High-end ‘SNIA’: The Partitioned System as Guidepost

Steve Reinhardt, SGI, 655F Lone Oak Drive, Eagan, MN
55123 USA
1-651-683-5664 Fax: 1-651-683-7145 spr@sgi.com

ABSTRACT: *At SGI, we are building a family of highly scalable servers and supercomputers which are differentiated by their capability performance and scalability (due to using next-generation NUMAlink™ interconnect technology), production strength OS based on Linux and the Open Source model, and high RAS (due to a highly modular construction). The notion of a ‘partition’ is fundamental to the ability to deliver performance and RAS simultaneously. A partition is a portion of the system which can be physically and logically isolated, and in which failures (hardware or software) are contained. Partitions cooperate to enable application-level use of low-latency hardware communication mechanisms, which in turn enable the capability performance required by the high-end market.*

This family of systems, code-named ‘SNIA’, is based on IA-64 processors from Intel, which provide very high performance as well as exceptional cost-effectiveness. SNIA systems are built on the strong scalable infrastructure first delivered with the sister family of MIPS-based systems (code-named ‘SN-MIPS). Acceptance of SNIA in the high-end marketplace will be gated by the maturity of the Linux operating system, and especially its ability to grow into a mature production OS for high-end systems. We propose SNIA as a strong system to supersede current highly scalable supercomputers such as the Cray T3E™.

Introduction

Partitions

Builders of production quality kiloprocessor systems have struggled to balance four key constraints on their design. First is the requirement to deliver low latency (1-3 microseconds) and high bandwidth communication across a kiloprocessor system. Second is the requirement to provide the system stability and resilience required by a large community of users. Third is the use of commodity building blocks to provide good price-performance as well as good performance. Fourth is providing the operating system, programming environment and resource management that enables the kiloprocessor to be run effectively as a single coherent system. The Cray T3E™ system balanced these constraints with heavy emphasis on communication, resilience, and system software, but with commodity use being limited to microprocessors and memory. The IBM SP-2™ system balances these constraints with emphasis on resilience and commodity use, but with communication latency, which makes scaling programs to kiloprocessor scales difficult and a less productive software environment.

The ‘SNIA’ family of systems from SGI bridges these constraints via extensions of key ideas from both the T3E and the SGI Origin2000 system. Principal among these ideas is the partition, which on one hand is the unit of failure containment and OS kernel span, and on the other hand is the ‘microkernel’ supporting access to hardware communication mechanisms which deliver the latent performance of the hardware interconnection network. Building on this union of performance and RAS, higher layers of the HPC Linux operating system on SNIA connect the partitions together into a single system with strong ease of use for both end-users and administrators.

Background

Gaining a broader audience for scalable systems expertise

SGI has delivered major innovations in scalable systems, notably the cache-coherent NUMA ('ccNUMA') infrastructure and scalable OS of the Origin system. Coupled with the expertise of those SGI employees who contributed to the extremely scalable Cray T3E system, SGI has a reservoir of technology and knowledge about scalable systems unequaled in the industry. The insatiable demand by the industrial and research communities for ever-greater amounts of computing is pressuring cost and price-performance to a degree never seen before. Two major opportunities in fundamental computing technology, the IA-64 or EPIC architecture from Intel, and the Open Source movement typified by the Linux operating system, offer the promise of greatly improved performance and price-performance, a promise which must be fulfilled to meet the needs of these high end customers. SGI is applying its proven expertise in scaling and production systems with these two technologies, which will allow it to reach a wider audience of customers with those core competencies.

Strategy

Microprocessors

Microprocessors as a technology have been converging over the last several years, as the performance differences between the best and worst processors (optimized for some key characteristic) continue to narrow. This, taken together with the increasing use of parallelization within a large part of the scientific and engineering application base, means that the single processor capability while still important is becoming less of a unique performance differentiator. As this technology begins to look like a commodity, it is essential that SGI apply its intellectual resources to areas where differentiation can be realized for its technical and creative user base. As a result, we are starting to use the IA-64 architecture, which we believe provides excellent absolute performance and exceptional price-performance, along with all the technical features to build the world's most powerful computers, which will be much more affordable.

Operating Systems

Because we are adopting a new instruction set architecture (ISA), it is an opportunity to consider a new operating system at the same time. While Irix provides many strong abilities, its application catalog outside the technical computing markets is limited. At the same time, the Linux operating system has exploded in use over the last few years, and is attracting new applications in many markets in which users want a Unix-like OS. Many of these new applications are in markets which fit well with SGI's expertise in systems architecture, visualization, digital media conveyance and manipulation, and scalable systems interconnect. Additionally, the benefits of the Open Source development model are becoming apparent to many of us for whom such an approach would have been preposterous even a few years ago. We see an opportunity to harness the collective insight of the high-end community while allowing the easy tailoring of the system software to individual site needs.

Strategy for Differentiation: NUMAlink scalability in dual platforms

The commodity ISA and OS allow SGI to focus more tightly on the areas where we will expend resources and differentiate our products. For large-system customers, our primary differentiation is the robust scalability of our systems, in all dimensions of processor count, memory size, and internal and I/O bandwidths. Our NUMAlink™ interconnection technology enables this robust scalability, and will be the fabric that supports both our 'SNIA' and next generation MIPS/Irix scalable system ('SN-MIPS') systems. NUMAlink2 was delivered in the Origin 2000, providing ccNUMA memory access with latency just over a microsecond, along with delivered bandwidth of over 300MB/s per processor, in 128P configurations with all processors communicating equally. NUMAlink3 will be delivered with the initial SN-MIPS and SNIA deliveries, and will provide reduced latency and a doubling of bandwidth per link within the network, coupled with improvements in network RAS. NUMAlink4 will be delivered with SN-Madison systems and will again reduce latency and double the bandwidth per link, along with advances in routing algorithms.

Because of these two major opportunities, we are able to deploy the strongly scalable infrastructure developed for SN-MIPS with a different ISA and OS, and hence reach some existing sets of customers and some significantly different audiences as well.

MIPS/Irix and IA-64/Linux will have different characteristics demanded by different market segments, we anticipate delivering the infrastructure in both flavors for some considerable length of time.

Technology and Implementation

Hardware – Rolling Upgrades

The SNIA hardware is based on the strongly scalable NUMA/ccNUMA infrastructure that first will be delivered with the 'SN-MIPS' system. This infrastructure builds on the innovations of the Origin2000 system, improving the performance with next-generation bandwidths and latencies, extending the reach to larger processor counts, and adding in some ideas pioneered by the T3D and T3E systems.

A system is built up by integrating various combinations of hardware modules or 'bricks':

- C-brick for computation
- R-brick for system interconnect ('router')
- I-/P-/D-bricks for I/O

Customers will create a balance their system to their applications by choosing appropriate amounts of compute, interconnect, and I/O. A key benefit of this infrastructure permits the key component technologies (processors and speed bumps, interconnect, and I/O) to change at their varied paces.

The first generation of SNIA ('SN-Itanium') will employ the SN-MIPS infrastructure with an adapter chip that permits the Itanium™ processor to be used. The Itanium processor (see http://developer.intel.com/design/ia-64/ for details) supports the IA-64 instruction set, as well as supporting the IA-32 instruction set in hardware. It is manufactured on a .18 micron process and will be delivered at 800MHz. This system will be delivered in the first half of 2001.

Within a year later, the 'SN-McKinley' system will deliver a next-generation 'hub' chip connecting microprocessors, memory, routers, and I/O. This infrastructure will support the 'McKinley' IA-64 processor from Intel, which is expected to deliver twice the sustained performance of Itanium primarily due to triple the front side bus bandwidth. An upgraded router will provide twice the bandwidth of the SN-Itanium router.

Within a further year, the 'SN-Madison' system will deliver the 'Madison' IA-64 processor, along with a next-generation router providing another doubling of bandwidth and reduced latency. The Madison processor will be further optimized for floating-point performance, and will be fabricated on a .13 micron process.

During this period, I/O infrastructure will be upgraded as well. First is the advent of PCI-X connectivity. Peer I/O will be the next major benefit, as I/O connections into the interconnection network are upgraded to full 'peer' status, allowing independent scaling of compute and I/O. As Infiniband™ matures toward the end of this era, it will be incorporated into systems as practical.

The discussion above does not adequately express the modularity of upgrades that are possible within this family of systems. Simply put, the processor modules lead the advance of technology, and the remainder of the infrastructure can be upgraded or not at a site's discretion. Thus an SN-Itanium system can be upgraded with SN-McKinley or SN-Madison compute modules, with no changes to the interconnect or I/O fabric. Or, customers wanting the most robust interconnect might upgrade both compute and router modules, with the I/O fabric remaining unchanged. All of this can be achieved within an existing (set of) chassis, with cables being moved to reflect new connectivity.

Software – Linux Strategy

Our embrace of the Open Source software model (see http://oss.sgi.com for more details) is a major change in approach from previous OS work. The Open Source model enables many Internet and Web-related development groups to move very rapidly on a stable OS base, creating and absorbing important new features. The Extreme Linux group (http://www.extremelinux.org/) has used this model for high performance computing, mainly among research and university users. SGI believes this model can provide big benefits to the larger HPC user base as well.

SGI will be a genuine participant in the Open Source community; that is, we will work within the community to get our ideas and features accepted, with the default being that kernel features not accepted by the community will not be delivered in SGI systems. We believe that fragmentation of the Linux kernel would be a catastrophe, and will do everything within our power to avoid that happening. We are placing SGI-developed Linux code into three categories: a) code which will be offered immediately under GPL; b) code which will be retained for a specified period as SGI proprietary, and code which will be retained for an undetermined period as SGI proprietary. Essentially all kernel code will fall in category a). Code that would fall into category b) and c) will be non-kernel code, with possible examples being file/tape management and resource management. We recognize the risks of holding the technology too tightly proprietary.

'HPC Linux'

Within the Open Source community, SGI will lead the motivation, design and development of features necessary for high-end production systems. Because the underlying kernel must support desktop and small server systems as well as large systems, the kernel gatekeepers will damp the set of changes that might ideally be made to support large systems. Hence the development of 'HPC Linux' will be focused outside the kernel proper, on layers of software which provide access to hardware features permitting extreme capability performance, single filename space, workload management (including job scheduling), high reliability and resilience, and global system administration.

The Partitioned System as Architectural Guidepost

Experience with kiloprocessor Cray T3D™, Cray T3E, and SGI Origin2000 systems has taught us some valuable lessons about the trade-offs necessary to meet the requirements of these customers.

- A single job must be permitted to occupy essentially the whole system (processors and memories), while providing application-level latency (measured by MPI one-sided or shmem) of 1-3 microseconds. Appropriate remote memory bandwidth, in the range of 15-30% of local memory bandwidth, must also be provided. The T3E provided this level of communications performance, and users were able to scale applications on the T3E that would not scale elsewhere.
- Having a unit of fault containment less than the entire system is required to achieve system reliability levels needed by large system customers. In the T3E, this unit was one PE, with its accompanying microkernel, and this permitted strong resilience.
- An efficient execution environment is needed for both achieving capability performance, and for achieving repeatability of performance. This requires efficient memory management (minimizing TLB faults, for example) and non-interference with memory, processors, and interconnection network being used by the application. The T3E provided this by the architecture of its routers, and by the gang scheduling of the processors attached to a job.
- The system must permit a portion of the system to be repaired, including physical replacement, while the remainder of the system continues to function. T3E was not ultimately able to provide this feature, and for large, high-availability sites (typified by operational weather centers), it proved a problem. The T3E's redundant PEs provided a related capability that was highly valued. The Origin 2000, in configurations like ASCI Blue Mountain, was able to provide this, and it proved highly valuable.
- Users require a single filename space for ease of use. On the T3E, this was provided by the work done to make the UNICOS/mk filesystem highly parallel.
- System administrators require a single means of administering the entire system. The T3E provided this, though the tools were not as simple to use as many would have liked.
- Large-system customers require strong price-performance from their big systems, as the price-performance directly effects how much equipment they can buy, and hence their aggregate computing capability. Thus the system must be built from near-commodity building blocks. The T3E provided good price-performance, but for some customers it proved expensive, partly due to a fixed rate of communication infrastructure per processor, which was sometimes overkill in small systems. The Origin 2000's separation of computing and communication infrastructure avoided this difficulty.

A close reader of this list will see that it combines the best attributes of 'single system image' systems (a la T3E) with the best attributes of 'cluster' systems (a la SP2, Beowulf). A partitioned system responds exactly to this apparent conflict. At the lowest levels of hardware and OS, a partition provides fault containment and the necessary OS primitives to serve as a building block for the system. At the higher levels of the system, access to robust communications and the necessary global services provide the performance and ease of use that are most visible. Both T3E and Origin2000 had some attributes of partitioned systems. The attributes they possessed proved to be strengths, and the attributes they lacked often proved to be weaknesses.

SNIA as a Partitioned System

A partition of an SNIA system will consist of a set of physical C-bricks, typically 16-32 processors, with a Linux kernel running in it. Faults, both hardware and software, will typically be contained within the partition. A partition communicates with other partitions via NUMAlink router s, using MPI, shmem, or TCP/IP, depending on the application. Within a partition, memory is ccNUMA: i.e., globally addressable, cache-coherent, and physically distributed. Between partitions, the memory model is typically NUMA; that is, globally addressable and physically distributed, but not cache-coherent.

A large SNIA system will comprise multiple partitions.

- Capability performance: Parallel jobs will often span multiple partitions, communicating in a NUMA style (MPI or shmem) between partitions at hardware speeds. NUMA communication between partitions will provide good user control over data placement while preserving the strong communications performance. Gang scheduling will contribute to the efficient execution environment.
- The failure of a partition typically will cause the abortion of any jobs running on that partition (with checkpoint/restart being an eventual remedy) as well as the notification of the workload manager and error manager. Almost all errors will be contained within the failing partition. Partitions can be downed electrically for physical replacement without effecting surrounding partitions.
- A single file-name space will be provided by CXFS, a version of the local XFS filesystem that provides a coherent view across multiple partitions. CXFS is already being delivered on Irix on the Origin2000, and will be ported to SNIA/Linux as it matures.
- System administrators will administer a single system via global tools. The administrator will accomplish upgrading a system of N partitions globally, rather than requiring N distinct upgrades. Some of these tools are already available from third-party vendors.

Performance and Delivery

The SNIA will be delivered incrementally in terms of both hardware and software. The first shipments of SN-Itanium in the first half of ’01 will support partitioned systems up to 128P, growing rapidly to 2048P of SN-McKinley in ’02. Alternately, viewing this as peak FLOPS available within a 3 microsecond latency domain (i.e., computing which is closer together in time is more valuable), SN-Itanium will support up to 1.5TF (512P at 3.2GF peak). This increases to about 8TF for SN-McKinley, and well beyond that with SN-Madison. Sustained performance improvements will be even more dramatic due to higher local and remote memory bandwidths.

For large-T3E customers who schedule their systems in relatively simple ways, the SNIA should be a credible T3E successor in the second half of 2001. For customers who schedule their systems in more complex ways, the first half of 2002 is a good target to consider an SNIA system.

Summary

The SNIA system responds to the demanding requirements of the high-end technical computing market by exploiting the knowledge gained from implementation of the T3E and Origin2000, commodity components from Intel and other vendors, and the SN-MIPS system infrastructure. A key mechanism for satisfying these often-conflicting requirements is the concept of a partitioned system. Partitions, each of which provides a near-commodity building block (hardware, OS, and cost), are combined with SGI’s differentiated NUMAlink interconnection network and the HPC Linux software to create a compelling system. This kiloprocessor system architecture provides capability performance, high resilience, excellent job scheduling, strong ease of use, and exceptional price-performance.

Literature Citations

Galimberti, Gustavo, “Parttioning in the SN-MIPS System”, in the Proceedings of the Summit2000 Meeting, Noordwijk, Netherlands.

Higgins, Dan, “Resource scheduling in the Irix System”

IBM SP-2, Special issue on SP-2 architecture. IBM Systems Journal, 34(2):144-322, 1995.

Johnson, Lynne, “HPC Linux Project Update”, in the Proceedings of the Summit2000 Meeting, Noordwijk, Netherlands.

Laudon, James, and Lanoski, Daniel: “The SGI Origin: A ccNUMA Highly Scalable Server.” ISCA 1997: 241-251

Scott, Steve, “Synchronization and Communication in the T3E Multiprocessor”, Proceedings of the VIIth Conference on Architectural Support for Programming Languages and Operating Systems”, Cambridge, MA, October 1996.