Uptime Institute® | INTELLIGENCE

UI INTELLIGENCE REPORT 74 • Q3 2022

Silicon heatwave: the looming change in data center climates

A rapid rise in the concentration of processor thermal power will have far-reaching consequences, not only for servers but for facility design and operations

Author Daniel Bizo, Research Director, Uptime Institute Intelligence

KEY POINTS

After a long period of relative stability, typical server power consumption has risen steeply since the middle of the 2010s — and is expected to keep rising.

The force behind this trend is an upsurge in silicon power, chiefly server processors, which is likely to continue.

High thermal power and lower temperature limits of next-generation server processors will challenge the practicality of air cooling and frustrate efficiency and sustainability drives.

Data center operators face some complex decisions around how to prepare their facilities (if at all) for the next generation of higher power servers.

The evolution of the data center has, by and large, been shaped by developments in IT. From their beginnings as a few small rooms with a handful of big and expensive computers, data centers have swelled in numbers and size to house large numbers of relatively small and inexpensive machines.

In many ways, the triumph of inexpensive volume servers over mainframes and UNIX-based systems was a boon for the data center. Scale and standardization in IT, including their form factors and environmental requirements, have allowed the data center industry to mature both in design and engineering.

Cooling is a good example: the dominance of affordable volume servers not only made it possible for climatic standard-setting body ASHRAE to create common thermal management guidelines for data centers, but encouraged a wider band of operating temperatures too.

A phase of relative stability in IT hardware requirements for power and cooling from the early 2000s to the middle of the 2010s allowed data center engineers to focus on honing their craft. Data center design power densities have remained virtually static for nearly two decades, reducing the need for the replacement of aging power and cooling equipment for performance reasons, and this has helped extend the useful life of many legacy facilities.

For example, replacing or adding more servers at the start of the 2010s was a nonevent for data center operations and capacity planning. With few exceptions, the new servers would all draw no more than 250 to 300 watts (W) of power at peak and typically stayed under 200 W (much like their older counterparts from the early 2000s). Another refresh cycle around the middle of the decade saw some uptick, but not a dramatic increase.

However, this well-established pattern has changed. A technology refresh at the start of the 2020s brought a marked increase in power for the new generation of highly performing and efficient servers that are the foundation of many IT infrastructures. Looking ahead, the rise in server power and cooling needs throws any past assumptions on future power densities and cooling technologies into the air. The most recent Uptime Institute Global Data Center Survey shows most businesses refresh their servers every three to five years, which means that by 2027 most servers currently in operation will have been replaced with more power-hungry generations.

In this report we explore the factors that drive this change and what the implications and options are for data center designers and operators.

The silicon road to the great power wall

The history of semiconductors is a history of ever smaller transistors making chips faster, more energy efficient and less expensive. It is also a history of increasing power densities: with every new generation of semiconductor technology, active power for a given area of silicon jumps. As a rule of thumb, for every doubling of the number of transistors on a chip, power per transistor typically only drops by 20% to 30% with existing complementary metal oxide semiconductor (CMOS) technology.

This two-speed innovation trend creates a steady rise in power density. There are irregular and temporary breaks in this trend thanks to innovations, such as the introduction of novel transistor materials and structures, and advances in circuit design. Even so, the underlying force of semiconductor physics that drives silicon power density up is inescapable over the long term.

The silicon power density problem will continue to worsen until the chip industry is able to adopt new semiconductor technologies that use much less switching energy than current CMOS transistors. This is unlikely to happen before 2030; it will require the addition of fundamentally different materials (likely to be gallium-based, rather than silicon) and major changes to the production process.

Another factor is associated with, and amplifies, the effect of silicon density: infrastructure economics. Servers equipped with high-performance, many-core processors and large amounts of memory tend to be more energy efficient if highly utilized. By being able to carry a much larger software payload, better performing servers also tend to be more cost-effective. This effect is clearly visible in the choices of hyperscale cloud operators: these operators prefer higher-end processors (but not top of the range models) that perform close to the fastest models yet cost considerably less. This demand for more processor performance, however, means chipmakers keep pushing the scale of integration through design and manufacturing, while driving up server processor power.

These trends are not new, but a confluence of technical and market forces is now changing the development trajectory for server chips. The advances in chip design (such as vastly improved core architectures, multi-core designs and circuit power management) and manufacturing innovations that kept power under control in the 2000s and muted power escalation in the 2010s, have run their course. Market demand has also shifted toward more powerful processors as workloads grow in both size and complexity. Chip vendors are locked in a fierce bout of competition for performance and efficiency supremacy — but efficiency does not mean less total power consumption.

As a result, the 2020s will see a continued increase in the size of processor power envelopes. How far chip vendors will go before (if ever) they hit a power wall, which is either too technically impractical or commercially untenable to push, remains to be seen.

Change is in the air

Silicon power consumption is fast becoming a problem for all infrastructure operators keen to keep pace with technical developments and reap the full benefits of next-generation servers

In data centers, power density is typically understood to be the maximum sustained power of an IT rack (kilowatts, or kW, per rack — also referred to as a cabinet). Generally, high-density racks differ from regular racks in the large amount of equipment they contain, such as servers, storage systems and networking switches (and how compactly they are designed), rather than in the type of components they use.

This does not mean that there are no high-density components. The growing adoption of high-performance accelerators used for a range of scientific and technical workloads, with the notable addition of training deep neural networks, does mean the power density of individual servers and components is drawing more attention. Top-end accelerators derived from graphics processing units (GPUs) can use as much as 500 W to 600 W each at their peak power.

These systems, however, remain rare — especially in commercial and enterprise data centers where they are mostly used for specific workloads within the domain of high-performance computing. For most operators, the cooling requirements of these high-performance workloads are handled on a project-by-project basis using, for example, either close-coupled air coolers (such as rear-door heat exchangers) or direct liquid cooling (DLC) of server components.

The power required by these more unusual IT systems has diverted attention from a more fundamental trend that is beginning to affect the wider data center industry. This is the trend of rising mainstream server and component power to levels, which would have been considered extreme not long ago. Silicon power consumption, and the requisite cooling, is fast becoming a problem for all infrastructure operators who are keen to keep pace with technical developments and reap the full benefits of next-generation servers.

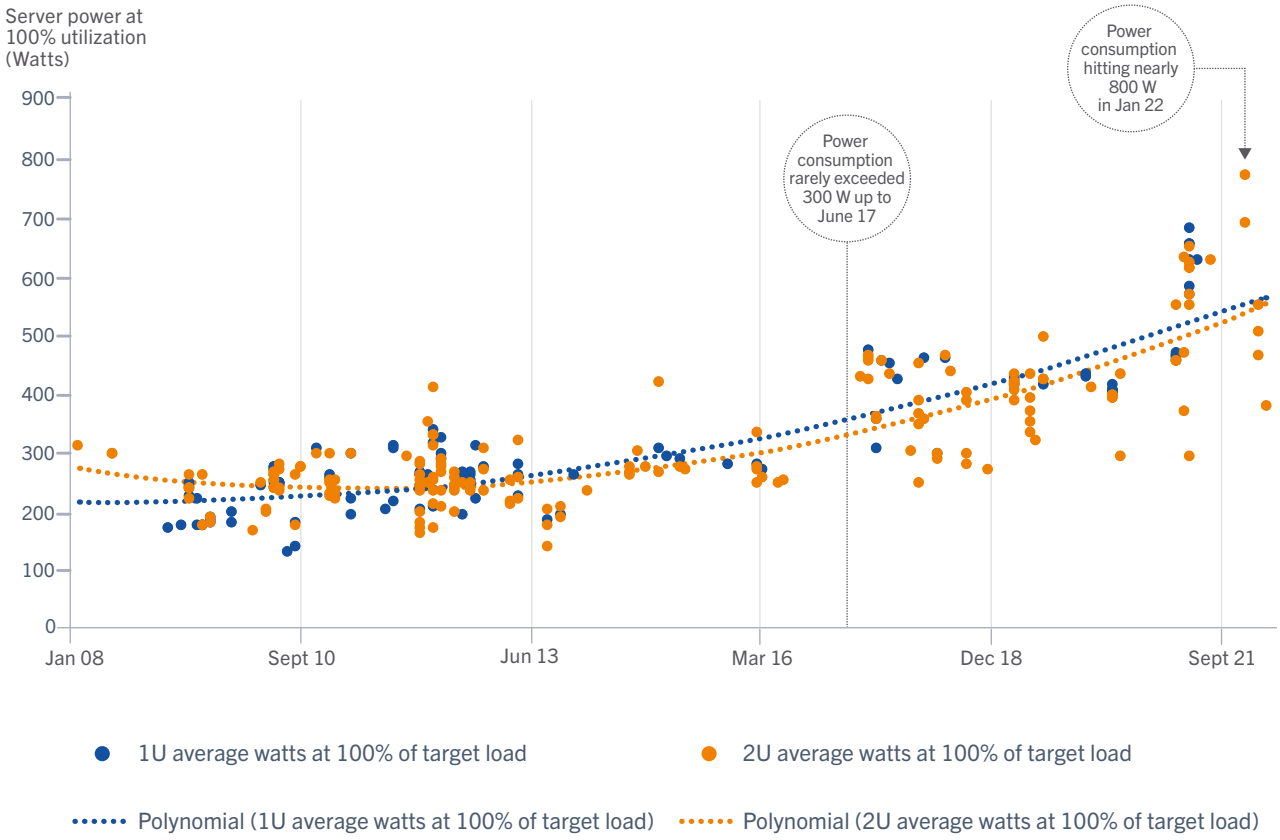
An added problem for data center designers and operators is a lack of publicly available guidance on likely power consumption requirements from chipmakers and IT systems vendors, who tend to share their information with only a few major buyers.

The trajectory of power use for the economically most attractive processors is perhaps best indicated by a cloud operators' catalogue of machine instances. Take Amazon Web Services (AWS), the world's largest provider of cloud infrastructure services, as an example. At the start of the 2010s, AWS used processor models that had a thermal power rating (thermal design power in processor parlance) in the range of 90 W to 120 W. By 2015, the thermal rating had crept up to 150 W for some high-frequency (highly clocking) models, but the processors for more mainstream instances remained less power hungry.

However, in 2017, processor thermal ratings shifted to between 180 W and 200 W in the contemporary generation of mainstream instances and up to 240 W for high-frequency instances. By 2020, thermal power for mainstream AWS processors had shifted to around 270 W to 280 W. In 10 years, thermal power for processors in the infrastructure economic sweet spot had doubled.

The industry standard power benchmark from the Standard Performance Evaluation Corporation (SPEC) corroborates this trend. Entries for dual-processor (two-socket) 1U volume servers show that sustained power consumption under full load (running a Java-based business logic) rarely exceeded 300 W up to 2017 and 2U systems topped out at around 400 W, even when configured for high-processing capacity (see **Figure 1**).

Figure 1 **Server power consumption is on a steep climb**



Data is showing sustained maximum power consumption of 2-socket servers when running the SPECpower_ssj2008, which simulates a Java-based business logic. Results for 1U and 2U form factors. Data as of June 27, 2022.

SPEC.ORG, COMPILED BY UPTIME INTELLIGENCE

Uptime Institute | INTELLIGENCE

Since 2017, burning more than 400 W of power has become typical for both Intel and AMD-based servers, even though these servers have been optimized for running the benchmark's workload as efficiently as possible. By 2021, entries with the best efficiency results showed power consumption above 600 W, before hitting nearly 800 W in 2022.

Uptime Intelligence expects further jumps in processor power consumption. Based on publicly disclosed plans and in-depth interviews, some next-generation mainstream server processors will move into the 350 W to 400 W range by 2023, which indicates that some high-volume server configurations will approach 1 kW of power at full load.

It won't stop there: by the middle of the decade, product roadmaps will call for 600 W processors. Beyond this the future is uncertain but another doubling of server power from current levels — impacting not just cooling requirements but the power density operators see in their facilities — seems almost inevitable.

From overcooling to overheating

Thermal power levels are fast approaching the practical limits of air cooling in servers

For many years, discussions of cooling requirements in the data center industry have centered on the gradual rise in rack densities and the need for DLC to handle high-power racks, which have become more prevalent. Rack power alone is enough to indicate what cooling performance is required from the data center but, due to the evolution of processor silicon, thermal issues will increasingly relate to server design and configuration and will be dictated by the cooling needs of individual components.

Thermal power levels are fast approaching the practical limits of air cooling in servers. To remove the heat from the components, the heat sinks will need to become bigger and use more expensive alloys with lower thermal resistance. More airflow means more fan power in both servers and the data center facility — pushing up energy consumption at a time when there is a concerted effort across the industry to reduce it.

In some instances, the practical limits are already visible: for an increasing number of server configurations, inlet air temperatures are severely restricted. Most servers available today conform to ASHRAE's Class A2 equipment thermal guidelines that allow for operating temperatures up to 35°C (95°F) as standard — and even higher under extended temperature support for a narrower set of configurations. However, depending on the components used in certain server configurations, operating temperatures may be restricted to lower levels — sometimes as low as 25°C (77°F) or below. Moreover, there is a growing number of instances when configurations with a certain combination of components are not supported at all due to thermal limitations. This may occur when there is a risk of high-performance processors or GPU accelerators overheating either each other or peripheral components, such as storage drives or active optical cabling systems.

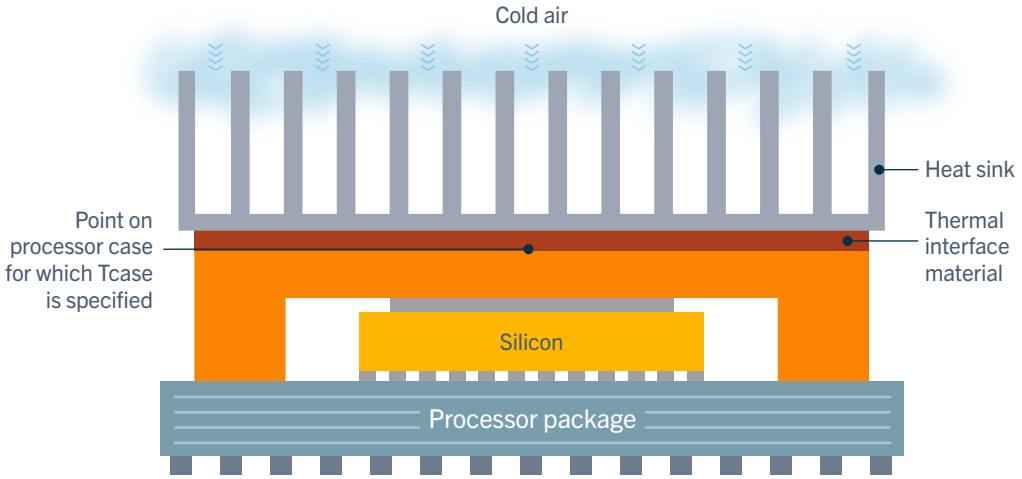
Mindful of this trend, ASHRAE updated its thermal guidance for data centers in 2021 (2021 Equipment Thermal Guidelines for Data Processing Environments) to include a new, temperature-restricted thermal standard. The new class, called H1 for high density, is applied to high-performance servers without enough room for the larger heat sinks and fans that would guarantee sufficient cooling for certain components when operating at the higher end of the regular inlet air temperature range.

For a Class H1 operation, ASHRAE recommends a lowered air supply temperature band between 18°C (64.4°F) to 22°C (71.6°F) — as opposed to the general recommendation of up to 27°C (80.6°F) — to meet cooling requirements. The allowable envelope's upper limit for excursions in a Class H1 environment is 25°C (77°F). ASHRAE also recommends that data center operators create a dedicated zone for this class of high-density air-cooled system in the interest of better managing overall cooling energy requirements. This is a reversal in the trend toward gradually relaxing temperature requirements in data centers, challenging the efficiency drive of the industry.

Given the overall trend in thermal power, Uptime Institute expects these temperature restrictions to become common and an operating requirement of ordinary server hardware, not just density-optimized systems targeting high-performance computing. As the typical processor silicon dissipates more power when heavily exercised, it is not only its thermal power that challenges server cooling, but the corresponding tightening of the processor's temperature limits.

With considerably more thermal energy moving out of the processor, the temperature difference between the actual silicon and the package's case (where it meets an air-cooled heat sink, see **Figure 2**) needs to widen to help increase the internal heat flux. Unless this is achieved, the silicon will not be able to stay within critical silicon temperature limits.

Figure 2 **Cross-section of a processor**



UPTIME INSTITUTE INTELLIGENCE 2022 Uptime Institute | INTELLIGENCE

To guide hardware vendors on the required cooling performance, chipmakers specify a case temperature (Tcase) for each processor model. If the server cooling is unable to maintain Tcase below the specified level (either because it is inadequately specified or the facility supplies insufficient volumes of cold air), the processor silicon will likely throttle (deliberately slow down) under high-intensity workloads to protect itself from thermal damage.

Processors with restricted Tcase limits — currently as low as 60°C (140°F) as opposed to the norm of around 80°C (176°F) or higher — are not a novelty. Intel, for example, has been offering processor models optimized for peak frequency with relatively fewer cores. These tend to use more power relative to the silicon area, which in turn requires a higher heat flux toward the case. However, these models currently also tend to have a relatively moderate overall thermal power rating compared with their current higher-end many-core siblings (which dissipate as much as 280 W), making life easy for server cooling engineers.

The combination of very high power and low Tcase limits is a more recent and challenging development for volume server processors. Handling larger amounts of lower-temperature heat, such as 280 W at 65°C (149°F), is thermodynamically disadvantageous (more difficult to eject) compared with a lower amount and hotter source of heat, such as 180 W at 80°C (176°F). Having multiple high-power / low-temperature heat sources, such as processors and accelerators, in proximity within the constraints of a server chassis makes cooling next-generation servers even more difficult.

Another factor that contributes to some operating temperature restrictions is the integration of dynamic memory chips on the processor package — a design innovation used to speed up data access for high-performance applications. Memory chips have lower operating temperature requirements than compute silicon, so may require more cooling. As the practice becomes more economical, we expect it to gain wider adoption because it boosts performance across a broad set of workloads.

Some next-generation server processors (which are due out by the end of 2022 / early 2023) will have a thermal power rating of 350 W, paired with a Tcase limit in the 55°C to 60°C (131°F to 140°F) range. Initially, servers equipped with these processors will be rare (most will exist in supercomputers), but they will gradually become readily available for the wider market.

These temperature requirements will lead to even more compromises in server configuration and potentially to a loss of server density due to larger form factors (e.g., 3U instead of 2U). Subsequent generations of processors will only exacerbate the issue.

Conclusions: the operator's quandary

There is a lack of information from chipmakers and their IT systems partners on what server densities and thermal ratings to expect in the future

Data center operators that want to make sure their facilities can keep up with IT hardware in the 2020s face a dilemma. Future requirements for power and cooling in five years (let alone 10 years) is unclear, bringing uncertainty and financial / business risk. There is a lack of information from chipmakers and their IT systems partners on what server densities and thermal ratings to expect in the coming years, even directionally.

This is further complicated by the variety of IT technology choices available to systems architects. Will IT infrastructure teams want highly performing servers? Will they also want to densify racks and, if so, how dense? What will be the sustained utilization level of next-generation IT in five years? Will future servers be predominantly liquid-cooled and, if so, will that be a requirement or a benefit?

The baseline option for operators is not to upgrade, but to consider their facility assets fixed and then see what they can support with little to no additional cost. Counterintuitively perhaps, many data centers with more traditional designs and operating envelopes, such as chilled water systems supplying air at 20°C (68°F), will be in a relatively better position compared with more recent builds that are highly optimized for cooling efficiency using economization and elevated operating temperatures.

Either way, the risk in this approach is the inability for a facility to support a growing number of server configurations in the future, potentially including highly performing "workhorse" servers that IT teams will likely want for a cloud-style infrastructure. This would mean losing out on the performance and energy efficiency benefits of such systems — or potentially losing business in the case of a colocation provider.

Continued reliance on air cooling for next-generation servers are likely to lead to higher parasitic IT power in the forms of server fans driving harder and electrical components (including silicon) exhibiting higher power losses due to elevated component temperatures. Even though the additional losses in parasitic power due to air cooling will vary greatly and are difficult to establish, modeling and anecdotal evidence suggests it could typically be between 10% and 20% of server power. Next-generation servers are likely to edge closer to the upper end of this range. Any airflow issues in the data hall, such as hot spots or insufficient inlet air pressure, will further amplify already sizeable parasitic losses.

On the other end of the scale of change (and cost) lies a vision of a highly densified, direct liquid-cooled infrastructure — either in the form of a new capacity expansion or a full refurbishment. While the objective of this strategy is to make sure the data center asset is future proof, there are currently no obvious specifications to plan and design around. If the usual assumptions around average rack density no longer anchor data center design, it is unclear what that number should look like in five to 10 years' time.

The population of ultra-high-density racks above 40 kW is growing fast, yet remains niche — only large computational problems justify such dense infrastructure. In the next decade, 40 kW racks are likely to be far more common, due to a combination of silicon power trends and an effort to reduce spatial needs and the cost of infrastructure.

For those building or designing data center assets to last 10 to 15 more years, that is a plausible scenario to plan for. Provisioning a facility build or upgrade to 40 kW per rack even in just power distribution capability seems like a costly excess, but a data center that cannot handle such demands may become prematurely obsolete.

There are also complexities around designing and operating DLC at scale. While the potential benefits in efficiency, sustainability and overall infrastructure performance are clear, scale adoption of DLC is still problematic. The root of the issue is the lack of standards for DLC. Unlike air cooling, where air acts both as the standard cooling medium and as the interface between facilities and IT, there is no comparable standard coolant or common means of mechanical integration with the server hardware, which creates material compatibility and technical support issues for IT hardware. This has prevented DLC from achieving mass adoption, and suppliers from scaling up. While Uptime Intelligence expects these barriers to gradually ease, it will take years for DLC to become dominant in server cooling.

The most attractive middle-ground option for operators may be to ensure the infrastructure is modified for future changes, but not fully commit to supporting them immediately. This option is probably the trickiest too: planning for and delivering an upgraded capacity in a live environment and installing new pipes to support DLC systems across data halls is likely to be challenging. Not all data center sites lend themselves well to such flexibility. A major benefit of the approach, however, is that it keeps technology options open. These include not just innovations in DLC, but developments of novel heat sinks that improve the thermal performance of air cooling sufficiently to meet future needs.

One thing is certain, whichever response an operator chooses the decision will be neither straightforward, nor without its potential downsides.

About the author

Daniel Bizo is Uptime Institute's Research Director. He has been covering the business and technology of enterprise IT and infrastructure in various roles for 15 years, including a decade as an industry analyst and advisor.
dbizo@uptimeinstitute.com

About Uptime Institute

Uptime Institute is the Global Digital Infrastructure Authority. Its Tier Standard is the IT industry's most trusted and adopted global standard for the proper design, construction, and operation of data centers – the backbone of the digital economy. For over 25 years, the company has served as the standard for data center reliability, sustainability, and efficiency, providing customers assurance that their digital infrastructure can perform at a level that is consistent with their business needs across a wide array of operating conditions. With its data center Tier Standard & Certifications, Management & Operations reviews, broad range of related risk and performance assessments, and accredited educational curriculum completed by over 10,000 data center professionals, Uptime Institute has helped thousands of companies, in over 100 countries to optimize critical IT assets while managing costs, resources, and efficiency.

Uptime Institute is headquartered in New York, NY, with offices in Seattle, London, Sao Paulo, Dubai, Singapore, and Taipei.

For more information, please visit **www.uptimeinstitute.com**

All general queries:

Uptime Institute
405 Lexington Avenue,
9th Floor, New York,
NY 10174, USA
+1 212 505 3030
info@uptimeinstitute.com