Diagram of NumaChip

The distributed switching reduces the cost of the system since there is no extra switch hardware to pay for. It also reduces the amount of rack space required to hold the system as well as the power consumption and heat dissipation from the switch hardware and the associated power supply energy loss and cooling requirements.

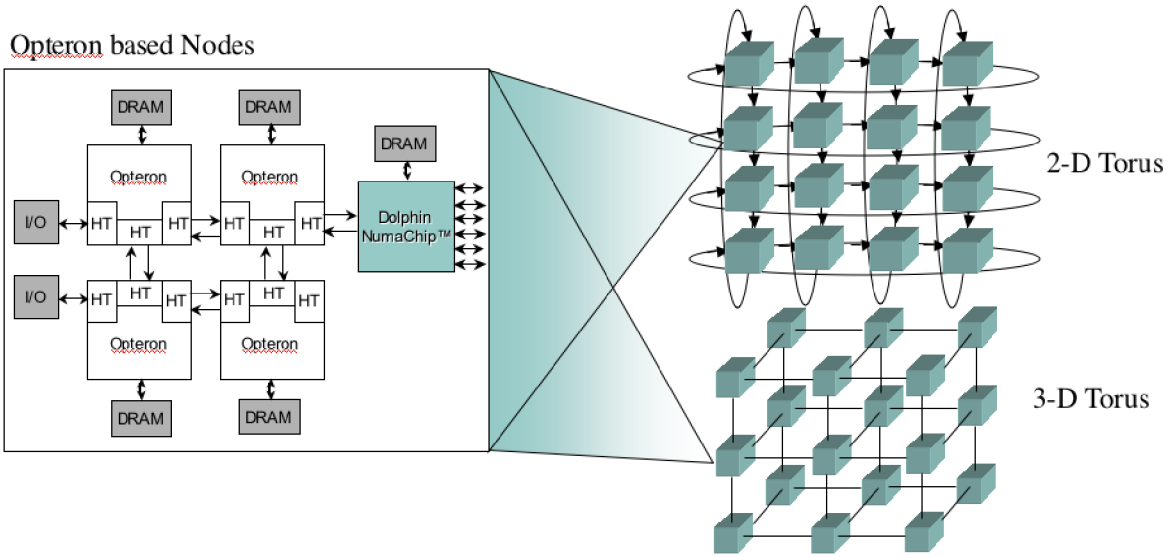

Figure 4, System Topology examples