Power Management in the Cisco Unified Computing System: An Integrated Approach

What You Will Learn

During the past decade, power and cooling have gone from being afterthoughts to core concerns in data center construction and operation. Industry participants have improved the efficiency of almost all the elements in the power and cooling chain, from individual chip efficiency, to system design, to efficient data center cooling. Currently, users can buy the most efficient server designs ever available in terms of throughput per watt, and the macro design of new data centers is vastly improved over those of even a decade ago.

The weakest link in the system remains the management of power at the system and rack and row levels, where much of the power consumption occurs. While system vendors can manage power within a single chassis, until now no vendor has provided a comprehensive and usable solution to rack- and row-level power management, especially for cases in which flexible management of power peaks is required. The Cisco Unified Computing System™ is the first solution to meet this challenge.

The Cisco Unified Computing System provides power management across groups of Cisco® UCS chassis and racks that can enable throughput improvement of up to 30 percent for a traditional data center at no additional capital cost. This capability is integrated into the Cisco UCS architecture.

This document explores the following topics:

- What is power capping and why is it a good idea?
- Why is power capping rarely used even though it has been available for years in limited implementations?
- How does the Cisco Unified Computing System solve the power management problem?

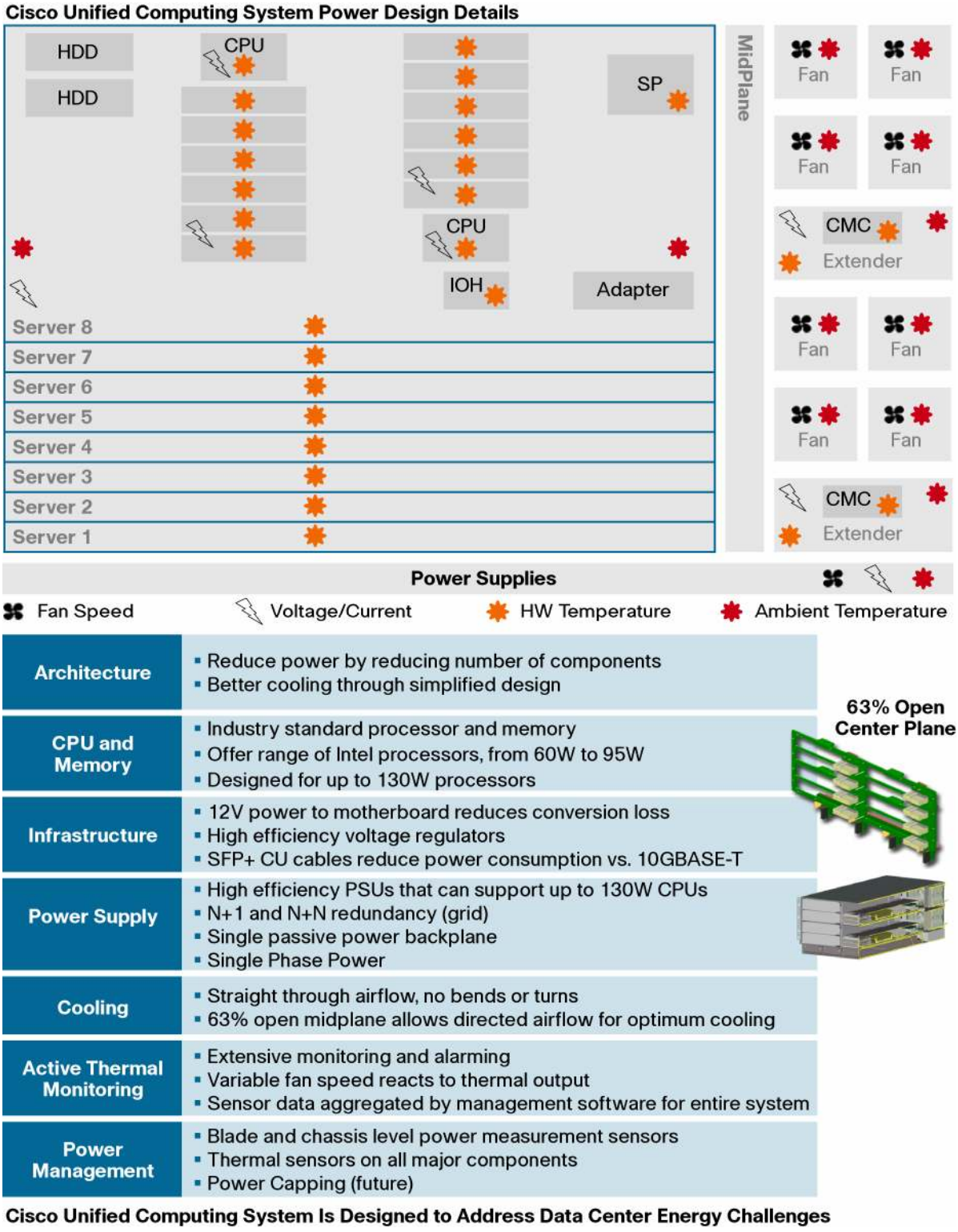
Multiple Places to Save Power: Not All Are Equal

Before looking in depth at power management and capping, you should understand the context within which Cisco has been working. The fundamental goal of reducing power consumption per workload unit has been the subject of multiple-year and multi-vendor effort, with innovations occurring at multiple points within the data center as well as within the computer systems themselves. While some elements of the power and cooling chain are outside the current purview of a system vendor such as Cisco, within the bounds of the system there are multiple opportunities for power and cooling efficiencies.[1] Figure 1 shows the technologies that the Cisco Unified Computing System has implemented to reduce power consumption at the server and chassis levels.[2]

[1] This document uses the terms “power” and “power and cooling” somewhat interchangeably in discussions of the overall problem, since cooling is additive to and directly correlated with power consumptions. When the document uses “power” in its literal sense to refer solely to electrical power, the meaning will be clear from the context.

[2] The power consumption challenge is even more complicated than what is shown here. The server exists within a larger context that includes power distribution and the data center’s power and cooling architecture. From best to worst cases, these extrinsic (to the server) factors can make almost a 2 to 1 difference in overall power utilization efficiency (PUE), a metric of data center power efficiency.

Figure 1. Cisco UCS Power and Cooling Technology



Some of the technologies shown in Figure 1, such as power supplies, voltage regulator modules (VRMs), CPU, memory, and fans, are available to all vendors, and while Cisco has taken good advantage of them, there is in fact little differentiation between vendors in these areas.[3] Innovation in other aspects of power management, such as overall cooling and airflow architecture and the capability to actively manage power among system modules, can yield greater benefits, and this is where Cisco adds value.

What Is Power Capping?

Power capping is one of the main differentiators of the Cisco Unified Computing System. This feature provides increasing benefits as each individual Cisco UCS instance scales. Power capping is the capability to limit the power consumption of a system, be it a blade server or a rack server, to some threshold that is less than or equal to the system's maximum rated power.

For example, if the maximum power rating of a blade server is 340 watts (W), but the power available to the chassis is only 3334W AC, which is sufficient to supply an average of 300W per blade, plus the chassis, in the Cisco UCS chassis, each blade can be capped at a maximum of 300W per blade to avoid exceeding the capacity of the power supply. This type of capping is known as *static power capping*. Although it helps ensure that the chassis will never draw more power than allowed, it does not take into account that the various blades may have varying loads, and at any given time one blade may not be using its full allotment of power while another may require more.

Another type of capping, *dynamic power capping*, allows the power management system to allocate the total pool of power across multiple blades in a chassis. With dynamic power capping, the system as a whole can conform to a specific power budget, but power can be steered to the blades that have higher load and require additional power.

To date, dynamic power capping offerings on the market have been limited to a single blade chassis or chassis as their managed power domain, as discussed previously. The following sections describe how Cisco has extended dynamic power capping across multiple blade chassis and implemented it in a fashion that is more useful to operations management than the other, traditional alternatives.

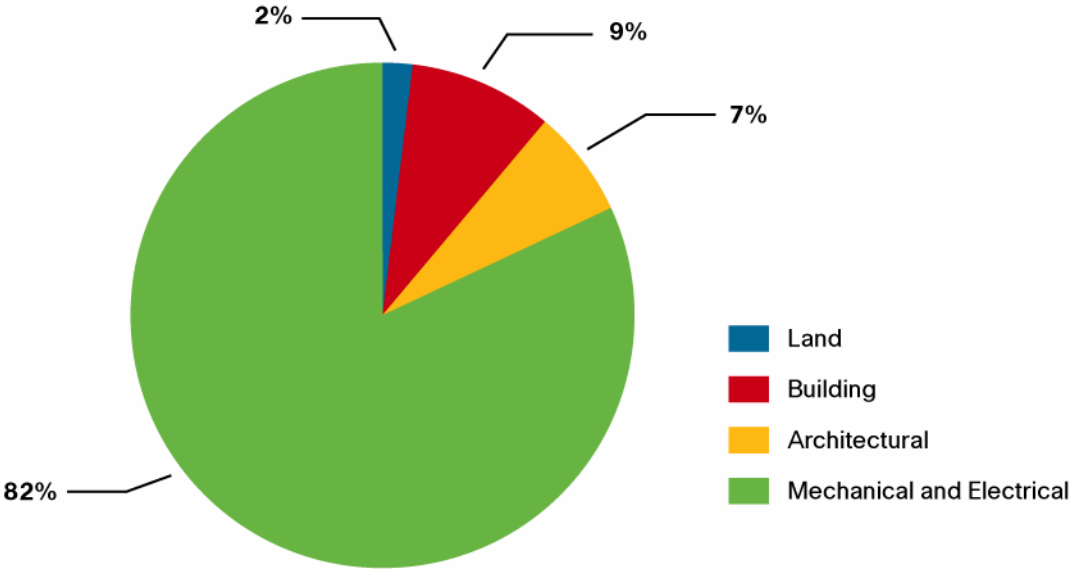
Why Use Power Capping?

The first step in understanding the importance of power capping is to look at where the money really goes in building and running a data center. Outside of the data center industry, many people believe that the greatest costs in a data center are associated with space: land, buildings, etc. This belief has led customers and vendors to increase computing density with such solutions as one-rack-unit (1RU) servers and blades, which simply pack more into the same floor space, reducing what is perceived as the most expensive data center item.

While space may have been the major cost factor a decade ago, as larger data centers have come online, it has become clear that the single greatest cost in a data center is actually the power and cooling infrastructure (Figure 2).

[3] For example, processor power states, or P-states, are built into the CPU by the vendor and allow the CPU to be switched between different power levels by the OS or other software utilities. While one vendor may have a slightly better set of tools to modify P-states, the differences are small and always transient.

Figure 2. Capital Cost Components in the Data Center



Source: C. Belady and G. Balakrishnan, Microsoft, Incenting the Right Behaviors in the Data Center, http://events.energetics.com/datacenters08/pdfs/Belady_Microsoft.pdf.

Looking at these cost metrics, you can immediately see that reducing overall power or doing more work within the design power envelope is probably the most efficient way to reduce data center capital costs.[4]

In regards to operating costs (including depreciation), while servers in a data center are the single greatest expense, the combined costs of the power infrastructure and power for the servers make up the next-greatest operating expense in a large data center. Details of data center operating costs are shown in Figures 3 and 4.

[4] The “Mechanical” component in the pie chart in Figure 2 is mostly related to cooling pumps, fans, computer room air conditioning (CRAC), adaptive airflow, etc.

Figure 3. Data Center Power Consumption Breakdown

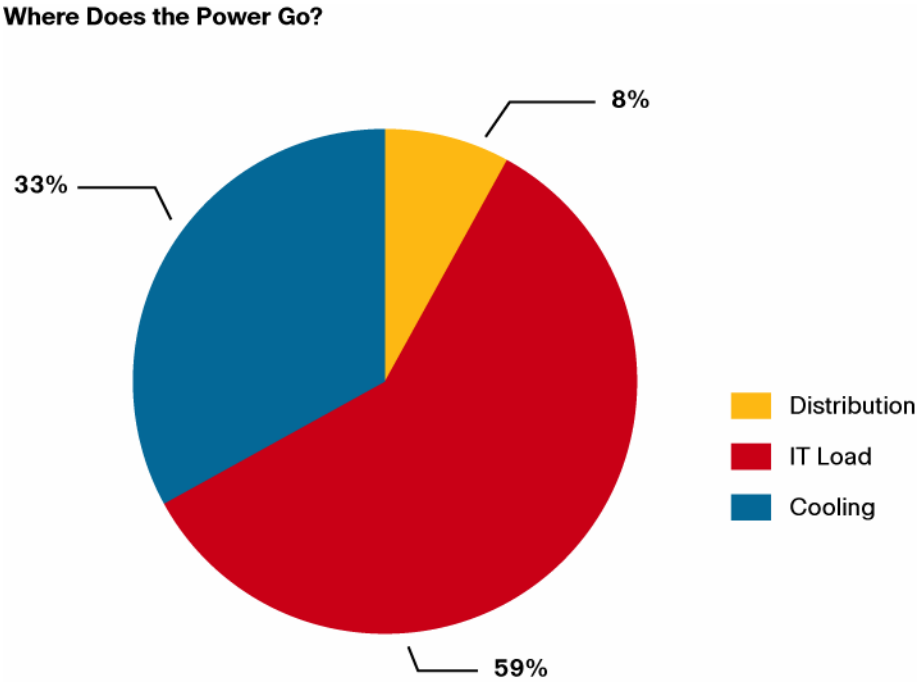
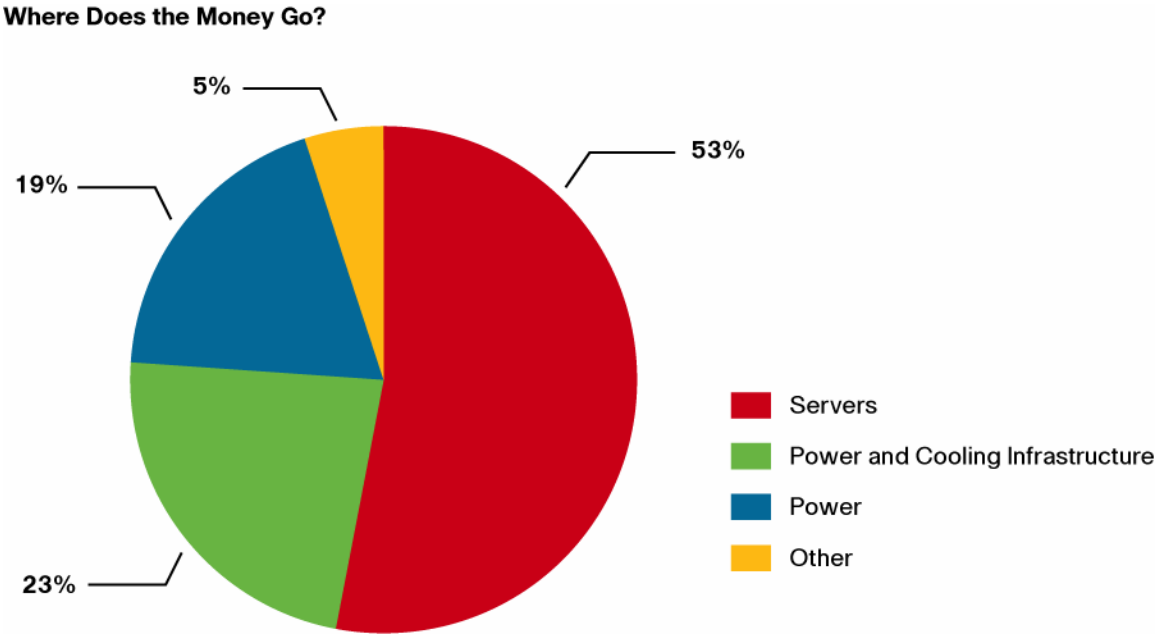


Figure 4. Operating Expense Contributions by Infrastructure Components in the Data Center

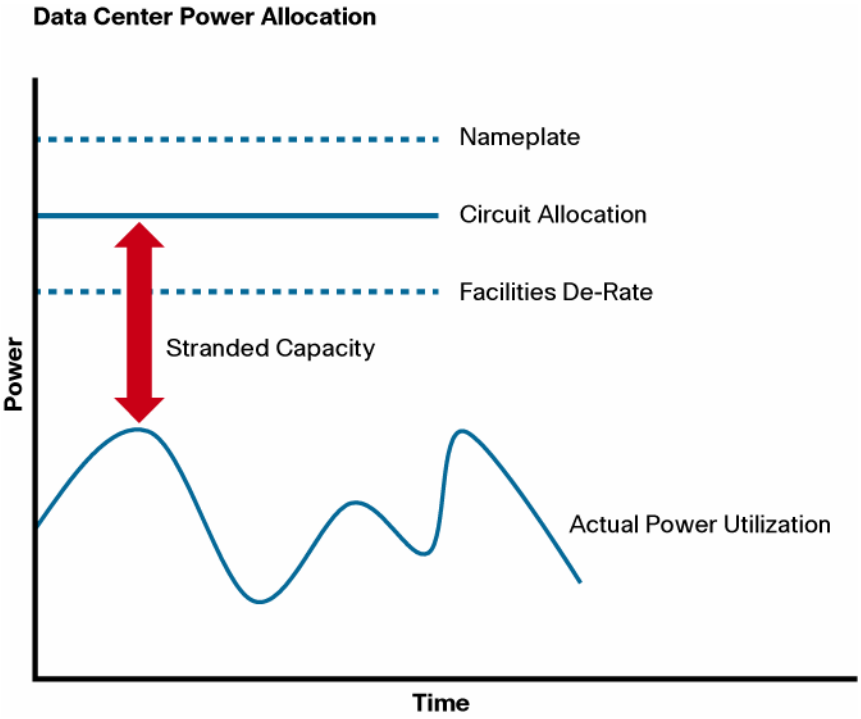


Source: James Hamilton http://perspectives.mvdirona.com/2010/09/18/OverallDataCenterCosts.aspx

Power Provisioning: An Exercise in Inefficient Allocation

Compounding its role as a major capital and operational cost component, power in traditional data centers has historically been inefficiently provisioned and allocated. Although facilities designers have moved beyond the practice of provisioning to nameplate ratings and have learned to provision based on estimates of actual power consumption, the inability to intelligently manage power consumption at the granularity of a desired circuit or chassis still leads to considerable stranded power capacity in modern data centers, as shown in Figure 5.

Figure 5. Data Center Power Inefficiently Allocated



Reading: http://loosebolts.wordpress.com/2009/06/02/chiller-side-chats-the-capacity-problem/

Given the costs related to power as the dominant contributor to capital expenditures (CapEx), its strong contribution to operating expenses (OpEx), and the relative inefficiency of its provisioning and management, today's data centers need to optimize the use of power.[5]

Data center efficiency is measured by power utilization effectiveness (PUE), which is the total power input to the facility divided by the power that reaches the IT load. Older data centers often had a PUE rating of 2 or greater, whereas the current goal is approximately 1.5 or lower for new construction. Unfortunately, most of the factors that influence PUE are extrinsic to the server design, although architectures such as the Cisco Unified Computing System, with their inherently extensible management framework, can more easily fit into future integrated data center management schemes than today's traditional servers can.

[5] When considering a holistic view of the data center, you must not lose sight of the fact that management is still the dominant operating expense for a collection of servers, which is addressed by the Cisco UCS management architecture. Power remains a contributor to OpEx and is a strong predictor of CapEx, as noted earlier.

Three major variables affect the optimization of power usage in a data center:

- **Cooling and distribution system efficiency:** In Figure 3 earlier in this document, you can see that the cost of cooling the system is a sizable fraction of the cost of the energy used to perform the workload.
- **Efficiency:** Given that overall PUE is outside the purview of the server designer, the main parameter that the server design can alter is the processing efficiency per watt. This parameter is largely influenced by two factors: how efficiently a single server node can perform as measured by a variety of standardized benchmarks, and the overhead of any shared infrastructure such as the fans and power supplies in a blade chassis.
- **Capacity utilization:** Utilization is where power capping comes into play, since the capability to intelligently ration power across a collection of servers increases overall efficiency by reducing the amount of over provisioning from power distribution equipment through system power supplies. As discussed in more detail later in this document, the amount of power required with a well-engineered power capping scheme and that required for a system without such a scheme can differ by a factor approaching 100 percent.

In addition to these efficiency considerations, important operational aspects to power capping need to be considered, most notably circuit protection. While IT users and server administrators may be concerned with the amount of power per server and per chassis, IT operations are more sensitive to total power per circuit[6] and the capability to protect circuit breakers from tripping.[7]

To optimize the use of power and cooling capacity, you need to understand the actual power use in the data center. In most enterprise data centers, most servers are idle most of time, and most servers never have a full power load. Therefore, similar to the way that airlines oversell seats because they know that in reality not everyone will turn up for their flights, a data center can assume that there is effectively no chance that all workloads in the data center will simultaneously go to full power load, and that even at 100 percent capacity most workloads are not consuming maximum power. Thus, data centers are able to oversubscribe available power to allow more computing resources to be used.

Unfortunately, this traditional and relatively predictable data center power model is changing in as a result of more scalable web-facing workloads. In traditional data centers running a mix of enterprise applications, it is statistically highly improbable that all servers will go to maximum load. However, as users increasingly implement web-facing applications on a large scale, the chances of power spikes based on sudden traffic fluctuations becomes highly likely. When such a workload spike hits, the data center may simply shut down unless there is a mechanism in place that can shed load and manage power in some way. This potential for wider swings in power consumption than in older environments complicates attempts to oversubscribe power allocations and recapture some of the stranded capacity.

To put this capacity oversubscription into practice, new technologies such as power capping are being introduced that provide the necessary technical features. However, simply having the technology available is not enough to implement this scheme effectively. Integration of power capping technology into the IT governance process is also required.[8]

[6] To be exact, operations people are concerned with the total current on a circuit, since circuit breakers are current-sensitive devices. In most cases current is linearly correlated with power, so the terms can be used interchangeably.
[7] The tripping of a circuit breaker in a data center is a major event. Aside from the worst-case scenario of a cascading failure that takes down the entire data center or a part of it, such an even can cause transient power surges that may disrupt other equipment. At best, it is highly visible as a major operational event, with attendant management visibility, and is to be avoided at costs.
[8] Unfortunately, this is also true of many other advanced technologies. Unless governance models and operational processes and sometimes business processes are changed to take advantage of them, much of their potential value will not be achieved.

Why Power Capping Is Not Widely Used

Currently neither fixed nor dynamic power capping is widely used, although the basic technology, static power capping, has been available for several years. There a number of reasons for this but the main ones relate to competing objectives and priorities of the major stakeholders with respect to implementation of power caps. The perspectives of the major stakeholders are summarized here:

- **Business end users and application owners:** Their objective is the best performance at all times for their applications at the least cost. They often have little visibility into the details of IT operations. The required governance change to educate users is a combination of better communications regarding their solution resource utilization and some form of chargeback so that their business units understand the true costs of providing their desired levels of performance and availability. In many cases, proper education will convince them that effective power capping is not a threat to their performance service-level agreements (SLAs).

  Additionally, most organizations already apply the concept of business-critical and noncritical applications. It is reasonable for the IT manager to be able to ask the business owner to define the priority of an application and use that information when looking at power capping. The application classification should also be the basis for differential charge backs: it is reasonable for users to pay more for critical applications.
- **Facilities and data center managers:** Their primary objective is to protect the facility. If a technology such as power capping were to fail, a group of servers that breaks its power cap could cause a circuit breaker to trip, causing outages and, in the worst case, a cascading failure in which the workload transfers to other servers, causing their breakers to trip, and so on.[9] Other objectives, including increasing data center capacity utilization within the constraints of power, cooling, and floor space resources, are secondary to uninterrupted operation. For these managers, power capping must be shown to reduce the risk to their facilities and to have unimpeachable reliability over and above any other benefits.
- **IT executives:** These stakeholders need to balance the application users' wants and needs against the facilities' capabilities to provide the best return to the business on its data center investment. With proper cost-focused metrics to support their decisions, senior IT managers can become powerful agents for change and balance the inherent conservatism of operations managers.

Other factors contributing to the underutilization of power capping are tools and complexity. Typically, the IT, facilities, and applications management groups have different tool sets that do not communicate or share any common data, making any unified monitoring, reporting, or event correlation and response strategy difficult to implement[10].

Due to this lack of a common tool set, setting the power caps is a significant and time-consuming operation. Setting a cap too low could significantly affect the performance of the workloads on the capped servers, so each server needs to be monitored over a relatively long period of time to determine where to set the power cap for that particular server. After the caps have been set, ongoing monitoring is required to help ensure that as workloads change, the caps still meet the required application SLAs. This requirement for ongoing monitoring and cap setting is strengthened as an organization moves to a more fluid and dynamic virtualized data center.

[9] From an operational perspective the objective is to manage the power cap at a circuit level. All the servers that are attached to a common circuit, typically through a rack-based PDU, should be managed so that even if all those servers were to peak in power consumption simultaneously, the breaker would not trip. The facilities manager will always want to control the absolute cap values expressed as watts or amps.

[10] Most companies cannot answer the fundamental question in power management: "What is the impact on power consumption of an X% increase in users, transaction rate, or other application metric?"

How Does the Cisco Unified Computing System Solve the Power Management Problem?

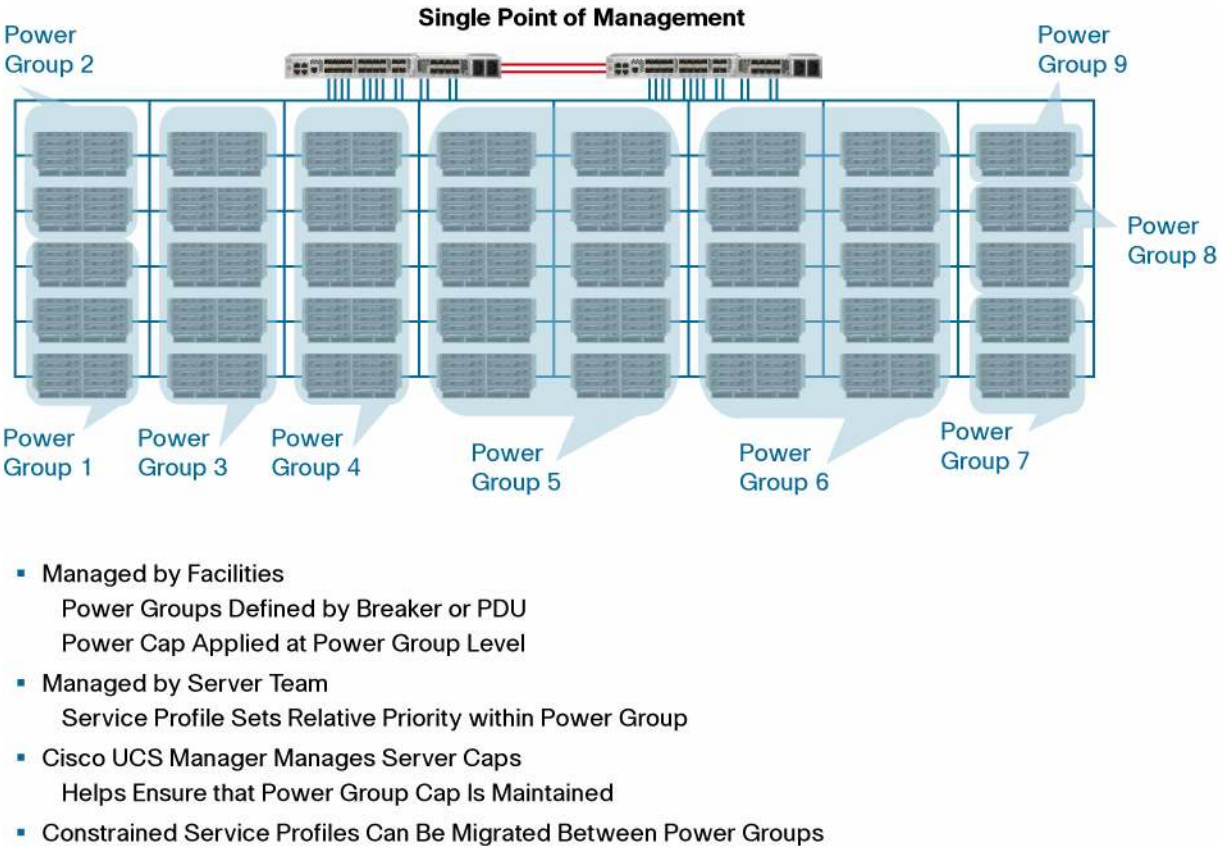
Power Cap Groups

Rather than implementing conventional power capping, either static or dynamic, on an individual chassis or server, the Cisco Unified Computing System implements a powerful new concept called a power cap group. A power cap group is a collection of servers or chassis that share a common power allocation or budget, which is then applied in a dynamic manner to servers within the group.

Power cap groups (or domains) follow a simple set of rules:

- All servers within a single chassis are part of the same power cap group. There can be multiple chassis in a single power cap group.
- The chassis within a given power cap group do not have to be physically contiguous and can be located anywhere as long as they are within the same logical Cisco Unified Computing System instance. This use case is not common, but it is available for management of spot cooling problems.
- All chassis in a power cap group do not have to be connected to the same distribution circuits. In the usual case, all members of a power capping group are connected to the same circuits, but this case can be used to manage to a global cooling limitation for which power is not a primary limiting factor.
- Any Cisco Unified Computing System instance can have multiple power cap groups, containing varying numbers of chassis (Figure 6).

Figure 6. Power Cap Groups Across Multiple Chassis



The Cisco Unified Computing System enables power cap groups through the Cisco UCS Manager embedded management software, as shown in Figure 6. The capability to implement power cap groups is unique in the industry, and the capability to define them on a power circuit or on arbitrary physical boundaries brings new features to IT operations management.

Associating Power Caps with Virtual Servers

Currently, all other products on the market, whether offering fixed or dynamic power capping, associate power caps with specific slots in a chassis or to a physical rack server, even in products in which there is some virtualization of the physical server itself. In an environment in which the identity and workload of a given physical sever can change, thus changing the server's power cap requirements, associating caps with physical entities or fixed locations is inadequate.

The Cisco Unified Computing System solves this problem by incorporating power cap information in the service profile associated with a server (see the brief discussion of service profiles later in this document), thus helping ensure that the power cap settings follow a workload as it moves, and that a new physical server inherits the correct power cap when it is associated with its service profile. Figure 6 shows the basic concept of power cap groups distributed across multiple Cisco UCS chassis.

Power Caps and Service Profiles

A Cisco UCS service profile is one of the fundamental abstractions upon which the Cisco Unified Computing System is built. A service profile is the logical encapsulation of the server identity and physical resource requirements, such as the LAN and SAN addresses, number of Ethernet and storage interfaces, firmware versions, boot order, network VLAN, physical port, and QoS policies. (More detail can be found in Understanding Cisco Unified Computing System Service Profiles.) By taking this base concept and adding power as another resource associated with the service profile, Cisco abstracts the management of power caps away from the individual server in the same way that physical identity management is abstracted away from the physical server.

Each service profile can be assigned a power cap policy that defines the relative power priority of a physical server that is associated with that power profile, and the power capping group to which the server belongs. When there is adequate power for all servers, the priorities do not come onto play. In the event that the servers in a given power cap group begin to exceed their group allocation, power is allocated according to the priorities defined in the attached power cap group policy, ensuring that critical loads are throttled last. Additionally, there is an option to designate a server as having no power cap, for workloads whose performance cannot be compromised even to the minor extent that power capping entails.

Management of power caps is distributed between Cisco UCS Manager, the chassis management controllers (CMC) in each chassis, and the Cisco Integrated Management Controller (Cisco IMC) in each blade.

For a single chassis cap UCS manager simply assigns the cap to chassis. If a multiple-chassis group cap is defined, Cisco UCS Manager intelligently divides the power amongst the chassis in the group based upon hardware configuration and service profile priority. This division of power is dynamic in that if configuration changes power may be re-apportioned among the chassis in the group.

Each CMC is then responsible for ensuring that the chassis maintains the power cap that it has been assigned by UCS Manager. The CMC for each chassis in the group will allocate a power cap to each blade based on the blade type, configuration, and relative priority. The blade Cisco IMC then is responsible for managing that particular blade to ensure that the blade's power cap is maintained.

Once a power cap has been assigned to each blade, the Cisco IMC control algorithm will ensure that any blade is brought under its power cap within 500 ms. Hence, the group power cap will be maintained and avoid tripping the circuit breaker.

If a blade in a chassis reaches its power cap and stays at the cap for 20 seconds[11], the CMC in the chassis will reallocate power if there are blades that are not currently using their allocated cap. Simply put, idle blades will have their power caps lowered and active blades that have reached their power caps will have their power caps increased by a corresponding amount (subject to the priorities defined in the associated service profile). In the event that all blades in the chassis are at their power cap, the power caps will be returned to their initial allocations. In future releases of UCS Manager, an additional level of reallocation will occur between chassis. This process will be less frequent if the high-priority blades are distributed across the group of chassis, and more frequent if they are located in the same chassis. Users need to be aware of this distinction, because the reallocation of power between chassis is a slower process, taking 30 seconds to 1 minute.

Separation of IT and Facilities Roles

Another critical element that must be addressed by an effective power management and control solution is the necessary separation of roles between the IT administrator and the facilities manager. Cisco UCS Manager can separate roles built into its fundamental architecture, giving the facility manager the control to set the absolute power cap values on individual chassis or groups of chassis. The IT administrator sets the relative priority of the power cap policy of each service profile so that as servers are associated with a service profile, the system knows how to allocate power between the servers automatically based on the absolute power cap values that have been set by the facility manager.

Allowing the facility manager to set the absolute values for the power cap group enables protection of the data center at the circuit level, protecting against isolated and cascading failures. At the same time, the IT administrator can work independently with the end user to allocate power for the workloads within the power cap group without further involvement from the facilities manager, delivering computing power where it is needed most and where it fits best in the overall business plan.

Power Capping and External Management

The Cisco Unified Computing System includes multiple power and thermal sensors that provide detailed power consumption and temperature information about the behavior of the systems. Each blade is instrumented with a power sensor that provides information about the power consumption of the blade, as well as with multiple thermal sensors. Additionally, the chassis and power supplies can provide the power consumption details for each chassis.

All this management and monitoring information is available through the XML interface that is built into Cisco UCS Manager. It can be consumed by third-party management software from vendors such as BMC, CA, HP, and IBM as well as by custom tools that IT and facilities management may already have in place, and it can be used to manage power intelligently across multiple management domains.

Power Capping and Cost Savings

The net effect of implementing proper power capping using Cisco UCS power cap groups is an increase in the overall computing capacity and throughput of the data center within the current power constraints. These increases result in both savings on monthly power bills and the avoidance of major capital investments because the lifespan of the data center is extended, delivering computing headroom to help meet both current and future needs.

Cisco has observed overall throughput improvements of 10 to 30 percent, depending on the details of the specific data center, the composition of the workload, and the policies used to provision power and operate the data center, through the capability to install more servers into the same power footprint using group power capping with the Cisco Unified Computing System.

[11] The response time is designed to prevent hysteresis problems, in which spikes cause a constant power reallocation. A sustained load increase is required to cause the Cisco Unified Computing System to reallocate power.

Conclusion

Cisco Unified Computing System technology has redefined the enterprise computing environment. By breaking the traditional data center model and redefining the data center infrastructure as pools of virtualized server, storage, and network resources, the Cisco Unified Computing System has delivered a new computing model with advantages in capital and operational cost, improving flexibility and availability, and reducing the amount of time needed for IT to respond to business changes.

By adding power capping technology to the management infrastructure of the Cisco Unified Computing System, data centers can provision power based on actual usage rather than on theoretical server maximums or power supply capability. In addition, by using priorities within the service profile, data centers can direct power to the most important workloads in environments where power is constrained.

For More Information

- **Cisco Unified Computing:** http://www.cisco.com/go/ucs
- **Cisco Unified Computing System (UCS) Manager:** http://www.cisco.com/en/US/products/ps10281/index.html
- **Cisco on Cisco IT:** http://www.cisco.com/web/about/ciscoitatwork/index.html
- **The Energy Efficient Data Center—Cisco Systems:** http://www.cisco.com/en/US/netsol/ns980/index.html#~products
- **Understanding Cisco Unified Computing System Service Profiles:** http://www.cisco.com/en/US/prod/collateral/ps10265/ps10281/white_paper_c11-590518.pdf

Appendix: The Truth About Server Power Consumption

Use care when correlating CPU utilization and power consumption. A server at 100 percent CPU utilization is not necessarily using its maximum power. Modern CPUs such as the Intel Xeon and AMD Opteron processors have hundreds of millions of transistors on a die, and what really affects power consumption is the number of those transistors that are actually active. When a program is running, the individual instructions activate a number of transistors on the CPU, and depending on what the instruction is actually doing, a different number of transistors will be activated. For example, a simple integer register add would use only a relatively small number of transistors. A complex instruction that streams data from memory to the cache and feeds it to the floating-point unit would activate millions of transistors simultaneously. Both sequences could potentially show 100 percent CPU utilization but would consume very different amounts of power—in some cases, more than 60 percent more for the more complex workload.

The net result of these differences is that application power utilization varies depending on what the application is actually doing and how it is written. Differences can even been seen when running the same benchmark depending on which compiler is used, whether the benchmark was optimized for a specific platform, and the exact instruction sequence that is run.

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **www.cisco.com/go/offices.**

Printed in USA C11-627731-01 02/11