WHITE PAPER

cādence®

Energy-Efficient SoCs for the Zettabyte Era Using Power-Saving IP and System Design Techniques

Leveraging protocol updates and implementation techniques to lower power

By Arif Khan, Cadence

As the modern world becomes increasingly connected, businesses and consumers alike are relying more and more on digital data. Behind the scenes, data centers that manage all of this digital data are a somewhat silent, yet impactful, part of this connectivity revolution. These data centers are lined with servers that process digital data for everything from social media status updates to analytics for scientific research. These servers require energy to run as well as energy to keep them cool, and the data centers are typically filled with additional machines for redundancy. In the US alone, data centers consume anywhere from 1.5% to 3% of the country's energy production. How can the chip design industry make data centers 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 data centers.

Contents

Introduction

Nearly eight years ago when the first version of this white paper was published, video and social media consumption was the primary driver of internet and data center traffic. Since then, advances in artificial intelligence (AI) and high-performance computing (HPC) have made a wide swath of new applications possible. The SARS-CoV-2 genome was mapped in weeks in 2020 and vaccine trials accelerated through the use of technology. There's an overwhelming growth in data—computed, stored, and transferred—and IT infrastructure and cloud computing are growing in response to this.

However, 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."[1] This and other studies have revealed that servers, storage, and networking equipment—the core elements of data centers—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® (PCIe®).

Increasing Traffic and Workloads Taxing Data Centers

According to the New York Times study, data centers—which numbered more than three million worldwide at the time of the 2012 examination—use about 30 billion watts of electricity.[1] As shown in Figure 1, data centers 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[2], Americans use over 3 petabytes of data each minute and Google receives over 3.8 million queries in the same period of time, close to twice the number from 2012.

Cisco Systems' 2020 Global Networking Trends Report[4] highlights major changes impacting information technology worldwide. Globalization, digital transformation, continued business automation, and sustainability are key drivers for new networking architectures. The new landscape encompasses cloud-native applications, internet of things, mobile, AI, immersive (AR/VR) applications, and new cybersecurity threads. Already, 50% of enterprise workloads are outside the traditional data center. By 2022, internet video will represent 82% of all business internet traffic. The Cisco study also highlighted that 72% of traffic is internal to data centers—between servers or between servers and storage, the east/west paradigm.

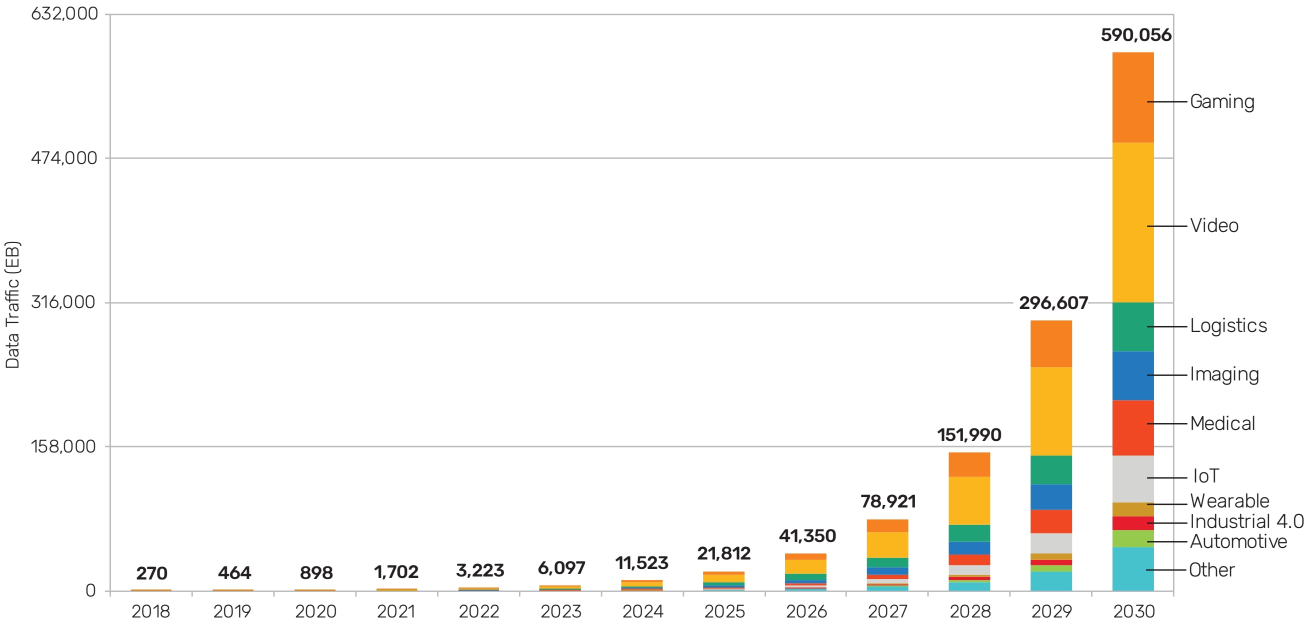

Figure 1: IBS Forecast for Data Traffic. 1EB = 10^{18} bytes[3]

The Energy Drain

Data centers 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 data centers use diesel backups that are significant polluters. Less than 20% of data center energy is used by active servers, according to published research.[1,5] Much energy is spent to cool the servers and also consumed while the machines are idle.

The waterfall chart in Figure 2 provides insight into areas where energy savings can be found. Savings can result when data centers are set for higher utilization—more compute activity at the same power consumption level. The same can happen when the data center 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 Barraso and Hölzle[6] that 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.

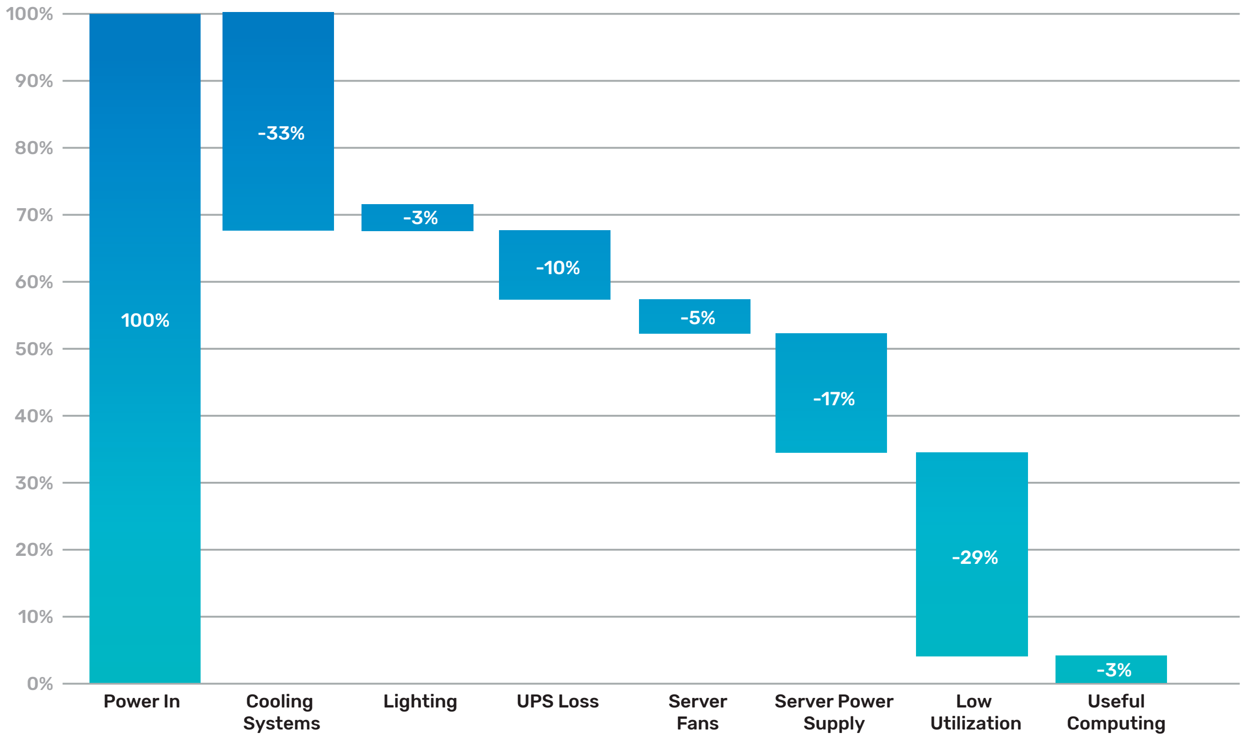

Figure 2: Energy consumption waterfall[7]: Less than 3% of the energy consumed in a data center is used for active, useful computing, nearly a third is wasted on idle computers, and two-thirds spent on physical infrastructure (including server fans)

The 9th Annual Uptime Institute Data Center Survey[8] surveyed 1600 participants and found that the biggest efficiency gains in data centers occurred in 2013-14, have flattened out since then, and to some degree even degraded in 2018-19, relying on further efficiency gains to come from core information technology (systems and software). The Power Usage Effectiveness as measured in this study was 1.67 (ideal = 1.0).

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 (Figure 3). Improved ASIC designs that enter much lower levels of energy consumption in low utilization conditions can help here. Recent work by Sen and Wood[9] expands on Barraso and Hölzle's[6] work demonstrating techniques to obtain power optimality in modern systems with reconfigurable resources using some of the methods outlined in the following sections.

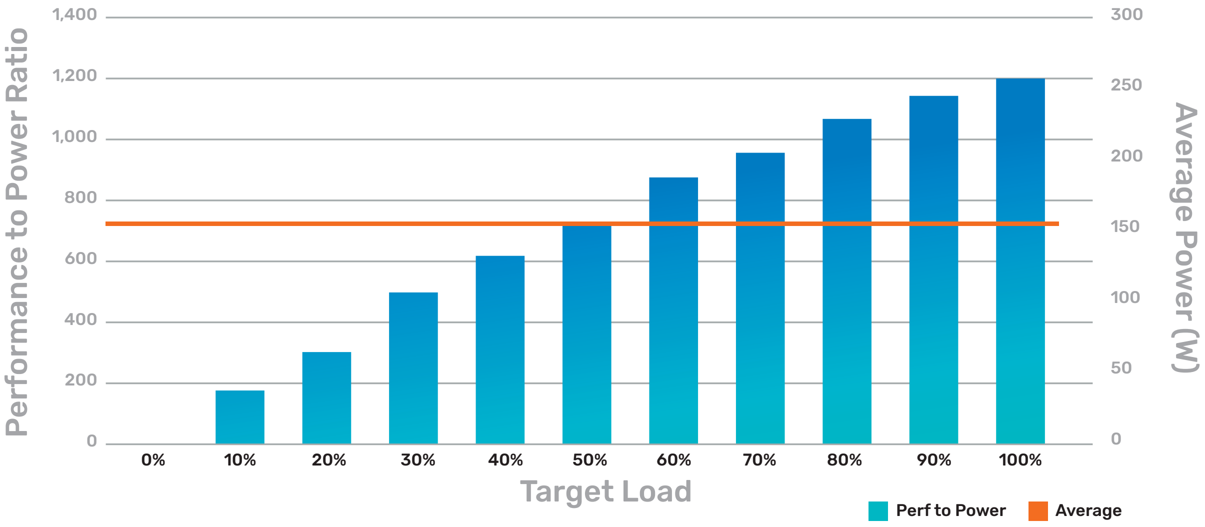

Figure 3: Barroso and Hölzle[6] found that average power declines nearly linearly with utilization, but performance to power (efficiency) degrades much faster. At utilizations of less than 50%, a significant gap exists due to idle power consumption in servers—for optimal use, operate in the region on 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 data centers. 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. Applications design has also evolved considerably from monolithic chunks of software to micro-services that rely heavily on containerized models, relying on virtualization extensively. Google's Kubernetes has been a foundational piece of infrastructure for container orchestration.

Virtualization does increase utilization, but there's still the risk of underutilization at non-peak loads. Another consideration is that providers, following the guidelines of strict service-level agreements (SLAs), often have to guarantee performance levels. As a result, data center 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?

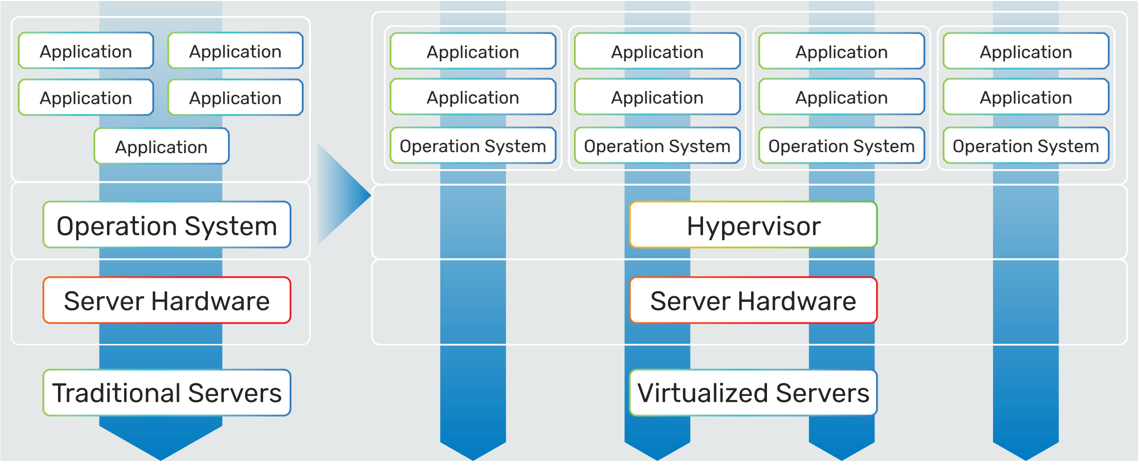

Figure 4: While traditional servers run a single operating system with applications running on it, virtualized servers allow for increased utilization on a common hardware platform through the use of hypervisors that permit multiple OS instances ("guest OSs") and their associated applications

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 data centers. With multi-core machines, data center 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 data center hardware.[10] 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 data centers 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.[6]

On the memory side, self-refresh assist can reduce power consumption by an order of magnitude.[10] 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, 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 PCIe specification defines protocol features, device states, and link states for power management.[13] The link power states are 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 state rises

The L0 (Stall) state offers relatively low power savings but allows applications to return to active state much faster. Driven by market and regulatory needs, the PCI-SIG proposed an ECN to lower power consumption (in milliWatts) in the L1 state. "L1 Power Mode Substates with CLKREQ"[11] redefines the L1 state as L1.0 and includes two sub-states defined as L1.1 and L1.2 (Table 2). With these two states, the standby state can reduce power consumption in its mode. Cadence was the first to implement 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 idle
L1.1	RX/TX off, common-mode voltage maintained
L1.2	RX/TX off, common-mode voltages off

Table 2: PCIe L1 substates further reduce power by turning off portions of the analog design; exiting from this state occurs with assertion of CLKREQ#

The two 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 low-power modes are critical since system performance can suffer from long latencies. Further, there may be functionality issues if Link Training Status State Machine (LTSSM) timers are not honored correctly. PHY designs are being pushed to the limit to reduce exit latency while providing low current consumption simultaneously. Maintaining this performance as PCIe designs progress into higher levels of the specification (5.0 and 6.0) poses new and unique challenges that IP providers must face to in order to provide the best solutions possible.

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.

Dynamic Power Allocation (DPA) allows applications to manage PCIe power management and power budgeting. Operating system device drivers use these capabilities as part of an overall power management strategy.[14,15] This feature has been available since the 2.0 version of the specification.

Dynamic Link Width Change allows the link width to be modulated based on bandwidth demand, a key aspect of energy proportionality. For example, a wider link of four lanes could be modulated to a single lane of traffic if the application determines that demand for bandwidth is lower, or vice versa. This feature requires the entire link to enter the recovery and configuration state, impacting performance overall at the expense of power saving.

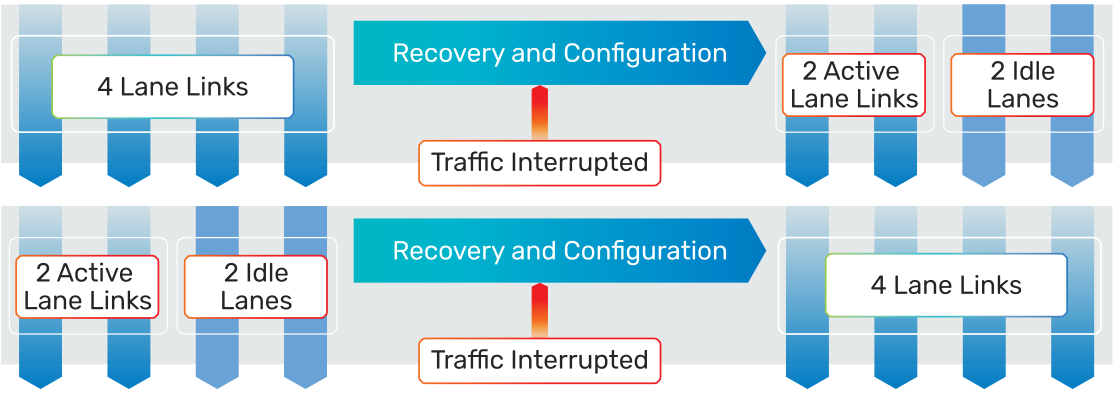

Figure 5: Dynamic Link Width Change

Dynamic Speed Change allows the PCIe link to run at a lower PCIe speed if the bandwidth required by the application is lower than the maximum bandwidth supported by the link. This typically saves less power than reducing the link width. The link goes through recovery and speed transition LTSSM states.

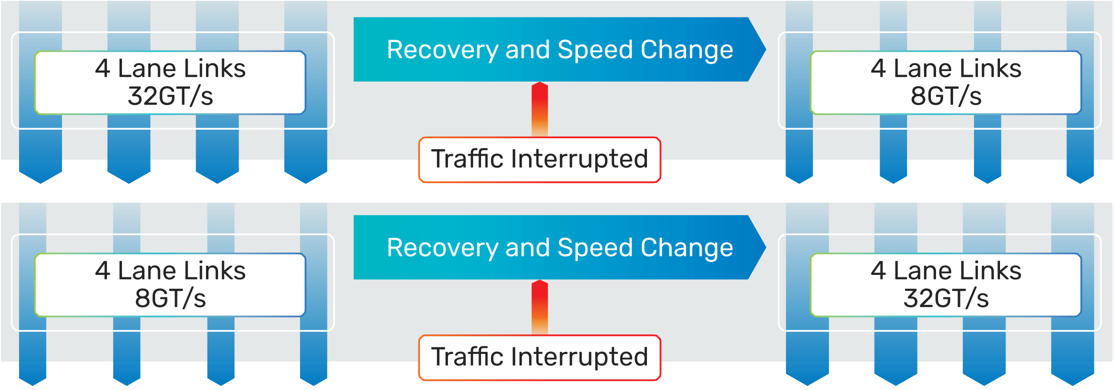

Figure 6: Dynamic Speed Change

With Latency Tolerance Reporting (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 estimate 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.

Optimized Buffer Flush/Fill (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 7), 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.

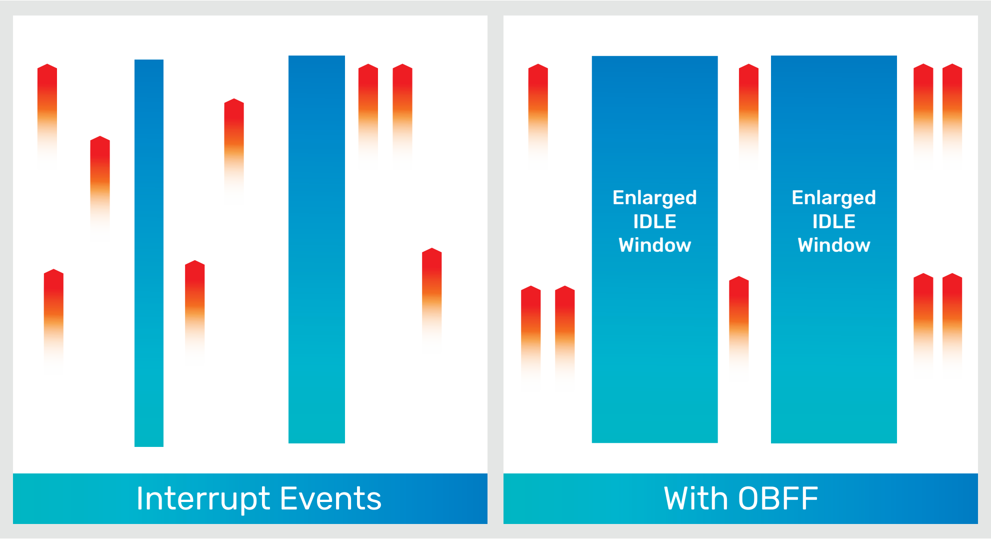

Figure 7: The premise for OBFF is that devices could optimize request patterns if they knew when they were allowed to interrupt the central system, which would allow the system to stay in a lower power state longer by expanding the idle window

PCIe 6.0 introduces a new state, L0p that enables link reconfiguration without the downside of the link retraining in Dynamic Link Width Change that put all lanes of the link into the recovery state. L0p allows active lanes to continue carrying traffic while other lanes retrain to higher or lower speeds based on bandwidth demand. Power savings are expected to match the L1 state for idle lanes with minimal delays while entering higher bandwidth modes. This new feature is only enabled when operating in the PCI 6.0 Flow Control Unit mode (FLIT mode).

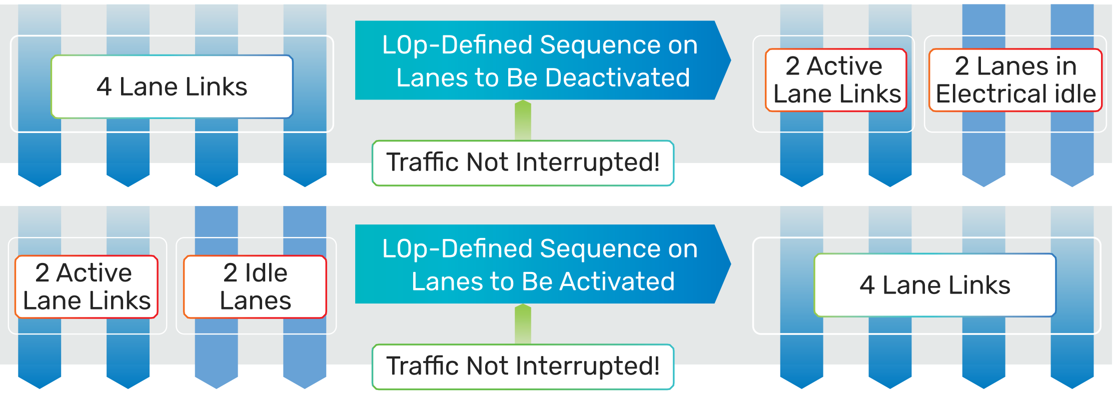

Figure 8: Example of link width change with PCIe 6.0 L0p allowing traffic to remain uninterrupted on active lanes—compare this change with that in Figure 5

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 but, superior PLL design techniques for fast lock times can reduce exit latency significantly, improving the resumption time and user experience.

Cadence was the first IP provider to optimize its PCIe controller and PHY to support these 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 was the first to market with a wire-bond PHY option and in flip-chip designs. Since then, this feature has been available in every generation of Cadence IP for PCIe and in advanced FinFET nodes, including its latest flagship IP for PCIe 5.0, which has stellar active and low-power performance. The Cadence solutions for PCIe 6.0 include support for the L0p state in order to realize the highest energy savings for all application segments.

Conclusion

To support our digital demands, data centers 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 Sub-states 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 (PPA). Vendors such as Cadence, with its low-power PHY IP for PCIe, can help turn this ideal into reality.

References

1 - Glanz, J. (2012, September 22). “The Cloud Factories: Power, Pollution, and the Internet”. Retrieved June 03, 2013: http://www.nytimes.com/2012/09/23/technology/data-centers-waste-vast-amounts-of-energy-belying-industry-image.html?pagewanted=all& _r= 0

2 - James, J. “Data Never Sleeps 6.0: How much data is generated every minute?”: https://web-assets.domo.com/blog/wp-content/uploads/2018/06/18-domo-data-never-sleeps-6.png

3 - International Business Strategies. (September 2021). Global Semiconductor Industry Service Report, Vol. 30, No. 9. “Data Center Activities”.

4 - Cisco Systems. “2020 Global Networking Trends Report”: https://www.cisco.com/c/dam/m/en_us/solutions/enterprise-networks/networking-report/files/GLBL-ENG_NB-06_0_NA_RPT_PDF_MOFU-no-NetworkingTrendsReport-NB_rpten018612_5.pdf

5 - Kay, R. (2012, October 18). “Taming Datacenter Power Usage”. (Endpoint Technologies Associates, Inc) Retrieved June 2, 2013: http://www.forbes.com/sites/rogerkay/2012/10/18/taming-datacenter-power-usage/

6 - 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.

7 - Rocky Mountain Institute of Technology. (2008, August 07). “Designing Radically Efficient and Profitable Datacenters”. Retrieved June 03, 2013: http://www.treehugger.com/gadgets/designing- radically-efficient-and-profitable-data-centers.html

8 - Uptime Institute (2019). “Annual Data Center Survey Results”: https://uptimeinstitute.com/resources/asset/2019-data-center-industry-survey

9 - Sen, R., and Wood, D. A.(2017). *Pareto Governors for Energy-Optimal Computing*. ACM Transactions on Architecture and Code Optimization 14, 1, Article 6. Retrieved March 15, 2021: https://research.cs.wisc.edu/multifacet/papers/taco2017_pareto_governors.pdf

10 - Meisner, D., Sadler, C. M., Barroso, L. A., Weber, W.-D., & Wenisch, T. F. *PowerNap: Eliminating Server Idle Power*. ISCA 2011. San Jose: ACM.

11 - PCI-SIG. (2012, August 23). “L1 PM Substates with CLKREQ”. Retrieved June 03, 2013: www.pcisig.com/specifications/pciexpress/specifications/ECN_L1_PM_Substates_with_CLKREQ_23_Aug_2012.pdf

12 - Hammond, T. (2013, April 8). “Toolkit: Calculate datacenter server power usage”. Retrieved June 9, 2013: http://www.zdnet.com/toolkit-calculate-data-center-server-power-usage-7000013699/

13 - Kwa, S and Cohen, D. (2002, November 8). “PCI Express Architecture Power Management Rev 1.1”. Intel Corporation. Retrieved July 14, 2021: https://www.intel.com/content/dam/doc/white-paper/pci-express-architecture-power-management-rev-1-1-paper.pdf

14 - Microsoft. (2017, June 16). “Device Low-Power States”. Retrieved June 14, 2021: https://docs.microsoft.com/en-us/windows-hardware/drivers/kernel/device-sleeping-states

15 - Wysocki, R. J. (2010). “PCI Power Management”. Retrieved June 14, 2021: https://www.kernel.org/doc/html/latest/power/pci.html

cādence®

Cadence is a pivotal leader in electronic systems design and computational expertise, using its Intelligent System Design strategy to turn design concepts into reality. Cadence customers are the world’s most creative and innovative companies, delivering extraordinary electronic products from chips to boards to complete systems for the most dynamic applications. **www.cadence.com**