



INTEGRATION SYSTEMS

## A COMPETITIVE VIEWPOINT



# NUMA versus UMA The IBM Shared Everything Advantage

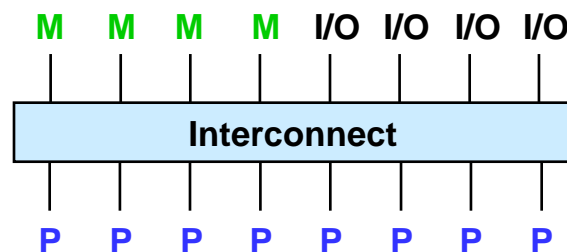
By  
Terry Keene

Integration Systems, LLC  
350 Fairway Drive  
Suite 110  
Deerfield Beach, FL 33441  
[www.e-isys.com](http://www.e-isys.com)

# A COMPETITIVE VIEWPOINT

UNIX system architects, system administrators and database administrators are being shocked every day by the performance that the IBM System p POWER6 based UNIX servers are achieving on real live production workloads. The difference in performance between the IBM System p and Sun's SPARC servers, the Fujitsu APL kit, HP's Itanium inventory and even the Intel/AMD 32/64 x86 offerings is that IBM System p has a 2X to 40X advantage in throughput and response time over all the rest. Many proof-of-concepts have been run and rerun in disbelief when there is such a broad gap in performance between what most in the technology business consider a "leap-frog" technology market. There are many significant differences between these servers; near 5GHz processor speeds for IBM versus 1GHz to 3GHz of the other offerings; 2X to 4X the L2 cache on chip versus other designs; distributed switch interconnect at 2.3GHz to near 5GHz versus 150MHz to 1GHz for others; and simultaneous multithreading versus "concurrent" or context switched multithreading just as a few examples. But the one truly significant advantage that IBM has in their POWER processor design is a shared everything architecture versus non-uniform memory access (NUMA) multi-level interconnects of the other servers. The NUMA architectures of Sun, HP, Intel and AMD limit throughput, increase the latency and restrict SMP scaling. The shared everything architecture of the IBM POWER design is a legacy transferred from the mainframe that IBM has been perfecting for 50 years.

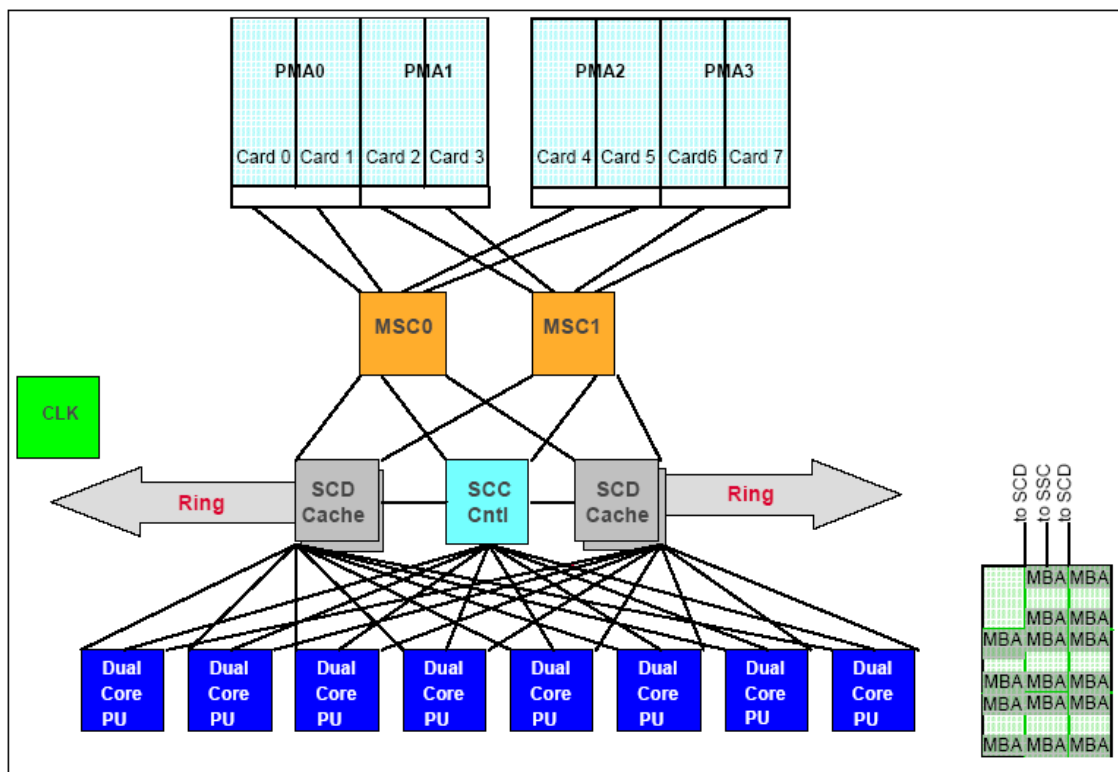
## IBM Shared Everything Architecture



# A COMPETITIVE VIEWPOINT

The IBM shared everything design of the IBM System z/p/i architectures allow near direct shared access to all of the system cache, memory, and I/O by every processor. The mainframe was the first architecture to adopt this design in order to scale to 54 processors and terabytes of memory without trading off latency. Every gate that data must traverse between the processor and any level of memory incurs latency and reduces response time and throughput. As a result the mainframe is built with an interconnect between processor books that provides every processor access to all of the system cache, memory, and I/O with no more than two connections in the worst case. Thus we refer to the mainframe as uniform memory access across the system.

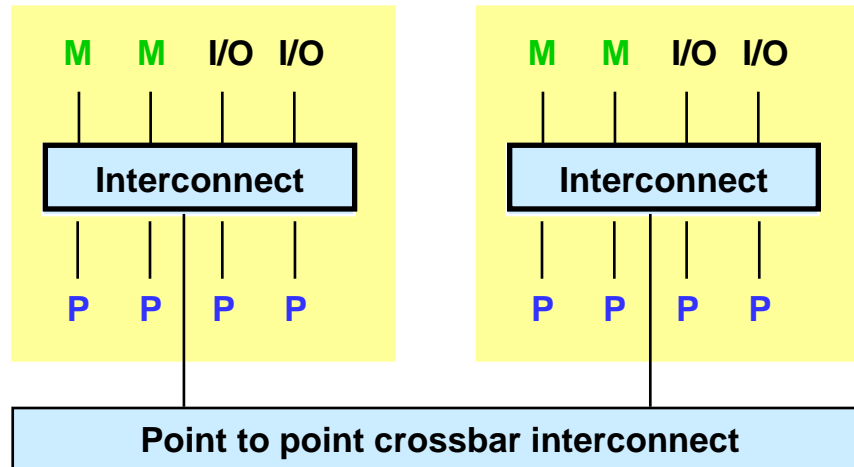
## IBM z9 Architecture



In addition to the shared everything design, the z9 mainframe gets much of its high throughput and low latency from 40MB of shared L2 cache per processor book. The interconnect between books is a 2.7GHz ring structure that connects up to 4 books in the same flat architecture.

# A COMPETITIVE VIEWPOINT

## NUMA Multi-Level Interconnect Architecture



Non-uniform memory access designs position local memory at each processor within a multi-layer hierarchical structure. This improves performance of local processor clusters or nodes at the expense of overall system performance. Most operating systems today support 4 processor (4 socket) NUMA structures and optimize application performance within that node or cluster. When multiprocessor systems scale beyond that design point, cache coherency and remote memory or I/O access becomes the limiting factor for growth and performance. A NUMA architecture also requires either smart application and data placement by the OS, something hard to achieve in many commercial applications, or a partitioning discipline that restricts a partition to a 4-socket group to avoid the NUMA design.

The multi-level NUMA interconnect architecture suffers multiple disadvantages as the number of processors are scaled outside each 4 socket node. Consider the Sun E25K system architecture below as an example.

# A COMPETITIVE VIEWPOINT

## SUN E25K Interconnect Architecture

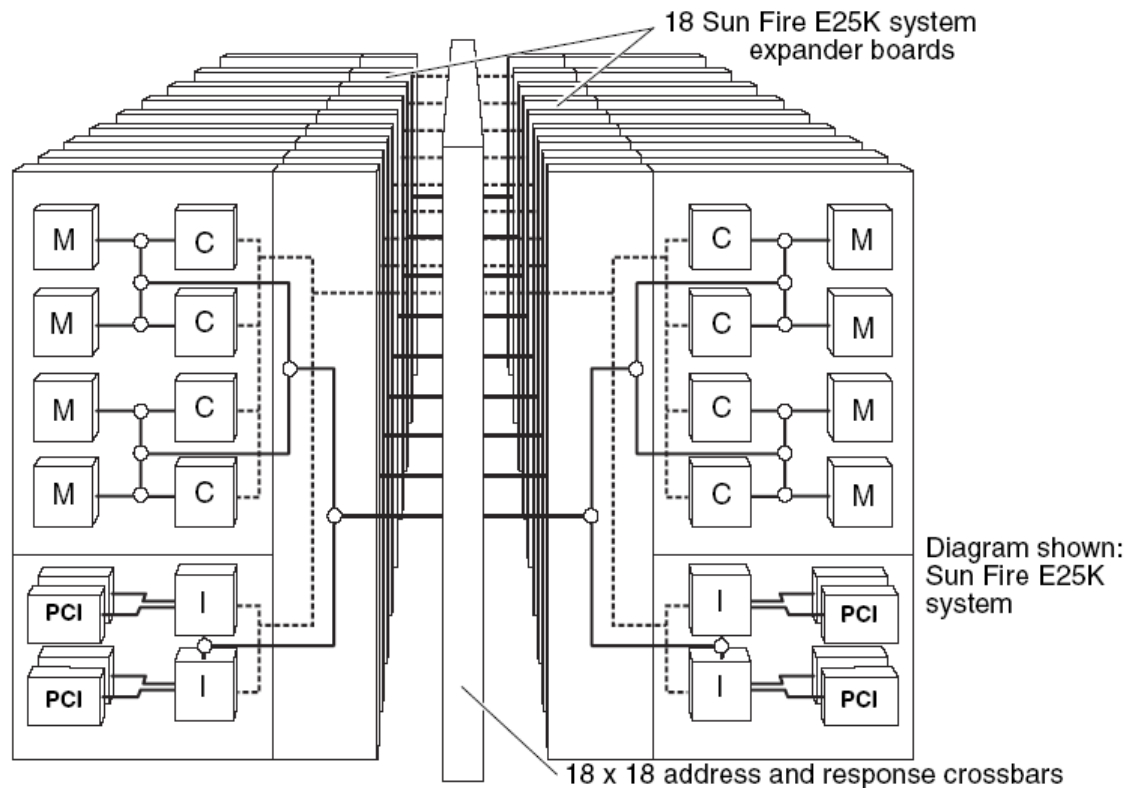


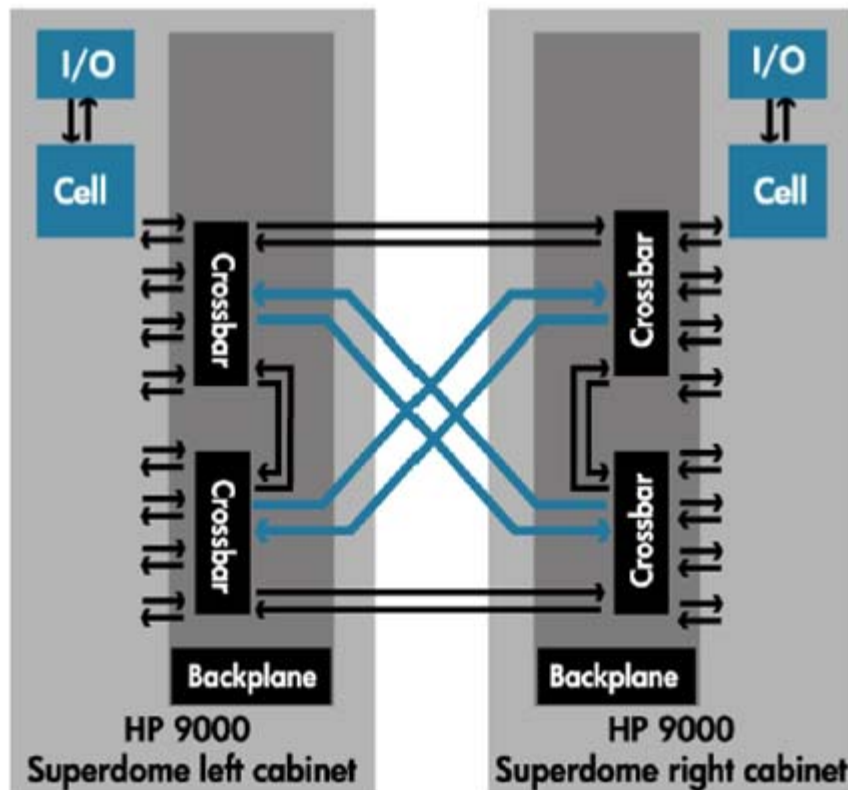
FIGURE 1-2 Sun Fireplane Interconnects

Each of the processors (C) is connected to memory on a single board, the Uniboard, through a set of gates that increases as the locality of the reference is extended. To access I/O even on the same Uniboard, two additional gates are encountered, and to access memory or I/O on a separate Uniboard, the backplane, running at 150MHz, becomes a significant bottleneck. As the processor speed increases from 800MHz to 2.1GHz, the processor to crossbar interconnect further degrades performance. As a result, NUMA across a 72 processor (144 core) E25K creates significant latency and performance actually begins to drop at some point. Smaller versions of the same Domain architecture, such as the E6900, suffer the same latencies.

HP suffers the same levels of latency with their NUMA design as you can see in this diagram of the Superdome architecture in use today.

# A COMPETITIVE VIEWPOINT

## HP Superdome Crossbar Interconnect



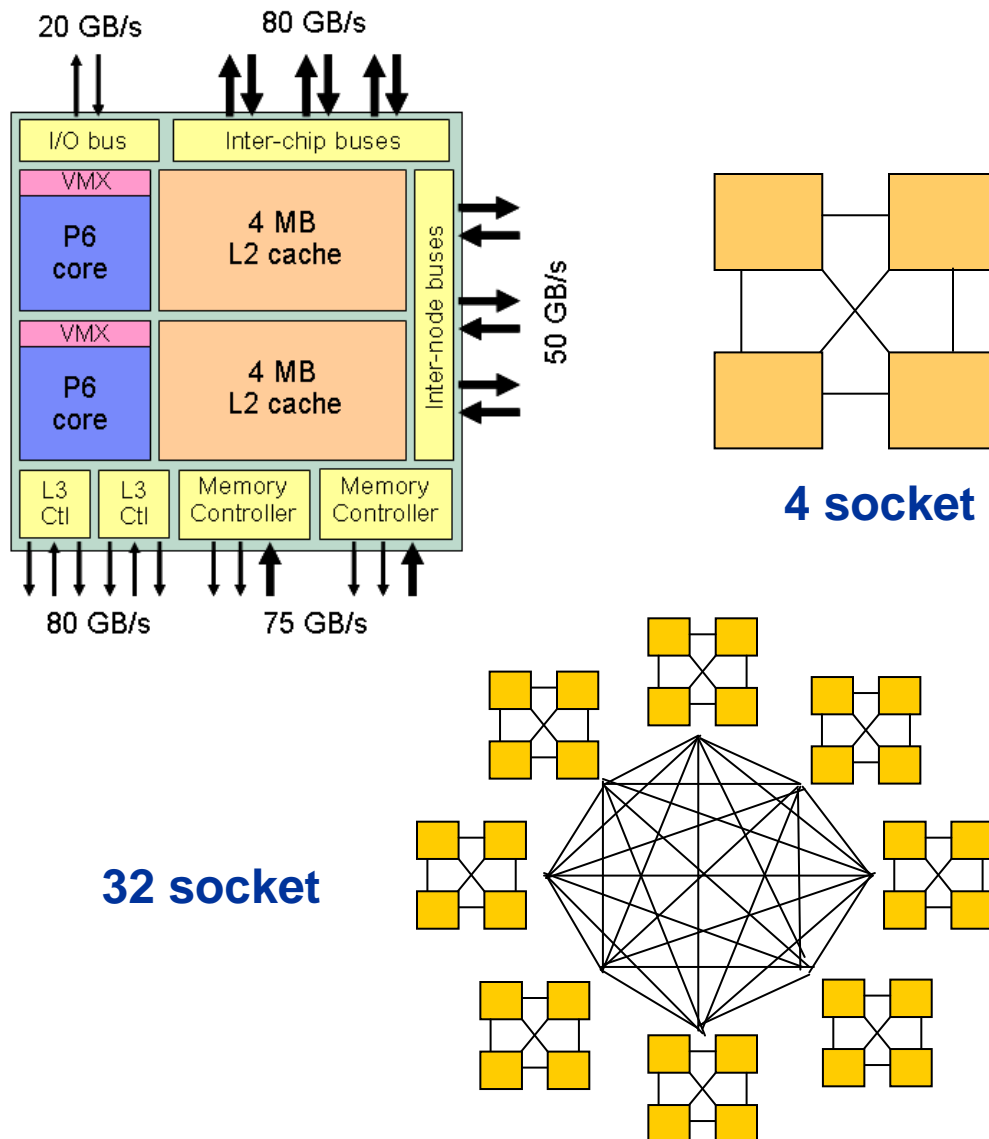
## Hierarchical Crossbar architecture

Each board is a cell of 4 processors, dual and now quad-core, that shares local memory and I/O across a 250MHz interconnect. HP's new cell architecture makes performance on a single cell board quite acceptable. However, to scale SMP up to 32 processors (64 to 128 cores depending on the processor chip) the crossbar connections introduce latency. To increase to 64 processors a second cabinet is required and NUMA "outside the box" using the flex cable becomes a significant bottleneck.

# A COMPETITIVE VIEWPOINT

## IBM POWER6 Chip Interconnect Architecture

The IBM System p POWER6 architecture has evolved from the original shared everything design adopted in the first IBM POWER4 systems from 2001. The interconnects are built into the silicon on the chip and provide a high speed interconnect not only on chip at processor speeds but also between processors and cache on separate chips, and between processor books housing additional processors and memory. The overall design is a shared everything architecture that dramatically reduces latency and improves throughput. The POWER6 design is shown below.





# A COMPETITIVE VIEWPOINT

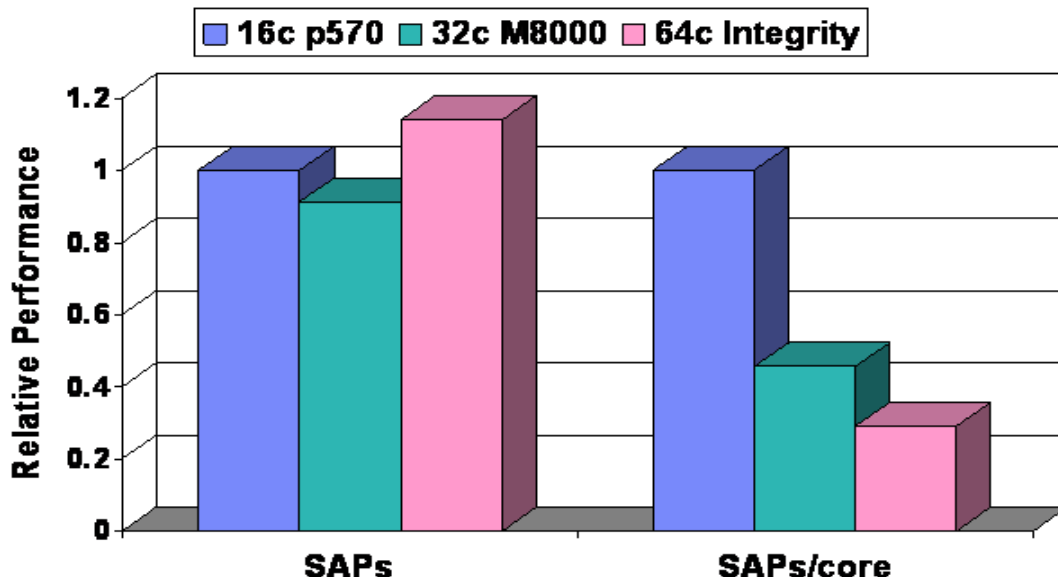
From the diagrams above you can see that the fabric bus controller, in essence a non-blocking crossbar distributed switch, is built into the silicon on the chip. This controller operates at the clock speed of the silicon between cores on chip, up to 4.7GHz in the current POWER6, and at half the processor speed between chips. The fabric bus controller on each processor socket is cross-connected to the controllers on neighboring chips and provides point-to-point connectivity from each processor in the system to every other processor within a four processor node, and point-to-point connect between every node in the server. This architecture exists in all IBM POWER Systems models from two socket p520 systems to 32 socket p595 servers supporting up to 64 cores. This kind of connectivity dramatically improves processor to cache performance as well as the performance of memory and I/O associated with other processors. There are several observations about this design that are significant advantages to this IBM POWER architecture. First, as the number of processors in a server increases, scaling from two sockets to 32, the number of fabric crossbar connections increases. Unlike the backplane multilayer NUMA designs in other systems where adding Cell boards or Uniboards to the backplane increases the overhead for a fixed capacity interconnect, the POWER interconnect adds capacity as each processor is added. Second, as the clock speed of the processor increases, from 3.5GHz to 4.2GHz to 4.7GHz, the clock speed of the interconnect fabric also increases. This removes the obvious bottleneck caused by a fixed speed interconnect backplane experienced by other vendors designs. It is notable that as other servers scale in classic SMP NUMA designs, the performance eventually reaches a peak and then actually falls off as saturation occurs in the backplane

## Performance Results from UMA design

The most effective way to demonstrate the impact of this design difference is to look at the results of head-to-head industry leading ISV benchmarks. SAP is a standard software application for sales and distribution. The SAP benchmark is designed to test the capacity of each processor architecture in order to size SAP implementations. The following chart shows the actual SAP 2-tier SD benchmark results for industry leading SMP servers from each of the three top vendors, IBM, HP and Sun

# A COMPETITIVE VIEWPOINT

## SAP 2-Tier



Notably the IBM server has 16 cores, the Sun M8000 Fujitsu server has 32 cores and the HP Integrity Itanium based Superdome has 64 cores. The overall performance is within a general margin of error that would indicate that they are comparable. As you can see, the performance from each processor core is significantly weighted in IBM's favor. This is the result of a high throughput, low latency shared everything design. IBM shows more than 2X the performance per core of the Sun (Fujitsu) M8000 and almost 4X the performance of the HP Integrity Itanium 1.6GHz 12MB L3 cache server,

## Real Life Security Trading Benchmark

In a recent head-to-head benchmark of a production electronic stock trading application, the performance difference was even more dramatic. The IBM POWER6 p570 with 8 cores ran a trading application where transactions from trading desks and automated trading programs are fed in multiple streams to a Sybase ASE database server that stores the trades and then makes them available for reporting and analysis. The application was running on SunFire E6900 servers with 48 cores that had recently been updated to the fastest available configuration. The E6900 peaked at a load of 2500 trades per minute.



# A COMPETITIVE VIEWPOINT

The IBM p570 POWER6 was able to execute almost 19000 trades per minute on only 7 of the 8 cores, the 8th being used as a virtual I/O server within the p570. That equates to more trades per minute per IBM core than the entire 48 core E6900 was able to process. On a per core basis, these results indicate a p6 core is 48X the performance of the Sun core for this application. The same trade streams were executed on a Sun M8000 and the maximum load rate was 12,500 trades per minute with 48 cores in the M8000. The list price of the 8 core p570 is well below \$300,000 as configured, while the list price of the E6900 and the M8000 are both over \$1M. Not only is the performance in a different league, but the total cost of ownership including hardware costs, software per core licensing costs, and maintenance over three years is significantly lower for the IBM p570 POWER6 server.

## Conclusion

The shared everything architecture of the IBM POWER6 server has significant advantages over multi-level NUMA designs being used by other vendors in the technology industry. The performance envelope of the POWER6 chip is 2X to 4X the performance of other microprocessors based on industry standard benchmarks and an even greater advantage on the customer benchmark exercise noted above against Sun's E6900. The UMA architecture of IBM provides the value of scalability not available to other servers. IBM has changed the competitive landscape of the market with the introduction of the POWER architecture. With each iteration from the POWER4, POWER5, and now the POWER6, IBM continues to demonstrate real competitive advantage. POWER7 is currently being developed and is expected to be introduced within the next two years. Sun's latest processor hope, the ROCK, which was scheduled for introduction earlier this decade, has been moved from 2008 to second-half 2009\*, which most probably means early 2010. Intel will introduce the next release of Itanium, the Montvale, later this year but the processor clock will only increase from 1.6GHz to 1.66GHz, well behind the near 5GHz processors that were announced by IBM last year. IBM has a five year technology lead on the rest of the market and it does not appear that anyone, not even Intel, will be able to produce a competitive offering. For the near future, commodity servers will sport the cost-effective Intel/AMD processors, The real work horses, the database and mission-critical application and infrastructure servers, will rely on IBM's mainframe expertise brought to the mid-range in the IBM POWER servers.

\*[http://www.theregister.co.uk/2008/02/04/sun\\_rock\\_2009/](http://www.theregister.co.uk/2008/02/04/sun_rock_2009/)