CHAPTER 4

Platform Power Management

Each CPU in a data center requires a large amount of support hardware. This support hardware, contained within the server chassis, is generally referred to as the *platform*. Over the years, more and more of the platform has been integrated into the CPU, such as memory controllers and PCIe connectivity. However, a large portion of the overall power in the data center is still consumed by the support infrastructure outside of the CPUs and memory. Storage (drives), networking, power delivery, and cooling all can contribute a significant amount to the overall cost of a data center. Some of these components (like the fans) have sophisticated algorithms that attempt to manage their power consumption, whereas others (like drives) tend to employ minimal power management techniques.

Platform Overview

A *platform* is conceptually everything (including the CPU) required for a CPU to operate. It includes the power delivery (which converts electricity from the power grid into something usable by the different platform components), cooling (fans, heat sinks, etc.), as well as the memory, drives, and networking that are connected to the CPU sockets.

Common Platform Components

A single platform is commonly referred to as a *node*, which generally incorporates from one to eight CPUs that are connected with coherency.[1] A wide range of platform designs are possible and available. However, some standard building blocks go into just about any platform design (see Table 4-1). This chapter investigates some of the power management characteristics of these various platform components.

[1] *Coherency* is a mechanism that allows different software threads running on different CPUs to share a large set of physical memory without requiring software management.

Table 4-1. *Common Platform Components*

Component	Description
CPU	These processors provide the computation and execution of user workloads. See Chapter 2.
Memory	Memory provides temporary storage for data being used by the CPUs. See Chapter 3.
Storage	Storage (drives) provides bulk storage of data. SAS (serial attached SCSI) and SATA (Serial ATA) are two common protocols for connecting drives to a storage controller.
Networking	Networking provides for communication between multiple nodes. Ethernet and InfiniBand (IB) are common networking interfaces. NICs (network interface cards) provide the connectivity between the CPU and the Ethernet/IB network.
Power delivery	Different components in the system require different voltages and types of current (AC/DC). VRs (voltage regulators) are DC to DC converters that take an input voltage and step it down to a lower operating voltage. PSUs (power supplies) take AC current and convert it to DC.
Cooling	When servers consume power, it is turned into heat. Fans and other cooling devices are used to extract that heat from the platform to maintain a safe operating temperature.

A wide range of platform designs are used in the industry. Some designs provide large amounts of data storage and connect a large number of drives. Others may be completely driveless and use the network to bring data into the node. Figure 4-1 provides an example of one potential platform node with two CPU sockets.

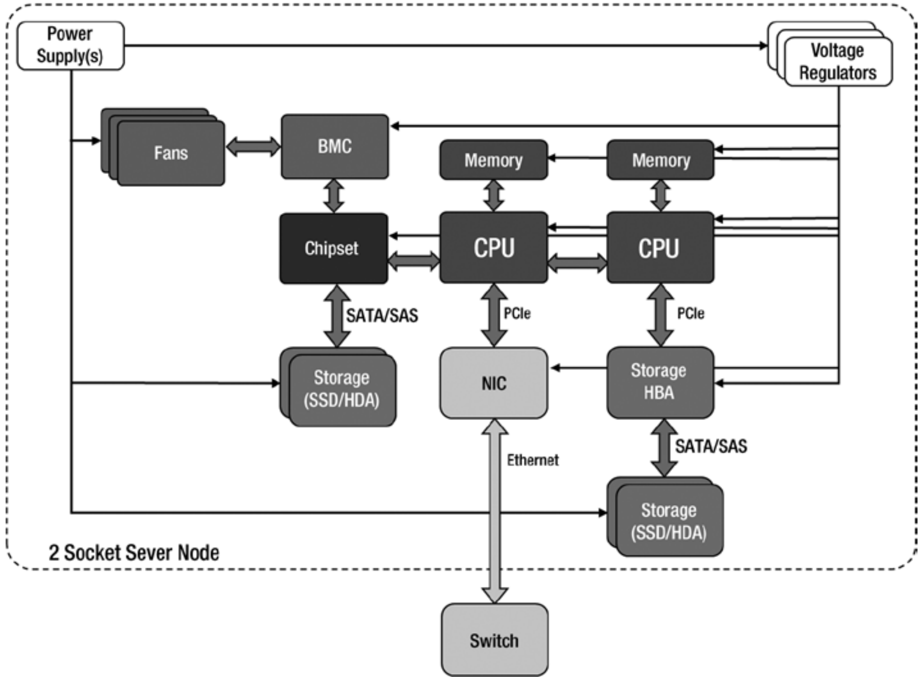

Figure 4-1. *Two socket platform node example*

Integration

As time has progressed, more pieces of the platform have been integrated into fewer discrete chips. This can save cost and power and even improve performance in some cases.

CPU Integration

For many years, the sole role of the CPU die was to provide one (or a couple of) cores and a supporting cache hierarchy. These were then connected to some system bus (front-side bus (FSB) on Intel systems), which then connected them to a chipset. This chipset provided a memory controller and PCI connectivity for devices like drive and network controllers. These busses consumed power, limit bandwidth, and increased latencies. As a result, more and more of the chipset began to get integrated into the CPU itself, both to reduce platform power and to increase performance. Table 4-2 provides an overview of some of the key integration milestones over Xeon processor generations.

Table 4-2. *Integration over Intel EP CPU Generations*

Generation	Integration
Nehalem	Memory controllers
Jasper Forest (Nehalem Derivative)	PCIe 2.0 (up to 16 lanes)
Sandy Bridge	PCIe 3.0 (up to 40 lanes) that can share L3 cache with cores
Haswell	Voltage regulators

Although power can be saved (fewer platform busses and I/Os driving them) and performance can be improved (on-die busses provide higher bandwidth and lower latency compared to off-chip interconnects), this integration is not free. The area of the CPU must increase to accommodate the additional components. CPU packages may need to accommodate more pins (which increases cost). This integration also moves power that was previously consumed out in the platform into a much closer physical location to the traditional CPU components. This either requires that more power (and cooling) be provided to the CPU or that less power be made available to the cores.

CPUs are typically built on the latest manufacturing process technology that provides the best power efficiency. Other devices in the platform are usually manufactured on older technologies. When they are integrated into the CPU, these capabilities get an immediate upgrade in power efficiency due to the process technology improvement.

Chipset Integration

The CPU absorbed the memory controller and some of the PCIe connectivity away from the chipset in the Nehalem and Sandy Bridge generations. However, the chipset has started integrating other components of the platform. Storage and network controllers are now standard on server chipsets. PCIe is still provided, although it is generally lower performance than the CPU links and is focused on low-bandwidth connectivity. Chipsets are discussed in more detail later in this chapter.

Microservers and Server SoCs

Server system on a chip (SoC) components are becoming more and more prevalent. In these designs, the chipset and CPU are integrated together into a single die or as a multi-chip package (MCP). The primary goal here is to reduce the costs of deploying a single CPU node. The concept of a microserver is where you target these lower-cost devices in mass quantities in a data center to provide adequate performance at reduced costs. Although microservers have received significant press in recent years, deploying these lower-cost, power efficient, highly integrated devices into embedded markets is arguably even more interesting.

Platform Manageability

Running a large data center requires capabilities for monitoring and managing the various components that go into a data center. Controlling fans, rebooting nodes that have crashed, monitoring power, and many other tasks are all critical to managing a typical data center. These concepts will be discussed in Chapters 5 and 9. Rather than having software running on the CPU cores to provide these capabilities, many server platforms have traditionally deployed dedicated management chips. These are commonly called baseboard management controllers (BMCs).

Server BMCs are OEM-proprietary devices with a small microcontroller at their heart. They have tentacles throughout the platform in order to monitor and control the various subsystems. Platform Environment Control Interface (PECI) is a standard used for interactions between BMCs and CPUs. System Management Bus (SMBus) protocol is also commonly used for providing telemetry information from platform devices (power supplies, etc.) to the BMC. Intelligent Platform Management Interface (IPMI) is an interface used for software to interact with the BMC for extracting the wealth of information of which the BMC is aware (see Chapter 7 for examples). A single platform with *N* coherent CPU sockets is generally paired with a single BMC, but this is not strictly required.

BMCs themselves do not consume a significant amount of power but can have a notable impact on the overall power draw of the system since they control the fans associated with a given platform node. Thermal management is discussed later in this chapter. Servers without BMCs have been investigated in order to reduce power consumption and save on integration costs, but thus far, such designs have not taken off.

CPU Sockets

Modern CPU nodes can support varied numbers of CPU sockets. Uni-processor (UP) and dual-processor (DP) servers make up the bulk of the server processor nodes sold today.

Multi-processor (MP) nodes commonly consist of four or eight processors, but other topologies are also possible. MP platforms have a higher procurement cost associated with them, and are frequently used in situations where large single-node performance or memory capacity is required. By moving to a larger number of CPU processors per node, the cost of some of the platform components can be amortized. For example, if each node requires a boot SSD and a network connection, one can potentially reduce the number of required SSDs and network connections by two times by going from a UP to a DP platform.

■ **Note** Due to the large procurement costs and usage models associated with MP systems, power efficiency and power savings are typically a lower priority for end users.

DP platforms provide an excellent cost/performance sweet spot. MP platforms have typically demanded a higher overall price per CPU, while UP platforms are not as effective at amortizing other platform costs (power and procurement). DP platforms also exhibit strong performance scaling for many workloads.

UP server systems have traditionally been relegated to situations that simply did not demand the performance of a DP or MP system. Rather than being deployed in a data center, they have been used in other lower-end server appliances such as small business NAS (network-attached storage). As single node performance continues to increase, UP systems cost amortization is improving. If a DP system requires two network connections in order to provide sufficient data to saturate the capabilities of the cores, then there is no additional savings by scaling to two sockets. Server SoCs (like microservers) that incorporate capabilities like networking also help reduce the power and procurement amortization benefits of multi-socket systems.

Platforms that directly connect two to eight processors coherently to each other are said to be *glueless*. A variety of glueless topologies have been developed over time. Figure 4-2 shows some examples from recent processor generations from Intel. Note that in each of these examples, every socket is either one or two "hops" from each other socket on the platform. It is possible to connect even more processors in a coherent network, but this generally requires special hardware (or glue) called node controllers. If all the processors are directly connected to each other through point to point links, the platform is said to be *fully connected*. Fully connected platforms generally have lower latencies, higher bandwidth, and better performance scaling than platforms that are not. There is a small power cost for the additional connectivity, but the return on investment (performance) is well worth the cost for most usage models.

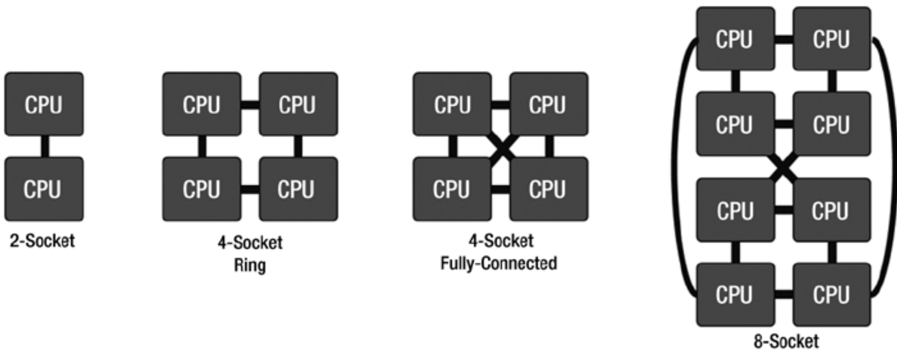

Figure 4-2. *Example glueless coherent platform topologies*

Node Controllers

Although the majority of platforms limit the number of coherent CPUs to a maximum of eight, it is possible to build much larger coherent systems using *node controllers (xNC)*. Node controllers are generally discrete chips that connect one, two, or four CPUs out to other node controllers through a proprietary fabric (see Figure 4-3). These systems are frequently used for building supercomputers and can connect hundreds of processors and thousands of cores into a single coherent domain running a single operating system.

Note that it is also possible to build large supercomputers without node controllers by connecting a large number of nodes non-coherently through a network. The differences between these designs are beyond the scope of this book.

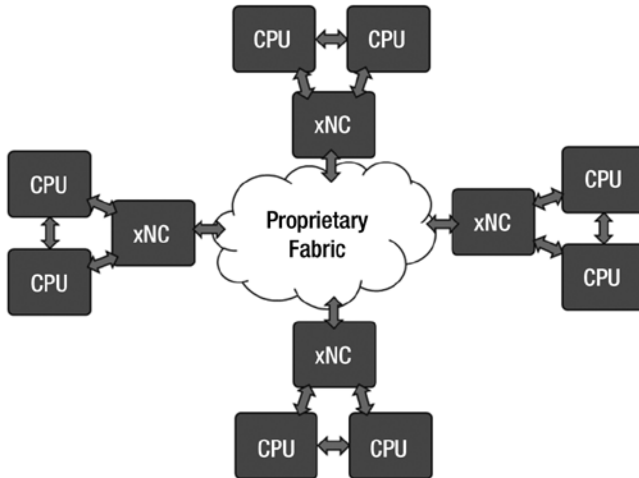

Figure 4-3. *Node controller example*

Memory Risers and Memory Buffer Chips

Certain high-end servers demand very large memory capacities. Databases are the most common example. Each CPU socket is generally limited in the amount of DDR memory to which it can directly attach. The number of DDR channels on a socket is constrained by packaging and die costs. The number of DIMMs on a channel is limited by electrical loading constraints. LR-DIMMs attempt to address some of these issues but can only go so far. In order to expand memory past the constraints imposed by the CPU socket, memory risers and memory buffer chips have been used on some high-end servers. Rather than connecting the CPU directly to memory, the CPU communicates with a discrete chip in the platform that is then able to communicate to the actual DDR memory. In these platforms, the memory is connected on separate riser cards, where a set of DIMMs is connected to a card, and then that card is connected into the motherboard. There have been various flavors of these technologies over the years. Intel has historically productized a memory buffer technology as part of its EX platforms (called Scalable Memory Buffer [SMB]), and other OEMs have deployed their own proprietary technologies to provide similar capabilities. These buffer chips do consume measurable power (usually a few watts), but they tend to be dwarfed by other power in such platforms (including the memory that they provide connectivity to).

Server Chipsets

Many server platforms employ discrete chipsets that are connected to the CPUs. These devices provide key legacy capabilities required for booting the platform, capabilities for manageability, and also integration of many features that otherwise would require discrete controllers (such as storage, network, and USB controllers). Some SoCs integrate the chipset functionality into the package with an MCP, while others (like Avoton) integrate the entire chipset into the same die as the CPU. The discrete chipsets used in many Xeon server designs at Intel are called a PCH (Platform Controller Hub). The PCH attaches to the CPU via a proprietary DMI (Direct Media Interface) link, and provides boot, manageability, and I/O services to the platform.

The PCH has been the south bridge of the two-chip Xeon Intel Architecture since the Nehalem/Tylersburg generation and is a companion to the CPU. This architecture succeeds the Intel Hub Architecture, which was a three-chip solution. Successive generations of PCH have advanced the I/O capability of IA platforms, with Gen2 PCIe, Gen3 SATA, and Gen3 USB now available on Wellsburg. A microcontroller-based power management controller (PMC) and a Management Engine (ME) were added to the PCH to support traditional power management features, along with several extended features.

The chipset serves a variety of purposes in the platform. Table 4-3 provides a high-level summary of some of the key capabilities. Figure 4-4 shows an example block diagram of such a system. Table 4-4 enumerates some of the integrated functionality of modern PCHs.

Table 4-3. *PCH High-Level Capabilities*

Capability	Description
High-performance I/O connectivity	This includes PCIe, storage (SATA and/or SAS), networking, etc. These capabilities are only available when the CPU is active and the system is in the S0 state.
Wake/boot	The PCH both detects wake events (like Wake on LAN) and sequences the platform to transition in and out of platform power states. The PCH also provides access to flash memory for BIOS.
Manageability	This provides interfaces for the data center to monitor and manage the node, such as reading temperatures.
Real-time clock (RTC)	This maintains the system clock that tracks clock time. If you unplug a desktop from the wall, your time and date is not lost since it is maintained on the RTC. The same capability exists in server PCHs.
Legacy I/O connectivity	This provides connectivity to low-performance platform connectivity that is generally required for system operation.

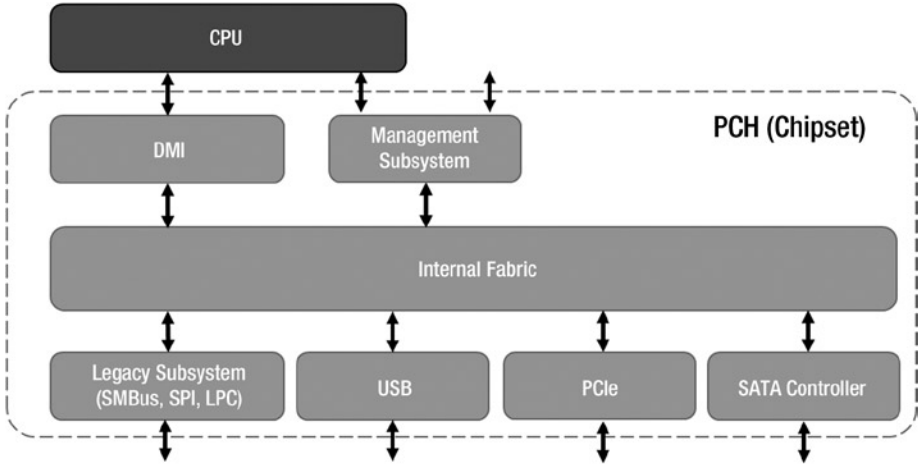

Figure 4-4. *A typical server PCH architecture block diagram*

Table 4-4. *Primary PCH Components*

Component	Description
On-die fabric	Interconnects exist on the PCH that are commonly called *on-die* or *on-chip* fabrics. These interconnects are conceptually similar to those in CPUs. These are not to be confused with fabrics that connect multiple CPUs together at the data center level.
DMI	DMI provides a mechanism to connect the PCH and components connected downstream from the PCH to the CPU. It operates very similar to PCIe.
PCIe	PCIe connectivity can be incorporated both into the CPU and the PCH and shares the same basic power management capabilities. The PCH is useful for high fanout, low bandwidth connectivity.
SATA	Storage connectivity is included on some PCHs and is discussed in the “Storage” section of this chapter.
USB	USB is primarily targeted and consumer usage models but is also present in servers (particularly for debug usage models). It is discussed later in this chapter.
Ethernet	Ethernet integration is also incorporated into the PCH. Networking power management is discussed in the “Networking” section of this chapter.

(*continued*)

Table 4-4. (*continued*)

Component	Description
SPI/LPC (legacy I/O)	Serial Peripheral Interface (SPI) and Low Pin Count Interface (LPC) provide connection points to platform boot devices that contain the BIOS/UEFI image, as well as firmware for other PCH components (e.g., ME, Ethernet).
ME	The Management Engine on the PCH provides platform management services, key management, and cryptographic services.
SMBus	SMBus provides a legacy mechanism for communication with platform peripherals for system and power management–related tasks.

The PCH has traditionally been the component that provides access to various high-speed I/Os (SATA, Ethernet, PCIe, USB), although these capabilities are increasingly being integrated into the CPU to create SoC components. The SATA/PCIe/Ethernet/USB interfaces provide access to external communication, including disk/solid-state storage, networking, USB ports, and manageability. Moving data with higher performance between two devices in a platform consumes non-trivial power, and integration is an effective way to significantly improve overall platform power consumption.

Internally, the PCH architecture is constructed with a mix of analog and digital components. Similar to the CPU uncore, analog design is used for designing the off-chip communication (e.g., PCIe/SATA/USB physical interface) and on-die memory (SRAM), while the bulk of the remaining system is built out of synchronous digital logic. Unlike traditional server CPUs, a large percentage of the chipset power is consumed by analog I/O circuitry (commonly called physical layer or PHY) and not the digital logic.

PCH and Platform Power Management

The PCH orchestrates many of the platform power states introduced in Chapter 2. In addition to this task, it is responsible for managing its own power states. Table 4-5 provides an overview of the power management states in which the PCH participates.

Table 4-5. *System States Supported by Server PCH*

State	Description
C-states	*Low-power states for PCH I/Os* See Chapter 2 for details on CPU C-states. PCH does not support all the traditional CPU C-states but places its I/Os in low-power states when the CPU is not active. • C0 is an active state when the PCH logic and I/O are functional. • Cx is a clock/power-gated state, during which PCH I/Os are transitioned to a lower power-managed state.
S-states	*Turning off the CPU package (sleep state)* See Chapter 2 for details on S-states. The PCH I/Os are turned off (except S0), but the PCH core logic remains active in all S-states. PCH can wake the CPU up from S3/S4/S5 states based on platform signaling. Waking from an S3 state takes seconds, whereas waking from S5 requires a full system boot and can take multiple minutes.
M-states	*Turning off the Management Engine* These states are related to the Management Engine. • M0: Active state, when platform is in S0 state. • M3: Active state, when platform is in S3/S4/S5 state, used for out-of-band platform management and diagnostics. • MOff: Management engine is turned off in S*x*.
G-states	*Global states* See Chapter 2 for details. The PCH is active in G0 to G2 and is only off in the G3 state (mechanical off).

Since the PCH controls the platform rails and clocks, it needs to remain powered on even in states where the CPU is powered off (G- and S-states). This is accomplished by using platform power rails that are successively powered off depending on the system state. The PCH provides a number of high-level capabilities that are successively disabled at lower power states. Table 4-4 provides an overview of these capabilities, and Table 4-6 shows how the capabilities are disabled in each power state.

Table 4-6. *Example PCH Power States and Capabilities*

G-State	M-State	S-States	Power Rails	Manageability	Wake Capability
G0	M0	S0	All available	Available	N/A (awake)
G0	M3	S3/S4/S5	Wake + Manageability + RTC	Available	Yes
G2	MOff	S4/S5	Wake + RTC	Disabled	Yes
G3	MOff	S4/S5	RTC	Disabled	None

Systems autonomously transition out of the G3 state and into the G2 state when power is supplied to the platform. From there, various wake events can be used to transition the state into a higher power operational mode as needed. As such, the G3 state is hidden from the user.

PCH Power Management

The PCH consumes a small percentage of a node's overall power. Because the bulk of the power consumed by the PCH exists in the I/O PHYs, the typical power consumed under load is very dependent on the number of connected devices. The TDP power of Patsburg (the chipset used with Sandy Bridge and Ivy Bridge E5 processors) was 8 W to 12 W when all of the high-speed I/O ports were connected (fully populated). Wellsburg (paired with Haswell E5) consumed a TDP of 7 W when fully populated. Notable power can be saved if certain I/Os are not populated. Table 4-7 provides an overview of four different usage configurations of the PCH and the corresponding TDP power for those configurations.

Table 4-7. *PCH TDP Power (W) Consumption with Various I/O Port Configurations*

	Workstation	Server	Low Power	Boot-Only
USB2 Ports	14	6	2 (detection)	0
USB3 Ports	4	4	1 (detection)	0
SATA3 Ports	8	5	2	0
SATA2 Ports	2	1	1	0
PCIe Lanes	8	4	2	0
TDP (W)	6.5	5	3.2	1

The Wellsburg PCH, which launched with the Haswell Server CPU, is built on a low-leakage process and does not implement techniques like voltage-frequency scaling or power gating to reduce the power consumed at runtime. Turbo is not available. In order to save power, clock gating is performed on logic features that are disabled or not currently in use. Since the PCH is I/O dominated, a sizable portion of the power is

consumed by the circuits that provide the physical interface to the platform. The voltage of these interfaces is generally static (as defined by industry standard specifications). Several power management states are defined for the links to opportunistically reduce power based on the operating state, as described in the following subsections.

If an entire section of logic is not being used, then the PLL (phase-locked loop) that drives that logic can be powered down. For example, if a user is building a compute node that lacks any local drives, the storage subsystem in the PCH can be completely powered down.

PCIe in Chipsets

Prior to Sandy Bridge and Jasper Forest, chipsets provided the PCIe connectivity in the platform. When Sandy Bridge integrated PCIe into the CPU, the chipsets continued to provide this capability. Today in platforms with discrete PCH devices, PCIe connectivity is offered on both the CPU die and the PCH. PCIe in the PCH provides the same power-saving capabilities that are described in Chapter 3 (L1, DLW).

PCIe on the CPU provides high performance and (relatively) low latency connectivity at the expense of limitations in the fanout (devices smaller than x4 consume four lanes). The PCH, on the other hand, provides lower bandwidth and longer latencies, but can be bifurcated down to x1 making it an excellent choice for low bandwidth devices.

PCH Thermal Management

The PCH contains thermal sensors in order to monitor the temperature and help guarantee that the PCH will not get to a dangerous temperature where reduced reliability or damage could occur. The PCH may throttle itself to stay under a target temperature or even initiate an immediate shutdown if temperature exceeds a catastrophic threshold. Like CPUs, PCHs are spec'd with a TDP rating that is used to design the thermal solution and an ICCMAX rating that is used to size the voltage regulators to power the voltage rails. They also contain similar thermal protection mechanisms such as shutting down the platform when catastrophic temperatures are detected. Platform thermal management is discussed in detail later in the chapter.

Networking

Network interfaces—both the local LAN adapter as well as network infrastructure devices—are the gateway for the server platform to the rest of the world. Network activity demonstrates unpredictable distribution of packet arrival times at multiple scales.
As a side effect, the network interfaces are never fully powered down. LAN adapters contribute ~5–10 W to the overall platform power. This power is not one of the primary power contributors in typical server platforms that deploy high-power CPUs and large amounts of memory. Although the LAN adapters themselves do not directly contribute a significant percentage of the platform power, their behavior and configuration can have a large impact on the power consumption of the CPU (and thus the platform).

■ **Note** Although network cards do not themselves contribute a significant percentage of the platform power consumption, their configuration and behavior can have significant impacts on CPU power (and thus platform power).

In typical usage, LAN component power is driven by five main factors. Table 4-8 provides a high-level summary of these factors (which are discussed in detail in the following pages.).

Table 4-8. *Primary Factors in LAN Controller Power*

Factor	Description
Ambient temperature	LAN devices have traditionally been manufactured with high-leakage process technologies, resulting in a significant power increase at higher temperatures.
Attached media	The type of connection (fiber optic, copper cable, etc.) can have a moderate impact on the power consumption.
Configured speed	LAN controllers can be configured by software to run at lower frequencies. This can save notable power.
Power management features	Various power management options are available that can trade off performance (latency) to save power.
Bandwidth	Packets per second have the biggest impact on NIC power (not raw bandwidth). However, on recent high-performance networking devices, there is not significant sensitivity to bandwidth.

Ambient Temperature, TDP, and Thermal Management

Many LAN vendors quote typical power numbers in their datasheets. However, there are no industry conventions as to what typical usage is, though many assume 25°C for ambient air temperature, and nominal voltage. An increase in temperature from 25°C to 70°C can increase the component power by 50% to 100% solely due to leakage (which itself is a function of the silicon process used to produce the device). As LAN controllers transition to lower leakage processes or are integrated into low-leakage SoC designs, the sensitivity to temperature will decrease.

Similar to CPU designs, the maximum quoted power of the LAN controller is measured assuming worst-case conditions, including high temperatures (~70°C ambient). The server platform thermal management—such as fan size and speed—is designed to cool to this maximum component thermal design point (TDP). LAN controllers are typically designed assuming passive cooling, and it is also common for these devices to exist in areas of limited airflow. Active cooling—such as fans—is discouraged because of server platform reliability concerns. The net result is, regardless of the functionality or media provided,

the server LAN component TDP must be 10 W or less (unless special design provisions are made at the platform level for additional fan cooling). Tables 4-9, 4-10, and 4-11 show some historical information about Intel LAN adapter TDP power.

Table 4-9. *Historical TDP Power of Single-Port 1 Gbps Intel LAN Adapters*

Year	Device	Ports/Speed	TDP (W)	TDP (W) / Gbps
2001	Intel 82544EI PCI-X	1x 1 Gbps	1.5 W	1.5 W
2004	Intel 82541 PCI	1x 1 Gbps	1.0 W	1.0 W
2005	Intel 82573 PCIe	1x 1 Gbps	1.3 W	1.3 W
2008	Intel 82574 PCI	1x 1 Gbps	0.7 W	0.7 W
2012	Intel I210 PCIe	1x 1 Gbps	0.7 W	0.7 W

Table 4-10. *Historical TDP Power of Multi-Port 1 GBps Intel LAN Adapters*

Year	Device	Ports/Speed	TDP (W)	TDP (W) / Gbps
2005	Intel 82571 PCIe	2x 1 Gbps	3.4 W	1.7 W
2009	Intel 82576 PCIe	2x 1 Gbps	2.8 W	1.4 W
2010	Intel 82580 PCIe	4x 1 Gbps	3.5 W	0.9 W
2011	Intel I350 PCIe	2x 1 Gbps	2.8 W	1.4 W
2011	Intel I350 PCIe	4x 1 Gbps	4.0 W	1.0 W

Table 4-11. *Historical TDP Power of 10 GBps Intel LAN Adapters*

Year	Device	Ports/Speed	TDP (W)	TDP (W) / Gbps
2001	Intel 82597 PCI-X	1x 10 Gbps	9.0 W	0.9 W
2007	Intel 82598 PCIe	2x 10 Gbps	6.5 W	0.3 W
2011	Intel 82599 PCIe	2x 10 Gbps	6.2 W	0.3 W
2012	Intel X540 PCIe w/ 10GBASE-T Phy[2]	2x 10 Gbps	12.5 W	0.6 W
2014	Intel X710 PCIe	4x 10 Gbps	7.0 W	0.17 W

[2]This device includes a 10GBASE-T attached media, increasing the TDP power. The other controllers listed must be paired with a separate attached media.

■ **Note** Typical power for NIC cards is well below their TDP specifications. NICs frequently operate at lower temperatures than their specifications, saving significant leakage power.

In typical usage, the LAN component does not operate at TDP. Some LAN devices include thermal sensor diodes, as well as management interfaces, to enable other platform components to query the component thermal state and adjust fan speed. In practice, many of these platform methods require additional calibration of the thermal sensors which, if not done, may limit the effectiveness of the fan speed algorithms.

Attached Media

Most LAN adapters can be paired with a variety of different interconnect types that provide the actual connectivity between the LAN adapter and network switches. These are called *attached media*.

Server LAN implementations have a greater variety of media types than those found on client systems. Whereas most equate Ethernet to the pervasive RJ-45 connector and 10BASE-T (10 Mbps), 100BASE-TX (Fast Ethernet, or 100 Mbps) and 1000BASE-T (1 Gbps Gigabit Ethernet), server platforms have employed several media types as summarized in Table 4-12.

Table 4-12. *Types of Attached Media*

Type	Max Distance	Power	Latency
Multi-mode short reach (SR) fiber optic	~400 m	~1 W	Slight increase
Single-mode long read (LR) fiber optic	~10 km	~1 W	Slight increase
KX/KX4/KR Backplane (copper)	Server backplane	100s of mW	Best
Direct Attach (DA)	3–10 m	100s of mW	Best
BASE-T	100+ m	2–3 W	Adds ~1 microsecond

■ **Note** Cost and distances are generally the deciding factors in attached media selection. Latency is important to a subset of customers.

Each of these media solutions have tradeoffs between cable cost, power, distance, and even propagation velocity (fiber is slightly slower than copper-based connections). Because of this diversity, many server LAN connections are shipped with an SFP or SFP+ cage, which accepts various media type pluggable modules.

LAN Power Management Features

A number of common features are used for reducing power of both the LAN devices and the CPU. In addition to these, higher-end server LAN adapters implement multiple queues and methods to balance network traffic across multiple CPU cores. As a result, CPU cores can operate at a reduced frequency and save power.

Media Speed

Some media types—such as BASE-T and backplane—support establishing a link at lower media rates than the maximum possible—such as a 1 Gbps adapter linked at 100 Mbps. Lowering the established link rate often reduces the component power, sometimes by as much as 50%. As the link speed drops, the internal synchronized media clock lowers in frequency, leading to a lower dynamic power. Another effect relates to effective packet rates, since LAN component power varies more as a function of packet rate than packet size. For each packet, the LAN controller performs various lookups on the packet headers. Reducing the media rate reduces the packet rates as well, again leading to lower dynamic power.

In practice, changing media speed is not applicable for most server usage models. The transition latency is slow, and the reduced speed results in significant peak throughput reductions and the potential for increased latencies. Although this can save notable power from the perspective of the LAN controller, it is generally not as significant as a percentage of the overall platform power.

Energy Efficient Ethernet

BASE-T and backplane media also support Energy Efficient Ethernet[3] (EEE). This is frequently called *triple-E* for short. EEE devices enter into a low power mode during idle periods, periodically sending idle sequences to keep the link active and sending a wakeup symbol to the peer when the link needs to be reactivated. Depending on the media, the link transitions from idle to active are less than about 16 microseconds. BASE-T devices can reduce their PHY idle power as much 400 mW with 1000BASE-T, and by 2 W with 10GBASE-T.

EEE is managed by the NIC driver and can be controlled at runtime. It is generally enabled by default. The latency cost of EEE is not noticeable in many usage models, but the power savings is also not particularly significant. Latency sensitive users may want to attempt to disable this capability.

[3]EEE is decribed in detail in IEEE Std 802.3az-2010.

Wake on LAN

Wake on LAN (WoL) is a common feature available on server LAN adapters. It is not a power-savings feature as much as a mechanism to wake the system from an S-state.

If the platform supports suspend or wake from soft-off modes, WoL allows remote administrators a simple method to remotely activate a server platform. Often, the LAN interface will reduce link speed automatically when entering this state to minimize power and await receipt of a wake pattern. A common pattern used is the *Magic Packet* pattern. Upon receipt, the LAN controller asserts a signal to the platform to bring the system out of the low-power state.

Active State Power Management (ASPM)

Discrete LAN controllers are connected using PCIe. As such, PCIe Active State Power Management is available to manage power on LAN controllers. PCIe power management is discussed in Chapter 3.

NIC ASPM L1 is frequently disabled in server deployments. This can frequently be performed in the system BIOS. The latency implications are frequently not worth the low amount of power savings. One common issue with ASPM L1 is that it blocks communication between a driver (running on the CPU) and the NIC device. When communication is required, the core is then stalled. This ends up wasting CPU power, which eats into the already small savings from L1.

Interrupt Moderation

Interrupt moderation is another common feature of LAN controllers. It limits the rate at which interrupt signals are delivered to the host CPU. This often reduces the CPU utilization with little to no observable impact to bandwidth. Interrupt moderation has little impact on NIC power, but the decreased CPU utilizations can significantly improve CPU power consumption. It can also make additional CPU cycles available for other processes, improving the throughput of the node. By rate limiting interrupts, the CPU is notified less often, resulting in an increase in latency and response time. The amount of latency impact can be tuned inside the NIC driver and is commonly configured to levels on the order of 100–200 microseconds. This feature is typically enabled by default, and can be disabled (or configured) inside the NIC driver configuration.

Interrupt moderation can have a significant impact on power and latency in systems, and is frequently overlooked. Tuning this feature should be a priority for anyone who is concerned about latency and response times.

DMA coalescing is a related feature that attempts to queue up data transfers inside the NIC and burst them into the CPU. The intention of this feature was to allow the CPU to get into a low-power idle state between bursts of activity. In practice this feature has shown minimal effectiveness in server environments while also significantly increasing network latencies. It is not enabled by default.

USB

USB connectivity is provided by server PCHs. Many large-scale data centers do not connect devices to USB under normal operation, but it is common for USB ports to be included on those platforms. “Crash cart” support is a common usage model, where USB is periodically used to connect a keyboard/mouse for local debug, or to connect a USB drive for similar purposes. USB can also be used in some low-end storage systems for connecting USB storage. Power management of USB can be very effective at saving power at minimal to no power cost due to these limited usage models.

Link Power States

The initial USB power management capabilities were very coarse grained. A *suspend/resume* scheme provided two levels—effectively “on” and “off.” These take milliseconds for transitions, making them inadequate for many power-efficiency usage models. USB devices are common in consumer usage models where achieving very low idle power is critical to achieving long battery life. As a result, USB has been a focus for power optimization in these environments. Much of these capabilities are unnecessary in server usage models.

USB 2.0 originally only supported these two levels but later added support for L-states that complemented the suspend state. On USB 2.0, the state of the link is tied to the power state of the device. These states are summarized in Table 4-13. Suspend can still be used for states that have no latency sensitivity (such as S3/S4).

Table 4-13. *USB 2.0 Power States*

State	Name	Link Savings	Device Savings	Exit Latency
L0	On	--	--	--
L1	Sleep	~100 mW	Device-specific	microseconds
L2	Suspend	~125 mW	Device draws almost no power	milliseconds
L3	Off/Disconnected	~140 mW	Device powered down	milliseconds

Note: Power savings are design dependent. These numbers provide a reference point.

On USB 3.0, the device power states were decoupled from the link power states. U-states were defined that control the power state of the link only. Table 4-14 provides an overview of the four USB 3.0 link power states.

Table 4-14. *USB 3.0 Link Power States*

State	Name	Link Savings	Exit Latency
U0	Link active	--	--
U1	Link down	~100 mW per lane	Microseconds
U2	Link down	~125 mW per lane	Milliseconds
U3	Link off	~140 mW per lane	Milliseconds

Note: Power savings are design dependent. These numbers provide a reference point.

Link Frequency/Voltage

USB has gone through three generations. Each generation has a single frequency at which the device runs (see Table 4-15). Multiple voltage/frequency points are not supported. Newer generation devices support (by rule) backward compatibility to the prior generation frequencies.

Table 4-15. *USB Generations*

Generation	Frequency	Duplex	Theoretical Bandwidth
USB 1.*x*	12 MHz	Half	1.5 MB/s
USB 2.0	480 MHz	Half	35 MB/s
USB 3.0	5 GHz	Full	500 MB/s (per direction)

USB 3.0 moved to a full-duplex design, effectively providing separate communication channels for both directions, increasing the peak throughput when data are transferred in both directions simultaneously. This required the addition of two more differential pairs, and is similar to a single lane of PCIe or a SATA connection.

Storage

Many modern data centers deploy storage in a variety of different ways. Some compute nodes have no local storage and depend entirely on the network to provide access to remote storage. It is also common to see compute nodes with a single drive (commonly an SSD) that provides for high-performance local storage. Other nodes can be targeted for storage and can provide access to a large number of drives. These nodes are connected to compute servers through high-performance interconnects to provide large pools of shared storage.

Traditionally, drives have been connected through PCIe-based controllers. Two standard interfaces exist for these controllers: SATA and SAS. Serial Advanced Technology Attachment (SATA) is the lower cost of the two, but it also provides lower peak performance. SATA can be used both in consumer and server usage models. Serial attached SCSI (SAS) is generally more expensive and higher performance and is targeted at enterprise usage models. SATA drives can be connected to a SAS infrastructure, but SAS drives cannot be connected to a SATA controller. SATA and SAS support both SSDs (solid state drives) and HDDs (hard disk drives). In addition to providing higher peak performance, SAS provides the ability to connect a large number of drives to a single controller, making it popular in very high capacity deployments.

In recent years, SSDs have begun to be directly connected on PCIe. Non-Volatile Memory Express (NVMe) is a specification for performing this direct connection. NVMe provides lower latency and higher performance than SAS and SATA. This is particularly well-suited for high performance compute nodes that require local storage. NVMe SSDs exhibit similar power characteristics to SATA and SAS SSDs. Their potential for higher performance also translates into higher power consumption.

Storage power consumption is generally not a significant component of the overall node power in traditional compute servers. However, the power consumption of the drives on a storage node can dwarf the other components on the node. Storage power is also more significant in low-power, low-performance servers where the drive power is not amortized across a high-power CPU node.

Storage Servers and Power Management

In a typical storage server the CPU complex manages tens to thousands of drives. Storage servers can use a mix of SSDs and HDDs, and the mix is determined by the performance needs. SSDs have higher procurement costs but provide improved performance. A significant amount of power in storage servers is consumed by the drives. Cooling a dense storage complex can also consume non-trivial power.

It is increasingly critical to manage the power consumed by the storage devices, without adversely affecting performance. Various power management schemes can be used depending on the performance requirements of the application. In a cold storage system where a massive amount of data is maintained but accessed rarely with low performance and latency requirements, aggressive power management can be used to reduce costs. On the other hand, limited power savings is used in performance- and latency-critical storage deployments. Aggressive power savings in such environments can be detrimental to performance but can also reduce overall data center power efficiency by forcing compute nodes to wait longer for data (wasting power in the process).

Power savings opportunities exist both within the drives and in the communication layer between the drives and their controllers. Storage power management schemes have been developed for both server usage models as well as consumer usage models. In consumer usage models, very low idle power is critical for battery life, and capabilities have been developed to address these concerns. These same capabilities may be available in the server space but can provide poor tradeoffs. A few hundred milliwatts of power savings is commonly a poor tradeoff if it could result in milliseconds of response time increase.

HDDs and SDDs

SSDs generally provide higher bandwidth and lower latencies than HDDs; however, this has traditionally come at increased power and cost per capacity. Actual power consumption is drive dependent, and can range anywhere from a couple of watts to more than 10 watts.

Traditionally, 3.5" HDDs have been extensively used in data centers and other server applications. They can provide the best cost per GB due to the larger platters. 2.5" HDDs do tend to exhibit lower power draw than 3.5" drives of the same capacity, but this benefit tends to be overshadowed by their overall lower maximum capacity in the same technology generation.

The amount of traffic that an HDD is servicing has minimal impact on the amount of power that is consumed. Rather, the state of the drive (Is it spinning? Are the heads loaded?) is the predominant component of HDD power consumption. SSDs, on the other hand, exhibit significant power dynamic range as a function of bandwidth. The act of reading and writing the cells itself consumes a significant percentage of the drive power. SSDs' power consumption also appears to scale with capacity. This is not because the cells themselves consume significant power, but because peak performance of these higher capacity drives is frequently higher, providing more potential for power consumption.

Power consumption during spin-up of an HDD is often the highest power draw of all of the different operating states of an HDD. It can dwarf the power consumption of normal operation. In storage servers with a large number of HDDs, staggered spin-up can be employed to prevent the excessive power consumption of spin-up, which may result in a power shortage. Staggered spin-up starts one drive at a time, either waiting for the drive to signal that it is ready or waiting a predefined amount of time prior to starting the next drive. Many data center designers are concerned about the *provisioned power* of each node, rack, and so on. This is the amount of power that the system must be designed to provide and is generally less than the sum of the worst-case power of every individual subcomponent in the system. When spin-up is staggered, platform power delivery does not need to be designed for this high-peak power across many drives simultaneously, which means that both cost and the potential for compute density improve. However, this comes at the cost of additional potential latency when drives are spinning down for power efficiency reasons.

SATA and SAS Drive Power Management

Power management of drives can be split into two categories: saving power on the actual drive and saving power on the interconnect (PHY). SAS and SATA have many similarities in their power management methodologies and terminology. Power management of the drive itself is significantly more important than the PHY.

SATA devices (drives) support four power states as shown in Table 4-16. These states can save significant power. The Sleep state is rarely used on servers. You would likely not want to use Standby with HDDs on a compute server when any activity is possible, but careful use is possible on large storage arrays.

***Table 4-16.** SATA Device Power Savings*[4]

Action	Description	Receives Commands	HDD	Wake Latency
Working (active)	Normal operation. Fully powered.	Yes	Spun up	N/A
Idle	Active power savings. May take longer to respond to commands.	Yes	Spun up	Milliseconds
Standby	Device still responds to typical commands, but response time may be significant.	Yes	Spun down	<= 30 seconds
Sleep	Device is put to sleep. Must be explicitly woken up.	No	Spun down	<= 30 seconds

SAS power management is conceptually very similar to that of SATA. The PHY supports both Partial and Slumber states with the same characteristics. In the T10 SAS standard, additional states are defined. Although the ATA8-ACS SATA standard only calls for the four states enumerated in Table 4-16, SATA drives may also support similar states to SAS. Table 4-17 provides a summary of these states with ballpark power savings estimates. Note that significant HDD power savings is only possible when significant exit latency costs are accepted. As a result, these power savings modes are typically only deployed after significant idle periods (if at all).

***Table 4-17.** SAS/SATA HDD Power Savings Modes*

State	Spinning	Heads	Power Savings	Exit Latency
Active	Full speed	Loaded	Baseline	N/A
Idle A	Full speed	Loaded	~10%	~100 ms
Idle B	Full speed	Unloaded	~20%	~200–400 ms
Idle C	Reduced speed	Unloaded	~50%	Seconds
Standby Y (SAS)	Spun down	Unloaded	~90%	Seconds
Standby Z (SAS) Standby (SATA)	Spun down	Unloaded	~90%	Seconds

[4]ATA8-ACS Standard. www.sata-io.org/sites/default/files/images/SATAPowerManagement_articleFINAL_4-3-12_1.pdf.

For the PHY, both SAS and SATA supports two low-power modes; Partial and Slumber (see Table 4-18). After some predetermined period of inactivity, either the host or the device can signal the PHY to enter its reduced power state. PHY power management has moderately long wakeup latencies, limiting the ability for fine-grained power savings. Slumber has quite long wake latencies, which preclude them from being used in some server usage models. Partial also achieves idle power on the order of ~100 mW, so further power savings at the expense of latency can be counter-productive at the platform level outside of deep idle platform states. DevSleep is predominantly a consumer device power state.

Table 4-18. *PHY Power States*[5]

Action	Wake Latency
Active (SAS)	N/A
PHY Ready (SATA)	
Partial	<10 µs
Slumber	<10 ms
DevSleep (SATA)	~1 s

SSD drives are common in both consumer and enterprise environments. However, these drives have different characteristics and optimization points. Low-power operation and idle power optimization is critical in the consumer space, and the drives have been optimized for those cases. On the other hand, this has been less of a focus in many enterprise drives. Unlike with HDDs, enterprise SSDs may consume only 25% (or less) of their peak read bandwidth power while running in an Active Idle state (and maintaining fast response times). Traditionally, enterprise SSD procurement costs have dwarfed power costs, and users were not likely to deploy SSDs into areas that would be exposed to significant idle periods. As time progresses, the cost per GB of SSDs is decreasing. This will enable SSDs to compete in usage models traditionally reserved for HDDs and is expected to make power management more of a focus.

Frequency/Voltage

SATA and SAS both have evolved over time by increasing frequencies to provide higher bandwidth. Table 4-19 illustrates how these generations have evolved. Newer generation devices support (by rule) backward compatibility to the prior generation frequencies, and the operating voltage has been held constant.

[5] www.sata-io.org/sites/default/files/documents/SATADevSleep-and-RTD3-WP-037-20120102-2_final.pdf

Table 4-19. *SATA and SAS Generations*

SATA Generation	SAS Generation	Link Frequency	Theoretical Peak Bandwidth
SATA 1.*x*	SAS 1.0	1.5 GHz	150 MB/s
SATA 2.*x*	SAS 1.0	3 GHz	300 MB/s
SATA 3.*x*	SAS 2.0	6 GHz	600 MB/s
-	SAS 3.0	12 GHz	1200 MB/s

SATA and SAS traditionally transfer data serially (1 bit of data at a time) and use 8b/10b[6] encoding, which consumes 20% of the data in order to support the high-speed transmission.

80% * (1.5 GHz / 8 bits per byte) = 150 MB/s

SATA 3.2 includes support for PCIe connected devices, which can leverage the parallel nature of PCIe to achieve even higher throughput. It includes the SATA 3 capabilities for traditional SATA connectivity (at the same bandwidth).

NVMe Drive Power Management

NVMe provides power management capabilities that allow the power to be scaled down at the cost of lower throughput and higher latency. Seven different states are defined (numbered 0 to 6) as shown in Table 4-20.

Table 4-20. *NVMe Power States*

State	Operational	Exit Latency	Performance
0	Yes	--	Peak
1 to 4	Yes	Microseconds	Degraded throughput and latency with increasing exit latency
5	No	~50 ms	
6	No	~500 ms	

NVMe allows the host to manage power statically or dynamically to complement autonomous power management performed by the NVMe drives. When power is managed statically, the host predetermines the power allocated to the NVMe drives and sets the NVMe power state of each drive. When the host manages power dynamically, the NVMe power state of each device is updated periodically to accommodate changing performance and power requirements of the host.

[6] *8b/10b* is an encoding scheme that takes 8 bits of data and transfers it using 10 bits. It is used to transmit data over some high-speed interfaces.

■ **Note** NVMe is frequently used in compute servers that demand peak performance. As a result, aggressive power management, particularly with states 5 and 6, may not be a good match.

NVMe is a relatively new technology. Power management of enterprise-class NVMe drives has not been a priority for many users or designs. Many of the initial enterprise offerings do not implement these power management states.

Power Delivery

There are a large number of components within a server platform, and each of them requires power to provide their necessary function. Different components in the system have different requirements for the type of power that they receive. Some need high voltages, others low. Some are very sensitive to operating at a very specific voltage, whereas others are able to tolerate a range of voltages.

Since the type of power provided to a server system is similar to the power you get from your home's wall outlet, the system has many power converters to convert this AC (alternating current) voltage to the many specific DC (direct current) voltages needed by all its components. The conversion of this AC voltage to the required DC voltages consumes power, referred to as *losses* in the converter. The typical measure of these losses in the power converters is expressed as an *efficiency*. Efficiency is expressed as the ratio of the output power to the input power. Since the input power equals the output power plus the power losses of the converter, the efficiency can be expressed as the following equation:

$$Efficiency = \frac{Output\ Power}{InputPower} = \frac{Output\ Power}{Output\ Power + Converter\ Power\ Losses}$$

How efficiently these power converters convert power from higher voltages to the lower voltages that are required by the loads is critical to the overall efficiency of the system. Even in systems with the best converter efficiency, these losses can make up 10%–20% of the power in the system. At low system utilizations, they can contribute an even higher percentage of the power. This section provides an overview of these power converter losses, basics of the different type of power converters used in the system, the various elements of the power conversions that contribute to their losses, and special features to help reduce losses in these power converters.

Overview of Power Delivery

Figure 4-5 illustrates an example power converter block diagram for a standard dual processor system. Block diagrams like this are commonly found in motherboard schematics.

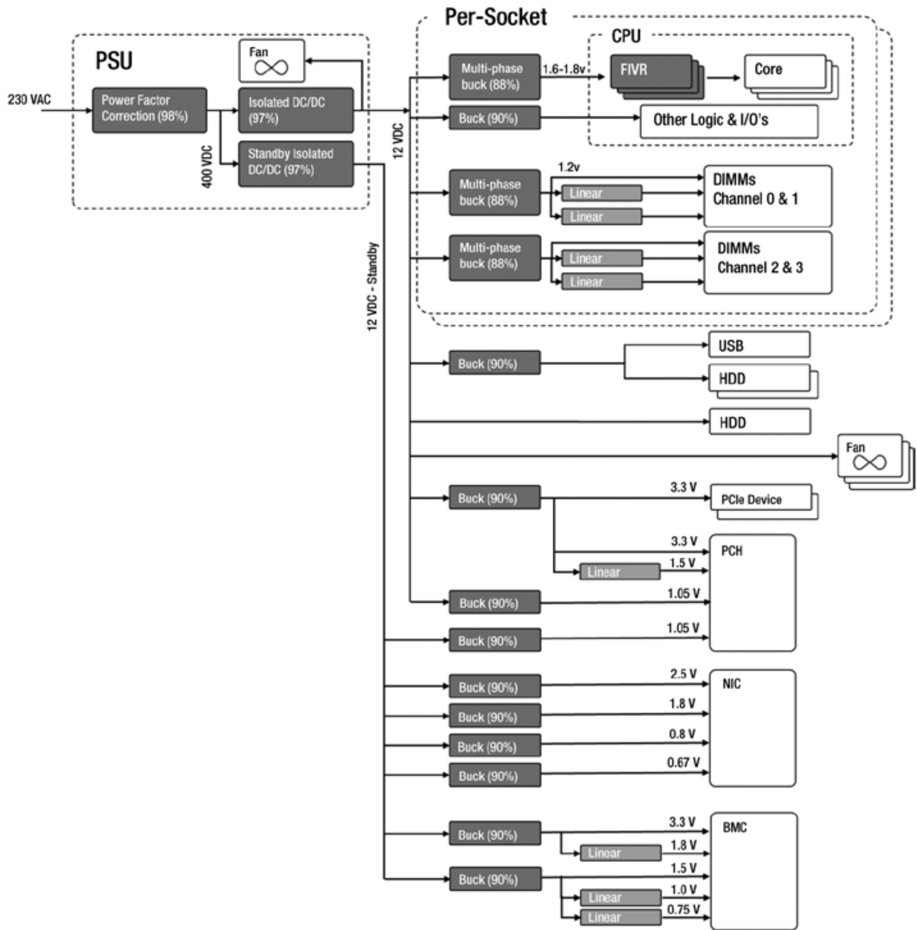

Figure 4-5. *Dual socket power conversion block diagram*

Power is first processed by a power supply and converted from AC to DC (see Table 4-21). The output DC power from the power supply is then converted to the various DC voltage levels required by different platform components (see Table 4-22). Power budgets must be determined for each component in the system so that sizing can be done at each stage of the power delivery network (see Table 4-23).

Table 4-21. *Components in a Power Supply*

Component	Type	Typical Efficiency	Description
Power factor correction	Power converter in system AC/DC power supply	~98%	This is a power conversion stage inside the system power supply whose primary function is to provide power factor correction. This is the first stage converter. It provides a 400 VDC output voltage to the second stage in the power supply.
Isolated DC/DC stage	Power converter in the system AC/DC power supply	~97%	This is a power conversion stage inside the system power supply whose primary functions are to provide safety isolation for the AC input and provide a regulated DC output voltage that can be used by the system.

Table 4-22. *Types of DC/DC Power Converters*

Component	Type	Typical Efficiency	Description
Multi-phase buck	DC/DC switching power converter on the motherboard	80–90%	Power converter used to provide high currents at low voltages. Frequently converts 12 V to 1–2 V.
Buck regulator	Simple DC/DC switching power converter on the motherboard	~90%	Simple converters used to power lower power devices on the motherboard. Typical inputs of 12 V/5 V/3.3 V converting to outputs of 5 V to <1 V.
Linear regulator	Power converter used to power very low power devices on the motherboard	$\frac{\textit{Output Voltage}}{\textit{Input Voltage}}$	Very simple and low-cost converter that provides poor efficiency and therefore is used only for very low-power loads. Their efficiency is determined by the ratio of the output voltage to the input voltage.

Table 4-23. *Power Block Diagram—Loads*

Component	Description
Cores, uncore, DIMMs	These loads in the system are the primary power consumers and provide the core computing capabilities of the system.
LAN, PCH, USB, PCIe	These are lower power loads in the system that provide input and output to the system compute capabilities.
HDD	These are medium power loads that provide storage capabilities.
Active cooling	These are medium power loads that are primarily axial fans in the system. Other types of exotic cooling, such as liquid cooling, are also possible.

The block diagram contains switching power converters, linear regulators, and the loads. Almost all of the system power passes through three to four stages of power conversion to get from the 230 VAC input to the points of load. There are multiple reasons why the power passes through these series stages:

- The first step converts from AC to DC to provide power factor correction. Most digital circuits require DC power for operation.
- It is more efficient to transmit higher voltages over longer distances. This is why power is kept at higher voltages as long as possible.
- Low-power loads are powered by linear regulators. While these regulators are less efficient than more complex voltage regulators, they also have lower cost due to their simple design and small number of components. The loss in efficiency is small in the overall power consumption.
- The easily accessible portions of the platform must not expose technicians to dangerous sources of electricity (such as the AC input).

■ **Note** Transmitting power using higher voltages is more efficient. Wires used for transmitting power have resistance in them. Power is consumed because of this resistance and is proportional to the square of the current (Watts = I^2R). By increasing voltage at a fixed power, current is reduced, thus reducing these quadratic losses (Power = I * R).

It's not uncommon to see a system with a total of 25–30 power converters in the system and on the motherboard. There are a few reasons for these many converters:

- Power efficiency can be improved by adding more regulators in order to reduce I^2R losses.
- Certain legacy functions and capabilities require specific voltage levels.
- System standard components (USB, HDD, PCIe adapters) require industry standard voltages that cannot be changed.
- Customized voltages are required for processors and other silicon devices in order to achieve high performance and power efficient designs.

Power Converter Basics

As discussed earlier, the system contains different types of power converters. Each of these has tradeoffs that can have a significant impact on the overall power efficiency of the platform.

First, energy can be transmitted either as AC (alternating current) or DC (direct current). Digital circuits require DC power to operate. AC power is commonly used to transmit power from power plants across the electrical grid because it is relatively easy (and inexpensive) to change AC voltages using transformers. AC power is used to distribute power within data centers as well, but it must be converted into DC power at some point in order to drive digital circuits.

AC/DC power conversion provides this mechanism. Different components in a platform require different voltages. DC/DC power converters change the voltage of DC power to match the requirements for each component. The main types of power converters used in standard server systems are boost converters, isolated buck converters, single and multiphase buck converters, and linear regulators.

System AC/DC Power Supply

The first components in the node power delivery network, the boost converter and the isolated buck converter, are integrated into the power supply. The basic schematic for these converters are shown in Figure 4-6. It is important to understand the basic functions of these converters to grasp the tradeoffs between efficiency and features.

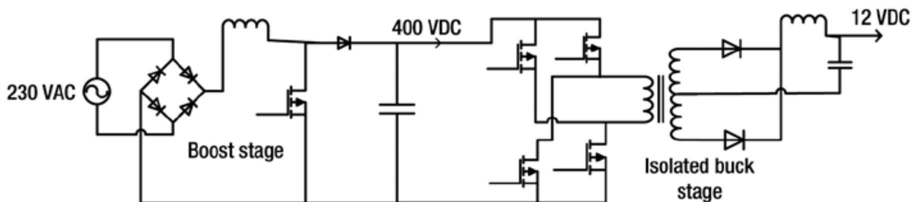

Figure 4-6. *Example AC/DC power supply schematic*

■ **Note** It is not a requirement that AC voltage be used as an input to a platform node and power supply. Other input voltages such as 380 VDC and 240 VDC are being used to help improve facility power distribution efficiency and availability. In these cases, the AC to DC conversion stage is not required and can be removed from the power supply to improve efficiency.

Both converters in the system AC/DC power supply are switching converters, which means MOSFETs are used to chop the input voltage into a square wave, and then they are filtered again to obtain a DC voltage. There is a PWM (pulse width modulation) controller that controls the duty cycle to maintain the required output voltage.

PSUs and the Boost Stage

The boost converter in the power supply maintains a regulated voltage to the isolated buck stage of the power supply. The boost's main purpose is to wave-shape the input current to provide power factor (PF) correction and lower current harmonic distortion (ITHD), resulting in improved power efficiency. Good power supplies achieve PF > 0.99 and ITHD < 5%. Since a boost converter requires that the output voltage always be greater than the input voltage, you typically see a boost output voltage of ~400 VDC ($>110\% * 240\,VAC * \sqrt{2} = 373\,Vpeak$).

■ **Note** Inputs other than 240 VAC are also possible. 277 VAC (one phase of a 480 VAC system) is becoming more common since it can be used in more efficient facility power delivery designs.

PSUs and the Isolated Buck Stage

The isolated buck stage of the PSU provides a few basic functions:

- A regulated output voltage (12 V in most server systems), which is used by down-stream DC to DC converters as well as fans
- *Galvanic isolation* (preventing current flow) between the AC input to the DC output as required by safety agencies
- *Ride-through capability*, which powers the system from its input bulk capacitor during short (½ to 1 cycle) loss of the AC input

The AC ride-through capability is important to keep in mind because this requires the isolated buck stage to maintain regulation on its output over a wider range of input as the bulk capacitor discharges. This tends to make the design of this stage less optimized for efficiency; however, it is required for reliability of the IT equipment.

Redundant Power Supplies

Historically, many servers deployed redundant power supplies and redundant AC feeds to the system in order to improve reliability. Both of these are supported by using multiple power supplies in the system. A common design uses two power supplies. One PSU has enough power to power the system (sometimes at a lower performance), and a second power supply of the same wattage is used in parallel to provide redundancy in case either one fails. This is referred to as a *1+1 design*. The redundant power supplies normally share the load of the system. Since each power supply has its own AC input, this also provides 1+1 redundant AC feeds to the system. With more power supplies in the system a 2+2 or 3+1 redundant configuration is sometimes used. This can scale up to *N*+*N* or *N*+1 number of power supplies; where *N* power supplies are needed to power the system.

In recent years, certain classes of server deployments have focused on improved software resiliency. This is particularly true of large cloud deployments. In such situations, system failure is expected (at some low rate) and the software that executes on the system is robust to handle occasional failures. Conceptually, a problem that used to be solved with additional hardware (and procurement costs) is now being handled in software. Redundant power supplies are not necessary in such designs.

Shared Power Supplies

A single power supply (or even redundant supplies) can be shared by multiple nodes in some designs. A good example of such a design is with microservers. In such a design, the output power of the CPU is routed to multiple sets of voltage regulators associated with different nodes in the rack. In this case, the definition of a platform is somewhat blurred since the PSU is now a shared resource. One drawback of this approach is the *blast radius*, which refers to the number of nodes/components that are impacted if one fails.

PMBus

The power supply has become a key power measurement device in the IT equipment and is used by the facility to see how much power the IT load is consuming. The accuracy of these embedded sensors has improved to +/– 2% over a typical loading range of the system by using special metering IC in the power supply and by using calibration techniques on the manufacturing line. The power sensor in the power supply is used by Intel Node Manager (see Chapter 5) in conjunction with the processor RAPL feature (see Chapter 2) to control the system power. This allows the user to protect the facility infrastructure by guaranteeing that the system does not exceed a predefined limit.

DC to DC Power Converters

Once the power has been converted from a high VAC down to ~12 VDC, a number of additional DC to DC conversion steps are required so that each component in the platform is supplied with the voltage (and current) that they require. There are a number of different types of DC to DC converters that can be used in different situations.

Single-Phase Buck Converters

For medium power loads, single-phase buck converters are used to convert power from 12 V to lower voltages in the system. These may range from < 1 A to 30 A outputs.

Figure 4-7 provides an example schematic for a single-phase buck converter. The PWM controller converts the 12 VDC input to output voltage by switching the high-side MOSFET to chop the 12 V input. Then a DC voltage is reconstituted with the LC filter on the output. A low-side MOSFET is used to allow the inductor current to keep flowing through the output filter. The industry has optimized special components in these converters taking 12 V input to low-output voltages. The high-side MOSFETs are specially designed to handle the high switching voltages with very low duty cycles, and the low-side MOSFETs are optimized for the high duty cycles and high RMS currents demanded by their special requirements in the converters.

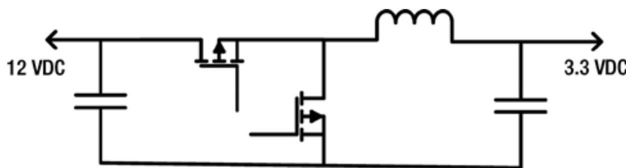

Figure 4-7. *Single-phase buck converter*

Motherboard Multiphase Buck Converters

In a standard server system, more than half of the system power goes to power the processors and memory. For mainstream motherboards, this power is supplied by multiphase buck converters to achieve high performance and small form factors. Requirements of these power converters drive their design to have very fast responses to load changes required by the processors and DIMMs, to maintain a tight voltage regulation on a low-voltage rail as silicon processes reduce their geometries, and to have a small footprint to fit on dense motherboards. All of these requirements challenge the efficiency of the motherboard VR designs. The use of multiphase buck converters has become the method for achieving the best design to meet all of these growing requirements and still maintain good efficiency.

Figure 4-8 shows a simplified schematic of the power stage for a multiphase buck converter. This example shows a four-phase buck converter. PWM controllers are available that have the flexibility to provide anywhere from two to six phases. Some can provide multiple output voltages. These multiphase converters have shared input and output capacitors. The PWM controller switches the phases similar to the single-phase buck; however, the controller switches one phase at a time. Therefore, for a four-phase buck converter, each phase is switched at 90 degrees from one another. This allows the controller to meet the high load transient and high current demands of the processors and DIMMs. The PWM controllers for these multiphase buck converters have added features to shed phases at lighter loads to save power (with no cost to software performance) and serial communication to communicate/manage these high power converters.

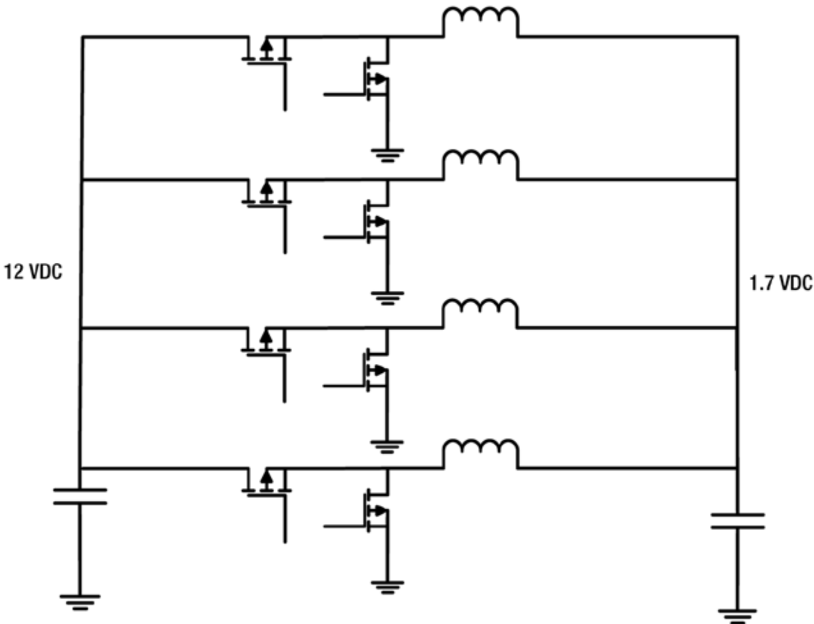

Figure 4-8. *Multiphase buck converter*

SVID

VRs used to power the main CPU voltage rails and DDR memory frequently support SVID (serial VID). SVID is a serial communication bus between the processor package and the voltage regulator controllers that is used for two main purposes to help improve the efficiency of the processor and manage power in the system:

- *Voltage set point:* The processor uses SVID to set the optimum voltage for the motherboard VR to power the processor. This can be used to set the static voltage for a given type of processor for certain rails. This can also be used to set the voltage to the cores in the processor dynamically as the P-state changes (see Chapter 2).
- *Power reporting:* The processor uses SVID to read the power from the VR on some systems. This way the processor can monitor how much power it and the memory is consuming, enabling RAPL (see Chapter 2).

One or more SVID busses can be used per CPU socket. They connect the CPU to all of the SVID-controller regulators that supply power to that CPU (or memory connected to that CPU). The SVID bus is a simple three-wire interface with a clock (the frequency can change across different platforms), an alert (interrupt), and a data wire. Multiple SVID busses may be required depending on the bandwidth demands of the bus for supporting both of the primary usage models just defined.

Motherboard Linear Regulators

Linear regulators convert a higher voltage to a lower voltage by dropping voltage across a series FET (field-effect transistor) operated in its linear range. The output is controlled by the FET's gate voltage. Linear regulators are used sparingly since they are lower efficiency. Their efficiency is determined by the ratio of output voltage to input voltage plus a small quiescent power. Linear regulators can be a good design choice for very low power supplies where losses are not significant in the big picture or when a very small voltage drop is required.

Integrated Voltage Regulators

The Intel Haswell processor integrated the last voltage regulator stages into the processor package with a new capability called Integrated Voltage Regulator (IVR). This added power conversion stage brings with it advantages that outweigh the disadvantage of adding another series power conversion stage.

- *Max current reduction.* Designing motherboard voltage regulators with high maximum currents can be cost prohibitive. By providing the die with a higher input voltage (and by using IVR to step the voltage down for use by the circuits), the max current provided to the die decreases.
- *Higher input voltage to the processor resulting in smaller power delivery losses in the platform.* IVR allows a higher voltage to be delivered to the processor package while still maintaining the required lower voltages at the chips since the IVR power converter controls the chip power. In Haswell, the package input voltage is maintained at about 1.8 V, about twice the voltage needed by the circuits in the package. By running at twice the voltage, the current required to provide a given level of power is cut in half. Lower current results in less voltage loss between the VR and the package (in the platform), improving the overall platform power efficiency.
- *Tighter voltage regulation resulting in lower voltage guardbands and lower power operation.* IVR brings with it tighter voltage regulation at the chips since it is physically closer to the chips. This means the voltage may be kept lower at the chips since less margin for parasitic inductive drops needs to be allowed for—the lower the voltage, the lower the power consumption of the silicon (lower leakage and lower active power).
- *IVR provides cost-effective voltage control of small subcomponents within a die such as an individual core.* This enables features like per-core P-states (see Chapter 2). It also enables the voltage levels to be optimized for each of these subcomponents.

One drawback of IVR is that the power losses that existed in motherboard VRs are moved into the IVR on the CPU die. Although this may result in a net power win for the platform, it does increase the power on the CPU die, which can lead to challenges with thermal density and cooling. When you compare similar SKUs on Haswell E5 (with IVR) and Ivy Bridge E5 (without IVR), you will notice that the TDP power has increased on Haswell. These changes were primarily driven by the increases in CPU power from the IVR integration.

Power Management Integrated Circuit

The term *Power Management Integrated Circuit (PMIC)* is applied to integrated circuits that have multiple power conversion controllers in one small package; they also may contain integrated switches for supporting switching buck converters. The integration of many converters into one package helps to reduce the size and has traditionally been used in the small form factors of mobile devices like phones and tablets. PMICs are now being applied in server computers to help keep the size small while still supporting the many lower power rails on the motherboard. These PMIC may be used to power LAN, PCH, and BMC devices on a standard server board or the memory and processor rails on a microserver with a SoC package. The reason to use PMIC is not to reduce losses in the system, but to reduce the size of the power converters.

Power Conversion Losses

Now that we have reviewed the various types of power converters and their applications, this section will take a holistic look at power converters to understand what causes losses. It will also explore system level design tradeoffs.

Some energy is always lost in transmission through wires because of the resistance in those wires. There are losses due to the currents passing through resistances; these are the condition losses and are proportional to the resistance and the square of the current (power = current^2 × resistance). If we consider only these resistive losses, we would expect the losses at very light loads (like those found when a system is at idle) to drop to very low power; however, this is not the case. We need to consider two other types of losses that occur in switching power converters; these have been commonly referred to as *proportional losses* and *fixed losses*.

Proportional losses are losses in the power converter that increase linearly with the output power of the converter (or output current, since power converters are voltage regulators). These are elements like diodes that have a loss equal to the forward drop of the diode times the current through the diode. The proportional losses of power converters are not very interesting since this is not a dominant element at any load.

The *fixed losses* of the power converter are very significant at light loads. Fixed losses in a power converter are caused by elements such as the switching of the MOSFET parasitic capacitances, the switching of currents through the power components, and the transformer core hysteresis losses in the power supply. These fixed losses are always present whenever the power converter is working. Table 4-24 provides an overview of the types of power delivery losses on a platform.

Table 4-24. *Types of Power Conversion Losses*

Type	Description
Conduction losses	Power losses that are caused by current passing through resistive elements.
Proportional losses	Power losses that are caused mainly by the forward drop of diodes. These are the least significant of the three types of losses.
Fixed losses	Power losses that do not change with the output load on the converter.

Motherboard VRs

Each of the different types of motherboard VRs exhibits different power efficiency profiles. The behavior and efficiency of different types of voltage regulators changes as the load (current demand) of the regulator changes. This section will evaluate different types of VRs by looking at their power losses both in terms of power (watts) as well as their efficiency.

■ **Note** Voltage regulator efficiencies typically appear poor at low utilizations. However, it is important to note that the actual power losses in these conditions are relatively small in absolute terms compared to the losses at higher utilizations.

Single-Phase Buck Converter

Figures 4-9 and 4-10 show the loss curve and efficiency curve for a single-phase buck converter that is capable of 30 A maximum output load and converts 12 V input to 1.7 V output. A second order polynomial trend line of the losses versus the output current closely fits the loss curve. These three coefficients represent the squared ($0.052x^2$), proportional ($0.045x$), and fixed (+ 1) losses of the single-phase buck converter where x is the output current of the VR. The plot also shows these three loss elements separately to see how they contribute across the output load. The fixed losses dominate at lighter loads less than 10 A and the squared losses dominate at heavier loads of greater than 20 A. The effects of the fixed losses cause the efficiency to drop very quickly as load decreases below 10 A, and the efficiency tails off at heavier loads due to the effects of the squared losses.

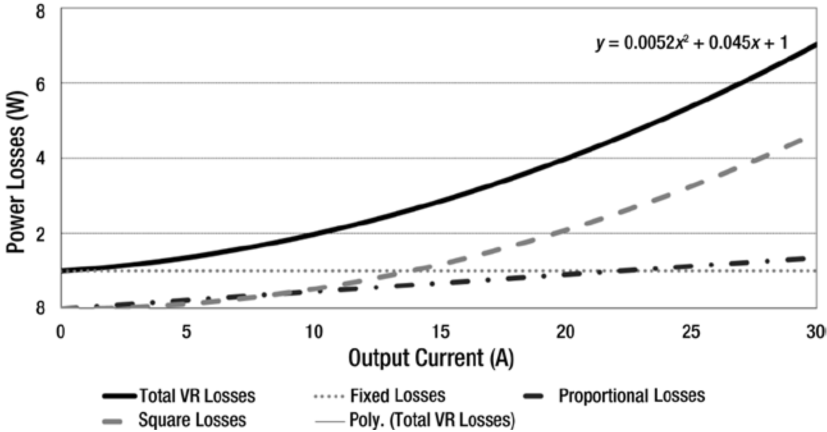

Figure 4-9. Example single-phase buck converter losses

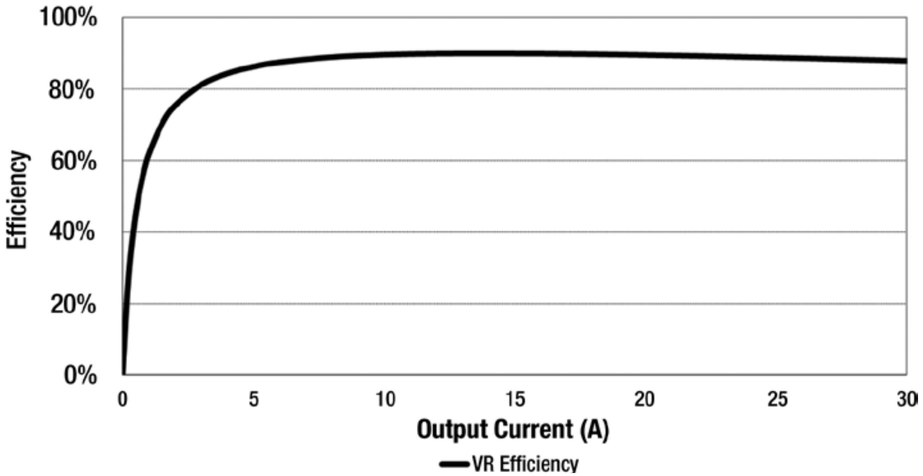

Figure 4-10. Example single-phase buck converter efficiency

■ **Note** VRs must be sized for the worst-case possible current demands of the loads they power in order to avoid system failure. However, typical steady-state current demands, even under heavy load, are common at much lower utilization levels. As a result, the efficiency tail-off that is observed at higher utilizations is generally less significant to the overall power delivery efficiency than the efficiency losses at low utilizations.

Multiphase VR Losses

Multiphase VRs are typically used for heavier loads on the motherboard, like memory DIMMs and processor cores. If we consider the same 12 V to 1.7 V buck converter, but expand it to three phases to support up to 90 A, we see a different loss curve compared to the single-phase converter. Figure 4-11 shows the losses in the three-phase converter are higher than the single-phase converter. This is due to the additional fixed losses from the additional phases and added switching losses of the extra phases.

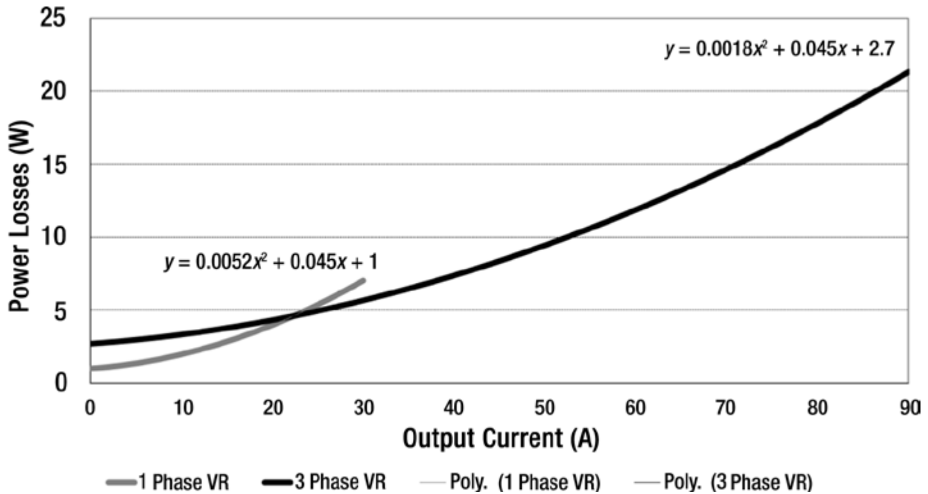

Figure 4-11. *Example multiphase VR power losses*

At loads greater than 20 A the three-phase converter starts to have lower losses. The squared component (from conduction losses) is smaller with the multiphase VR (0.0018 vs. 0.0052). This is because the current now has about a third of the resistance to pass through.

Comparing the efficiency curves shows how the single-phase converter is better at lighter loads and the three-phase converter is better at heavier loads. Note that this example is provided without phase shedding (discussed momentarily), which can improve the efficiency of multiphase VRs at low utilizations.

Phase Shedding

Figure 4-12 illustrates that a single-phase VR can provide better efficiency at low current demands when compared to a multiphase VR that is capable of much higher max current. *Phase shedding* is a feature on some multiphase VRs that is intended to provide the best of both worlds: good, light load efficiency of a single-phase buck converter and the power/efficiency advantages of the multiphase buck converter at heavier loads. In present systems, the phase shedding has mostly been controlled by the processors; the phases are shed when the processors know their power requirements are less than a single-phase

capability. In newer PWM controllers, the controller automatically turns off the phases as it senses the load dropping and turns on phases as loads increase. This is referred to as *auto-phase shedding*. Auto-phase shedding can be far more efficient than CPU-managed phase shedding, because the CPU is not always aware of the immediate current demands and must request phases assuming some worst-case condition.

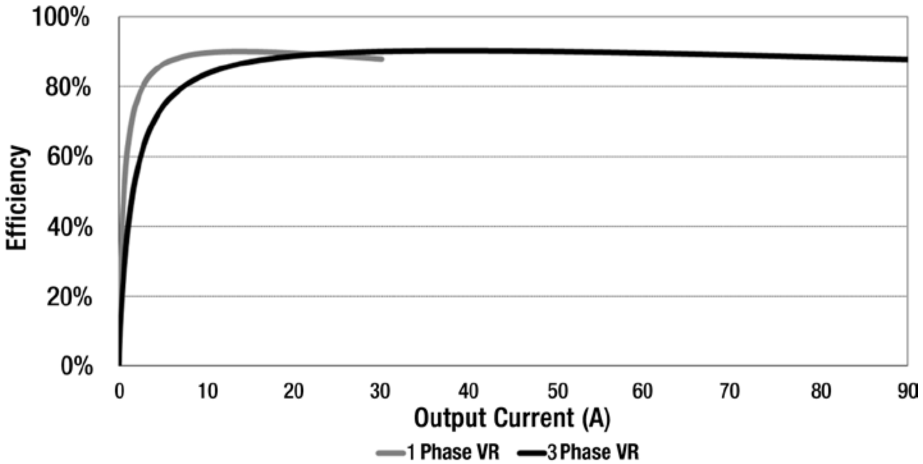

Figure 4-12. *Example multiphase VR efficiency*

Diode Emulation and Burst Mode

Two other methods used to help reduce VR losses at very light loads are *diode emulation mode* and *burst mode*. Diode emulation mode turns off the low-side FET switch and instead uses the body diode in the FET, saving the switching losses in the low-side FET. This is used only at very light current demands. Burst mode reduces losses by skipping switching cycles to effectively reduce the switching frequency of the converter, helping to further reduce switching losses, again at very light current demands.

■ **Note** In typical servers, the current demand does not drop low enough to take advantage of diode emulation, even when the system is completely idle. As a result, this feature is not as commonly supported. Phase shedding provides the bulk of the efficiency improvements at low utilizations in server VR designs.

System Power Supplies (AC/DC)

System power supplies have similar power losses and also can be accurately modeled as a second order polynomial made up of squared losses, proportional losses, and fixed losses. In most server systems, manufacturers offer multiple power supply wattage ratings that can be used in the same system. This allows the users to optimize the cost of the power supply for the configuration they plan to use in the system (e.g., processor performance, memory size, storage size). The selection of the power supply wattage also affects the power consumption of the system. Using a properly sized power supply in the system can help reduce the system power. The use of redundant power supplies in the system to improve availability also affects the efficiency of the system.

Figure 4-13 shows the losses in a 750 W power supply with the fixed, proportional, and squared losses broken out. This is an 80-Plus platinum-level-efficient power supply.[7] As with the motherboard VR, at low loads (less than ~30% of peak), the fixed losses dominate and cause the efficiency to drop off. At moderate to high loads (greater than ~50% of peak), the squared losses dominate and cause the efficiency to drop off.

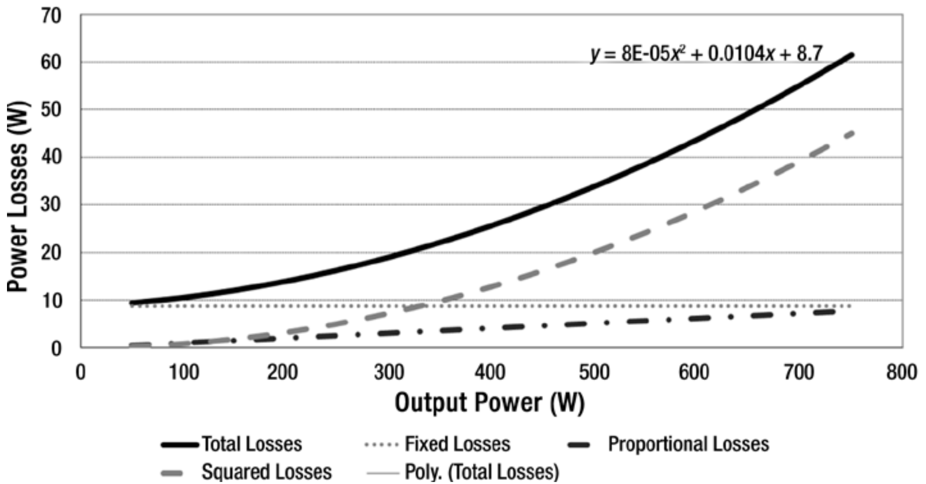

Figure 4-13. *Example 750 W PSU losses (230 VAC)*

■ **Note** PSU losses increase with lower input voltages. The charts in Figure 4-14 are measured with a 230 VAC input (high line). At lower input voltage, such as 120 VAC, the efficiency is reduced by about 2%. This is due to the higher currents in the power factor correction stage of the power supply.

---

[7] >94% efficiency at 50% load based on requirements documented at www.80plus.com.

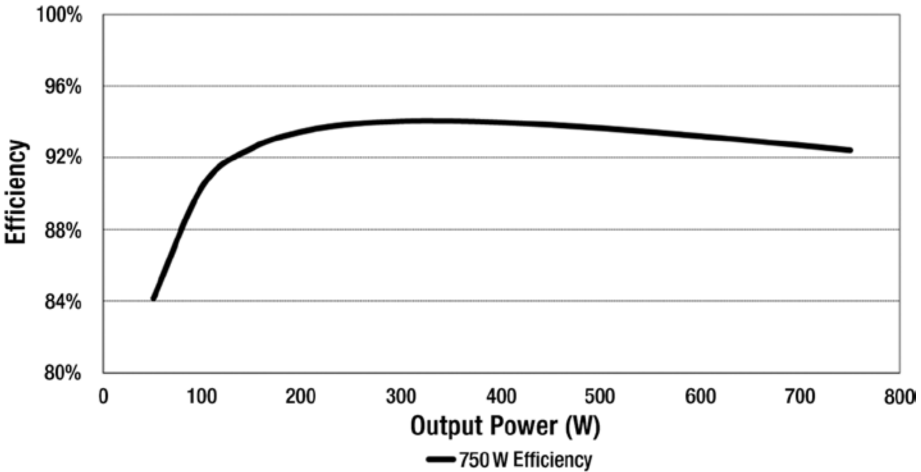

Figure 4-14. *Example 750 W PSU efficiency (230 VAC)*

Right-Sizing Power Supplies

Next we will consider the tradeoffs in using power supplies with higher and lower power ratings. Figure 4-15 shows the loss curves for four different power supplies: 460 W, 750 W, 1200 W, and 1600 W. These are all platinum efficient per 80 Plus.[8] The squared, proportional, and fixed loss coefficients are also shown as a comparison. Two notable conclusions can be drawn from this chart:

- Power supplies with lower wattage ratings have lower fixed losses.
- Power supplies with higher wattage have lower squared losses.

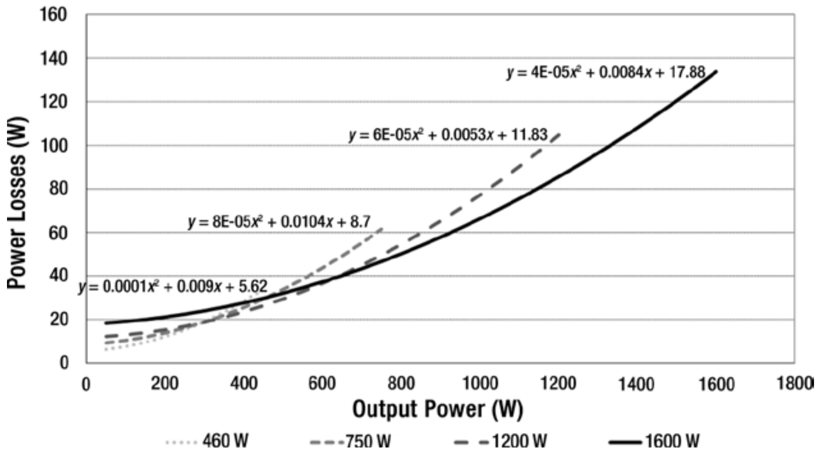

Figure 4-15. *Example PSU losses for different power ratings*

[8]80 Plus is a voluntary certification program for PSUs.

■ **Note** Power supply selection can have an impact on the power consumption of a system. Systems that run at low utilizations will experience the best power efficiency using power supplies that are just large enough for the system. On the other hand, power can be saved by selecting an oversized power supply if system utilization tends to be heavy.

The following two examples show how differently sized power supplies can provide benefits to power efficiency depending on the typical load of the system. Larger power supplies can be more power efficient in systems that, on average, run at higher utilizations.

> *Example 1:* A system load of 100 W on the output of the power supply produces 7.7 W losses in the 460 W power supply, whereas the 1200 W power supply produces 13.0 W losses, a 5.3 W savings using the smaller power supply.
>
> *Example 2:* A system load of 700 W on the output of the power supply produces 55.2 W losses in the 750 W power supply whereas the 1600 W power supply produces 43.4 W losses, a 8.2 W savings with the larger power supply.

Closed Loop System Throttling (CLST)

When right-sizing your power supply to the system configuration and the workloads you plan to run, you must consider the system reliability. If some abnormal condition occurs on the system (like running a higher power workload) the system cannot shut down due to an overload on the system power supply. Many systems running Intel Node Manager and a PMBus power supply have a protection feature called Closed Loop System Throttling (CLST). This feature will throttle the system power/performance if the power supply senses an overload warning condition. This quick reduction in load will protect the power supply from shutting down. Therefore, CLST provides protection against unexpectedly higher system power consumption. This gives you the protections needed to maintain good system reliability while using a lower power supply rating. This throttling is very aggressive and can result in significant performance loss. As a result, it is important that you budget sufficient headroom in the power supply selection to compensate for increases in power demand that may occur over the life of the server (software updates resulting in higher power draw, increased temperatures, etc).

Losses in Redundant Power Supplies

When considering redundant power supplies, remember that the system power supplies will share the system load, which will change the overall power supply losses in the system. Figure 4-16 shows an example of the power supply losses for a 750 W in a non-redundant single PSU configuration along with a redundant 1+1 PSU configuration. The x-axis in the plot is the total load on all power supplies in the

system. You can again see that for heavier loaded systems (> ~500 W, in this example), the redundant 1+1 system with two power supplies in the system have lower losses in the power supplies. And for lighter loads (< ~500 W) the redundant power supply system has higher losses in the power supplies than the system powered by a single power supply.

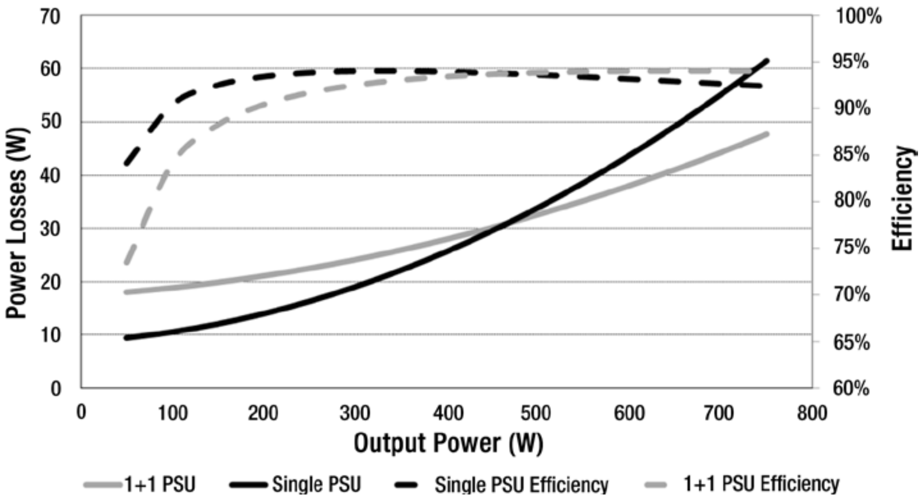

Figure 4-16. *Example PSU losses with redundant power supplies*

■ **Note** System availability and reliability is generally a high priority for server customers. The efficiency loss associated with redundant power supplies is generally an acceptable cost, and many systems make use of redundant power supplies as a result.

The relative efficiency of redundant power supplies versus single power supplies is dependent on the specifics of the power supplies in question, as shown in the following two examples: one shows a more efficient single power supply, and the other lists a more efficient redundant power supply.

> *Example 1:* At a load of 200 W on the system, the single 750 W losses are 14.0 W and the 1+1 power supply total losses are 21.1 W, a 7.1 W savings in losses for the system with a single power supply.
>
> *Example 2:* At a load of 650 W on the system, the single 750 W losses are 49.3 W and the 1+1 power supply total losses are 41.1 W, a 8.2 W savings in the losses for the system with the 1+1 configuration.

Power Supply Cold Redundancy

You can see by the preceding power supply redundancy loss examples that to achieve the best power efficiency in all cases, it would be best to have something similar to what was discussed for the motherboard VR phase-shedding feature. So, at lighter loads, the system runs from one power supply (but still maintains redundancy), and at heavier loads, the system runs from both power supplies in a load-sharing mode. This can be achieved by a feature supported by many server systems today, which is sometimes referred to as *cold redundancy*. In this case, one power supply is powered off into a cold standby state automatically at lighter loads. The cold standby state still allows the power supply to power on quickly if the active power supply fails. This maintains the system power supply redundancy feature. Then, at heavier loads, the cold standby power supply powers on automatically to share the load and maintain the lowest possible losses in the power supplies.

Thermal Management

Typical servers in data centers consume a large amount of electricity, turning it into heat. Extracting this heat from the data center consumes a non-trivial amount of overall data center electricity. Over time, the efficiency of cooling has improved significantly, reducing the overall contribution to power. However, it is still a major factor in energy consumption.

A server cooling system must ensure that each and every component meets its specification. Most components have damage, functional, and reliability temperature specifications as seen in Figure 4-17. A well-designed thermal management scheme must ensure compliance to the specifications while also not over-cooling and wasting power. In most cases it is impractical to design a system to handle every possible workload under all possible combinations of extreme conditions, including fan failures, high-room ambient temperatures, and altitude.

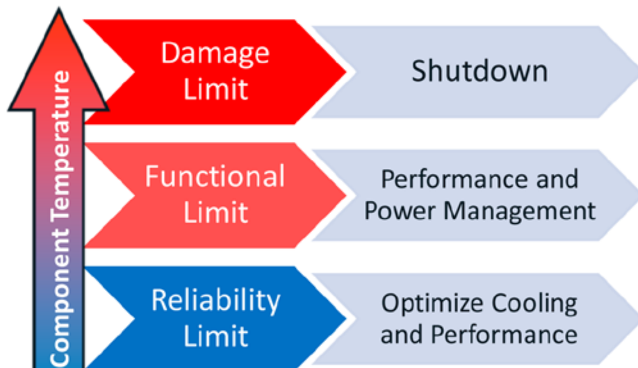

Figure 4-17. *Component temperature specifications and thermal management*

The functional limit is normally aligned with maximum system utilization whereas the server is exposed to a worst-case corner of the allowable range of the environmental class (temperature, altitude, humidity) for which the server has been designed. At lower utilizations, the system is maintained at a lower temperature in order to reduce component wearout that can occur if higher temperatures (at or near the functional limit) are sustained for long periods of time.

A well-designed server will have thermal management to ensure compliance to those specifications either directly through the cooling design implementation, or in combination with the thermal management system. Component temperature is driven by three factors in an air-cooled system defined in Table 4-25: system ambient, air heating, and self heating. These are illustrated in Figure 4-18. Table 4-26 provides some common terms used for heat transfer.

Table 4-25. *Types of Heating*

Type	Description
System ambient[9]	*Inlet temperature of the system* This includes any rack effects, which can increase the temperature delivered to the platform node.
Air heating	*Increase in air temperature due to upstream heat sources in the platform* This is affected by component placement, upstream component power dissipation, air movers, and local air delivery.
Self heating	*Increase in component temperature above local ambient due to the heat dissipated on the device of interest* This is driven by component packaging, power dissipation, and thermal solution (e.g., heat sink).

[9]Defined in the ASHRAE (American Society of Heating, Refrigeration, and Air-conditioning Engineers) Thermal Guidelines for Data Processing Environments.

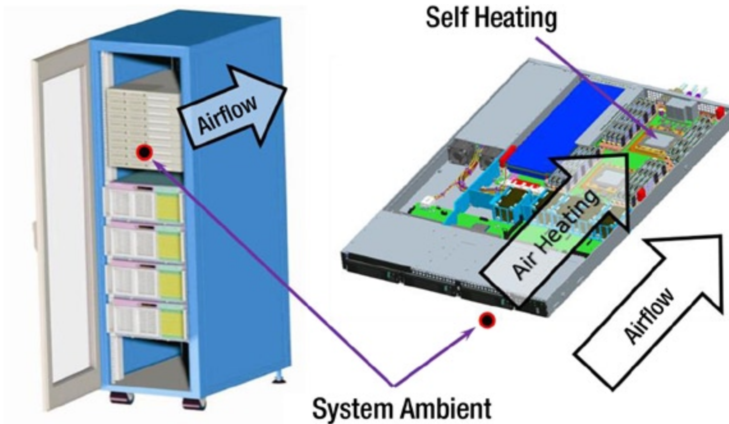

Figure 4-18. *Local ambient, air heating, and self-heating of second socket*

Table 4-26. *Common Heat Transfer Terms*

Type	Description
Conduction	The heat transfer at the molecular level between adjacent particles. Heat from a die is conducted through the surrounding packaging until it is delivered to a heat transfer surface (like a heat sink).
Convection	The heat transfer through random molecular motion and bulk movement of a fluid (or air). Airflow in a server is an example of convection.
Radiation	The heat transfer through electromagnetic waves; generally negligible in server heat transfer due to the dominance of forced convection.

Most server processor dies are connected to a substrate made of FR4 (a glass-reinforced epoxy laminate) enabling simple integration using a socket that enables removal and replacement of the processor. To facilitate heat sink attachment, an integrated heat spreader is attached on top of the package. Component heat is transferred by conduction to the heat spreader and removed by forced convection.

When a heat sink is used on a component, a thermal interface material (TIM) is required to fill the air gaps between the component and the heat sink. The TIM has much higher thermal conductivity than air.

Figure 4-19 shows the typical packaging of a processor with an integrated heat sink (IHS). The IHS serves to protect the die, spread the heat, and provide a mounting surface for a heat sink. Heat is primarily conducted through the first TIM (TIM 1) to the IHS and out through the second TIM (TIM 2) to the heat sink. Thermal paste between the CPU package and the heat sink is an example of a TIM 2. Most servers use forced convection created by a fan to provide higher local velocities, thereby enhancing the convective heat transfer out of the heat sink.

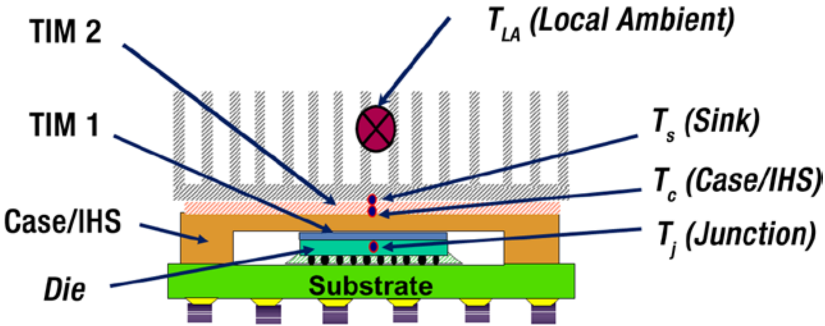

Figure 4-19. *CPU packaging thermal terminology*

When designing an air-cooled system, the thermal engineer must consider a number of factors contributing to the component temperature. Through a careful understanding of the critical components, their specifications, and placement requirements, the thermal engineer can optimize layouts to maintain the lowest cost, highest efficiency, and highest performance solution. So-called shadowing of components results in significantly increased cooling difficulty and the lowest cooling capability. Shadowing implies that the air heating in the following component temperature equation will be relatively high, resulting in costly thermal solutions and high fan power.

Components with high power density (power/area) require thermal enhancement, such as a heat sink or heat spreader. Either of these devices spreads the heat to a larger surface area enabling significantly improved convective heat transfer.

The following equation describes how power, air heating, and ambient temperatures impact the temperature that is exposed to the package. The actual silicon die (and transistors) are exposed to even higher temperatures than the T-case.

$T_C = \Psi_{CA}$ x Power + System Pre-heating + External Ambient

where

- T_C (T-case) is the case temperature of the component.
- Ψ_{CA} (psi-CA) is the thermal characteristic of the heat sink as measured from case (C) to ambient (A) and in units of °C/W. The lower the Ψ_{CA}, the better the thermal performance of the cooling solution, since the component (case) temperature will be closer to the ambient temperature at a given power consumption.
- Power is the power dissipated (consumed) by the component.

Component and heat sink convective thermal performance is proportional to the inverse of airflow, as shown in the example characteristics shown in Figure 4-20. This means that the cooling efficiency (Ψ_{CA}) improves significantly with airflow up until a point (somewhere between 10 and 15 CFM in the illustration). After that point, significant increases in airflow (at high power cost) will only provide small benefits to the actual cooling. Fan power is proportional to the cube of airflow (and fan speed). Operating in the conduction-dominated region of a heat sink can significantly increase the power consumption (and inefficiency) of a system.

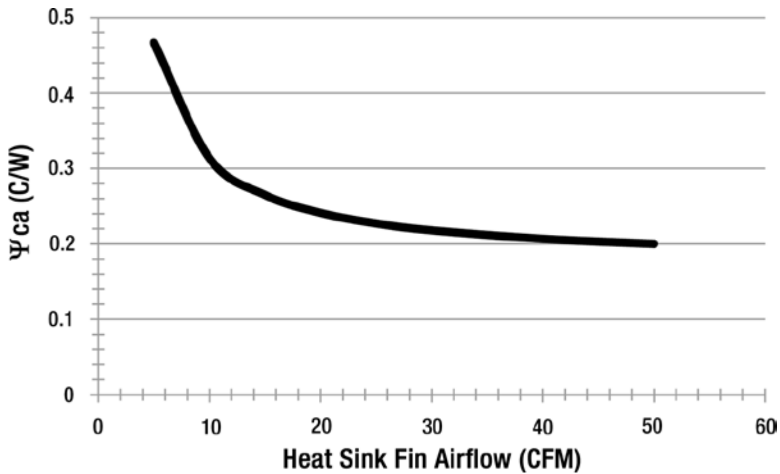

Figure 4-20. *Example heat sink performance*

■ **Note** Cooling efficiency is non-linear with airflow. Significant increases in fan speed (and fan power) may only yield slight improvements in cooling once a heat sink has reached its maximum capabilities.

System Considerations

The platform design team must carefully consider the components, configurations, usage models, environmental conditions, and the system-, rack-, and room-level airflow protocols to achieve an optimal cooling solution. These design considerations must be evaluated against the cost, performance, and energy objectives of the solution.

■ **Note** Running a system at higher temperatures will increase the leakage power of the CPU (and other devices in the platform). However, the power savings from running with reduced cooling typically far exceed the increases in device leakage power.

Component selection and placement detail will drive the design and consequently are the most critical elements to consider during the design phase. One example is the selection of memory technology to be supported. An entry-level server designed to support the highest capacity and frequency memory could burden the system design with expensive fans that are never needed by most customers. All components must be similarly considered including the power range under which the components must function.

■ **Note** Increased CPU leakage power from higher temperatures can reduce turbo performance. However, the performance returns from over-cooling a platform are generally not large and can be cost prohibitive. Reducing the temperature by 20°C may only increase performance by a few percent (if at all).

In many platform designs, component placement is primarily driven by electrical routing considerations. Lengths between key components must be minimized to ensure signal integrity and meet timing requirements. Placement for thermal considerations matters but is not the foremost driver during the board layout process. The thermal engineer must provide the guidance to the board design team to enable solutions that have a reasonable chance for success while not necessarily being thermally optimal. Examples of systems that vary in cooling difficulty are shown in Figure 4-21, where the system on the left has thermally shadowed memory and processors whereas the system on the right does not. *Thermally shadowed* refers to a component being downstream from another component in the airflow. In such a design, the shadowed components are exposed to higher temperatures.

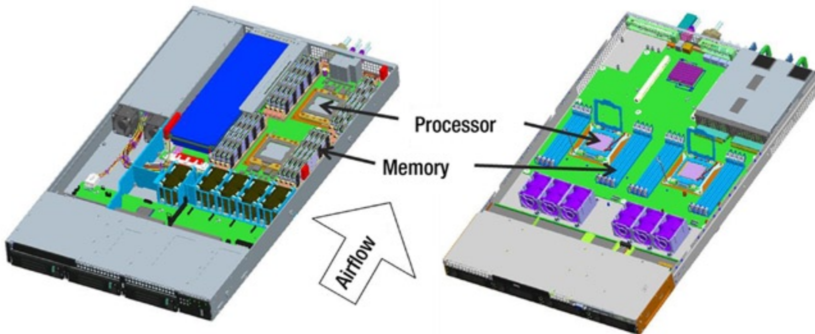

Figure 4-21. *Example board layouts*

Thermal shadowing is commonly used in dense multi-socket platform designs. Because the first processor heats the air before it gets to the second processor, the ambient temperature of the second processor is higher. The cooling solution must compensate for this increase in ambient temperature. This frequently results in more expensive heat sinks and higher fan speeds, which increases both procurement costs and power consumption. The thermal requirements of higher densities come at a power/performance efficiency cost.

■ **Note** Components that are in the thermal shadow of other high-power components frequently operate at higher temperatures and therefore consume more leakage power. With CPUs, this increase in power can result in different levels of turbo being achieved if equal power is allocated across the two sockets.

The design engineer must thoroughly understand the expected airflow paths and optimize the airflow delivery accordingly to maximize energy efficiency of the thermal subsystem. Selection and usage of the air moving devices must be matched and designed to the server design. Tradeoffs between air movers and heat sink design must be performed to find the optimal design points for both. The cooling performance, power consumption, acoustic signature, fan reliability, and redundancy features are important characteristics that factor into the overall solution.

Component Thermal Management Features

Power management features are used to perform power-performance limiting that enables a component to stay within temperature limits. Sensors create the data necessary to trigger power management. Processors, memory, and some chipset components contain sophisticated thermal management capability and are discussed in the following sections.

Processors

Processors have three high-level temperature points as shown in Table 4-27. These temperature values vary across different products, and the values shown provide an example of typical values.

Table 4-27. *Notable Processor Temperature Levels*

Level	Description	Typical Temperature
TCONTROL	Above this temperature, fans should be running at full speed in order to ensure the long-term reliability of the processor.	~5°C–10°C below prochot
	Between TCONTROL and PROCHOT, the fans will all be running at full speed. These conditions typically occur when the processor is running at full utilization and ambient temperature is high.	
PROCHOT (DTSMAX)	Max temperature at which the processor functionality is guaranteed. Autonomous thermal management algorithms inside the processor (see Chapter 2) will work to ensure that this level is not exceeded.	~80°C–100°C
	Between PROCHOT and THERMTRIP, processor functionality is not guaranteed. Data corruption (silent or detected) or system hang may occur.	
THERMTRIP	Catastrophic shutdown temperature. Above this temperature, irreversible physical damage may occur to the processor. This is protected by a combination of the processor and the platform.	~125°C

Memory

Similar to processors, power management features are used to manage potential excursions above unsupportable temperature limits on DIMMs. Because the memory controller is now contained in the processor, the processor determines the memory's thermal state and activates power management features. Thermal sensors on the DIMMs are accessed by the processor, and memory traffic regulation can be activated as needed.

The data retention time of DRAM devices used on DIMMs is temperature dependent. Increasing the memory refresh rate allows operation at higher temperatures. Operation at that higher temperature is called extended temperature range (ETR) and is supported by most volume DRAM manufacturers. By enabling higher temperature operation, one can reduce cooling costs of the platform. This does come at a small cost of DIMM power from the extra refreshes and some small performance loss, but the resulting fan power savings implies higher power efficiency at the platform level. As memory temperatures increase beyond the ETR threshold, memory thermal throttling features in the CPU will engage. See Chapter 3 for more details.

Platform Thermal Management

Thermal control enables optimization of system performance as a function of usage, configuration, cooling capability, and environment. Underlying this optimization is the parallel use of fan speed control and power management to meet the customer's requirements. Some customers may desire maximum performance and may not want to be concerned with cooling costs, while others may be willing to trade off a small amount of performance under certain circumstances in order to achieve better power efficiency (and lower cooling costs).

Components and their specifications are the primary drivers in a server's thermal design (e.g., heat sink, fan selection, and airflow management). The thermal engineer can create a superior thermal design, but without a thermal management system to provide real-time optimization, that design may be acoustically unacceptable or highly inefficient. True superiority quite often lies in the thermal management scheme and its capability for delivering precisely the performance needed and meeting the component specifications while consuming the lowest amount of power.

Platform thermal management enables optimization in four areas:

- Operation within component thermal limits
- Maximization of performance
- Minimization of acoustic output
- Minimization of wall power consumption

All server components are designed to handle thousands of thermal cycles due to the natural temperature variation that occurs as a result of workload demands. Servers can go from idle to high usage many times a day and must be capable of years of operation under this type of variation, resulting in wide temperature extremes on the components.

The thermal management (TM) system manages component temperatures, performance, power consumption, and acoustics using two primary mechanisms:

- Fan speed control (cooling delivery)
- Component power-limiting features (e.g., P-states and memory throttling)

With some servers the initial setup during boot time enables the end user to configure the system to preferentially favor acoustics, power efficiency, or performance.

Thermal Control Inputs—Sensors

Sensors provide the inputs to the control scheme. Table 4-28 provides an overview of some types of sensors used for platform thermal management.

Table 4-28. *Types of Sensors Used for Thermal Management*

Sensor Type	Description
Direct temperature	On-component sensor(s) found on processors, memory, hard disk drives (HDDs), chipset, GPGPU, etc.
Indirect temperature	Off-component, discrete sensor used to directly measure air or board temperature. This information can be correlated to components without sensors.
Power or activity	Power can be used to estimate the temperature of different components of a platform (in conjunction with platform characterization and other temperature sensors). It can also be used by algorithms to optimize the overall platform power.
Fan speed	Used for ensuring that a fan is operating within design parameters.
Fan presence	Used for detecting whether a fan is populated (e.g., redundant configuration).

Different components in the platform use different types of sensors for monitoring temperature (see Table 4-29).

Table 4-29. *Platform Component Sensor Types*

Component	Sensor Description
Processor	A server processor has many thermal sensors but only exposes the max temperature to the platform thermal management. Multiple sensors are strategically placed to enable the processor's own power management features to engage as necessary to ensure that silicon temperature does not exceed the point to which the processor was qualified and tested, but also to eliminate inaccuracy in determining actual die hot spot temperature.
Memory	Most server DIMMs have an on-PCB (printed circuit board), discrete thermal sensor. Thermal sensor temperature is highly *correlated* with DRAM temperature, thereby enabling a single sensor to cover all components on the DIMM. Some DIMMs have a buffer, which may also have a separately accessible thermal sensor.
Chipset (and other silicon devices)	Many silicon devices have an accessible sensor for use in TM. Some limited thermal management may be available locally on these devices, but they are used primarily for fan speed control and catastrophic protection.
Hard drives	Hard drives contain thermal sensors that are accessible through a drive or RAID (redundant array of inexpensive drives) controller.
Voltage regulators	Nearly all high-powered voltage regulators have a local thermal sensor to protect the components in the VR region. Historically this has been primarily for high-temperature protection.
PCIe cards	In some cases the PCIe card supports sensor capability, which is available to the server for thermal management. However, this is not common and, as a result, cooling must be sized to handle any possible card that can be installed. Indirect sensors are sometimes used to infer PCIe temperatures.
Power supplies	Most power supplies have their own cooling (internal fans) and manage their cooling without system intervention.

Figure 4-22 provides an example of how thermal sensors are distributed across a platform.

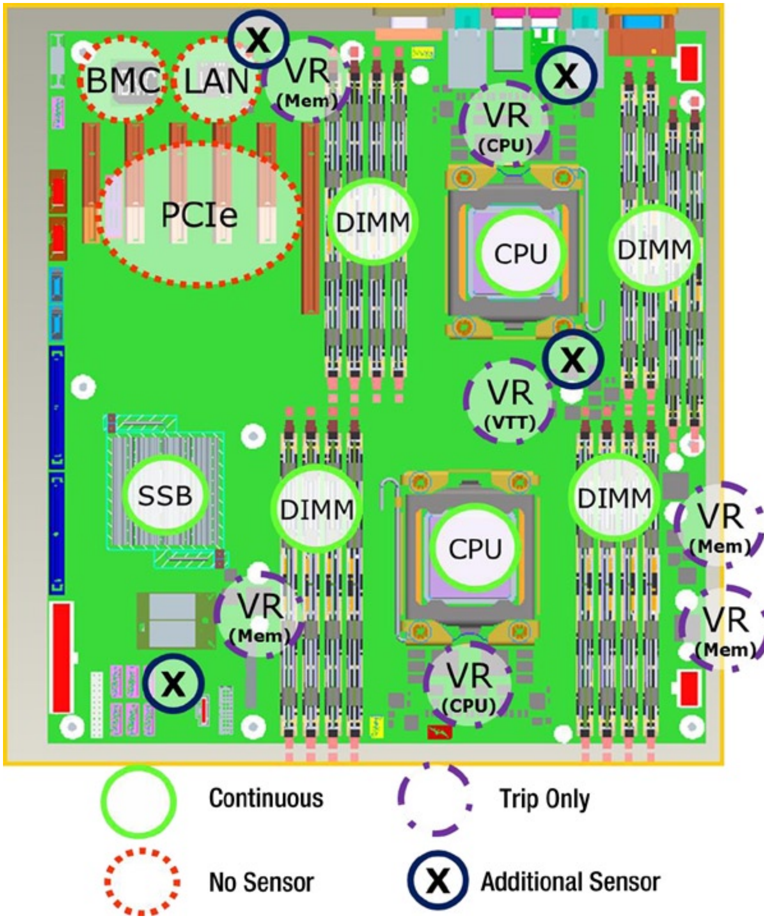

Figure 4-22. *Example platform thermal sensor layout*

Voltage Regulators

Voltage regulators (VRs) can be made of multiple discrete components on a board, and a thermal sensor is generally placed near the component that is expected to exhibit the highest temperature. In the past, the sensor primarily provided functional protection with little available usage for fan speed control. More fine-grained TM implementations have the capability for using the VR sensors in fan speed control algorithms. VRs can support

- OTP (over-temperature protection), which results in an immediate shutdown
- Prochot, which is a connection to platform Prochot to cause the CPU to engage in heavy throttling when the VR is hot
- VRHOT, which is an alert on SVID to tell the CPU to throttle

Power Supplies

Power supplies have their own thermal sensors and fans that are used for autonomous thermal management. They have temperature protection mechanisms that can shut down the system when a catastrophic condition is detected. The power supply's fans can supplement the server's cooling in certain conditions. As a result, the platform TM system can sometimes override the power supply fan control in order to drive higher fan speed as necessary to cool system components.

Fan Speed Control and Design

Optimizing the speed of fans in a system can result in significant improvements in power efficiency. Simply running the fans at max speed is an easy way to ensure that the system operates within its specifications and provides the maximum performance, but this comes at a significant energy cost.

System designers have proprietary fan speed control algorithms that run in their BMCs; these attempt to minimize fan speeds while staying within the component specifications. Multiple algorithms can be used simultaneously with the final fan speed to be determined by comparingthe results of these algorithms.

Multiple (i.e., tens of) sensors are used in the algorithm with the required fan speeds set based on the components with the least margin to their specifications. The algorithm must ensure that unacceptable fan oscillations do not occur, even at low fan speeds. These could be just as annoying to a customer as a continuous loud noise.

■ **Note** It is possible for a system to transition from low-power consumption and corresponding low fan speeds to a very high-power workload in microseconds. Although the CPU die does not heat up instantaneously due to the higher power utilization, it may heat up faster than the fan speed control subsystem can increase the fan speeds, resulting in a short period of CPU thermal throttling. Fan speed control algorithms that are heavily optimized for energy efficiency can be more exposed to this type of behavior.

Fan or cooling zones are often used to precisely adjust specific fans to the needs of components most coupled with those fans. Cooling zones can be proximity based, or physically separated. The extent of optimization versus cost is considered when designing cooling segmentation created through cooling zones. Using a fan zone

implementation enables total fan power and acoustic to be minimized. Fans in a non-stressed zone can run at lower speeds than those needed in a more highly stressed zone. A stressed zone implies that at least one component is approaching its temperature limit. Figure 4-23 shows two examples of fan zones mapped to two different boards designed for use in a 1U chassis.

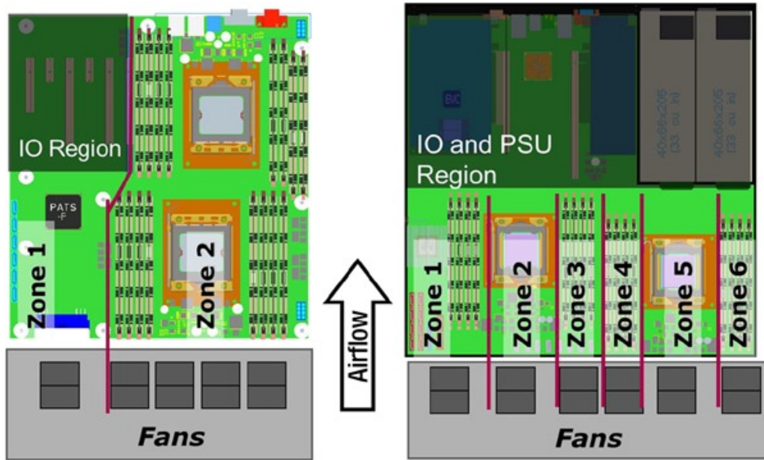

Figure 4-23. *Fan zone mapping*

Each sensor is mapped to the fan zones depending on its thermal connection to that zone. A single sensor could impact a single zone or multiple zones depending on positioning and ducting. By mapping specific components to specific fan zones, more granularity in fan control can be obtained, thereby reducing total fan power.

A variation on a proportional, integral, derivative (PID) controller is commonly used for fan speed control. For each thermal sensor or group of thermal sensors, a separate PID algorithm is running. At each time step, a new fan speed setting is determined from the PID controller using temperature value from each sensor. The management controller determines the actual fan speed setting based on the maximum calculated fan speed setting from all the simultaneously operating PID algorithms.

Fan speed settings normally have a *floor*, preventing operation below the levels necessary to cool components without sensors where the real-time temperatures are unknown.

Summary

Each CPU requires a large amount of support hardware in order to complete its tasks. Data centers are generally made up of a large of number of racks. Each rack contains a number of separate platforms (or chassis) that provide one (or a few) compute nodes. In addition to the CPUs, a platform includes the memory, drives, networking, power delivery, cooling, and more that is required for enabling a small number of CPUs to operate. Similar problems and tasks must also be managed at larger scales in a rack or even across a data center. As an example, each platform likely has dedicated cooling hardware and thermal monitoring and management capabilities. However, additional cooling is necessary for extracting heat from the rack and ultimately the data center. This topic is discussed in Chapter 9.

A wide range of platform designs are possible, with different optimization points for different usages. A storage platform may include a massive number of drives all connected to a single two-socket (DP) node. A compute node may have a single high-speed network connection, no drives, and some amount of memory that has been selected for the types of workloads that run on that node. Large EX platforms tend to have high CPU TDP powers. However, their support hardware also tends to have high costs (both for power and procurement), amortizing the cost of that additional power and making it very cost efficient. Similarly, low power offerings can be very effective if high performance is not needed.

The power/performance efficiency of a CPU is heavily dependent on the platform around it and the demands of the workload in question on that platform. Measuring performance per TDP watt of a CPU can be a very misleading statistic due to the many platform components that contribute to power. If the task that is required is bottlenecked by drives, adding more compute performance (and increasing power) will be inefficient. Similarly, if a platform includes significant power outside the CPU, and increasing the CPU power results in an increase in overall platform performance, it may be more power efficient for the platform to increase the CPU power even if the CPU performance per CPU watt decreases. Building a power efficient platform requires balancing the compute needs with the capabilities of the supporting platform devices.