MECCAH: The Mathematics Extreme Computation Cluster At Harvard

William Stein

August 2002

In Spring 2002 I assembled and configured MECCAH, which is a rack-mounted cluster of six fast computers for use by mathematicians who are doing demanding computational work. This article is about my experience building and maintaining MECCAH. It should be of use to anyone considering undertaking or funding a similar project at their institution.

As a graduate student at Berkeley and a faculty member at Harvard, the computational resources available to me at my host universities consisted of scattered Sun workstations running at about one-fourth the raw speed of the current Pentium processors. These machines spent much of their time running Netscape and texing documents, so they were not suitable for demanding computations that could easily use all available resources. Sure, at each institution a senior faculty member had a powerful computer (McMullen at Berkeley, Elkies at Harvard), but that was for his own personal use.

In 2001, the Harvard sysadmin, Arthur Gaer, mentioned that the department was tentatively considering spending several tens of thousands of dollars (that they didn't yet have) on a single multi-processor Sun workstation to support computation-intensive work. My opinion was that such a workstation would be solid but hardly useful; the raw computational power would scarcely touch what two cheap Intel-based Linux boxes could do, though the Linux boxes would likely be less reliable.

I decided to build a cluster of dual processor machines running Linux. I did research and discussed possible configurations with Berkeley grad student Wayne Whitney and a Harvard undergrad named Alex Healy, and requested money. Finally, I secured a grant of $6000 from Harvard, and Harvard alumnus William Randolph Hearst III gave me an additional $14000, which made the budget $20000.

I decided to assemble an Athlon-based system. The Athlon 2000MP is a multi-processor-ready Pentium-like CPU that Athlon claims has performance that is similar to a 2GHz Pentium IV. I selected the Athlon 2000MP processor in March because it was the fastest available budget-priced multiprocessor capable CPU on the market. Intel's only fast multi-processor capable CPU was the Xeon, which was then much more expensive (the Xeon might be a good choice today). Six months later, Athlon has just announced the 2200MP, so I don't feel like Athlon 2000MPs are out of date.

In February 2002, I ordered six custom-built Athlon 2000MP machines in 2U-sized rack-mount cases from `www.pcsforeveryone.com` which is a local Cambridge "chop shop". They ordered the parts I wanted, assembled them, tested them, found surprisingly often that they were defective, got replacements, and finally delivered the individual computers. I still have occasional hardware reliability problems with one of the nodes, even after returning it for service under warranty, and it is currently off (a CPU fan had failed, so they replaced the CPU fan, but not the CPU, which is a cheap "solution" that didn't work).

Unwrapping the rack and putting the computers in it took Alex Healy a full afternoon. Once assembled, I had to keep the machine in my office, because the math department's server closet was tiny and currently full of equipment. It would be several months until we made room in the server closet for the cluster. In the meantime, I kept a rack of noisy and hot computers running in my office. When students came to see me during office hours, they had to shout over the 30 cooling fans in MECCAH.

And, the fuses kept blowing! My neighbor's office is on the same circuit as mine and when he returned from vacation and turned his computer on, the circuit breaker blew, so I had to call the electricians out to switch it back. I moved back to running only four machines, then once increased to five, again blowing the circuit.

MECCAH's operating system is Redhat 7.2 with Linux kernel version 2.4.16 on all six nodes. MECCAH also uses openMosix, which makes the rack of six computers appear to the user as a single computer with 12 processors and 13GB memory (though a single process should not use more memory than on any node). Under openMosix, jobs are automatically migrated from one node to another to dynamically balance the overall system load. Users only have accounts and login privileges for the master node, and never worry about logging into other nodes. I also configured MECCAH to use the ext3 journaled filing system, so, e.g., I can pull the plug from the wall, plug it back in, and have MECCAH back up in five minutes with absolutely no file system corruption.

For computations, people mainly use MAGMA, PARI, Python, C++, and Mathematica. Though Harvard has a Mathematica site license, I HATE administering Mathematica because the licenses regularly expire and limit the number of copies of Mathematica that can be run at once (there should be a way around the latter problem). MAGMA for Linux, on the other hand, requires no license and is free to me because I'm a MAGMA developer. Evidently, Maple is expensive, so we have only a limited Sun license for Maple in the math department.

Here is how I organize computation of a basis for the space of modular forms with level N and weight 2 for N between 1 and 1000. I run 12 jobs simultaneously that each look to see the next level that hasn't been computed, compute that level, and save the result. If it took 1 day to do this computation on my 1Ghz Pentium III last year, it will take only 1 hour to do it on MECCAH. When I am in the throws of a big computation, having this kind of computational resource available to me is extremely exciting. Instead of waiting 1 day, I wait only an hour to generate more than enough data to stimulate theorem proving!

I've given MECCAH accounts to nearly 80 mathematicians all over the world. Abuse of the system by users is rare but not unheard of. Somewhat surprisingly, the usage pattern comes in bursts. There are almost always at least two or three jobs running, but every so often many mathematicians simultaneously become inspired to run lots of computations all at once.

I am the only systems administrator of MECCAH, and I typically spend under five hours a week on administrative responsibilities. I still haven't upgraded the Linux kernel or openMosix since March, but I probably should since there have been a few unexplained problems that might be fixed by a Linux and openMosix upgrade. I use a 30GB Onstream ADRx2 tape drive to make regular backups.

If I were to build a similar cluster from scratch again, I would probably buy more expensive and better warrantied pre-configured dual-processor rack mount nodes instead of custom designing the nodes myself. I definitely would not have kept the computer in my office. When first designing MECCAH, I thought long about whether or not to stack a bunch of conventional cases on shelves or to buy a rack and rack-mount cases. A rack costs nearly $1000 and rack-mount cases cost more than double what ordinary cases cost. In retrospect, it would have been madness to buy conventional cases and shelves, because I've had to move the cluster around many times, and it barely fits in the tiny server closet. The $1500 premium for a rack-mounted system was well worth it. I also deliberated between a fancy serial console or a KVM (keyboard, video, mouse) switch; I went with the $500 KVM, which turned out to be an excellent choice.

The six nodes are networked via a switched 100Mbps ethernet network. I wish the network were faster, because it takes a few minutes to transfer 1 GB from one computer to another. Since user programs migrate between machines and frequently do use in excess of 1GB memory, this transfer time is significant. I purchased 100Mbps ethernet instead of 1Gbps ethernet, because I read that 1Gbps ethernet with Linux is not very reliable, and there can be significant latency problems. Since I didn't have the resources to experiment with many configurations, I opted for 100Mbps, which is very easy.