White Paper

The Power of Now: Accelerate the Datacenter

Sponsored by: Supermicro and AMD

Chris Drake
February 2024

IDC OPINION

Datacenter service providers currently face a confluence of challenges, which require that they adapt and modernize to cope with upcoming requirements of their principal customers. These include the growth and proliferation of new compute-intensive and storage-intensive workloads and applications, such as those that leverage AI. They also include applications whose viability and success depends on the ability to process application data in real time. Examples include customer service chatbots, manufacturing quality inspection, and machinery maintenance.

While a growing number of service providers need to be able to support applications with more demanding compute and storage requirements, many also face the need to handle ever larger volumes of data from a range of sources. These include operational data related to things such as customer orders and transactions, business performance metrics such as data from their organization's own IT systems, and data that is specific to the performance of consumer and business applications.

Alongside the challenges posed by the changing nature of IT workloads are the pressures datacenter service providers face to ensure their facilities comply with a wide range of environmental standards and commitments. As the need for more powerful compute and storage systems increases, this has implications for a datacenter's essential resources, including space, power, and cooling. Traditional cooling technologies fall short of being able to maintain datacenter equipment within their operating envelope, leading to subpar performance. This is driving a focus on ways of increasing a datacenter's compute capabilities while minimizing the environmental impact of cooling the IT equipment.

To adequately address the challenges posed by the impact of new workloads requiring higher levels of processing and storage, it is imperative that datacenter service providers adopt a new infrastructure stack that spans hardware and accelerated computing, rack-scale integration, a software-defined architecture, and the use of a microservices application environment.

It is now widely recognized that many enterprises will continue the need to retain mission-critical and highly proprietary workloads and data sets in on-premises and private cloud environments for cost, security, and performance reasons. Although some will leverage their own facilities, at least partially, many will choose to rely on datacenter service providers for a range of hosted IT requirements. Therefore, to ensure that their own datacenters are fully prepared to support changing customer requirements, service providers will need to deploy an entire new infrastructure stack for datacenter modernization.

This new infrastructure stack includes:

- Higher-performing CPU cores, as well as the use of distributed and composable hardware architectures, which enable a more efficient use of shared resources, as well as more scalable compute performance
- Accelerated computing, which includes GPUs for AI workloads and CPU-based accelerators for storage-intensive, security, and network workloads
- Rack-scale integration, which essentially involves the use of a converged infrastructure approach for building and scaling the datacenter, one where pretested racks facilitate faster deployment, integration, and expansion
- Software-defined datacenter technology, which extends the tenets and benefits of server virtualization to a datacenter's compute, storage, and networking capabilities for increased efficiency, management, and flexibility
- The use of a microservices application architecture, which allows for a more flexible and agile way of developing, maintaining, and upgrading applications

SITUATION OVERVIEW

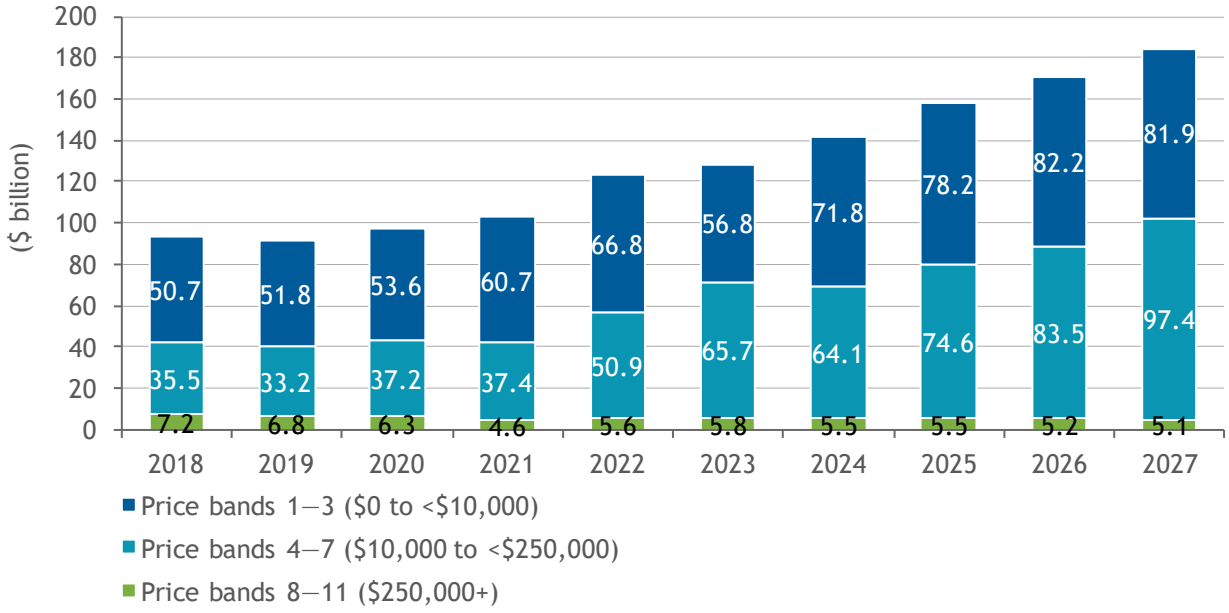
Recent and ongoing innovations in datacenter technology have important implications for enterprises and service providers, which will influence their ability to succeed in an increasingly data-driven world. We already have a glimpse into what this new data-driven world will look like. Native digital companies already collect vast data sets from customer interactions, as well as from service delivery, sales, and marketing activities. Meanwhile, a growing number of enterprises from traditional industries such as consumer packaged goods, automotive, and healthcare are also now digitizing data from a range of sources. These include customer orders and transactions, sales reporting, and billing systems, as well as data from business management applications such as customer relationship management (CRM) and enterprise resource management (ERP) platforms and data from a wide range of applications, including those that leverage IoT sensors and GPS capabilities.

In addition to collecting data from a wide range of sources, many enterprises are now applying and using that data in applications from AI-driven customer service and marketing to product and service analytics and R&D. It is the ability to manage these highly valuable sets of enterprise data and apply them in ways that make organizations more efficient, responsive, and innovative that will distinguish the next generation of leading companies.

IDC research shows that demand for compute power has been growing steadily for years. With analytics tools, real-time performance needs, media streaming, and now a rush to deploy AI/ML models in a variety of contexts, increased performance is seen by most customers, including service providers, as a key reason driving platform selection. According to IDC's Quarterly Server Tracker, demand for servers priced from $10,000 to <$250,000 (price bands 4–7) is increasing, thanks to the growth of AI-based applications and accelerated server configurations (see Figure 1). From 2022 to 2027, spending on these servers is forecast to rise from $50.9 billion to $97.4 billion, reflecting a CAGR of 13.9%. Demand for servers in price bands 1–3 will also exhibit moderate growth, whereas demand for servers in price bands 8–11 is expected to decline over the same period.

FIGURE 1

AI Drives Demand for Richly Configured Servers, 2018-2027



Note: The data is based on IDC's Enterprise Infrastructure Tracker.

Source: IDC's 3Q23 Quarterly Server Tracker, December 2023

Despite this, growing volumes of data and new data-intensive, real-time, and low-latency applications pose a real challenge for traditional datacenters, which can lack flexibility in the way compute, storage, and other resources are provisioned and allocated and in their ability to flexibly scale capacity in response to fluctuating demand.

Many of the key challenges associated with operating a datacenter relate to its use of essential resources such as power, cooling, and space. Others relate to the need to scale the datacenter in response to growing demand for compute and storage capacity.

Power use continues to be one of the largest sources of cost associated with operating a datacenter. Costs include the use of energy to power the IT equipment and, in many cases, to power necessary cooling technologies.

Unless properly planned and managed, increased demand for IT resources can result in escalating power costs, with the deployment of new compute and storage infrastructure pushing up overall power consumption. Additional challenges can arise when datacenters have power usage caps in place, and in regions where energy utilities have limited the amount of new power that can be allocated to datacenters.

Poor planning and management when expanding a datacenter's IT footprint also has potential to negatively disrupt power usage effectiveness (PUE) targets while having implications for datacenter sustainability. Data from IDC's 2023 *Service Provider View Survey* shows that service provider sustainability initiatives have a broad mixture of drivers, including national, regional, or industry-

specific regulations; the expectations of customers, investors, and ecosystem partners; and the desire to achieve more efficient operations and improved cost efficiencies. An effective power usage strategy will be central to achieving these goals.

The deployment of additional compute and storage capabilities, if poorly planned and managed, can also result in growing demand for costly colocation and on-premises enterprise datacenter space as rack space footprints expand.

Alongside the impact an expanded IT footprint can have on a datacenter's primary resources are the implications this has for datacenter scalability, particularly the time, cost, and management complexity involved in installing and integrating new IT systems. The more complex new rack systems are to deploy, the higher the risk of longer deployment time frames, increased datacenter labor costs, and the emergence of other unforeseen hurdles.

Technology Responses to the Challenges of the Data-Driven Era

Technology is already responding to the challenges confronting traditional datacenters, particularly with regard to the use of power and power efficiency, methods for cooling the IT equipment, the efficient use of a datacenter's IT space (rack space), and ways of deploying and scaling new server and storage capabilities. Several areas that stand out where technology is evolving to address these key challenges are discussed in the sections that follow.

Power Efficiency

As workloads become more demanding in terms of the amount of power required by the underlying compute and storage systems, it is imperative that those systems become more efficient in the way they consume power — if costs are to be kept under control and environmental targets met. For HPC and AI workloads, and workloads that rely on GPUs and other accelerators, the introduction of more efficient power management systems can ensure that power is used efficiently and that PUE levels are kept low. One way to achieve greater power efficiency is to cap the power usage capabilities of servers so that consumption remains within provisioned budgets. However, service-level agreements (SLAs) must still be met. In instances where power use is dynamic and unpredictable, simple power capping can be combined with additional power management techniques to achieve greater stability and flexibility. These include mechanisms for minimizing the provisioned power for each server by setting a precise power budget based on an application's specific SLA.

Datacenter Cooling

Another way to reduce power consumption and increase overall power efficiency (lower the PUE) within the datacenter is to address the type of IT cooling methods that can be used. The use of rack-level liquid cooling has the potential to considerably reduce opex compared with air cooling methods. This is largely because liquid cooling removes the need for power-hungry datacenter fans and capacity-consuming air conditioning units. Meanwhile, the denser nature of liquid means it has significantly more heat carrying capacity than air, while further efficiency is achieved by ensuring that heat is removed at the level of the rack itself. Together, these features mean that servers with CPUs and GPUs can be packaged with greater density, thereby reducing a datacenter's PUE (see Figure 2).

FIGURE 2

Supermicro Enterprise Rack with Liquid Cooling

Source: Supermicro, 2024

Datacenter Density

Tackling datacenter sprawl and the inefficient use of IT space can be a major challenge, especially when it is necessary to rapidly bring new server systems into operation. One way of ensuring that additional capacity can be added while minimizing the impact on a datacenter's rack space is to deploy higher-density servers (those containing a higher number of CPUs), GPU servers, and storage systems. Within cloud and colocation facilities, high-density racks have been deployed to accommodate the rising power requirements of AI and other HPC workloads. Enterprise datacenters need to follow suit with similar innovations to maximize the efficient use of datacenter rack space and minimize unplanned, costly expansion (see Figure 3).

FIGURE 3

Supermicro Rack-Scale Liquid Cooling Solution

Source: Supermicro, 2024

Deployment Efficiency

Another challenge that arises when scaling datacenter capacity and adding new compute and storage capabilities relates to the speed and efficiency with which new deployments are carried out. Upgrades and new deployments often encounter delays, which can result in additional costs for datacenter operators in terms of lost business. Projects to add new IT capacity also involve labor costs and these can also escalate when projects encounter delays. A more efficient way of deploying new datacenter capacity involves the use of factory-assembled, factory-tested, rack-scale systems that minimize the need for complex and error-prone onsite installation, cabling, and setup. This results in a faster time to deployment, as well as lower labor costs, and reduced risk of provisioning error.

Deploying a "Modern Technology Stack" for the Data-Driven Datacenter

To ensure that datacenters are not only properly prepared for the data-driven era but actually able to become engines for driving growth across the wider business organization, a holistic technology response is needed. While innovations in the power consumption of server systems, rack-level liquid cooling, higher-density servers and racks, and the use of pre-integrated systems (at the manufacturer's facilities) can all make a difference in their own right, a datacenter modernization strategy that combines all of these technologies as part of a full technology stack will be best positioned to capitalize on the opportunities of the data-driven era. A new technology stack for the data-driven datacenter is made up of several elements that are discussed in the sections that follow.

Distributed and Composable Hardware

Key innovations to the hardware layer include the deployment of more CPUs and higher-performing CPU cores. These combine efficiency and performance cores as well as chiplet-based CPU architectures such as Universal Chiplet Interconnect Express (UCIe), high bandwidth memory (HBM), and high-performance I/O and interconnects using PCIe Gen 5 and CXL. This enables the deployment of distributed and composable hardware architectures, which allow a more efficient utilization of shared memory resources, as well as higher-scalar compute performance, higher-core count CPUs, and faster memory and I/O bandwidth.

Compute Express Link (CXL) is a cache coherent interconnect for CPUs, memory expansion, and accelerators. This is a fundamental-enabling technology that supports distributed composable systems where memory can be a pooled resource that can be shared by different CPUs. The current CPU architecture using direct attached memory often results in stranded memory if more memory is installed than needed. According to the Semiconductor Research Consortium (SRC, January 2021), an average of 49% of the compute node cost is memory. The pooled memory approach allows the dynamic reallocation of memory to the CPU that requires it. CXL and pooled memory are still in the early phases of technology adoption, but it may become a key technology in the future similar to the adoption of virtualization. CXL can also increase the efficiency of the datacenter.

Diversified Processors

To support a more diverse range of high-performance and data-intensive workloads, core CPU processing capabilities should be supplemented with the use of hardware accelerators, including accelerators for AI workloads. The use of supplementary hardware accelerators allows CPU utilization to be freed up and reserved for end-user workloads, rather than much-needed capacity going to support infrastructure workloads. It therefore facilities a more effective way of managing the compute-intensive applications that use accelerators. Recent product releases from AMD are specifically designed to target emerging AI workloads in the datacenter. In December, AMD announced new products, including the AMD Instinct MI300X GPU, which is designed to support generative AI workloads and large language model training and inferencing, and the MI300A accelerated processing unit (APU), which combines AMD CDNA 3 GPU cores with the vendor's latest Zen 4 x86-based CPU cores to provide performance for HPC and AI workloads.

AMD's approach to supporting emerging high-performance workloads in the datacenter includes a focus on both GPU accelerators and CPUs and solutions such as the AMD Instinct MI300A, which combine both GPU and CPU technology. On the CPU front, the company continues to evolve and enhance its AMD Ryzen Embedded processor range, a scalable x86 CPU portfolio that is designed to support a wide range of high-performance workloads.

Converged Rack-Scale Infrastructure

To support more efficient deployment and a more manageable and scalable datacenter environment, enterprises should aim to emulate the rack-scale integration and standardization principles already widely established by the cloud service providers. An IT environment built on preassembled, pre-integrated racks is best placed to prepare datacenter operators for the data-driven era, thanks to its ability to support faster and more efficient deployment, integration, and scalability.

One best practice used by many cloud providers is to standardize a few rack-level hardware configurations, which fit their workloads and scale to different numbers of concurrent users and applications. This allows easier maintenance and logistics, ordering, testing, and deployment.

Software-Defined Systems

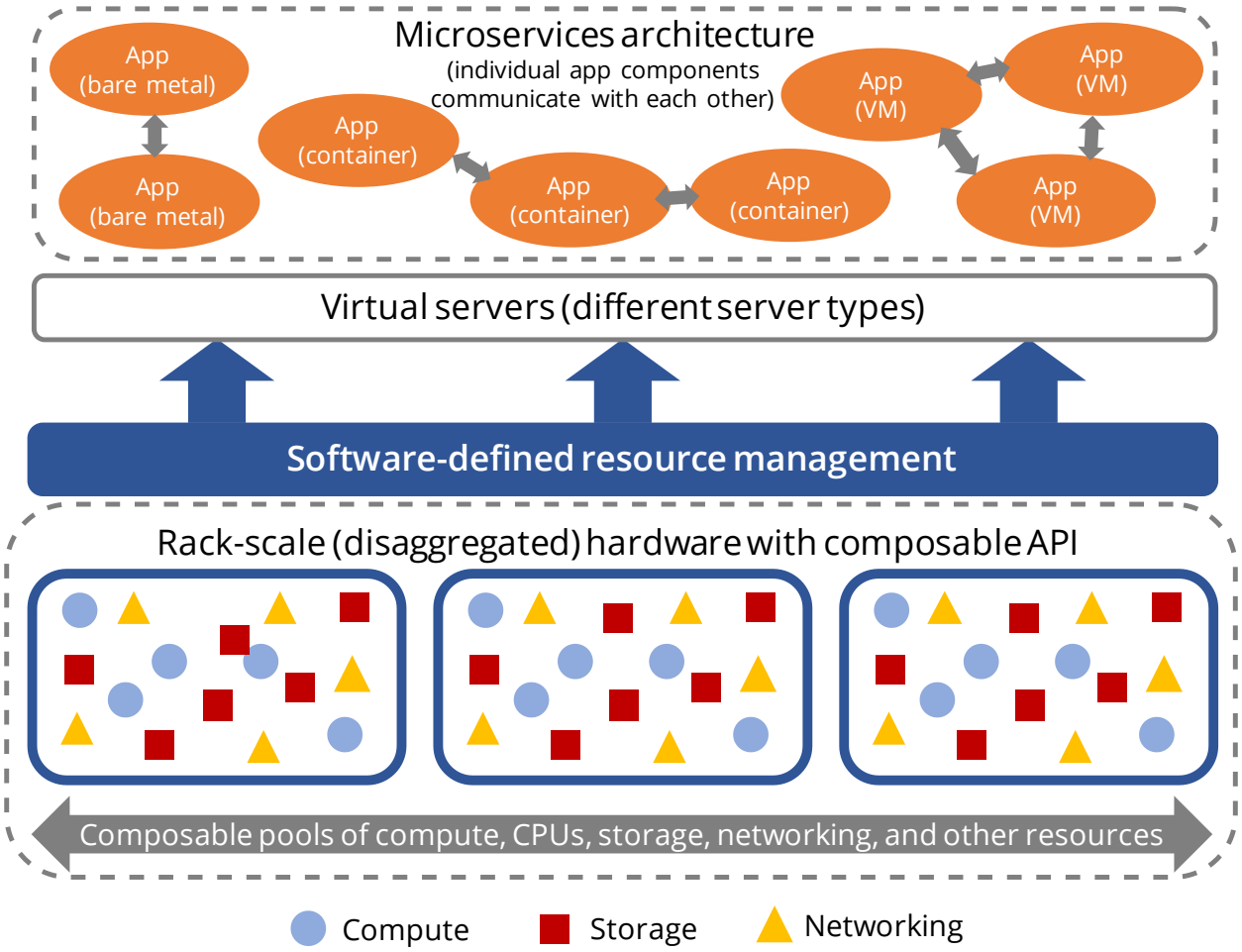
Within datacenters designed for the data-driven era, software ensures that the principles of manageability, flexibility, and simplified scalability can be applied across all of a datacenter's essential resources. A software-defined datacenter is one in which virtualization concepts such as abstraction, pooling, and automation are extended to all of a datacenter's compute, storage, networking, and other resources with the goal of achieving a full IT-as-a-service environment within the datacenter and can improve the overall efficiency of the datacenter. A software-defined datacenter allows service providers to simplify and centralize the management of their datacenter assets, ensuring that compute, storage, and networking resources can be allocated with increased agility and flexibility. Additional benefits include enhanced datacenter security and the creation of more finely grained security policies that are application centric.

Microservices Application Architecture

The next generation of enterprise software applications will use a microservices architecture instead of the previous generation of client/server and web application architecture. Microservices is an architectural approach that involves dividing large applications into smaller, functional units capable of functioning and communicating independently (see Figure 4). Microservices enables a more modular and agile continuous integration/continuous delivery (CI/CD) approach to application development and the allocation of underlying hardware resources. It also allows for easier maintenance and non-forklift upgrades through separate service modules that use technologies such as Kubernetes or Docker, open source software and frameworks, and containerized applications. The adoption of a microservices architecture can help service providers manage and allocate their datacenter resources more effectively and efficiently. Older application-based infrastructures can become more challenging to maintain over time due to expanding infrastructure capabilities and the need for more frequent patches and system updates. This makes it harder to make smaller, more frequent changes to software and ensure quick update cycles. Updating these aging software models is therefore a key benefit of microservices. Microservices can be used in conjunction with standard underlying servers. The use of serverless tools, which allow developers to build and deploy an application without the responsibility of having to manage the underlying servers, ensures flexibility and mobility for the various microservices within an application by eliminating server dependency.

FIGURE 4

A Technology Stack for the Data-Driven Datacenter



Source: IDC, 2024

EMBRACING THE MODERN DATACENTER — AN ACTION PLAN FOR CHANGE

Many service providers have already started engaging with one or more elements of the previously mentioned technology stack for modernizing their datacenter environments. However, many continue to have a piecemeal and disconnected approach to datacenter modernization that underachieves on things such as automation, infrastructure management, and the delivery of datacenter resources as a service.

Service providers continue to face numerous hurdles to longer-term plans to modernize their datacenter portfolios. These include the challenges associated with integrating new technologies, such as the need for specific skill sets among IT personnel and the need to test and certify new technologies – or new versions of existing technologies – prior to integration. Other challenges include the need to juggle datacenter modernization efforts alongside other IT investment priorities, and the fact that, if

existing technologies work, there will often be a reluctance to make, what appear to be, unnecessary changes.

Ultimately, a joined-up and aligned approach to datacenter modernization will better serve service providers' longer-term objectives to achieve greater IT efficiency, agility, and increased innovation. However, it is also important that datacenter service providers have a clear road map for moving to new platforms and systems. Service providers need to ensure that service availability does not suffer during the migration from legacy systems. They also need to ensure that customer data is protected via a technology ecosystem that guarantees both cybersecurity and physical security.

As they look to implement a successful strategy for datacenter modernization — one that will properly prepare them for the age of data-driven applications — service providers should consider the following:

- **Transformation road map.** Service providers should develop a road map for transforming their datacenter assets and adopting a technology stack that prepares them to harness the opportunities of the data-driven era. An effective road map needs to be realistic with regard to time frames and costs and needs to prioritize the safety of customer data, as well as things such as service-level agreements. A useful starting point may be developing a full inventory of existing datacenter assets (hardware and software) and identifying those that are in greatest need of being upgraded. Wherever possible, a holistic approach should be pursued.
- **Full-stack portfolio provider.** Datacenter service providers should consider working with IT vendors that are able to offer them a complete solution portfolio that spans servers, hardware accelerators, storage, and networking equipment. Working with technology suppliers with a broad selection of product options can help ensure that service providers benefit a more tailored solution for their specific workload requirements. In instances where not all elements of the technology stack are not directly available from a specific provider, service providers should ascertain how a supplier can work with partner suppliers to deliver a full solution. Supermicro has a strong track record of working with service providers to help them transform their datacenter environments. A total IT solutions manufacturer offering servers (including AI servers), storage, switching systems, software, and support services, the company is able to provide fully integrated racks and clusters from various manufacturing locations worldwide. Customer application software can be installed and tested and liquid cooling solutions from the CPUs to the rack and cluster level can also be tested and validated.
- **Accelerator workload match.** Service providers have access to a continually growing market of hardware accelerator options, and it is essential that they deploy the right accelerators for the sort of workloads they are likely to support. This will mostly require a full assessment of how the requirements of key customers are expected to evolve over the next two to five years. In identifying the right accelerator mix for specific workloads, service providers should look to work with suppliers that can help them evaluate the benefits of new technologies through testing. AMD's expanding portfolio of hardware accelerators and CPU-based products is designed to support a wide range of emerging high-performance workloads in the datacenter. AMD works both directly and via its partners to help its customers select the right processors for specific environments and workloads.

- **Service provider experience.** As they look to embark on their datacenter modernization strategy, service providers should look for technology suppliers that have a proven track record of helping transform and scale service provider architectures. Learning from the experience of others should be a key part of this process, and service providers should look for suppliers with experience in designing, testing, building, and deploying of rack-scale solutions for service providers, including workload proof-of-concept testing, software testing, and onsite deployment and support.
- **Supply chain ownership.** Datacenter service providers should consider technology suppliers that have full ownership of the board, system, and rack design (liquid cooling technologies) and that are able to manage the complete supply chain — including the essential management of the vertical integration needed to combine the power supply equipment, cooling systems, and rack infrastructure. Technology suppliers with diverse regional manufacturing locations and high economies of scale could also prove to be beneficial to buyers by reducing both deployment costs and time to market.
- **Long-term vision.** Together with planning for the initial transformation of their existing datacenter assets, service providers should put in place a long-term plan for maintaining their datacenters and ensuring they keep pace with the latest technological developments. Service providers should plan for technology refreshes and ensure they have both an appreciation of and a road map for necessary configurations and upgrades. They should also have a clear support plan that includes an understanding of what can be dealt with in-house versus what needs to be handled by partners.

CONCLUSION

To ensure that their datacenters are able to support the AI and other high-performance workloads that will be an inevitable feature of the data-driven era, it is important that service providers embrace a modern technology stack for their datacenters as soon as possible. A modern datacenter technology stack represents a holistic model for ensuring that datacenters are agile, flexible, resilient, and scalable. Such a stack includes the use of a distributed and composable approach to managing and allocating a datacenter's chief hardware resources, including servers and processors, storage, and networking resources. Building on the benefits of server virtualization, it also leverages a software-defined approach to resource management, one that also extends to the application layer and to the use of a microservices architecture for developing and delivering applications. The use of a microservices architecture can equip service providers with a more agile, flexible, and scalable solution that enables continuous application deployment and faster release cycles.

Alongside the use of software to manage both a datacenter's infrastructure resources and the applications they support, it is inevitable that datacenter modernization will include the use of hardware accelerators and other advanced processors — including the latest range of CPU-based processors. Service providers are presented with an ever-expanding choice of accelerator and processor options, making it necessary to work with IT vendors, partners, and customers to establish the correct processor mix for the workloads they are likely to support.

The deployment of new servers and processors needs to be done in ways that safeguard a service provider's commitment to sustainability goals. Innovations to server design, including the use of rack-level liquid cooling, can help ensure that datacenter expansion and enhancement does not compromise sustainability.

Last, a modernized datacenter needs to be scalable, ensuring that additional datacenter capacity and resources can be deployed in a quick, flexible, and cost-effective manner. Datacenter service providers should make use of converged, rack-scale infrastructure that comes factory assembled and pretested. This approach will ensure faster and more efficient deployment, integration, and scalability.

As they transition toward adopting a modern datacenter architecture for the data-driven era, service providers may find it beneficial to work with an IT solutions vendor that can help them address the different layers of the technology stack simultaneously. Supermicro is well positioned in this regard, with its ability to provide fully integrated racks and clusters that combine a datacenter's server, accelerator, and other resources, and which include the option to install and test customer application software and liquid cooling options.

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