IDC

SPOTLIGHT

Sponsored by: Dell

Power availability, efficient cooling, and related environmental metrics are becoming critical for datacenter planning. CIOs and IT decision-makers must invest in efficient, highly performant, and secure server infrastructure as the foundation of a sustainable hybrid infrastructure strategy.

Sustainable Infrastructure for the AI-Driven Era

July 2024

Written by: Lara Greden, Senior Research Director, Infrastructure-as-a-Service Solutions, Flexible Consumption, and Circular Economy, and Ashish Nadkarni, GVP/GM, Worldwide Infrastructure and BuyerView Research

Introduction

The AI-driven era is upon us. Business stakeholders are demanding that their CIOs and IT decision-makers (ITDMs) make infrastructure investments that enable newer and faster ways to deliver AI-driven insights. This demand puts a strain on datacenter power, space, and cooling requirements — for both public cloud and on-premises datacenters — at a time when corporate sustainability goals are also front and center.

Power availability, efficient cooling, and related environmental metrics are becoming critical bottlenecks for datacenter planning. Recognizing power consumption as the driving factor for datacenter demand, IDC recently updated its primary metric for forecasting datacenter capacity from square footage to power (megawatt). This shift reflects the changing landscape of higher-density computing, where power more accurately correlates with the capacity to support and sustain computing operations.

AT A GLANCE

KEY STATS

- Datacenter energy consumption is now of material importance. IDC forecasts that global datacenter electricity consumption will grow at a CAGR of 22.6% from 2022 to 2027, growing from 320TWh to 887TWh in 2027.
- Server infrastructure and specifically CPU choices can make a big difference. Over 40% of end-user organizations identify processors (CPUs) as the source of resource bottlenecks or limitations for their on-premises server infrastructure.

For CIOs and ITDMs, the decision to invest in datacenter infrastructure is often driven by a need for data security, which can make it difficult to go to the public cloud for many of their AI initiatives. Consequently, many are shifting to a hybrid infrastructure strategy as they modernize their workloads and invest in AI and generative AI (GenAI) initiatives. Given the need for data security, IDC's research finds that for CIOs and ITDMs, on-premises private clouds remain the preferred location for performance-intensive workloads, including AI, high-performance computing (HPC), and analytics environments.

However, the need for increased investments in on-premises infrastructure means additional pressure on already stressed IT budgets and datacenter capacity. With a hybrid infrastructure strategy, CIOs and ITDMs can consider high-performance hardware to achieve greater computing capacity while minimizing the need for additional power, cooling, and datacenter floor space. In taking a hybrid infrastructure approach that includes high-performance hardware, IT leaders can expand computing capacity while reducing total cost of ownership (TCO) and maintaining focus on data security and datacenter sustainability objectives.

Addressing sustainability at the datacenter level has a facilities angle — that is, choices made for procuring power from sustainable energy sources, efficient cooling solutions, and energy-efficient (LEED-certified) facilities. However, the most efficient kilowatt or megawatt is the one that was never needed in the first place. With datacenter power supply capacity under growing pressure, IT organizations are looking closer at their infrastructure investments. A capable server infrastructure, including x86-based servers with high core count and memory bandwidth, is not just operationally efficient but also enables consolidation of workloads, delivers fit-for-purpose performance for AI workloads, allows more efficient use of datacenter floorspace and cooling capacity, and meets the strategic imperatives for today's CIOs.

Why Should Sustainability Matter to CIOs and ITDMs?

The need to manage datacenter sustainability in the AI-driven era is urgent. With more workloads requiring high-performance compute, storage, and networking, the IT industry will face challenges related to resource scarcity and rising costs, particularly around power consumption and greenhouse emissions. IDC forecasts that global datacenter electricity consumption will grow at a CAGR of 22.6% from 2022 to 2027, rising from 320TWh to 887TWh in 2027. Owing to this significant growth in resource requirements, sustainability considerations have risen prominently. In IDC's March 2024 *Datacenter Operations and Sustainability Survey*, enterprise and service provider datacenter operators indicated that sustainability is a top 3 initiative and expect it to remain a top initiative two years from now.

How Can CIOs and ITDMs Address Sustainability in Datacenters?

CIOs and ITDMs can address their sustainability objectives by the datacenter in two principal ways. First, they do that at the infrastructure level with efficient, fit-for-purpose solutions complemented by a hybrid cloud decision framework for their enterprise and especially for their performance-intensive workloads. Second, they invest in appropriate power and cooling solutions to ensure that datacenter consumption remains well within their TCO objectives for infrastructure.

Together, these two approaches ensure the organization can increase datacenter efficiency, reduce emissions, and meet organizational sustainability goals in a budget-friendly manner.

Sustainable Facilities

Investments in datacenter facilities — either owned, leased, or hosted — that are certified as energy efficient includes consideration of rack design, advanced cooling systems, and renewable energy sources. For example, LEED-certified facilities are deemed sustainable by design, according to the U.S. Green Building Council, and they can include:

- Intelligent and modular rack layout and design to increase cooling efficiency
- Advanced and efficient datacenter cooling (HVAC) systems
- Monitoring, analysis, and actuation of power usage in real time
- A clean backup power system with a goal of reducing emissions, noise pollution, and fuel consumption
- Renewable energy sources such as solar and wind power to reduce dependence on the grid and fossil fuels

Efficient Server Infrastructure

Investments in efficient server infrastructure enable workload consolidation, resulting in improved capacity utilization. An efficient server infrastructure environment can:

- Deliver on efficiency and scaling objectives with workload consolidation and modernization initiatives.
- Enable seamless deployment of performance-intensive AI workloads.
- Further drive cooling efficiency at the rack level with thermal design characteristics including control systems.

How Does Server Design Influence Datacenter Sustainability Objectives?

Server platforms with highly efficient CPUs drive overall datacenter efficiency, delivering maximized outcomes while minimizing power, space, and cooling requirements. By increasing the utility of a given CPU, enterprises can run more performance-intensive (e.g., AI-enabled and AI-centric) applications and workloads on significantly fewer servers in the datacenter, helping reduce power consumption. Finally, highly efficient CPUs also drive power and cooling efficiencies at the rack level, further supporting datacenter sustainability objectives.

Efficiency at Scale

While virtualized and containerized enterprise workloads benefit from highly efficient CPUs, performance-intensive AI workloads require performance that can scale on demand. IDC research finds that a key reason behind the failure of these initiatives is that IT organizations underestimate the role of server infrastructure for these workloads, resulting in speed and reliability bottlenecks. On the other hand, over-provisioning can lead to higher TCO. Not all workloads require high-performance infrastructure; by taking a nuanced approach, ITDMs can ensure efficient utilization of their infrastructure.

A well-designed, fit-for-purpose infrastructure with a capable processor (CPU) serves as a foundation for a higher-density footprint that services the spectrum of enterprise and performance-intensive workloads. In the case of AI, on-premises deployments are more cost-effective in cases where existing models need to be optimized, retrained, or fine-tuned on data sets that are too sensitive or large to move to the public cloud.

Workload Consolidation and Modernization

For an organization pursuing a hybrid infrastructure strategy, the choice of CPU is of paramount importance. IDC's *Enterprise Infrastructure Pulse Survey* finds that over 40% of end-user organizations identify CPUs as the source of resource bottlenecks or limitations for their on-premises server infrastructure. CPU speeds can be affected by different factors, including transmission delays, heat accumulation, memory limitations, and challenges with networking/power and cooling requirements.

The use of x86-based servers with high core count and memory bandwidth CPUs enables in-place workload modernization and consolidation. Furthermore:

- Workload modernization — a multipronged approach — can take many different paths. A virtualized environment built with servers running an efficient x86 processor (CPU) platform can deliver a seamless experience for replatforming and refactoring initiatives. Businesses can modernize many of their enterprise workloads in place, reducing costs and time as they usher in AI-centric operations.

- Workload consolidation, usually focused on reducing silos and infrastructure islands, requires the server infrastructure to scale to handle mixed workload profiles. Servers running an efficient x86 processor platform can deliver a consistent experience for workload consolidation initiatives.
- Other considerations include capex and opex costs and reduced TCO. Opex costs include software licensing costs, where the savings can be in the form of a reduction in core or socket-based licenses. Capex costs include those related to datacenter floor space buildout. IT organizations can reduce operational TCO by investing in efficient server infrastructure.

Power and Cooling Considerations

Cooling is fundamental for getting the highest performance out of server systems. It is also a primary consumer of energy in the datacenter and thus a primary contributor to sustainability impact. Innovations in cooling efficiency, to benefit both server performance and sustainability, begin at the server and rack level. By combining control systems with physical design at the server level, greater cooling capacity and efficiency gains are possible for both air-cooled and direct-to-chip liquid cooling systems.

One advantage of server-level gains in cooling efficiency is the ability to use air cooling systems for higher-performance, higher-capacity server infrastructure. Air cooling systems (as opposed to liquid cooling systems) are prevalent in datacenters. They often have the most favorable TCO terms including initial costs, maintainability, and ability to utilize existing datacenter operator skill sets. The ability to plug and play higher-density, more performant CPUs in existing rack layouts in air-cooled datacenters is an attractive option for many datacenter operators. Advanced thermal design and control systems at the server level make this possible.

However, the most demanding use cases may require direct-to-chip liquid cooling. Design at the server packaging level, including thermal control systems, is critical to gaining both heat dissipation and performance efficiency from the server infrastructure. Given the highly specialized datacenter skill sets involved in direct-to-chip liquid cooling systems, ITDMs that focus on the efficacy of the server infrastructure design have the opportunity to improve maintainability and reduce TCO while ensuring the infrastructure performs as needed for the most demanding use cases.

Other Considerations for CIOs and ITDMs

CIOs and ITDMs must take a holistic and multipronged approach to ensure that their hybrid infrastructure strategy can meet the needs of the business while also meeting budget constraints and sustainability objectives. They must start by taking stock of their investments in datacenter facilities, the power and cooling requirements of those facilities, and the infrastructure solutions housed in them. Many of these initiatives require net-new capital investments that necessitate careful planning and ROI analysis before implementation. Implementing workload consolidation and placement strategies also requires careful planning (e.g., reliance on "burst" infrastructure) to minimize disruption to the business. In addition, they must focus on zero-trust security, server life-cycle management and refresh, and server automation.

Zero-Trust Security

IDC's research finds that around 60% of end-user organizations allocate 3–10% of their annual IT infrastructure budget to server security, a figure that will increase in the next 12 months. Although 30% of organizations have a compliance-focused approach to security for server infrastructure, 27% describe their server security strategy as "reactive." Unsurprisingly, 15% have an "ad hoc" approach. A secure infrastructure provides a good foundation for an organization's

cyber-resiliency strategy. By protecting data in use (i.e., memory encryption), IT can ward off hostile actors that seek to exploit code execution vulnerabilities. By protecting data at rest, IT can create a barrier against malicious software. Servers with hardware-assisted security can deliver a complete confidential computing experience at scale. Organizations are not forced to compromise efficiency or performance to deliver a secure computing experience.

Server Life-Cycle Management and Refresh

IDC research shows that while server life spans are increasing, a strategic refresh of existing infrastructure can also provide investment capacity for new, fit-for-purpose infrastructure. This is true in capex and opex/flexible consumption spending scenarios. Regardless of the procurement model, IT asset refreshes will heighten the focus on server life-cycle management and IT asset disposition (ITAD).

IDC observes that vendors are including services for secure and environmentally sustainable ITAD as part of early-stage strategic advisory assessments. Those that can deliver on ITAD have supply chains for end-of-life processing, redeployment, recycling, and refurbished equipment sales in place. They also typically have a strong foundation in flexible consumption models and the requisite go-to-market playbooks for on-premises IT infrastructure deployment. Such capabilities solidify the ability to be a trusted partner for enterprise customers with board-level sustainability targets.

Server Automation

Automation capabilities for routine server management tasks within systems management software from a server vendor can complement the core functionality of a CPU, leading to various benefits. IDC's *Enterprise Infrastructure Pulse Survey* found that nearly 40% of end-user organizations identify improved security as one of the main benefits of server automation. This same study found that around one-quarter of respondents identify operating cost savings, improved infrastructure resiliency, and sustainability as top benefits. Server automation can contribute to operating cost savings by enabling simplified management and improved productivity, as well as making it easier to scale and optimize the performance of servers. By improving server efficiency, automation can enhance sustainability by reducing the datacenter's carbon footprint.

Choosing a Trusted Partner for the Journey

CIOs and ITDMs are better served by seeking out trusted partners that can help with long-term planning and execution. While do-it-yourself approaches can seem attractive, they can be perilous, especially with larger environments. IT staff availability and skills can also influence these decisions. IDC research shows that partnering IT staff with a trusted and experienced partner can help an organization with decision-making. Partnering may also increase the speed at which the benefits of infrastructure investments start accruing.

Considering Dell

Dell PowerEdge Servers with AMD EPYC processors (CPUs) are designed to deliver efficiency, performance, cyber-resiliency, and TCO objectives in hybrid infrastructure environments. They build on the capabilities of the AMD EPYC CPU family to deliver power-efficient performance for demanding enterprise workloads, including AI. With a trusted partner like Dell, businesses can gain consistent and assured service quality in their environment.

Dell PowerEdge rack servers powered by AMD EPYC CPUs are designed to address existing and future enterprise and emerging workload requirements. The servers are paired with Dell's OpenManage integrated IT management system. They offer the following capabilities:

- **Accelerated AI innovation:** The servers are designed to deliver business agility and time to market and support transformational workloads such as databases and analytics, virtualization, software-defined storage, virtual desktop infrastructure, containerization, HPC, AI, and ML.
- **Advancing sustainability:** Energy efficiency and sustainability are top priorities. Dell PowerEdge servers — with advanced thermals and cooling options — are efficient and performant and can serve as a foundation for a sustainable datacenter. With the help of tools like Dell OpenManage Enterprise, IT organizations can gain nearly 5:1 consolidation (as claimed by Dell) in their environment with an EPYC-based Dell server infrastructure.
- **Zero-trust security:** The servers are designed for secure interactions with the ability to predict potential threats. Dell PowerEdge servers feature cryptographically verified hardware integrity, dynamic system lockdown, and robust boot and firmware protection anchored by a silicon root of trust.
- **Intuitive systems management:** The servers are designed to increase the observability and automation of IT infrastructure and provide visibility into key operational metrics. Dell OpenManage helps discover, deploy, monitor, manage, and upkeep the PowerEdge server infrastructure.

Challenges and Opportunities for Dell and AMD

CIOs and ITDMs consider IT vendors as partners in their digital transformation journey. The level of trust a company places in an infrastructure provider is related to its ability to support efficient datacenter infrastructure. For Dell, delivering AMD EPYC-powered server infrastructure provides the following differentiation:

- **Efficiency and sustainability:** CIOs and ITDMs are looking to invest in on-premises and design-efficient infrastructure. This enables the organization to achieve or exceed its sustainability goals while reducing its datacenter footprint, including through workload consolidation.
- **Fit-for-purpose performance:** This includes the ability to host performance-intensive AI workloads alongside other business- and mission-critical enterprise workloads. The server infrastructure must be capable of managing latency and bandwidth-sensitive workloads alongside memory and compute-intensive workloads.
- **Delivering a secure-by-design infrastructure:** Incorporating security features in the hardware, starting with the CPU, minimizes the risk of malicious attacks. Dell can add other hardware-level security, such as silicon root of trust, secure boot, and other firmware protections.

AMD and Dell should continue to articulate their value proposition in a way that resonates with CIOs and ITDMs. The differentiation for Dell and AMD is to deliver efficient, sustainable, and secure infrastructure solutions that transform their relationship with CIOs and ITDMs into that of strategic and reliable partner.

Conclusion

An efficient, highly performant, and secure server infrastructure is at the heart of a scalable hybrid infrastructure strategy in today's power-constrained datacenter industry. More and more businesses see on-premises infrastructure as the foundation for their hybrid infrastructure operating model. A capable CPU that powers this server infrastructure can enable the business to consolidate its workloads onto a smaller footprint, gain datacenter efficiency, and meet organizational sustainability objectives. The organization can invest in AI-enabled server automation to gain visibility into server operations and carbon footprint and to reduce its TCO.

A capable CPU that powers this server infrastructure can enable the business to consolidate its workloads onto a smaller footprint, gain datacenter efficiency, and meet organizational sustainability objectives.

About the Analysts

Ashish Nadkarni, *Group Vice President and General Manager, Worldwide Infrastructure and BuyerView Research*

Ashish Nadkarni leads IDC's worldwide research on compute and storage infrastructure systems, platforms and technologies, enterprise, emerging and performance-intensive workloads, cloud and edge infrastructure and infrastructure services, and infrastructure software platforms. He also manages IDC's BuyerView research portfolio.

Lara Greden, *Senior Research Director, Infrastructure-as-a-Service Solutions, Flexible Consumption, and Circular Economy*

Lara Greden leads IDC's worldwide research on IT infrastructure-as-a-service (aaS) solutions, flexible consumption models, leasing markets, and circular economy sustainability strategies. Her analysis provides insight from both a supply-side and a buyer's point of view, with core research *coverage* including circular economy and sustainability for IT assets and the evolution of procurement strategies for better operating models from purchasing, leasing, and financing to as-a-service, flexible consumption models.

MESSAGE FROM THE SPONSOR

Together, Dell Technologies and AMD redefine data center excellence with unparalleled efficiency, reducing the need for servers, racks, and power consumption while delivering top-tier performance.

Dell PowerEdge servers powered by AMD EPYC processors push workload boundaries with tailored IT and business solutions, all while helping your business lower energy consumption and meet sustainability goals. AMD's data center solutions, including their EPYC CPUs, are designed with power efficiency in mind, utilizing advanced technologies such as 7nm process technology and a high-performance architecture to minimize energy consumption while maintaining high levels of performance. AMD EPYC processors provide 50% more core density with up to 47% better performance per watt over the previous generation — based on Dell Technologies' internal benchmark testing (2022) — enabling a highly efficient data center that helps you reduce your business's carbon footprint.

Learn more at dell.com/servers/AMD.

IDC Custom Solutions

The content in this paper was adapted from existing IDC research published on www.idc.com.

IDC Research, Inc.
140 Kendrick Street
Building B
Needham, MA 02494, USA
T 508.872.8200
F 508.935.4015
Twitter @IDC
idc-insights-community.com
www.idc.com