Energy Conservation for Servers

P. Bohrer, D. Cohn, E. Elnozahy, T. Keller, M. Kistler, C. Lefurgy, R. Rajamony, F. Rawson, E. Van Hensbergen
IBM Austin Research Lab
mootaz@us.ibm.com

Introduction

Power management has been thoroughly studied for applications running on battery-powered platforms [3,5]. We take the position that power management is equally important for server environments, where high performance and reliability have traditionally been the most important design and evaluation factors. This position breaks with tradition and argues for considering energy consumption on the same footing. We base our opinion on several observed trends in the technology and the marketplace.

Technology trends for server-class processors predict ever increasing performance at the expense of a rising energy budget [4]. Recent advances also have allowed manufacturers to pack and integrate unprecedented numbers of transistors on a chip, with a corresponding increase in power consumption and cooling requirements. These technology trends coincide with a growing trend in the marketplace to “consolidate” computing services into large data centers, which use the economy of scale to amortize ownership and management costs over a large number of machines. A typical data center thus deploys hundreds or thousands of computers densely packed to maximize floor space utilization, providing the customer with a more cost-effective approach than the alternative of operating the same services in-house. Server consolidation, however, pushes the limits of power supply and cooling systems in a data center. Anecdotal evidence from data center operators already points out to the large contribution of power consumption and cooling to operation costs, and the intermittent failures of computing nodes in densely packed systems due to insufficient cooling.

Given these trends, the case for reducing the energy consumption in servers becomes clear: Energy-efficient servers can be packed in denser configurations than inefficient ones, leading to better utilization of floor space, lower energy bills and cooling requirements. These factors thus have a direct financial impact on the successful operation of a data center. . Additionally, by reducing the power requirements for data centers, energy-efficient servers reduce the need for new power generation and distribution facilities and allow the more rapid deployment of servers without waiting for utility company upgrades.

Server Workloads and Power Management

Servers are typically configured with sufficient processing capacity to handle the expected *maximum* workload. Additional capacity may also be planned in clustered environments for high availability, where it may be necessary to distribute the load of a failed server over the functioning ones. But *actual* workloads vary widely depending on the time of the day, time of the year, and nature of the application among other factors. For example, a Web server that provides banking services will likely be idle during nighttime, while a file server farm in an academic campus tends to be busy until the early morning hours. Likewise, retail Web sites in the western world experience their maximum load only during the last two months of the year. Previous work corroborates these observations[2].

The variability in the actual workload in any system creates opportunities for conserving power. In battery-powered platforms, for example, the display and disk are typically turned off after a specified period of inactivity, and the entire system can be put in hybernation mode if desired [1]. In server workloads, however, it is not clear whether the knowledge about power management that we have gained from client systems will be applicable. For instance, servers do not typically deal with workloads that include substantial interactions with only one individual like client systems, where the individual is ready to tolerate some loss in performance in return for longer operation out of the battery. Additionally, servers are more complex systems generally with multiple disks and often multiple processors or network interfaces, which may require per device power management for the optimal trade-off between performance and power reduction.

Many aspects of power management in client systems are thus not applicable to servers. For example, there is a large latency in state transitions between different modes of operation[1], which could result in unacceptable response time if a server receives a request and has to move from hybernation mode to maximum speed operation. Similar arguments could be made for the storage system. These and other factors make the case that new research is necessary both in hardware and software to accommodate the needs of servers and their expected workloads.

New Ideas for Server Power Management

Voltage and frequency scaling can be used to control the performance of the processor. This can be exploited in two ways. A server can reduce its operating frequncy under software control during periods of low load, so that it executes the requests at a low speed provided that the response time remains within acceptable limits. Alternatively, a server can adjust its mode of operation to match the workload, operating at maximum speed whenever it receives a request, and enters a low-power idle state in between requests. This alternative ensures the minimum response time that the server can provide. Both alternatives will need some hardware support to enable significant energy conservation during idle time and fast state transitions between different power consumption operating modes. On multiprocessor servers an additional set of choices arise including doing voltage and frequency scaling to different degrees on different processors or running with a subset of the processors when the load is light enough that the performance requirements can be met without using all of them.

New storage structures can also be designed to reduce power consumption. In particular, disk arrays typically used for performance and high availability can be modified to take power consumption into account. For example, a straightforward approach would be to replicate or stripe popular files only on a subset of the disks that will be turned on most of the time. Under this scheme, infrequently used files will be stored on a different subset of disks that will be turned off or slowed down most of the time. An alternative design would vary the number of disks that are turned on depending on the workloads, using more sophisticated ECC codes to enable reading files while some of the disks are intentionally turned off. Other design points may possible in a tradeoff between performance, power consumption, and reliability. These design points will need hardware support also to enable effective energy conservation without compromising disk longevity or performance. Surprisingly, existing disks are not suitable for these purposes. Currently, laptop disks are designed for intermittent operation, where the design is optimized to withstand the mechanical stress resulting from frequent turn on and offs. However, laptop disks have relatively low performance when server workloads are considered, and their mean time to failure (MTTF) is typically an order of magnitude lower than those found in server disks. On the other hand, server disks are designed for long MTTF and high performance, but they cannot withstand the mechanical stresses of power ups or downs.

Other than the processor and disk, areas that could be investigated include whether and how a portion of the system memory can be turned off when the workload can run acceptably, and how best to manage multiple network interfaces, especially ones that are being used to increase connection bandwidth to reduce power without unduely impacting performance.

Power management can extend beyond the server boundary in cluster-based environments. Cluster-based power management can play a role in dense server configurations, allowing groups of servers to be turned off if the aggregate workload permits. This requires clever power-aware request distribution (PARD) to reduce energy consumption while retaining performance and reliability levels that are expected from servers. Many of these ideas are currently being investigated in our lab.

Challenges Ahead

There are several challenges that must be overcome through research and other means to realize the vision that we outline in this paper. We cite a few:

Cultural Hurdles: Our experience so far has shown that acknowledging the importance of energy conservation in servers faces some major cultural obstacles. System administrators responsible for dense server configurations understand the need for power management and acknowledge the severity of the problem. The problem has been brought in the mainstream news lately given the energy shortages in California. But outside this group, there is no awareness of the problem or the need for solving it. Servers are specified using metrics describing their performance, functionality and reliability. Examples of such specifications include the TPC benchmark suite for commercial applications and the SPEC suite for scientific ones. None of these benchmarks includes considerations for power consumption. In fact, few servers specify their actual power consumption, reporting only the maximum power that the server's power supply can deliver. The same is true for the component parts used in server construction: the maximum power requirement is specified by the manufacturer but the typical power profile is unknown. This makes it difficult to configure servers with power consumption as a factor. Also, designers of server software still haven't considered power consumption as a factor, and performance optimization remains the most important goal.

Technology and Inadequate Standards: Development of power-aware software for servers will be hampered in the near future by the lack of hardware support for power management in server platforms. Server processors and disks are designed for performance and reliability. Power management thus is not an option on many servers. And even if one configures a high-end client machine as a server, there is little support that can be obtained from the power management features that may be available. The existing standard for power management[1] is designed primarily for battery-powered platforms, which we argue have different requirements than servers. Bridging the gap between server and client platforms will require moving some of the existing technology aimed for portable machines into the server area (e.g. voltage and frequency scaling in processors), and developing new technology to meet the needs of servers (e.g. high-performance disks that have both a high MTTF and a large rating for power cycling).

Legacy Applications and Software: Server software tends to evolve slowly, and because of the low-volume, the cost of development tends to be high. Therefore, it is understandable that data centers will need to preserve the existing investments made in software. Therefore, application profiling tools and operating system support will be necessary to accommodate the existing investment in server software, so as to preserve the existing software infrastructure.

Performance Impact: More research is necessary to understand the relationship between power consumption and the system's overall performance (and not just the processor). Currently, the variation of processor performance with the consumed energy has been well studied. But server workloads require a lot of interactions between the processor and other system components, impacting performance in intricate ways. Similar understanding of the combined effects of these interactions with power consumption and performance need to be understood.

Concluding Remarks

Over the last several years, the dominant focus of power management research has been on portable and handheld devices. In this paper, we have presented some issues conserving energy for servers. Servers are designed to operate at a fraction of their peak operating capacity, and this over-engineering creates opportunities for energy conservation. We believe that exploiting these opportunities will be most needed in data centers where dense packaging lead to severe constraints on cooling and electricity delivery. We predict that technology, business and social trends will intensify the emerging need for energy-efficient servers, and we therefore suggest that a fundamental change is necessary in the way we design and configure servers today. Power management features commonly found in processors intended for mobile computing should be adapted and incorporated as standard features for server processors. Indeed, the performance-centric view of designing servers today must give way to a more balanced view in which energy consumption is as important as other goals of the system.

Bibliography

[1] Compaq et al., *Advanced Configuration and Power Interface Specificatio*n, version 2.0, 2000. Available at http://www.teleport.com/~acpi/.

[2] M. Crovella and A. Bestavros, "Self-Similarity in World Wide Web Traffic Evidence and Possible Causes", *Proceedings of the 1996 ACM SIGMETRICS Intl. Conferenc*e, 1996.

[3] F. Douglis, P. Krishnan, B. Bershad, " Adaptive Disk Spin-down Policies for Mobile Computers", *Proceedings of the 2nd USENIX Symposium on Mobile and Location-Independent Computin*g, April 1995.

[4] Intel Corp., *Pentium III Technical Specification*s, 2000. Available at http://www.intel.com/design/Pentium III/datashts/24445208.pdf

[5] J. Lorch and A. J. Smith, "Software strategies for portable computer energy management," *IEEE Personal Communications Magazin*e, 5(3):60–73, June 1998.