White Paper, sponsored by Lenovo and AMD | April 2025

Transforming Cloud Service Provider (CSP) Datacenters for Next-Generation AI and HPC Workloads

How CSPs Can Benefit from Lenovo ThinkSystem Servers Featuring AMD Instinct GPUs

Chris Drake

Senior Research Director,
Compute Infrastructure and Service Provider Trends,
Worldwide Infrastructure Research, IDC

White Paper

Transforming Cloud Service Provider Datacenters for Next-Generation AI and HPC Workloads

How Cloud SPs Can Benefit from Lenovo ThinkSystem Servers Featuring AMD Instinct GPUs

Sponsored by: Lenovo and AMD

Chris Drake
April 2025

IDC OPINION

A new generation of servers is needed to support growing demand for artificial intelligence (AI)/machine learning (ML), hybrid cloud, and high-performance computing (HPC) capabilities. Faced with increasing demand from their customers, (non-hyperscaler) cloud service providers (SPs) are responding by expanding their GPU-as-a-service offerings and the infrastructure that supports them. (For the purposes of this paper, cloud service providers refers to all infrastructure-as-a-service and software-as-a service providers that are non-hyperscalers.)

As they look to support the growth of their GPU-as-a-service businesses, cloud SPs need to address a range of challenges that include differentiating in a competitive market, ensuring sustainable growth, and addressing fluctuating customer expectations and evolving skill set requirements. Key investment priorities will include the deployment of more sophisticated compute platforms, the expansion of GPU capacity and capabilities, the transformation of datacenter environments, and the enhancement of other supporting infrastructure such as networking and security architectures.

SITUATION OVERVIEW

Many cloud SPs report strong customer growth for their GPU-based service offerings and expect that growth to continue over the next few years, thanks to customer demand for artificial intelligence/machine learning, hybrid cloud, and high-performance computing capabilities. However, growing demand for accelerated compute capabilities will require ongoing (and in some cases significant) enhancements to cloud SPs' accelerated compute architectures. These enhancements will include the adoption of a

new server and GPU infrastructure, along with the transformation of service provider datacenter environments.

Investments in new server platforms and GPU capabilities will focus on ensuring necessary performance enhancements, and especially performance at scale while enabling cloud SPs to achieve important operational and cost efficiencies and enhanced ROI.

Meanwhile, the transformation of datacenter environments will include investments in new datacenter power and cooling technologies, especially liquid cooling. It will also involve the application of automation technology and smart management systems to optimize datacenter resource usage.

Over the next two years, cloud SPs also plan to make significant investments to their network infrastructure to support the growing demands of their compute infrastructure environments. In addition to the adoption of 800GbE for AI and HPC clusters, specific investments will focus on strengthening network security and the increased use of automation to improve network management and operational efficiency.

Key Challenges for Cloud Service Providers

Many cloud SPs recognize that, in order to successfully grow their GPU-as-a-service business, they will need to navigate and tackle several challenges that include differentiating in a competitive market, preparing for potential supply chain disruptions, managing changing customer expectations and the need for new and evolving skill sets among their employees, and ensuring that they can sustainably grow their IT infrastructure.

Differentiating in a Competitive Market

To remain competitive in a growing but challenging market environment, cloud SPs emphasize the importance of several different strategies that include offering competitive pricing while protecting margins, capturing market share via the acquisition of new customers or users, and looking for opportunities to gain early market entry within emerging sectors.

However, many cloud SPs believe that one of the most important strategic initiatives for ensuring competitiveness involves differentiating their existing products, services, and business models. Key focus areas for cloud SP differentiation include specializing in tailored industry solutions, the ability to offer hybrid and other advanced solutions, and the strengthening of customer service.

Some cloud SPs will also look to establish or expand their relationship with strategic partners as a way to help them differentiate. These include partnerships with systems integrators that can bring valuable skill sets and specialized expertise. They also include partnerships with specific chip vendors, whose reputation can give cloud SPs a unique selling proposition.

> We differentiate ourselves by focusing on hybrid cloud and AI solutions, emphasizing industry-specific expertise in the financial services, healthcare, and telecommunications sectors. — Technical manager, large cloud service provider, the United States

Potential for Supply Chain Disruptions

Business operating environments continue to be a challenge for cloud SPs, which express concerns about how geopolitical shifts, tariffs and trade wars, semiconductor shortages, and inflation could all potentially affect supply chains. Some expect increased costs and delays in logistics and production, while others focus on specific industry impacts, such as energy availability for datacenters, resource scarcity, and chip shortages affecting automation device production. However, many cloud SPs recognize the need to establish measures that can help them mitigate the impact of potential supply chain disruptions. Proactive supply chain management, strategic inventory planning, vendor diversification, and technological investments are all widely seen as essential for navigating the evolving supply chain landscape.

For cloud SPs that are ramping up investments in AI infrastructure and GPU-based service offerings, the need to mitigate and minimize potential supply chain disruptions becomes even more important. Having a choice of infrastructure provider and a business model based on multiple vendor suppliers is seen by many cloud SPs as a key risk mitigation strategy.

Changing Customer Expectations

Many cloud SPs report experiencing significant changes in the expectations of their customers over the past five years, which include an expectation of faster times to deployment, as well as the ability to leverage self-service capabilities and automation tools that can help them reduce the operational burden across a range of monitoring, remediation, and orchestration tasks.

> Customers increasingly demand faster deployment times, expecting our solutions to be readily available and easily integrated into their existing systems and workflows. — Technical manager, large cloud service provider, Australia

Other cloud SPs say their customers expect more professional support to help them enhance operational efficiency and navigate complex implementations, such as hybrid cloud adoption and other solutions. Customers also want access to new technologies such as real-time data analytics to help them with strategic planning and proactive customer support. To address these challenges, cloud SPs will prioritize the increased use of automation and data-driven insights, as well as the development of support channels such as online portals and ongoing investments in staff training.

Overcoming the Skills Challenge

Many GPU-as-a-service providers talk about the challenge of securing and cultivating in-house GPU expertise, and this is partly due to the speed of hardware development and the specialized nature of skills. GPU technology continues to evolve at a rapid pace and cloud SPs report that, with each new platform version, there is a need to ensure that their technical, sales, and other personnel are up to speed with the latest capabilities. Cloud SPs report particularly high demand for specialized AI/ML and GPU programming skills.

Skill shortages in high-demand areas such as AI, cloud computing, and cybersecurity are cited by many cloud SPs as pressing issues that can lead to increased competition for talent, pressure for higher salaries, and high levels of staff turnover, which can pose challenges for productivity and project completion.

To address these challenges, some cloud SPs rely on external support, which includes working directly with systems integrators. Meanwhile, others are partnering with technical universities to help them recruit the latest skilled graduates.

Sustainable Growth

A further challenge for cloud SPs is the need to ensure that they grow their businesses in ways that maximize the efficient use of datacenter and other resources. Here they face a host of intertwining requirements that include the need to manage total cost of ownership (TCO) with regard to their IT footprint and resources. Meanwhile, almost everywhere cloud SPs face pressure from regulators, customers, investors, and ecosystem partners to demonstrate their sustainability credentials, which include a commitment to reducing IT waste and carbon emissions.

To address the sustainability challenge, cloud SPs are promoting energy efficiency in their accelerated computing environments, prioritizing the use of sustainable hardware, and optimizing resource utilization.

Optimizing resource utilization includes the use of techniques such as workload consolidation and the rightsizing of datacenter and other resources. Meanwhile, many cloud SPs are committed to using modern, energy-efficient datacenters that leverage advanced cooling technologies and high-efficiency power systems.

> Our approach to sustainability focuses on reducing carbon emissions and promoting energy efficiency in our infrastructure, particularly in accelerated computing. We also prioritize sustainable hardware. — Technical manager, large cloud service provider, United States

Service Provider Investment Priorities

As they focus on expanding their GPU-as-a-service businesses, cloud SPs identify several investments priorities. These priorities span the deployment of more sophisticated compute offerings, the expansion of GPU capacity and capabilities, which for many include an emphasis on vendor choice, the transformation of datacenter environments, and the enhancement of other supporting infrastructure such as network architectures.

More Sophisticated Compute Offerings

Cloud SPs report "steady" to "substantial" and even "dramatic" customer demand for their GPU offerings, and many expect this growth to continue over the next 12–24 months. Demand for GPU capabilities is linked to the growing use of AI/ML, as well as HPC. In addition to GenAI initiatives and the use of ever-growing large language models (LLMs), specific AI initiatives involving GPU clusters include training deep learning (DL) models, running neural networks, and accelerating inferencing tests. Cloud SPs identify finance, manufacturing, healthcare, and retail as industries where demand for GPU solutions is particularly strong.

> Our typical GPU customers are in industries like finance, healthcare, manufacturing, and retail, using GPUs for AI training, inference, data analytics, and HPC. — IT manager, large cloud service provider, United States

As their customers look to deploy more demanding workloads that incorporate AI, HPC, and advanced data analytics, cloud SPs need to invest in the expansion and development of their own compute offerings. This includes being able to ensure that their servers offer maximum GPU capacity and the latest hardware and software GPU capabilities at price/performance levels that are attractive to customers. They also need to be able to offer their customers flexible and scalable accelerated compute offerings. Important investment priorities for cloud SPs include the ability to offer customers a

choice of deployment, consumption, and configuration options — including a choice of vendor solution options.

Although some cloud SPs specialize in offering private, dedicated GPU offerings, many offer their customers a choice of both private and general, multitenant options. Private GPU instances are tailored for individual clients requiring dedicated resources, enhanced security, or customized configurations, while general (multitenant) instances offer cost efficiency while maintaining performance. Shared instances also cater to customers wanting to experiment with small-scale workloads, with the option to later move to private instances as their workloads grow.

Increasingly, cloud SPs are also prioritizing the ability to offer diverse pricing structures, flexible consumption options, and service and support measures that accommodate different customer needs and budgets. Specific initiatives include the use of pay per use and capacity-based models, as well as subscription-based pricing with discounted pricing for longer-term commitments. Some cloud SPs also offer financing solutions to help customers reduce up-front capital expenses and align costs with business growth, and cost management and consumption monitoring tools that include usage alerts, budget notifications, and recommendations for optimizing GPU resources.

Diversification of GPU Platforms

Although many cloud SPs continue to rely on a single GPU vendor, they recognize the benefits of having more than one supplier. In addition to being able to offer customers a choice of GPU options, vendor diversification is seen by cloud SPs as a way of mitigating supply chain disruptions, price volatility, and vendor lock-in. However, those that currently rely on a single provider argue that the move to diversify must be demand driven and based on growing customer demand for GPU services and capabilities. As their GPU business grows, this strengthens the case for expanding the choice of GPU options (see Figure 1).

> While our current demand may not justify large-scale adoption of alternative GPUs, we're considering a gradual expansion of our vendor base. — IT manager, large cloud service provider, India

FIGURE 1

Cloud SP Arguments for Diversifying GPU Supplier

Supply chain risk

- Diversifying can help organizations mitigate the risks associated with supply chain disruptions and price volatility.

Cost-performance balance

- Diversifying can help ensure cost-performance balance, cost-effectiveness, and reduced dependency on a single supplier.

Deployments at scale

- Growing demand for GPU services increases the scale of infrastructure deployments, strengthening the case for a diverse product range.

Source: IDC, 2025

For cloud SPs, the purchase attractions and inhibitors of different GPU platforms focus on a range of touchpoints, including price, performance, supply chain dynamics, ecosystem support, and software integration. For some cloud SPs, a hardware accelerator's reliance on a proprietary software ecosystem is a potential risk that can increase the possibility of being locked-in to a specific vendor ecosystem. However, AMD offers support for open source software, including AMD ROCm software and the OpenCL software stack (often identified as a major attraction of the AMD platform).

Many cloud SPs recognize the benefits of an open source software ecosystem for GPUs. These include increased interoperability, innovation, collaboration, and cost-effectiveness while reducing vendor lock-in.

> An open software ecosystem for GPUs could be a game changer, promoting broader innovation and compatibility across different hardware vendors. If the ecosystem improves software portability, ease of integration, and reduces vendor lock-in, it's definitely something we'd consider supporting. — IT manager, large cloud service provider, India

Transformation of Datacenter Environments

Cloud SPs generally rely on a mixture of their own datacenters and colocation environments to deliver their GPU-based service offerings, with the actual use of datacenter infrastructure varying according to a cloud SP's size, geographic distribution, service portfolio, and business model. Many cloud SPs are currently focused on rightsizing their use of datacenter infrastructure to support evolving workload requirements, with many expecting infrastructure spending to increase in 2025. Investment priorities include the adoption of renewable energy sources and energy-efficient servers. They also include the use of automation to enhance datacenter and datacenter resource management and the implementation of modular design to facilitate easier expansion in response to rising demand.

Specific cloud SP strategies to enable more efficient datacenter management include the use of process automation and smart management systems to optimize resource allocation and energy use in real time, as well as the use of AI to monitor and optimize consumption. As workloads and data processing requirements become more demanding, cloud SPs also need to ensure that their datacenters have a reliable and sustainable source of energy. Specific challenges for cloud SPs to navigate include rising energy costs, inefficient energy use, and a dependence on nonrenewable energy sources. In some places, concerns include electricity grid constraints and power quality issues.

> As energy costs rise globally, the operational costs of running a datacenter also increase. This is especially impactful for data-heavy tasks like AI model training where power consumption is high. — Datacenter manager, large cloud service provider, Canada

Many cloud SPs also recognize the benefits that liquid cooling systems can offer. As GPU clusters continue to become more powerful and as datacenter rack densities increase, many traditional air cooling technologies could, by themselves, be insufficient to sustainably cool datacenter equipment. Meanwhile, liquid cooling is increasingly being regarded as an important way of supporting the next generation of GPU-powered IT infrastructure. Many cloud SPs are expected to use a hybrid approach to cooling their accelerated compute infrastructure, with liquid cooling used to support GPU-based platforms and traditional air cooling for CPU and networking equipment.

> We intend to increase the use of liquid cooling to better meet the demands of AI/ML and HPC loads, as well as support greater energy efficiency, which will reduce costs. — Technical manager, large cloud service provider, United States

Enhancing Network Architectures

Cloud SPs are aware of the need to invest in their network infrastructure to support growing accelerated compute environments with powerful new GPU capabilities. The evolution of cloud SP networks will focus on increasing bandwidth, reducing latency, and enhancing scalability, ensuring that networks can support the intensifying compute workloads across AI training, real-time data processing, and data-heavy applications. Investments in high-speed, low-latency networking solutions such as InfiniBand can help facilitate rapid data transfer between GPUs and storage systems, enhancing overall performance. In addition, the use of redundant network paths can help ensure uptime and reliability, critical for service-level agreements (SLAs) with clients.

Cloud SPs identify a mixture of network investment priorities, which include upgrading to 400/800GbE for AI and HPC clusters and 400Gbps InfiniBand for GPU-to-GPU communication within AI clusters. They will also include enhancing network security and mobile connectivity and increasing the use of automation to streamline network management, reduce manual interventions, and improve operational efficiency.

Some cloud SPs also anticipate increased use of edge computing, which will require a distributed network architecture to process data closer to where it's generated, thereby reducing latency and bandwidth use. Meanwhile, the adoption of software-defined networking (SDN) will allow for more flexible and efficient network management, enabling dynamic allocation of resources based on real-time needs.

> We plan to make significant investments in our network infrastructure to support the growing demands of our compute infrastructure environment. Key investments include 5G integration, edge computing, network automation, and network function virtualization. — IT manager, large cloud service provider, Australia

Other Investment Priorities

To support the development of a successful GPU-as-a-service business, cloud SPs recognize the need for additional infrastructure investments across several areas. These include the adoption of high-performance storage systems to ensure fast data access and retrieval for GPU workloads. They also include distributed file systems that can enable efficient data management across multiple GPUs and nodes. Other investment priorities for cloud SPs will include the adoption of technologies that support GPU orchestration and management, containerization and Kubernetes, and AI/ML tools and frameworks. They will also include enhancing cloud-to-edge connectivity.

Cloud SPs also recognize the need to implement robust security measures, including encryption and access controls, to protect sensitive data processed on GPU resources, and measures to ensure compliance with industry standards and regulations, especially in sectors like healthcare and finance.

Lenovo ThinkSystem Servers with AMD Instinct GPUs

Lenovo's ThinkSystem servers featuring AMD EPYC processors and AMD Instinct GPUs are designed to address many of the challenges and investment priorities identified by cloud SPs looking to expand their GPU-as-a-service businesses. These servers support modularity and flexibility with a range of configuration options. They also incorporate various capabilities that maximize energy efficiency, as well as platform storage, security, and GPU networking connectivity.

Lenovo ThinkSystem servers featuring AMD EPYC processors and AMD Instinct GPUs enhance the cloud SP's ability to support the most demanding of workloads, including AI, HPC, and big data analytics. Additional support from Lenovo includes AI and HPC services, which aim to help cloud SPs with AI systems planning, design and implementation, and management, and Lenovo AI Fast Start, which helps customers establish the business value of specific use cases with a ready-for-product AI solution that uses customer data and focuses on their specific needs. Additional features include the ability to flexibly consume all of these server platforms via Lenovo's TruScale Infrastructure-as-a-Service business.

Lenovo ThinkSystem servers support AMD technology including:

- **AMD EPYC Processors** are designed to elevate performance, boost energy efficiency, and prepare for AI-driven innovation, and they are among the world's most powerful x86 processors.
- **AMD Instinct MI300 Series GPUs** are designed to deliver the highest levels of performance and efficiency for AI training and inference workloads. Ideal for training, fine-tuning, and inferencing large AI models and HPC workloads, AMD Instinct GPUs are powered by AMD Compute DNA (CDNA) architecture, offering performance with high memory capacity and bandwidth, scalability, and energy efficiency. AMD Instinct MI300 Series GPUs are designed to hold today's large AI models in fewer GPUs, allowing for lower overall cost.
- **AMD Instinct MI200 Series GPUs** provide advanced I/O capabilities and are designed to support enterprise, research, and academic HPC and AI workloads for single-server solutions and more.

Lenovo ThinkSystem servers configured with AMD Instinct GPUs include:

- **Lenovo ThinkSystem SR685a V3** is an 8U rack server that features two AMD 9000 Series EPYC processors and eight AMD Instinct MI300X GPUs. The server features air cooling and is designed to support the most demanding AI workloads, including GenAI applications. It supports up to 16x PCIe 5.0 NVMe drives for high-speed internal storage and 8 NDR 400Gb/s InfiniBand adapters with high-speed GPU Direct connectivity[1]. Security features include an integrated hardware Trusted Platform Module (TPM), which supports TPM 2.0 and enables advanced cryptographic functionality, such as digital signatures and remote attestation.

  The server also incorporates several energy efficiency features to help save energy, reduce operational costs, and increase energy availability. These include energy-efficient planar components to help lower operational costs, high-efficiency power supplies with 80 PLUS Titanium certifications, and the optional Lenovo XClarity Energy Manager provides advanced datacenter power notification and analysis to help achieve lower heat output and reduced cooling needs.

  The AMD ROCm open source software platform is optimized to extract the best HPC and AI workload performance from AMD Instinct MI300 accelerators while maintaining compatibility with industry software frameworks. ROCm consists of a collection of drivers, development tools, and APIs that enable GPU programming from low-level kernel to end-user applications and is customizable to meet specific end-user requirements. Once designed, the software is portable, allowing movement between accelerators from different vendors or between different inter-GPU connectivity architectures, regardless of the underlying device. ROCm is particularly well suited to GPU-accelerated HPC, AI, scientific computing, and computer-aided design.
- **Lenovo ThinkSystem SR675 V3** is a 3U rack server that supports up to eight double-wide and single-wide GPUs. The platform features a modular design to maximize deployment flexibility and support multiple configurations, including up to two AMD EPYC 9000 Series processors and up to eight AMD Instinct MI200 Series GPUs.

  The solution also offers a choice of front or rear high-speed networking and a choice of local high-speed NVMe storage with NVMe drives directly connected to the GPUs, to maximize storage performance. Security features include AMD Secure Root-of-Trust, Secure Run, and Secure Move, which are designed to minimize potential attacks and protect data as the OS is booted, as applications are run, and as applications are migrated from server to server.

[1] Lenovo ThinkSystem SR685a V3 Server product guide (**lenovopress.lenovo.com/lp1910-thinksystem-sr685a-v3-server**)

The SR675 V3 also offers energy efficiency features including energy-efficient system board components that help lower operational costs, high-efficiency power supplies with 80 PLUS Titanium or Platinum certification, solid state drives, which consume up to 80% less power than traditional spinning 2.5in. HDDs, and the aforementioned optional Lenovo XClarity Energy Manager[2].

OPPORTUNITIES FOR LENOVO

As it looks to support its cloud SP customers with its rapidly expanding GPU-as-a-service offerings, Lenovo can emphasize the benefits of numerous features and capabilities associated with its ThinkSystem servers featuring AMD EPYC processors and AMD Instinct GPUs. These systems are specifically designed to support the latest AI, HPC, and data analytics workloads, precisely the workloads that cloud SPs identify as driving significant demand for GPU capabilities among their own customers. With these servers, Lenovo can support a range of configuration and flexible deployment options — both key priorities for cloud SPs. In addition to expanding the capacity of their GPU service offerings, cloud SPs identify the need to make several adjacent infrastructure investments, including those related to networking, storage, and security. Lenovo ThinkSystem servers featuring AMD EPYC processors and AMD Instinct GPUs are also designed to address these adjacent infrastructure requirements.

In addition to these platform capabilities, Lenovo can point to several ways in which it maximizes support and flexibility for its cloud SP customers. These include the support offered through AI and HPC services and AI Fast Start, and the range of flexible consumption options that are available via Lenovo TruScale Infrastructure-as-a-Service.

Cloud SPs that currently rely on alternative GPUs sometimes point to the propriety nature of their supporting software stack and ecosystem as factors that can increase potential for vendor lock-in and higher TCO. By contrast, Lenovo can emphasize the open source software stack on which AMD GPUs are based, which can offer cloud SPs a range of benefits that include flexibility, customization, collaborative development, and cost reductions, since open source may remove the need for licensing fees.

Lenovo has a long history of developing liquid cooling technology, having first introduced its Neptune Liquid Cooling solution in 2012. Although Lenovo ThinkSystem servers featuring AMD EPYC processors and AMD Instinct GPUs currently leverage air and liquid-to-air cooling technologies, there could be opportunities for Lenovo to add more liquid cooling AMD-powered servers.

[2] Lenovo ThinkSystem SR675 V3 Server product guide (**lenovopress.lenovo.com/lp1611-thinksystem-sr675-v3-server%23key-features**)

CONCLUSION

Cloud SPs are experiencing growing demand from their customers for GPU-as-a-service capabilities to support their expanding investments in AI, HPC, and big data analytics workloads. To support this growing demand, cloud SPs are prioritizing investments in developing more sophisticated compute platforms, expanding GPU capacity and capabilities, transforming their datacenter environments, and enhancing other supporting infrastructure such as networking and security.

At the same time, cloud SPs recognize the need to address various challenges that include differentiating in a competitive market, navigating potential supply chain hurdles, ensuring sustainable growth, and addressing changing customer expectations and evolving skill set requirements.

With its ThinkSystem servers featuring AMD EPYC processors and AMD Instinct GPUs, Lenovo aims to address many of the challenges and investment priorities identified by cloud SPs as they expand their GPU-as-a-service businesses. In addition to being designed to support the latest AI and HPC workloads, these server platforms offer a range of configuration and flexible deployment options and capabilities that support cloud SPs' evolving networking, storage, and security requirements. Additional attractions include AMD and Lenovo's own support for an open source software ecosystem for GPUs — with benefits including greater flexibility, customization, collaborative development, and reduced costs for cloud SPs — and a wide range of support and flexible consumption options offered by Lenovo.

ABOUT IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. With more than 1,300 analysts worldwide, IDC offers global, regional, and local expertise on technology, IT benchmarking and sourcing, and industry opportunities and trends in over 110 countries. IDC's analysis and insight helps IT professionals, business executives, and the investment community to make fact-based technology decisions and to achieve their key business objectives. Founded in 1964, IDC is a wholly owned subsidiary of International Data Group (IDG, Inc.).

Global Headquarters

140 Kendrick Street
Building B
Needham, MA 02494
USA
508.872.8200
Twitter: @IDC
blogs.idc.com
www.idc.com