Data Intensive Computing

How SGI® Altix® ICE and Intel® Xeon® Processor 5500 Series (Code-named Nehalem) Help Sustain HPC Efficiency Amid Explosive Data Growth

Christian Tanasescu & Thomas Reed
Silicon Graphics Inc.

SiliconGraphics

TABLE OF CONTENTS

Executive Summary

Data is a problem that just keeps on growing. Swelling at more than four times the rate of Moore's Law, large data is placing unprecedented challenges on critical IT infrastructures. Yet in the face of this unrelenting growth, users still must find a way to manage, transport and process that information. They must use that data to rapidly achieve insights and take action.

Via a new computing methodology called Data Intensive Computing, enterprises and high-performance computing (HPC) users can achieve a sustainable infrastructure that allows all users, no matter where they are located or what systems they use, to exploit this data to its greatest advantage.

With Data Intensive Computing, organizations can progressively filter and reduce massive data volumes into information that helps people make better decisions sooner. An essential part of implementing a Data Intensive Computing environment involves centralizing data processing and analysis – moving processing to data, where possible, rather than data to processing.

It's also critical to find ways to maximize the efficiency of data movement between discrete devices within a system. In fact, that's more important than ever, due to a widening gap between memory speed and the accelerating growth of data volumes and CPU performance.

This paper examines how the ***SGI Altix ICE*** integrated blade platform and the recently launched ***Intel® Xeon® Processor 5500 Series (code-named Nehalem)*** address the enormous processing requirements of Data Intensive Computing. The paper includes real-world performance data to show how the SGI Altix ICE architecture, and the Intel® Xeon® Processor 5500 Series combine to deliver improvements in *memory bandwidth*, *I/O communications* and *scalability* that provide unprecedented and reliable performance efficiency.

> *"To achieve breakthroughs in designing and testing next-generation materials and weapons systems, DOD researchers require high-performance systems that are scalable and reliable. Throughout our evaluation process, Silicon Graphics clearly demonstrated that it more than met these criteria, particularly with SGI Altix ICE systems featuring next-generation Intel® Xeon® (Nehalem) processors."*
>
> —Cray Henry
> Director, HPCMP, Department of Defense

The IT Infrastructure Problem

We are rapidly surpassing the capability of the current generation of IT infrastructure to manage data. IDC estimates the size of our digital universe exceeds 281 exabytes and is anticipated to increase tenfold before 2011.

The problem is already apparent. Users report they can't find data fast enough. And due to massive file sizes, data can't be moved across the network to where it must be processed fast enough to enable timely insight and action.

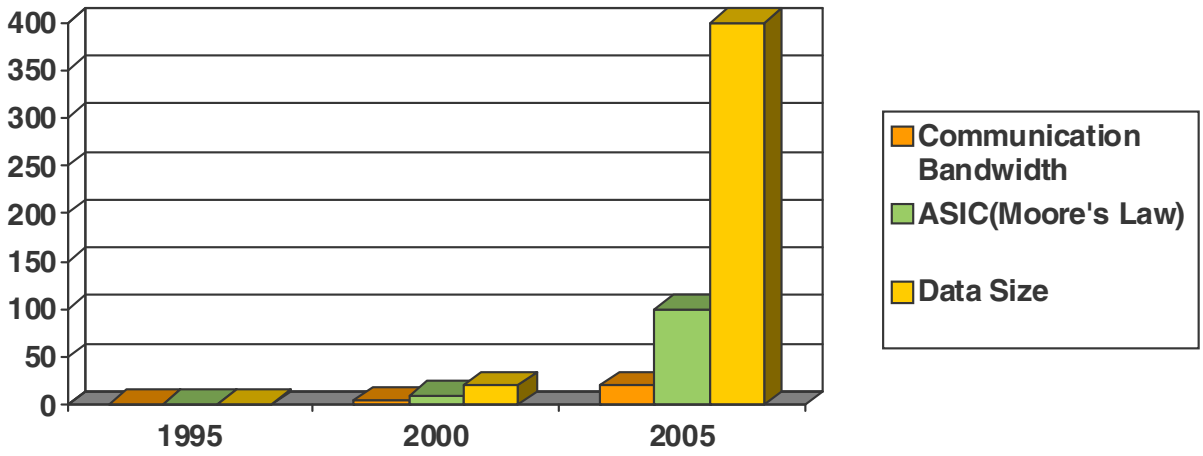


Figure 1 The evolution of communication bandwidth technologies hasn't kept pace with Moore's Law, even as the data they carry is growing at a rate that outpaces Moore's Law by four times.

A key source of this problem is communication bandwidth. Figure 1 illustrates how communication speed and the amount of data that today's enterprises must manage are not on the same path. The chart reveals two trends:

Even as the evolution of discreet devices (CPUs, GPUs, RAID controllers, Network Adaptors, etc.) has tracked the curve of Moore's Law (the benchmark for computer technology evolution), data is growing at approximately four times the rate defined by Moore's Law. This has made communication speed between discreet devices more relevant than ever to overall processing efficiency.

With communication speed evolving at just one-fifth the pace of Moore's Law, data is growing approximately 20 times faster than the ability of systems to move data between components. That means application performance will even suffer more tomorrow than today.

For the past decade, parallel systems have decomposed and distributed data into chunks that both the discreet devices and communication bandwidth could handle. But in the face of soaring data volumes, parallel architectures alone cannot sustain processing efficiency. What's called for is a new paradigm.

Data Intensive Computing

In Data Intensive Computing, data size is the long pole in the performance tent. What once was the exclusive domain of HPC environments for science and engineering, big data now permeates a range of data centers. Whether in fraud detection for e-commerce, improved security through better intelligence gathering, or 3D high-definition television or movies over the Web, the age of Data Intensive Computing is upon us.

To better understand Data Intensive Computing, let's look at a highly conceptual view of the end-to-end problem.

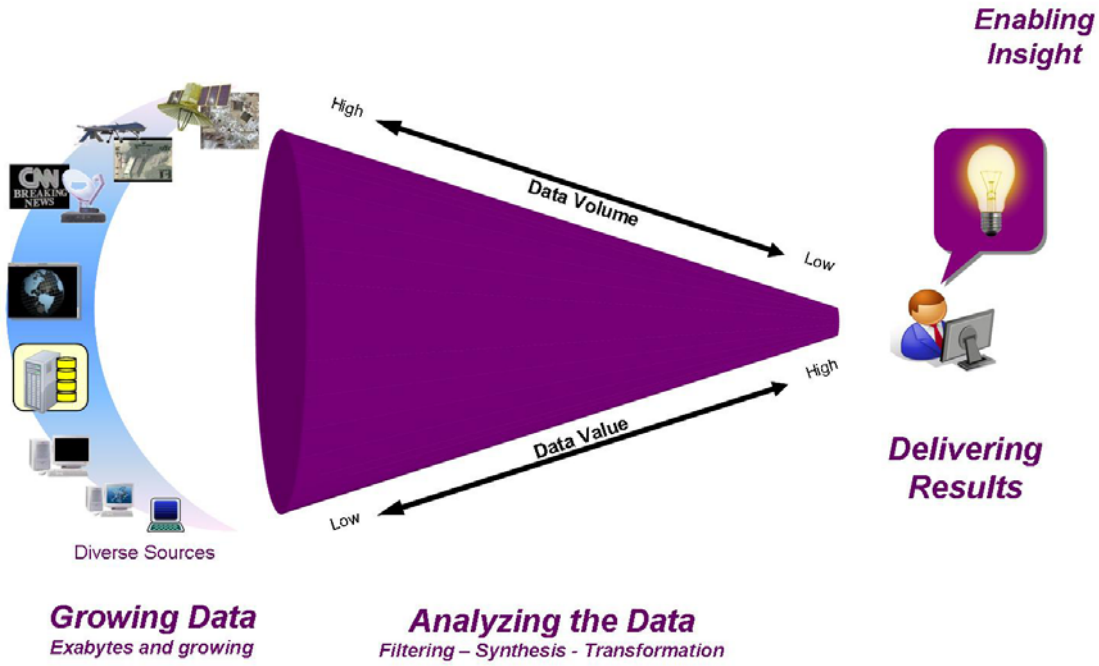


Figure 2 Conceptually, the data intensive computing problem is a data filtering problem.

Think of the Data Intensive Computing problem as a filtering or funneling process. It starts with vast amounts of data and ends with what amounts to a one-bit decision: yes or no, buy or sell, live or die. This requires taking vast amounts of structured or unstructured data through a series of processing steps – a filtering process that transforms raw information and enables insight that leads to intelligence, knowledge, experience, and ultimately better decision making. Doing this efficiently requires an environment that:

- Centralizes data processing and analysis
- Consolidates application workflows around data
- Maximizes the efficiency of data movement
- Visually renders data in one place
- Fuses data from disparate sources into a single view
- Delivers visual representations of data to users, rather than the data itself

Ideally we want this filtering process to occur in sub-seconds, or real time. That requires a highly efficient, reliable and scalable platform designed for data intensive applications.

A High-Efficiency Platform for Data Intensive Applications

With each new generation of HPC systems, regardless of architectural approach, the gap between theoretical peak performance and actual application performance widens dramatically.

Generally speaking, compute intensive applications exhibit a range of computational resource demands across a wide range of market segments. Applications in Computational Biology and Explicit Computational Structural Mechanics reuse data read from memory to a large degree and are considered cache-friendly. Contrast these to Computational Fluid Dynamics , Implicit Computational Structural Mechanics (in particular dynamic problems like Noise, Vibration, Harshness), Seismic Processing and Reservoir Simulation – all of which consume the data read from memory and have to load continuously new data from memory. To keep the floating point units (FPU) busy, these applications require computer architectures with high memory bandwidth, mainly due to the data addressing patterns and heavy I/O activities.

Computer architecture requirements may differ if the application is scalable. While CFD applications can scale to a higher processor count and thus reduces the memory bandwidth requirement per processor, Computational Structural Mechanics Implicit (dynamic problems) exposes reduced scalability, so the memory bandwidth per processor determines the overall performance. Seismic and Climate/Weather/Ocean applications require high floating point (FP) processing speed, while Bioinformatics is dominated by integer intensive algorithms.

Innovations in Memory, I/O and Scalability

To realize a Data Intensive Computing model, it's necessary to move processing to the data – rather than data to the processing. This approach is particularly beneficial to applications that suffer the most from the data communications bottleneck.

Following is a detailed look at how the Intel® Xeon® Processor 5500 Series (code-named Nehalem) and the SGI Altix ICE platform achieve unprecedented application performance through innovations in:

- ***Memory infrastructure***, providing the bandwidth necessary to enable unfettered communication between discreet devices
- ***I/O***, which maximizes the efficiency of data movement
- ***Scalability***, which enables organizations to take maximum advantage of large-scale systems to handle data-intensive problems

High-Bandwidth Memory Design of Intel® Xeon® Processor 5500 Series (Code-named Nehalem)

HPC application performance suffers because CPU performance is increasing eight times faster than memory speed. Compared with processor performance, which improves at 60% per year, DRAM speed is improving much slower, at the rate of 7% per year.

This disparity results in a relative increase in DRAM latency when expressed in terms of instructions processed while waiting for a DRAM access, or in terms of DRAM words accessed while waiting for a DRAM access. This "memory wall" means that Load and Store is slow, but Add/Multiply is fast. This slow scaling results in an increase in memory latency when measured in floating-point operations.

The next-generation Intel® Xeon® Processor 5500 Series (code-named Nehalem) put enormous emphasis on adding features to improve the byte/flop ratio – also known as system balance. Among the key features enabling this balance are the Intel® QuickPath Interconnect (Intel® QPI) and an integrated memory controller, which together result in unprecedented amounts of aggregate bandwidth. The innovative memory system and processor interconnect, therefore, directly addresses the memory gap problem.

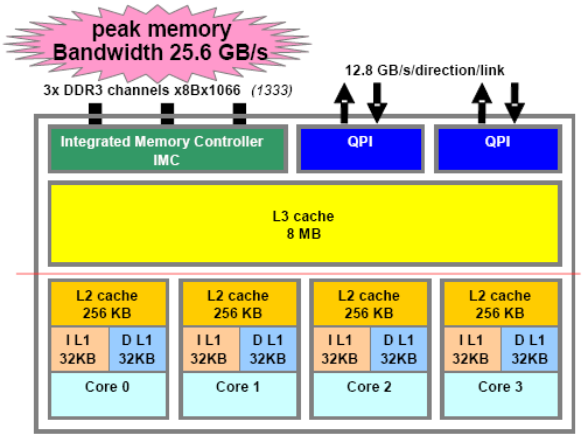


Figure 3 The innovative architecture of the Intel® Xeon® Processor 5500 Series (code-named Nehalem) addresses the memory gap problem.

With this new Intel® quad-core processor, the bandwidth-related problems specific to the current generation of multi-core processors are extensively resolved. The Intel® Xeon® Processor 5500 Series provides for a better balance between computational units and the memory path characteristics because:

- The socket of the Intel® Xeon® Processor 5500 Series has a memory controller device directly on the die that allows doubling of the available bandwidth. Depending on memory type, memory configuration and the processor chosen, the hardware bandwidth is in the range of 20-25GB/sec.
- Each core has its own L1 and L2 cache that in turn helps to decrease the number of stalls in a data path.
- The data pre-fetch algorithm for L2 and L3 caches has been substantially reworked to achieve more effective data load.

The improvements designed into this new processor result in substantial boosts in application performance and their scalability by supporting the data requirements of multiple processing threads running simultaneously.

The Impact of Memory Bandwidth on Application Performance

For most HPC applications, the memory system drives overall performance more than any other component. So how does the Intel® Xeon® Processor 5500 Series stack up in terms of HPC application performance?

Two general-purpose benchmarks, Linpack and STREAM, provide orthogonal ways to specify performance of a node. Together these measurements make it possible to understand the application performance potential of a node.

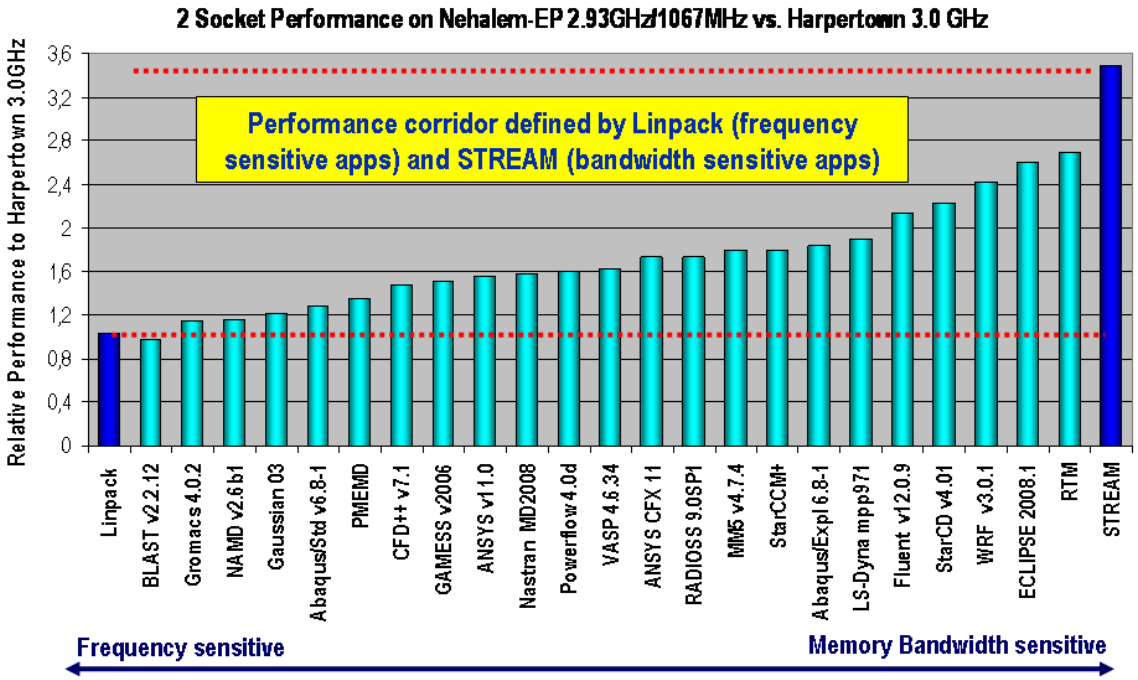


Figure 4 Along with the STREAM benchmark, applications sensitive to memory bandwidth excel on Intel® Xeon® Processors 5500 Series (code-named Nehalem). Applications that are clock frequency sensitive and cache friendly tend to see less of a performance improvement. All runs on Altix ICE 8200EX with Intel® Xeon® Processor 5500 Series at 2.93 GHz.

From a general benchmark perspective, when comparing performance on Intel® Xeon® Processor 5500 Series against the previous Intel® Xeon® Processor 5400 Series (Code-named Harpertown), the lower limit of HPC application performance is measured by Linpack and the higher limit by STREAM.

- If the application is clock frequency sensitive and cache friendly, then it behaves like Linpack and the performance boost on the Intel® Xeon® Processor 5500 Series is minimal.
- Memory bandwidth sensitive applications, like the STREAM benchmark, benefit most from the Intel® Xeon® Processor 5500 Series architecture. These include RTM (Reverse Time Migration), Eclipse, StarCD, Fluent, and WRF.

I/O Innovations on SGI Altix ICE

Commodity Linux® clusters use the free memory on the board to cache all I/O files within an I/O buffer maintained by the OS. But HPC data sizes are fast becoming too large for the node memory available in commodity clusters.

As performance decreases due to the time spent in I/O operation, out-of-core memory solvers come into play. One approach is flexible file I/O (FFIO), a patent pending technology from Silicon Graphics, which features an I/O accelerator library. In essence, FFIO is an I/O buffer cache that allows the user to specify the I/O page size, the number of pages in the I/O buffer, the number of I/O read-ahead and write-behind operations, the I/O stride, and more. FFIO helps to avoid memory thrashing on a system that is running I/O- and compute-intensive applications. The FFIO library has a “linkless” design, which means an application doesn’t have to be linked against the FFIO library to use it. Only an environment variable has to be set up.

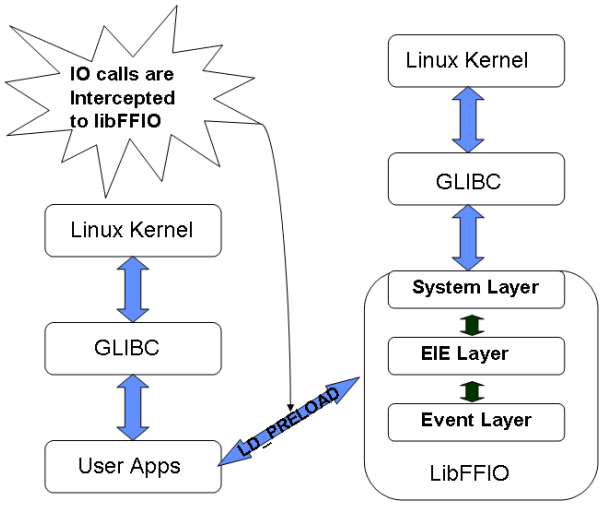


Figure 5 FFIO usage is application transparent (linkless)

The value of FFIO in real simulation is illustrated in Figure 6. The figure shows a comparison of a 10 million degrees of freedom power train model ran with Abaqus/Standard 6.8-1 on an eight-core SGI Altix ICE 8200EX with Intel® Xeon® Processor 5500 Series 2.93GHz, 48GB memory and a good local file system with and without FFIO. When FFIO is used, the application performance improves by 19.8%. This feature is available on Silicon Graphics® clusters with SGI® ProPack™ for Linux® software.

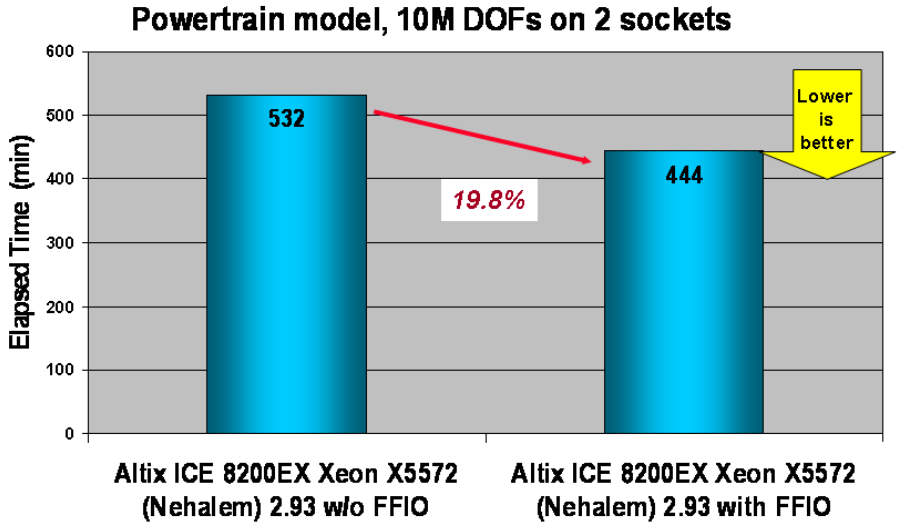


Figure 6 On a power train model with 10 million degrees of freedom in Abaqus/Standard 6.8-1, running FFIO improves performance by 19.8%.

Scalability of HPC Applications

With Intel® Xeon® Processor 5500 Series as its multi-core engine, the SGI Altix ICE platform is capable of achieving industry-leading scalability without sacrificing application performance efficiency. The platform offers a variety of interconnect options that enable users to efficiently scale their applications across hundreds or thousands of processor cores. These include support for multi-rail InfiniBand networks, and a choice of Fat Tree or Hypercube network topologies.

Multi-Rail Networks

Fast interconnect speeds are a must for Data Intensive Computing. One way to optimize communication traffic is through multi-rail networks. These networks can improve MPI (Message Passing Interface) communication performance by dividing large messages into chunks and distributing those chunks across multiple independent InfiniBand (IB) networks, or rails. With the dual-plane network topology of SGI Altix ICE clusters, MPI communications can make use of both IB rails.

A multi-rail network may not automatically benefit an application performance. It does depend on several details and importantly, on the relations between HCA and PCIe bus capabilities. For example, each compute blade of an SGI Altix ICE 8200 EX cluster has a single dual-port ConnectX HCA on a PCIe x8 bus. So using dual-rail to stripe a single large message across both ports can almost double the MPI communication bandwidth, depending on the application communication patterns and limited by the PICe x8 bus bandwidth.

The other benefit of the dual-plane InfiniBand network is that the system can separate the MPI communication from the I/O traffic, dedicating separate network for MPI respectively I/O.

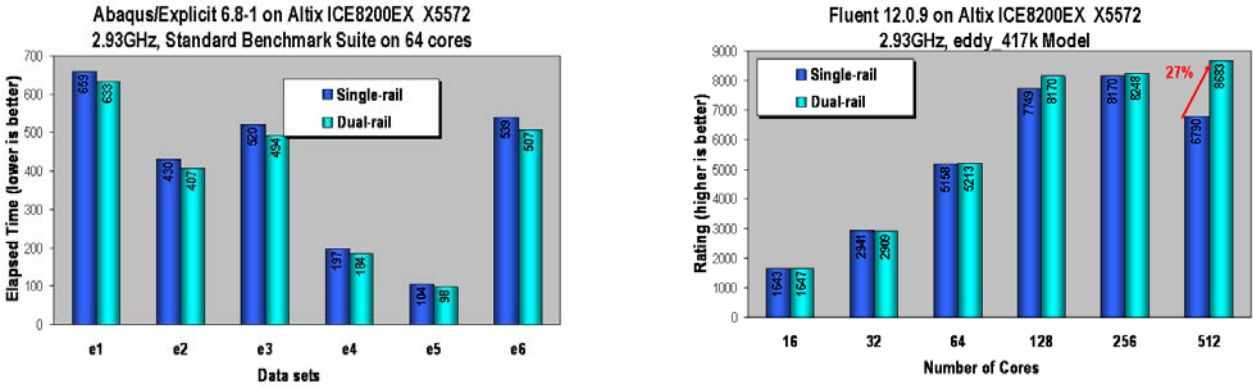


Figure 7 Scalable, communication bandwidth-sensitive applications see increasing benefits from multi-rail IB networks.

Figure 7 compares how an HPC application performs on SGI Altix ICE 8200EX with Intel® Xeon® Processor 5500 Series using dual-rail vs. single-rail IB networks. Even on only 64 cores the benefit is 6%. On scalable applications that are communication bandwidth sensitive, the benefit is much higher. For example, a 512-core Fluent job runs on a dual-rail network 27% faster than on single-rail.

Fat Tree Network vs. Hypercube Topology

SGI Altix ICE systems also allow users to choose between Hypercube (best for larger node count MPI jobs) or non-blocking Fat Tree network topology (suited for smaller node count MPI jobs). Figure 8 shows how each topology affects the performance of key HPC applications.

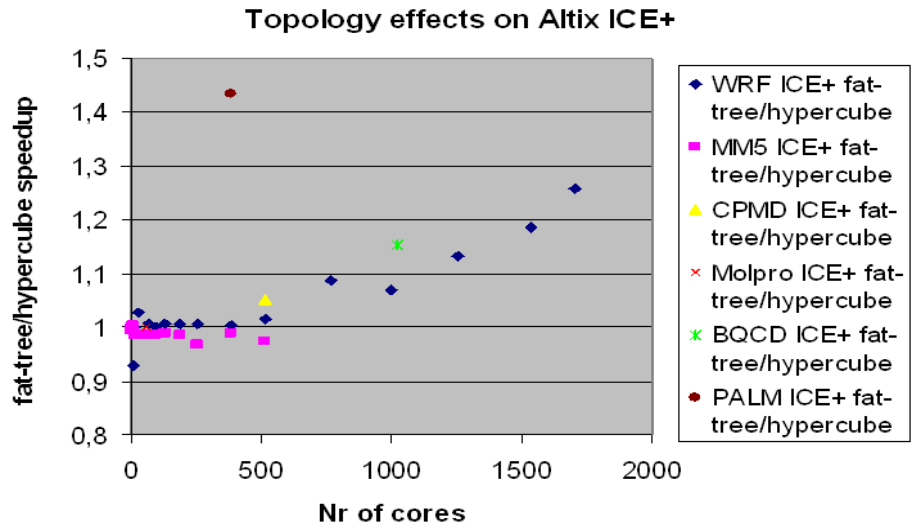


Figure 8 By examining the effect of topology on HPC applications, users can determine which option – Fat Tree network or Hypercube – best serves their needs. The chart shows that WRF, BQCD, PALM are interconnect bandwidth sensitive, and that MM5, CPMD and Molpro are interconnect latency sensitive. Both topology options are supported by SGI Altix ICE systems.

Real-World Scalability Results

Unprecedented scalability can be observed in the performance of a variety of HPC applications, which in recent years have proliferated with the adoption of low-cost processors designed to work in parallel on a computational task.

One such application class is CFD, where high-fidelity models and simulations are increasingly critical in reducing the need for expensive physical testing. At the same time, ever-increasing numbers of simulations enable engineers to consider multiple design ideas and even perform automated multiple discipline optimization.

CFD software takes advantage of multiprocessor and multi-core systems by employing domain decomposition, which divides the simulation model into sub-domains. Each sub-domain is then computed on a separate processor (core), while all processors work in parallel to speed up the computation.

Fluent. Superior large-model parallel scalability can be demonstrated on SGI Altix ICE 8200EX with Intel® Xeon® Processor 5500 Series. In tests of the popular ANSYS Fluent CFD, a large truck model of 111 million cells of mixed type is used to understand the benefits of multi-rail network and appropriate tuning of the hardware features. The problem involves external flow over the truck body and uses a discrete-event simulation (DES) model with a segregated implicit solver.

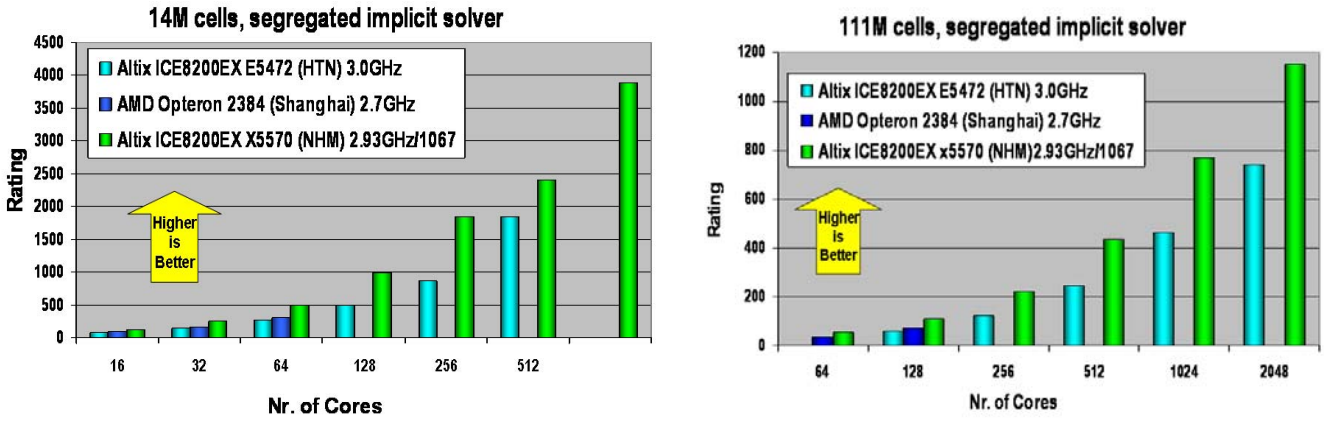


Figure 9. Large and very large Fluent models achieve unprecedented scalability across 2,048 cores on the SGI Altix ICE 8200EX platform with Intel® Xeon® Processor 5500 Series (code-named Nehalem) at 2.93GHz. Scalability actually improves as the model size increases, due to the redesigned Intel® Xeon® processor's memory bandwidth. Source: http://www.fluent.com/software/fluent/fl6bench/fl6bench_6.4.x/problems/truck_111m.htm and SGI internal measurements

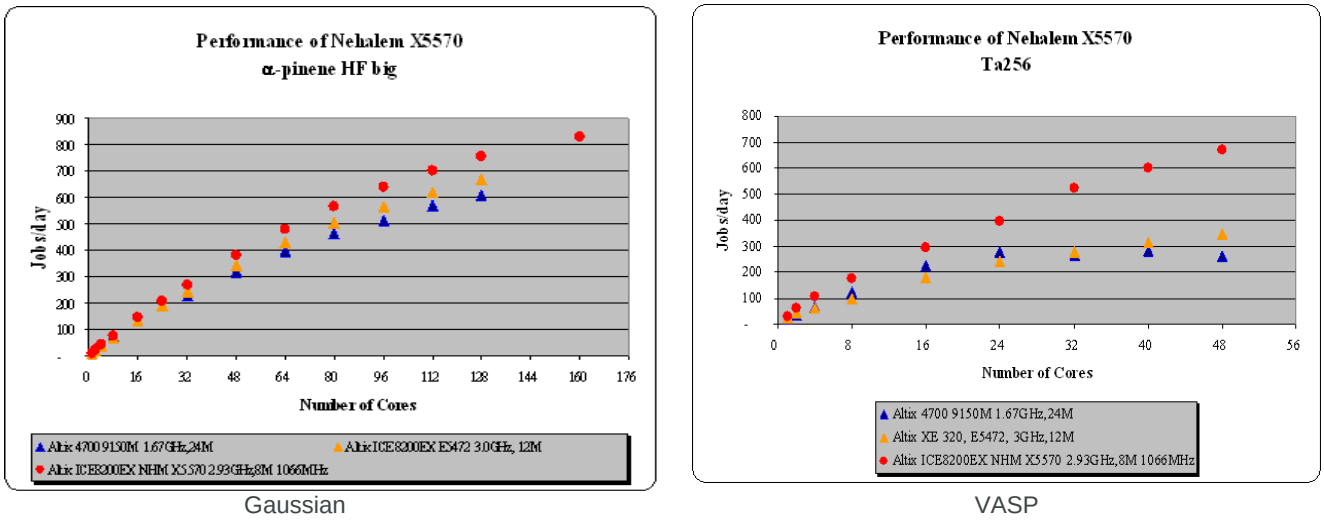
As illustrated in Figure 9, Fluent achieves unprecedented scalability on 2,048 cores of the SGI Altix ICE 8200EX platform running Intel® Xeon® Processor 5500 Series at 2.93GHz. At left, a large model of the external flow of a truck body amounts to a computational mesh of 14 million cells. A much larger version of the model, at 111 million cells, appears at right. The results are telling:

- On the large 14 million cell model, Altix ICE 8200EX with Intel® Xeon® Processor 5500 Series at 2.93GHz/1067 is ***1.59x faster*** than AMD Shanghai 2.7GHz and ***1.73x faster*** than Altix ICE 8200EX with Intel® Xeon® X5472 (Harpertown) 3.0GHz on 64 cores. Scalability is near linear on this model.
- On the very large 111 million cell model, the performance improvement is even higher, due to the Nehalem processor's memory bandwidth. Altix ICE 8200EX with Intel® Xeon® Processor 5500 Series at 2.93GHz/1067 is ***1.64x faster*** than AMD Shanghai on 64 cores and ***1.81x faster*** than Altix ICE 8200EX with Intel® Xeon® X5472 (Harpertown) 3.0GHz on 128 cores. While not linear, scalability on this larger model is still impressive.

Other fields that benefit from the scalability offered by this platform are computational chemistry and molecular modeling. Data-intensive applications in this field include:

- **Gaussian**, which helps predict the energies, molecular structures, and vibrational frequencies, and properties of molecular systems.
- **VASP** (Vienna Ab-initio Simulation Package) simulates the properties of systems at the atomic scale.
- **GAMESS** (General Atomic and Molecular Electronic Structure System), a general ab initio quantum chemistry package.
- **PMEMD** (Particle Mesh Ewald Molecular Dynamics), a new version of Sander, a core component of the AMBER of molecular dynamics packages.

Figure 10 shows the superior performance of SGI Altix ICE 8200EX with Intel® Xeon® Processor 5500 Series at 2.93GHz in running Gaussian (at left) and VASP (at right). The Ta256 mode used in VASP is based on Fast-Fourier Transformations (FFT), which mainly stress local memory bandwidth



Gaussian VASP

Figure 10 In terms of absolute performance and scalability, SGI Altix ICE 8200EX with Intel® Xeon® Processor 5500 Series (code-named Nehalem) at 2.93GHz/1067 outperforms the same system equipped with Intel® Xeon® X5470 (Harpertown) processors.

GAMESS. In Figure 11, the SGI Altix ICE 8200EX with Intel® Xeon® Processor 5500 Series at 2.93GHz delivers superior performance on GAMESS. On Linux® clusters in particular, GAMESS starts two processes per processor core, which communicate with each other through a socket. One process conducts the actual computation, while the other (the data server process) provides a mechanism for accessing memory located on remote processors and or nodes. This can typically lead to data congestion for data intensive workloads, with the compute thread and the data server communication competing for available CPU resources and reducing computational efficiency. The Intel® Xeon® Processor 5500 Series supports hyper-threading (HT) to enable two threads per core. Mapping computational and data server threads on the same core improves performance considerably on ICE 8200EX with Intel® Xeon® Processor 5500 Series when compared to a Harpertown cluster and even with a shared memory system like SGI Altix 4700. In addition, GAMESS uses SHMEM for communication, which is optimized on SGI Altix ICE 8200EX.

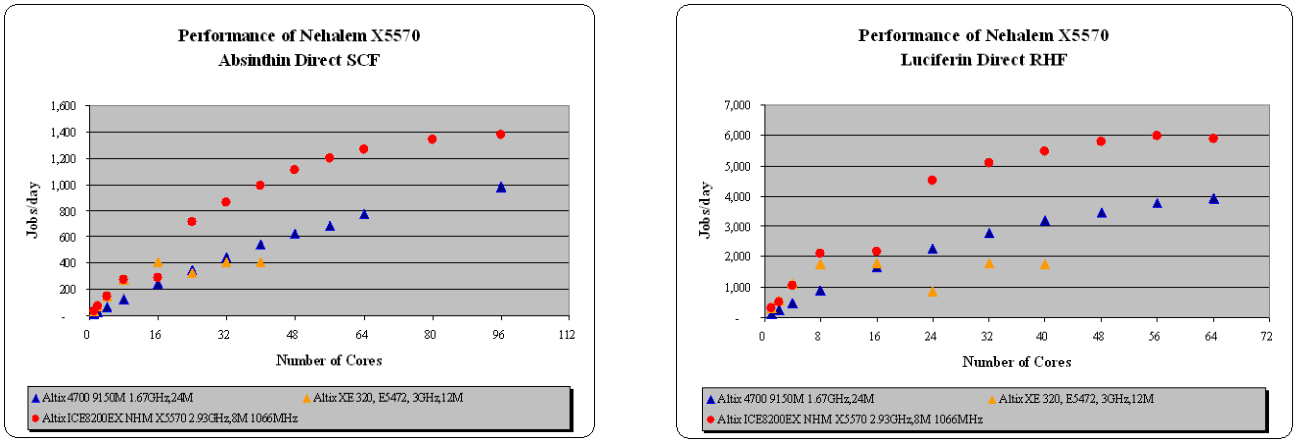


Figure 11 Running on SGI ICE 8200EX with Intel® Xeon® Processor 5500 Series, GAMESS two typical GAMESS workloads – self consistent field (SCF) and Restricted Hartree Fock (RHF) calculations – scale reliable across up to 64 and 96 cores, respectively.

Conclusion

In the face of exploding data volumes, the need to maximize the efficiency of data movement – and where possible to move processing to data, rather than data to processing – is an essential requirement of Data Intensive Computing.

When combined with the memory bandwidth intensive design of the Intel® Xeon® Processor 5500 Series (code-named Nehalem), SGI Altix ICE systems enable today's users to achieve unprecedented application performance and scalability.

This standards-based, industry-leading platform is a core component of a sustainable IT infrastructure. In addition to innovations in I/O architecture and scalability, the SGI Altix ICE integrated blade platform offers other unique benefits important to any HPC organization or enterprise:

- **Integration**. SGI Altix ICE systems are designed for HPC. Altix ICE systems feature performance-dense architecture that combines an innovative board design, cable-free board enclosures, integrated switches, and a high-performance, InfiniBand-focused interconnect architecture that can efficiently scale to thousands of nodes.
- **Reliability**. Altix ICE systems feature an advanced reliability architecture with redundant power and fans, diskless blades and hot-swappable blades. The platform's proven power/cooling architecture delivers rack power efficiency of greater than 75%, and the optional water-cooled door design lowers cooling costs and extends the life of dense configurations.
- **Immediate Productivity.** The SGI Altix ICE system's "Power Up and Go" design, with systems that are fully integrated and tested in the factory, enables customers to deploy scalable cluster solutions in hours, rather than weeks. The standards-based Linux system also features a comprehensive cluster solution stack that includes SGI® Tempo management software for quick deployment and easy management.

Combined with high-performance storage, innovative visualization software and a unified standards-based operating environment, this platform helps organizations transform big data from a computing problem to a competitive advantage.

For more information on SGI Altix ICE systems and Intel® Xeon® Processor 5500 Series (code-named Nehalem) processors, visit www.sgi.com/products/servers/altix/ice/

Performance test configurations used in this paper:

SGI Altix ICE 8200EX with Intel® Xeon ®Processor 5500 series (code-named Nehalem) with core frequency of 2.93 GHz: 2 sockets per node , 4 cores per socket, 11.72 GFLOPs peak/core; 8MB cache per socket ; 48GB 1333MHz DDR3 memory per node; STREAM-Triad at 36557MB/sec on 8 cores; 256 nodes; dual InfiniBand DDR interconnect planes with ConnectX HCA at 2x2GB/sec; Bristle (16) Hypercube topology; 32GB/sec bisection bandwidth of 32GB/sec.

SGI Altix ICE 8200EX with Intel® Xeon® processor 5400 series (code-named Harpertown) with core frequency of 3.0GHz: 2 sockets per node, 4 cores per socket, 12.0GFLOPs peak/core; 2x6MB cache per socket with 2x2 cores; 32GB 1600MHz DDR2 memory per node; STREAM-Triad at 9750MB/sec on 8 cores; 256 nodes; dual InfiniBand DDR interconnect planes with ConnectX HCA at 2x2GB/sec; Bristle (16) Hypercube topology; 32GB/sec bisection bandwidth of 32GB/sec.

Corporate Office
1140 E. Arques Avenue
Sunnyvale, CA 94085
(408) 524-1980
www.sgi.com

North America +1 800 800 7441
Latin America +55 11 5185 2860
Europe +44 118 912 7500
Japan +81 3 5488 1811
Asia Pacific +61 29448 1463

SiliconGraphics