

SeaMicro Technology Overview

Anil Rao, May 2010
anil@seamicro.com, www.seamicro.com

Overview

Data centers consume 1.5 percent of the total electricity in the United States. In the data center, 68 percent of the power consumed by IT infrastructure is consumed by volume servers. Volume servers use seven times as much power as the second leading consumer of power: data-networking equipment. Between 2000 and 2006, the power used by volume servers in the nation's data centers more than doubled.¹

SeaMicro has developed a revolutionary server that reduces the power and space used by volume servers by 75 percent when working on the most prevalent workloads in the data center.

SeaMicro reconceived the volume server as an ultra high density low power single box cluster, a "rack in a box." The system is built around a parallel array of 512 independent ultra low power Intel Atom processors and has a programmable traffic controller front end. The system integrates into a single box—compute, switching, server management, and load balancing.

Three primary technical innovations are key.

- SeaMicro invented and patented technology called CPU I/O virtualization, which dramatically reduces the power draw of the non-CPU portion of a server by eliminating 90 percent of the components from the motherboard. CPU I/O virtualization allows SeaMicro to shrink the mother board to the size of a credit card thereby enabling hundreds of more power efficient CPUs to replace traditional power hungry multi-core processors.
- SeaMicro also designed a super computer style interconnected fabric that can link 512 of these credit-card-sized motherboards into a single system with an order-of-magnitude reduction in power draw. SeaMicro developed dynamic routing algorithms to move traffic across the fabric to avoid congestion and to route around failure. This fabric provides an order of magnitude reduction in power over today's communication technologies while providing lower latency, lower cost and more bandwidth. The fabric enables the SeaMicro architecture to support any CPU instruction set and protocol—Ethernet, fibre channel, data center Ethernet, etc.
- SeaMicro invented technology that combines CPU management and load balancing, allowing us to dynamically allocate workloads to specific CPUs on the basis of power-usage metrics. This ensures that the active CPUs operate in the most energy efficient range of utilization. In addition it allows the user to create pools of CPUs for a given application and can then dynamically add compute resources to the pool based on predefined utilization thresholds.

¹ 2007, EPA Report to Congress on Server and Data Center Efficiency, Public Law 109-431.

Background

The power consumed by servers is a major issue at the micro and macro economic levels. For the data center owner, power as it is often the single largest single Operating Expense line item accounting for more than 30 percent of spend. At the national level, the problem is just as severe as detailed in the 2007 EPA Report to Congress on Server and Data Center Efficiency, Public Law 109-431.

The energy used by the nation's servers and data centers is significant. It is estimated that this sector consumed about 61 billion kilowatt-hours (kWh) in 2006 (1.5 percent of total U.S. electricity consumption) for a total electricity cost of about \$4.5 billion.

One type of server—the volume server—was responsible for the majority (68 percent) of the electricity consumed by IT equipment in data centers in 2006. The energy used by this type of server more than doubled from 2000 to 2006, which was the largest increase among different types of servers. The power and cooling infrastructure that supports IT equipment in data centers also uses significant energy as well, accounting for 50 percent of the total consumption of data centers. Among the different types of data centers, more than one-third (38 percent) of electricity use is attributable to the nation's largest (i.e., enterprise-class) and most rapidly growing data centers.

Designed to replace volume servers, the SeaMicro compute appliance delivers significant advantages in power, size, and total cost of ownership.

- Power efficiency: The SeaMicro appliance reduces by 75 percent the power used by the best-in-class volume server when working on the primary workload in the data center.
- Size: The SeaMicro compute appliance takes one-quarter of the space to do the same work as the best-in-class volume servers.
- Total Cost of Ownership: The SeaMicro compute appliance is less costly to buy, own operate, and manage, than any server currently on the market.

To begin, SeaMicro made several key observations about changes in the computational landscape and weaknesses in the existing architecture of volume servers.

First, SeaMicro recognized that in the data center, and in particular for volume servers in the data center, the computational requirements had significantly diverged from other parts of the computational landscape. Historically, the challenge in computing was increasingly taxing workloads. Indeed, this led the makers of CPUs to focus on "single-thread performance"—roughly speaking, the ability to speed up work on a single hard problem. But with the rise of the Internet and the rapid growth of the data center, this has changed. Rather than relatively few complex problems, the computational challenge, in the part of the data center dominated by volume servers, became how to handle the huge volume of relatively

modest computational workloads. These workloads are generated by the millions of independent users each wanting to search, view web pages, check email, and read the news, all for free.

With companies such as Yahoo!, Google, Facebook, Walmart, New York Times, and others growing to hundreds of millions of users, the challenge for the volume server in their primary application became “how to handle small computational workloads at a scale that has never before been encountered?” While this challenge appeared first in Internet sites, it now has come to dominate corporate data centers, as large enterprises increasingly provide services to their customers and employees over the Internet.

The exponential growth of this workload produces a challenge for today’s volume servers: their CPUs are not well matched for the workload that has come to dominate the data center. Volume servers have historically used large, complex, high speed multi-core CPUs. These CPUs are designed for complex computational challenges and, in terms of computation per dollar, are best-in-class at complex workloads. But these same large, complex, high speed multi-core CPUs are particularly inefficient at small simple computational workloads. Put simply, large, complex, high speed multi-core CPUs are “overkill.” Indeed, the mismatch between the CPU in volume servers and the primary workload in the data center is a fundamental underlying cause of the rapid rise in power consumption by volume servers.

A second driving force underpinning the power issue in the data center is the packaging of the volume servers themselves. Volume servers are also called “rack servers” because they are stacked like pizza boxes one on top of another in metal racks. These servers are usually 1.75”–3.5” tall, 19” across, and up to 30” deep. Each server is a discrete unit—it is individually managed, controlled, powered, and cooled. These individual servers are linked together with Gigabit Ethernet switches and routers. Herein lies the second source of the power issue in the data center—massive replication of components—as each server has the overhead necessary for individual operation, management, and connectivity but is never used in isolation. It is as if a train had been constructed with each train car having its own engine and its own caboose and then linked together to form the train. Obviously, we do not move freight that way. Rather we use one engine and one caboose, and we amortize these “overhead costs” over hundreds of train cars.

To summarize, the power issue in the data center was born of the confluence of several factors:

- (a) the rise to dominance of a new and different workload;
- (b) the fact that the CPUs that had historically been the data-center workhorse were suddenly ill-suited for the predominant and fastest-growing workload;
- (c) the inherent inefficiency in server packaging that left enormous duplication in components.

The SeaMicro compute appliance systematically addresses each of these factors to better ensure that the computation delivered is in complete alignment with the most prevalent workload. To execute on this architecture, SeaMicro innovations span multiple domains: communication networks, supercomputer design, routing, ASIC design, and motherboard design, to name a few.

Technology

The observations previously described laid the groundwork for the SeaMicro solution. These observations also shed light on the

breadth of the innovations required in order to successfully research, develop, and bring to market this solution.

First, SeaMicro identified a more efficient CPU, one that was better aligned to handle the most common workloads in the data center. SeaMicro then developed technologies that enable these new low-power CPUs to replace the large, complex, high speed multicore CPUs in volume-server applications. This meant rethinking the architecture of the server and reconstituting it as a delivery system for low-power CPU cycles.

Second, SeaMicro developed technology that links hundreds of these CPUs together so that the overhead costs of management and connectivity were amortized over hundreds of CPUs rather than two or four as had been done traditionally—thereby eliminating the replication of components inherent in clusters of volume servers.

Third, SeaMicro developed software and integrated circuit technologies that would allocate load dynamically across its array of parallel-processing CPUs, in order to minimize power usage by ensuring that the CPUs were operating in their maximally efficient range (CPUs, like many electrical systems, are particularly inefficient at low utilization).

For the first challenge, identifying the most efficient CPUs in terms of computation per watt for the most common workload in the data center, it turns out that the most efficient CPUs were not intended for servers at all. Rather, they were designed for handheld devices and the smallest of laptop computers; they are simpler designs, they use less power, and they are significantly more power-efficient at web-tier workloads. In fact, these CPUs offer more than three times the performance per watt of the large multi-core CPUs for this workload. To be more specific, they provide half the single-thread performance for a sixth of the power draw, a dramatic improvement in computation per unit of power. It is important to note that these CPUs are smaller and slower and do less work per CPU in absolute terms, but *for the less demanding computation needs that dominate the Internet data center, they offer fundamental improvements in computation per unit of power.* Although smaller, simpler, and slower often means better performance per unit of energy, it also presented substantial engineering challenges at the system level.

For example, when these low-power CPUs are placed into the existing volume-server architecture, the power consumed to do a unit of work actually increases. The CPU in a server uses approximately a third of the total power consumed by the server. As a thought experiment, assume a magical CPU that delivers half the performance of a traditional CPU but uses no power at all. If that CPU were placed into an existing architecture, then the new server would offer half the performance, for two-thirds the power. In other words, it would deliver lower performance per watt than did the original server. In some sense, this is not surprising. In the 70’s and early 80’s, car makers did not take smaller, more-efficient motors and put them in the same cars that had historically struggled with gas mileage. Rather, they had to rethink the rest of the car as well, and make it smaller, lighter, etc., in order to be well matched to the more efficient motor.

It is the interaction between power reduction in the CPU and the power reduction in the rest of the system that produces the dramatic gains in computation per watt that SeaMicro has been able to achieve. The key observation is that if one wants to use these small, more-efficient CPUs, then technology must be developed that reduces the non-CPU “two-thirds of the total power” drawn by the server. That is, one needs to scale the reduction in power draw from

the CPU across all components. SeaMicro accomplishes this through its technological breakthrough called hardware-based CPU I/O virtualization.

Hardware-based CPU I/O virtualization technology. SeaMicro has developed I/O (input/output) virtualization technology that removes all non-CPU/memory components from the motherboard, while allowing the CPUs to run standard operating systems and software without requiring modification or recompilation. The CPU I/O virtualization technology also allows common components to be shared across hundreds of CPUs, rather than being duplicated on each motherboard. Thus, components such as basic input/output system (BIOS), external network access, storage, and console are instantiated once and then are amortized over the entire system. Hardware-based CPU I/O virtualization enables SeaMicro to eliminate 90 percent of the components from the server and to shrink the server to the size of a credit card.

Supercomputer-style interconnected fabric. Once the server had been shrunk to the size of a credit card, technology needed to be developed that could link together hundreds of these card-sized computational nodes, with significant reductions in power, cost, and latency. For answers, SeaMicro looked to the techniques used to interconnect the CPUs of the largest and most complicated supercomputers and set about scaling the technology down for data-center applications. The result is the SeaMicro interconnect fabric, which ties together 512 computational nodes. It is a three-dimensional torus, with both path redundancy and diversity. The fabric is FLIT-based and wormhole-routed, with integrated virtual-channel technology to manage congestion, and has a throughput of 1.2 Terabits. These technologies combine to produce a low-latency, high-bandwidth, redundant fabric at very low cost. While the fabric has its origins in the supercomputer world, SeaMicro tuned the design of the fabric, and optimized it for the requirements of the data center.

Dynamic Compute Allocation Technology™ (DCAT). DCAT adds a layer of intelligence to the distribution of work to the 512 CPUs by combining CPU management and stateful load balancing. The SeaMicro management software is aware of the health of, and the workload on, all the CPUs in the system. Based on this knowledge, the management software can transparently provision the CPUs and dynamically program the field programmable gate arrays (FPGAs) to direct traffic to one group of CPUs and away from another. The technology works by creating Virtual IP addresses which can be assigned to pools of compute as small as one server and as large as 512 servers. The stateful hardware load balancer then distributes flows across these pools of servers using various load-management algorithms including round-robin, least connections, and max connections to enable the internal servers to be used effectively at the most optimal power consumption levels without degradation of end user performance. The pools of compute are accessed as if they were a single CPU. CPUs can be added or removed from a Virtual IP pool dynamically based on predetermined rules. So for example, traffic can be directed to a pool of CPUs to ensure they are operating in the maximally efficient range, while allowing other CPUs to enter deep sleep mode or even to be turned off. Similarly, a utilization threshold for a pool of compute can be set, and if met, CPUs can be dynamically provisioned and added or removed from the pool.

ASIC and custom silicon. SeaMicro instantiates its fabric and CPU I/O virtualization technology in custom application-specific inte-

grated circuits (ASICs), which are paired with each CPU. External I/O and storage-system components are implemented in custom-designed FPGAs. The FPGAs implement both the SeaMicro fabric technology and standards-based I/O technology. The fabric side of the FPGAs connects to the ASICs to form the fabric, while the standards-based I/O side connects to external interfaces providing standards-based interfaces for the network and storage cards.

To summarize, hardware-based CPU I/O virtualization enables SeaMicro to eliminate 90 percent of the components from the server and to shrink the motherboard to the size of a credit card. Hundreds of these low-power, card-sized computational units are tied together with a supercomputer-style fabric to create a massive array of independent but linked computational units. Work is then distributed over these hundreds of CPUs via hardware- and software-based load-balancing technology that dynamically directs load to ensure that each of the CPUs is either in its most efficient zone of performance or is sleeping. The key technologies reside in three chips of SeaMicro's design, one ASIC and two FPGAs, and in the management, routing, and load-balancing software that directs traffic across the fabric.

This combination of technologies produces a 75 percent reduction in power usage and space for the most common workload in the data center.

Recommended Reading

Report to Congress on Server and Data Center Energy Efficiency, Public Law 109-431, U.S. Environmental Protection Agency ENERGY STAR Program, August 2, 2007.