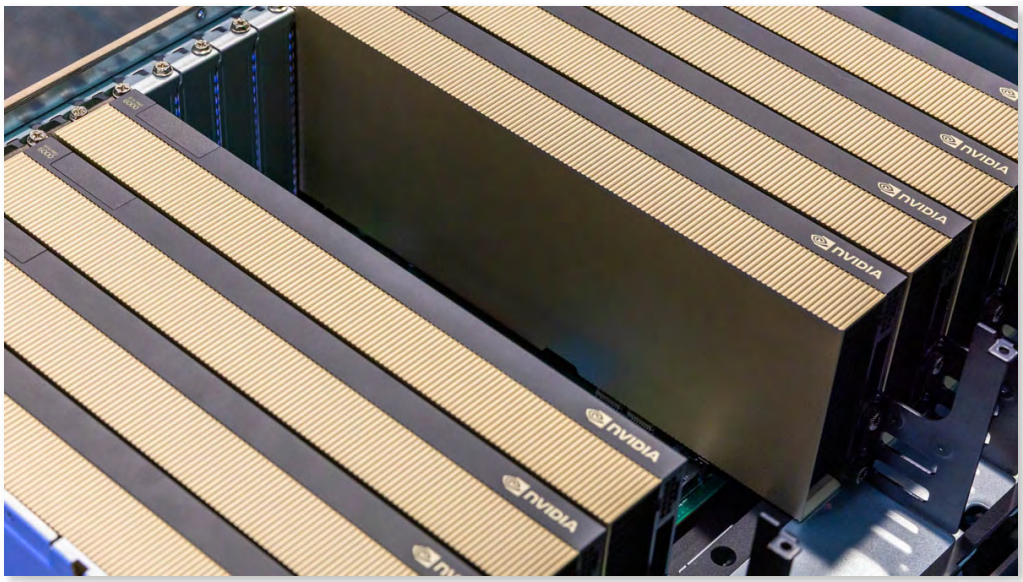
The 2025 PCIe GPU in Server Guide



Supermicro 2x PCIe GPU NVIDIA RTX PRO 6000 Blackwell Cover

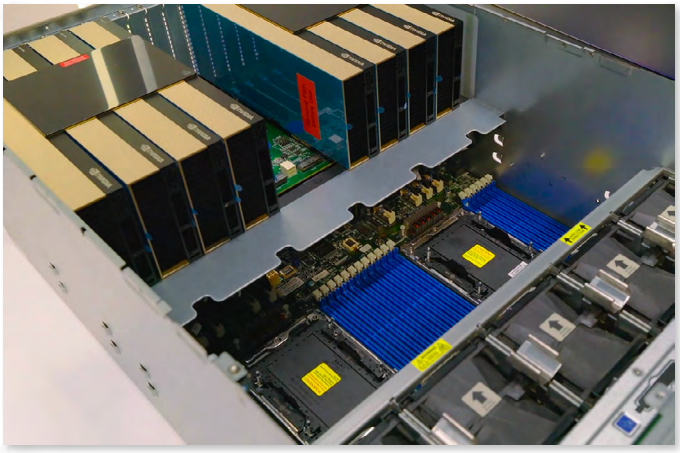
We often discuss the AI GPUs in the context of high-end integrated racks involving the NVIDIA GB200 NVL72 or HGX 8-GPU platforms that are used for AI factory applications. Still, there is a lot of AI that happens outside of these huge installations. Today, we are going to talk about those, both in the context of an "Enterprise AI Factory" but also down to edge servers and workstations. I have been wanting to do this piece for several months, and Supermicro managed to get all of the components together in one room so we can show folks the answers to simple questions like "which PCIe GPU should I use, where and when?"

Of course, for this one we have a video:

As always, we suggest watching it in its own tab or app for the best viewing experience. We also need to say thank you to Supermicro who got all of these systems together. We are going to say they are sponsoring this. With that, let us get to it.

8x PCIe GPU Systems

The 8x PCIe GPU systems are in some ways parallel to the SXM-based systems, with a few big caveats. PCIe GPU systems typically have GPUs between 300W and 600W per GPU, making them lower power than the SXM-based solutions. Beyond that, we typically see ratios of two East-West 400GbE NICs per PCIe GPU while in SXM-based systems, that is more of a 1:1 ratio. Also, removing the NVLink switch architecture means that the systems can be produced at a lower cost (albeit without that capability) and at lower power. While it may sound like these systems are simply lower power versions of the SXM-based platforms, that is not necessarily the case. There are additional options to customize the GPUs used in the platform and add additional graphics capabilities.



Supermicro SYS-522GA-NRT with 8x NVIDIA H200 NVL and NVLink Bridges

Typical PCIe GPUs we see used are:

- NVIDIA H100 NVL / H200 NVL
- NVIDIA RTX PRO 6000 Blackwell
- NVIDIA L40S

The NVIDIA H100 NVL and H200 NVL solutions are designed to expose NVLink interconnects across up to four GPUs. These are solutions typically used for re-training models and for AI inference, potentially at a lower power per GPU than the SXM systems. Perhaps the biggest reason to choose the H200 NVL over the H100 NVL is the newer HBM memory subsystem, which is improved for memory-bound workloads.

NVIDIA H100 L40S A100 Stack Top 1

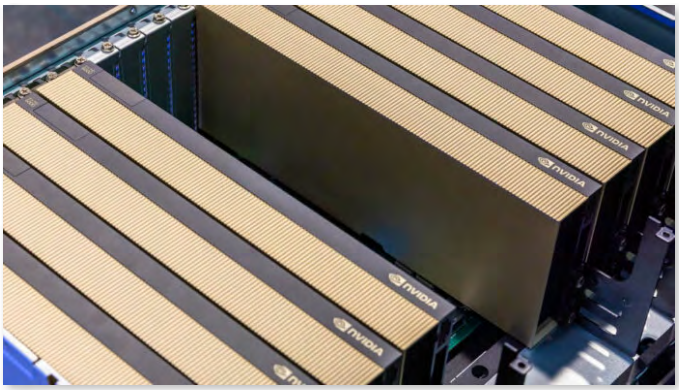
The NVIDIA RTX PRO 6000 Blackwell is used for something slightly different. This is NVIDIA's solution for those running a large mix of workloads. While these cards do not have the high bandwidth memory, they gain the RT cores, encoders, and even video outputs. That means the RTX PRO 6000 can be used for graphics workloads such as for engineering, VDI, rendering, and so forth. They can also be used for AI inference, with each card packing 96GB of GDDR7 memory per GPU. In an 8-GPU system, one can partition these GPUs into four instances using Multi-Instance GPU (MIG) for up to 32 logical GPUs. In total, eight of these GPUs offer 768GB of combined GPU memory for AI inference applications. In an eight GPU system, one can use the GPUs for different applications based on the time of day (e.g. VDI in the day, and AI inference in the evening.) One can also use the GPUs for different tasks during different times of the day. The key here is the flexibility on what you can do with these since each GPU has a lot of memory, but critically the NVIDIA RTX graphics capabilities that AI-focused GPUs do not have.

The NVIDIA L40S is effectively a lower-cost GPU for this platform based on the Ada Lovelace architecture. These GPUs have 48GB of memory and graphics pipelines, but do not have some of the newer features like MIG.

Supermicro has the SYS-522GA-NRT that is the company's 8x PCIe GPU platform. Inside the platform, we have two PCIe switches along with two CPUs, 32 DDR5 DIMMs, room for multiple NICs, and SSDs. Power consumption varies widely based on the configuration, but the advantage of these platforms is that they tend to use less power than the SXM systems leading to lower operating costs. Acquisition costs are often lower than the SXM-based systems as well.

Something new for 2025 is the new NVIDIA MGX PCIe Switch Board with ConnectX-8 for 8x PCIe GPU Servers. This is a big change to the platform that Supermicro is adopting in its Supermicro SYS-422GL-NR. Instead of using two or four larger switches, the new platform utilizes four NVIDIA ConnectX-8 NICs and their built-in switches to provide high-speed networking for the GPUs. This is, by far, the biggest change in the platform in many years.

Supermicro SYS 522GA NRT PCIe Switch Board



Supermicro SYS 522GA NRT With 8x NVIDIA RTX PRO 6000 Blackwell

Supermicro NVIDIA MGX PCIe Switch Board With NVIDIA ConnectX 8 SuperNICs For 8x PCIe GPU Servers Slots Large

Supermicro NVIDIA MGX PCIe Switch Board With NVIDIA ConnectX 8 SuperNICs For 8x PCIe GPU Servers NIC Ports Large

Standard Servers with PCIe GPUs

While the 8-GPU platform is designed primarily for GPU compute, the prospects for AI extend beyond those types of servers. A great example of why organizations are deploying these types of platforms is that if you believe that AI will be in virtually every workflow, then the question becomes how to address that. Deploying servers without GPUs today means that the only option is to go off of a server and to an AI server. An alternative model is to add GPUs into traditional servers so that they can be used during parts of the workload that are best accelerated.

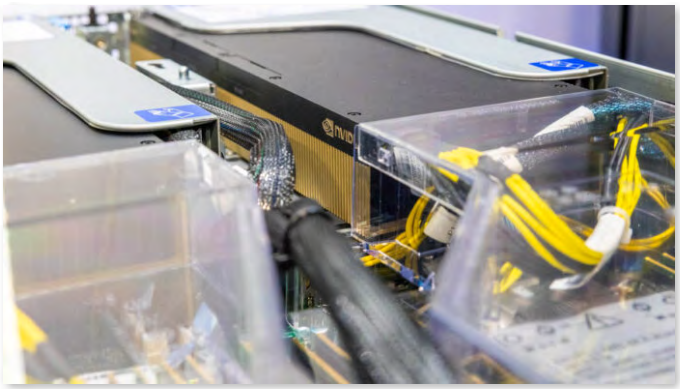
Like the 8-GPU systems, the GPUs used often range between the NVIDIA H100 NVL, H200 NVL, RTX PRO 6000 Blackwell and L40S. A big difference with this is that typically in a 2U server, one can only fit two GPUs side-by-side.

As a result, 4-way NVLink is less common in traditional servers as compared to finding one or two GPUs in each domain. Some are also deploying lower-power GPUs like the NVIDIA L4 to add a smaller amount of GPU compute and GPU memory, but at a lower power consumption and cost.
As an example, we showed the Supermicro SYS-212GB-NR. This is one of Supermicro's Hyper line of high-end servers where one can add many different types of GPUs.

Supermicro also has 2U GPU servers that are inspired by the NVIDIA MGX architecture. We have looked at a number of these previously, but we saw a new Xeon-based design to house multiple GPUs while we were in the demo room.

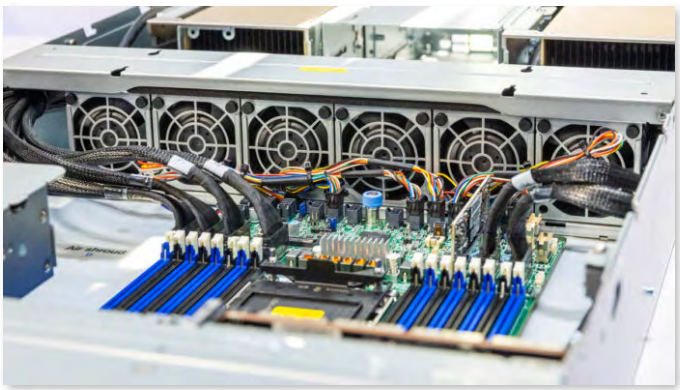
Supermicro SYS 212GB NR Front

Supermicro SYS 212GB NR Rear



Supermicro SYS 212GB NR Server With Airflow Routing To NVIDIA H100 NVL

Supermicro MGX Inspired 2U GPU Server Front



Supermicro MGX Inspired 2U GPU Server CPU Memory Fans And GPUs

High-Density Servers with PCIe GPUs

In the video, we also showed high-density servers that can utilize PCIe GPUs. An example we showed was the Supermicro SuperBlade with NVIDIA L4 GPUs. The L4 is nice because it is a low-profile GPU with minimal cooling needs.

Over the years, we have seen SuperBlade and other high-density Supermicro platforms take a range of GPUs from single-width low-profile GPUs to double-width GPUs. The reasoning is usually the same as with the Standard Servers, but just in a higher-density design. Both 6U and 8U SuperBlade servers will support a range of NVIDIA GPUs but that support will vary based on the SKU.

Supermicro SuperBlade with NVIDIA L4

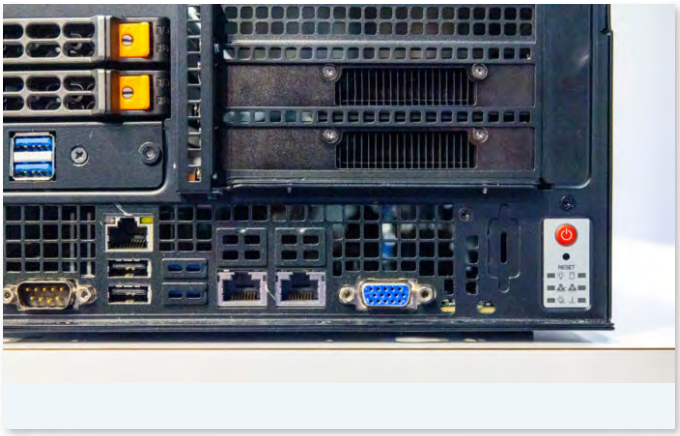
Edge Servers with PCIe GPUs

Edge servers present a different opportunity. Applications such as computer vision become more prevalent at the edge. A great example of this is in many retail locations where self-checkout is powered by edge servers with GPUs. Other typical applications in retail can be inventory analytics, shopper analytics, and more.

In the video, we showed an example with a NVIDIA L4 GPU in the Supermicro SYS-E403-14BFRN2T. These servers are often highly power and space constrained so having a 75W TDP or lower single-width low-profile GPU is the go-to. Beyond the L4 GPUs, there are other edge use cases from network infrastructure to smart cities where larger GPUs can be utilized, often with higher-end networking.

Supermicro SYS E403 14B FRN2T Front

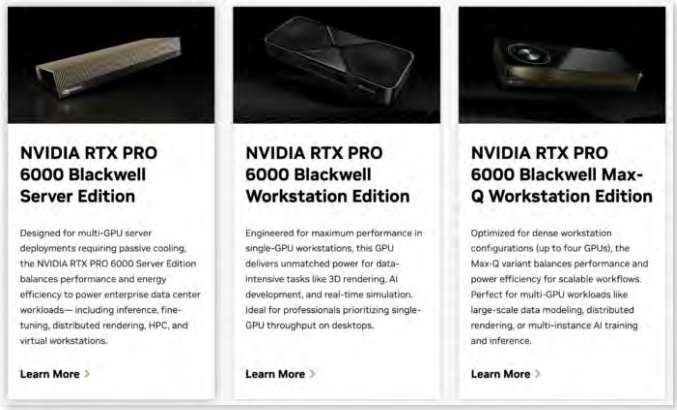
Supermicro SYS E403 14B FRN2T With Two NVIDIA L4 GPUs In Risers



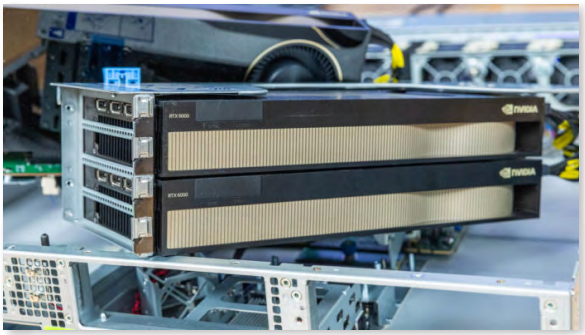
Supermicro SYS E403 14B FRN2T Front PCIe Risers

Workstations with PCIe GPUs

Workstations have become hot topics in the AI era. Folks want to develop AI tools locally. Perhaps the bigger shift will be that as AI becomes a bigger part of everyday workflows, having bigger GPUs can yield outsized productivity gains. When NVIDIA launched the RTX 6000 PRO Blackwell, there were three versions. One was a 600W version that was designed to provide the maximum performance in a single PCIe card slot. There were then two double-width cards. One is a 300W actively cooled version. The other is the passively cooled version that we often see in the 8-GPU systems.
Recently, we reviewed the Supermicro AS-2115HV-TNRT, a 2U server that can handle up to four double-width GPUs. The innovation here is that most other workstations on the market, even if they can be converted to 4U or 5U workstations, only handle up to three GPUs. With this system, we can get up to four GPUs in a system, along with IPMI for management, and then pack them into data center racks. Supermicro also has other options such as the AS-531AW-TC and SYS-532AW-C that are designed to handle either a single 600W RTX PRO 6000 Blackwell Workstation GPU or multiple 300W versions like the Max-Q edition.



NVIDIA RTX PRO 6000 Blackwell Series Versions


Supermicro AS 2115HV TNRT 2x NVIDIA RTX 6000 Ada In Riser 1


Supermicro SYS 532AW C With NVIDIA RTX PRO 6000 Blackwell

Final Words

Ultimately, if you believe in AI and are using new tools daily, then the concept that AI is going to be part of most workflows going forward will seem very familiar. We managed to show a number of GPUs and their use cases, and how to deploy them. AI is not just going to happen in enormous AI factories. Latency needs, workflow needs, data security requirements, and even deployment preferences are going to push GPUs into most servers going forward.

We spend a lot of time focusing on the large AI cluster deployments, but the writing is on the wall. As we transition to an era of AI, GPUs are coming to many other locations and server form factors. It felt like it was time to show folks a sample of different options in different categories. Over time, there will be new GPUs, new networking, and new architectures, but hopefully, this helps frame some of the common use cases and deployments for STH readers today.