Green Supercomputing Comes of Age

Wu-chun Feng, Xizhou Feng, and Rong Ge

Energy-efficient (green) supercomputing has traditionally been viewed as passé, even to the point of public ridicule. But today, it's finally coming into vogue. This article describes the authors' view of this evolution.

In 2002 the Japanese Earth Simulator supercomputer shattered US domination of high-end computing (HEC) with a 400 percent increase in computational speed, as measured by the TOP500 List's Linpack benchmark. Soon thereafter, the Council on Competitiveness, sponsored by the US National Science Foundation and the Department of Energy, conducted a study that found 97 percent of businesses could not compete (or exist) without HEC (www.compete.org/pdf/HPC_Users_Survey.pdf). These businesses viewed HEC as essential to economic competitiveness: "to out-compete is to out-compute."

Although HEC might elicit an image of a gargantuan supercomputer that occupies a huge amount of space and consumes an outrageous amount of power to solve large-scale scientific and engineering problems, it is clearly important to the business and enterprise community. But rather than referring to an HEC resource as being a supercomputer, the business and enterprise communities refer to HEC in general as *large-scale data center computing*; in this article, we refer to supercomputing as HEC, large-scale data center computing, or both.

Until recently, such systems enjoyed a "free ride" on institutional infrastructures. However, with the annual costs of both power consumption and cooling projected to exceed annual server spending in data centers in 2007, institutions with data center supercomputers have

 Published by the IEEE Computer Society 1520-9209/08/$25.00©2008 IEEE

become particularly sensitive to energy efficiency—so-called "green" issues. We explore those issues in this article and describe how the supercomputing industry has evolved from viewing power and cooling as a secondary concern to a primary design constraint.

The Problem with Supercomputing

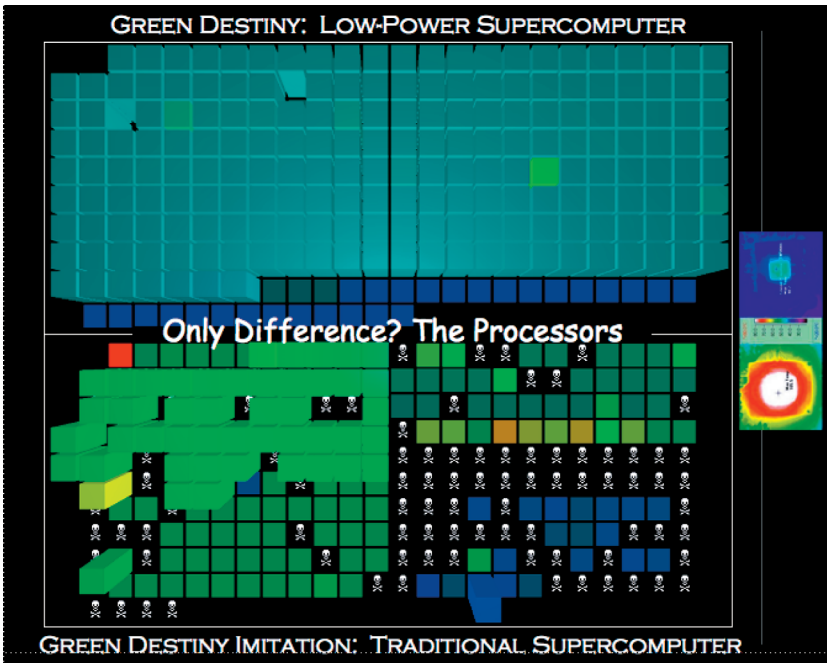
Although supercomputers provide an unparalleled level of computational horsepower for solving challenging problems across a wide spectrum of fields—from scientific inquiry, engineering design, and financial analysis to national defense and disaster prediction—such horsepower usually comes at the expense of enormous power consumption, not only to run the supercomputer but also to cool it. This, in turn, results in extremely large electricity bills and reduced system reliability.[1,2] Accordingly, the HEC research community started exploring green supercomputing as a way to achieve autonomic energy and power savings with little to no impact on performance. However, the notion of green supercomputing is still viewed as an oxymoron: a supercomputer summons up images of speed, a Formula One race car of computing, whereas green or energy-efficiency computing evokes images of the more practical Toyota Prius.

For decades, the supercomputing community has focused on performance and occasionally price/performance, where performance is defined as speed. Examples include the TOP500 list of the world's fastest supercomputers (www.top500.org), which calculates the speed metric as floating-point operations per second (flops), and the annual Gordon Bell Awards for Performance and Price/Performance at the Supercomputing Conference (www.sc-conference.org). As with computers in general, supercomputers' raw speed has increased tremendously over the past decade—for instance, the top system's speed on the TOP500 list has grown from 59.7 Gflops in 1993 to 280.6 Tflops in 2007, roughly a 4,690-fold speedup. We can attribute such huge performance jumps to increases in three different dimensions: the number of transistors per processor, each processor's operating frequency, and the system's number of processors. Over the 1993 to 2007 period, we've observed more than a 100-fold increase along each of these dimensions.

However, this focus on performance, as defined by speed, has let other evaluation metrics go unchecked. As we mentioned earlier, supercomputers consume a huge amount of electrical power and generate a tremendous amount of heat. Consequently, keeping a large-scale supercomputer reliably functioning requires continuous cooling in a large machine room, resulting in substantial operating costs. As a current rule of thumb, 1 megawatt (MW) of power consumed by a supercomputer today typically requires another 0.7 MW of cooling to offset the heat generated—and each megawatt of electric power costs approximately US$1 million per year. The Japanese Earth Simulator, for example, ranked as the top supercomputer on the TOP500 list from 2002 to 2004, consumed 12 MW of power, resulting in US$10 million per year in operating costs just for powering and cooling.

Green supercomputing is still viewed as an oxymoron: a supercomputer summons up images of speed, a Formula One race car of computing, whereas green or energy-efficiency computing evokes images of the more practical Toyota Prius.

Moreover, when supercomputing nodes consume and dissipate more power, they must be spaced out and aggressively cooled; otherwise, the system temperature rapidly increases and results in a higher system failure rate. According to Arrhenius' equation as applied to computer hardware, every 10°C increase in temperature results in a doubling of the system failure rate. Several sets of informal empirical data indirectly support this equation—for example, a 128-node Beowulf supercomputing cluster that resided in a warehouse at Los Alamos National Laboratory failed once per week during the winter months when the temperature inside the warehouse hovered at 21- to 23°C; the same cluster failed twice per week when the temperature reached 30 to 32°C.[1]



1 Green Destiny's reliability versus its imitation. Both systems have the same number of nodes, as represented by each square-shaped area, and run the same applications without any special cooling facilities. The only difference is that Green Destiny uses ultra low-power processors whereas the imitation uses higher-powered "low-power" processors. Green Destiny has no failed nodes, whereas more than one-third of the traditional imitation nodes are dead (shown with skull and crossbones).

The lesson learned is that by keeping the power draw lower, we can lower a supercomputer's system temperature, thus improving system reliability, which, in turn, contributes to better availability and productivity.

The Rise of Energy-Efficient Supercomputing

Green Destiny, the first major instantiation of the Supercomputing in Small Spaces Project at Los Alamos National Laboratory (http://sss.lanl.gov and now at http://sss.cs.vt.edu), was arguably the first supercomputing system built with energy efficiency as its guiding principle.[3–5] Green Destiny was a 240-processor Linux-based cluster with a footprint of only 5 square feet and a power appetite of as little as 3.2 kW when booted diskless. Its extraordinary power efficiency also made it extremely reliable. It ran in a dusty 85°F warehouse at 7,400 feet above sea level with no unscheduled downtime over its lifetime (from 2002 to 2004), and it did so without any cooling, air filtration, or air humidification.

How did Green Destiny achieve such energy efficiency and system reliability? It used low-power components whose performance could be optimized for supercomputing. For example, the Transmeta Crusoe processor used in Green Destiny was a nontraditional and very low-power hardware-software hybrid; by optimizing the software portion of the Crusoe processor for supercomputing, the flops in Green Destiny improved by 50 to 100 percent, and power consumption remained an order of magnitude better than in a traditional processor. With lower power consumption—and thus a lower temperature—the compute nodes were more reliable and could be packed more densely. Figure 1 shows Green Destiny's reliability versus an imitation cluster, which was identical to Green Destiny except for the processor.

Green Destiny heralded a new age in supercomputing, one that focused more on efficiency, reliability, and availability than on just raw speed. As noted earlier, a traditional supercomputer is viewed more like a Formula One race car, which wins the raw performance metric with respect to top-end speed, but the reliability can be so poor that it requires frequent maintenance and results in low throughput. In contrast, a green supercomputer such as Green Destiny is more like a turbocharged Toyota Prius: it loses relative to raw speed, but its better reliability and efficiency results in higher throughput. Table 1 indirectly captures this trade-off between efficiency and raw speed, showing raw configuration and execution numbers for Green Destiny and ASCI White (the top supercomputer on the TOP500 list in 2001) on the Linpack benchmark. As we would expect, the ASCI White supercomputer led all the raw performance categories (shown in italics), whereas Green Destiny won the efficiency categories (shown in bold). Specifically, Green Destiny's memory density, storage density, and computing efficiencies relative to power and space are all one or two orders of magnitude better than ASCI White's.

Media coverage in hundreds of news outlets demonstrated that Green Destiny had captured a tremendous amount of interest in both the com-

puting and business industries. However, when Green Destiny debuted, the supercomputing community thought that being green was synonymous with low performance. So even though Green Destiny ultimately produced 101 Gflops on the Linpack benchmark, which was as fast as a 1,024-processor Cray T3D supercomputer on the TOP500 list at the time, many people still felt that Green Destiny sacrificed too much performance for energy efficiency and system reliability. Consequently, Green Destiny received both ridicule and scorn.

Table 1. Comparison of supercomputing systems on the Linpack benchmark. ASCI White's top performance is shown in italics; Green Destiny's appears in bold.

Performance Metric	ASCI White	Green Destiny
Year	2000	2002
Number of processors	8,192	240
Performance (Gflops)	*7,226*	101
Space (ft^2)	*9,920*	*5*
Power (kW)	*2,000*	*5*
DRAM (Gbytes)	*6,200*	150
Disk (Tbytes)	*160.0*	*48*
DRAM density (Mbytes/ft^2)	625	**30,000**
Disk density (Gbytes/ft^2)	16.1	**960.0**
Perf/space (Gflops/ft^2)	0.7	**20.2**
Perf/space (Gflops/kW)	4	**20**
Reliability (hours)	5.0 hours (2001), 40 hours (2003)	**No unscheduled downtime**

Evolution and Success?

The pendulum eventually shifted, though, and the notion of green supercomputing brought on by Green Destiny bifurcated into two directions: a low-power supercomputing approach that balances performance and power at system integration time and a power-aware supercomputing approach that adapts power to performance needs when the system is running. Both approaches aim to reduce power consumption and improve energy efficiency.

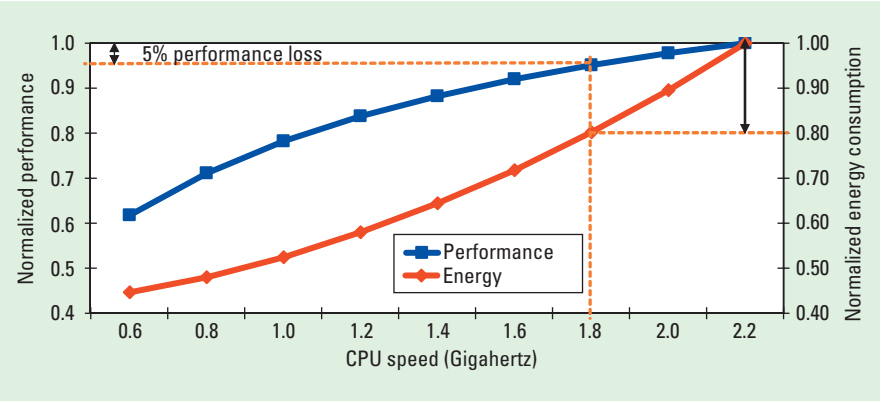
Like Green Destiny, the low-power approach focuses on energy efficiency at system integration time, selects low-power chips as the building blocks for power reduction and system reliability, and achieves better performance by scaling up to a larger number of processors. Since Green Destiny, many projects using low-power approaches have emerged. The MegaScale Computing Project, for example, has the ambitious goal of building future computing systems with more than a million processing elements in total. It aims to simultaneously achieve high performance and low power consumption via high-density packing and low-power processors, with lofty design objectives of one Tflop/rack, 10 kW/rack, and 100 Mflops/W. The first MegaScale prototype, called MegaProto, debuted at Supercomputing 2004 as a 16-processor low-power cluster with dual Gigabit Ethernet for data communication, all in a 1U chassis that consumed only 330 W.[6]

The significant power reduction and efficiency improvements that Green Destiny pioneered have also appeared in commercial products, including a desktop supercomputing system called the Orion Multisystems DT-12. Orion Multisystems aimed to bridge the widening performance gap between PCs and supercomputers with the DT-12 (12-processor desktop supercomputer) and its bigger sibling, the DS-96 (96-processor deskside supercomputer), both of which could be plugged into a standard electrical wall outlet in any office: the former consumed as much power as an overhead light with two 75-W light bulbs, and the latter consumed as much power as a hairdryer.[7]

However, the most prominent architectural approach for low-power supercomputing is the IBM Blue Gene/L (BG/L),[8] which debuted in November 2004 as the fastest supercomputer in the world relative to the Linpack benchmark. BG/L is based on IBM's low-power, embedded processor PowerPC 440 and system-on-chip technology. Each BG/L rack consists of 1,024 processors that collectively consume 28.14 kW and produce 3 Tflops on Linpack. With the IBM BG/L's success, being green finally came into vogue. Another recent development in low-power supercomputing comes from SiCortex, a startup company that produces the SC5832, which is comprised of 972 six-way symmetric multiprocessing nodes connected by a low-latency, high-bandwidth interconnect fabric.

Power-Aware Supercomputing

Although low-power supercomputing has resulted in impressive efficiency and competitive performance, many system researchers argue that it still has two disadvantages. First,



2 Energy–performance trade-off. For certain workloads, we can scale down the processor to save 20 percent in energy costs while losing only 5 percent in performance. This 5 percent loss means that an application that normally takes 57 minutes to run would take one hour but would have a 20 percent reduction in energy consumption, which translates to $2 million in energy savings.

most low-power supercomputing solutions require architectural modifications—for example, the BG/L processor is a stripped-down version of the 700-MHz embedded PowerPC 440 processor, and Green Destiny relies on a high-performance customization of the Crusoe processor's software portion. Second, the tremendous growth in supercomputing performance has largely been spurred by commodity parts designed for PCs and servers. Because current low-power supercomputers don't rely entirely on commodity technology, they aren't as cost-effective, and, consequently, it would be difficult to sustain continuous growth over a long period.

To address these issues, an alternative approach uses high-end but still commodity-based hardware to build a supercomputer and then layers power-aware systems software on top to reduce the hardware's power consumption with minimal impact on performance. As Figure 2 shows, the key idea here is to "green" the compute node. We're building a system with multiple power and performance modes and then dynamically adapting the system's power and performance mode to match its current workload—the aim is to reduce power while maintaining performance. For instance, a processor-bounded workload requires the highest clock frequency (and voltage) to maintain performance, whereas a memory-bound workload can use a lower frequency and voltage for better energy efficiency while still maintaining performance.

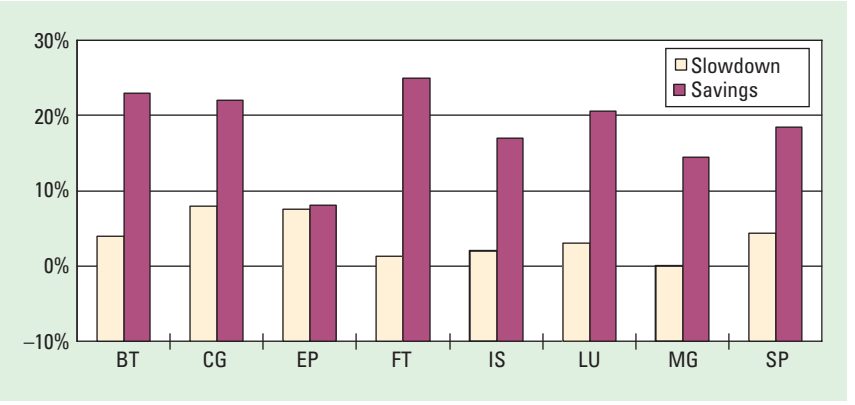
Researchers have implemented many power-aware software prototypes for commodity supercomputing to demonstrate this approach's feasibility.[1,9–11] Most of them start with a cluster of high-performance, high-power processors that support a mechanism called dynamic voltage and frequency scaling (DVFS) and then create a power-aware algorithm (or policy) that conserves power by scaling down the processor's supply voltage and frequency at appropriate times.

Ideally, the appropriate time to scale down the processor's voltage and frequency is whenever the processors is waiting for data from memory accesses or I/O operations because there's no reason for it to "sit and spin its wheels" at the maximum voltage and frequency while doing nothing. However, switching the processor from one voltage and frequency to another at the system level currently takes on the order of milliseconds (that is, millions of clock cycles). This power/performance transition overhead plus the large penalty for false scheduling can significantly reduce performance and consume more energy. The challenge for power-aware algorithms is thus to place processors in low-power mode only when doing so won't reduce performance.

Researchers have studied three distinct power-aware approaches for parallel applications. The first one aims to identify compute nodes and execution phases that aren't on the critical execution path and then scale them down while still meeting the time constraint.[12] However, this approach requires significant effort with respect to performance profiling and program analysis and is often application-dependent; studies have also shown that the benefit here is smaller than what people might expect. The second approach leverages the identification of execution phases in an application and then schedules appropriate voltage and frequency for each phase. The programmer or compiler can insert DVFS control commands into the source code or binary execution file through compiler-di-

rected technology.[10,11] The third approach is similar to the second one, but the DVFS control is implemented as an autonomic, performance-directed, runtime controller independent of the applications running.[1,9] This approach is transparent to both users and applications, and requires no code profiling, code modifications, or user intervention.

In various efforts to advance this last approach, researchers have developed rigorous theories and intelligent algorithms in power awareness for autonomic DVFS-based runtime systems. Examples include the β power-aware algorithm,[1] which served as the basis for EnergyFit, and CPU MISER.[9] These systems work on any commodity platform that supports DVFS and are capable of reducing power and energy consumption while minimizing the impact on performance. Figure 3, for example, shows that the β power-aware, runtime system saves an average of 20 percent of the processor's energy while impacting performance by only 3 percent.[1]



3 Energy savings and performance slowdown of the EnergyFit DVFS algorithm. Overall, we observed an average of 20 percent energy savings and 3 percent performance slowdown. For the MG benchmark, we observed a 15 percent energy savings and 1 percent performance speedup.

Industry-Wide Awareness

The ever-growing concerns related to global warming, coupled with the exponential growth of data center and supercomputing installations, means that the energy efficiency of servers and supercomputers has become one of today's major IT issues, as evidenced by the following series of events:

- The first workshop on High-Performance, Power-Aware Computing debuted in April 2005 to provide a timely forum for the exchange and dissemination of new ideas, techniques, and research in high-performance, power-aware computing (http://hpac.cs.vt.edu).
- The SPEC Power and Performance Committee began developing benchmarks for evaluating energy efficiency in server-class computers in January 2006 (www.spec.org/specpower/).
- The Green500 list's Web site (www.green500.org) launched in November 2006 to provide a ranking of the most energy-efficient supercomputers in the world.
- The US Congress passed public law 109-431 in December 2006 to call for the study and promotion of energy-efficient computers and servers in the US (www.gpoaccess.gov/plaws/).
- The Green Grid Consortium (http://www.thegreengrid.org/) formed in February 2007 as a global group of companies dedicated to advancing energy efficiency in data centers and computing ecosystems.
- The US Environmental Protection Agency (EPA) released its "Report to Congress on Server and Data Center Energy Efficiency" in August 2007 to prioritize efficiency opportunities and policies that could lead to additional savings (www.energystar.gov/index.cfm?c=prod_development.server_efficiency_study).

In addition to these events, we've also seen a significant response to green issues by major vendors in the supercomputing industry. Both Intel and AMD, for example, released roadmaps aimed at improving the energy efficiency of their flagship processors. IBM, Hewlett-Packard, Sun, and Dell have also adjusted their server strategies to promote more efficient computer systems and technologies.

Based on our own research and experiences, we believe that the future of energy-efficient supercomputing will evolve in the direction of holistic power-aware systems. In particular, we think that

- the low-power and power-aware approaches will converge and be integrated into supercomputing system design and integration;
- the power-aware approach will be exploited at different levels in the whole system, especially in microarchitecture, operating, and runtime systems;
- power-aware features will be available to all major parts of the system (memory, disk, video cards, coprocessors, network, storage, and even power supplies will be made either more energy efficient or power aware); and
- at a high level, job scheduling, workload migration, and server consolidation will provide additional opportunities for energy savings.

Ignoring power consumption as a design constraint will result in supercomputing systems with high operational costs and diminished reliability. Without technology innovation, future petaflop machines will need approximately 75 to 100 MW to power up and cool down, resulting in a bill on the order of US$100 million per year. Furthermore, the expected mean time between failures for such gigantic systems will be on the order of seconds,[13] indicating an unacceptable availability and low productivity.

Fortunately, improving energy efficiency in supercomputers (and computer systems in general) has become an emergent task for the IT industry. Hopefully, this article will inspire further innovations and breakthroughs by providing a historical perspective. From Green Destiny to BG/L to innovative power-aware supercomputing prototypes, we envision that holistic power-aware technologies will be available and largely exploited in most if not all future supercomputing systems.

References

1. C. Hsu and W. Feng, "A Power-Aware Run-Time System for High-Performance Computing," *Proc. ACM/IEEE SC2005 Conf. High Performance Networking and Computing*, IEEE CS, 2005, p. 1.
2. W. Feng, "Making the Case for Efficient Supercomputing," *ACM Queue*, vol. 1, no. 7, 2003, pp. 54–64.
3. W. Feng et al., "Honey, I Shrunk the Beowulf!" *Proc. 2002 Int'l Conf. Parallel Processing* (ICPP 02), IEEE CS, 2002, pp. 141–149.
4. W. Feng et al., "The Bladed Beowulf: A Cost-Effective Alternative to Traditional Beowulfs," *Proc. IEEE Int'l Conf. Cluster Computing* (CLUSTER 02), IEEE CS, 2002, pp. 245–257.
5. M.S. Warren et al., "High-Density Computing: A 240-Processor Beowulf in One Cubic Meter," *Proc. IEEE/ACM SC2002 Conf.*, IEEE CS, 2002, pp. 61–72.
6. H. Nakashima et al., "MegaProto: 1 TFlops/10kW Rack Is Feasible Even with Only Commodity Technology," *Proc. 2005 ACM/IEEE Conf. Supercomputing*, ACM Press IEEE CS, 2005, pp. 26–28.
7. W. Feng et al., "Green Supercomputing in a Desktop Box," *Proc. 21st IEEE Int'l Parallel and Distributed Processing Symp.* (IPDPS 2007), IEEE CS, 2007, pp. 1-8.
8. N.R. Adiga et al., "An Overview of the BlueGene/L Supercomputer," *Proc. IEEE/ACM SC2002: High-Performance Computing, Networking, and Storage Conf.*, IEEE CS, 2002, pp. 1-22.
9. R. Ge et al., "CPU MISER: A Performance-Directed, Run-Time System for Power-Aware Clusters," *Proc. 36th Int'l Conf. Parallel Processing* (ICPP 07), IEEE CS, 2007, p. 18.
10. V.W. Freeh et al., "Using Multiple Energy Gears in MPI Programs on a Power-Scalable Cluster," *Proc. 10th ACM Symp. Principles and Practice of Parallel Programming* (PPoPP), ACM Press, 2005, pp. 164–173.
11. K.W. Cameron et al., "High-Performance, Power-Aware Distributed Computing for Scientific Applications," *Computer*, vol. 38, no. 11, 2005, pp. 40–47.
12. G. Chen et al., "Reducing Power with Performance Constraints for Parallel Sparse Applications," *Proc. 19th IEEE Intl Parallel and Distributed Processing Symp.* (IPDPS 05), IEEE CS, 2005, p.231a.
13. N.R.C. Committee on the Future of Supercomputing, *Getting Up to Speed: The Future of Supercomputing*, Nat'l Academy Press, 2004.

Wu-chun Feng *is an associate professor of computer science and electrical and computer engineering at Virginia Tech. Contact him at feng@cs.vt.edu.*

Xizhou Feng *is an senior research associate in the Network Dynamics and Simulation Science Laboratory at Virginia Tech. Contact him at fengx@vt.edu.*

Rong Ge *is a postdoctoral fellow at the Center for High-End Computing Systems at Virginia Tech. Contact her at ge@cs.vt.edu.*