



November 2002

HPC Times

IN THIS ISSUE:

- ▶ **What's New at Microway®**
 - ▼ A detailed look at **NodeWatch™** [Pg. 1]
 - ▼ Our participation in Intel's **Premier Partner Board of Advisors** [Pg. 2]
- ▶ **SC2002 Bulletin:** This year will be one of the most exciting for new technology at SC2002. Microway will be showcasing InfiniBand®, Intel® Itanium® 2, NodeWatch, MCMS™ (Microway Cluster Management Software) and Clusterware™. Visit Microway at Booth 1022 and others to discover leading edge connectivity, cluster monitoring and management, and grid computing. [Pg. 2]
- ▶ **As I See It by Stephen Fried, CTO:** History of the HPC Market. Transputers and the history of distributed memory parallel processing. [Pg. 3]
- ▶ **Parallel Thoughts:** 20th Anniversary Research Grant-You could win! [Pg. 8]
- ▶ **HPC Community Feedback:** Creating a community of questions and answers. [Pg. 8]

What's New at Microway

Microway NodeWatch™ - Remote Hardware Monitoring Solution

Now available on Microway 1U Athlon and 1U Xeon clusters!

NodeWatch measures, reports and records the physical health of each node in your Microway cluster. NodeWatch provides fan speeds, temperatures, and power supply voltages, in addition to providing physical control over front-panel power and reset. NodeWatch software delivers all this capability remotely via a web-based GUI.

NodeWatch features:

- ▶ Web-enabled remote cluster status and physical power and reset control. Automatically monitors up to 5 temperatures, 3 voltages, and 11 fans, including power supply fans.
- ▶ Out-of-bounds values are displayed uniquely.
- ▶ Monitors nodes, even those that are powered off.
- ▶ Cluster or individual nodes can be switched on and off, rebooted, and reset remotely.
- ▶ Physical measurements integrated by Microway into Ganglia reporting software.
- ▶ Common operating system tasks, such as shutdown, are integrated into the user interface.
- ▶ Completely motherboard and interconnect independent. Monitored information collected via independent 18MHz embedded NodeWatch board and transmitted via physically independent RS485 serial network. All software resides on master node.
- ▶ NodeWatch is fully customer configurable and information is delivered securely through MCMS™ (Microway Cluster Management Software).
- ▶ Two-year warranty.

NodeWatch is the affordable option that gives you peace-of-mind through remote monitoring and control of your Microway cluster. Call Microway for a live Web Demo!

Chairperson Appointed to Intel® Premier Partner Board of Advisors

Ann Fried, Chairperson and Cofounder of Microway, Inc., accepted an invitation to join the Intel® Premier Partner Board of Advisors.

Microway is one of only 15 organizations from North America selected. This appointment reflects the quality and level of products and customer service Microway has provided to the high performance computing industry for 20 years.

"The ongoing mission of this Board is to provide input to Intel for improving its Premier Partner program, which is offered to its top 350 vendors in the US and Canada. The recommendations we make will strengthen this program and result in better products and more timely information for both Microway and our customers," said Ann Fried after accepting the appointment.

Microway is an Intel Premier Partner – Education Specialist and utilizes genuine Intel processors, chipsets, and motherboards in many of its server and Beowulf cluster products. Microway is also a value-added reseller of Intel compilers and software development tools.

Ann attended the initial board meeting in Orlando, Florida on October 27 – 29, 2002.

**SC2002
Bulletin**

Microway Exhibits at SC2002, November 18-21, Baltimore, MD

Find Microway solutions at:
Microway Booth 1022
Intel® Booth 1835
Raytheon® Booth 1319
Platform True Grid Demonstration

MICROWAY BOOTH 1022

This year will be one of the most exciting for new technology at SC2002. Microway will be showcasing its new InfiniBand products, NodeWatch, MCMS (Microway Cluster Management Software), Intel® Itanium® 2, and Clusterware™. Visit Microway at Booth 1022 and others to discover leading edge connectivity, cluster monitoring and management, and grid computing.

- ▶ Microway will be showcasing exciting products based on InfiniBand technology. These include a 12-node Xeon cluster interconnected using InfiniBand switches and HCA's. Eight of the Xeon nodes are housed in a 5U Rack mounted chassis that can hold up to 12 Microway InfinaBlade™ Xeon Processor nodes and a pair of redundant InfiniBand Switches. The blades and switches are interconnected by a common backplane and powered by a 2 KW industrial-grade power supply. Four of the nodes are mounted in 1U Microway Xeon chassis and are interconnected to the InfinaBlade Chassis using four InfiniBand 4X connections and a Mellanox 8-port switch.
- ▶ Intel® Itanium® 2 is the emerging leader in the 64-bit processor market. Microway has been conducting product evaluations since last summer and will have a quad Itanium 2 based system on display.
- ▶ Microway's NodeWatch remote cluster environmental monitoring and control hardware/software combination. NodeWatch is capable of managing 64 nodes per serial port on the master. It monitors fan speeds, five temperatures, and power supply voltages, in addition to providing physical control over power and reset switches via a secure, web-based GUI.

- ▶ The 8-node heterogeneous cluster displayed is built with Microway's unique 1U chassis design. Included are four nodes configured with dual Athlon™ 2200+ MPs and four nodes with dual 2.8 GHz Intel® Xeons™. Two primary applications will be demonstrated on the cluster.
 - ▼ NAMD integrated with Platform's Clusterware. NAMD is a parallel, object-oriented molecular dynamics code designed for high-performance simulation of large biomolecular systems. Clusterware is a modular application that provides load sharing, consolidated cluster-wide node status and batch job scheduling capabilities across clusters. These will run on the AMD nodes.
 - ▼ An MPI version of the Persistence of Vision Ray-Tracer (POV-Ray). POV-Ray is a rendering package that creates high-resolution 3-D images including detailed texturing through the use of ray-tracing. By running an MPI version of POV-Ray, the Xeon nodes will mimic the techniques employed in a render farm.

At Booth 1022 you will discover the uncompromising value that Microway's repeat customers have come to expect for 20 years.

INTEL® BOOTH 1835

In Intel's Booth, Microway will be demonstrating grid-computing solutions for the Life Sciences industry.

- ▶ Process critical data 100X faster with significant cost reductions
- ▶ Accelerate key applications that drive your business
- ▶ Leverage common organizational computing assets
- ▶ High Density Technology tuned for Grid Computing

RAYTHEON® BOOTH 1319

The answers to tough questions do not emerge randomly and creating any technical solution is a complex task involving the right architectures, technologies, standards, suppliers, and products.

The answer that will be highlighted for attendees at SC2002 is a solution which incorporates ideas, products, and programs from Raytheon, EMC, Microway, Brocade, CNT, and ADIC. The demonstration will focus on business continuity and potential implementations for Linux clusters as metadata servers for storage environments.

PLATFORM TRUE GRID DEMONSTRATION

Platform's *True Grid*: Enabling Extreme Interoperability Through Grid Computing

Microway, along with Platform Computing and six other corporate participants are collaborating to showcase an extreme degree of interoperability at SC2002. Platform MultiCluster, Platform Globus and The Globus Toolkit collectively enable compute resources from Platform and seven partners to be virtualized into "True Grid". In addition to show-floor resources in Baltimore, "True Grid" incorporates resources across North America.

History of the HPC Market

Over the last 20 years, Moore's law has had a dramatic effect on computer performance and HPC in particular. During this period, the primary bottleneck in HPC machines has moved from the FPU (floating point unit) to the BIU (bus interface unit, which controls a CPU's access to caches and memory). New concepts like vector registers and caches came into being along with numerous ways to use them. The history of HPC is riddled with technologies and companies that came to market, lasted two to four years, and then disappeared. However, it has also spawned a number of technologies that are still with us today, although their current uses don't precisely agree with the original innovators' concepts. Today's four HPC architectures include shared memory machines, vector machines, distributed memory machines and SIMD (single instruction multiple data) machines, which use a single instruction to simultaneously process four or more sets of data. As memory becomes an increasingly severe bottleneck, the shared memory machines are increasingly being used as business servers and not HPC building blocks.

In the mid 80s, HPC problems were limited by the speed of the FPU's available. Microway played an important role in today's modern cluster market, by writing the software that made it possible to use an 8087 in the IBM-PC in 1982. Microway continued acting as a catalyst in this market, introducing the first 32-bit Fortran that was used to port applications like MATLAB to the PC and the first Transputer and i860 PC add in cards. However, the distributed memory branch of the HPC tree we sat on did not really start to bear fruit until the mid 90's. For most of this period another branch flourished, that which was cultivated by Seymour Cray. Cray eliminated the FPU bottleneck using a pipelined vector unit that ran a factor of ten faster than the FPU's on the fastest scalar mainframes of the age. Having sped up his FPU, his next challenge was keeping it fed. He was the first one to hit the problem that now dominates our industry, keeping your pipes full. He kept the pipelines on his FPU full by adding a specialized set of registers that he implemented with high speed SRAMS (the devices used in LII caches today) and which he used to hold a portion of a large array (typically several thousand elements of double precision numbers). He called these registers vector registers. They were fed by a memory system that contained many parallel banks of DRAMs which were designed to fire in succession and avoid another problem: DRAMs of his day had latencies on the order of 150 nanoseconds. Finally, Cray needed software to drive his machine. It came from Pacific Sierra Research, which wrote VAST, a vectorizer that is still used to this day. VAST took a Fortran program and translated it into a new program, which called a library of vector operations that were perfectly matched to the Cray architecture. The combination of custom FPU's being fed by custom registers and custom memory, resulted in machines that ran an order of magnitude faster than the mainframes of the day. However, this approach lacked universal applicability because the hardware was expensive, it was tuned to solve problems where data had to be stored in arrays, and computing operations needed to be pipelined to run fast.

Not all HPC problems fit the vector model. Some turn out to be scalar bound. One of Cray's competitors was Gene Amdahl, as in Amdahl's law. This law states that after you have succeeded in vectorizing your problem, you will discover that the scalar bound portion of your application will now become a significant contributor to execution speed. If you don't do a good job optimizing that as well, you will leave a lot on the table, performance wise. Because there was a large set of problems that did not vectorize at all, or that were scalar bound, the stage was set for alternative technologies to develop. It turns out that many problems can be broken into embarrassingly parallel problems, each of which is scalar bound. This led to machines in which less powerful processors were tied together using serial interfaces, with each CPU being used to solve a small piece of a big problem. The work was coordinated by one of the CPU's that also had access to the file system and became known as the master node. The master node distributed work to the processors it controlled and gathered back the results that it then stored on its file

system. The term for this paradigm coined by Inmos, the first company to enter the distributed memory parallel processing business, was “farming.” Later it was called “fork and join.” And, in some cases, it can be accomplished without even a formal parallel processing structure by using techniques like queued batch processing. Before Inmos came on the scene in the mid 80’s, there were crude experiments being performed in a number of university labs using general-purpose microprocessors including the Motorola 68000 and Intel 80286. One of them in the US ended up being influential, and that was the Hyper-Cube group at CAL Tech. Another was at Oxford. The market we now call “cluster technology,” actually is based on three threads, which were intertwined and eventually merged into one.

Thread 1 – Transputers, i860, Alpha and PowerPC DSP

Tony Hoare of Oxford had a different idea. Instead of using off the shelf components, he built a processor from scratch that had everything on chip required to execute “distributed memory parallel processing.” In addition, he developed a programming language designed from the outset to write parallel code. The processor was called a Transputer and the language was called Occam.

The Transputer contained most of the components required to build a single node of a parallel system. Its processor core was a 32-bit CPU that ran as fast as an Intel 80486 when the 80386 was Intel’s benchmark device. It also had an on-chip memory controller and a DMA link engine that connected four devices together. Using these links, it was easy to build large arrays of processors using a number of different topologies. (This differed from the Cal Tech approach in which a single topology became the preferred way to hook up processors.) This also made it possible to place four processors on a single card that plugged into a PC’s ISA bus.

The British government funded Inmos to convert Tony Hoare’s idea into a product. Most distributed memory parallel machines that we see today, which we commonly call Beowulf clusters, are descendants of the technology developed at Inmos. This includes the switches used to link them together and the library routines used to write parallel code. In 1986, Inmos hired the English subsidiary of Microway, Microway Europe, to popularize Transputers in the PC market, which we did. A number of the companies that we competed with in this market have legacies that survive to this day. Meiko was founded by several Inmos engineers and took the high road, building Transputer based machines that they sold to government research labs in Europe and the United States. One of its founders, Gerry Talbot, led the development work on the extensible instruction set used on the T8 Transputer. He also ran API, the Samsung subsidiary that helped popularize the use of Alphas in clusters and invented the HyperTransport Bus, a part of the new AMD Opteron™ processor.

The R&D money at Inmos eventually dried up and the company was sold to SGS-Thompson, which cut things like its next generation T9000 products. However, before that happened, Inmos also invented the switch technology which stays with us to this day and uses a device called a wormhole router. The next significant processor was the Intel i860, which was modeled on the Cray 1. This new CPU had features that stayed with the Intel product line of Pentiums for close to 10 years. It included a four deep vector unit that could be run in two modes, scalar or vector. It also had an LI cache that did double duty as either a cache or a vector register and a pipelined memory unit that was perfectly matched to its FPU and the DRAMs of its day. To properly take advantage of its features you needed to purchase a copy of VAST. The i860 was the first processor to include SIMD instructions, which it used to implement special graphics instructions that were built into a pipelined graphics unit. This made it a winner in the HPC and graphics acceleration markets. Performing operations like dot products, the i860 hit 80 megaflops

at a time when the 486 hit just one megaflop (even equipped with a Weitek coprocessor)! Following Inmos's lead in 1990, Intel bowed out of the HPC business in 1992, by freezing i860 clock speeds at 50 MHz and taking many of its features and folding them into the Pentium. Microway was the only i860 vendor outside of Intel to build a board that hit the full 400 MB/Sec bandwidth of Intel's i860-XP processor. The i860 BIU was incorporated into the Pentium and became known as the FSB (front side bus). Eventually memories doubled in speed and memory controllers were built into chipsets, making it possible to hit this bandwidth with off the shelf motherboards. Intel stuck with this bus and other i860 features for many years opening the door to Digital Equipment who eventually came out with Alpha that featured a 128-bit wide memory bus and a Superscalar architecture that obsoleted both the i860 and the Pentium in the HPC market. Microway was the lone hold out of the original HPC Transputer companies which made the jump to the Alpha. However, there were companies like Alta and Alpha Data (which was a spin off of a Transputer Software company) that survived and entered the distributed memory business with Alphas. And, a spin off of Meiko, Quadrics, is alive and well to this day selling Inmos - like switch technology.

PowerPC / DSP Branch

It is important to note that a parallel branch of this thread got going around Transputers and i860's that persists to this day in the DSP world. This thread is centered on companies like Floating Point Systems, Alliant, CSPI, Mercury, Allacron, etc. All of these vendors sooner or later ended up concentrating on building data processing solutions that performed some sort of DSP. Most of them started life building custom vector boxes that attached to machines like the VAX. Alliant started life as a true HPC parallel processing solution and then moved to the i860. Floating point systems used Transputers to tie custom hardware together along with i860's, which was also the route taken by Meiko. Some of the solutions died on the vine because they relied on the Transputer interconnect too long. Some of these concerns also developed proprietary connection schemes, such as the Mercury Raceway or CSPI, used off the shelf technology that they bought from Myricom. When Intel pulled the plug on the i860, many of them died, the ones that survived moved to the IBM PowerPC or Analog Devices Sharc. The PowerPC was essentially an Alpha look alike which recently acquired a SIMD back end called AltiVec, which makes it an ideal low power DSP device. The Analog Devices Sharc is essentially a SIMD version of the Transputer that includes links for connecting processors together. Both TI and Motorola have subsequently developed DSP devices that integrate Transputer technology with modern high speed floating point cores that are aimed at the DSP market. One of the characteristics of these devices is that they excel at performing 16 to 32-bit arithmetic, which basically makes them fine for DSP applications, but which falls far short of the precision needed to solve general purpose HPC problems. Ironically, the Xeon also contains a SIMD back end, and could be used for DSP, but it takes a lot of power to run, making it a poor candidate for military applications. In the last few years, FPGA's (field programmable gate arrays) have also entered into this DSP market. These devices make it possible to build a custom machine that can perform DSP operations at very high speeds. FPGA's with up to 6 million gates, which can be programmed by software, are currently available. These devices are programmed by the tools used by EE's to design circuits, aided by libraries with off the shelf designs for "cells" that make it possible to implement adders, multipliers and interconnects. These devices typically cost \$10K and up, per board, and plug into the PCI bus of a PC. Heterogeneous clusters that contain high speed switches and FPGA's are now starting to come into this market, competing with solutions from the current leader in the business, Mercury, which connects together PowerPC's with its Raceway bus.

Thread 2 - Hypercubes

In addition to the Transputer groups, which were heavily funded in Europe, there was also a strong parallel processing effort that went on in the US. A group at Cal Tech that

included Geoffrey Fox and Adam Kolawa worked on gadgets called Hypercubes – an architecture used to hook eight processors together. Fox developed HP Fortran, a language with explicit parallel hooks built into it that was supposed to simplify coding parallel problems (for a number of reasons neither Occam or HP Fortran made it mainstream). Adam went off to found ParaSoft, a company that competed with 3L in the T8 OS business. Intel Supercomputing (a company which lasted about six years until it decided there was no money to be made in the field) productized Hypercubes out of 386's hooked to Weitek 1167 numeric Coprocessing cards in 1987. Microway also participated in the Hypercube market, building 1167 cards for Intel, selling over a million dollars worth of these cards into Intel Hypercubes during a two-year period. This indicates that Intel sold roughly 1,000 parallel nodes into the HPC market before moving on to the 80486 and eventually standardizing on the Intel i860 and Pentium. 1987 was also the year Microway started to move large number of Transputer based QuadPuters and came out with a version of NDP Fortran-386 that supported both the Weitek and Cyrix math coprocessors. The real impact of the Cal Tech effort on the HPC movement turned out to be Parasoft and more recently Myricom, which was started by Cal Tech people with a switch based on the wormhole router concept of Inmos.

Thread 3 – MPI based PC Clusters

Three years after Transputers entered the HPC market, Parasoft showed up at the 1990 Supercomputing in the IBM booth, running a wall of RS-6000's connected together using Parasoft Express – a communications library they originally sold into the Transputer market, which was an outgrowth of Cal Tech HyperCube research. This turned out to be the real impetus behind distributed memory parallel processing in the US. The following year a whole slew of universities showed up at Supercomputing with 386/486 based systems linked together using Ethernet cards driving their own communications libraries. The first standardized communication library to make it big in the US was PVM, which started to show up during 1992 – 1993. In 1992, Intel was awarded the contract to build the largest cluster of all time, which linked thousands of i860's together at Sandia and at the same time announced they were getting out of the i860 business. This was also the year Digital announced the Alpha.

MPI came along in 1994, and quickly replaced PVM as the library of choice, because it was portable, public domain and easily ported to a number of different platforms, making the code produced with it portable across architectures. However, even MPI owes a portion of its legacy to the Transputer. There currently are two versions of MPI, MPICH and LAM MPI. LAM MPI came out of the Ohio State Supercomputing group. It turns out that both of the original implementers of MPI based their product on older message passing protocols and libraries. In the case of MPICH, it was a library of routines called Chamelleon, in the case of LAM it was Trollius, a Transputer OS developed at Ohio State.

The last piece of hardware developed by Inmos, the T9000, featured an interconnect that ran at 40 MB/Sec and connected to a switch using the “wormhole router.” This device was built from a matrix of switches each of which contained a cross bar and eight links. With these devices you could build as large a switch as your heart desired, by simply plunking more devices down on a PCB in a rectangular grid. (The term wormhole exists router because as a message passed through the grid of devices, it could reside on several of the devices at once, with its head on one device and its tail on another. And as the message moved along, the path it followed would also close up behind the tail the same way a wormhole closes up behind the tail of the worm creating it.) One of its other characteristics was a header to the message that contained a series of addresses in it. As the message passed down the wormhole, the next address was used to specify where the worm went next, and was then stripped off. This is the technique that has been copied by

a number of builders of the switches we used today, including the switches that come from Myricom (former Cal Tech people) and Quadrics (former Inmos and Meiko people)!

InfiniBand – The Next Frontier

The next generation switch that is now gaining momentum and follows the Inmos Thread was co-invented by a consortium of Tier 1 vendors. This device falls under the broad label of InfiniBand, and is defined in a 2,000-page specification that is available on the web. InfiniBand provides three kinds of devices: HCA's (host channel adapters), TCA's (target channel adapters) and Switches. HCA's plug into a system's primary bus and make it possible to communicate with a switch. TCA's are similar to HCA's, but are designed to make it easy to implement things like hard disks that can talk to an InfiniBand switch. And finally, the switch makes it possible to connect together HCA's and TCA's. This makes it possible to have processors, hard disks and links to the outside world (i.e., Ethernet) communicate and exchange information over a single fabric. This architecture is being supported by a number of major silicon vendors along with a number of new players. For example, IBM Microelectronics and Agilent (a spin off of HP) are both providing silicon for creating HCA's and Switches. A new company founded by former Intel engineers based in California and Israel, Mellanox, also provides both types of silicon and appears to be an early leader in the business. Finally, several software vendors who provide products that make it easier to write applications that run on all IB hardware have come into existence as well, along with a plethora of companies who are providing switches, HCA's and specialized silicon (which translates things like GigE or FC to IB) to the market place. Altogether, close to \$500 million has gone into developing this new technology. A number of the players are coming out with silicon that mates to future AMD and Intel interconnects, such as the PCI Express and the HyperTransport bus. This provides new opportunities to companies like Microway, which for twenty years has specialized in adopting leading edge silicon to the HPC market.

Parallel Thoughts

20th Anniversary Research Grant – You Could Win!

To celebrate 20 years of providing innovative HPC solutions, Microway is offering an opportunity for you to turn your purchase of a 16-node or larger Microway cluster into a research grant for your choice of Microway products.

The research grant is our way of saying thank you to all our repeat and new customers. We couldn't have done it without you!

Visit http://www.microway.com/contest_20th.htm for full details.

HPC Community Feedback

What would you consider to be your greatest triumph in 2002? What about your greatest obstacle?

Share your thoughts with us by emailing info@microway.com and we'll post some of the answers in next month's newsletter. Creating a community of questions and answers.

To view Microway's HPC News online or receive it via email – visit www.microway.com/newsletter.htm