White Paper

The Numascale Solution: Extreme BIG DATA Computing

By: Einar Rustad

ABOUT THE AUTHOR
Einar Rustad is CTO of Numascale and has a background as CPU, Computer Systems and HPC Systems De-signer and R&D manager. He has worked in HPC System Software companies and with Business Development

“Annual NGS capacity now exceeds 13 quadrillion base pairs (the As, Ts, Gs, and Cs that make up a DNA sequence). Each base pair represents roughly 100 bytes of data (raw, analyzed, and interpreted).”

ABSTRACT

Reports indicate that we generate “2.5 quintillion bytes of data [every day] — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This is big data.”

Traditional clusters based on distributed memory cannot adequately handle this crush of data. Shared memory approaches are required. Numascales technology provides an affordable solution. It delivers all the advantages of shared memory computing – streamlined application development, the ability to compute on large datasets, the ability to run more rigorous algorithms, enhanced scalability, etc. – at a substantially reduced cost.

In fact, early Numascale customers such as Statoil and the University of Oslo are already proving the case and gaining performance and cost advantages and Supermicro has deliverd a new record sized system with 5184 CPU cores and 20.7 TBytes of shared main memory.

Big Data and Data Analytics

Big Data applications – once limited to a few exotic disciplines – are steadily becoming the dominant feature of modern computing. In industry after industry, massive datasets are being generated by advanced instruments and sensor technology. Consider just one example, next generation DNA sequencing (NGS). Annual NGS capacity now exceeds 13 quadrillion base pairs (the As, Ts, Gs, and Cs that make up a DNA sequence). Each base pair represents roughly 100 bytes of data (raw, analyzed, and interpreted). Turning the swelling sea of genom-ic data into useful biomedical information is a classic Big Data challenge, one of many, that didn’t exist a decade ago.

This mainstreaming of Big Data is an important transformational moment in computation. Datasets in the 10-to-20 Terabytes (TB) range are increasingly common. New and advanced algorithms for memory intensive applications in Energy (Power Grid Analytics, Oil & Gas e.g. seismic data processing), finance (real-time trading), social media (database), and science (simulation and data analysis), to name but a few, are hard or impossible to run efficiently on commodity clusters.

> *“Any application requiring a large memory footprint can benefit from a shared memory computing environment”*
>
> William W. Thigpen, Chief, Engineering Branch, NASA Advanced Supercomputing (NAS) Division.

The challenge is that traditional cluster computing based on distributed memory – which was so successful in bringing down the cost of high performance computing (HPC) – struggles when forced to run applications where memory requirements exceed the capacity of a single node. Increased interconnect latencies, longer and more complicated software development, inefficient system utilization, and additional administrative overhead are all adverse factors. Conversely, traditional mainframes running shared memory architecture and a single instance of the OS have always coped well with Big Data Crunching jobs.

“Any application requiring a large memory footprint can benefit from a shared memory computing environment,” says William W. Thigpen, Chief, Engineering Branch, NASA Advanced Supercomputing (NAS) Division. “We first became interested in shared memory to simplify the program-ming paradigm. So much of what you must do to run on a traditional system is pack up the messages and the data and account for what happens if those messages don’t get there successfully and things like that - there is a lot of error processing that occurs.”

> *“If you truly take advantage of the shared memory architecture you can throw away a lot of the code you have to develop to run on a more traditional system. I think we are going to see a lot more people looking at this type of environment”*
>
> William W. Thigpen, Chief, Engineering Branch, NASA Advanced Supercomputing (NAS) Division.

“If you truly take advantage of the shared memory architecture you can throw away a lot of the code you have to develop to run on a more traditional system. I think we are going to see a lot more people look-ing at this type of environment,” Thigpen says. Not only is development eased, but throughput and accuracy are also improved, the latter by allowing execution of more computationally demanding algorithms.

Numascale’s Solution

Until now, the biggest obstacle to wider use of shared memory computing has been the high cost of mainframes and high-end ‘super-servers’. Given the ongoing proliferation of Big Data applications, a more efficient and costeffective approach to shared memory computing is needed. Now, Numascale has developed a technology, which turns a collection of stan-dard servers with separate memories and I/O into a unified system that delivers the functionality of high-end enterprise servers and mainframes at a fraction of the cost.

Numascale Technology snapshot (board and chip):

- Numascale hardware links commodity servers together to form a single unified system where all processors can coherently access and share all memory and I/O. The combined system runs a single instance of a standard operating system like Linux. At the heart of NumaConnect is NumaChip – a single chip that combines the cache coherent shared memory control logic with an onchip 7-way switch. This eliminates the need for a separate, central switch and enables linear capacity and cost scaling in 2D or 3D Torus topologies.

- Numascale systems 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 only by the 48-bit physical address range provided by the Opteron processors resulting in a record-breaking total system main memory of 256 TBytes. (For details of Numascale technol-ogy see http://www.numascale.com/numa_ pdfs/ numaconnect-white-paper.pdf)

"Numascale systems running with entire data sets in memory are orders of magnitude faster than clusters or systems based on any form of existing mass- storage devices and will enable data analysis and decision support applications to be applied in new and innovative ways"

Morten Toverud, Numascale CEO.

The result is an affordable, shared memory computing option to tackle data-intensive applications. Numascale systems running with entire data sets in memory are "orders of magnitude faster than clusters or systems based on any form of existing mass-storage devices and will enable data analy-sis and decision support applications to be applied in new and innovative ways," says Morten Toverud, Numascale CEO.

The big differentiator for Numascle 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 and tens to hundreds of terabytes of main memory.

Early adopters are already demonstrating performance gains and costs savings. A good example is Statoil, the global energy company based in Norway. Processing seismic data requires massive amounts of floating point operations and is normally performed on clusters. Broadly speaking, this kind of processing is done by programs developed for a messagepassing paradigm (MPI).

Not all algorithms are suited for the message passing paradigm and the amount of code required is huge and the development process and debugging task are complex.

Shorten Time To Solution

“We have used development funds to cre-ate a foundation for a simpler program-ming model. The goal is to reduce the time it takes to implement new mathematical models for the computer,” says Knut Sebastian Tungland Chief Engineer IT, Statoil. To address this issue, Statoil has set up a joint research project with Numascale who has developed technology to interconnect multiple computers to form a single system with cache coherent shared memory.

“The goal is to reduce the time it takes to implement new mathematical models for the computer”

Knut Sebastian Tungland, Statoil

Statoil was able to run a preferred application to analyze large seismic datasets on a Numascale system – something that wasn’t practical on a traditional cluster because of the application’s access pattern to memory. Not only did use of the more rigorous application produce more accurate results, but the system completed the task more quickly. Statoil is gaining improved accuracy.

“Time is an expensive resource,” says Trond Jarl Suul, senior manager for High Perfor-mance Computing in Statoil. Many applications would benefit from shared memory architecture since the task of programming for distributed memory is much more complicated and time consuming he says.

The technology from Numascale is based on a chip that handles all the complexity of managing the coherency of the entire memory hierarchy of many interconnected computers to make it appear as one giant memory for all processor cores in the system.

“A lot of time is lost by having to move data in and out of the machine. We have memory hungry algorithms that can make better pictures of the geology faster given proper memory and processing capacity”

Trond Jarl Suul, senior manager for High Perfor-mance Computing in Statoil

Even though the time it takes to access data that resides in a part of the memory that is not local to the requesting processor may be up to an order of magnitude longer than a local access (1µs vs. 100ns) it is still 3-5 orders of magnitude faster than fetching data from mass storage for distribution to separate memories of nodes in a traditional cluster, according to Jarl Suul.

“A lot of time is lost by having to move data in and out of the machine. We have memory hungry algorithms that can make better pictures of the geology faster given proper memory and processing capacity,” says Jarl Suul. This is the reason for Statoil to try out the Numascale technology. A pilot system installed in Numascale’s lab is already being used for this purpose and the next step is to install a larger system where Statoil can run and experiment with a multitude of algorithms and applications that face limitations in cluster environments.

"A lot of time is lost by having to move data in and out of the machine. We have memory hungry algorithms that can make better pictures of the geology faster given proper memory and processing capacity"

Jarl Suul, Statoil

A second example is deployment of a large NumaConnect-based system at the University of Oslo. In this instance, the effort is being funded by the EU project PRACE (Partnership for Advanced Computing in Europe) and includes a 72-node system in 3x6x4 topology. Some of the main applica-tions planned in Oslo include bioscience and computational chemistry. The overall goal is to broadly enable Big Data comput-ing at the university.

"We focus on providing our users with flexible computing resources including capa-bilities for handling very large datasets like those found in applications for next generation sequencing for life sciences" says Dr. Ole W. Saastad, Senior Analyst and HPC expert at USIT, the University of Oslo's central IT resource department. "Our new system with NumaConnect contains 1728 processor cores and 4.6TBytes of memory. The system can be used as one single system or partitioned in smaller systems where each partition runs one instance of the OS. With proper Numa-awareness, applications with high bandwidth requirements will be able to utilize the combined bandwidth of all the memory controllers and still be able to share data with low latency access through the coherent shared memory."

Dr. Saastad continues to say that the impact of clusters with the requirement of coding parallel applications with message passing limits the productivity of the scientists that are not trained in MPI programming. This has limited the amount of new applica-tions with large data sets that run well on clusters. Systems with Numascale technology now provide shared memory capabilities with the same cost structure as a cluster. This represents a compelling solution for scientists that are used to work with their SMP codes on x86 desktops and laptops to scale up their datasets without any extra effort within a familiar standard OS environment.

The applications that will run most frequently on the NumaConnect system will be genome sequence assemblers and other related applications like BLAST where fast access to an in-memory database is key for overall application performance.

Numascale's Advantages

Significantly more affordable than comparable highend enterprise servers, Numascale systems deliver all the benefits and performance gains of shared memory architecture. Here are a few of the advantages shared memory machines 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

Another advantage shared memory computing delivers is significantly improved utilization. At data centers running many different applications on large clusters, those applications with large memory requirements tend to oversubscribe the nodes with large memory leaving smaller nodes undersubscribed. Indeed, many organizations with a diverse set of applications that switched from mainframes and 'superservers' to less expensive clusters have been hit with declining utilization rates.

Other important benefits include:

- Reduced Administration. Less effort is required to administer a unified system compared to one with many separate images in a cluster. In a system with 100Tflops computing power, the number of system images can be reduced from approximately 600 to 6, a reduction factor of 100.

- MPI Performance. If necessary you can still run MPI programs, and NumaConnect provides superior MPI latency performance– on the order of microseconds for all com-mon message sizes.

Numascale's implementation of shared memory has many innovations. Its on-chip switch, for example, can connect systems in one, two, or three dimensions (e.g. 2D and 3D Torus). 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 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 associ-ated power supply energy loss and cooling requirements.

Scalability

Several applications have already been scaled far beyond the previous limitations with current NumaConnect systems. The large Supermicro system is expected to take this further with its 5184 CPU cores and 20.7 TBytes of shared main memory within a single image Linux OS instance.

Conclusion

Reports indicate that we generate "2.5 quintillion bytes of data [every day] — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This data is big data."

Traditional clusters based on distributed memory cannot adequately handle this crush of data. Shared memory approaches are required. The NumaConnect technology provides an affordable solution. It deliv-ers all the advantages of shared memory computing – streamlined application de-velopment, the ability to compute on large datasets, the ability to run more rigorous algorithms, enhanced scalability, etc. – at a substantially reduced cost.

In fact, early Numascale customers such as Statoil and the University of Oslo are already proving the case and gaining perfor-mance and cost advantages.

For more infor-mation about Numascale and NumaConnect-based systems contact: Morten Toverud (mt@numascale.com) or
Einar Rustad (er@numascale.com) or visit Numascale at
http://www.numascale.com.

Footnote: Portions of the Statoil material are taken from Computerworld, Norway, Feb 2013, http://www.idg.no/computerworld/