cadence®

Reducing Datacenter Energy Usage Via Power-Saving IP and System Design Techniques

Virtualization and PCI Express updates contribute to lower power

By Arif Khan and Osman Javed, Cadence Design Systems

As the modern world becomes increasingly connected, businesses and consumers alike are relying more and more on digital data—from documents to video to audio. Behind the scenes, datacenters that manage all of this digital data are a somewhat silent, yet impactful, part of this connectivity revolution. These centers are lined with servers that process digital data for everything from social media status updates to web searches to cloud computing. Each of these servers requires energy to run as well as energy to keep them cool, and the centers are typically filled with additional machines for redundancy. In the US alone, datacenters consume anywhere from 1.5% to 3% of the country's energy production. How can the chip design industry make datacenters more energy efficient? This paper examines system techniques, including virtualization, as well as design considerations and protocol improvements that can lower the energy utilization of datacenters.

Contents

Introduction

In a modern world where mobility, cloud computing, and the Internet of Things (IoT) are becoming increasingly pervasive, businesses are under pressure to increase the energy efficiency of the datacenters that power these digital data-driven technologies. In 2012, the New York Times conducted a year-long study that showed how "the information industry is sharply at odds with its image of sleek efficiency and environmental friendliness (Glanz, 2012)."

There's an overwhelming growth in data—computed, stored, and transferred. IT infrastructure and cloud computing are growing in response to this. Studies have revealed that servers, storage, and networking equipment—the core elements of datacenters—typically run in low-utilization conditions, where they operate at an inefficient ratio of computing performance to energy consumption.

From the perspective of the chips that process all of this data, there are a number of techniques available to increase energy efficiency. One is virtualization to enhance system utilization, though it may not be the most sustainable option. Other techniques include design considerations and protocol improvements via key interface intellectual property (IP) components used in servers, such as PCI Express.

Increasing Traffic and Workloads Taxing Datacenters

According to the New York Times study, worldwide, datacenters—which numbered more than three million at the time of the 2012 examination—use about 30 billion watts of electricity (Glanz, 2012). Datacenters continue to grow in number and performance level because of the push of new applications, which call for faster delivery of larger amounts of data. According to a vivid infographic posted by blogger Josh James at Domo.com (James, 2012), nearly 700,000 Facebook updates occur each minute and Google receives over 2 million queries in the same period of time. Mobile and video applications

are amongst the biggest drivers of Internet data (Cisco Systems, 2013). Consider this comparison: in 1992, daily Internet traffic was in the 100GB range; in 2012, more than 12,000GB of data cross the Internet every second, and this number is projected to triple by 2017.

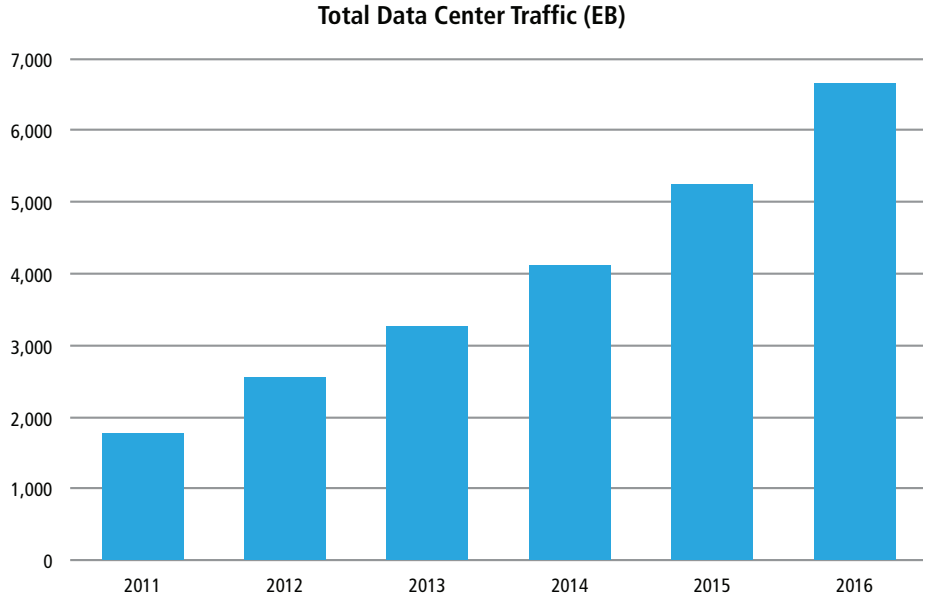

Figure 1: Cisco Systems forecasts that annual datacenter traffic will reach 6.6 Zettabytes (1021 bytes) by the end of 2016 with a CAGR of 31% from 2011 to 2016. The chart uses Exabytes (1 Exabyte = 1018 bytes) on its Y-axis (Cisco Systems Global Cloud Index).

Currently, the trend is for traditional datacenters to move to cloud datacenters, which provide application and data access on multiple platforms to increasingly mobile business users and consumers. At the same time, the amount of digital data being stored has gone up tremendously because storage costs are dropping, providers are offering "unlimited" email services, and there are an array of cloud-based solutions that depend on stored data.

Experts estimate that by 2014, two-thirds of traffic will be cloud-based, displacing traditional datacenter traffic. While compute workloads on traditional servers will grow by approximately 33%, virtualized workloads on cloud servers are projected to increase by 200% (Cisco Systems, 2012). The Cisco Systems study also finds that nearly three-fourths of datacenter traffic is internal to the datacenter, inter-server, or between servers and storage.

The Energy Drain

Datacenters and cloud computing facilities are typically designed to handle peak usage that is unpredictable in nature. For example, while breaking news can bring servers down, e-tailers build out capacity to handle the peak loads of infrequent events such as Cyber Monday. In addition, businesses need to guard against downtime when a power failure occurs; because of this, many datacenters use diesel backups that are significant polluters. Less than 20% of datacenter energy is used by active servers, according to published research (Glanz, 2012), (Kay, 2012). Much energy is spent to cool the servers and also consumed while the machines are idle.

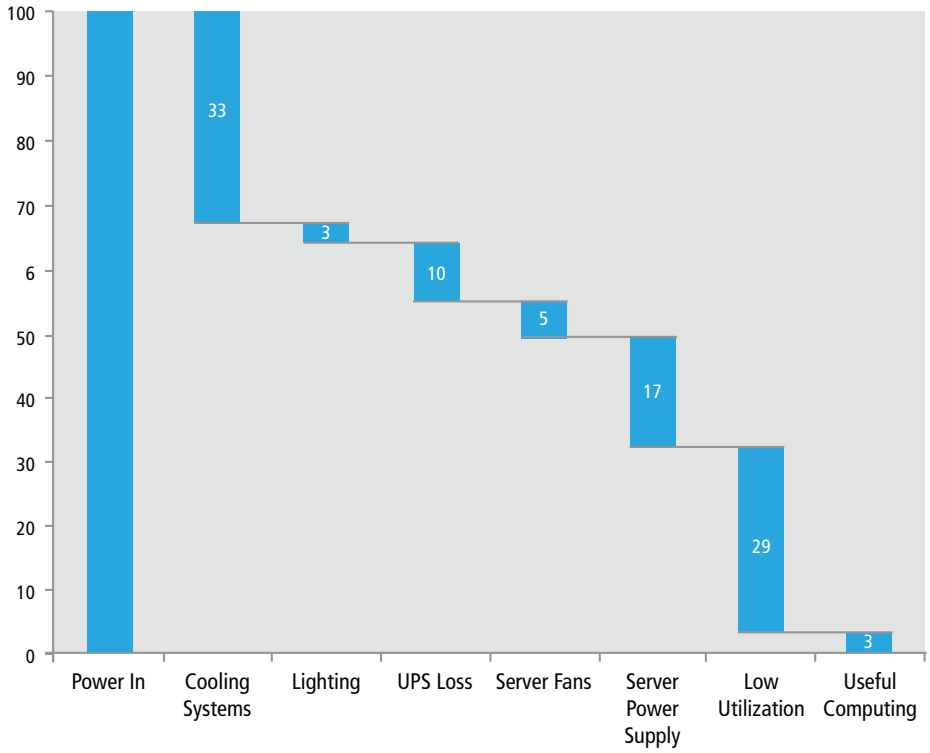

Figure 2: Energy consumption waterfall (Rocky Mountain Institute of Technology, 2008). Less than 3% of the energy consumed in a datacenter is used for active, useful computing. Nearly a third is wasted on idle computers and two-thirds spent on physical infrastructure (including server fans).

The waterfall chart in Figure 2 provides insight into areas where energy savings can be found. Savings can result when datacenters are set for higher utilization—more compute activity at the same power consumption level. The same can happen when the datacenter is designed for better power efficiency while in idle states. What's more, these conservation methods can also reduce system cooling costs. As an interesting point, consider the results from an important study by Google (Barroso & Hölzle, 2009), which showed that servers are never really idle; instead, most of the time is spent in the low utilization regions in the 10% to 50% range.

Reducing Energy Waste

To reduce energy waste in servers, designers can tap into software and operating system scheduling to improve utilization through virtualization and batching of compute loads. Systems designers can enhance system energy consumption under low utilization. Improved ASIC designs that enter much lower levels of energy consumption in low utilization conditions can help here.

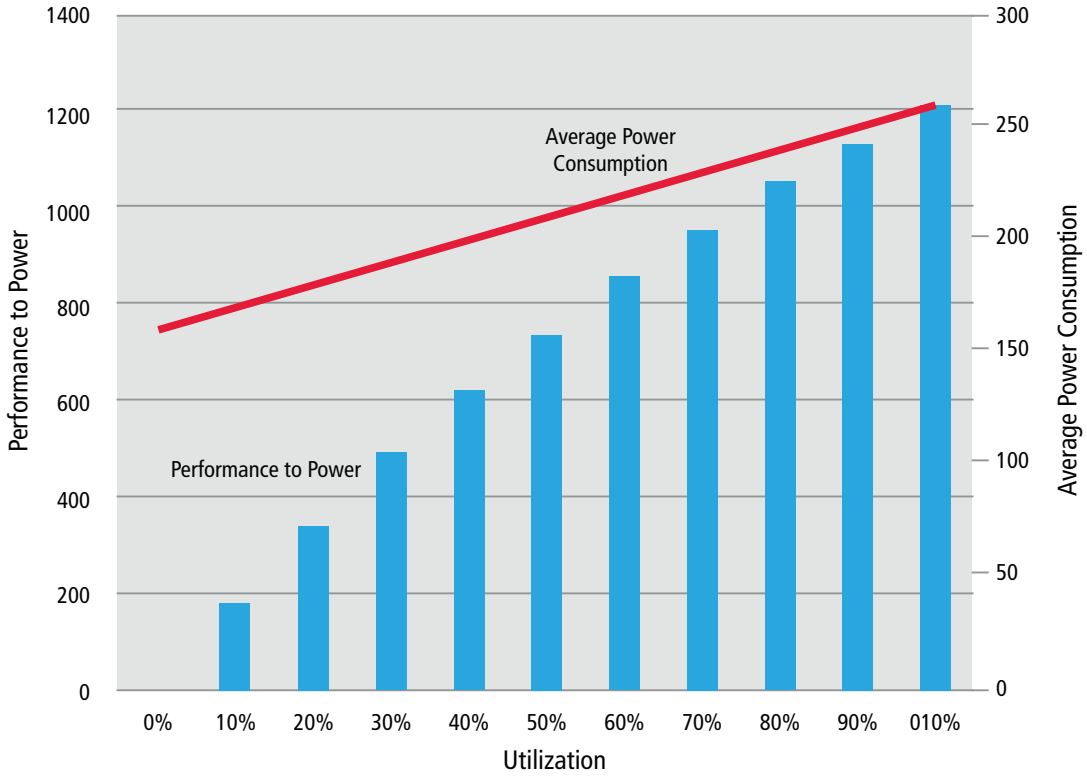

Figure 3: Barroso and Hölzle found that average power declines nearly linearly with utilization, but performance to power (efficiency) degrades much faster. At utilizations of less than 50%, a significant gap exists due to idle power consumption in servers. For optimal use, operate in the region on the right side of the chart.

Improving Utilization

Batch processing has long been a workhorse in the IT world. But newer applications in today's connected world—applications that are commonly interactive and require instant results—do not lend themselves to batch processing. Virtualization—a key driver of cloud computing services—can raise utilization in datacenters. Targeting underutilized servers, virtualization reduces energy waste by allowing a single computer to run multiple "guest" operating systems by abstracting the hardware resources (CPU, memory, I/O) through a hypervisor, also known as a "virtual machine manager".

In a virtualized system (Figure 4), a guest virtual machine (VM) can be migrated from one hardware system to another. If the VM runs out of memory or other resources during a peak usage period, then it can be migrated to an underutilized server. Behind the scenes, hardware and other resources, such as network addresses, must be managed so that the migration is transparent to the VM.

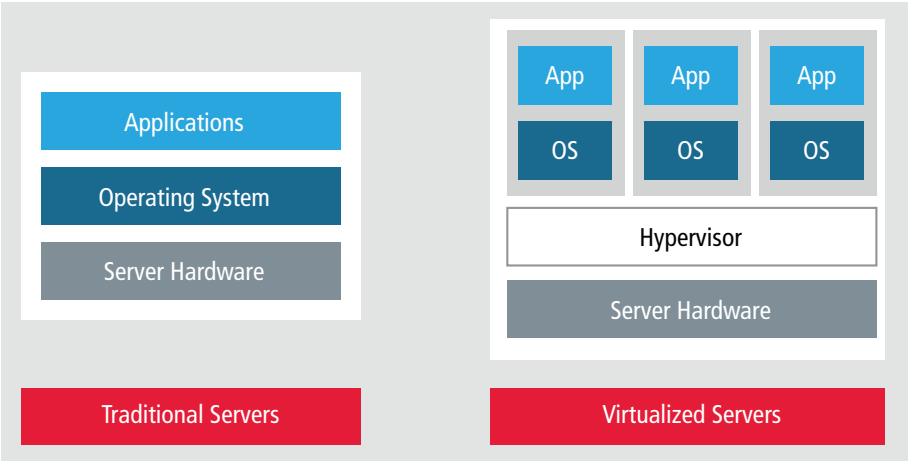

Figure 4: Traditional servers run a single operating system with applications running on it. Virtualized servers allow for increased utilization on a common hardware platform through the use of hypervisors that permit multiple OS instances ("guest OSs") and their associated applications.

Virtualization does increase utilization, but there's still the risk of underutilization at non-peak loads. Another consideration: providers, following the guidelines of strict service-level agreements (SLAs), often have to guarantee performance levels. As a result, datacenter operators maintain redundant equipment to serve as backup in the case of equipment failures. As these machines stand idle, they add to the overall energy consumption. For power efficiency to occur, utilization must be maintained at 70% level or higher (Figure 3). How can we address low utilization states?

Looking to System Solutions

Inside the server, the most power-hungry components are the processor cores, memories, disk drives, and I/O network. As performance and bandwidth demands increase, so too does the complexity of the hardware used in datacenters. With multi-core machines, datacenter operators can increase efficiency of threads/watt and reduce the cost per unit of performance. These designs demand very-high-bandwidth coherency links between sockets. Coherency is required to maintain consistent data between processors. These designs also require suitable bandwidth in the I/O subsystem to feed the inputs and outputs of the system (Ethernet, storage, etc.).

Various techniques have emerged to lower power consumption in datacenter hardware (Meisner, Sadler, Barroso, Weber, & Wenisch). Dynamic voltage frequency scaling (DVFS), which lowers power while the CPU is active, reduces power consumption, though at some performance cost. As process technology evolves and the gap between circuit nominal voltages and threshold voltages shrinks, DVFS advantages will shrink simultaneously. Deep scaling impacts performance significantly.

Within the core, CPU clock gating is an option. However, shared caches and on-chip memory controllers typically remain active as long as any core in the system is active for coherency reasons. Optimizing for a complete system idle state is impractical, since datacenters are rarely in full system idle mode (Figure 4).

There are opportunities to apply active low-power techniques to the memory and I/O subsystem. CPUs have a dynamic range greater than 3.0X (i.e., the power varies 3X over the activity range, making it fairly proportional to the usage). In contrast, the dynamic range for memory is 2.0X and for storage and networking, it's 1.2-1.3X (Barroso and Hölzle).

On the memory side, self-refresh assist can reduce power consumption by an order of magnitude (Meisner, Sadler, Barroso, Weber, & Wenisch). This technique allows DRAM refreshes while the memory bus clock, phase-locked loop (PLL), and DRAM interface circuitry are disabled.

Interface links can consume a substantial amount of power in idle and peak usage modes. Among the interface protocols, PCI Express (PCIe) is a ubiquitous technology used for storage, graphics, networking, and other connectivity applications. The PCI-SIG has actively focused on platform power-reduction enhancements to the specification. These techniques line up with the energy-proportionality concept: to reduce power consumption in response to lower utilization levels.

Power-Reduction Enhancements in PCIe

The PCI-SIG has updated the PCIe spec with engineering change notices (ECNs) that help increase the dynamic range for power consumption on PCIe devices based on activity and utilization. This, in turn, enhances the energy proportionality of systems. Among the recent ECNs are latency tolerance reporting (LTR), optimized buffer flush/fill (OBFF), and further power reduction in the L1 state.

With LTR, a host can manage interrupt service (by scheduling tasks intelligently) to optimize the time it stays in a low-power mode. The host can still service the device within the device's window for tolerating service latency. Platform power management (PM) policies guesstimate when devices are idle, with approximations from inactivity timers. Incorrect estimation of idle windows can cause performance issues, or even hardware failures in extreme cases. As a result, PM settings often result in sub-optimal power savings. To preserve correct functionality, PM is sometimes completely disabled.

Using LTR, the endpoint sends a message to the root complex to indicate its required service latency. The message encompasses values for both snooped and non-snooped transactions. Multi-function devices and switches coalesce LTR messages and send them on to the root port. The LTR ECN also allows endpoints to change their latency tolerance when service requirements change (for example, when sustained burst transfers need to be maintained). In an LTR-enabled system, the endpoints provide actual service intervals to the root complex. The platform PM software can use the reported activity level to gate entry into low-power modes. LTR enables dynamic power versus performance tradeoffs with the low overhead cost of an LTR message.

OBFF allows the host to share system-state information with devices, so that devices can schedule their activity and optimize the time spent in low-power states. In a typical platform, PCIe devices in the system aren't aware of where central resources are in terms of power states. Therefore, the engineer can't optimally manage CPU, root complex, and memory components because device interrupts are asynchronous, fragmenting the idle window. With OBFF (as shown in Figure 5), devices receive power management hints, so they can optimize request patterns because they know when they can interrupt the central system. As a result, the system can expand the idle window and stay in a lower power state longer. OBFF can be implemented with expanded meanings of the WAKE# signal or with a message.

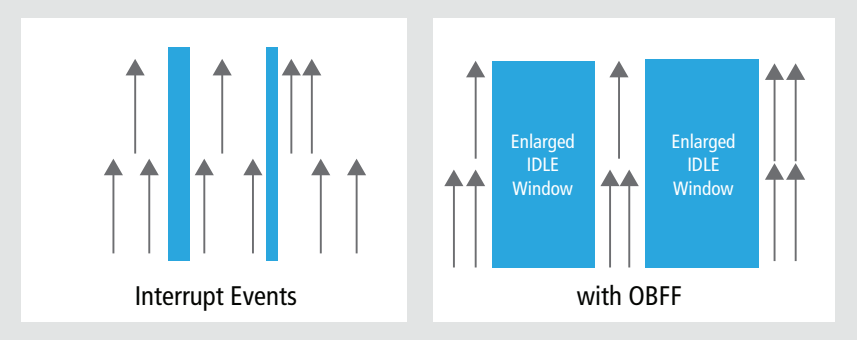

Figure 5: The premise for Optimized Buffer Flush/Fill is that devices could optimize request patterns if they knew when they were allowed to interrupt the central system. This would allow the system to stay in a lower power state longer by expanding the idle window.

The PCIe protocol defines a number of link power states as shown in Table 1.

PCIe Link States	
Link State	Description
L0	Active state
L0s	Stall
L1	Low-power standby
L2	Auxiliary power, deep standby
L3	Off

Table 1: PCIe link states: as power savings increase, exit latency from the low-power state to active rises.

Driven by market and regulatory needs, the PCI-SIG proposed an ECN to lower power consumption in the L1 state. The current low-power state (L1) has power consumption in milliWatts. "L1 Power Mode Substates with CLKREQ" (PCI-SIG, 2012) redefines the L1 state as L1.0 and includes two sub-states defined as L1.1 and L1.2 (Table 2). With these two new states, the standby state can reduce power consumption in its mode. IP providers have already announced implementation of these states in PCIe controllers and PHYs, dropping standby power consumption by two orders of magnitude to low microWatts while providing compelling transition times between active and idle states.

PCIe L1 Substates	
Link State	Description
L1.0	Standby: Rx/Tx off or idel
L1.1	Rx/Tx off, common-mode voltage maintained
L1.2	Rx/Tx off, common-mode voltages off

Table 2: PCIe L1 substates: reducing power more by turning off portions of the analog design. Exit from this state occurs with assertion of CLKREQ#.

The two new substates define lower power states that accomplish this by disabling circuitry not required by the protocol. In both L1.1 and L1.2, detection of electrical idle is not required, and the states are controlled by CLKREQ#. L1.2 reduces power further by turning off link common mode voltages.

Exit latencies from these new low-power modes are critical since system performance can suffer from long latencies. Further, there may be functionality issues if LTSSM timers are not honored correctly. PHY designs are being pushed to the limit to reduce exit latency while providing low current consumption simultaneously.

How IP Design Can Contribute to Lower Power

Applying techniques that provide coverage for process, voltage, and temperature variation can reduce active-mode PCIe PHY power. Without these techniques, PHYs must be designed with greater overhead that increases power consumption significantly. Clock gating and power islands can significantly reduce leakage current, which optimizes static power consumption.

Entering deep low-power states has a deleterious impact on exit times from these states. However, superior PLL design techniques for fast lock times can reduce exit latency significantly, improving the resumption time and user experience.

Cadence has optimized its PCIe controller and PHY to support these new L1 power saving states. The Cadence® implementation of these low-power states is available in x1 to x16 configurations—all applications that use these devices can benefit from the power savings. The 28HPM implementation of the Cadence controller is first to market with a wire-bond PHY option and is also available in flip-chip designs.

Conclusion

To support our digital demands, datacenters will continue to grow, along with the need to reduce the energy consumption of these digital warehouses. Engineering design continues to uncover low-power techniques for semiconductor and system implementations. From an energy proportionality perspective, designs should be optimized for energy efficiency in idle and low-utilization states. Virtualization can support efficient system utilization and operation, though most systems persist in operating in the sub-optimal low-utilization region of the power-performance spectrum. Protocol enhancements such as the PCI-SIG's L1 Power Mode Substates with CLKREQ ECN can be more effective in bringing energy efficiency to low-utilization and idle states—particularly when chip designs are optimized for power, performance, and area. Vendors such as Cadence, with its low-power PCIe PHY, can help turn this ideal into reality.

Works Cited

- Barroso, L. A., & Hölzle, U. (2009). The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines. (M. D. Hill, Ed.) Madison, WI: Morgan & Claypool.
- Cisco Systems. (2012, October). Cisco Global Cloud Index: Forecast and Methodology, 2011–2016. Retrieved June 3, 2013, from Cisco.com: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns1175/Cloud_Index_White_Paper.html

- Cisco Systems. (2013, May 29). The Zettabyte Era—Trends and Analysis. Retrieved June 3, 2013, from Cisco.com: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/VNI_Hyperconnectivity_WP.html
- Glanz, J. (2012, September 22). The Cloud Factories: Power, Pollution, and the Internet. Retrieved June 03, 2013, from The New York Times: http://www.nytimes.com/2012/09/23/technology/data-centers-waste-vast-amounts-of-energy-belying-industry-image.html?pagewanted=all&_r=0
- Hammond, T. (2013, April 8). Toolkit: Calculate datacenter server power usage. Retrieved June 9, 2013, from ZDNet: http://www.zdnet.com/toolkit-calculate-data-center-server-power-usage-7000013699/
- James, J. (2012, June 22). How much data is created every minute? Retrieved June 5, 2012, from domo.com: http://www.domo.com/blog/2012/06/how-much-data-is-created-every-minute/
- Kay, R. (2012, October 18). Taming Datacenter Power Usage. (Endpoint Technologies Associates, Inc) Retrieved June 2, 2013, from Forbes: http://www.forbes.com/sites/rogerkay/2012/10/18/taming-datacenter-power-usage/
- Meisner, D., Sadler, C. M., Barroso, L. A., Weber, W.-D., & Wenisch, T. F. PowerNap: Eliminating Server Idle Power. ISCA 2011. San Jose: ACM.
- PCI-SIG. (2012, August 23). L1 PM Substates with CLKREQ. Retrieved June 03, 2013, from pcisig.com: http://www.pcisig.com/specifications/pciexpress/specifications/ECN_L1_PM_Substates_with_CLKREQ_23_Aug_2012.pdf
- Rocky Mountain Institute of Technology. (2008, August 07). Designing Radically Efficient and Profitable Datacenters. Retrieved June 03, 2013, from treehugger.com: http://www.treehugger.com/gadgets/designing-radically-efficient-and-profitable-data-centers.html