NVIDIA MGX: Data Center Architecture for Accelerated Computing

Modular Hardware Built for Evolving Technology

White Paper

Table of Contents

Modular Design for Easier Data Center Technology Configuration

NVIDIA MGX is an open, multi-generational accelerated computing reference architecture built to allow rapid time-to-market adoption of key platform technologies including CPUs, GPUs, and DPUs with minimal non-recurring engineering investment. NVIDIA MGX server specification enables over 100 different system configurations.

The world's $1 trillion of data centers are populated primarily by unaccelerated CPU systems with basic networking infrastructure. With traditional compute architectures only improving by approximately 10% every year, and data centers being inherently power-limited, accelerated computing is the path forward. Over the coming years, the world's data centers will need to be re-architected to enable accelerated computing and generative AI. CPUs alone are no longer sufficient to meet the demands of today's data centers, due to overall growth in traditional compute needs, as well as new applications such as AI and digital twins which are expected to drive demand further.

These use cases require hardware acceleration via GPUs to be performant, and the amount of data that needs to be transferred for these applications necessitates high speed networking devices like DPUs, and closely optimized ratios of connectivity, CPU resources, I/O and disk, and more to deliver the best workload performance and total cost of ownership (TCO).

Many systems today limit form factor sizes and GPUs and other technologies, particularly in the context of future requirements around GPU size, network interface card (NIC) size, and airflow.

System Flexibility: Any Data Center Environment

Beyond the core accelerator, CPU, storage, and networking choices within the node, hyperscale, edge, high-performance computing (HPC), and traditional data centers have rapidly evolving and diverging needs, even between differing regions and hyperscale technology providers, a few examples of which are listed below.

The optimal compute infrastructure for each workload places varying demands on server design. For example, platforms that deliver the best performance for real-time visualization or cloud gaming may look substantially different than those for large language model inference or training.

Standard CPU servers are not designed for the unique thermal, power delivery, and mechanical requirements of accelerated computing.

Thermal

The cooling capabilities of the NVIDIA MGX can support higher GPU thermals through a suite of solutions not available in standard CPU servers including:

> Support for current and future GPU/ DPU/CPU
> Support air cooling and liquid cooling in a single architecture
> Special design to support higher TDP GPU (ex 400W+ GPU TDP)

Mechanical

The NVIDIA MGX solves for the mechanical issue that current CPU designs often find problematic as they are less flexible for PCIe layouts and may not support 12.3" GPUs. The NVIDIA MGX is flexible enough to support both through a wide range of solutions including:

> By re-allocating the modules bays, the NVIDIA MDX architecture can support more than 100 configurations
> Reduce the product development resource and time for partners to bring new products to market
> Support the future GPU form factors like 12.3" GPU
> Support both cold and hot aisle cabling to accommodate enterprise (hot aisle) and CSP (cold aisle) data center design

Power

As multiple power delivery designs are required to meet different system requirements, the NVIDIA MGX provides:

> Single architecture to support different power delivery to enable both enterprise (EIA) and CSP (OCP)
> Supports both power busbar and power supply

This paper introduces a family of mechanical chassis, Host Processor Module (HPM), and bay components used to build a MGX compliant server for any application.

Figure 1. The NVIDIA MGX chassis enables over 100 configurations

NVIDIA MGX Overview

The MGX architecture is modular in nature. It consists of configurable bays on both aisles, hot and cold, which can be populated with modules to achieve a desired configuration. Internally, the system designates zones for major subsystems: HPM, power distribution boards, cooling solutions, and module bays.

One example of a two rack unit (2RU) system layout portraying the bay design is shown in Figure 2. When the MGX architecture is followed, it can support multiple server applications, such as hyperscale, edge, HPC, and cloud. NVIDIA MGX supports both Arm and x86 CPUs and is designed to be multi-generational, so that once you make an investment, the next generation GPU, CPUs,and DPUs will be easily configured to give the best TTM and longevity from the investment.

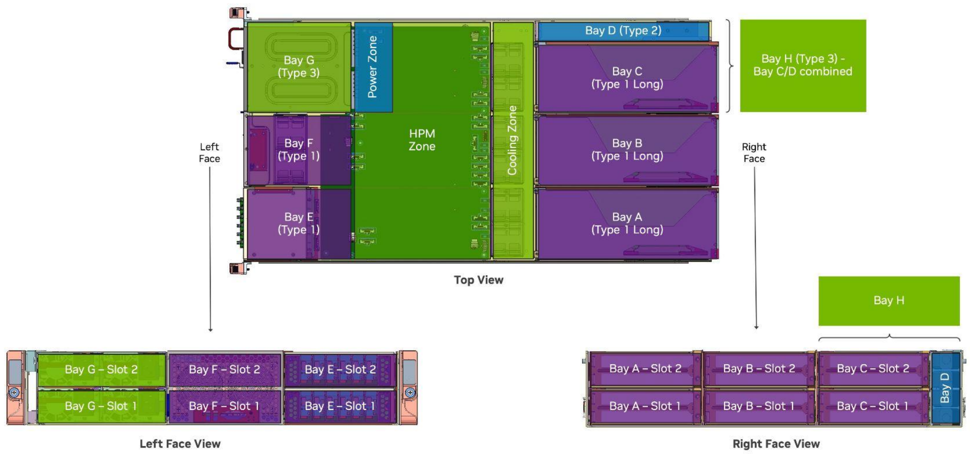

Figure 2. System with Modular Bays That Can Be Configured Differently Per Use Case Example

The NVIDIA MGX specification defines key design parameters, allowing reuse of the subsystems in a wide range of applications and across generations. Examples of design attributes covered are mechanical form factors, mechanical attachment guidance, feature placement, general thermal guidelines, electrical requirements, and interconnect definitions.

Accelerated computing requires hardware acceleration via GPUs to be performant, and the amount of data that needs to be transferred for these applications necessitates high speed networking devices like DPUs. Many systems today limit growth of accelerators and other technologies, particularly in the context of future evolutions of GPUs, DPUs, and cooling. For example, most current servers today can only accommodate 10.5” PCIe Add-in Cards (AIC), whereas MGX can also support 12.3” PCIe AICs.

In addition, the balance between these accelerators and the CPU, I/O, and storage needs to be properly tuned for each workload to deliver the best workload performance and minimize the total cost of ownership. NVIDIA MGX’s modular nature allows for easy tuning and optimization based on specific applications.

At the data center level, hyperscale, edge, HPC, and other markets have rapidly evolving and diverging needs beyond the CPU, storage, and networking choices — a few examples of which are listed below.

Rack Form Factors

> EIA racks use 19” wide chassis
> OCP racks use 21” wide chassis
> Edge and telecommunications racks have additional chassis requirements

Power Delivery

> AC-DC PSUs
> 54V DC input
> 12V DC input

Cabling and Cooling

> Hot aisle cabling vs cold aisle cabling
> Air vs liquid cooling
> Altitude and maximum operating temperature

Building Blocks Are Key For Hardware Composability

NVIDIA MGX aims to standardize the various hardware building blocks using a modular approach and to create an ecosystem for ODMs and OEMs to create accelerated computing servers.

To maximize flexibility, a proper set of basic building blocks are required. The intent of the building blocks is to reserve a volumetric and logical space within the system’s floorplan. These blocks allow for forward compatibility with evolving technologies in the domain of accelerated computing, such as processors, memory, storage, networking, power delivery, and thermal solutions.

Host Processor Module

The HPM consists of the fundamental functional blocks listed below:

> CPU(s) + system memory
> Removable BMC module
> On-board storage devices

- High-speed interconnect for cabled peripheral devices
- Various interconnect for power and system controls

The MGX specification currently has two HPM form factors: micro-MGX and MGX. A common board pan supports both form factors at various locations.

Micro-MGX Host Processor Module

The micro-MGX HPM outline and mounting holes are compliant to the OCP DC-MHS DNO Type 2 specification. This form factor may fit two boards side-by-side within a single EIA 19" wide chassis. The following figure shows the approximate locations of major parts of the micro-MGX HPM.

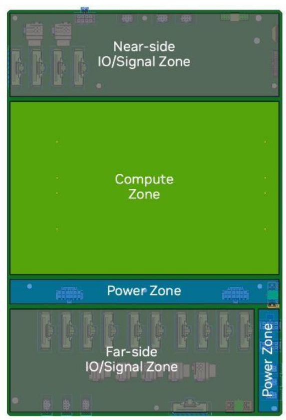

Figure 3. Micro-MGX Host Processor Module Overview

MGX Host Processing Module

The MGX HPM outline takes advantage of the entire HPM zone in the 19" wide chassis. The following figure shows the approximate locations of major parts of the MGX HPM.

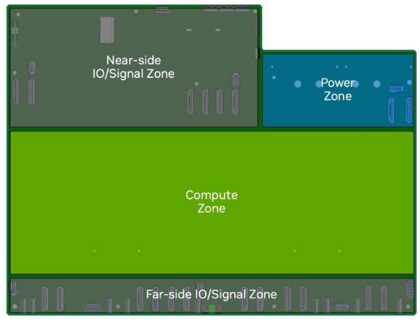

Figure 4. NVIDIA MGX Host Processor Module Overview

Modular Bays

The modular bay is an essential enabler for this architecture. The bay modules are sized to support current technology form factors that are common across the industry with intent to enable future form factors as they emerge and become available.

There are several sizes of bay modules to cover different chassis lengths, chassis heights, and desired system technology layout configurations. Current bay definitions and sizes are shown in Table 1, Table 2, Table 3, and Figure 5.

Type	Description
Type 1	Supports full-height PCIe AIC with standard faceplate.
Type 2	Supports EDSFF storage or power input connectors.
Type 3	Spans the width of Type 1, Type 2 and the divider. Supports two CRPS PSUs side by side with room for pass-through power cable.

Table 1. Bay Module Width Classifications

Length	Description
Short	Supports half length PCIe AICs, E1.S, E2.S and U.2 drives.
Long	Supports a full length (up to 12.3”) PCIe AIC

Table 2. Bay Module Lengths

Height	Description
1RU	Fits within a 1RU system height.
2RU	Fits within a 2RU system height.
4RU	Fits within a 4RU system height.

Table 3. Bay Module Heights

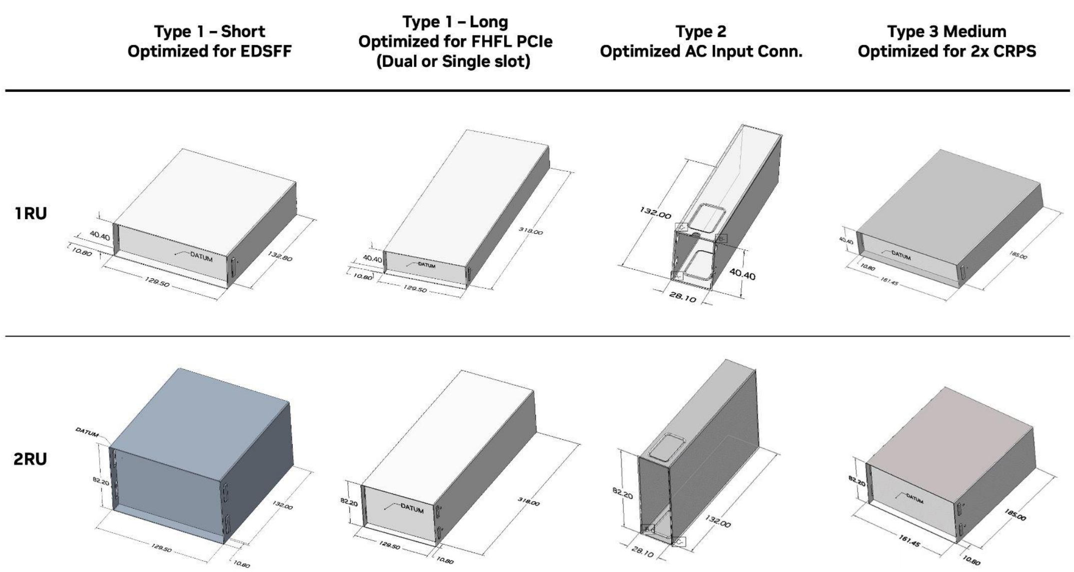

Figure 5. Examples of Modular Bay Types

Figure 5 shows some examples of some applications of the various bay types.

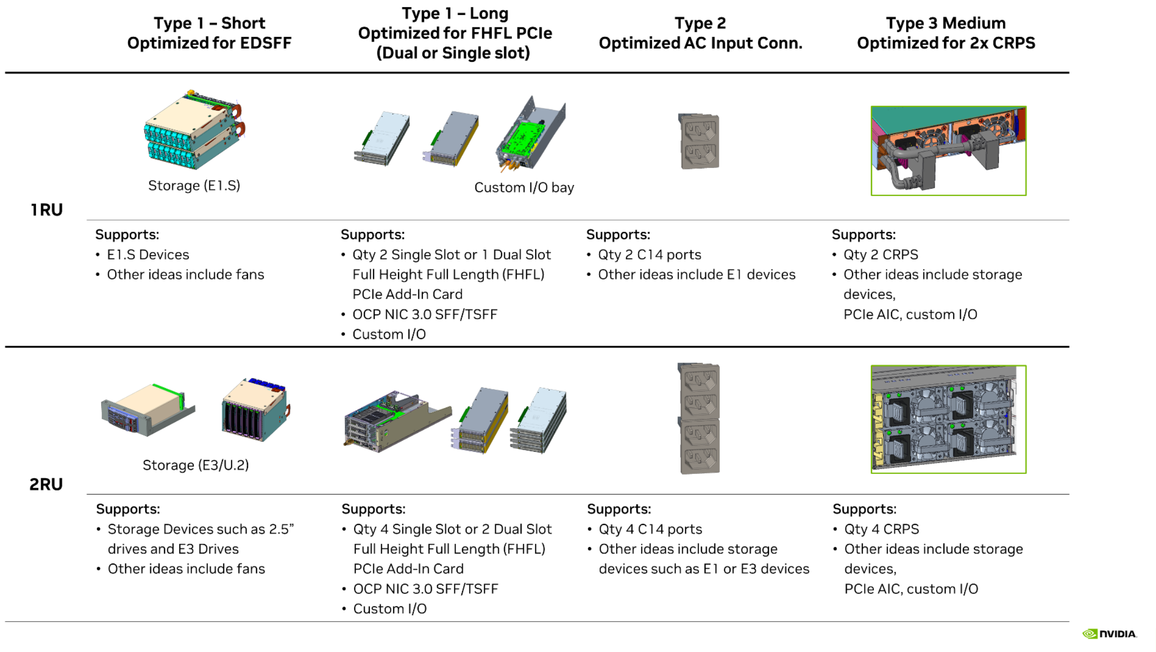

Figure 6. Example of Bay Modules Hardware

The chassis has standard locking and shelving mechanisms that are compatible with all bay module types. For example, a 2RU bay module may be used interchangeably with two 1RU bay modules. Similarly, a Type 3 bay may be used in place of Type 1 bay and a Type 2 bay.

The 1RU bay's height is optimized to fit within a 1RU system. In systems taller than 1RU, multiple 1RU bays may be stacked. For example, two 1RU bay modules will fit within the 2RU system chassis and cage, see Figure 7.

The 2RU bay is a height optimized to fit within a 2RU system. It allows for devices to freely utilize the space that would otherwise be partitioned into two sections constrained by the 1RU bays. For example, a half-height PCIe card or E3.S can be placed vertically in a 2RU bay. A 2RU bay can take the place of any two adjacent 1RU bays.

Bay Module to Chassis Interface

Each bay module is secured to the chassis during configuration assembly using one of two methodologies. A single spring latch and catch, internal to the system, or a camming latch handle, external to the system, retains the bay in place. Each bay in the chassis may be designed to support a short bay, long bay, or both by allowing for spring or camming latch positions.

Interconnect

To enable interoperability across different HPM types and bays, the NVIDIA MGX specification standardizes common elements and interfaces by defining the following:

> Input Power and Control Signals
> Cabled Peripheral Device I/O
> BMC Connectivity
> Service and Control Panel I/O
> Far-Side I/O Board Connectivity
> Fan connectivity to HPM

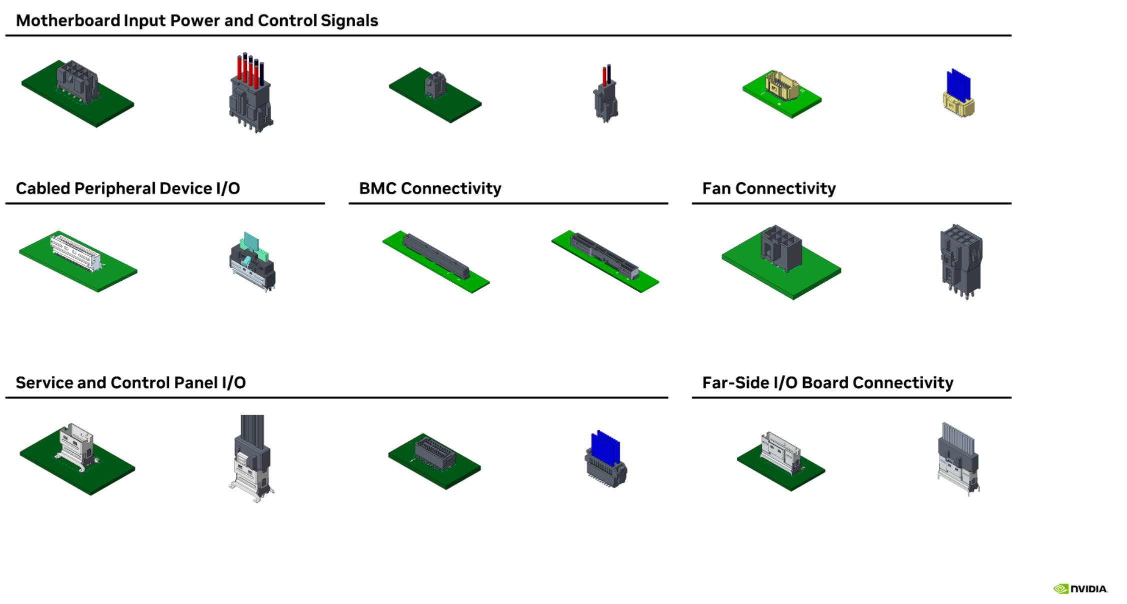

Figure 7. Examples of various interconnect on HPM

The defined interconnects leverage industry standard and other commonly available connectors. Standardized interface references include SFF-8654, SFF-TA-1002, SFF-TA-1016, and SFF-TA-1020. Where possible, existing electrical interface standards were leveraged. New electrical interface definitions were introduced to ensure interoperability.

NVIDIA MGX Architecture Ecosystem

The MGX architecture produces an ecosystem of basic building blocks, which consist of modular bays, HPMs, PDBs, cooling solutions, and tailored to fit in a 19" chassis. These building blocks can be assembled in over 100 ways to achieve unique system designs tailored to meet the needs of the server as shown in Figure 8.

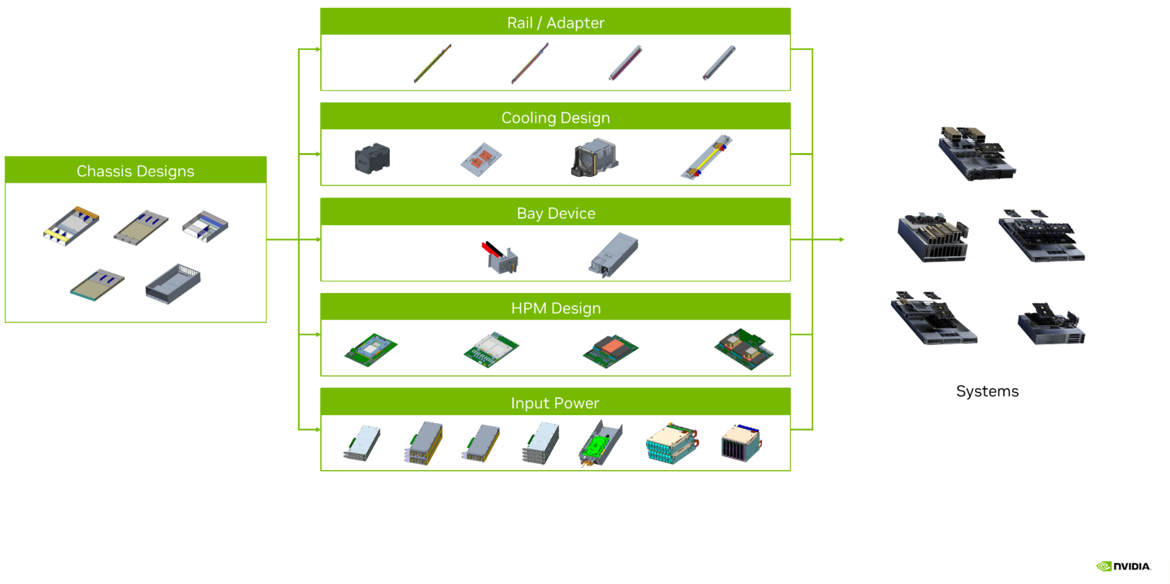

Figure 8. Modular building block architecture and example system configurations

MGX Partner Benefits

NVIDIA MGX offers significant benefits to original design manufacturers (ODMs) and original equipment manufacturers (OEMs) who are designing and building accelerated computing systems. ODMs and OEMs who adopt the NVIDIA MGX architecture can benefit from reduced design and research and development investment costs and faster time to market.

> Support for both x86 and Arm processors and the full lineup of NVIDIA GPUs, CPUs, DPUs, allowing for hundreds of different configurations for different markets.
> Support for 1U, 2U, and 4U chassis designs with many storage and power options.
> Forward compatibility for the next generations of GPUs, CPUs, and DPUs, ensuring investment longevity.
> Faster time to market for major NVIDIA product launches.

> MGX partners receive a full set of design files including: electrical schematics, mechanical drawings, and thermal simulations.
> NVIDIA Technical support on MGX related design issues.

MGX differs from NVIDIA HGX™ in that it offers flexible, multi-generational compatibility with NVIDIA products to ensure that system builders can reuse existing designs and easily adopt next-generation products without expensive redesigns. In contrast, HGX is based on an NVLink®-connected, multi-GPU baseboard tailored to scale to create the ultimate in AI and HPC systems.

Example System Designs

There are over 100 different core MGX configurations, and numerous customizations including storage, CPU, GPU, and DPU– yielding almost limitless flexibility.

A few selected example systems are shown, with potential use cases below.

Though optimized for specific workloads, accelerated data center servers are extremely flexible by nature, and it would be expected that there is some overlap in addressable scenarios between some of these designs.

For instance, an NVIDIA Omniverse OVX™ server, though optimized for its core use case, might be deployed for midrange inference scenarios in a number of circumstances.

This flexibility is core to NVIDIA MGX's value and allows operators and system builders to economically right-size trade-offs between scenario-specific optimizations and fleet or design fungibility.

Omniverse OVX Server

The Omniverse OVX server is a data center remote visualization platform suitable for building and deploying Omniverse applications, and other demanding visualization applications where traditional x86 CPUs may be desired for compatibility.

> 2U Dual CPU x86 server
> Four NVIDIA L40 GPUs, BlueField-3 DPU®, and two ConnectX-7s in hot aisle
> Up to sixteen E1.S drives in cold aisle

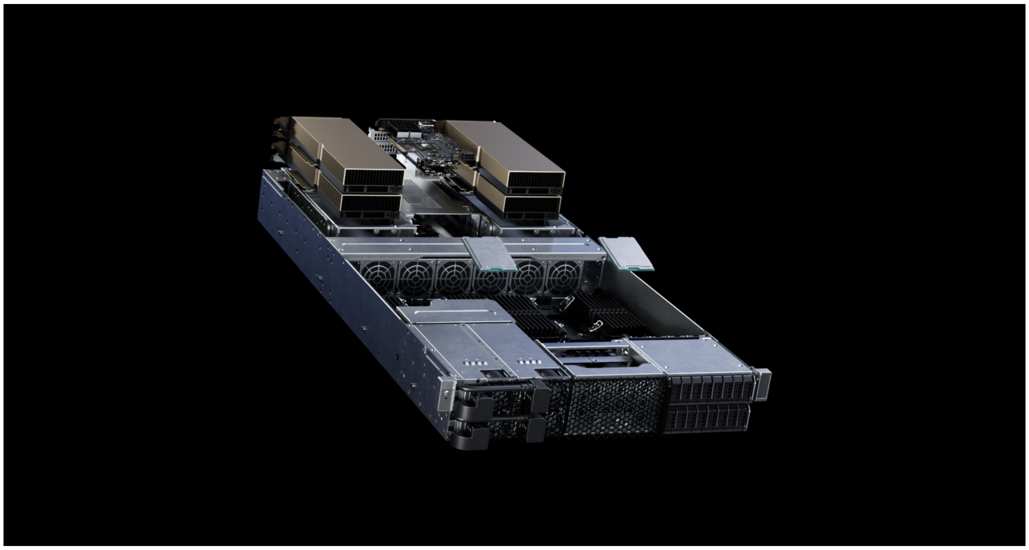

Figure 9. NVIDIA Omniverse OVX Server

Grace Omniverse OVX Server

The Grace™ Omniverse OVX Server extends the Omniverse OVX server with the new NVIDIA Grace CPU Superchip to deliver an all-NVIDIA scalable visualization platform powered by high-performance energy-efficient processors suitable for demanding use cases in dense deployments.

- 2U Grace CPU Superchip server
- Four NVIDIA L40 GPUs, a BlueField-3 DPU, and a ConnectX-7 on the hot aisle
- Up to sixteen E1.S drives on the cold aisle

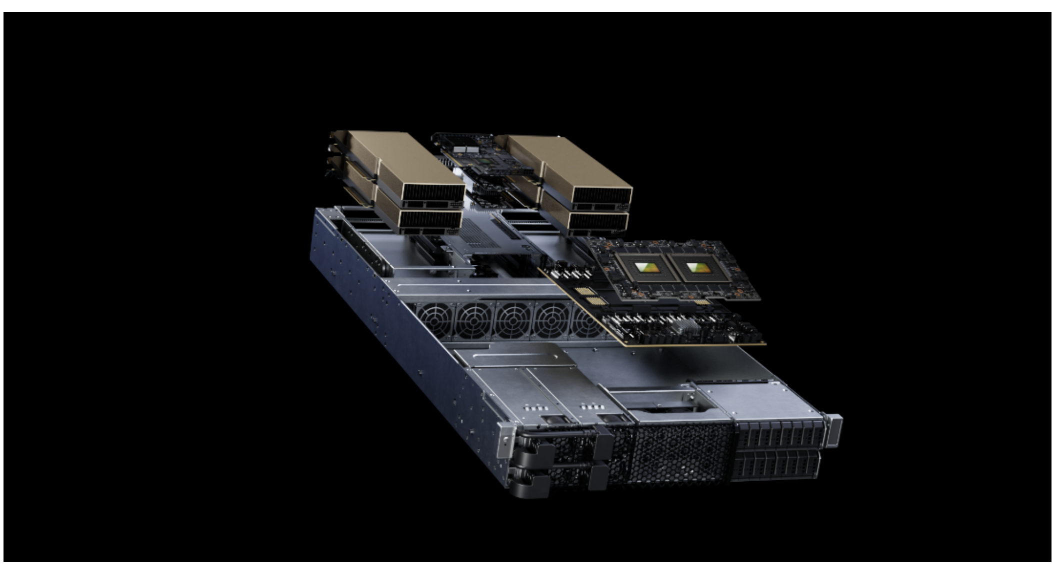

Figure 10. Grace CPU Omniverse OVX Server

Grace CPU Cloud Gaming Server

Remote gaming's demanding latency goals mean operators must maximize in-demand data center locations close to user hotspots. The Grace Cloud Gaming Server optimizes the number of gaming sessions that can be handled in a given data center location by combining ultra low-power high-performance Grace CPUs with ten power-efficient L4 GPUs to dedicate the right amount of resources to each session without compromising on energy efficiency.

> 2U Grace CPU Superchip server
> BlueField-3 DPU, a ConnectX-7 on the hot aisle
> Ten NVIDIA L4 Tensor Core GPUs and up to eight U.2 drives on the cold aisle

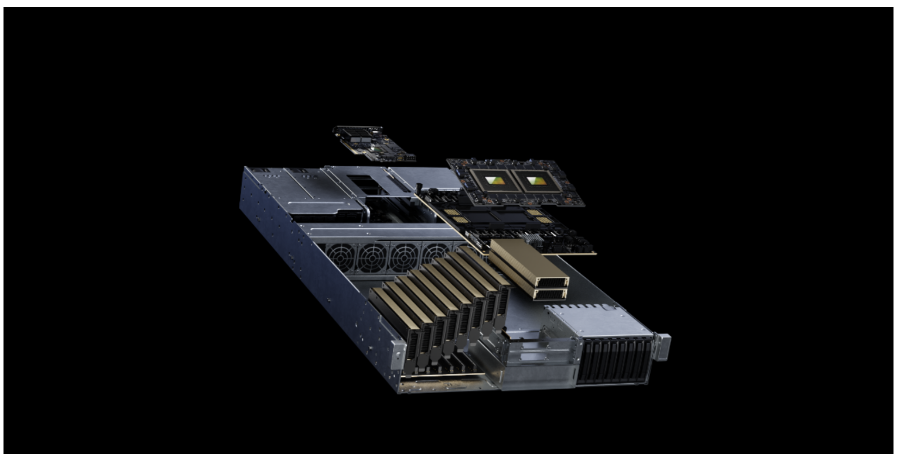

Figure 11. Grace Cloud Gaming Server

H100 NVL Inference Server

H100 NVL Inference Servers provide optimum inference TCO for large, demanding generative models including LLMs – and high-throughput batch inference use cases with tight latency requirements where NVLINK attached memory and GPU scale-up are key to unlocking performance, but the scale-out and MaxP TDP configurations offered by HGX products may not be required.

> 4U x86 server with eight NVIDIA H100 NVL GPUs

> Two BlueField-3 DPUs on the cold aisle
> Hot aisle bays unpopulated

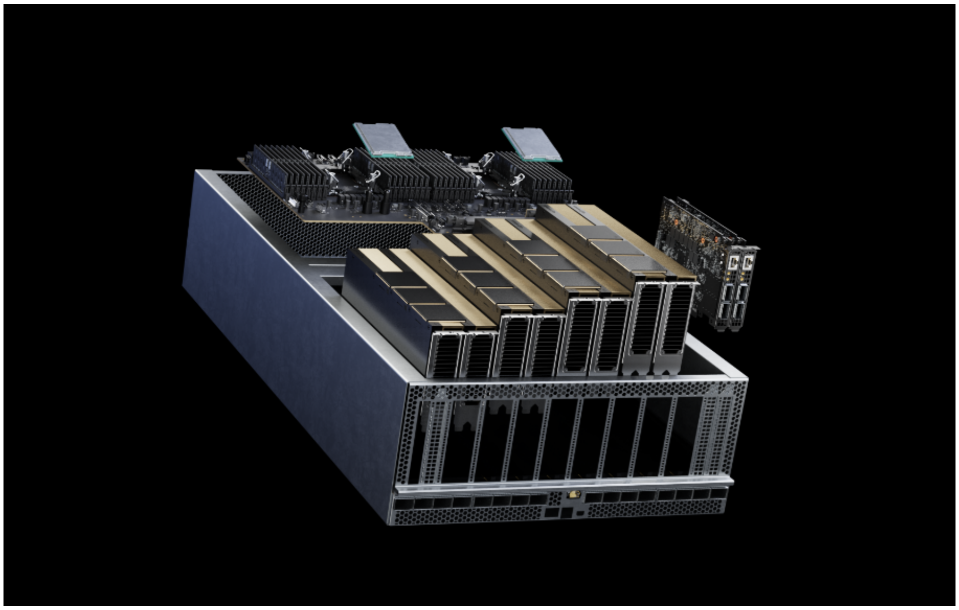

Figure 12. H100 NVL Inference Server

Grace-Hopper Aerial Server

The Grace-Hopper Aerial server provides an ideal converged 5G network and AI server in a form factor suitable for large, centralized telecommunications data centers.

> 1U Grace-Hopper Superchip server with two BlueField-3 DPUs and a NVIDIA L4 Tensor Core GPU on the hot aisle
> Additional fans on the cold aisle

Figure 13. Grace-Hopper Aerial Server

Grace-Hopper Aerial Server Short Depth

The Grace-Hopper Aerial Short-Depth Server allows moving 5G network processing and AI applications closer to end-users, radio, and backhaul infrastructure by accommodating the latest high-performance NVIDIA technologies in a edge-friendly platform that can meet traditional telephony rack depth requirements. Short-depth platforms can be deployed in a wide array of existing facilities.

> 1U Grace-Hopper Superchip server in a short chassis
> BlueField-3 DPUs and a ConnectX-7 on the cold aisle
> Fans on the hot aisle

Figure 14. Grace-Hopper Aerial Server Short Depth

High-Density General-Purpose Grace CPU Server

The High-Density General-Purpose Grace CPU Server enables groundbreaking compute densities and energy efficiency in hyperscale and edge data centers for traditional CPU workloads by integrating two Grace CPU nodes in a traditional 1U air-cooled form factor, improving overall data center efficiency.

Each Grace CPU node provides similar performance to a flagship 2-socket x86 system, while consuming close to half the power – allowing the high-density Grace CPU server to incorporate 2X the node count and 2X the performance of a traditional 2-socket 1U system within the same rack space and power requirements.

> 1U server with two Grace CPU Superchips
> Two BlueField-3 DPUs on the hot aisle
> Additional fans on the cold aisle

Figure 15. High-Density General-Purpose Grace CPU Server

Data Center Flexibility for AI and Accelerated Computing

The NVIDIA MGX architecture stands at the forefront of a transformative era for data centers worldwide.

With its open and multi-generational design, it offers a seamless transition towards accelerated computing, empowering data centers to unlock their full potential. In a landscape dominated by unaccelerated CPU systems, the need for change is undeniable. With traditional compute architectures experiencing incremental improvements of merely 10% annually, the limitations become apparent.

Now, NVIDIA MGX provides a solution that combines the power of CPUs, GPUs, and DPUs in over 100 different system configurations, tailored to meet the evolving needs of data centers. By leveraging hardware acceleration and high-speed networking, this architecture paves the way for unprecedented workload performance and superior total cost of ownership. Data centers can enjoy enhanced modular flexibility, unencumbered by the constraints of form factors and technology limitations. It is a bold step towards a new era of data center excellence.

Visit the NVIDIA MGX homepage to learn more about the features and capabilities that will help you take your work to the next level.

NVIDIA Corporation | 2788 San Tomas Expressway, Santa Clara, CA 95051
http://www.nvidia.com

.