



HPCTimes

March 2003

IN THIS ISSUE:

- ▶ **News from Microway®** [Pg. 1]
 - ▼ *GSA Schedule Contract Awarded to Microway*
 - ▼ *In a Rush*. Testimonial from Roy Kimbrell, Northrop Grumman IT
 - ▼ *Good Things Can Come In Pairs* by Jay Owen, Vice President Business Development
- ▶ **You Can Count on It** by Steve Fried, President and CTO [Pg. 2]
 - ▼ *64 bit Processors in the HPC Cluster Market*
- ▶ **Parallel Thoughts** by Bob Condon, Storage Business Development Manager [Pg. 4]
 - ▼ *Storage Considerations for HPC and the Enterprise Applications*
- ▶ **Microway Spotlight:** Paul Howard, Chief Scientist [Pg. 6]

New From Microway

Microway Receives GSA Federal Supply Contract

Microway is proud to announce that it has officially been awarded a GSA Federal Supply Schedule contract under agreement number GS-35F-0431N. This allows Microway to sell products and services at pre-negotiated prices to Federal Government customers.

In a Rush

Roy Kimbrell, Technical Director, Space and Intelligence Systems, Northrop Grumman IT

I had a rush job to build a prototype for part of an enterprise system. On a Tuesday, I called Microway - this happened to be during the snow storm that paralyzed most of the East coast (Microway is in Plymouth, MA). Despite that, I spoke to a friendly person who wrote down my request for five 1U compute nodes plus other related equipment to be delivered in a week. Bill Applegate called me back to verify the particulars and generate a quote.

The next day, I got preliminary approval from my boss. Based on that, Microway started building and configuring my compute nodes. I started the lengthy process of getting a rush-rush corporate purchase order drafted and approved. By Friday, Microway had the purchase order. The equipment arrived the following Monday - less than a week from first contact to delivery.

The equipment works great. One minor problem was quickly resolved by their tech support. I'm impressed by how quickly Microway was able to deliver high quality equipment at a very reasonable price and by how competent and caring their tech support is.

Good Things Can Come in Pairs

Trends in two complementary technologies for HPC

Jay Owen, Vice President, Business Development

Interest is growing in the use of 64-bit processors for HPC applications. As vendors sort out performance and pricing issues with the processors and compilers, we expect to see a renewed interest in the capabilities provided. Red Hat, SuSe and other Linux vendors are releasing 64-bit versions. 64-bit technology has been around for some time (Microway has been selling Alpha-based systems in 1995). Users are exploring new methodologies to problem solving that take advantage of capabilities they offer (like very large memory address spaces).

The second trend is in the area of cluster interconnect. We are seeing growing interest in large (96 port) non-blocking InfiniBand switches. There is also potential for a family of InfiniBand switches with Gigabit Ethernet or Fiber Channel ports integrated into the fabric. These products offer users the potential to meet the original design goal for InfiniBand – simplification of transports within the data center. For the HPC market, the ability to do RDMA makes the large address space for 64-bit processors all the more intriguing.

Combining 64-bit processors with ultra low latency, high bandwidth interconnects is an opportunity for some HPC users to tackle problems that are too unwieldy today. Microway's mission is to design and integrate leading edge technology for early adopters in the HPC community. We see a real opportunity in the convergence of 64-bit processors, high speed cluster interconnects and the applications that virtually eliminate data storage as a bottleneck.

We would love to hear your thoughts on the matter. Please feel free to share them with me at jowen@microway.com

**You Can
Count On It
by Stephen
Fried, CTO**

64 bit Processors in the HPC Cluster Market

In the mid 90's one of the first markets captured by Digital's Alpha processor was the high performance computing (HPC) market. By 1996, Alpha was beating IA-32 processors by a factor of ten in typical HPC applications. (It gained a factor of two from frequency, another factor of two because it could issue two floating-point instructions per clock, another factor of two from the memory/cache interface and a final factor of two from compilers.) Alpha

offered competitive price performance (as measured in dollars/MFLOP) and the benefit of addressing large amounts of on-board memory. High profile HPC users even influenced the inclusion of specific instructions in the Alpha instruction set because they were planning to use a large number of Alpha-based systems.

By 2000, the price/performance advantage of Alpha was largely mitigated due to the battles for bragging rights between AMD and Intel. IA-32 processors improved in frequency, could issue multiple floating point instructions per clock cycle and had vastly improved compilers. Today the majority of HPC problems can be handled with a Xeon-based system provided the computation is not bound by the size of the L II cache or the need to address an extensive amount of on board memory. There has also been erosion in the memory bandwidth advantage Alpha had over its IA-32 competitors.

With the Intel® Itanium® 2 offering and the imminent announcement of the AMD Opteron™ processor we see interesting prospects emerging for 64-bit processors in the HPC market. Two factors will drive how widely these technologies will be adopted. They are price/performance and the ability of the hardware to perform tasks that are not possible with today's 32-bit architecture.

One potential for 64-bit based systems is to address problems currently too complex to handle cost effectively on 32-bit machines. There are many applications requiring manipulation of very large data sets. These would run much faster when as much of the data set as possible has been loaded into either cache or memory. While modern parallel processing techniques make it possible to effectively parallelize many problems, there exist a large set of users who don't have the time to use these techniques. As a result, most of the problems that get run on clusters turn out to be embarrassingly parallel rather than elegant parallel implementations of problems which don't parallelize efficiently.

Based on load distribution in a typical research lab, about 70% of the work performed by clusters often turns out to be scalar bound and is embarrassingly parallel by nature. These problems do not require access to large caches. *However, this does not necessarily mean that these applications do not benefit from access to 64 bit addressable memory.* At Microway we are doing research into DSM (distributed shared memory) and how DSM maps to technologies like *InfiniBand* (IB). What we have been examining is the use of InfiniBand E2E contexts in parallel processing. These constructs make it possible to map the memory of one processor on one node into the memory of a processor on another node (i.e., set up shared memory between nodes). They do not require the shared memory coherency that a typical shared memory system possesses.

When one starts to investigate the implications of using DSM, one quickly observes what appears to be a "32-bit wall." For example, suppose we wanted to perform a large 2D FFT on a cluster that contained 32 nodes. Assume we are willing to devote 400 MB of each node's memory to the storage of a section of this array and that we would use another 400 MB as an input buffer for the second half of the problem. During the first phase of the algorithm, each of the processors in our cluster will perform column wise FFT's on the columns in their 400 MB store. During the next phase, they will use the DSM feature of the IB hardware to

perform operations like a matrix Transpose. The right way to perform this Transpose is to have each of the processors write their data into the input buffers of the other processors. This can be done easily using E2E IB contexts and mapping these contexts into the virtual address space of the application. In writing the algorithm, the user does not have to worry about the details of where things are in the cluster, etc. The DSM characteristics of the interconnect handle these details. All the user has to worry about is performing a Transpose efficiently which in FORTRAN boils down to writing the data to DSM memory in column order. Here's the rub. To get this job done without the use of special compilers and other tricks requires the processor to be able to address 12.8 GB of linear address space, a feat that can only be carried out by a 64-bit processor. This is only one example of the use of DSM.

Another problem in today's parallel processing paradigm is MPI. While easy to use, MPI is not that efficient. In the days when 10/100 Ethernet took 130 microseconds to transmit a small packet between processors (i.e., it had a latency of 130 microseconds), the overhead associated with MPI did not seem significant, especially if the coarse grain problem being run was one of those in which, once every 60 seconds or so, a processor would require a small amount of data to start the next iteration of a problem. However, with hardware latencies now approaching 2 to 3 microseconds, we are discovering that the 10 to 12 microsecond MPI latencies that result from porting MPICH to new hardware devices are becoming the major bottleneck in fine grain parallel problems, such as the matrix Transpose just mentioned.

To run these problems efficiently, one requires a communication paradigm which simplifies the addressability of arrays which are distributed across a cluster. MPI is not the right paradigm for this problem, although the 2.0 specification is adding features which take advantage of "Remote Addressability." Other ancient paradigms, such as the Cray SHMEM approach, are ideal for these kinds of problems, and will probably become increasingly important in the future.

Another 64-bit application is mapping very large data sets to large distributed file systems. We know that only a small percentage of the genes in the human genome actually get used to characterize us as a species. Geneticists and microbiologist are now starting to unravel the importance of these genes in the specifications of proteins, which in turn end up becoming the "handles" which control human subsystems. These handles are much more complex than the genes used to encode them. The chemical attributes which determine their functionality will end up dwarfing the human genome itself, thereby requiring large disk farms for storage. For this storage to be meaningful and fast, it might have to be content addressable. Again, we find 64-bit address spaces coming to the rescue.

Users should be aware that these capabilities will come at a cost. One mitigating factor is performance/watt. Dense clusters can be built into a single 44U cabinet with one Teraflop throughput. A cluster with that computational capacity can consume nearly 25KW of power. Typically systems with this type of computing power use hard disks and large memory banks that further tax the power budget. Air conditioning and power become key considerations. 64-bit processor based systems will likely be even more expensive to operate than their 32-bit predecessors.

The bottom line is that we are on the cusp of a new computer revolution, like every other technology revolution that has come along about every twenty years ago or so. The phase we are currently entering is going to require more than 32-bit addressability, and the problems that require this addressability really won't be known until they present themselves. All that is known for certain is that we are approaching the 32-bit limit rapidly in today's mainstream processors, and that the next generation is going to have to be able to address much larger regions of memory to solve tomorrow's HPC problems.

**Parallel
Thoughts
by Bob
Condon**

Storage Considerations for HPC and the Enterprise Applications

Independent market research studies forecast the Linux cluster market will reach \$640M in 2003 with growth to \$1.8B by 2005. These studies also suggest that Linux clusters will account for 80% of High Performance Computing (HPC) sales by 2005. The reason can be summed up in two words...price and performance.

It is no wonder that corporate IT groups are starting to deploy Linux clusters in enterprise applications such as data mining, parallel databases, image processing and analysis, statistical analysis and batch processing. However, before this transition can develop to its fullest potential, Linux cluster solutions must incorporate enterprise features that offer availability, scalability and manageability. These features are largely centered on data storage, data integrity and reducing downtime.

As the old adage goes... time is money. Superior price/performance is of no value if the system is unavailable or critical data is lost. Recreating or reloading data is an expensive and time-consuming proposition assuming the data can be recovered at all.

By employing software that monitors independent nodes in a cluster and automatically fails over (in the event of a node failure) downtime can be reduced or eliminated. Policies can be set to configure the fail over from one-to-one, one-to-many, or many-to-one depending on customer preference. Once cluster nodes are protected against failure it's time to turn attention to storage. Using a Storage Area Network (SAN), redundant Fibre Channel switches can be configured in front of dual RAID controllers creating a highly available, shared storage pool. Each node in the cluster connects to each of the two switches by way of a dual ported host bus adapter.

No one wants to buy a point solution that cannot scale and must be upgraded by means of a 'forklift'. Scalability can be achieved through the use of load-balancing software, scaleable database offerings (such as Oracle 9i RAC), and/or Global File Systems (GFS) combined with a SAN that offer seamless growth as requirements demand. As requirements grow the system can be incrementally scaled with little to no disruption. Networking technologies such as Gigabit Ethernet, Myrinet and InfiniBand provide sufficient bandwidth for both infrastructure and application software.

Manageability issues can take many forms, but the most common problems with cluster solutions center on monitoring nodes and the provisioning and protection of data storage.

Microway cluster solutions solve the problem of monitoring with our Microway Cluster Management Software (MCMS™) and NodeWatch™ remote monitoring and management tools. For more on NodeWatch and MCMS please visit www.microway.com/mcms.htm

Managing storage may be more difficult than managing the cluster itself. In recent years the storage industry has been moving at a furious pace away from Direct Attached Storage (DAS) toward SAN or Network Attached Storage (NAS). Although SAN and NAS use different approaches to solving the problems of ‘islands of storage’ there are two basic benefits both technologies provide: *better utilization of resources* by ‘pooling’ storage to be shared by heterogeneous systems and *ease of management* by centralizing storage. A SAN is a block based storage pool using the Fibre Channel protocol similar to SCSI. NAS is a file based shared storage server using protocols like NFS, CIFS or http to access data via a network such as Gigabit Ethernet. Microway cluster solutions offer SAN, NAS or a combination of both, depending on customer requirements.

A cluster that uses a load-balancing product such as LSF from Platform Computing is essentially a grouping of individual computers loosely coupled using a networking backbone. Without the use of a SAN or NAS each node would have dedicated storage resulting in wasted capacity and time consuming management. Making physical connections to a SAN is only half of the answer since each node would be allocated a partition of the SAN using techniques such as zoning or LUN masking. To solve this problem Microway includes a GFS with a distributed lock manager. The GFS allows sharing of storage at a file level by locking files when they are in use by other nodes in the cluster. With read or read/write permissions enabled files can also be shared between nodes in the cluster, or with other systems connected to the SAN and supported by the GFS. This sharing eliminates the need to transfer files between nodes across the network, saving bandwidth and latency in addition to more efficiently managing storage resources.

Alternatively, a dedicated NAS server from Microway could be used to pool storage. Microway’s NAS offerings scale from entry level to multi-Terabyte offerings. The NAS approach does not require a GFS since data is stored and accessed at the file level.

Finally, Microway offers a combined solution that uses a NAS front-end and a SAN backend. The benefit of using this approach is lower cost per node connection using Gigabit Ethernet vs. Fibre Channel and unlimited scalability on the back-end by virtue of the SAN.

Affordable Linux clusters offer unparalleled price/performance as evidenced by their rapid growth in the HPC market. As Linux moves up the food chain customers will demand total solutions including robust storage designed to meet their unique needs. As usual, Microway continues to offer leading edge technology. Since pioneering Linux clusters in the late 90’s we have anticipated the need for ‘bullet proof’ availability, scalability and manageability by our customers as they replace expensive proprietary systems with Microway clusters.

**Microway
Spotlight**

Microway Spotlight on Paul Howard, Chief Scientist

Paul is a world expert in arithmetic coding for data compression, and holds eight compression related patents. His other skills include compiler and user interface design. Before re-joining Microway in 2002, Dr. Howard worked for eight years at Bell Labs and AT&T Labs. In the mid 1980's, he developed several software products at Microway before completing his Sc.M. and Ph.D. in Computer Science at Brown University. He also holds a B.S. in Computer Science from MIT. He is involved in technical product development and oversees the Microway customer benchmark program.

Tell us a little about your background and experiences over the past years.

I went to MIT as an undergraduate. They say your goal at MIT is not so much to win as to survive; well, I survived, despite changing my major from meteorology to computer science after my third year. After that, I never wanted to go to school again, so of course a decade later I left Microway to enroll in the Ph.D. program at Brown. Brown's CS department was small enough that I didn't get lost, and my advisor was Jeff Vitter, a brilliant researcher and a demanding advisor. He helped me to become an expert on arithmetic coding for data compression. I had a NASA fellowship, so I also did some work on lossless image compression. One of the algorithms I developed in graduate school is now being implemented in hardware by SEAKR Engineering for the Mars Reconnaissance Orbiter, scheduled for launch in 2005.

During my education, I worked at a number of different and interesting places. At Tracor in Falls Church, Virginia, I designed all the software for a laboratory prototype of a cell phone system for the Inland Waterway ... this was in 1980! I spent a year as a short term missionary for Wycliffe Bible Translators in Dallas doing software work to help translators with their linguistics and literacy work in the field. I was with Microway in its early days, from 1984 to 1987. Among other things I wrote 87BASIC/INLINE, a compiler postprocessor that for at least five years generated the fastest floating-point code for the PC from any high-level language. After graduate school I joined AT&T Bell Labs (later renamed AT&T Labs) in New Jersey. While at AT&T I was editor of the JBIG2 International Standard for bi-level (black and white) images. I was granted a number of patents (nine at last count) for various aspects of text, image, and video compression.

How has high performance computing evolved over the past 20 years?

After a number of years doing research at the application level, I've found that the world closer to the machine has changed immensely! Instead of just counting machine cycles, now we're intimately concerned with cache effects, latencies, and instruction level parallelism. And of course MPI has made parallel computing readily available and usable for all kinds of scientific research. In addition, 64-bit architectures will allow scientists to solve problems of magnitudes unimaginable up to now.

Your ideas on what prospects HPC holds for the future

When Moore's Law finally slows down, as it eventually must, the only way to achieve increased performance will be through massive parallelism. The problem now is that tools like MPI, useful though they are, are still rather primitive, like assembly language for communication. As computing and communication hardware become faster, the challenge

will be to develop software methodologies to take advantage of the raw power. I suspect this will come out of a marriage of industrial and academic research.

You said earlier that you left Microway to go back to school -- why did you come back to work at Microway?

Well, the obvious answer is that after massive layoffs at AT&T Labs Research last year, I needed a job! Microway is a great place to work. Ever since the early 1980's it has provided hardware and software that the scientific community has needed for the fastest possible computation. What distinguished Microway in the 1980's was its philosophy of strong customer support ... if something went wrong, we fixed it fast. I've been pleased to see that that hasn't changed a bit. The result is a strong sense of loyalty: our customers know that they can count on us, and our employees know that they're producing a quality product that will be well-supported in the field.

From a technical point of view, my work is very interesting. I'm fortunate to have the opportunity to see first-hand how high-performance clusters work, and to be able to tune the software to improve performance on new network architectures. It's also satisfying to be able to help customers run their application benchmarks ... it gives me a sense of the kind of real-world problems our customers have to solve.

To receive HPC Times via email or to tell a friend about it visit: www.microway.com