White Paper

Cases for Operator-Owned and Managed Infrastructure: Interviews with Three Leading Companies

Sponsored by: Supermicro

Natalya Yezhkova Ashish Nadkarni
June 2022

IN THIS WHITE PAPER

This white paper provides overview business cases of enterprises choosing to deploy and manage their own on-premises compute and storage infrastructure. These business use cases summarize IDC's in-depth interviews with Intel, Twitter, and Preferred Networks (PFN) (a deep learning software company based in Japan). All three enterprises are customers of Supermicro, sponsor of the white paper. This white paper explores the benefits these customers recognized by running on-premises IT operations and their long-term IT strategies and challenges.

SITUATION OVERVIEW

Digitization of business and consumer services accelerated in the past years, putting pressure on enterprise IT infrastructure already overwhelmed by increasing growth in digital data and information that needed to be processed, stored, analyzed, distributed, and protected. Every business has some form of a digital footprint to deliver its services, sell products, and communicate with its customers. Emerging workloads that help businesses automate operations, expedite product development, enhance customer experience, and boost employee productivity all rely on agile and resilient IT operations. IT departments have a multitude of choices for building such operations while maintaining control over budgets and attracting IT talent to support business growth and transformation.

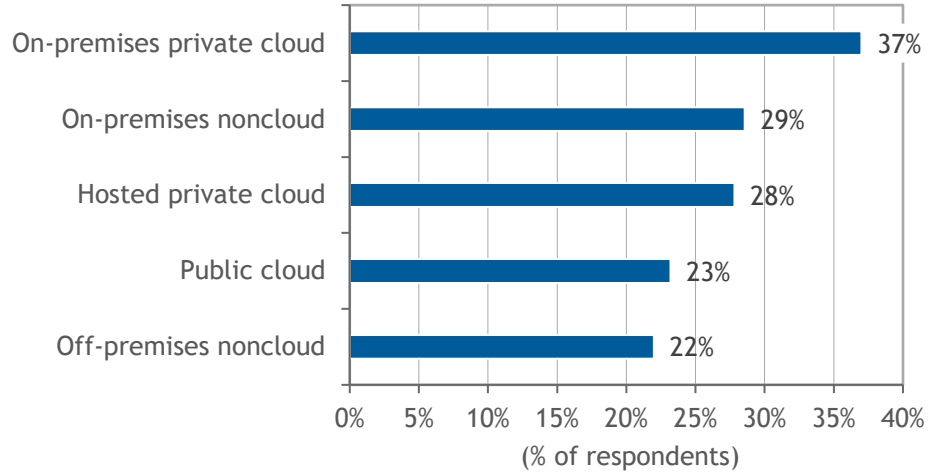
In this environment, owning and operating IT infrastructure continues to be one of the major areas of investments: IDC estimates that in 2021, half of the spending on server and storage infrastructure was driven by on-premises deployments. Further, IDC expects that these investments will continue to grow in the next five years at a compound annual growth rate of 2.9% and will reach $77.5 billion in 2026. In a recent IDC survey on server and storage workload deployments, respondents identified on-premises environments among top choices for running their workloads in the next two years (see Figure 1).

June 2022, IDC #US49236122

FIGURE 1

End-User Preferences for Workload Deployment Environments

Q. *Thinking about your workload infrastructure in the next two years, in what environments do you expect your workloads will run?*



n = 2,325

Note: Multiple responses were allowed.

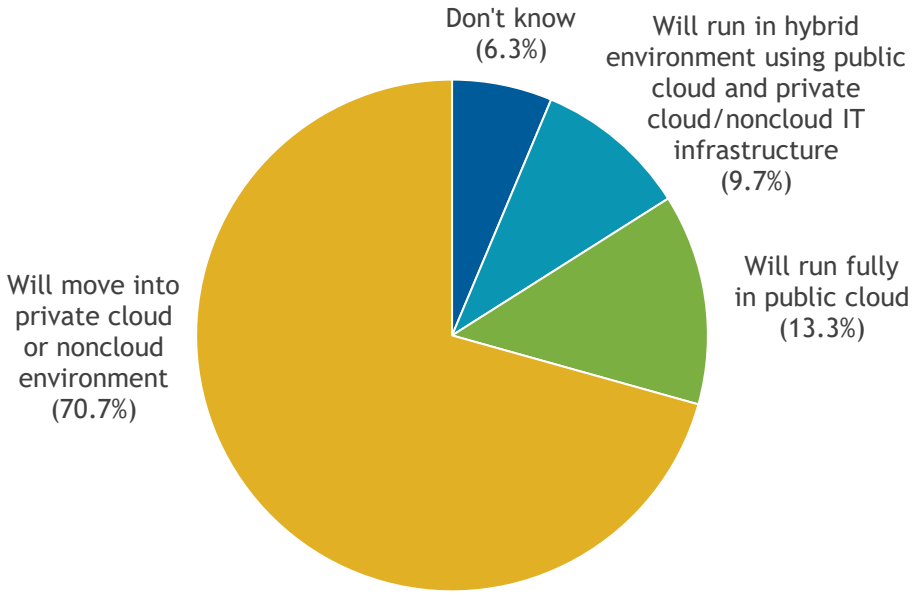
Source: IDC's 1H21 *Servers and Storage Workloads Survey,* August 2021

In the digital economy, workload placement isn't always a permanent decision. Depending on workload performance in different environments, changing needs, and monetary and other considerations, end users want to have flexibility to move workloads among different IT environments or distribute various functions of a particular workload (e.g., data generation, production, backup, and archiving) across multiple locations. This flexibility is indicated by workload repatriation activity, a process of moving workloads from public cloud into dedicated on-premises and off-premises environments. IDC observes this activity since the public cloud became a mainstream option for running IT operations, and it's not slowing down over time. In the aforementioned survey, 71% of respondents expected to move, partially or fully, their workloads currently running in public cloud into a dedicated IT environment in the next two years (see Figure 2).

FIGURE 2

Workload Repatriation Activity

Q. *Thinking about your workloads that currently run in the public cloud (fully or partially), do you expect any of them will be repatriated and moved to private cloud or noncloud infrastructure in the next two years?*



n = 2,325

Note: Survey includes 7,487 workloads across 2,325 respondents.

Source: IDC's 1H21 *Servers and Storage Workloads Survey,* August 2021

What drives the continuous demand for dedicated infrastructure and managing own IT operations? Cost is always a dominant factor in IT investment decisions, whether it impacts selection of IT consumption models or selecting vendors and solutions. To explore other factors behind investments in dedicated and self-managed IT infrastructure, IDC conducted interviews with Intel, Twitter, and Preferred Networks (a deep learning software company) to learn what benefits they recognized as a result of these investments. The use cases represent different types of businesses, but they all have something in common – these organizations partnered with Supermicro to deploy computing hardware optimized to the needs of their core workloads. By doing so, all three organizations maintained full control over their infrastructure to gain the most value out of IT investments and achieve their business goals.

Leading Technology Company: Efficiency and Sustainability

Intel is one of the largest technology companies in the world. Besides its core expertise in infrastructure component design and manufacturing, Intel focuses on technology innovations that enable safer, healthier, and greener environments across the globe. To fuel its innovation efforts, Intel runs its own global infrastructure that includes 16 datacenter sites with 56 datacenter modules, hosting more than 380,000 servers, more than 3.6 million cores, 787PB of storage capacity, and more than 725,000 network ports. About 95% of the servers within this massive infrastructure deployment

are used for chip design in high-performance computing and 3% are used for traditional enterprise and office workloads, while the remaining 2% are used in the manufacturing computing space, which includes fabrication and assembly test manufacturing plants.

In the past two years, Intel observed increasing rates of growth in its infrastructure needs: previously, the number of cores was growing at 21%, and in the past two years, growth accelerated to 38% year over year; the actual compute demand, measured in EDA MIPS (electronic design automation million instructions per second), was growing at 31%, and in the past two years, this rate of growth jumped to 43% year over year. Since 2003, Intel reduced the number of its datacenter modules from 152 to 56 by closing inefficient legacy datacenters and building modern, extremely energy-efficient, high-density, and hyperscale datacenters. In fact, with this move, Intel more than doubled the amount of power in its datacenters and estimates that its support for up to 43kW per rack at scale capability is the highest in the industry.

As a global technology leader with strong R&D operations, Intel designs a variety of its own datacenter and IT infrastructure elements. In 2016, Intel IT designed a disaggregated server architecture, which now accounts for the majority of server deployments in Intel's datacenters: "The disaggregated server is what we want … In traditional datacenters, every four to five years, when people replace hardware, they replace everything including rack. There is no reason to do it. Within a server, if you really look at it, there are many things like power supplies, fan units, drives, many other components … they don't change technology wise, they change infrequently." The premise of disaggregated server is to optimize the life of its components without replacing them all based on replacement cycles for CPUs or memory modules, whose capabilities are upgraded more often.

> "*It [disaggregated architecture] serves two purposes. One is business advantage: it reduces the amount of money you need to upgrade; you only pay for what you want to upgrade to realize the value. And the second one is significant ewaste reduction: if you look at a 3U chassis, nearly 82% of the weight of the material you don't need to change.*"

Since 2016, Intel collaborated with Supermicro on the development of disaggregated servers. Intel designs CPU complex – a unit that provides most of the performance benefits by upgrading more frequently compared with other server components, allowing users to save on upgrade costs: "If I'm getting two times more cores, then at one-third of the cost, I can upgrade the CPU complex." Memory upgrades, for example, can go on a separate cycle as they deliver lower boost to the performance of the computing system while being about as costly as CPU upgrades: "If I add memory, then it takes around 55–56% of the money to upgrade."

While Intel maintains a small fleet of standard servers in its datacenters for running different workloads, all its storage needs are fulfilled by systems from major storage OEMs. Intel's message to them is to also look at disaggregated designs: "We have a greater role in the industry – we want to influence sustainability, reduction of ewaste. Green computing is not just being energy efficient, green computing is reducing ewaste." Intel is passionate about the benefits of disaggregated architecture it introduced in its own datacenter operations and computing infrastructure design.

> "*Intel has huge goals on sustainability. It's not just economical value. It has an ethical value, it has a social responsibility, it has a commercial value. It has many, many aspects of it. Intel believes in it.*"

Another area, in which Intel has made progress in its on-premises IT operations, is infrastructure management: "Going back in time, we had around 800 people to manage [Intel's IT infrastructure] in early 2000s … Now we have around 550 people to manage 10 times more capacity. We have moved our employee mindset from IT operational to engineering, more automation, more AI and ML workloads."

Intel's own infrastructure is a core to the company's future innovation plans: "Going forward, we want to make sure that we do more innovation, more breakthroughs from our end. From a business standpoint, we need to take the cost out and automate more and, also, get more predictability in about when a particular server may fail, or when storage can fail, when the network [may fail] … so we have a preemptive way of managing and maintaining those things."

Social Network Provider: Cost and Flexibility

Twitter, one of the largest social network providers, operates three datacenters housing hundreds of thousands of servers and multiple petabytes of storage in the United States. All datacenters are hosted in colocation facilities, but the company operates and manages its own compute and storage infrastructure while working closely with the datacenter operations team to address any issues on the datacenter infrastructure management side. These datacenters account for the majority of Twitter's IT infrastructure; however, the company also utilizes public cloud resources for certain workloads and data (e.g., cold storage) and targets to achieve close integration between its own dedicated and shared cloud IT resources.

> *"We can tailor our hardware solutions to fit what we need and strip out what we don't need and make that pretty cost effective. We can do things like that we wouldn't normally be able to do with just an off-the-shelf solution or in the cloud without paying extra money for it."*

Similar to the other two companies that participated in the study, Twitter opted from using general-purpose servers in its datacenters: "We find that the off-the-shelf solutions are going to be more expensive because they are more versatile and more general-purpose based while we need systems that can be tailored specifically to our needs."

Another benefit of utilizing on-premises infrastructure is ensuring adequate levels of security and data privacy: "[Security and data privacy] are important, but not a huge concern because we deploy into our datacenters that we control, so we don't have access from external sources directly into there. So we have a lot more control over the entire infrastructure … each individual server doesn't need to have as much resiliency built into it."

Twitter is prone to handle spikes in online traffic during big events, and hence its dedicated infrastructure is built to accommodate these spikes without interruptions or adding additional hardware. However, the company sees how it can utilize off-premises cloud resources for offloading infrastructure demands during such events.

While tuning characteristics of servers to the needs of a variety of services run by Twitter's teams and infrastructure ownership delivers significant cost savings, this comes with all the challenges of infrastructure self-management and life cycling: "We have to monitor our hardware. We have to check quality and make sure we are monitoring failure rates and things like that in order to come to root cause of issues on our own or with the help of integrators and vendors … so we have to absorb that cost as well … we have to deal with the old hardware, whether tech refreshing it, whether reselling it, or

recycling it … We have to maintain a hardware team to actually design servers. We have to maintain a supply team … to procure parts and orchestrate all of the building of the servers … "Collaboration between the teams and with colocation service providers is critical for ensuring smooth infrastructure operations. For example, the engineering team, which manages datacenter infrastructure, works closely with the colocation provider to distribute servers across racks to accommodate power and space requirements, while the hardware management team works with the datacenter operations team to provide the technical expertise needed to identify the need and perform equipment repairs.

Despite the typical challenges associated with managing and maintaining its own IT infrastructure, Twitter isn't planning to scale down on its usage: "… we feel the on prem will still be the backbone of our service." In fact, when Twitter acquires a company that utilizes public cloud for its IT operations, the IT team evaluates whether public cloud is the best option or Twitter's own infrastructure is more beneficial and repatriates workloads as needed. Working with internal customers, the Twitter IT team has the ability to track whether current hardware meets their needs and project how these needs, whether it's related to compute, storage, or network bandwidth characteristics, might change in the next two to three years. This process provides a base for new requirements to server configuration and technical characteristics.

Looking forward in the future of its IT, Twitter realizes that flexibility and choices are prevalent factors in infrastructure consumption: "We need to have a better relationship between the on-prem and cloud solutions … The way it's going to evolve is having tools to be able to move back and forth and having tools to compare what the differences are, and to make them more similar, but also be able to highlight differences so that decisions can be made about what is the best."

> *"The ownership of the on-prem work is obviously our bread and butter. Most of our services and workloads are on prem. Our job is to provide our services with options. Price is obviously a big factor of that, but it may not really be THE factor."*

Deep Learning Software Company: Infrastructure Control and Performance Predictability

Based in Japan, Preferred Networks is an eight-year-old company that designs and develops one of the most powerful supercomputers in the world. Its primary business is developing and implementing projects requiring deep learning techniques. In fact, the company developed its own deep learning framework and libraries, making them available as open source to the developer community. The company also has expertise in robotics, pharma, and other business domains.

The nature of PFN's core business dictates computing requirements, which, as the company realized, it can only create and manage by itself. Yusuke Doi, VP of Computing Infrastructure at PFN, said commenting on this decision: "The GPU computing on the cloud can be very difficult to use, expensive, and uncontrollable … The cloud computer is not ours so we cannot dig into the hardware details, including the network and storage … "Six years ago, the company started its first supercomputer project building a GPU cluster housing 1,024 NVIDIA GPUs, and by now, it has already developed the third generation of its supercomputer expanding its capacity to 2,000 GPUs.

> *"Cost [of on-premises infrastructure] is one of the benefits, and the other is performance."*

With extensive compute capabilities came high power requirements, which couldn't be satisfied by ordinary datacenters: "Prebuilt datacenters cannot give us good kilowatt power per rack." The company decided to host its GPU cluster at one of the country's major supercomputer centers that delivers adequate power and cooling. The trade-off was in a lack of redundant power and security, so the company had to build its own security perimeter within the datacenter.

Besides the computing cluster, PFN also designs and maintains its own storage. "Deep learning data access is more uniform, there is not much variation — there are many random-access operations, but it's mostly really a read-intensive task." To serve these tasks, in addition to NFS, the company introduced Hadoop cluster and uses HDFS as its large-scale object storage. For data protection and availability purposes, PFN utilizes the three times redundancy strategy.

Supermicro is the company's longtime major technology partner: "We can make an ASIC, but we cannot make a server. We asked Supermicro to help us build a server to place our deep learning accelerator into the computing system." Being young, the company doesn't yet have an established strategy on hardware replacement but expects to see at least five years of lifetime of its infrastructure in the future while preparing for its first cycle of hardware retirement and expansion: "Our latest accelerator is much faster than an accelerator that was shipped five or six years ago. We have to buy better GPUs, so we have to build better processes for ourselves."

PFN manages its extensive compute, storage, and networking infrastructure by itself. The company's engineers are involved in planning new infrastructure, purchasing and managing the components, and troubleshooting the problems. For identifying and resolving the most common problems, PFN introduced rule-based automation, which tries to resolve minor infrastructure issues automatically or notify an administrator about more serious issues: "Automation is the nature of our business; of course, on every corner, we use automation techniques." This approach is working well, and the company is satisfied with its current level of automation, not targeting yet more advanced, AI-based capabilities: "We are actually an AI company and would be expected to use AI for infrastructure management, but funny part, we cannot find anything that is complex in that to make use of AI in our infrastructure."

> *"We have a subsidiary … and they use plenty of computing power [in public cloud]. If we open the door of our own infrastructure to them, it will be good for our business because we have a fast accelerator, we can provide* ***more stable*** *and* ***predictable*** *GPU resources … we can provide a predictable facility to the subsidiary."*

The company expects to, at least, double the number of its on-premises server deployments in the next five years: "We have to expand our infrastructure to maintain the operational efficiency." However, decisions about infrastructure expansions need to be weighted to ensure proper utilization: "This is not the service infrastructure, this is R&D infrastructure, and the empty R&D infrastructure is just a waste of money." To assess the needs for expansion, the company constantly monitors dominant resource share to watch usages of resources by user role. However, business value has a superior impact on the decision making, and the profitability of a project defines further infrastructure investments into this project.

While cost efficiency and performance predictability are two major benefits of maintaining and expanding on-premises infrastructure, the company's goal is also to achieve greater level of infrastructure control: "In theory, computer infrastructure should be uniform and predictable, but in reality, we sometimes have a hard time controlling the behavior of our infrastructure when it doesn't return the expected output due to all kinds of complex factors — they could be the internal state, differences in settings, or some hard-to-solve failures. Automation of monitoring and information sharing with the vendors are the keys, and our vendors are helping us a lot in this regard. I wish we could uniformly manage the whole infrastructure regardless of vendors or components, but it's not the case, it's a long way to go, not only for us but for everyone."

CONCLUSION

Summarizing takeaways from the three interviews, the following are the common drivers for organizations to keep investing in on-premises IT infrastructure:

- **Control:** All three interviewed companies emphasized that they define specifications for their compute and, in some cases, storage systems to optimize the architecture and system performance for the needs of core business workloads — a task that requires highly configurable solutions.
- **Cost:** By optimizing system configurations, all three organizations also optimized their investments by bypassing investments in system functionality that they don't utilize or, in case with Intel, by disaggregating system components and investing at the component rather than the system level.
- **Predictability:** By maintaining control over system design and infrastructure management, all three companies introduced predictability of infrastructure performance enabling smoother IT operations and reducing disruptions.

While not all workloads benefit substantially from compute or storage system optimization, in primary research conducted by IDC in the past years, these benefits were also mentioned among the top benefits recognized by end users who operate their own dedicated IT environments built with general-purpose systems. Other factors included security and data privacy — a factor also confirmed by one of the interviews.

However, these benefits come at a cost of another element of on-premises operations — infrastructure management. As IT operations grow, finding the right skill set and providing training to IT staff become critical. The effectiveness of on-premises IT goes hand in hand with the company's ability to invest in more advanced forms of IT automation to increase employee productivity and reduce disruptions in operations. In fact, in the past two years of IDC's surveys of compute and storage professionals, IT automation was ranked a top characteristic driving decisions about purchase of infrastructure products, surpassing performance and reliability. In this demand for autonomous operations, end users need to consider what integrated functionality, including systems management, monitoring, and provisioning tools, vendors offer on their hardware platforms.

Looking at the evolution of IT departments within enterprise organizations in the past five years and into the future, there is a steady transformation of IT from a cost center supporting back-end operations and focused on day-to-day system management to an enabler of companies' growth and capturing new business opportunities. In this journey, organizations need to consider all the pros and cons of different approaches to IT infrastructure operations learning from the experiences of other organizations while adjusting to their business goals and circumstances. Business cases featured in

this white paper demonstrate how by partnering with a system vendor organizations can build their own IT infrastructure optimized to the needs of their core workloads and run them most efficiently from both functional and monetary perspectives, which couldn't have been achieved with more generalistic approaches, while maintaining control over IT operations by adopting automated operations and streamlining IT management processes.

MESSAGE FROM THE SPONSOR

About Supermicro

Supermicro is a global technology company, which designs and develops a range of highly configurable enterprise IT solutions, including components, hardware systems, and integrated systems management software. Supermicro's core portfolio includes a broad range of application-optimized server solutions for variety of target markets such as cloud computing, data center, enterprise, high-performance computing, AI, IoT, 5G, embedded and edge computing. Working closely with its enterprise customers to stay on top of end user demands and partnering with technology providers, Supermicro tuned its business model to enable fast integration of innovative technologies into its server, storage, and networking platforms.

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. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

Global Headquarters

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