IDC
Analyze the Future

WHITE PAPER

Solutions for the Datacenter's Thermal Challenges

Sponsored by: IBM and Intel

Jed Scaramella Matthew Eastwood

January 2007

IDC OPINION

Today's datacenter managers are required to place an increased emphasis on the power and thermal characteristics of their IT systems. The electricity cost of operating the infrastructure can no longer be viewed simply as a cost of doing business, as the expense to power servers is rapidly growing and the thermal capacity of the infrastructure is a limiting factor in the expansion ability of the datacenter.

The emerging concern of power and cooling has added to the complexity of the datacenter in terms of both operational and management costs. As such, datacenter managers have a real requirement for energy-efficient systems and power management solutions.

The following are key considerations that datacenter managers should consider regarding power and cooling costs associated with hardware in the datacenter:

- While capital expenditures for server hardware have moderated, operational expenses have escalated. The expense to power and cool is becoming a significant cost factor within the datacenter. Today, for every $1.00 spent on new hardware, an additional $0.50 is spent on power and cooling, more than double the amount of five years ago.

- Datacenters at their power and cooling thresholds are unable to support new server deployments, a fact that severely limits the expansion of IT resources. Customers in these situations must undertake costly retrofitting or new buildouts to deploy additional systems.

- Current environments do not provide optimal conversion of power and cooling into actual compute capacity. Customers may be surprised to learn that the percentage of energy drawn from the utilities that is actually transferred to compute resources is quite low. Advanced solutions and proper planning are required to enable datacenters to maximize the conversion of electricity. The cost savings from the improvement in efficiency can be reinvested in IT resources or used to lower IT expenses.

Global Headquarters: 5 Speen Street Framingham, MA 01701 USA P.508.872.8200 F.508.935.4015 www.idc.com

SITUATION OVERVIEW

IDC has been extensively researching the issue of datacenter power and cooling, monitoring datacenter trends and surveying customers on priorities and challenges. Recently, IDC has observed that the customer demand for computation — and hence the pace of advancement in the server industry — is impacting datacenter facilities. Through working with customers and applying IDC data sets against industry standards around heat and power consumption, IDC has concluded that the time frames for datacenter evolution are grossly out of sync with the rate of change in server technology.

Key factors including the escalating performance of server components, the expanding IT footprint within the organization, and the adopting of high-density form factors have increased the importance of managing power and cooling. Additionally, IT managers experience challenges in the datacenter converting all available electrical energy into compute performance.

Increasing Computing Demand

The first factor in the mounting power consumption within the datacenter is the continual increases in processor performance. Customers have continually demanded improvements in computing to manage the ever-increasing complexity of applications and data sets. Semiconductor advancements and advanced server packing to date have met customer demands for increased compute capability, internal memory, and disk drives. According to IDC research, the consequence has been higher amounts of power consumed and heat dissipated, albeit resulting in increased throughput. In the past decade, the average server power draw has increased fourfold. This rise has challenged datacenters both to bring more power to the server and to cool the corresponding increased heat load.

High-Density Computing

To maximize the available floor space in the datacenter, customers have increasingly deployed smaller, more compact servers. IDC estimates that server density in the datacenter has increased by an annual rate of 15%. This shift toward high-density computing was first led by the transition from pedestal to rackmounted servers and has escalated further with the mainstream adoption of blade servers. According to IDC's *Server Installed Base Tracker*, customers averaged 7 servers per rack in 1996 and are expected to average 20 servers per rack by 2010. Further still, customers with large blade deployments expect to exceed 60 blades in a rack.

IDC estimates that the drive for density combined with the escalating speeds of server processors has led to an increase of power consumption at the rack level by a factor of eight since 1996. While customers have been able to increase the compute ability within the finite space of the datacenter, they have done so at the expense of power and cooling capacity. According to IDC surveys, 40% of datacenter customers have reported that power demand has outstripped the current supply. Additionally, the increased density creates hotspots in the datacenter that are subject to failures and reliability issues. This issue illustrates not only that financial implications are related to energy savings but also that the matter of IT downtime can impact business performance.

Server Installed Base

According to IDC's *Server Installed Base Tracker*, the server installed base has increased by 15% annually. IDC projects there will be over 35 million servers worldwide by 2009. This massive expansion is being driven by two factors: the expanding IT footprint within organizations and the type of systems being deployed. It is common practice today to deploy one or more servers to support a single business process. Workloads such as Web applications and email are increasing the number of servers that datacenter managers must power and cool on what seems to be a daily basis. Additionally, as business is increasingly being done on a 24 x 7 basis, IT resources are required to be available around the clock. This situation places a significant emphasis on the availability and reliability of IT assets. As a result, customers are building IT infrastructures with increased redundancy or an N+1 approach.

The second factor involves the migration to x86 servers. Faced with the challenge of improving performance levels while maintaining (or even reducing) expenses, IT managers have deployed an increasing number of x86 servers. IT has been able to migrate workloads to these lower-cost x86 systems due to the increased processing power of the servers and lower capital expenditures. With technological advances, such as 64-bit computing, x86 servers are now able to do work that was previously the domain of mainframe computers. The result to date has been IT infrastructures that are distributed environments with single applications dedicated to single systems.

Conversion of Power Load into Computation

While IT purchasers work closely with line-of-business managers in sizing and selecting server equipment, the facilities department and datacenter managers have the responsibility to power, cool, and pay the utility bill for the company's IT equipment. IDC has learned that there is often a disconnect within organizations between the IT purchasers and the datacenter managers. As an example, customers have recounted several stories of datacenter managers who are unaware of new servers they must support until the servers are delivered to the datacenter.

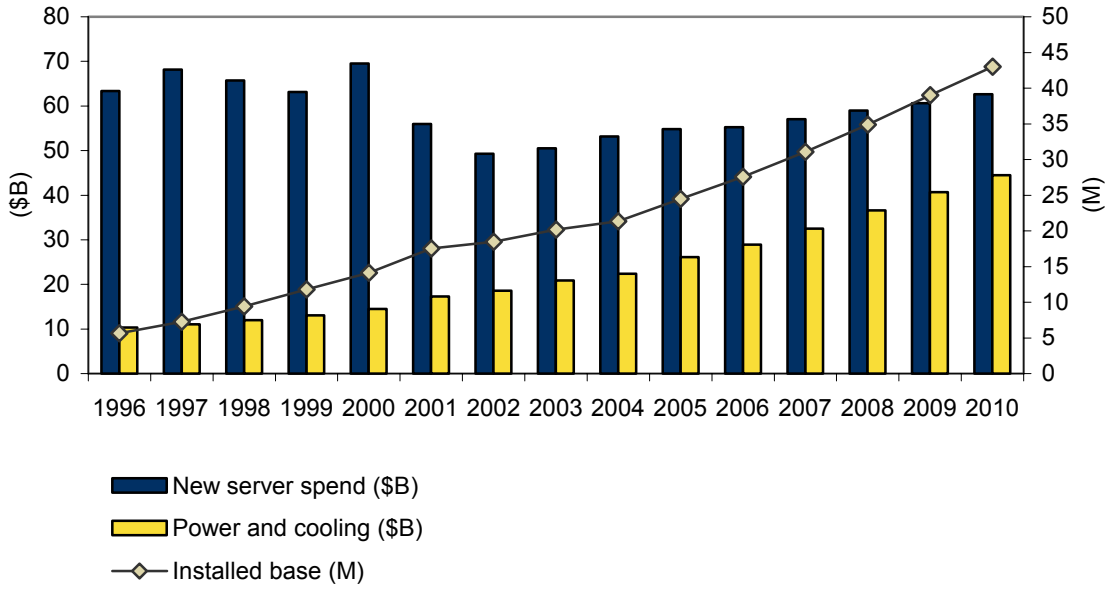
IDC has discovered that it is even more common that IT managers do not know how much power their individual systems consume. Since a solution is not available to accurately measure power consumption, datacenter managers must allocate power to server systems based on the expected peak loads. The downside of this method of allocating power is that datacenters do not consume electricity efficiently, as individual servers can quickly reserve the total available wattage in the datacenter — often reaching the power and cooling thresholds while operating at low utilization rates.

FUTURE OUTLOOK

The previously mentioned technological factors, combined with the shift in customer purchasing patterns, have dramatically increased the expense to power and cool the server infrastructure. Even though customers have been able to control the capital expenditures on IT equipment, the operational expenses have become the greater burden. In addition, ensuring that ever-increasing power is available in the datacenter may be difficult in some areas of the world. Over the next several years, the expense to power and cool the installed base of servers is projected to increase at four times the growth rate for new server spending (see Figure 1).

FIGURE 1

Worldwide Cost to Power and Cool Server Installed Base, 1996–2010



Source: IDC, 2007

In working with customers, IDC identified four strategic drivers motivating companies to control power and cooling in the datacenter:

- **Scale-out server environments.** The enterprise's reliance on IT and the computation requirement will continue to increase. As business applications become increasingly complex, the compute performance of the infrastructure will need to improve concurrently.
- **Financial impact of corporate strategy.** As noted, the electrical expense to power and cool servers has become a significant cost factor. Progressive companies comprehend that considerable financial savings can be realized by employing a corporate strategy to reduce power consumption.

- **Expansion limited by thresholds.** IDC believes the economy has come out of the recent downturn and companies are beginning to reinvest in IT resources. Customers are seeking to expand their IT infrastructure without the costly renovation of the existing datacenter or the building of new facilities.

- **Civic responsibility to the environment.** Companies are driven by environmental concerns, in addition to economic reasons. Several customers have even begun to measure their "carbon footprint" and their impact on the environment.

IBM'S "COOL BLUE" PORTFOLIO

IBM developed the Cool Blue portfolio of technologies and solutions to address the challenges of power and cooling in the System x server line. These solutions are designed to improve power efficiency and management in the datacenter without compromising throughput or compute performance. IBM offers a range of products and services through the four stages it has identified for a successful corporate power management strategy.

Presales

Even before the server hardware is procured and installed in the datacenter, IBM provides a solution to analyze and properly plan server deployment. The IBM Power Configurator is an assessment application that integrates facilities and IT management. End users input the computing requirements of their datacenter and receive a customized report of utilization rates and power and electrical requirements specific to their IT environment.

Installation and Planning

Up-front planning and careful considerations during the initial stages can ensure optimal efficiency in the datacenter and avoid costly retrofits in the future. To assist customers in maximizing the datacenter's potential, IBM offers services that include assessment, budgeting, and planning of power consumption. Factors including existing infrastructure, types of workloads, required availability, and future growth predictions are taken into consideration. These services are available for customers' existing datacenters or new installations, or both.

Management

The signature product offering of Cool Blue is IBM PowerExecutive, an automated energy management solution available on the new models in the System x server line. PowerExecutive provides customers with more visibility and control of the power dynamics of their systems. Customers are able to measure real-time power consumption and heat emission by individual server or server rack. Through the graphic display, customers can trend power usage over a time period, as well as compare actual rack power draw versus projected or allocated power. Additionally, PowerExecutive allows customers to cap power consumption to specific servers. This capability enables customers to use the datacenter's current power and cooling

capabilities more effectively to yield a better ROI on their IT investment. Customers are able to maximize performance per watt by converting all available power and cooling in the datacenter into computing. PowerExecutive enables customers to reduce the power and cooling margins placed upon the datacenter when actual consumption of power by each server either is not known or cannot be controlled. In addition, lowering power consumption in periods of low utilization will provide additional cost savings.

Cost Savings

Utilizing the detailed reports from PowerExecutive, customers are able to calculate the potential cost benefits through implementing the Cool Blue portfolio of power and cooling solutions. Additionally, the IBM Cool Blue portfolio offers innovative cooling technology such as the Rear Door Heat eXchanger, a cost-effective solution for datacenters that are at the cooling capacity threshold yet still have available rack space. A low-cost alternative to procuring additional air conditioning or cooling units, the Rear Door Heat eXchanger mounts to the back of a server rack to provide 15kW of cooling capacity. The system draws from the chilled water supply to cool the hot air exiting from the chassis before it enters the room.

ECOSYSTEM COLLABORATION

Understanding that the challenges confronting the datacenter are complex and caused not by one single factor but by a multitude of challenges, IBM and Intel offer individual as well as collaborative solutions to provide end-to-end energy management innovation from the processor to the server to the rack to the datacenter:

- In response to customer demands for a more energy-efficient server that does not compromise performance, Intel introduced the new Dual-Core Intel Xeon processor 5100 series based on Intel's Core micro architecture. Intel has stated that this technology provides three times the server performance and more than three times the energy efficiency of the fastest single-core Intel Xeon processors. Furthermore, Intel recently introduced the first quad-core processor in the x86 industry. Intel's data indicates that the Quad-Core Intel Xeon processor offers 50% greater performance than the Dual-Core Intel Xeon processor in the same power envelope.

- Intel's Demand-Based Switching (DBS) operates conjunctively with IBM's PowerExecutive to manage power consumption based on utilization. Intel DBS allows customers to lower utility costs while maintaining optimal performance levels. With OS support, DBS utilizes the Enhanced Intel SpeedStep technology to minimize wasted energy by dynamically ratcheting down processor power states whenever peak performance is not required. This capability can substantially reduce average power consumption for servers operating at the typical low utilization rates in the datacenter. Rightsizing power loads to specific workloads can be one of the most effective solutions toward better power management.

- To serve the high-density business segment, IBM and Intel collaborated on IBM's blade server offering. The IBM BladeCenter Ultra Low Power HS20 and HS21 blades combine IBM's power and cooling technology and Intel's low-voltage Dual-Core Intel Xeon processors. The maximum processor consumption rate of only 40W offers customers a high-density yet low-power configured system.

IDC ANALYSIS

The emergence of power and cooling concerns within datacenters offers a new opportunity in the market. Vendors that are able to successfully develop solutions that deliver on customer demands will be able to gain further traction within the customer base. As the topic is relatively new, it is critical that vendors first develop a clear and concise message that educates customers on the current challenges and articulates the benefits of their product offerings.

It is also important to develop a message that resonates with each of the customer stakeholders involved. A successful message will speak differently to customers at different points in the purchase decision-making process, be they C-level executives, datacenter architects, facility managers, or IT directors and managers. Server systems are part of a larger datacenter ecosystem; therefore, server vendors must also consider their relationships and potential partnerships with other datacenter suppliers in order to offer the broadest range of benefits to customers.

Server vendors must work closely with their customers to fully understand their needs and also demonstrate the full potential of their solutions. For instance, IBM should make customers understand that PowerExecutive not only is a power measurement solution but also enables customers to add IT equipment to the datacenter without increasing their existing power or thermal capabilities. Currently in the datacenter, power is allocated to systems based on expected peak load, a type of reservation process that can quickly use up the overall capacity. Moreover, this expected load is derived from the specification labels and not from real performance data. PowerExecutive enables power supplies to be pooled and provisioned based on actual demand, ensuring all available power is converted into computing.

The involvement of traditional power and cooling solutions companies addressing these challenges in the datacenter introduces new competitors to server vendors. Such vendors as Emerson/Liebert and APC have a history of working with datacenter and facilities managers. These competitors have recently stepped up their own product offerings and have even begun to partner with server vendors to offer integrated solutions.

IDC believes that power and cooling is and will be a key factor in the server market, but it is still emerging and not fully defined yet. Any vendor that can establish itself first in the market and be at the forefront of defining standards for power and cooling in the datacenter can gain a significant foothold for years to come.

CONCLUSION

The issue of power and cooling poses a real challenge to today's datacenter, in both financial and operational terms, particularly as customers are beginning to integrate IT departments and corporate facilities and continuing to drive initiatives to consolidate datacenter resources. The resulting high-density environment has an increased requirement for predictive analytics and comprehensive metrics on power and cooling.

To achieve optimal utilization of compute resources, customers must be able to measure, analyze, and, most importantly, dynamically provision power supply and cooling capacity to match utilization rates. Customers need to be able to break away from the practice of overprovisioning power and cooling supplies. Additionally, vendors need to supply solutions that enable datacenters to convert a higher percentage of energy to compute capacity. The capability to rightsize power to workload is a key element in an overall power and cooling strategy that encompasses solutions from processor to server to datacenter.