

SGI's Linux

Record breaker!

SGI's latest server launch turned heads at the recent LinuxWorld Expo in New York. It wasn't the groovy rack case or the brightly coloured trim that attracted visitors, but the promise of the best performing Linux system that money can buy that piqued interest.

The Altix 3000 family of servers and clusters is the culmination of several years hard work by SGI and the Linux community in general.

Designed for high performance tasks, the Itanium 2 based servers have a number of unique features which have enabled a leap in performance and opened up new avenues for high performance computing on Linux.



NICK VEITCH breaks his own records to bring you a report on the SGI Altix.

High performance computing (or HPC) is a big area for Linux. With commercial software and operating systems often being licensed on a processor basis, building large clusters has been Linux territory for some time. Sideways scaling (running large numbers of independent but interconnected nodes) is one thing, but Linux was never built for vertical scaling, and performs badly when run as a single instance on a large multiprocessor system. At least, that's what Linux's detractors have always tried to have us believe. But the Altix 3000 series may just change all that. So, how does it work?

Memory

One of the key considerations in multiprocessor/cluster environments is memory handling. Since any memory location on the system may be required by any individual processor, this is usually one of the performance bottlenecks – the memory may be local to processor number 1, but if processor number 12 needs it, the time taken to transfer the data is usually far greater than the time to perform the actual processor operation. The usual system architecture solution is NUMA – Non-Uniform Memory Access, which allows for the fact that some data is stored in 'local' memory, while some is stored elsewhere on the system. SGI have addressed this problem with their NUMAflex technology. Data from remote memory nodes still has to be transferred, but SGI has redefined the interconnect fabric which connects the processor units together. The new NUMAlink interconnect fabric carries memory data and network information between units and nodes with latencies as low as 50 nanoseconds – a figure so low that to all intents and purposes, all memory on the system can be considered part of a global pool.

Numaflex isn't a new technology – it was first introduced to the SGI Origin series on MIPS 64-bit processors in 1996, but this is the first time it has been used on an Itanium system. The same principles apply to the construction of a system. Components are packaged together into 'bricks' which are then put together to build a system.

XFS

The Altix range of systems also implements the XFS filesystem developed by SGI. Who needs another filesystem? We already have ext2, ext3, Reiser, JFFS and many more. XFS was developed by SGI to provide a high performance 64-bit compatible filesystem that was journaled, but also gave a similar performance to a traditional filesystem like ext2.

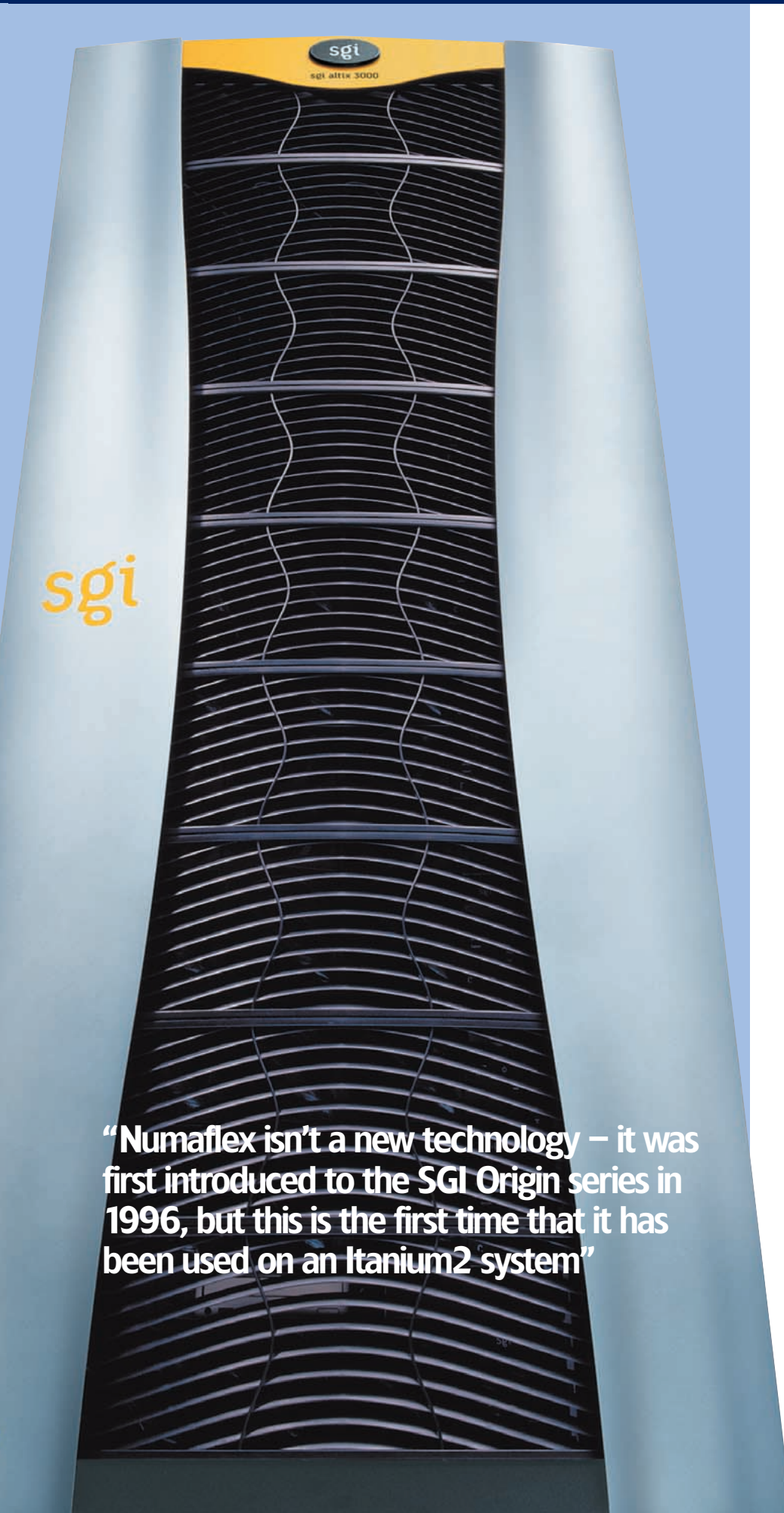
On massively multi-user systems, the journalling overhead becomes more of a hindrance, and ext2 easily outperforms ext3, Reiser and others by a wide margin. By contrast, XFS can actually outperform ext2 in some circumstances, and gives better overall performance than ext3, while still providing robust journalling.

There are also further useful additions in functionality. The XLV volume manager allows flexible attribution of partitions on the system, with an number of additional options such as concatenation (pooling disk

space), striping (optimising disk access for speed) and plexing or mirroring data (for extra security). There is support for dynamically altering the XLV space – very useful for high availability systems that are running out of storage!

To help programmers, XFS can also operate in a guaranteed I/O rate mode. Rather like USB2 and IEEE1394, this means that a certain data bandwidth can be guaranteed by the system for a time, enabling the confident ability to handle streaming media or data acquisition software whenever it's needed.

For backup purposes, XFS also supports active disk dumping – you can dump an image of the disk without unmounting it, and get a consistent, useable image, even when the device is actively being used. This certainly makes backups an easier task, and SGI provide utilities to make best use of this feature.



“Numaflex isn’t a new technology – it was first introduced to the SGI Origin series in 1996, but this is the first time that it has been used on an Itanium2 system”

Bricks

Each brick is a 1 or 2U box containing one of five different system building components. Usually the most populous of these in any given system is the ‘C-brick’, which contains processors and memory. A single C-brick can contain up to 4 Itanium 2 processors and 32GB of shared memory. Importantly, it also contains two specialist ‘SHUB’ ASIC devices, used to communicate between the processors and memory within the brick, and other C-bricks within the node or in remote nodes. As each SHUB has a maximum of two processor located on its Front Side Bus (FSB) it is capable of providing a bandwidth of up to 6.4GB/s. Limiting the number of processors on the SHUB FSB was a design decision to maximise the available bandwidth on each C-brick and throughout the system to minimise the effect of bottlenecks.


The other brick types provide storage and interconnect functionality. The IX brick is the base I/O module, the R-brick provides router interconnect facilities, the larger D-brick provides hard disk storage and the PX-brick adds space for up to 12 PCI-X hotswappable cards.

The flexibility of the brick system means that servers can be customised for their tasks. Some applications may largely require processing power, in which case a single rack can house 32 processors and the require interconnect. For a system where large volumes of storage are required, a 24-processor system could be built with space for a D-brick in a single rack.

While the NUMA interconnect system blurs the line between nodes and components of the same node in terms of connection speeds, there is a limit of 64-processors for running a single system image – ie for the purposes of processes and the OS, a 64-processor system can be seen as a single entity.

Linux doesn’t scale?

A 64-processor Linux system seems a little adventurous. Aren’t we supposed to believe that Linux doesn’t scale well? While it is still true of any useful operating system that you can’t get linear scalability in real performance, Linux isn’t that far off the mark these days.

Linear scaling would be if a task on a single processor machine could be performed 64 times faster on a 64-processor machine. Due to the overheads of inter-processor communication, this is an ideal rather than an achievable result. But improvements to the kernel can make a difference to scalability, and a lot of work on this has already been done, thanks to the profile of scalability as an issue. SGI has been able to contribute to kernel development here, as well as build on the work of groups such as the Linux Scalability Effort. 

◀ The LSE is really a collection of subprojects that cover matters such as scheduling and support for NUMA. Kernel support for NUMA systems is important, because scheduler awareness of the physical layout of the memory is crucial to making sure that processes are dispatched to CPUs adjacent to the memory they will access. This obviously results in a more efficient system, where processes can utilise memory directly through a local SHUB. Without this kind of system being present, the process could be running on a processor far removed from the memory allocated to it. In a system like the Altix the memory latency is very low, even in this sort of scenario, but it is still an inefficient use of the resources and will undoubtedly lead to bottlenecks as more and more processes are allocated.

SGI has further expanded on this functionality by adding kernel additions such as CPUMemSets, which allows processes to be directed to run and use the resources of specific groups or single processors. This allows for a finer grained management of the system, with priority given to the processes that require it.



The 'entry level' Altix 3300 server is a half rack device which can accommodate up to 12 Itanium 2 processors and 96GB of pooled memory.

As well as functionality improvements, SGI has also implemented a number of monitoring techniques that are of special interest to high performance computing. One of these is the Linux Kernel Crash Dump module, which is aimed at providing more useful crash information. The LKCD software will save the kernel memory image in the event of a crash and analyse it on restart to determine the exact nature of the failure.

Interestingly, the construction of 64-processor systems gave SGI to the opportunity to test a lot of other system software and its suitability for this sort of high stress work. For example, the Altix systems implement the devfs device filesystem for managing device names. One of the key advantages here is that the device names are persistent across reboots, and don't vary with whatever interface was registered first – a great boon when dealing with tens of disks spread across different fibre channels and controllers.

One further improvement to Linux on the driver side is the implementation of the SGI SCSI subsystem, originally developed for IRIX. While much work is going on in reinventing the SCSI subsystem for Linux to get over some limitations (especially when it comes to multiprocessor systems); the IRIX SCSI system running on the Altix delivers much better performance overall.

DEVELOPING SOFTWARE

Having the fastest Linux box in the world isn't much use if you haven't got any software to run on it. Fortunately, there is a rich environment of development tools available. Intel produced their own compiler for Linux, which will optimise software for the Itanium series. The good old workhorse GCC can also compile optimised code to an Itanium2 target. Both of these will also work with Fortran source code, which is still in wide use in the field of computational mathematics.

As well as compilers there is a choice of debuggers. The Intel debugger partners

the range of compilers and provides support for MPI and OpenMP systems debugging. The Gnu debugger is also available and will work equally happily on 64-bit code.

A range of support libraries is also provided by SGI. These are essential for making use of some of the features of the Altix environment, such as the MPT, CPU sets, array services for building clusters, and the SGI SCSL (Scientific Computing Software Library) which provides a huge range of optimised, useful scientific and mathematical functions.

ITANIUM 2

Work on the Altix servers began on Intel's original Itanium processor. The original Itanium was bulkier, more costly and generated a lot more heat, so building huge superclusters out of them wasn't a particularly easy task. Although the Itanium never sold in great volumes, it was seen very much as a developer release to give hardware builders and software developers the 64-bit Itanium environment to work with. Although some Itanium-based servers did appear, it wasn't until the launch of the Itanium 2 in 2002 that most server manufacturers began to introduce devices based around this processor.



Available in 900MHz or 1GHz clock speeds, the Itanium 2 features three levels of cache – a 32K level 1 cache, 256K level 2 and up to 3MB integrated level 3 cache. Combined with a 6.4GB/s system bus bandwidth, this gives the Itanium 2 impressive performance, twice the computing power of the original Itanium. New technologies in the Itanium 2

also include a Machine Check Architecture, which handles various system errors intelligently to enable highly available systems.

As a result, the Itanium 2 features widely in high-performance solutions from suppliers such as NEC, HP, Toshiba and Unisys, as well as SGI themselves.

Linux and the Itanium have a long history. Intel encouraged the development of 64-bit tools for the Itanium and kernel support for the processor long before launch. That's now paying off, as the Itanium 2 is able to leverage a host of tools and software already ported to 64-bit Itanium Linux.

Consequently, reported throughputs are up to 1000MB/s, around five times faster than the current kernel drivers.

Breaking records

So, what records have been broken? Well, the result of all this effort is the best-performing single instance of Linux currently available. In SPECfp_rate2000 tests, a 64-processor Altix 3000 recorded a score of 862. By comparison, HP's Superdome, a PA 8700 based supercomputer, scored 267. That's not breaking a record – it's smashing it spectacularly. Similarly, a 32-processor based Altix outperformed IBM's eServer to record a SPEC CFP2000 score of 443, more than 1.5 times the performance of the IBM machine.

The Altix set another record for memory bandwidth, recording an amazing 125GB/second on a single instance of Linux running across 64 processors. That's over 4 times the performance of the Superdome.

Of course, benchmarks are benchmarks, and the real world is a completely different kettle of bottlenecks; but the Altix range is already able to address High Performance Computing demands with real-world applications. SGI have tested the Altix with a range of software commonly used in HPC environments and have been pretty pleased with the results. The STAR-CD computational fluid dynamics software for example (used in the automotive industry) managed a 45 times speedup when running on 64 processors compared to a single processor running the task.

All SGI has to do now is watch the order book fill up – there's no better advert than good results.. In the meantime



The Altix 3300 is designed for smaller tasks, as a supercluster node or development platform for the 3700.

“As well as functionality improvements, SGI has implemented monitoring techniques of interest to high performance computing.”

they have not only proved the Altix concept, but also thoroughly discredited the notion that Linux doesn't scale for high performance tasks. ■