Why Are Datacenter Power and Cooling Still Challenging?

With current understanding of datacenter efficiency and low power components, why are power and cooling still under pressure?

Executive Summary

CPU transistors have gotten smaller, and server component power has dropped consistently with each new generation. But datacenters are still up against the proverbial wall with pressures on both power and cooling, as they attempt to maximize their use of datacenter floor space and compute resources. Compute density has increased dramatically over the last decade, pressuring datacenter infrastructure. Virtualization and system density drove rack-level power consumption up at the same time that platforms were trying to drive it down. To make a serious change in power and cooling profiles, datacenters (and vendors) need to re-think the servers that they are deploying and alter their datacenter strategies.

We don't believe that most datacenters today are ready for a complete transition, but instead these new form factors can help augment existing strategies and drive a better overall operational mix.

Resistance is Futile

At the start of this millennium, datacenters were being pressured by the need for better power and cooling efficiency. Since that time, we have seen tremendous improvements in the power efficiency of CPUs, memory, power supplies, fans, and even hard drives. But despite these advancements, businesses still are not able to keep pace. Why do datacenters still see power and cooling as such a problem more than ten years later? The answer, in short form, is that in the battle of datacenter architectures, power and cooling efficiency has taken a backseat to the business needs for greater performance and density. Heat and power continue to be an issue, but competitive pressures have taken the foreground and driven most of the strategy decisions.

OEMs and component vendors have spent much time trying to reduce power consumption, while at the same time boosting performance. Despite the fact that the results in performance-per-watt are impressive over each generation of server, the net raw power consumption remains troubling. Most of the advancements in performance-per-watt have been driven more by higher performance than lower power consumption.

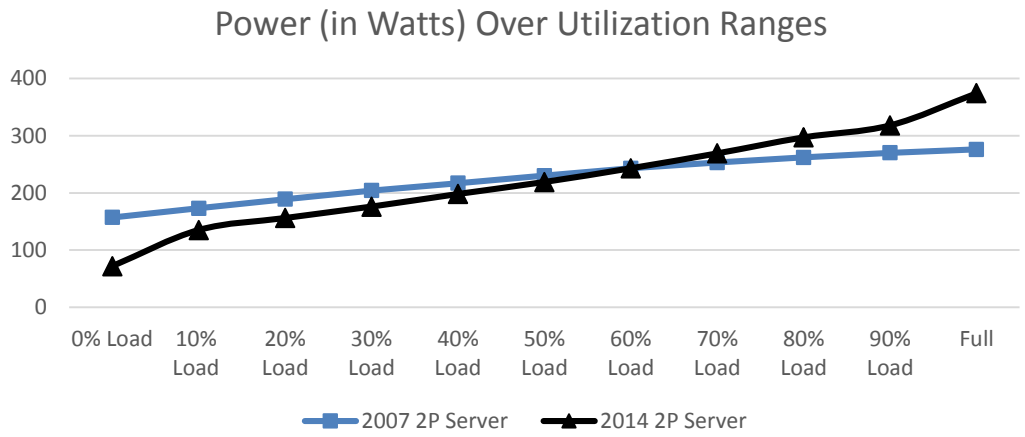
Heat is an unfortunate side effect of any datacenter. This is because IT equipment consumes massive amounts of electricity, and as the electricity flows through the equipment, resistance prevents some of the electrons from reaching their destination. Any electrical energy not consumed by the server is converted to heat energy. Even with the high efficiency power supplies, low power CPUs, and low voltage memory that

have sprung up in servers over the past decade, the typical 2P server still emits 2,000-4,000 BTUs per hour, depending on configuration and workload. This is not much of an improvement from where systems have been running over the past decade.

The advent of virtualization has only complicated the matter. Server utilization is an average of 10-20% on the typical standalone enterprise workload, but it is pushed to 60-80% on a typical mix of virtualized workloads. As utilization increases, the circuits in the CPU fire more often, memory cells charge more often, and drive seeks increase. So the stress on the system is pushing up net power consumption.

Examining power consumption and performance-per-watt over the last 7 years shows, on a relatively "apples to apples" comparison, that while server performance has gone up almost tenfold, the average server is still consuming about the same amount of power. Where platforms have become more efficient has been at the low end of the utilization curve. Unfortunately due to virtualization, cloud, and other technologies designed to drive better efficiencies, there are fewer servers operating at the lower end of the utilization curve. This situation is akin to a hybrid car where fuel efficiency at low speeds is greatly diminished if most of the driving is done on the highway. In fact, for higher utilization loads (above 60%), today's servers are often consuming **more** power along with the higher performance.

Figure 1: Power Consumption at Different Utilization Levels[1]



With customers demanding better power efficiency, how did we end up here? Today, much of the industry is still being driven on benchmarks, so both CPU vendors and OEMs battle it out to show how power efficient they are. Unfortunately this has resulted in tuning to the benchmark. If you observe the power results in Figure 1 above, you'll see that idle and low power loads (<20%) show significant power reductions, giving vendors bragging rights on power efficiency. Unfortunately, that's only a great data point if you are running circa-2004 workloads. To better comprehend where we are from a

[1] Source SPECpower_ssj2008 Benchmark.
2007 Server http://www.spec.org/power_ssj2008/results/res2007q4/power_ssj2008-20071128-00013.html
2014 Server http://www.spec.org/power_ssj2008/results/res2014q2/power_ssj2008-20140401-00654.html

power efficiency standpoint, the industry needs a different, more accurate accounting of power consumption, as we have discussed before. The SPECpower_ssj2008 Benchmark used by most vendors weighs performance and power across a range which includes the idle state. If customers were to focus only on the 60-80% utilization range (which is the typical target for virtualized environments), they'd see that power consumption varied from equal to 13% **higher** for the new "modern" servers with the smaller CPU transistors and lower power memory. Clearly, replacing aging servers with new servers based on power-efficient components would bring greater performance and better performance-per-watt, but it would not necessarily bring down overall power in the datacenter.

Most customers provision power based on watts/ft^2 of floor space. The average datacenter has a power density of well under 10kW provisioned for the typical rack (which occupies approximately 6ft^2 of floor space). Highly dense scale-out datacenters employing systems like the OpenCompute Project designed racks (specified by Facebook) may reach power density of 15kW per rack or higher.

Operating today's workloads clearly consumes more power, but the real devil is in the cooling details. The Green Grid, an industry consortium focused on power efficiency in the datacenter, has a measurement called Power Usage Effectiveness (PUE) that measures the ratio of power used to run IT equipment vs. total facility power. PUE is typically ~1.5 to ~1.8 for most businesses (with the very efficiency-minded hyperscale datacenters bringing that number down into the 1.1 range). This means that for most businesses, for every dollar spent to power actual compute cycles, an additional $0.50 to $0.80 is spent just cooling the systems down from that work.

Further exacerbating the power/cooling situation—and adding an additional level of complexity—is the compression in system density: denser systems have less room to move air through the chassis, complicating the cooling process.

It's My Density

Engineers are fond of saying "features, schedule, or price: pick any two". In the datacenter, it's density, performance, and low OPEX: pick any two. The laws of physics and the laws of economics conspire to make it difficult for someone to optimize on all three vectors.

Just a few years ago, the typical server was 2 rack units (2U) high, allowing customers to get an average of 12-15 servers in a rack while still leaving room for networking, storage, and other devices. Today's rack servers are denser, and conventional deployment options have been expanded to include blades, twins, and other modular dense form factors. These deployments are pushing density of servers up to a point where typical deployments may have 40-60 servers (or more) in a 42U rack.

Although there are now several "conventional" choices for server deployment, these different form factors have had less impact on the amount of power-per-server; their largest impact is on compute density, which simply invites more power density. One of

the remaining challenges is the physical footprint of power and cooling within the server itself; dense deployments simply repeat this over and over throughout the rack with hundreds of power supplies and cooling fans.

Table 1: Density & Power Consumption by Form Factor

Form Factor	Typical Servers Per Rack	Typical Theoretical Max Power* Per Rack	Per Server
2U 2P servers	21	15,750W	1500W
1U 2P servers	42	19,320W	920W
2P Blades	64	31,800W	993W
Modular (4 x 2P nodes in 2U)	84	31,500W	375W

**Assumes redundant power, and see Figure 1*

As companies grapple with power consumption, cooling, and datacenter density issues, there is only so much that can be done with the traditional server architectures and traditional datacenter cooling techniques. A different approach is required.

Chilling Out

Joule's first law dictates as electric current passes through a conductor it releases heat. But dense systems magnify this problem, as tight form factors increase the complexity of efficiently removing the heat. So a different approach is required. Luckily on the cooling side, there are more straightforward technologies that we know can address the cooling challenge. The issue has always been the cost and time to deploy, so to address cooling we mainly need to tackle the laws of economics not the laws of physics.

Cooling is so critical to datacenters that the ability to cool systems will drive other purchase decisions. In a recent conversation with a Fortune 500 technology company, they estimated that the loss of cooling would result in "losing a floor" in under 3 minutes, as the heat spike would require systems to go into emergency shutdown mode. Just a few years ago this wasn't even on the radar, but density and higher utilization has pushed this to the foreground as an issue. Knowing that the company has less than 3 minutes of runtime in an outage means provisioning for 15 minutes of emergency power was unnecessarily tying up IT capital: systems would be unable to ride out 15 minutes without cooling and would be shut down long before the 15 minutes of UPS power ran out.

Most datacenters are cooled by forced air. While this is hardly the most efficient mechanism, it is clearly the easiest to implement, most practical, and potentially the most cost effective from an infrastructure standpoint. But there is a significant operational cost that comes along with the extra power required for chillers, blowers, and the other components to extract the heat. Forced air is a tradeoff in economic savings up front but greater long term costs down the road.

Depending on location, some companies can take advantage of "free air" cooling that involves both choosing the right climate (something not every company has the luxury of doing) and installing a more specialized system. Facebook's datacenter in Prineville,

OR is a great example, and other companies have chosen locations in cooler climes like Sweden or Scotland. With free air cooling, chillers are not required, and power consumption over time is much lower than with forced air. There is a higher capital cost up front, but it is worth the tradeoff; the lower long term operational costs over time will help drive a total cost of ownership (TCO) savings.

Water cooling is the most efficient mechanism for removing heat from a datacenter—up to 1000 times more effective than air cooling[2]—but it is more difficult to deploy operationally, and that complexity also drives a higher cost. Much of the cost is borne in the risks of mixing both water and electricity and the extra safety precautions that need to be employed; even a simple leak could turn deadly to multiple systems. To reduce the risk to the datacenter, more expensive infrastructure and backup equipment is required. Because of the cost and complexity, water cooling was typically reserved for higher-end proprietary systems, as their heat density, high cost, and small numbers within the typical datacenter would lend themselves to liquid cooling. For the thousands of less expensive x86 nodes packed throughout the datacenter, cost and logistics prevented liquid cooling from being a practical solution.

However, as density for certain workloads continues to escalate, a cost-effective and easier-to-deploy liquid cooling solution could help drive better efficiency for datacenters.

HP's Illiquid and Liquid Assets

To address both the power and cooling challenges that density brings to the datacenter, HP has introduced a pair of products, the HP Apollo 6000 and 8000 Systems that bring density, power efficiency and cooling efficiency into focus.

The HP Apollo 6000: Power Efficiency for Scaling Out

The HP Apollo 6000 system is based on a dense, single socket server design for hyperscale datacenters and scale-out workloads. The single socket design of these systems lends itself well to the environments where the number of individual instances-per-square-foot in the datacenter is critical and the workload itself is lightly threaded. By reducing the footprint in the data center, customers can reduce both their capital budgets as well as their operational budgets by freeing up floor space and requiring fewer racks to house their servers.

The HP Apollo 6000 System is a modular chassis that employs 10 server trays with two 1P nodes each, all contained in a 5U chassis. This delivers a total rack density of up to 160 servers in only six square feet of floor space (using a 47U rack) or 140 nodes (in a 42U rack). The Apollo 6000 System is ideal for workloads like Electronic Design

[2] http://www8.hp.com/h20195/v2/GetDocument.aspx?docname=4AA5-0069ENW

Automation (EDA), financial risk modeling, or life sciences applications that tend to be more single-threaded, requiring higher speed with fewer threads per server.

Based on Intel's Xeon E3 processor, this server bridges the gap between the high performance ProLiant servers mainly built around Xeon E5 processors (larger cores, dual socket) and the HP Moonshot servers that deliver higher density but use Atom and ARM processors (smaller cores, single socket). This product is clearly targeted at those applications that need to balance the density of scale-out with the individual thread performance that Xeon delivers.

> "We are seeing up to 35% performance increase in our Electronic Design Automation application workloads. We have deployed more than 5,000 of these servers, achieving better rack density and power efficiency, while delivering higher application performance to Intel silicon design engineers."
>
> *Kim Stevenson, CIO, Intel*

The 5U chassis is designed with individual trays that slide in like blades, with each tray housing two individual servers. Because of the single CPU architecture and smaller system footprint, two servers can be installed in the space that is normally reserved for a single 2P server on a blade—doubling the overall system density. The chassis also uses an "I/O innovation zone" which is a modular unit that plugs into the back of each of the server trays, allowing for a multitude of I/O options to be added, matching the I/O capability with the needs each individual server/workload.

Figure 2: HP Apollo 6000 Compute Tray

For density, the power supplies are not located inside the chassis. Instead, they are in a separate unit that is 1.5U high and can support up to 4 chassis. This brings the effective power footprint down to as low as .375U per chassis (with each chassis containing up to 20 servers.

The Apollo 6000 System uses the same power supplies (and configurations) as the HP BladeSystem C7000 with up to 6 power supplies that can be installed. By using the same supplies, customers who also use BladeSystem blades for their traditional workloads can reduce costs and increase flexibility by utilizing the same service spares for both systems. But the clear difference is that the power shelf for the Apollo 6000 can drive up to four systems (80 total servers) versus the 16 servers and I/O networking that are driven by the C7000 chassis. With a maximum of 169W per tray (with two servers), the net power-per-server in an Apollo 6000 configuration should be about 85W, which is a dramatic difference from the other form factors cited above.

Figure 3: HP Apollo 6000 Chassis with Compute Trays

As the vast majority of workloads in the datacenter are running on dual CPU platforms, the Apollo 6000 is not necessarily a replacement for all of the servers in a datacenter. But it is an important additive element to help drive a better blended efficiency across workloads in the datacenter. By introducing Apollo 6000 platforms into the overall compute mix, customers can better blend both their capital and operation budgets with an additive boost that helps reduce the overall cost of acquisition and cost of operation for the data center as a whole.

For hosters and service providers, single socket servers are the platform of choice, as they provide a better mechanism for managing customers' virtually hosted servers. The ability to add incremental capacity in more granular units is also important, providing a smoother scaling upwards without the sawtooth capital expenditures that occur with larger dual socket platforms. Additionally, SLAs are easier to enforce with the fewer virtual variables that a single socket server maintains. Being able to characterize workloads effectively allows a business to gain better efficiencies, more accurately matching the workload to the platform.

The HP Apollo 8000: Remaining Cool Under Load

For heavier technical workloads that require greater scalability at the node level, a dual socket system with up to 10 cores (20 threads) per CPU, significant memory scalability,

and accelerators are required. But, unfortunately, trying to get that amount of performance and scalability into a very dense form factor usually presents a large cooling challenge.

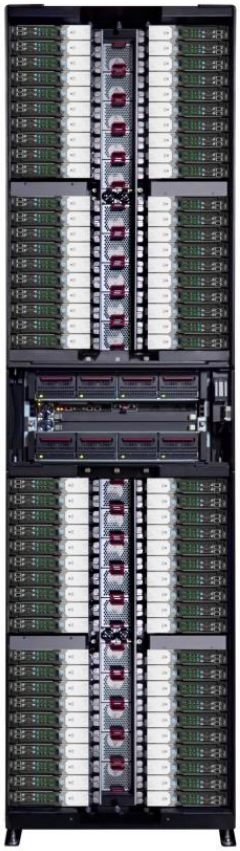
For High Performance Computing (HPC) or HPC cloud workloads, the new HP Apollo 8000 System delivers a scalable 2P platform with the threading and memory scalability required to tackle complex, highly parallel, high performance applications. The dense design allows for up to 144 dual socket servers to be deployed in a rack (with 288 total cores), squeezing exascale-level performance into a single rack. The HP Apollo 8000 System is designed for these technical environments; it brings the first commercially viable liquid cooling solution to large-scale HPC server deployments. By utilizing liquid cooling, the Apollo 8000 can have a significant reduction in the cost of operating an HPC system.

Traditionally, liquid cooling was very complicated to deploy for x86 HPC clusters. Performance-constrained HPC workloads were not good targets for liquid cooling because of the high node count and the fact that node failure resulted in complicated servicing procedures. But HP has developed an innovative mechanism for using the more efficient liquid cooling within these modular x86 servers. It also allows servers to be removed and replaced without exposing the liquid cooling elements or endangering the system/datacenter.

Previous attempts at liquid cooling for x86 HPC systems were relegated to special rack enclosures that consumed more datacenter floor space, essentially reducing density. However, this new HP system, while requiring a very specific rack, still meets the same width footprint as most traditional racks.

To deliver the power density that this rack will need, the HP Apollo 8000 system starts with a 480V AC input that is then stepped down to the right voltages for the servers through a series of up to 8 power rectifiers. The 480V power is actually less expensive of an infrastructure than 220V or 110V power; normally trying to bring this level of power density into the rack with a smaller input voltage would result in much larger and harder to manage cables.

Figure 4: Fully Loaded HP Apollo 8000 Rack



The water cooling system begins at the rack level with a “water wall”, which runs up the center of the rack, circulating cool water to draw heat off of the nodes on either side. This is a completely sealed environment, and there is no liquid “connect/disconnect” required in servicing a server tray. Each of the server trays (holding two 2P servers) slides into the rack on either side of the water wall. The trays feature a “dry disconnect” mechanism that disengages the heat transfer unit (aluminum thermal block on the side of a tray matched to an aluminum pin fin on a rack’s thermal bus bar). The server trays have a Dry IT loop and have virtually no liquid inside; heat is extracted from the CPU and memory inside the server tray via sealed copper heat pipes. This unique design allows servers to be removed in seconds and easily maintained without having to deal with the liquid (either in disconnecting or reconnecting).

The Apollo 8000 System can actually cool the entire rack with water as warm as 30C (86F), and there is no ionized or distilled water needed; the water only needs to adhere to ASHRAE spec.

Once removed from the server, the heat then can be cycled out of the rack and into the building. One customer, the National Renewable Energy Lab (NREL), actually used the output of their system to help heat the facility’s water and provide energy to other parts

of the building. Their HP-powered liquid cooled system is projected to save them up to $1M USD per year.[3] NREL's "Peregrine" system includes 720 nodes based on the new Apollo 8000 System as part of the overall supercomputing system.

To help drive better operational costs, the entire rack is managed by a single HP Apollo 8000 System management unit that connects to all of the servers through an out-of-band connection. A single cable is all that is required to manage the 144 servers in the rack. This helps reduce the typical cabling load for dense systems and gives administrators access to all of the management functions that are critical to ensure smooth operations.

The HP Apollo 8000 System attaches directly to standard pre-fabricated cooling circuits, making it easier to install, yet still delivering redundancy and serviceability to the servers in the rack. The quick connect modular circuits are designed to help customers easily and quickly deploy liquid cooled systems in only a few days, compared to weeks or months that traditional systems require.

> "When Steve Hammond [Director of NREL's Computational Science Center] told me he wanted to put in a petaflop-scale high performance computer, and that he wanted to cool it with warm water and use that water that comes off to heat the building, and then use the excess heat to warm the cement outside, I said, 'Riiight'.
> "We were doing something very unique and different—breaking boundaries. But the bottom line is, we nailed it. And we're making it commercially available, so others can take advantage."
>
> *Paul Santeler, Vice President of the Hyperscale Business Group at Hewlett-Packard*

Figure 5: HP Apollo 8000 Compute Tray

Through the combination of density, water cooling (which is significantly more efficient), and high voltage DC power, the Apollo 8000 system can have a large impact on both the cooling costs through higher efficiency as well as the time required to provision the

[3] http://www8.hp.com/h20195/v2/GetDocument.aspx?docname=4AA5-0069ENW

servers. Ultimately this leads to lower operating expenses which have a strong reduction in the total cost of ownership for the system as a whole.

Through its innovative and easily-deployable cooling system and its high density, high performance compute, the HP Apollo 8000 System helps businesses with complex technical workloads overcome some of the challenges with driving better density and performance in their datacenters without requiring a highly complex cooling system.

Summary

Clearly, as server utilization and density have increased, many of the benefits from lower-power components have been more than offset by the increase in power consumption that was needed to power these units and the corresponding increase in heat as a result of that power increase.

With workload complexity continuing to scale-up and deployments continuing to scale-out, the problems of power and heat will only increase. HP is addressing these challenges with a pair of new platforms that are designed to bring greater density/power efficiency for scale-out workloads and greater density/cooling efficiency for high performance compute workloads.

As customers try to balance their power needs across different workloads, they need to give serious consideration to the HP Apollo 6000 and 8000 series products which bring a new perspective on tackling the power and cooling issues that plagues dense datacenters. For more information: http://www.hp.com/go/Apollo

Important Information About This Paper

Author
John Fruehe, Senior Analyst at Moor Insights & Strategy

Review
Patrick Moorhead, President & Principal Analyst at Moor Insights & Strategy
Paul Teich, CTO & Senior Analyst at Moor Insights & Strategy

Editor
Scott McCutcheon, Director of Research at Moor Insights & Strategy

Inquiries
Please contact us here if you would like to discuss this report and Moor Insights & Strategy will promptly respond.

Citations
This note or paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title and "Moor Insights & Strategy". Non-press and non-analysts must receive prior written permission by Moor Insights & Strategy for any citations.

Disclosures
HP is a research client of Moor Insights & Strategy and this paper was commissioned by HP. No employees at the firm hold any equity positions with any companies cited in this documented.

DISCLAIMER
The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors. Moor Insights & Strategy disclaims all warranties as to the accuracy, completeness or adequacy of such information and shall have no liability for errors, omissions or inadequacies in such information. This document consists of the opinions of Moor Insights & Strategy and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Moor Insights & Strategy provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and

forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.