



HPC Times

June / July 2003

IN THIS ISSUE:

- ▶ **News from Microway®** [Pg. 1]
 - ▼ *Maximizing Cluster Price/Performance*
 - ▼ *NodeWatch™ Enhancements*
 - ▼ *Cluster News*
- ▶ **You Can Count on It** by Stephen Fried, President and CTO [Pg. 3]
 - ▼ *Cluster Robustness, Total Cost of Ownership and the Price of Nails*
- ▶ **Parallel Thoughts** by Bob Condon, Storage Specialist, Microway [Pg. 4]
 - ▼ *Storage Considerations for HPC Cluster Design*
- ▶ **Microway Spotlight:** Stephen Fried, President and CTO [Pg. 6]

VISIT MICROWAY AT CLUSTERWORLD (BOOTH 618)
June 24 – 26, 2003 in San Jose

**New From
Microway**

Maximizing Cluster Price/Performance

By Jay Owen, VP Business Development, Microway

In our industry one often hears the statement, “users need the maximum amount of computing capability at the lowest possible cost.” At least two questions should almost immediately come to mind. They are:

- How is the user defining cost? Is it the acquisition cost, operating cost or total operating cost of the cluster?
- How does the user define computing capability? Is that shorthand for “fastest processors” or does it include interconnects, data I/O, cluster I/O and such?

This issue of the HPC Times Newsletter addresses these two questions in detail. I hope you will gain some insights that will be useful in assessing the factors that are most relevant to you.

Microway’s NodeWatch™ and MCMST™ cluster management tools are designed to provide a cluster-wide view with control capabilities to the individual node. Environmental problems, like air conditioning failures during odd hours, could result in catastrophic failures on a large percentage of the cluster nodes. The cost of such an event, adjusted by the probability of its occurrence, varies by facility. While it is not a hard expenditure, this adjusted cost should be factored into the equation. NodeWatch™ as an optional management tool, may be a more cost effective alternative to unmonitored nodes.

Microway's engineers and technical support personnel are well versed in the computing capability discussion. Our work with high speed interconnects has demonstrated that applications which take advantage of low latency, high bandwidth interconnects can run faster on Linux clusters than proprietary UNIX SMP machines. However, those same applications on clusters with Gigabit Ethernet may perform poorly due to the high CPU overhead used by the communication protocol. Furthermore CPU cycles could be idle, due to the lack of data bandwidth. This problem is addressed in more detail in an article by Bob Condon on Scalable NFS Servers elsewhere in this newsletter.

By taking a balanced look at the solution architecture, the user can invest in such a way that long term costs are minimized and research work is not compromised by one-dimensional views.

NodeWatch™ Enhancements

By Michael Jacknis, Electrical/Embedded Engineer, Microway

Microway's NodeWatch features the ability to monitor voltages, temperatures and fan speeds within each node of a Microway cluster. The data is fed to a web-based display and to the Ganglia system performance logging system, allowing trends to be tracked and graphed over time. This enables the system administrator to chart environmental degradation and anticipate failures before they occur. In addition to highlighting out-of-band measurements, Microway's NodeWatch web-based user interface also allows the administrator to execute commands on any or all nodes of the cluster, and to shutdown, reboot, power on or off, or hard reset any node. The power and reset functions are equivalent to actually pressing the front-panel switches, a feature available elsewhere only on much more expensive retrofit monitoring systems.

Major enhancements to the NodeWatch software are currently in final test. The new version sends email to the system administrator if data is consistently seen to be out-of-range. The new software has an additional set of administrator established limits, which, if exceeded, cause the affected machines to be automatically shut down and powered off, greatly reducing the chance of physical damage. The algorithm for emailing a warning or shutting down a node includes extensive re-checks of the measurement to be sure it is valid. Considerable work has been done to avoid false positives, without significantly increasing the chance for a false negative.

Additional enhancements include user-adjustable configuration files listing all nodes in the cluster and the failsafe limits associated with them. When customers upgrade their NodeWatch-enabled cluster with additional nodes, the software will configure easily to include those new nodes in the monitoring cycle.

HPC clusters are typically operated unattended in close quarters. Unlike your home or office desktop PC, a failure in a cluster can go unnoticed, compounding the issue. NodeWatch is the inexpensive solution for automatic monitoring, failsafe warning and shutdown of failing nodes.

Cluster News

The University of Arizona has purchased three clusters built with dual Xeon processors and Microway MCMS. One cluster will be used to perform calculations on the photophysics and device physics of organic conjugated polymers.

The second cluster will be used for the solution of two fundamental problems in modern astrophysics. The first is related to the rapid flickering of black holes and neutron stars in our galaxy and requires the solution of the time-dependent three-dimensional equations of magneto-hydrodynamics. The second problem is related to the structure of rapidly rotating neutron stars in general relativity and the development of modern test of theories of gravity.

The third cluster will be used to run high-performance computations on the mathematical physics of precipitative pattern formation, and for the analysis of experimental data.

“We acquired our three clusters from Microway as they offered an aggressive proposal with a very attractive price. Their technical support team has been very helpful in making sure the clusters are ready to run when they arrive in Tucson. They have paid attention to the small details that if overlooked can cause a lot of frustration on the user’s part,” commented Mike Eklund, Computer Systems Manager, Physics Department, at the University of Arizona.

You Can
Count On It
by Stephen
Fried, CTO

Cluster Robustness, Total Cost of Ownership and the Price of Nails

At a time when HPC users are becoming increasingly concerned with total cost of ownership, Microway sees more and more inferior equipment appearing in the market. The total savings from purchasing this “stuff” are much less than customers imagine! MIT recently did a study that showed that 85% of the cost associated with fifty of their clusters was incurred after the purchase and installation. In other words, the actual cost to run a cluster is roughly six times that of its initial purchase price. From our experience competing against vendors selling equipment with marginal components, we have concluded that the difference in acquisition cost between a robust cluster and one that is built from low-end products is typically about 10%. In essence, the gamble is that a 2% reduction in TCO will not measurably impact available compute cycles. In choosing systems so closely margined, in all likelihood, users will lose this bet because they will never get 100% of their nodes up and running at the same time!

Why is this so? A large percentage of the subassemblies in equipment all HPC vendors sell are built in Taiwan and China. From those same locations, vendors can source PC class material as well. For a portion of the PC market, high reliability is not a major issue. Unlike HPC clusters, PCs don’t have to run 24/7 and stay up for months at a time without crashing. An occasional crash, say once every three weeks on a PC, is hardly an earth-shattering event. However, in a 128 node cluster, a crash rate per node of once every three weeks will basically keep the cluster from running, as it will result in node failures almost hourly.

There are several “corner cutting” techniques with hardware that are undesirable for HPC applications. For example, a typical Xeon cluster has anywhere from six to ten fans. It turns out that the cost of fans can vary by about a factor of two, along with their throughput and expected life. Use the cheapest fans, and you “save” \$60 per node! However, failure of an

inexpensive fan can lead to the catastrophic loss of a cluster node. The same concept also applies to the quality and number of capacitors used on motherboard power converters and power supplies. Altogether, a typical node has about 10 DC to DC power converters, including those in the power supply. At the lowest voltage, which turns out to be the 1.2V rail which drives the CPU, the power requirements are over 100 Watts, which is to say the current density on the motherboard will be over 100 Amps! Consistently supporting this much current requires high quality components. The same holds true for PCB material (the better material is not as brittle and can endure more temperature cycles). In short, if jobs are to run reliably, hardware in HPC cluster nodes must be much more robust than in a typical PC.

In addition to hardware, the software is just as crucial. Simple mistakes, like using the wrong device driver or Linux kernel, can result in failures that appear to be hardware related, but are actually software driven. So, along with high quality hardware, it is very important to run validated operating systems. The only way you can really be sure an OS validates is to run it for a week or more on a reasonably large cluster, with MPI applications and validation suites which stress both the hardware and software. It takes competently configured hardware and software to produce a robust production quality cluster. Microway has made significant investments in this area.

The bottom line is that on an HPC computational node that has close to 2,000 parts, there are nearly 2,000 ways to cut corners. Moreover, it only takes one of these 2,000 parts to end up bringing down a node – “for the price of a nail, the shoe was lost...”

Our primary responsibility as your vendor is to make sure that only the highest quality nails make it into a Microway Cluster.

Parallel
Thoughts by
Bob Condon,
Storage
Specialist

Storage Considerations for HPC Cluster Design

File Systems, NAS and SAN Background

Network File System (NFS) is a common protocol for accessing files and sharing storage using an IP network. NFS was popularized when packaged into a ‘turn-key’ storage solution by Network Appliance in the early 1990’s. Network Appliance and others coined the term, Network Attached Storage (NAS), which is used to refer to a dedicated packaged appliance that serves files using protocols such as NFS, CIFS or HTTP.

Storage Area Network (SAN) is a block level storage pool that uses a Fibre Channel protocol and switched technology referred to as a fabric. SANs have become popular as a way to pool heterogeneous RAID arrays, which can be used by heterogeneous systems. SANs have proven to be effective in commercial applications with large servers. The server itself is connected to a SAN switch using a Host Bus Adapter (HBA) connected via the server PCI bus. HBA’s can be fairly expensive (up to \$1,000).

Application for HPC Cluster Users

SANs offer the advantage of scalability in capacity and bandwidth. However, without the use of a virtualization front-end or Cluster File System (CFS) a SAN must be partitioned into local disks assigned to specific servers. In some respects, this defeats one advantage of shared

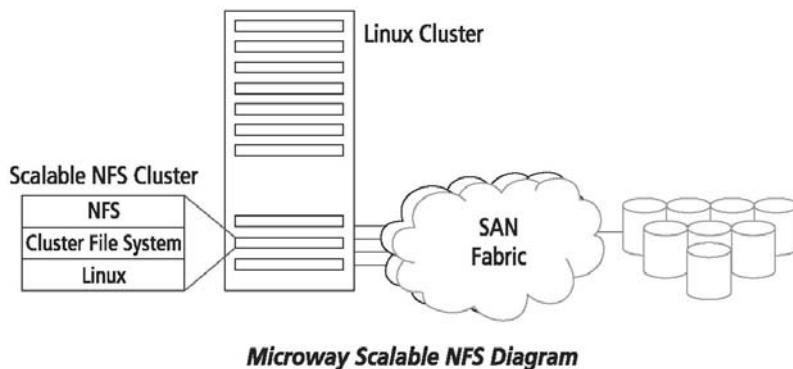
storage across multiple cluster nodes. The combination of a dedicated storage partition and the cost of an HBA make a SAN impractical to support many nodes in an HPC cluster.

On the other hand, NFS running on NAS offers significant benefits versus direct attached SCSI storage. In direct attached storage, I/O bottlenecks can occur during the read and write process when using a RAID array on the master node and internal storage in each slave node. For example, a typical BLAST application uses large data sets ranging from 40 to 80 GB's. These must be loaded through the master node to each slave node. However, only a fraction of this data may be used on any given node during computational analysis cycles. This creates unnecessary network loading and performance degradation. By using a diskless cluster where each node would have a direct connection to shared storage through a dedicated NFS server, the volume of data transfer is reduced and a network bottleneck is avoided since each node connects through its own Gigabit Ethernet connection. Unfortunately, most NFS/NAS solutions are stand alone "islands" of storage that do not scale well when adding I/O bandwidth or storage capacity.

Getting the Benefits of SAN and NAS in an HPC Cluster

Combining SAN and NAS creates a more cost effective option for a diskless cluster solution. In this case a dedicated low-cost server acts as a diskless NFS server (instead of a NAS appliance). The dedicated server is connected to a SAN via a HBA. Multiple NFS servers with HBAs can be added to the cluster by using a DNS round robin method of load balancing data queries.

Microway takes this one step further by adding a software product called a Cluster File System (CFS). The CFS adds dynamic file locking and file sharing. The CFS eliminates the need to partition the SAN since files are locked and unlocked as they are accessed by one of the NFS servers. As the ratio of cluster nodes to NFS servers increases, the cost of the HBAs can be amortized across many systems. This arrangement takes advantage of the accessibility and cost effectiveness of NAS with the scalability of SAN.



Microway has launched an initiative to create storage solutions designed to enhance HPC cluster performance and manageability. We use various technologies, as appropriate, to

integrate the right solution for each customer's application. To learn more about your options, please call your Microway account manager or email storage@microway.com.

**Microway
Spotlight**

Microway Spotlight on Stephen Fried, President and CTO

Stephen Fried is a former defense scientist, musician, sailplane pilot, writer and a pioneer in the development of PCs. He holds a Bachelor of Science degree in Physics from Brown University, with advanced work in chemistry at MIT.

For more than ten years he was a researcher in atomic and laser physics at the Avco Everett Research Laboratory. At Avco, he was the co-developer of the HF chemical laser, the chief ingredient of President Reagan's Strategic Defense Initiative (also known as "SDI Star Wars"). Pursuing a hobby, he went on to run an FAA certified flight school and Part 135 Air Charter business, before returning to professional life in 1979 as a chemical physicist.

How was Microway started?

In 1981 I recognized the importance of high-speed numerics for accelerating scientific applications being run on 16-bit PCs. Subsequently, I developed software tools, which made it possible to use an 80-bit Intel 8087 math co-processor on the IBM-PC. From there, my wife Ann and I founded Microway to sell the software and coprocessors. By 1985 Microway was Intel's sixth largest customer in the world in their microprocessor division. One thing that distinguished Microway from other companies early on was the truth of our slogan: "You can talk to us." Even today, I am available, along with other senior Microway technical people, to talk to customers.

How has Microway's focus changed over 21 years?

During the mid-1980's, we developed the first 32-bit compilers to run on the Intel 80386. I then designed and built accelerator boards, before focusing on parallel processing and small Supercomputers based on Transputers, Intel i860's, and more recently Digital's Alpha CPUs. Currently, I am interested in signal integrity, which is required to design high speed motherboards; code generation techniques, required to produce highly optimized code for CISC and RISC processors; and the physics of cooling and its impact on high density cluster design.

What are your interests other than the hottest new HPC technologies?

I play bassoon for the Brockton Symphony Orchestra and am an avid reader on all topics in current events and world history. I spend many hours annually skiing and flying my ASH-26E sailplane in the Green Mountains of Vermont. I also save time for my wife, children and grandchild.

Where do you see the most potential for improvement in high performance computing in the next eighteen months?

These are exciting times in our field. With the advent of very high performance commodity microprocessors and inexpensive large memory subsystems, HPC has come of age. However, these devices will not solely dictate the future direction of HPC technology. While today's processors are very powerful, they are hamstrung by bottlenecks that started to creep into the technology in the last three to five years. And, because they are built from commodity components, many solutions are not nearly as sophisticated or efficient as the dedicated

devices available for parallel processing 15 years ago! Those pioneering devices had built-in memory controllers and DMA engines that made it possible for processors to communicate with each other, without having the data pass through a Northbridge, Southbridge, PCI bus and a PCI interface card. We are starting to see new devices that emulate the very efficient Inmos Transputer. The AMD Opteron and the DEC Alpha EV79 have to compete with new Intel technology which keeps pushing the edges of scalar performance, hitting peak speeds of over 6 gigaflops per CPU, in certain applications. As a result, these newer devices will not get the kind of acceptance that they ought to, even though they hold out the promise of being able to operate at higher efficiencies and might also be easier to develop interface libraries for.

The new technologies contain 64-bit address busses and registers which will make it easier in the future to write large applications and take advantage of features like RDMA (Remote Direct Memory Access). RDMA is an especially interesting approach to reducing latency between processors. Using this technique, a region of each processor's address space is made available to the other processors in the system. Software will need to be written which maps the logical address space of an application to the global address space of the interconnect being used to connect processors. With this accomplished, it will become possible to directly transfer data between the memory subsystems of processors. One name for this technique is "distributed shared memory parallel processing." Microway designed and manufactured an i860 card in 1992 that featured DSM and also had inter-processor transfer latencies that were lower than those being achieved using interconnects like Myrinet and InfiniBand. These are the same sorts of latencies we see today from the Opteron processor which transfers data between local CPU's in less than 140 ns, a factor of 35 faster than the lowest latency inter-processor connections available today. However, for these low latency devices to really hit full speed, they need to be unencumbered by things like an MPI layer. The correct approach in the future will be non-coherent Shared Memory (SHMEM). This methodology will also require 64-bits.

Imagine you are a user writing code for a cluster, each node of which has 2 Gigabytes of local data allocated to the storage of large arrays. In a 128 node system, that amounts to 256 Gigabytes of data. This is 64 times as much memory as can easily be accessed (using the flat memory model, which is the only one available with today's compilers) with a processor like the Xeon. The bottom line is that to take proper advantage of the future low latency interconnects that feature RDMA, it will pay to own 64-bit processors. For today's users, who are working with very large arrays, the choice will be either AMD Opteron or Intel Itanium2. The Itanium2 holds the edge today in performance, but when the Opteron finally gets all of its ducks in order, its low latency interconnects and architecture should end up giving Intel a run for the money and help to set the stage for the next revolution in HPC.

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