CTWatch QUARTERLY

ISSN 1555-9874 Volume 1 Number 3 August 2005

THE COMING ERA OF LOW POWER, HIGH-PERFORMANCE COMPUTING

TRENDS, PROMISES, AND CHALLENGES

GUEST EDITOR: SATOSHI MATSUOKA, TOKYO INSTITUTE OF TECHNOLOGY

FEATURE ARTICLES

NOTES AND COMMENTARY

AVAILABLE ON-LINE AT
http://www.ctwatch.org/quarterly/

Cyberinfrastructure Technology Watch
http://www.ctwatch.org/

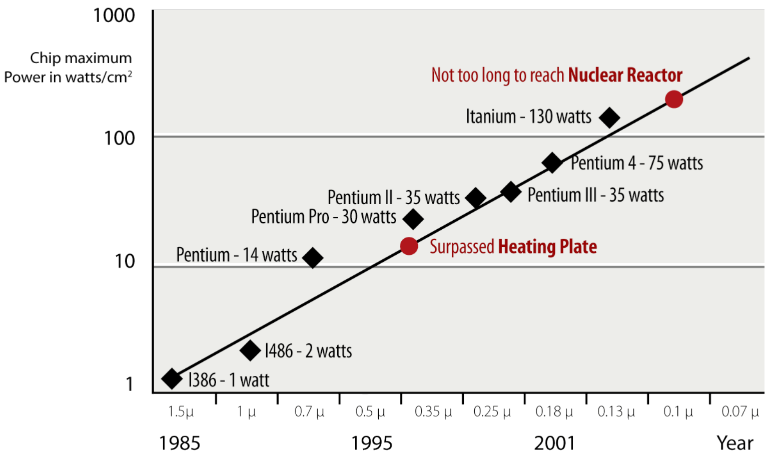
The Importance of Being Low Power in High-Performance Computing

Wu-chun Feng
Los Alamos National Laboratory

Introduction

Why should the high-performance computing community even care about (low) power consumption? The reasons are at least two-fold: (1) efficiency, particularly with respect to cost, and (2) reliability.

For decades, we have focused on performance, performance, and occasionally, price/performance, as evidenced by the Top500 Supercomputer List[1] as well as the Gordon Bell Awards for Performance and Price/Performance at SC.[2] So, to achieve better performance per compute node, microprocessor vendors have not only doubled the number of transistors (and speed) every 18-24 months, but they have also doubled the power densities, as shown in Figure 1. Consequently, keeping a large-scale high-performance computing (HPC) system functioning properly requires continual cooling in a large machine room, or even a new building, thus resulting in substantial operational costs. For instance, given that the cooling bill alone at Lawrence Livermore National Laboratory (LLNL) is $6M/year and given that for every watt (W) of power consumed by an HPC system at LLNL, 0.7 W of cooling is needed to dissipate the power; the annual cost to both power and cool HPC systems at LLNL amounts to a total of $14.6M per year, and this does not even include the costs of acquisition, integration, upgrading, and maintenance.[3] Furthermore, when nodes consume and dissipate more power, they must be spaced out and aggressively cooled; otherwise, such power causes the temperature of a system to increase rapidly enough that for every 10° C increase in temperature, the failure rate doubles, as per Arrhenius' equation as applied to microelectronics.[4]



Soure: Intel

Figure 1. Moore's Law for Power Consumption

Our own informal empirical data from late 2000 to early 2002 indirectly supports Arrenhius' equation. In the winter, when the temperature inside our warehouse-based work environment at Los Alomas National Laboratory (LANL) hovered around 21-23° C, our 128-CPU Beowulf cluster — Little Blue Penguin (LBP) — failed approximately once per week. In contrast, the LBP cluster failed roughly twice per week in the summer when the temperature in the warehouse reached 30-32° C. Such failures led to expensive operational and maintenance costs relative to technical staff working to fix the failures and the cost of replacement parts. Furthermore, there is the lost productivity of technical staff due to the failures.

[1] http://www.top500.org/

[2] http://www.sc-conference.org

[3] M. Seager, "What Are The Future Trends in High-Performance Interconnects for Parallel Computers?" *IEEE Symp. on High-Performance Interconnects Panel*, August 2004.

[4] W. Feng, "Making a Case for Efficient Supercomputing," *ACM Queue*, 1(7):54-64, October 2003.

Perhaps more disconcerting is how our warehouse environment affected the results of the Linpack benchmark when running on a dense Beowulf cluster back in 2002: The cluster produced an answer outside the residual (i.e., a silent error) after only ten minutes of execution. Yet when the same cluster was placed in an 18-19° C machine-cooled room, it produced the correct answer. This experience loosely corroborated a prediction made by Graham, et al — "In the near future, soft errors will occur not just in memory but also in logic circuits."[5]

Power (and its affect on reliability) is even more of an issue for larger-scale HPC systems, such as those shown in Table 1. Despite having exotic cooling facilities in place, the reliability of these large-scale HPC systems is measured in hours,[6] and in all cases, the leading source of outage is hardware, with the cause often being attributed to excessive heat. Consequently, as noted by Eric Schmidt, CEO of Google, what matters most to Google "is not speed but power — low power, because data centers can consume as much electricity as a city."[7] That is, though speed is important, power consumption (and hence, reliability) is more so. By analogy, what Google, and arguably application scientists in HPC, desires is the fuel-efficient, highly reliable, low-maintenance Toyota Camry of supercomputing, not the Formula One race car of supercomputing with its energy inefficiency, unreliability, and exorbitant operational and maintenance costs. In addition, extrapolating today's failure rates to an HPC system with 100,000 processors suggests that such a system would "spend most of its time checkpointing and restarting. Worse yet, since many failures are heat related, the [failure] rates are likely to increase as processors consume more power."[5]

System	CPUs	Reliability
ASCI Q	8,192	MTBI: 6.5 hours. Leading outage sources: storage, CPU, memory.
ASCI White	8,192	MTBF: 5.0 hours ('01) and 40 hours ('03). Leading outage sources: storage, CPU, 3rd-party HW.
PSC Lemieux	3,016	MTBI: 9.7 hours.

MTBI: mean time between interrupts = wall clock hours / # downtime periods

MTBF: mean time between failures (measured)

Table 1. Reliability of Leading-Edge HPC Systems

Low-Power HPC: The Past

Based on the above evidence, I would argue that although performance and price/performance are important, we need to focus more attention on efficiency and reliability in the coming decades. And as contended above, this translates into a substantial reduction in the power consumption of HPC systems via low-power (or power-aware) approaches. Our Green Destiny cluster was arguably one of the first such systems,[4][8][9] designed in late 2001 and debuting in early 2002 as the first major instantiation of the *Supercomputing in Small Spaces* project.[10]

Green Destiny, as shown in Figure 2a, was a 240-CPU Linux-based cluster with a footprint of only five square feet and a power appetite of as little as 3.2 kW (i.e., two hairdryers). Performance-wise, it produced 101 Gflops on the Linpack benchmark, which was as fast as a 256-CPU SGI Origin 2000 at the time.[11] Despite its competitive performance then,[12] many still felt that Green Destiny sacrificed too much performance *to achieve low power consumption, and consequently, high efficiency and unprecedented reliability, i.e., no unscheduled downtime*

[5] S. Graham, M. Snir, and C. Patterson, eds., Getting Up to Speed: *The Future of Supercomputing*, National Research Council, Committee on the Future of Supercomputing, National Academies Press, 2005.

[6] D. Reed, "High-End Computing: The Challenge of Scale," *Director's Colloquium*, Los Alamos National Laboratory, May 2004.

[7] J. Markoff and S. Lohr, "Intel's Huge Bet Turns Iffy," The New York Times, September 29, 2002.

[8] W. Feng, M. Warren, and E. Weigle, "The Bladed Beowulf: A Cost-Effective Alternative to Traditional Beowulfs," *4th IEEE International Conference on Cluster Computing (IEEE Cluster)*, Chicago, IL, September 2002.

[9] G. Johnson, "At Los Alamos, Two Visions of Supercomputing," *The New York Times*, June 25, 2002.

[10] http://sss.lanl.gov; At SC2001 in November, we demonstrated a small-scale 24-node prototype dubbed MetaBlade, running a simulation of a 10-million-body galaxy formation.

[11] http://www.top500.org/list/2001/11

[12] The original performance of Green Destiny on the Linpack benchmark was indeed "low performance" at about 68 Gflops. However, given that the Transmeta CPU was a hardware-software hybrid, we were able to optimize its floating-point performance (in system software) by 50%, resulting in a Linpack rating of 101 Gflops.

in its 24-month lifetime while running at 7,400 feet above sea level in a dusty 85° F warehouse without any cooling, air filtration, or air humidification.

The above tradeoff is captured (in part) in Table 2, where we present the raw configuration and execution numbers of four HPC systems as well their efficiency numbers with respect to memory density, storage density, and computational efficiency relative to space and power consumption.[13] As one would expect from a Formula One race car for supercomputing, the ASCI White supercomputer leads all the raw performance categories (shown in red). On the other hand, given that Green Destiny was specifically designed with low power and high efficiency in mind, it handily "wins" all the efficiency categories: Memory density, storage density, and computational efficiency relative to space and power are all two orders of magnitude better (or nearly so) than the other HPC systems, as shown in red in Table 2.

Metric / HPC System	Avalon Beowulf	ASCI Red	ASCI White	Green Destiny	
Year	1996	1996	2000	2002	
# CPUs	140	9298	8192	240	
Performance (Gflops)	18	600	2500	58	
Space (ft^2)	120	1600	9920	5	
Power (kW)	18	1200	2000	5	
DRAM (GB)	36	585	6200	150	(270 max)
Disk (TB)	0.4	2.0	160.0	4.8	(38.4 max)
DRAM Density (MB/ft^2)	300	366	625	30000	(54000 max)
Disk Density (GB/ft^2)	3.3	1.3	16.1	960.0	(7680 max)
Perf/Space (Mflops/ft^2)	150	375	252	11600	
Perf/Power (Mflops/W)	1.0	0.5	1.3	11.6	

Table 2. Comparison of HPC Systems on an n-body Astrophysics Code for Galaxy Formation

Low-Power HPC (and Power-Aware HPC): The Present

The preceding work has now bifurcated into two different directions but both are still oriented towards reducing power consumption: (1) a low-power, architectural approach and (2) a power-aware, software-based approach.

Low-Power, Architectural Approach

In the arena of low-power architectures for HPC, there exist three related but distinct approaches. The first, and most natural, evolution of Green Destiny is the MegaScale Computing project whose goals are more ambitious than Green Destiny's were. The MegaScale Computing project[14] is a multi-institutional project that is looking towards building future computing systems with over a million processing elements in total. Like the Supercomputing in Small Spaces project, the MegaScale Computing project aims to simultaneously achieve high performance and low power consumption via high-density packaging and adopting low-power CPUs, but with the loftier design goals of one Tflop/rack, 10 kW/rack, and 100 Mflops/W. Similar to Green Destiny, their first prototype of an ultra low-power

[13] We note up-front that the comparison is an "apples-to-oranges" one given that the HPC systems are from different eras and have different architectures. The choice of HPC systems was motivated by the fact that we had complete configuration information of the systems and complete and unencumbered access to the systems to tune our n-body code.

[14] http://www.para.tutics.tut.ac.jp/megascale/r_mproto.html

MegaScale system, called MegaProto, also leverages Transmeta CPUs, which deliver very low power but reasonable HPC performance, resulting in extraordinary performance-power ratios.[15] A picture of their MegaProto prototype that was demonstrated at SC2004 is shown in Figure 2b; it is a 16-CPU low-power cluster with dual Gigabit Ethernet for data communication and Fast Ethernet for management and control — all in a compact 1U chassis that consumes only 330 W. (As a point of reference, a traditional dual-CPU compute node consumes 250 W of power. Thus, for 16 CPUs, the aggregate power consumption would run on the order of 2000 W and would then need an additional 1400 W of power to cool the system for a total of 3400 W, or over ten times more power consumption.)

The second and more modest architectural approach to low power is a commercial evolution of Green Destiny, as embodied by Orion Multisystems.[16] The company has two offerings: the DT-12 (i.e., DeskTop-12 nodes) and DS-96 (i.e., DeskSide-96 nodes), as shown in Figure 2c. Their offerings are intended to fill the widening performance gap between PCs and supercomputers, as shown in Figure 3, whereas the ultimate goal of the MegaScale Computing project is to create the capability of constructing a supercomputer with one-million processing elements.

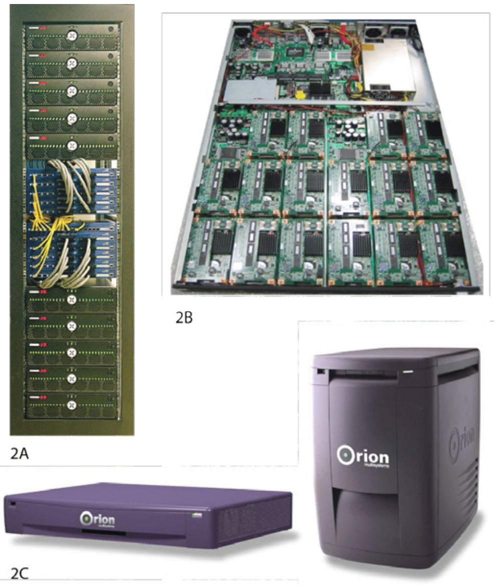


Figure 2a. Green Destiny
Figure 2b. MegaProto: An Ultra Low-Power Prototype of the Megascale Computing Project
Figure 2c. Orion Multisystems DT-12 and DS-96

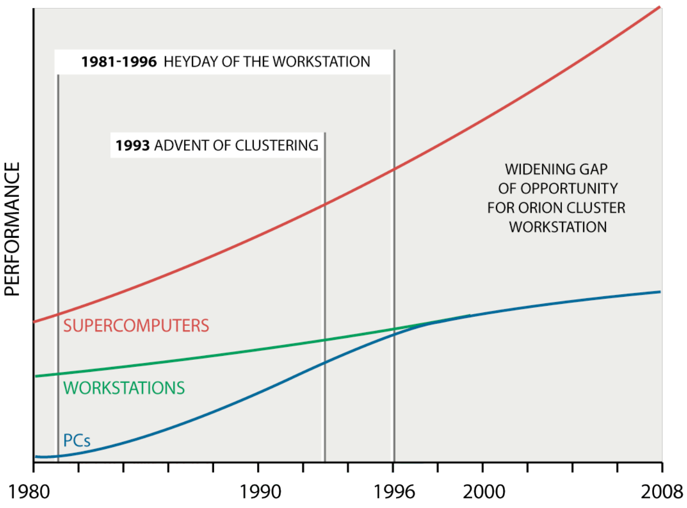


Figure 3. The Widening Performance Gap Between PCs and Supercomputers

Orion Multisystems identified three technology trends that make their offerings ideally positioned as the cluster workstation of the future: (1) the rise of cluster-based high-performance computers, (2) the maturity of open-source cluster software, and (3) the rapid decline of the traditional workstation. By placing a cluster workstation at the hands of an applications scientist, it can be more naturally used as a dedicated personal resource — application debugging with scalability at the desktop, redundancy possibilities whenever the datacenter HPC resource is down and unavailable, and no more scheduling conflicts or long queues for access to a datacenter HPC resource. And perhaps most importantly, by leveraging low-power components, both the DT-12 and DS-96 can be plugged into a standard electrical wall outlet

[15] H. Nakashima, H. Nakamura, M. Sato, T. Boku, S. Matsuoka, D. Takahashi, and Y. Hotta, "MegaProto: A Low-Power and Compact Cluster for High-Performance Computing," *IEEE Workshop on High-Performance, Power-Aware Computing (in conjunction with the IEEE Parallel & Distributed Processing Symposium)*, Denver, CO, April 2005.

[16] http://www.orionmulti.com

in any office, as the former only consumes as much power as an overhead light with two 75-W light bulbs and the latter consumes as much as a typical hairdryer, i.e., 1.5 kW.

Finally, the most prominent architectural approach to low-power supercomputing is IBM BlueGene/L, which debuted nine months ago on the Top500 Supercomputer List[1] as the fastest supercomputer in the world, relative to the Linpack benchmark. For an overview of the IBM BlueGene/L architecture and system software, see respective notes.[17] [18] Initial performance evaluations of IBM BlueGene/L can also be found in notes.[19] [20] [21] In short, IBM Blue Gene/L is a very large-scale, low-power (for its size) supercomputer. Its 65,536 CPUs, which are PowerPC 440s, are organized into 64 racks of 1024 CPUs per rack, where each rack of 1024 CPUs consumes only 28.14 kW, resulting in an aggregate power consumption of 1.8 MW.

Given that the only program that has been run across the aforementioned systems is the Linpack benchmark, Table 3 presents the same evaluation metrics as in Table 2 but for the Linpack benchmark.[22] And as in Table 2, Table 3 highlights the leader for a given metric in red.[23] One of the most striking aspects of this table is that IBM Blue Gene/L does *not* use the most amount of space or power despite having the most number of CPUs. Its resulting performance-space and performance-power ratios are consequently astounding, at least relative to Linpack. As an additional reference point, the Japanese Earth Simulator, which has been argued to be the most powerful supercomputer in the world relative to executing *real applications*, reaches 35,860 Gflops for Linpack while occupying 17,222 ft^2 and consuming 7,000 kW. This translates to performance-space and performance-power ratios of 2,082 Mflops/ft^2 and 5.13 Mflops/W, respectively.

Metric \ HPC System	ASCI Red	ASCI White	Green Destiny	MegaProto	Orion DS-96	IBM Blue Gene/L
Year	1996	2000	2002	2004	2005	2005
Performance (Gflops)	2379	7226	101	5.62	110	136800
Space (ft^2)	1600	9920	5	3.52	2.95	2500
Power (kW)	1200	2000	5	0.33	1.58	1800
DRAM (GB)	585	6200	150	4	96	32768
Disk (TB)	2.0	160.0	4.8	n/a	7.68	n/a
DRAM Density (MB/ft^2)	366	625	30000	1136	32542	13107
Disk Density (GB/ft^2)	1	16	960	n/a	2603	n/a
Perf/Space (Mflops/ft^2)	1487	728	20202	1597	37228	54720
Perf/Power (Mflops/W)	2	4	20	17	70	76

Table 3. Comparison of HPC Systems on the LINPACK Benchmark

Despite the performance of HPC systems such as Green Destiny, MegaProto, Orion Multisystems DT-12 and DS-96, and IBM Blue Gene/L, many HPC researchers gripe about the raw performance per compute node, which then requires additional compute nodes to compensate for the lower per-node performance. This, of course, is in contrast to using fewer but more powerful and more power-hungry server processors, e.g., Power5 in ASC Purple, which is slated to require 7.5 MW to power and cool its 12,000+ CPU system. The full system is expected to generate more than 16,000,000 BTU/h in heat, thus requiring new air-handling

[17] IBM and Lawrence Livermore National Laboratory, "An Overview of the BlueGene/L Supercomputer," *IEEE/ACM SC2002: High-Performance Networking & Computing Conference*, Baltimore, MD, November 2002.

[18] G. Almasi, R. Bellofatto, J. Brunheroto, C. Cascaval, J. G. Castanos, L. Ceze, P. Crumley, C. C. Erway, J. Gagliano, D. Lieber, X. Martorell, J. Moreira, A. Sanomiya, and K. Strauss, "An Overview of the Blue Gene/L System Software Organization," *Euro-Par 2003 Conference*, Klagenfurt, Austria, August 2003.

[19] V. Bulatov, W. Cai, J. Fier, M. Hiratani, G. Hommes, T. Pierce, M. Tang, M. Rhee, K. Yates, and T. Arsenlis, "Scalable Line Dynamics in ParaDiS," *IEEE/ACM SC2004: High-Performance Computing, Networking, and Storage Conference*, Pittsburgh, PA, November 2004.

[20] K. Davis, A. Hoisie, G. Johnson, D. Kerbyson, M. Lang, S. Pakin, and F. Petrini, "A Performance and Scalability Analysis of the BlueGene/L Architecture," *IEEE/ACM SC2004: High-Performance Computing, Networking, and Storage Conference*, Pittsburgh, PA, November 2004.

[21] G. Almasi, S. Chatterjee, A. Gara, J. Gunnels, M. Gupta, A. Henning, J. Moreira, and B. Walkup, " Unlocking the Performance of the BlueGene/L Supercomputer," *IEEE/ACM SC2004: High-Performance Computing, Networking, and Storage Conference*, Pittsburgh, PA, November 2004.

[22] We note that in addition to the differences in machine architectures and eras (which makes direct comparisons difficult) that power and space consumption do not scale linearly. So, the presented data should only be taken as ballpark figures.

[23] None of the power numbers include the wattage needed for cooling. This means that for ASCI Red, ASCI White, and IBM Blue Gene/L that the power numbers would increase by a factor of 1.7 to 2.0 times. Furthermore, none of the space numbers include the extra floor(s) needed to cool the HPC systems.

designs and specifications. Furthermore, all the above solutions do *not* rely entirely on commodity technologies, and hence, may not be cost-effective. For instance, Blue Gene/L is a stripped-down version of the 700-MHz PowerPC 400 embedded CPU while Green Destiny relies on a customized high-performance version of Transmeta's code-morphing software (CMS)[24] that improves floating-point performance between 50% and 100%, e.g., 12.6 Gflops on 24 CPUs. In contrast, the 16-processor MegaProto cluster is a custom hardware solution that uses the same processor that Green Destiny did but *without* the high-performance code-morphing software (HP-CMS). Consequently, its 16 CPUs only achieve 5.62 Gflops on Linpack. To address the criticisms with respect to non-commodity parts and low performance, the next section proposes an alternative approach for reducing power consumption, one that is largely architecture-independent and based on high-end commodity hardware.

Power-Aware, Software-Based Approach

Because many systems researchers argue that the low-power architectural approach sacrifices too much performance for low power consumption and high reliability, an alternative approach in HPC has recently emerged — one that is more architecture-independent than the low-power, architectural approach and one that takes the "middle ground" relative to the tradeoff between performance and low power consumption. This alternative approach is a power-aware, software-based one, as described in the cited feasibility studies[25] [26] [27] [28] [29] and autonomic systems.[30] [31] [32] [33] The basic idea is to start with a high-performance, high-power CPU that supports a mechanism called *dynamic voltage and frequency scaling* (e.g., an AMD Opteron with support for PowerNow!) and then to create a power-aware algorithm (i.e., policy) that conserves power by scaling down the CPU supply voltage and frequency at appropriate times, as power draw is directly proportional to the CPU frequency and the square of the CPU supply voltage.

Ideally, the appropriate time to scale down the CPU voltage and frequency is whenever there is an off-chip access that the CPU is blocking-on, e.g., memory access, as the CPU has no reason to "sit and spin its wheels" at the maximum voltage and frequency while waiting for the off-chip accesses to complete. In practice, however, knowing *when* to scale the voltage and frequency and *what* to scale them to are difficult tasksfor the following reasons. First, off-chip memory accesses are done in hardware, thus power-aware software would have no way of knowing that the CPU is waiting on a memory access. Second, changing the voltage and frequency settings must be done judiciously, because at the system level, it currently takes on the order of *milliseconds* (i.e., millions of clock cycles) for the voltage and frequency to transition and stabilize at their new settings.

The current and most ubiquitous approach for power-awareness is based primarily on CPU utilization and is meant to extend the battery life in a laptop computer. When the CPU utilization drops below some threshold, the CPU voltage and frequency are lowered to conserve energy; when the CPU utilization exceeds some threshold, the CPU voltage and frequency are raised to improve performance. While this simple approach is both application and input independent as well as transparent to the end user, it is only effective for interactive use, e.g., laptop usage of Microsoft Office, and depends critically upon the choice of the threshold values.[34] For scientific applications, the approach is ineffective as such applications do not have an abundance of CPU idle time that can be taken advantage of.[32] Therefore, there exists a need for a power-aware algorithm that works effectively on scientific applications.

[24] Each Transmeta processor has a software layer, called code-morphing software, that dynamically morphs x86 instructions into VLIW instructions. This provides x86 software with the impression that it is being run on native x86 hardware.

[25] X. Feng, R. Ge, and K. Cameron, "Power and Energy Profiling of Scientific Applications on Distributed Systems," *19th IEEE International Parallel & Distributed Processing Symposium*, Denver, CO, April 2005.

[26] V. Freeh, D. Lowenthal, F. Pan, and N. Kappiah, "Using Multiple Energy Gears in MPI Programs on a Power-Scalable Cluster," *ACM Symposium on Principles and Practices of Parallel Programming (PPoPP'05)*, June 2005.

[27] V. Freeh, D. Lowenthal, R. Springer, F. Pan, and N. Kappiah, "Exploring the Energy-Time Tradeoff in MPI Programs on a Power-Scalable Cluster," *19th IEEE International Parallel & Distributed Processing Symposium*, Denver, CO, April 2005.

[28] R. Ge, X. Feng, and K. Cameron, "Improvement of Power-Performance Efficiency for High-End Computing," *1st IEEE Workshop on High-Performance, Power-Aware Computing (in conjunction with the 19th IEEE International Parallel & Distributed Processing Symposium)*, Denver, CO, April 2005.

[29] C. Hsu and U. Kremer, "The Design, Implementation, and Evaluation of a Compiler Algorithm for CPU Energy Reduction," *ACM Conference on Programming Languages Design and Implementation (PLDI'03)*, June 2003.

[30] W. Feng and C. Hsu, "The Origin and Evolution of Green Destiny," *IEEE Cool Chips VII: An International Symposium on Low-Power and High-Speed Chips*, Yokohama, Japan, April 2004.

[31] W. Feng and C. Hsu, "Green Destiny and Its Evolving Parts," Innovative Supercomputer Architecture Award, *19th International Supercomputer Conference*, Heidelberg, Germany, June 2004.

[32] C. Hsu and W. Feng, "Effective Dynamic Voltage Scaling Through CPU-Boundedness Detection," *4th ACM Workshop on Power-Aware Computer Systems*, Portland, OR, December 2004.

[33] C. Hsu and W. Feng, "A Power-Aware Run-Time System for High-Performance Computing," *ACM/IEEE SC2005: The International Conference on High-Performance Computing, Networking, and Storage*, Seattle, WA, November 2005.

[34] D. Grunwald, P. Levis, K. Farkas, C. Morrey, and M. Neufeld, "Policies for Dynamic Clock Scheduling," *4th Symposium on Operating System Design and Implementation (OSDI'00)*, Oct. 2000.

We propose such a power-aware algorithm called β-adaptation, which works on any commodity platform that supports dynamic voltage and frequency scaling (DVFS), [33] e.g., AMD Opteron with PowerNow! Implementing the algorithm in the run-time system results in a power-aware runtime system that transparently and automatically adapts CPU voltage and frequency in order to reduce power and energy consumption while minimizing impact on performance. For example, Figure 4 shows that our power-aware run-time system running NAS-MPI Class C on a four-node, 16-CPU Opteron-based cluster saves nearly an average of 20% CPU energy while impacting performance by only 3% on average. (Note: For the MG benchmark, our β-adaptation algorithm not only reduces energy consumption by 14% but it also *improves* performance slightly.)

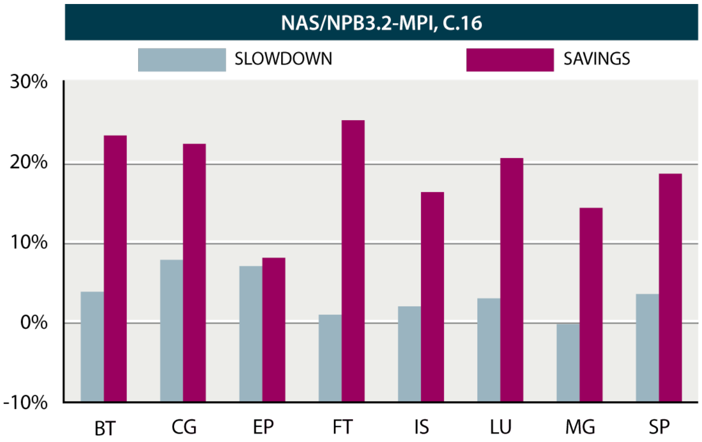


Figure 4. NAS-MPI benchmarks for Class C Workload on a Four-Node, 16-CPU Opteron-based Cluster — http://www.nas.nasa.gov/Software/MPB

Low-Power HPC (and Power-Aware HPC): The Future

Implicit in the preceding discussion is the distinction between capability and capacity computing. According to Graham et al,[5] capability computing applies maximum processing power to solve a large problem in a short period of time — with the main figure of merit being "time to solution." Another important facet to capability computing is the ability to solve problems of a magnitude that have never been solved before. Examples of such systems are the DOE ASCI-class supercomputers such as ASCI White and the recently demonstrated ASC Purple supercomputer — the Formula One race cars of supercomputing.

In contrast, capacity systems are typically cheaper and less performance-capable than capability systems on a per-node basis as well as relative to the entire system. Capacity systems allow scientists to explore design alternatives that are often needed to prepare for larger-scale runs on capability systems. In addition, capacity systems typically solve a multitude of smaller problems simultaneously. Systems such as Green Destiny, MegaProto, Orion Multisystems DS-96, and arguably Blue Gene/L fit into this category.

Because low-power HPC generally sacrifices a measurable amount of performance (e.g., 3.6-GHz Intel Xeon CPU versus 1.4-GHz Transmeta Efficeon CPU) to achieve substantially lower power consumption per node (e.g., 151 W versus 7 W), and hence, better efficiency and reliability, low-power HPC will be confined to capacity computing for the foreseeable future. See citations[35 36] for the latest results in low-power HPC.

[35] C. Hsu, W. Feng, and J. Archuleta, "Towards Efficient Supercomputing: A Quest for the Right Metric," *1st IEEE Workshop on High-Performance, Power-Aware Computing (in conjunction with the 19th International Parallel & Distributed Processing Symposium)*, Denver, CO, April 2005.

[36] H. Nakashima, M. Sato, T. Boku, S. Matuoka, D. Takahashi, and Y. Hotta, "MegaProto: 1 Tflops/ 10kW Rack Is Feasible Even with Only Commodity Technology," *ACM/IEEE SC2005: The International Conference on High-Performance Computing, Networking, and Storage*, Seattle, WA, November 2005.

But what about capability computing? HPC vendors now realize that in building capability systems, power consumption is becoming a primary design constraint because of the exorbitant operational costs associated with such systems due to their inefficiency and because of its effect of reliability, as noted in Table 1. Excessive power consumption is becoming such a dominant issue that ASC Purple requires new air-handling designs and specifications because of the 7.5-MW required to power the system and the cooling equipment. This 7.5-MW appetite equates to powering 7,500 typical homes.

With low-power HPC unable to support the requirements of capability computing and too much power being consumed by traditional capability systems, what the HPC community should expect to see over the next decade is the emergence of power-aware solutions for capability computing. These solutions will ultimately reduce operational costs and improve reliability and availability, particularly in capacity systems, while minimizing impact on overall performance. We are already seeing indications of this trend at SC2005 where the following three technical papers will be presented on power-aware HPC:

1. R. Ge, X. Feng, and K. Cameron, "Performance-Constrained, Distributed DVS Scheduling for Scientific Applications on Power-Aware Clusters." *Describes a software framework for implementing and evaluating dynamic voltage and frequency scaling, where performance-directed scheduling is of particular interest.*

2. C. Hsu and W. Feng, "A Power-Aware Run-Time System for High-Performance Computing." *Presents a power-aware run-time system on a high-end commodity cluster that automatically and transparently adapts its voltage and frequency settings to achieve about 20% energy savings on average with minimal impact on performance.*

3. N. Kappiah, V. Freeh, and D. Lowenthal, "Just-in-Time Dynamic Voltage Scaling: Exploiting Inter-Node Slack to Save Energy in MPI Programs." *Saves energy by taking advantage of the slack time that exists when the computational load is not perfectly balanced across a HPC system.*

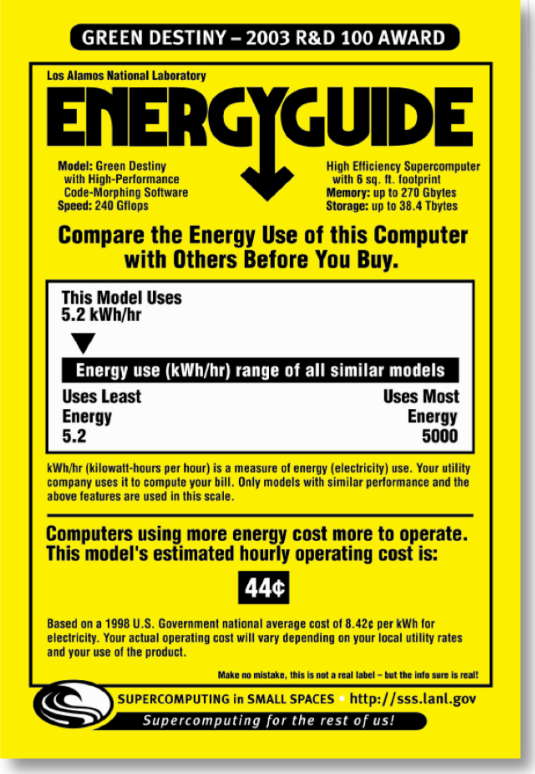
As noted earlier, a power-aware approach makes use of commodity processors (e.g., AMD Opteron[33]) with dynamic voltage and frequency scaling (e.g., PowerNow![33]) to ensure high-end capability performance while reducing power consumption. For the capability supercomputer called ASC Purple, using our power-aware run-time system would reduce the power envelope by 1.3 MW on average, thus reducing its electrical bill by $1.37M/year, when assuming a rate of $0.12/kWh. Furthermore, such a dramatic reduction in power consumption would lengthen the life of system components in the supercomputer, and hence, improve overall reliability of the supercomputer as well as those presented in Table 1.

Conclusion

Power consumption has become an increasingly important issue in HPC. Ignoring power consumption as a design constraint results in a HPC system with high operational costs and diminished reliability, which translates into lost productivity. Examples of such (capability) systems include ASCI White, ASC Q, and the recently unveiled ASC Purple.

Specifically, due to the exorbitant power consumption of ASC Purple, the facility that houses ASC Purple requires new air-handling designs and specifications to deal with ASC Purple's gargantuan 7.5-MW appetite. With an average utility rate of $0.12/kWh, the electrical bill alone for this system would run nearly $8M/year. If we scale this architecture up to a petaflop machine, it would need approximately 75 MW to power up and cool down the machine. The power bill for such a system would then be on the order of $80M/year, assuming energy costs stay at $0.12/kWh. In addition, the expected mean time between failures for systems of this size is forecasted to be on the order of hours rather than days; further scaling of such capability supercomputers would result in HPC systems that would have several failures per hour by 2010.[5]

For the above reasons, this article presented a case for low-power (and power-aware) HPC in order to significantly improve reliability and efficiency, particularly with respect to operational costs. However, the main issue with low-power HPC is that it sacrifices too much raw performance in order to achieve its goals. Perhaps what the HPC community needs is an EnergyGuide sticker for HPC systems, like the one shown in Figure 5 for Green Destiny. Or more seriously, perhaps we should remember that our attitude towards energy contributed to the massive rolling blackouts in the summers of 2000, 2001, and 2003 and cost the U.S. billions of dollars and disrupted millions of lives, as noted this month by President George W. Bush when signing the 10-year, $12.3-billion Energy Policy Act of 2005.



GREEN DESTINY – 2003 R&D 100 AWARD

Los Alamos National Laboratory

ENERGYGUIDE

Model: Green Destiny
with High-Performance
Code-Morphing Software
Speed: 240 Gflops

High Efficiency Supercomputer
with 6 sq. ft. footprint
Memory: up to 270 Gbytes
Storage: up to 38.4 Tbytes

Compare the Energy Use of this Computer with Others Before You Buy.

This Model Uses
5.2 kWh/hr

▼

Energy use (kWh/hr) range of all similar models

Uses Least Energy 5.2	Uses Most Energy 5000

kWh/hr (kilowatt-hours per hour) is a measure of energy (electricity) use. Your utility company uses it to compute your bill. Only models with similar performance and the above features are used in this scale.

Computers using more energy cost more to operate.
This model's estimated hourly operating cost is:

44¢

Based on a 1998 U.S. Government national average cost of 8.42¢ per kWh for electricity. Your actual operating cost will vary depending on your local utility rates and your use of the product.

Make no mistake, this is not a real label – but the info sure is real!

SUPERCOMPUTING in SMALL SPACES • http://sss.lanl.gov
Supercomputing for the rest of us!

Figure 5. EnergyGuide Sticker for Green Destiny

As a compromise, there exists an emerging body of research in power-aware HPC. The basic idea is to start with a high-performance, high-power CPU that supports a mechanism called dynamic voltage and frequency scaling and then to create a power-aware algorithm that conserves power by scaling down the CPU supply voltage and frequency at appropriate times, as power draw is directly proportional to the CPU frequency and the square of the CPU supply voltage. Because the CPU consumes the largest percentage of power in a HPC node, this technique has been shown to be highly effective in reducing the overall power and energy consumption in an HPC system.

In the longer term, e.g., by 2020 when the failure rate is expected to reach several failures *per minute*,[5] we will need the continued proactive approach towards power consumption espoused here in order to stave off the aforementioned forecast as well as reactive fault detection and fault handling in order to give the user the illusion of a fault-free machine.

This work was supported by the DOE LDRD Exploratory Research Program through Los Alamos National Laboratory contract W-7405-ENG-36.

PUBLISHERS

Fran Berman, Director of SDSC

Thom Dunning, Director of NCSA

EDITOR-IN-CHIEF

Jack Dongarra, UTK/ORNL

MANAGING EDITOR

Terry Moore, UTK

CENTER SUPPORT

Greg Lund, SDSC

Karen Green, NCSA

PRODUCTION EDITOR

Scott Wells, UTK

GRAPHIC DESIGNER

David Rogers, UTK

CTWatch QUARTERLY

Volume 1 Number 3 August 2005

THE COMING ERA OF LOW POWER, HIGH-PERFORMANCE COMPUTING

TRENDS, PROMISES, AND CHALLENGES

GUEST EDITOR: SATOSHI MATSUOKA, TOKYO INSTITUTE OF TECHNOLOGY

AVAILABLE ON-LINE:
www.ctwatch.org/quarterly/

CTWatch
http://www.ctwatch.org/

E-MAIL:
quarterly@ctwatch.org

http://icl.cs.utk.edu/

http://www.ncsa.uiuc.edu/

http://www.sdsc.edu/

CTWatch Quarterly is a publication of the CyberInfrastruture Partnership (CIP).