



Virginia Polytechnic Institute and State University

Crib Notes

Challenges

- Build a world-class supercomputer in less than three months
- Construct a system using only nonproprietary or “off-the-shelf” components
- Complete project in time for fall production runs, place on 2003 “Top 500” supercomputer ranking list

Solution

- 1100 dual processor Power Mac G5 computers at 2.0GHz
- Mac OS X
- Primary interconnect: InfiniBand* communications fabric over Mellanox host channel adapters and switches
- Secondary interconnect: Gigabit Ethernet over Cisco switches
- Custom-designed racks and new cooling system from Liebert Corporation
- Déjà vu fault-tolerant checkpoint/restart and migration software
- Wide variety of open source software

Benefits









- Supercomputer was fully functional in less than three months, and cost one-tenth as much as similar systems
- As the world’s third-fastest supercomputer, it provides massive computational power to scientists at Virginia Tech
- University now eligible to compete for a portion of \$1 billion in National Science Foundation Cyberinfrastructure funding
- Implementation further solidifies position of Virginia Tech’s world-class computer science program

Blacksburg, VA—Dr. Srinidhi Varadarajan knew that he wanted to build a world-class supercomputer. Also, he wanted to solidify the position of Virginia Tech’s world-class computer science program. But since Varadarajan had only a fraction of most supercomputing budgets to spend, it seemed like a hopeless dream. He crunched numbers, solicited every likely vendor, examined and ultimately discarded all possible options using other platforms and chips. Then in June 2003, news of Apple’s Power Mac G5 hit the airwaves. At last, Varadarajan realized, he’d have as much 64-bit processing as he needed to power his dream, without overtaxing his budget.

Soon after the Power Mac G5 announcement, Varadarajan took delivery of his very first PowerBook laptop running Mac OS X. Within days, he placed an order for the 1100 dual processor, 2.0GHz Power Mac G5 computers that now drive Virginia Tech’s new supercomputer. Smart choice: In November 2003, the giant system—named System X—became the third-fastest supercomputer in the world.

System X is radically different from traditional, high-performance supercomputers. Unlike most, it is based on a “supercluster” of Power Mac G5 computers, each of which has 4GB of main memory and 160GB of serial ATA storage. Not only is System X the world’s fastest, most powerful “home-built” supercomputer, it quite possibly has the best price/performance ratio of any supercomputer on the “TOP500 List” compiled by researchers at the University of Mannheim, Germany, Lawrence Berkeley National Laboratory, and the University of Tennessee.

Contents

 The People	Page 2
 The Science	Page 3
 Architecture	Page 3
 Logistics	Page 4
 Interconnects	Page 5
 Cooling Infrastructure	Page 6
 Scheduling	Page 7
 Optimization	Page 7
 The Future	Page 8

Profile in Success

Virginia Polytechnic Institute and State University



Advice to Other Universities

- Do your homework, and engage your vendors as early as you can.
- Solicit support from the uppermost levels of your administration as soon as possible—they'll be critical to your success.
- Recruit the most talented team of people that you can; keep them as informed and motivated as possible.
- Communicate! Ensure that you keep your higher-ups informed as to your progress; keep peers and other team members apprised of milestones met and goals still outstanding (but don't overload them with minutiae).

"I'd never used a Mac before in my life. Within three days I'd made up my mind that Mac OS X and the new Power Mac G5 were the tools I wanted to use."

Dr. Srinidhi Varadarajan, Associate Professor of Computer Science and Terascale Computing Center Director, Virginia Tech

The People

It Takes a Team

Dr. Varadarajan, assistant professor of computer science; Jason Lockhart, director of the College of Engineering's High-Performance Computing and Technology Innovation group at Virginia Tech; and Kevin Shinpaugh, director of Cluster Computing for the university, headed up the massive international effort to develop the cluster. In less than three months, Varadarajan, Lockhart, Shinpaugh, and hundreds of associates, advisers, vendors, and volunteers took the huge system from plans to prototype to implementation at approximately one-tenth the cost of similar high-performance machines.

Lockhart says that Virginia Tech would have been hard-pressed to even begin the supercomputer project without the eager participation of countless volunteers around the globe. From research scientists to consultants to students to carpenters, the project brought together an eclectic mix of skill sets. According to Lockhart, everyone shared the same zeal to break entirely new ground.

"It was amazing to see how folks rallied around this project," he says. "We had support from the highest levels of our administration, to the guys who built out the facility, and everyone in between. Once we explained what the project was, and what it would do for the university, everyone's spirits immediately lifted and they gave us higher than normal-quality work.

"When we finished the project, we gave out awards to everyone," Lockhart continues. "When the construction guys heard that we were ranked number three in the world, they all let out a huge roar. It was definitely a team effort, and it was fantastic to see it all come together."

A Quantum Leap

According to Varadarajan, now director of Virginia Tech's [Terascale Computing Facility](#) (where the supercomputer is housed), a fortunate confluence of events made the groundbreaking project possible. Hassan Aref, a strong advocate of constructing the supercomputer, had just assumed his post as dean of the College of Engineering. A new program—called Computational Sciences and Engineering (CSE)—made a talented pool of faculty and student computer experts available. But it was Varadarajan's first PowerBook, and the announcement of the new Power Mac G5, that really got the project rolling.

"I'd never used a Mac before in my life," admits Varadarajan. "I was a Linux/UNIX kind of guy, and used Windows to run the Office productivity suite. But I [bought] a 17-inch PowerBook laptop, and found that it was quite easy to do everything I'd been doing under the other platforms. I could compile the software application I'd been developing; it just worked under Mac OS X. Plus I could use the productivity applications I needed, without having to reboot or launch an emulator. Within three days I'd made up my mind that Mac OS X and the new Power Mac G5 were the tools I wanted to use."



“I realized that the PowerPC 970 processor would be ideal for us. Its fused multiply-add operation gives it a floating-point performance equal to—if not better than—Intel’s Itanium2 solution.”

Dr. Srinidhi Varadarajan

The Science

Harnessing “Big Science”

The supercomputer, which was constructed exclusively of off-the-shelf technologies from Apple and other partners, will significantly enhance the university’s research capabilities. “Virginia Tech’s idea was to develop a supercomputer of national prominence, based on a homegrown cluster,” says Aref, who is also a former chief scientist at the San Diego Supercomputer Center. Adds Varadarajan, “Expect to see a lot more Power Mac G5 clusters in the future.”

Supercomputing, or terascale computing, is mandated by the need to solve problems too large for an individual computer. The need for this computing power arises often in almost every field of study that uses quantitative methods.

In the past, these fields have relied on a combination of theory and experiment. But now computational science allows researchers to simulate the behavior of natural or human-engineered systems, instead of merely observing a system or building a physical model of it. This powerful approach enables glimpses into places that are too small, too large, too dangerous, too short-lived, or too long-lasting for typical experiments.

Obtaining information about the galaxy, global warming, computational biology, drug design, and homeland security are just a few of the grand challenges that computational scientists can address through supercomputing. The Virginia Tech supercomputer will also explore such topics as nanoscale electronics, quantum chemistry, computational chemistry and biochemistry, and aerodynamics through multidisciplinary design optimization, molecular statics, cell cycle modeling, and computational acoustics. Enabled in part by the versatility of the G5 processor, the cluster is truly an interdisciplinary tool.

Architecture

The Power Mac G5: A New Era in Supercomputing

Virginia Tech had already built a number of high-powered clusters, although nothing of this scale. Then, when the Power Mac G5 was introduced in June 2003—boasting dual 2GHz G5 processors and up to 8GB of main system memory—the landscape of supercomputing changed for the better. It wasn’t long before Virginia Tech had done the math: The Power Mac G5 provided higher performance at a lower price than any other system on the market.

“We already had discussions in progress with another hardware provider,” recalls Lockhart. “But obviously the IBM PowerPC 970 processor, with its supercomputing pedigree, was really appealing to us. We’d been tracking the ‘buzz’ for several months, and we were very anxious to see if the Power Mac G5 would give us another platform option. When we heard the announcement in June, we basically wrapped up our other discussions and went down the new path with Apple.”

Record-Breaking Processing Speed

Lockhart says several features of the new Power Mac G5 computers were especially attractive to Virginia Tech. “AMD’s Opteron can execute only two

“This project never would have been possible at this price, while getting this performance, with any other [platform].”

Jason Lockhart, Director of the College of Engineering's High-Performance Computing and Technology Innovation Group, Virginia Tech

double-precision floating-point instructions per clock cycle, which kind of limits system performance,” he notes. “The G5 processor has two floating-point units. Thus, unlike the Opteron, it can perform fused multiply-adds, giving four operations per clock cycle and achieving a theoretical limit of 8 gigaflops from a 2GHz processor. All this is in addition to the G5 processor’s Velocity Engine floating-point and integer units.”

Says Lockhart, “We would have had to buy more nodes to equal the amount of performance needed to hit our 10-teraflop goal. But the hypertransport memory architecture of the new Macs gives us lots of [speed]. And having the PCI-X bus in the system gives us the bandwidth we need for our communications fabric.”

Adds Varadarajan, “When we were evaluating machines and platforms, I realized that the PowerPC 970 processor would be ideal for us. Its fused multiply-add operation gives it a floating-point performance equal to—if not better than—Intel’s Itanium2 solution. Pretty quickly, I knew that the Power Mac G5 machines would help us reach the goals we had in mind for our supercomputer.”

The PowerPC G5 gets its smarts from the [execution core](#) of IBM’s 64-bit POWER4 processor—recipient of *Microprocessor Report’s* 2001 Analyst’s Choice Award for Best Workstation/Server Processor. Apple collaborated with IBM to leverage this industry-leading design for the Power Mac G5, combining an optimized Velocity Engine with a new superscalar, superpipelined execution core that supports more than 200 simultaneous in-flight instructions.

For more information on the G5 processor, visit [Apple’s G5 processor page](#).

Logistics

Priced Right, Right Off the Shelf

In addition to its processing speed, the new Mac computers met two other crucial criteria: pricing and availability. For Virginia Tech, it was important to create the new supercomputer exclusively with off-the-shelf components. And, as with any project in higher education, budget was a concern. The Power Mac G5 computers ably addressed both considerations.

Says Lockhart, “In addition to our requirements to have a 64-bit, high-bandwidth memory and communications platform, we were committed to using [off-the-shelf] components from start to finish. We knew that a system of this size will have failures, and we wanted to minimize the expense of replacing failed components.

“Also, while a similar system built in Japan cost \$350 to \$400 million for the machine alone, and other monies for facilities and personnel ran their total up to \$1 billion, we had only \$4 to \$5 million to spend,” Lockhart adds. “The Madison generation of Intel’s Itanium2 chips was really expensive—just one processor was as much as \$7500! But by going with the Power Mac G5 computers—because they were cost-effective, they needed minimal tweaking (aside from adding some additional RAM), and they would scale really well—we were able to hit our budget and meet all of our performance goals.”

A Safe “Port” in the Storm

Vu Pham, Dror Goldenberg, and Edward Bortnikov, software architects at Mellanox, were charged with porting the InfiniBand driver to Mac OS X. Since none of them had ever booted up a Mac before, let alone do any development on the Mac platform, all say their assignment was very ambitious. Factor in the need to complete the migration of a very sophisticated application to the new platform in little more than a month, and all players on the team were collectively holding their breath.

“The InfiniBand driver is a very nontrivial product—the latest specification is more than 2000 pages long,” Goldenberg explains. “Because of the robust capabilities of InfiniBand, it’s much more complex than standard network drivers. Fortunately, our software is architected in a modular way that requires only reimplementing the library and kernel extensions that provide the OS-specific services for each new platform. Plus the assistance we received from Apple was invaluable in helping us meet our deadline.”

Having visited Cupertino earlier, Goldenberg and Bortnikov returned to Israel with their plan in mind and completed their initial development on Power Mac G4 servers running Mac OS X. This plan required a development model that was strictly “follow the sun”: As one team finished a long day’s work, they would hand off their code to the next team to keep the project moving forward. When the coding was completed and the initial tests were passed, the code was ready to be installed and tested on the Power Mac G5 systems.

Just-in-Time Transition

A week before the software went into production, the Mellanox team received their first Power Mac G5 computers. From then on, says Goldenberg, the transition was “very smooth and painless.” The development effort also included the porting of Mellanox’s comprehensive testing and benchmarking suite to the Mac OS X platform, ensuring the stabilization of the InfiniBand driver.

(cont. on page 6)

No Time to Spare

Developing a supercomputer is a colossal undertaking under any circumstances, and normally requires several years. For Virginia Tech, the goal seemed almost impossible: The system had to be complete and fully functional before October 1, 2003—less than three months from the project’s start date. At that time, researchers from Germany’s University of Mannheim, the University of Tennessee, and Lawrence Berkeley National Laboratory would publish their semiannual [“TOP500 List” of supercomputers](#), ranking them by processing speed.

Winning a berth near the top of the list would qualify Virginia Tech to compete for funding from the National Science Foundation (NSF), through its [Cyberinfrastructure program](#). With a proposed budget of \$1 billion, the program aims to develop and support an enhanced cyberinfrastructure for science and engineering research and education.

“The Cyberinfrastructure program has been ramping up with what they call the ‘Extensible Terascale Facility,’ or the distributed terascale facility,” explains Lockhart. “Basically, it’s a large-scale computational network that’s being established across the United States, between large supercomputer centers. The NSF has been supplying funding to national labs and supercomputer centers to develop resources based on processing speeds of one trillion floating-point operations per second, or FLOPS.

“One of our original goals,” continues Lockhart, “was to position the university to compete in a much larger research arena. We felt that the combination of the Cyberinfrastructure funding and the supercomputing technology would enable us to bring in the ‘big science’ types of research projects.”

[Apple’s Developer Relations](#) helped with the hardware and software development. With this assistance from Apple, as well as from other volunteers at Virginia Tech, the cluster surpassed the 10-teraflop mark in time to qualify as the third fastest in the world.



Interconnects

Technology Partners Make the Vision a Reality

With this many computers, how can they all communicate effectively? The sheer volume of traffic among 1100 computers is immense, and in a supercomputer, information has to travel quickly. Virginia Tech turned to cutting-edge technology from Mellanox Technologies for this critically important component.

From Israel to Cupertino

Israel-based Mellanox Technologies signed on to provide InfiniBand technology for the cluster as the primary communications fabric, as well as all necessary drivers, cards, and switches for the project. Since Mellanox had not yet ported their driver to the Mac OS X platform, that lent still more urgency to the project. But three engineers from Mellanox—Vu Pham from Mellanox U.S. and Dror Goldenberg and Edward Bortnikov from Mellanox’s Israel headquarters—set up camp in Apple’s Cupertino, California, labs and went to work.

“Both Apple and Mellanox had identified the project as a strategic one,” says Goldenberg, software architect at Mellanox. “We were pleasantly surprised by the abundance of online documentation, and the willingness of Apple’s

(cont. from page 5)

While Bortnikov, Pham, and Goldenberg were busy with the driver porting project, Varadarajan was porting the Message Passing Interface (MPI) stack and the benchmarks to Mac OS X. Combining the new driver with MPI and the benchmark worked with little effort, says Bortnikov. But he, Pham, and Goldenberg still had one last, nontrivial issue to deal with: scaling the application to 1100 nodes and delivering 10-teraflop performance in two weeks.

“The Terascale cluster was the most significant scalability trial to date for both our hardware and Apple’s software—previously, our largest cluster was 256 nodes,” observes Bortnikov. “It was a tremendous logistical and human effort to assemble the 1100-computer cluster. But with unlimited 24-hour pizza, soda, and student volunteers from Virginia Tech, we did it!

“Then,” he continues, “working around the clock at the Virginia Tech Terascale Computing Facility with our FAE (Field Application Engineer) team of Gene Crossley and Jeff Kirk leading the way, we managed to stabilize the network, the MPI port, and the benchmarks, and make everything work together. The driver code was quite solid from the day it was ported to Mac OS X. All in all, we were very pleased to work on this software development project with Apple and Virginia Tech.”

Developer Technical Support (DTS) team to help us accomplish our aggressive goals. In less than two weeks, we had a solution outline.

“However,” Goldenberg continues, “it was clear that there was no chance to succeed unless we worked face-to-face with Apple developers. Therefore, we headed to Apple headquarters in Cupertino, where we spent two weeks studying the new platform, working on the Power Mac G5 prototypes, consulting with the DTS team, and writing sample drivers. Then we moved to the real code. The effort was very fruitful, and we all were very excited when the 1100-node cluster was ranked as the third most powerful supercomputer in the world.”

Turbo-Powered Communications

InfiniBand is now available on Mac OS X for any user who wishes to use it in a high-performance environment.

“By adding an InfiniBand card to each Power Mac G5,” says Lockhart, “we get a 10-gigabit-per-second communications fabric. It’s full duplex and capable of ensuring some very low latencies in transmitting data—on the order of 4.5 microseconds for certain packet sizes. The nice things about the InfiniBand solution were that its price point matched other comparable products, it was five times as fast, and it had lower latencies, so that was a great coup for us.”

Five Gigabit Ethernet switches from Cisco were also added to the supercomputer’s infrastructure, providing a secondary communications fabric for management, data transfer, and other less latency-sensitive tasks. With built-in Gigabit Ethernet standard on every system, putting each node on the network was a snap.

Cooling Infrastructure

Liebert Keeps It Cool

With 1100 Power Mac G5 machines—with a total of 2200 G5 processors—running 24/7 in an enclosed space, things can really heat up. But a custom solution from Liebert Corporation keeps the supercomputer cool, says Kevin Shinpaugh, who oversaw all logistics and facilities for the System X implementation.

Working within a seemingly impossible timeframe, Liebert constructed the rows of racks that would support the 36 tons of hardware, and designed a special chilling system to ensure optimum system operation. In all, the configuration takes up 3000 square feet, with six rows of racks holding nearly 200 computers each. Liebert also supplied rack-mounted heat exchangers and a special refrigerant-based cooling system that uses chilled water.

“Liebert was fantastic to work with,” Shinpaugh says. “We were the first installation of ‘XDV’ [extreme-density vertical cooling system]. Liebert really pushed their capabilities to get this stuff to us on time and get it installed. They bent over backwards to help us out, and we’ve been thrilled with the results.”

Chilling Out

Based on the expected heat load for Virginia Tech’s supercomputer, normal air conditioning units were deemed insufficient. But Liebert Corporation, known

“By going with the Power Mac G5 computers, we were able to hit our budget and meet all of our performance goals.”

Jason Lockhart

for its comprehensive range of protection systems for sensitive electronics, provided a high-density, rack-mounted cooling system that met the budget and time constraints of the project. In addition, Liebert designed the computer racks and provided the power distribution equipment.

Liebert’s engineers constructed a system that uses the natural convection that occurs in rack-mounted computers to generate a cooling airflow. Virginia Tech’s Power Mac G5 computers were first set up in a “hot” and “cold” aisle configuration, in which the exhaust fans from the Macs would blow into the “hot” aisles.

The XDV was then installed on top of the rack enclosure. This unit takes the hot air directly from the enclosure and cools the air before it is blown into the “cold” aisles, which the systems all face. If needed, the XDV can draw discharged “hot spot” air from the room through a cooling coil and distribute it to the cold aisle as well.

Scheduling

Déjà vu: No-Fault Performance

Virginia Tech’s new supercomputer showcases the industry’s first solution to the problem of transparent fault tolerance, a decades-old challenge in parallel computing. A single component failure can cripple the completion of the types of jobs that typically run for days, weeks, or even months on large-scale systems. One faulty node, and a project that has been running for two weeks might require a full restart. But thanks to Varadarajan and his new Déjà vu software, Virginia Tech and other supercomputer centers will waste precious processing time far less frequently.

Developed in partnership with the Pittsburgh Supercomputing Center (PSC) and with funding from the NSF, Déjà vu allows IT professionals to set parameters and options to identify various “checkpoints” while a job is running. If any node in the system fails, the software will automatically find another node and restart the job from the last safe, checkpointed state. If needed, the application also allows any jobs currently running to migrate to another resource of similar architecture.

“All of this takes place within milliseconds,” reveals Lockhart. “If a job fails, it will be restarted almost instantaneously. This is huge for large-scale computing— not only do we have this robust fault tolerance built in, we can actually move a job while it’s running! With Déjà vu, developers can integrate the technology directly into their applications at compile time, and deliver uninterrupted processing.”

Optimization

Pulling Out Incredible Efficiencies

Kazushige Goto, a visiting scientist at the University of Texas, Austin, added still more customization to Virginia Tech’s supercomputer. Goto optimized the Basic Linear Algebra Subprograms (BLAS) libraries, a function that Lockhart says was instrumental in achieving the supercomputer’s 10.28-teraflop processing speed benchmark.

Notes Lockhart, “These libraries have to be optimized by hand for each processor architecture at the assembly-code level. It’s usually a very long, tedious process, trying to eke out every gigaflop possible. But Mr. Goto got some early access to the Power Mac G5 systems, and he was able to pull some incredible efficiencies out of them—in the mid-90 percent range—in only five or six weeks. The theoretical limit of a dual processor Power Mac G5 is 16 gigaflops, or 8 gigaflops per processor. He was getting 95 percent efficiency on a single processor, which was amazing for the short amount of time he had.”

Dr. Dhabaleswar Panda of Ohio State University supplied the interface library (called MVAPICH) that enables the InfiniBand support. Since the library was developed under a Linux variant, Varadarajan had the straightforward task of bringing it over to Mac OS X, which is a variant of BSD. “Srinidhi did the porting work himself, with support from Dr. Panda’s grad students,” says Lockhart. “They were very instrumental in helping us pull the library into Mac OS X, so we could create our cluster of machines. Without it, we’d have been hard-pressed to get anywhere with the project.”

Since Apple provides all of its processor optimization tools for free, Virginia Tech was able to quickly access the powerful optimization capabilities built into the PowerPC G5. Collectively named CHUD (Computer Hardware Understanding and Development), these capabilities include profilers, processor simulators, and visualization tools. All are included as part of Apple’s free [Xcode development environment](#).

The Future

Success Breeds Success

Now that the results are in, Virginia Tech is enjoying international recognition for its work in supercomputing. Lockhart says the university is being inundated with requests for access to the system, from some very well-known organizations. Among the long list of those wishing to book time at Virginia Tech: the Department of Energy, NASA, the National Institute of Aerospace, the National Security Agency, and many others. With this type of interest, Virginia Tech’s faculty and administration are confident that the new system will more than pay for itself in the near future.

“Virginia Tech now has one of the top-ranked supercomputing facilities in the world, supporting ‘big science’ research,” confirms Glenda Scales, assistant dean of distance learning and computing in the College of Engineering at Virginia Tech. “It is anticipated that the university will realize at least a five to one return on investment, in terms of annual research grant and contract activity.”

“In the past, we’d typically have gone after smaller pieces of larger programs that our systems could handle,” Lockhart says. “But now that we have an asset like our supercomputer, it’s generating lots of interest. In addition, we believe it will help us attract higher-caliber faculty members and scientists in the future.”

The Clear Choice

Varadarajan, Lockhart, Shinpaugh, Scales, and the hundreds of colleagues who helped with the supercomputer project have every right to be proud of their achievement. All plan to share their successes with other educational institutions in the near future. Dean Aref has announced that Virginia Tech will

publish a “how-to” kit, which details the process from start to finish. Included in the bundle will be the Déjà vu software, as well as information about the lessons learned from the project.

Yet for Lockhart, the key to the group’s success is simple: “I’ve always said that people will have to pry my Mac away from my cold, dead hands,” he laughs. “But it’s true—Apple clearly has the right tools for the job. And I say that as someone who has very broad experience with every platform and operating system out there.

“This project never would have been possible at this price, while getting this performance, with any other [platform],” Lockhart finishes. “So here’s my advice to anyone else who’s thinking about building a supercomputer: Know that the experience will be very taxing, both mentally and physically. But if you make your choices wisely, and engage the vendors who will support you at every step, you’ll find it to be an amazing bonding experience. Everyone should build a supercomputer!”

For More Information

For more information about Apple education products, visit www.apple.com/education on the World Wide Web or call 800-800-2775.

Visit the Macintosh Products Guide at www.apple.com/guide for the latest information on thousands of hardware and software products for your Mac.

* The mention of third-party products is for informational purposes only and represents neither an endorsement nor a recommendation.

© 2004 Apple Computer, Inc. All rights reserved. Apple, the Apple logo, Mac, Macintosh, Mac OS, PowerBook, and Power Mac are trademarks of Apple Computer, Inc., registered in the U.S. and other countries. Velocity Engine and Xcode are trademarks of Apple Computer, Inc. PowerPC is a trademark of International Business Machines Corporation, used under license therefrom.