INTELLIGENCE UPDATE

Maximizing server efficiency is becoming more difficult


Daniel Bizo

19 Nov 2024

The power efficiency of IT infrastructure largely defines the overall energy performance of a data center, an observation that Uptime Intelligence reports have repeatedly made (see *Efficient servers hold the key to energy efficient data centers*) — and will continue to make in the context of sustainability.

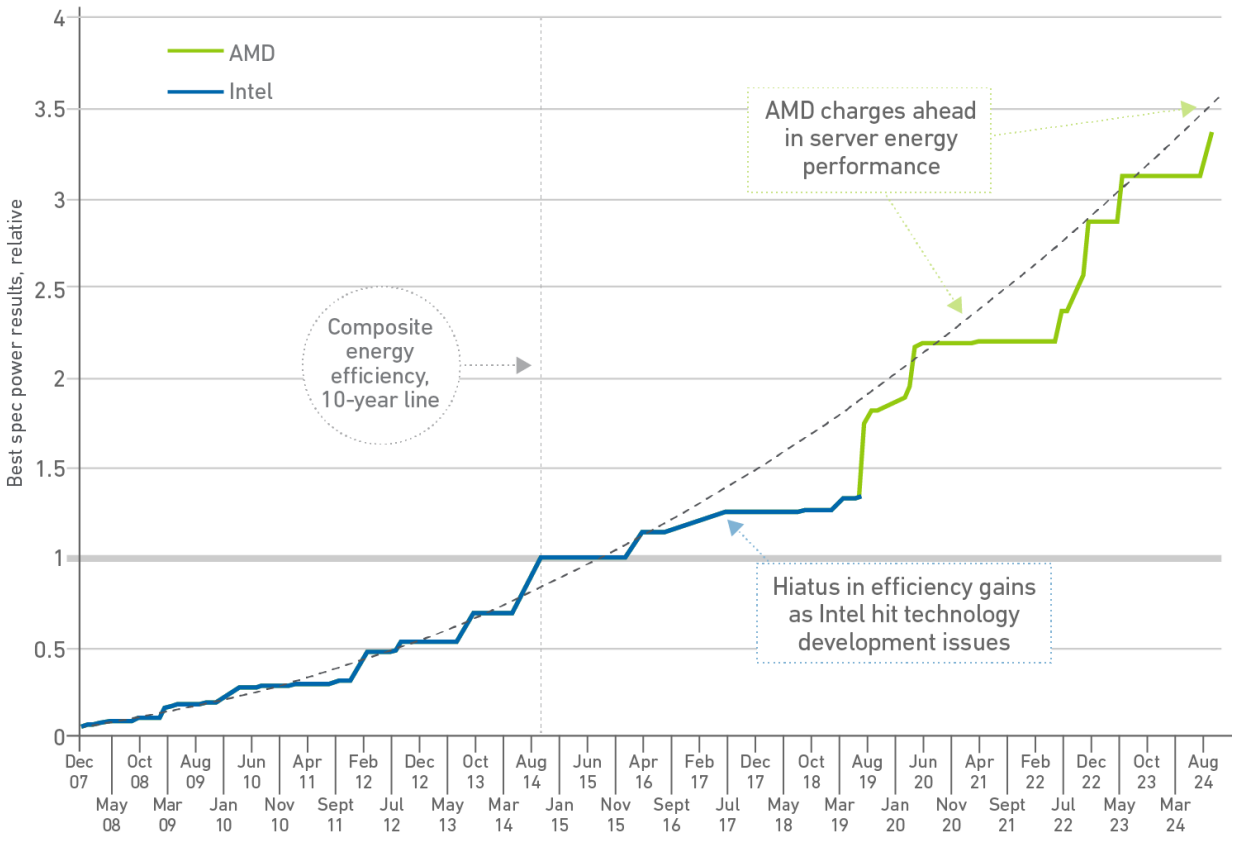
While there are numerous elements in play when it comes to total IT efficiency, the crux of the issue can be reduced to a couple of major factors. The most effective dial that operators can turn to improve work per unit of energy is server utilization, which can be improved by consolidating workloads onto fewer servers. For large and growing workloads, replacing systems with new servers can further boost efficiencies.

The long-term trajectory of server power efficiency continues to be shaped by the evolution of processor silicon. Colloquially (if inaccurately) often referred to as Moore's law, this evolution encapsulates the dynamics of semiconductor technology development, manufacturing economics and market forces. As semiconductor technology progresses, more components can be economically integrated onto silicon, which in turn means more freedom for chip architects to make design choices. Crucially, IT buyers tend to find this increase in silicon functionality valuable enough to pay for it, creating the market force that propels chipmakers forward.

Slow down in server efficiency gains

The latest generation of server platforms from Intel and AMD, which were launched in the third quarter of 2024, continue along the historical trajectory. Since 2019, AMD has been leading in overall power efficiency for most server workloads. Its 2024 server processors, codenamed Turin, lift power efficiency closely in line with long-term historical trends, as measured by the Standard Performance Evaluation Corporation's (SPEC) Power benchmark, the industry standard (see **Figure 1**). According to this data, the power efficiency potential of server technology has increased by a factor of nearly 3.4 in 10 years. The benchmark runs a Java-based business logic and measures performance and average power at regular 10-percentage point increments.

Figure 1. Best-in-class server efficiency (long-term trend), October 2014 = 1



STANDARD PERFORMANCE EVALUATION CORPORATION, COMPILED BY UPTIME INTELLIGENCE

Meeting this trend of year-to-year performance improvement still represents a gradual slowdown in power efficiency gains over the long run. The slowdown seems inevitable with current semiconductor technology unless there are major innovations in processor and system architecture. Also, extra efficiency tends to come at the cost of increased power use: top-end Intel and AMD processors are now thermally rated at 500 W with a configurable option to 600 W.

Uptime Intelligence has noted before that not only are manufacturing-driven power efficiency gains slowing down (see *Server efficiency increases again — but so do the caveats*), but that it is becoming less straightforward to extract the efficiency potential from newer systems. This is because those gains only materialize when the system is relatively well utilized (at least 20% to 30% on average), which puts an emphasis on workload consolidation for relatively light applications. At the same time, running severely underutilized machines is becoming costlier: the idle power consumption of servers released in the past five years has leapt when compared with older systems from the 2010s.

It has become even more difficult to realize the full power efficiency potential of systems launched in 2024. Initial submissions to the SPEC Power benchmark database suggest that the latest servers should be utilized well above 50% to eke out any substantial efficiency advantage over the previous generation of systems. Achieving such high levels of utilization is difficult for most organizations (outside the high-performance computing domain), however, it may be possible in some public cloud and large-scale web services environments where workload

placement is dynamic.

Outlook

Arguably, this narrowing of power efficiency opportunity does not pose an imminent threat to gains in infrastructure energy performance. Few organizations will consider replacing servers that are only a couple of years into their life cycle, yet, when compared with aging servers older than five years, the latest machines can significantly outperform them in both performance and efficiency. However, if this trend continues, it will become a concern for the data center industry because other methods of extracting infrastructure efficiencies are fraught with complexity — such as the use of server power and performance management, or software tuning (e.g., code changes, caching and resource management).

On the competitive dynamics side, Intel's 2024 server processors (codenamed Granite Rapids for performance-oriented chips) took a major step toward closing the performance and power efficiency gap with AMD. Intel can now match the core count (up to 128 performance cores in a socket) of comparable AMD products and has regained competitiveness across a range of workloads, particularly in high-performance computing.

For IT buyers, the market picture is complicated by swings in relative performance between Intel and AMD chips from workload to workload. These depend on differences between processor microarchitectures, memory speeds and software characteristics — including varying speed-ups from performance optimizations of code. This is another important point to remember for power efficiency too: matching the right hardware with the right software and applying performance tuning are becoming a must to achieve superior power efficiency compared with previous generation systems.

Importantly, AMD's latest server processors retain a considerable frequency advantage, often clocking 20% to 30% higher than Intel products at comparable core counts and thermal ratings, owing mostly to TSMC's lead in manufacturing technology. This should give AMD an edge in a wider range of workloads that respond well to higher clock speeds. Intel continues its long, forced march to catch up, ramping up new manufacturing technology nodes and server products in 2025 and 2026. How far competitive dynamics will push processor power levels and core counts before an inevitable change in design direction remains unclear.

An upcoming Uptime Intelligence report will detail the performance and efficiency of the latest Intel and AMD chips as more benchmarking data becomes available.

ABOUT THE AUTHOR

Daniel Bizo

Daniel serves as Senior Research Director at Uptime Institute. Over the past 15 years, he has covered the business and technology of enterprise IT and infrastructure in various roles, including industry analyst and advisor. His research includes sustainability, operations, and energy efficiency within the data center, on topics like emerging battery technologies, thermal operation guidelines, and processor chip technology.

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.