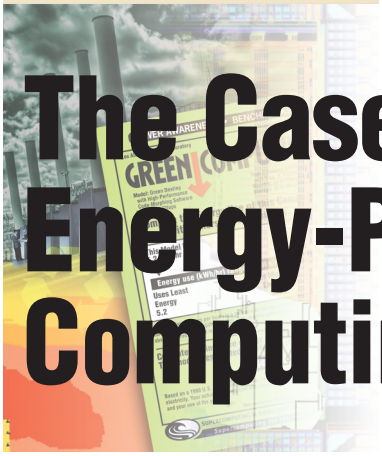

The Case for Energy-Proportional Computing

Luiz André Barroso and Urs Hölzle

Google

Energy-proportional designs would enable large energy savings in servers, potentially doubling their efficiency in real-life use. Achieving energy proportionality will require significant improvements in the energy usage profile of every system component, particularly the memory and disk subsystems.

Energy efficiency, a new focus for general-purpose computing, has been a major technology driver in the mobile and embedded areas for some time. Earlier work emphasized extending battery life, but it has since expanded to include peak power reduction because thermal constraints began to limit further CPU performance improvements.

Energy management has now become a key issue for servers and data center operations, focusing on the reduction of all energy-related costs, including capital, operating expenses, and environmental impacts. Many energy-saving techniques developed for mobile devices became natural candidates for tackling this new problem space. Although servers clearly provide many parallels to the mobile space, we believe that they require additional energy-efficiency innovations.

In current servers, the lowest energy-efficiency region corresponds to their most common operating mode. Addressing this mismatch will require significant rethinking of components and systems. To that end, we propose that energy proportionality should become a primary design goal. Although our experience in the server space motivates these observations, we believe that energy-proportional computing also will benefit other types of computing devices.

DOLLARS & CO_2

Recent reports[1,2] highlight a growing concern with computer-energy consumption and show how current trends could make energy a dominant factor in the total cost of ownership.[3] Besides the server electricity bill, TCO includes other energy-dependent components such as the cost of energy for the cooling infrastructure and provisioning costs, specifically the data center infrastructure's cost. To a first-order approximation, both cooling and provisioning costs are proportional to the average energy that servers consume, therefore energy efficiency improvements should benefit all energy-dependent TCO components.

Efforts such as the Climate Savers Computing Initiative (www.climatesaverscomputing.org) could help lower worldwide computer energy consumption by promoting widespread adoption of high-efficiency power supplies and encouraging the use of power-savings features already present in users' equipment. The introduction of more efficient CPUs based on chip multiprocessing has also contributed positively toward more energy-efficient servers.[3] However, long-term technology trends invariably indicate that higher performance means increased energy usage. As a result, energy efficiency must improve as fast as computing performance to avoid a significant growth in computers' energy footprint.

SERVERS VERSUS LAPTOPS

Many of the low-power techniques developed for mobile devices directly benefit general-purpose servers, including multiple voltage planes, an array of energy-efficient circuit techniques, clock gating, and dynamic

0018-9162/07/$25.00 © 2007 IEEE

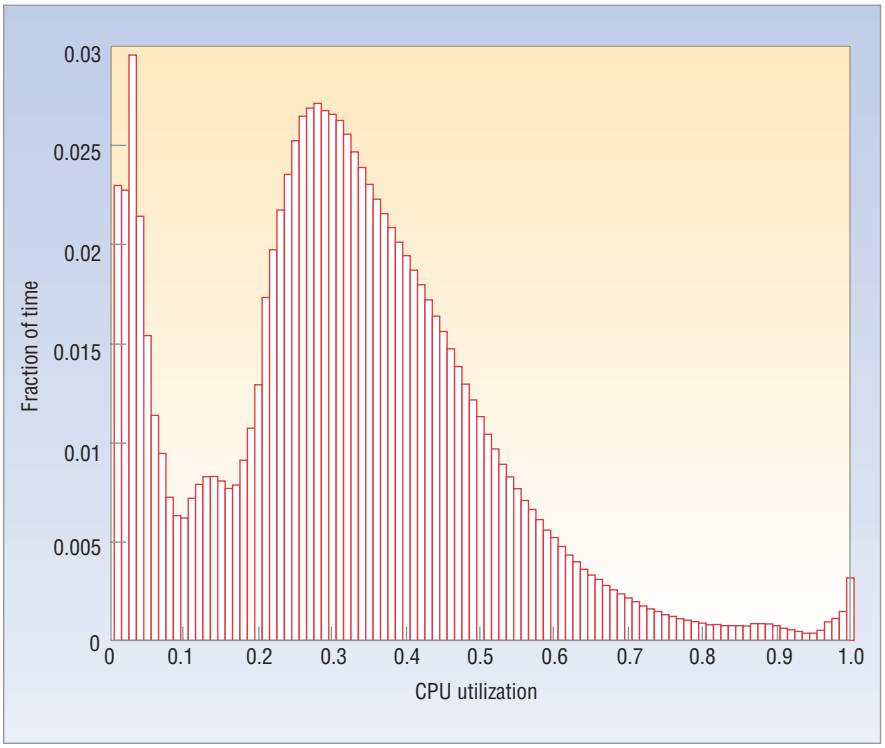

Figure 1. Average CPU utilization of more than 5,000 servers during a six-month period. Servers are rarely completely idle and seldom operate near their maximum utilization, instead operating most of the time at between 10 and 50 percent of their maximum utilization levels.

voltage-frequency scaling. Mobile devices require high performance for short periods while the user awaits a response, followed by relatively long idle intervals of seconds or minutes. Many embedded computers, such as sensor network agents, present a similar bimodal usage model.[4]

This kind of activity pattern steers designers to emphasize high energy efficiency at peak performance levels and in idle mode, supporting inactive low-energy states, such as sleep or standby, that consume near-zero energy. However, the usage model for servers, especially those used in large-scale Internet services, has very different characteristics.

Figure 1 shows the distribution of CPU utilization levels for thousands of servers during a six-month interval.[5] Although the actual shape of the distribution varies significantly across services, two key observations from Figure 1 can be generalized: Servers are rarely completely idle and seldom operate near their maximum utilization. Instead, servers operate most of the time at between 10 and 50 percent of their maximum utilization levels. Such behavior is not accidental, but results from observing sound service provisioning and distributed systems design principles.

An Internet service provisioned such that the average load approaches 100 percent will likely have difficulty meeting throughput and latency service-level agreements because minor traffic fluctuations or any internal disruption, such as hardware or software faults, could tip it over the edge. Moreover, the lack of a reasonable amount of slack makes regular operations exceedingly complex because any maintenance task has the potential to cause serious service disruptions. Similarly, well-provisioned services are unlikely to spend significant amounts of time completely idle because doing so would represent a substantial waste of capital.

Even during periods of low service demand, servers are unlikely to be fully idle. Large-scale services usually require hundreds of servers and distribute the load over these machines. In some cases, it might be possible to completely idle a subset of servers during low-activity periods by, for example, shrinking the number of active front ends. Often, though, this is hard to accomplish because data, not just computation, is distributed among machines. For example, common practice calls for spreading user data across many databases to eliminate the bottleneck that a central database holding all users poses.

Spreading data across multiple machines improves data availability as well because it reduces the likelihood that a crash will cause data loss. It can also help hasten recovery from crashes by spreading the recovery load across a greater number of nodes, as is done in the Google File System.[6] As a result, all servers must be available, even during low-load periods. In addition, networked servers frequently perform many small background tasks that make it impossible for them to enter a sleep state.

With few windows of complete idleness, servers cannot take advantage of the existing inactive energy-savings modes that mobile devices otherwise find so effective. Although developers can sometimes restructure applications to create useful idle intervals during periods of reduced load, in practice this is often difficult and even harder to maintain. The Tickless kernel[7] exemplifies some of the challenges involved in creating and maintaining idleness. Moreover, the most attractive inactive energy-savings modes tend to be those with the highest wake-up penalties, such as disk spin-up time, and thus their use complicates application deployment and greatly reduces their practicality.

ENERGY EFFICIENCY AT VARYING UTILIZATION LEVELS

Server power consumption responds differently to varying utilization levels. We loosely define utilization as a measure of the application performance—such as requests per second on a Web server—normalized to the performance at peak load levels. Figure 2 shows the power usage of a typical energy-efficient server, normalized to its maximum power, as a function of utilization. Essentially, even an energy-efficient server still consumes about half its full power when doing virtually no work. Servers designed with less attention to energy efficiency often idle at even higher power levels.

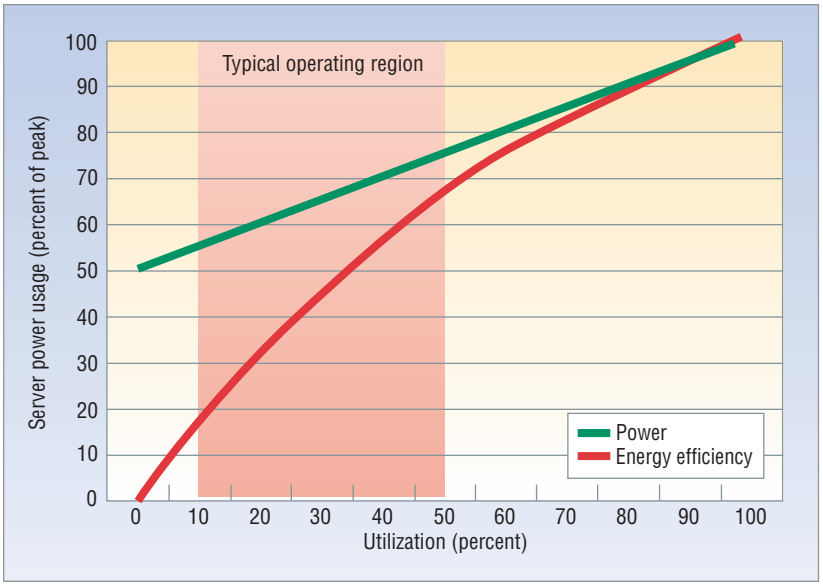

Figure 2. Server power usage and energy efficiency at varying utilization levels, from idle to peak performance. Even an energy-efficient server still consumes about half its full power when doing virtually no work.

Seeing the effect this narrow dynamic power range has on such a system's energy efficiency—represented by the red curve in Figure 2—is both enlightening and discouraging. To derive power efficiency, we simply divide utilization by its corresponding power value. We see that peak energy efficiency occurs at peak utilization and drops quickly as utilization decreases. Notably, energy efficiency in the 20 to 30 percent utilization range—the point at which servers spend most of their time—has dropped to *less than half* the energy efficiency at peak performance. Clearly, such a profile matches poorly with the usage characteristics of server-class applications.

TOWARD ENERGY-PROPORTIONAL MACHINES

Addressing the mismatch between the servers' energy-efficiency characteristics and the behavior of server-class workloads is primarily the responsibility of component and system designers. They should aim to develop machines that consume energy in proportion to the amount of work performed. Such energy-proportional machines would ideally consume no power when idle (easy with inactive power modes), nearly no power when very little work is performed (harder), and gradually more power as the activity level increases (also harder).

Energy-proportional machines would exhibit a wide dynamic power range—a property that might be rare today in computing equipment but is not unprecedented in other domains. Humans, for example, have an average daily energy consumption approaching that of an old personal computer: about 120 W. However, humans at rest can consume as little as 70 W,[8] while being able to sustain peaks of well over 1 kW for tens of minutes, with elite athletes reportedly approaching 2 kW.[9]

Breaking down server power consumption into its main components can be useful in helping to better understand the key challenges for achieving energy proportionality. Figure 3 shows the fraction of total server power consumed by the CPU in two generations of Google servers built in 2005 and 2007.

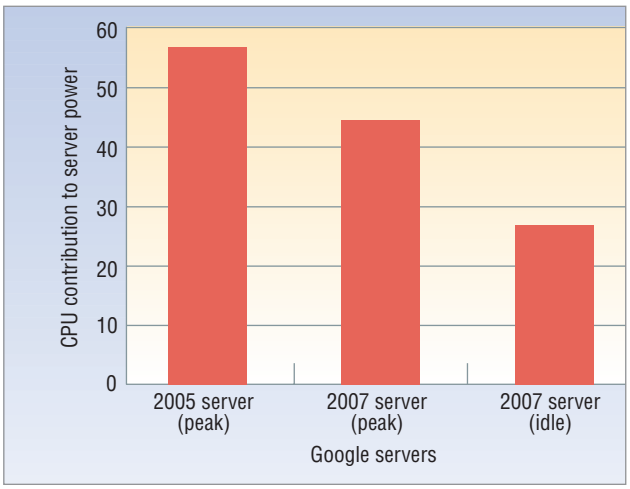

Figure 3. CPU contribution to total server power for two generations of Google servers at peak performance (the first two bars) and for the later generation at idle (the rightmost bar).

The CPU no longer dominates platform power at peak usage in modern servers, and since processors are adopting energy-efficiency techniques more aggressively than other system components, we would expect CPUs to contribute an even smaller fraction of peak power in future systems. Comparing the second and third bars in Figure 3 provides useful insights. In the same platform, the 2007 server, the CPU represents an even smaller fraction of total power when the system

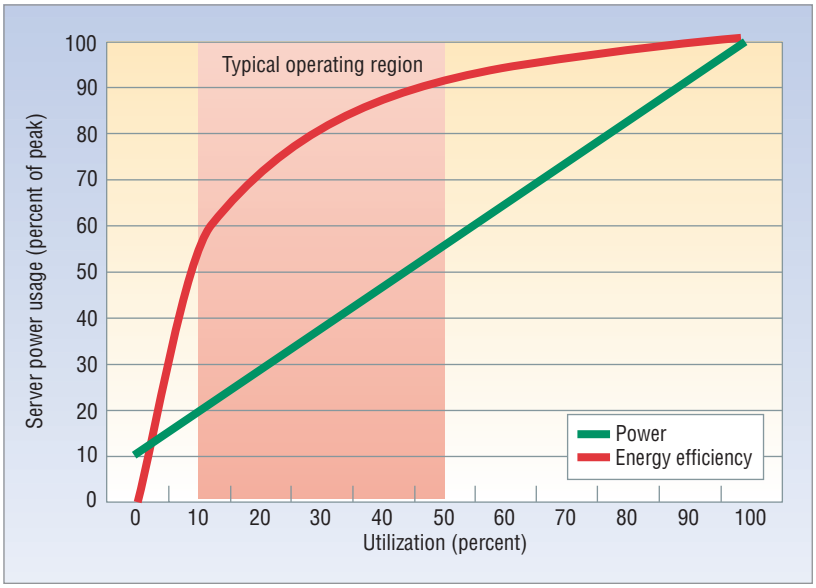

Figure 4. Power usage and energy efficiency in a more energy-proportional server. This server has a power efficiency of more than 80 percent of its peak value for utilizations of 30 percent and above, with efficiency remaining above 50 percent for utilization levels as low as 10 percent.

is idle, suggesting that processors are closer to exhibiting the energy-proportional behavior we seek.

Two key CPU features are particularly useful for achieving energy proportionality and are worthy of imitation by other components.

Wide dynamic power range

Current desktop and server processors can consume less than one-third of their peak power at very-low activity modes, creating a dynamic range of more than 70 percent of peak power. CPUs targeted at the mobile or embedded markets can do even better, with idle power often reaching one-tenth or less of peak power.[10] They achieve this even when not using any performance-impacting—or software-visible—energy-saving modes.

In our experience, the dynamic power range of all other components is much narrower: less than 50 percent for DRAM, 25 percent for disk drives, and 15 percent for networking switches.

Active low-power modes

A processor running at a lower voltage-frequency mode can still execute instructions without requiring a performance-impacting mode transition. It is still active. There are no other components in the system with active low-power modes. Networking equipment rarely offers any low-power modes, and the only low-power modes currently available in mainstream DRAM and disks are fully inactive. That is, using the device requires paying a latency and energy penalty for an inactive-to-active mode transition. Such penalties can significantly degrade the performance of systems idle only at submillisecond time scales.

Compared to today's machines, servers with a dynamic power range of 90 percent, shown in Figure 4, could cut by one-half the energy used in data center operations.[5] They would also lower peak power at the facility level by more than 30 percent, based on simulations of real-world data center workloads. These are dramatic improvements, especially considering that they arise from optimizations that leave peak server power unchanged. The power efficiency curve in Figure 4 fundamentally explains these gains. This server has a power efficiency of more than 80 percent of its peak value for utilizations of 30 percent and above, with efficiency remaining above 50 percent for utilization levels as low as 10 percent.

In addition to its energy-savings potential, energy-proportional hardware could obviate the need for power management software, or at least simplify it substantially, reducing power management to managing utilization.

Fundamentally, the latency and energy penalties incurred to transition to the active state when starting an operation make an inactive energy-savings mode less useful for servers. For example, a disk drive in a spun-down, deep-sleep state might use almost no energy, but a transition to active mode incurs a latency penalty 1,000 times higher than a regular access latency. Spinning up the platters also carriers a large energy penalty. Such a huge activation penalty restricts spin-down modes to situations in which the device will be idle for several minutes; this rarely occurs in servers. On the other hand, inactive energy-savings modes with wake-up penalties of only a small fraction of the regular operations' latency are more likely to benefit the server space, even if their low-energy state operates at relatively higher energy levels than would be possible in deep-sleep modes.

Active energy-savings schemes, by contrast, are useful even when the latency and energy penalties to transition to a high-performance mode are significant. Since active modes are operational, systems can remain in low-energy states for as long as they remain below certain load thresholds. Given that periods of low activity are more common and longer than periods of full idleness, the overheads of transitioning between active energy-savings modes amortize more effectively.

Servers and desktop computers benefit from much of the energy-efficiency research and development that was initially driven by mobile devices' needs. However, unlike mobile devices, which idle for long

periods, servers spend most of their time at moderate utilizations of 10 to 50 percent and exhibit poor efficiency at these levels. Energy-proportional computers would enable large additional energy savings, potentially doubling the efficiency of a typical server. Some CPUs already exhibit reasonably energy-proportional profiles, but most other server components do not.

We need significant improvements in memory and disk subsystems, as these components are responsible for an increasing fraction of the system energy usage. Developers should make better energy proportionality a primary design objective for future components and systems. To this end, we urge energy-efficiency benchmark developers to report measurements at nonpeak activity levels for a more complete characterization of a system's energy behavior.

Acknowledgments

We thank Xiaobo Fan and Wolf-Dietrich Weber for coauthoring the power provisioning study that motivated this work, and Catherine Warner for her comments on the manuscript.

References

1. US Environmental Protection Agency, "Report to Congress on Server and Data Center Energy Efficiency: Public Law 109-431"; www.energystar.gov/ia/partners/prod_development/downloads/EPA_Datacenter_Report_Congress_Final1.pdf.
2. J.G. Koomey, "Estimating Total Power Consumption by Servers in the U.S. and the World"; http://enterprise.amd.com/Downloads/svrpwrusecompletefinal.pdf.
3. L.A. Barroso, "The Price of Performance: An Economic Case for Chip Multiprocessing," *ACM Queue*, Sept. 2005, pp. 48-53.
4. J. Hill et al., "System Architecture Directions for Networked Sensors," *Proc. SIGOPS Oper. Syst. Rev.*, ACM Press, vol. 34, no. 5, 2000, pp. 93-104.
5. X. Fan, W.-D. Weber, and L.A. Barroso, "Power Provisioning for a Warehouse-Sized Computer"; http://research.google.com/archive/power_provisioning.pdf.
6. S. Ghemawat, H. Gobioff and S.-T. Leung, "The Google File System"; www.cs.rochester.edu/meetings/sosp2003/papers/p125-ghemawat.pdf.
7. S. Siddha, V. Pallipadi, and A. Van De Ven, "Getting Maximum Mileage Out of Tickless," *Proc. 2007 Linux Symp.*, 2007, pp. 201-208.
8. E. Ravussin et al., "Determinants of 24-Hour Energy Expenditure in Man: Methods and Results Using a Respiratory Chamber"; http://www.pubmedcentral.nih.gov/picrender.fcgi?artid=423919&blobtype=pdf.
9. E.F. Coyle, "Improved Muscular Efficiency Displayed as Tour de France Champion Matures"; http://jap.physiology.org/cgi/reprint/98/6/2191.
10. Z. Chen et al., "A 25W(max) SoC with Dual 2GHz Power Cores and Integrated Memory and I/O Subsystems"; www.pasemi.com/downloads/PA_Semi_ISSCC_2007.pdf.

Luiz André Barroso *is a distinguished engineer at Google. His interests range from distributed system software infrastructure to the design of Google's computing platform. Barroso received a PhD in computer engineering from the University of Southern California. Contact him at luiz@google.com.*

Urs Hölzle *is the senior vice president of operations at Google and a Google Fellow. His interests include large-scale clusters, cluster networking, Internet performance, and data center design. Hölzle received a PhD in computer science from Stanford University. Contact him at urs@google.com.*