

Ready for Deployment: A Deep Dive into the New Architecture of NVIDIA DGX B300 Systems

Technical Brief

Table of Contents

Table of Contents 2
List of Figures 3
NVIDIA DGX B300 Architecture 4
- Introduction 4
- Redesigned DGX Chassis 4
- Blackwell Ultra GPU Tray 6
- New Power Supply Options 7
  - Traditional Racks with AC Power Distribution Units 7
  - High-Density Racks with DC Busbar and Power Shelves 8
- Power Connectivity Differences 9
  - Integrated PSUs 9
  - New Busbar Design 10
- Inside the NVIDIA DGX B300 system 12
- Conclusion 14

List of Figures

Figure 1: NVIDIA DGX B300 system, front view
Figure 2: NVIDIA GPU tray
Figure 3: PSU for DGX B300 system AC version, front view
Figure 4: DGX B300 system DC version, front view
Figure 5: Busbar connectivity and power shelves for DGX B300 systems, DC version
Figure 6: Rear of DGX B300 system, AC version
Figure 7: Rear of DGX B300 system, DC version
Figure 8: DGX B300 system motherboard, front view
Figure 9: DGX B300 system motherboard, top view
Figure 10: NVIDIA ConnectX-8 network interfaces in the lower section of the GPU tray

NVIDIA DGX B300 Architecture

Introduction

AI adoption is rapidly growing, and organizations are embracing new business opportunities that require increased scale of AI infrastructure while continuing to manage costs and expertise gaps.

NVIDIA DGX B300, introduced at GTC 2025 in San Jose, CA, is a re-engineered infrastructure solution for enterprise AI. It addresses scaling demands, energy consumption, and complexity for next-generation AI models. Powered by the NVIDIA Blackwell Ultra architecture, DGX B300 features eight NVIDIA Blackwell Ultra GPUs in an air-cooled 10 Rack Unit (RU) chassis, offering breakthrough performance for training and inference. Its architecture includes 5th generation NVLink fabric for 14.4 TB/s bandwidth and PCIe Gen6 for direct GPU-to-network interface connectivity, ensuring maximum data throughput and minimizing I/O bottlenecks in scaled-out clusters.

DGX B300 offers flexible deployment with two power options: traditional AC power via integrated PSUs or efficient DC power via external busbars in NVIDIA MGX-compliant racks. This technical brief details the architectural innovations and strategic capabilities that make DGX B300 a significant advancement for high-performance AI workloads.

Redesigned DGX Chassis

The evolution from NVIDIA DGX B200 to NVIDIA DGX B300 incorporates a new design to accommodate organizations looking at denser deployments suited for different types of data center designs. At first glance, the changes are immediately apparent because of the change in color scheme and component layout. Let's walk through the changes starting from the outside and make our way inside the redesigned chassis.



Figure 1: NVIDIA DGX B300 system, front view

Looking at a single node from the front, the DGX B300 system still has the gold foam bezel that has become a hallmark of the NVIDIA DGX Platform. In contrast to previous generations, however, the DGX B300 bezel does not cover the entire front of the system, which allows the required network cabling to now connect directly to the node from the front. In this lower section of the system, we can also identify the storage drives used for local data caching, the IO ports including MiniDV and USB (for easy access in the event of a crash cart being connected for management), and the boot drives.

NVIDIA Blackwell Ultra GPU Tray

Now that we've covered the outer front-facing section of the system, we can remove the bezel and reveal the NVIDIA GPU tray. The GPU tray is equipped with eight 800Gbps NVIDIA ConnectX-8 OSFP network ports, marking a significant architectural change from previous versions as these were previously located in the motherboard tray. This development increases the bandwidth available from the DGX systems' eight NVIDIA Blackwell Ultra GPUs to the compute network, allowing faster RDMA transactions across all the systems in a DGX cluster.

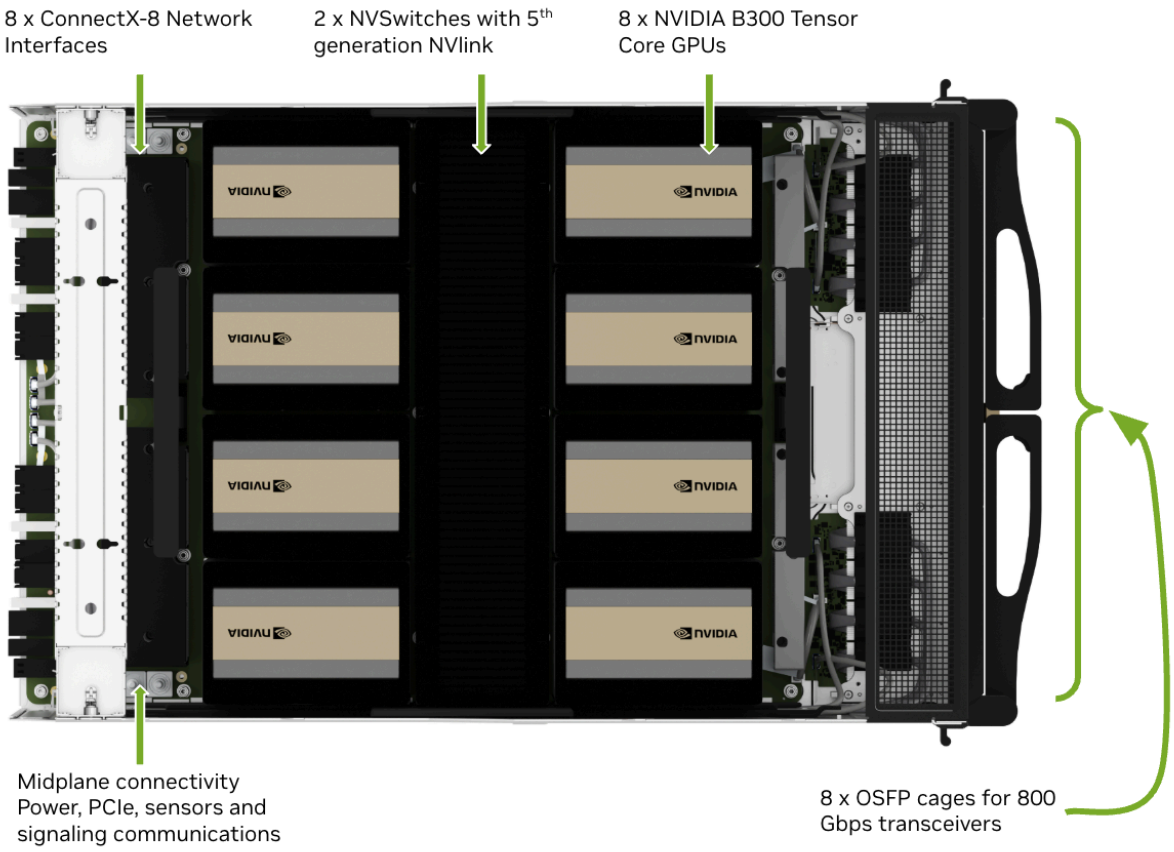


Figure 2: NVIDIA GPU tray

New Power Supply Options

In addition to these improvements to GPU computing, networking and cabling, customers now have the option to choose between integrated power supply units (PSUs) or an external busbar to power their DGX B300 systems.

Traditional Racks with AC Power Distribution Units

When planning for these systems, one of the first choices to make is whether the systems should come ready to connect to an AC or DC power source. If deploying in a traditional data center with AC power provisioned to the rack, users should look to systems with integrated PSUs that can convert 200-240 AC to DC in each system.

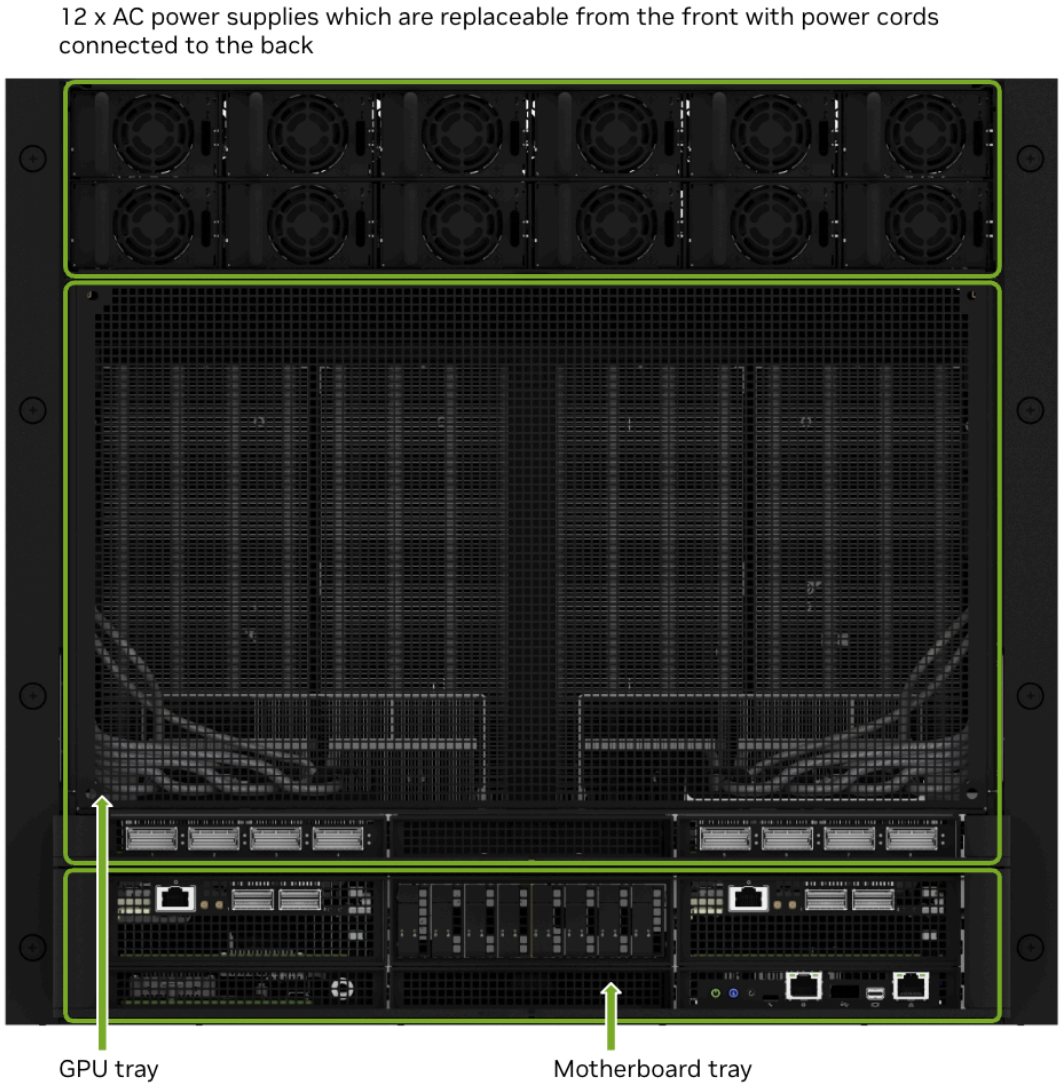


Figure 3: PSU for DGX B300 system AC version, front view

High-Density Racks with DC Busbar and Power Shelves

For customers who are considering higher densities of compute, and are capable of delivering 400-480V AC to the rack, then the newly-designed racks that incorporate power shelves and busbars are the best choice. These new NVIDIA MGX-compliant racks are designed to provide flexibility in configuration like including a busbar for DC-powered systems. They are designed to withstand the weight of multiple systems, and using power shelves, they can deliver 54V of DC power to each DGX B300. With this new option, customers convert AC power to DC power for all systems in a given rack, so there is no need for power supplies in the system.

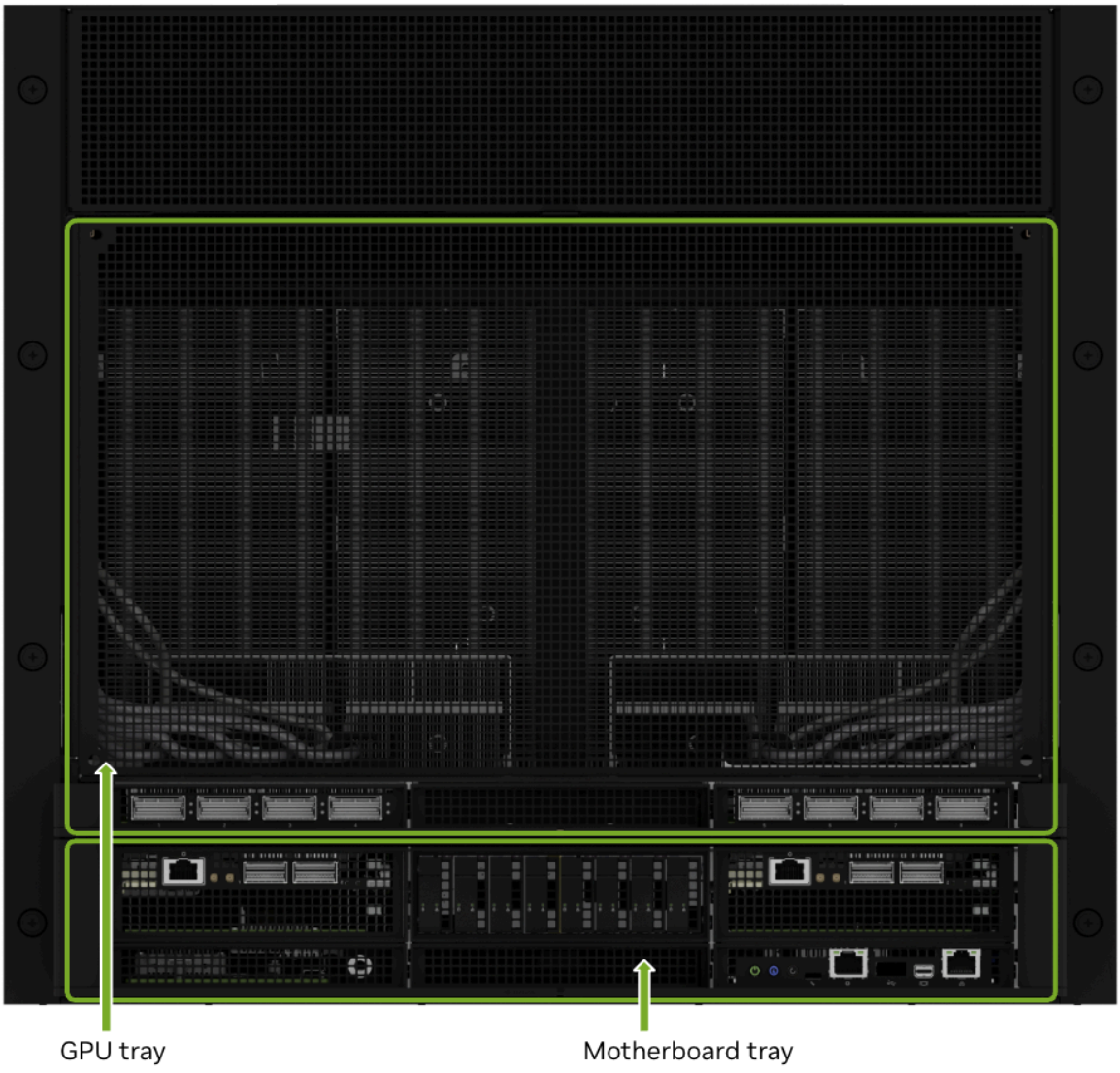


Figure 4: DGX B300 system DC version, front view

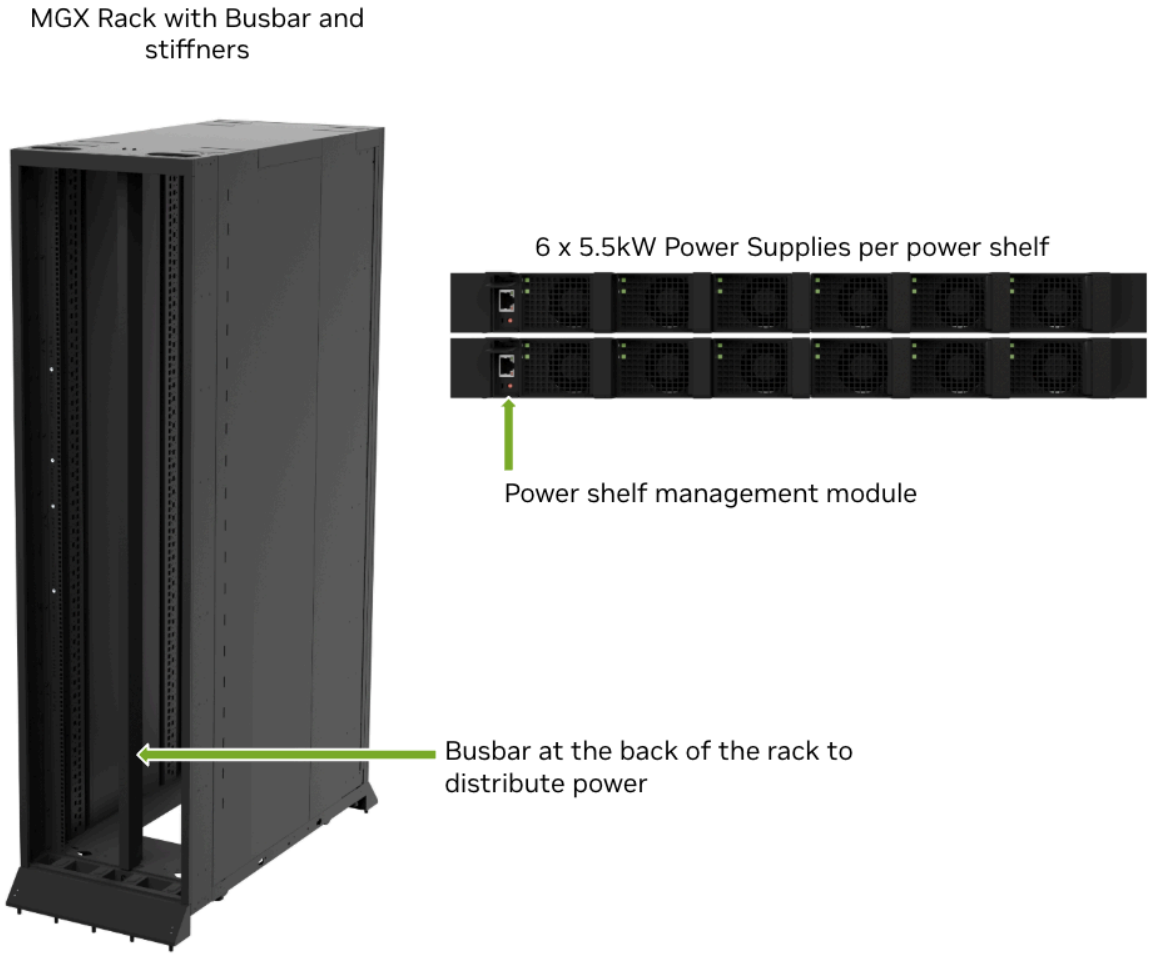


Figure 5: Busbar connectivity and power shelves for DGX B300 systems, DC version

Power Connectivity Differences

Integrated PSUs

Power supply considerations bring us to the rear of the DGX B300 system, where power connectivity is delivered. With the traditional AC PSU version of the DGX B300 system, twelve power cords are required for each system to connect to the power distribution units in the rack. This configuration provides fully redundant power connectivity because the DGX system needs six power supplies to function at maximum performance. It also allows data centers to balance the electrical load across three phases evenly across the twelve power cords.

12 x AC power supplies which are replaceable from the front with power cords connected to the back

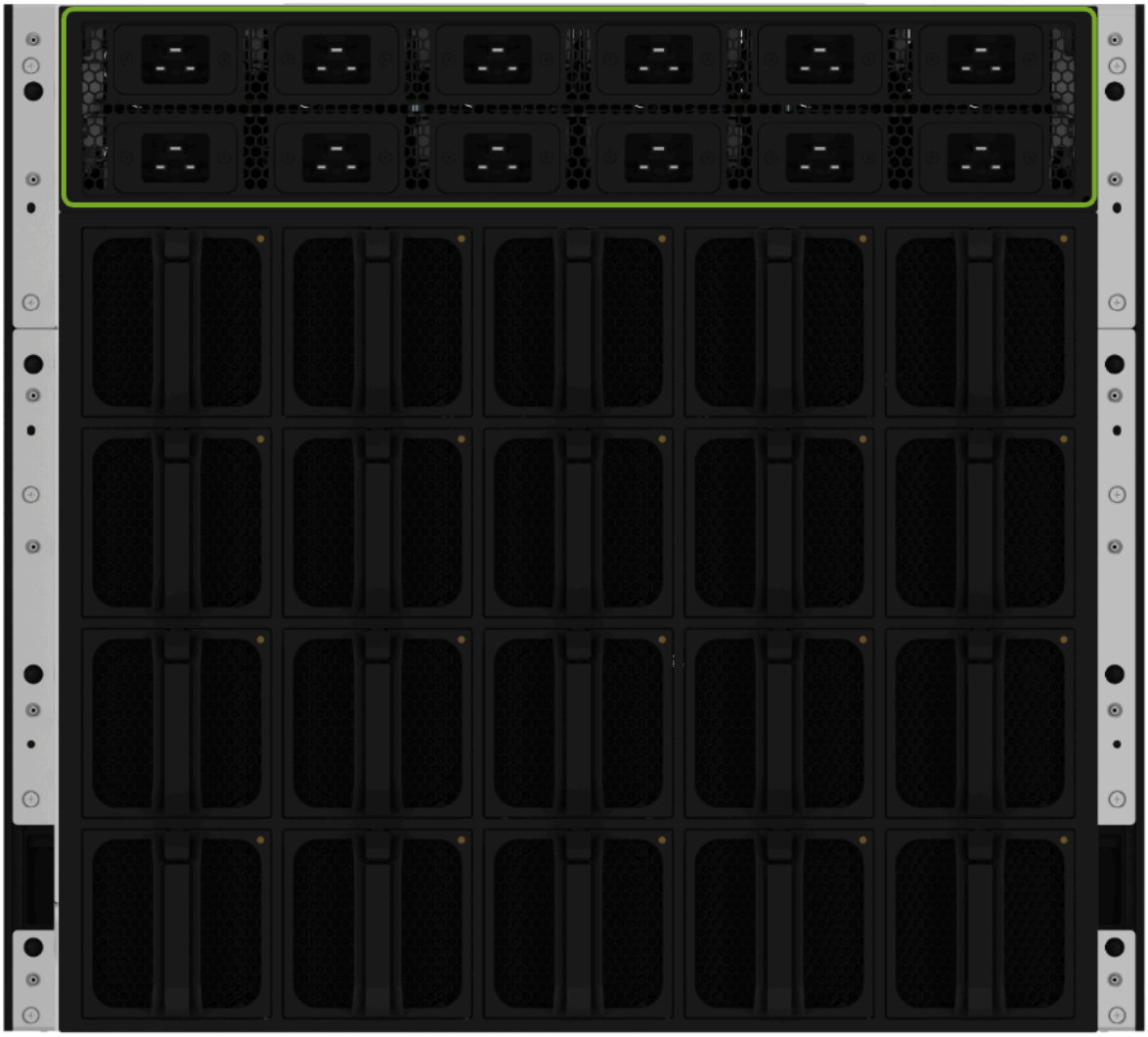


Figure 6: Rear of DGX B300, AC version

New Busbar Design

DGX B300 systems that require DC power come equipped with a clip that connects directly to a busbar in the rack. The power redundancy is taken care of by the power shelves in the rack, giving customers the flexibility to configure differing levels up to full N+N redundancy.

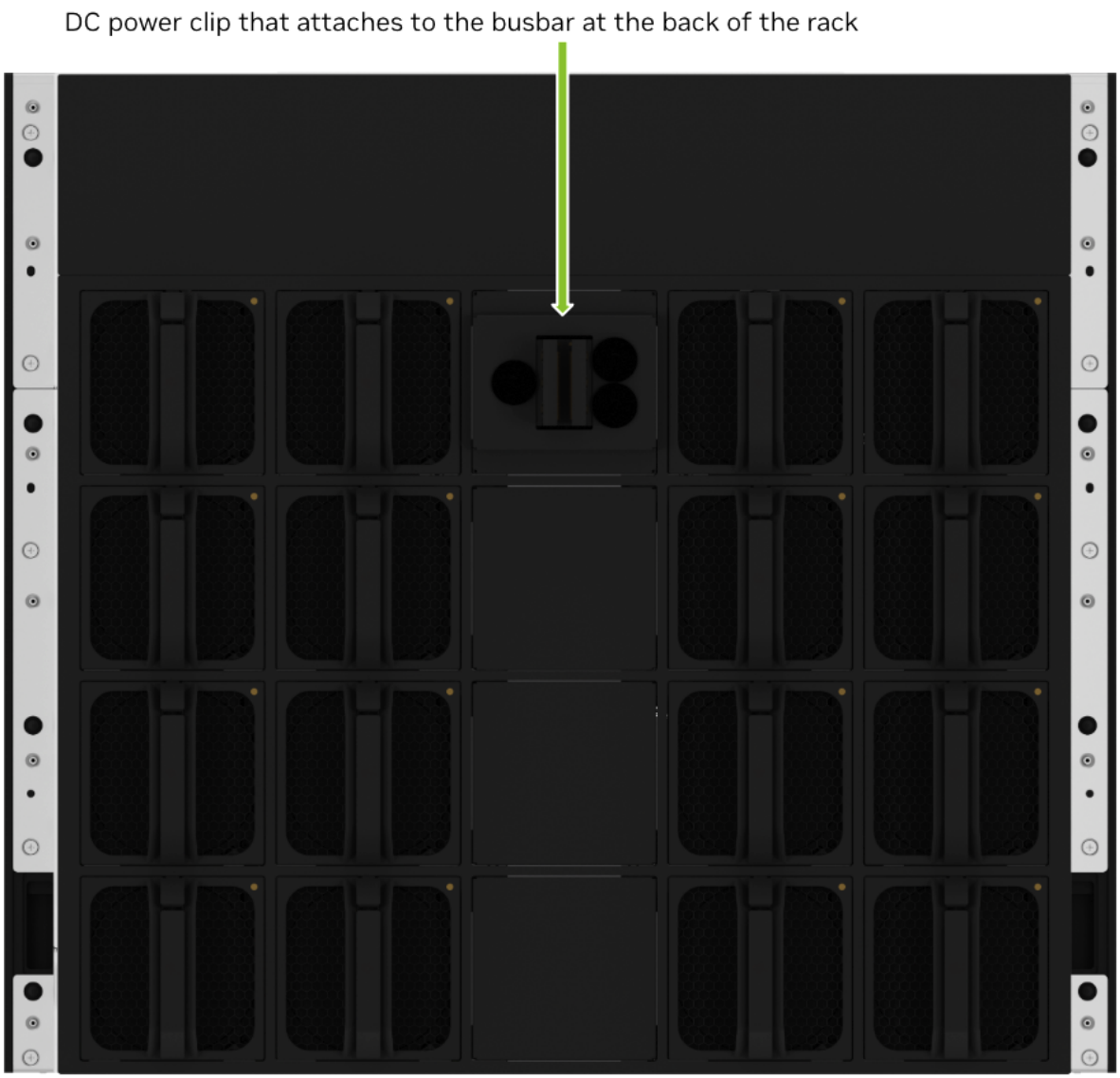


Figure 7: Rear of DGX B300, DC version

Allowing the data center to provide DC power using power shelves to the DGX system enables the infrastructure provider to configure redundancy to their comfort level based on local power outage history and maintenance schedules. It can also help make the data center run more efficiently because they reduce the number of places where AC to DC power is being converted, which usually uses energy and produces heat.

The benefit of these power configuration options is that data centers with the appropriate infrastructure can provision busbar versions of DGX clusters in denser data center environments while the integrated PSU systems can be installed in more traditional data centers.

Inside the NVIDIA DGX B300 system

Starting with the motherboard, we've discussed how all the IO elements were moved to the front of the tray. This not only moves the cabling to the front, but it also makes it easy to service the various components that use the cables. In addition to moving the NVIDIA BlueField-3 SuperNICs to the front, the dual M.2 NVMe boot drives are also in a special OPC carrier that connects them to PCIe. This allows the network interface cards and the boot drives to be serviceable from the front after unlocking the modules.

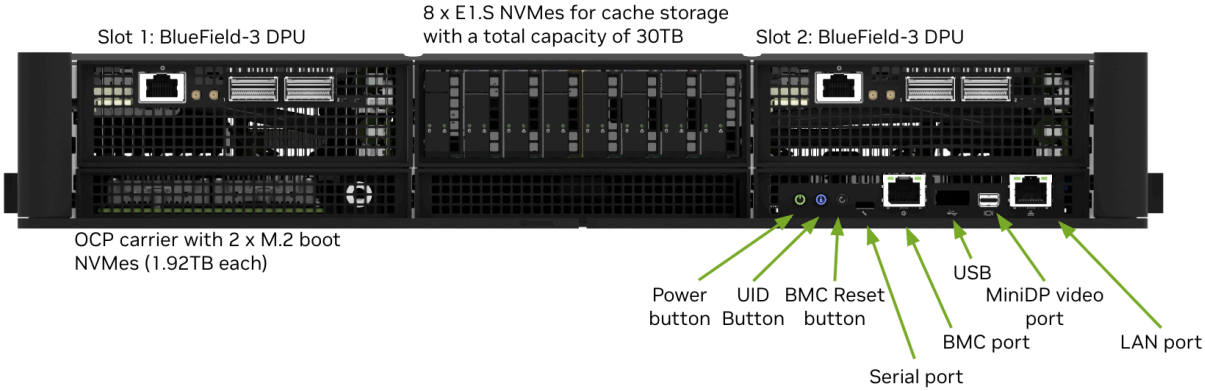


Figure 8: DGX B300 system motherboard, front view

The motherboard tray is on sliding rails, which makes it very easy to pull out of the DGX chassis. After removing the tray cover, the user can get access to any component that may need servicing, such as the DIMMs or the Data Center Secure Control Module (DC-SCM) which includes the baseboard management console (BMC) and the trusted platform module (TPM), among other devices and ports.

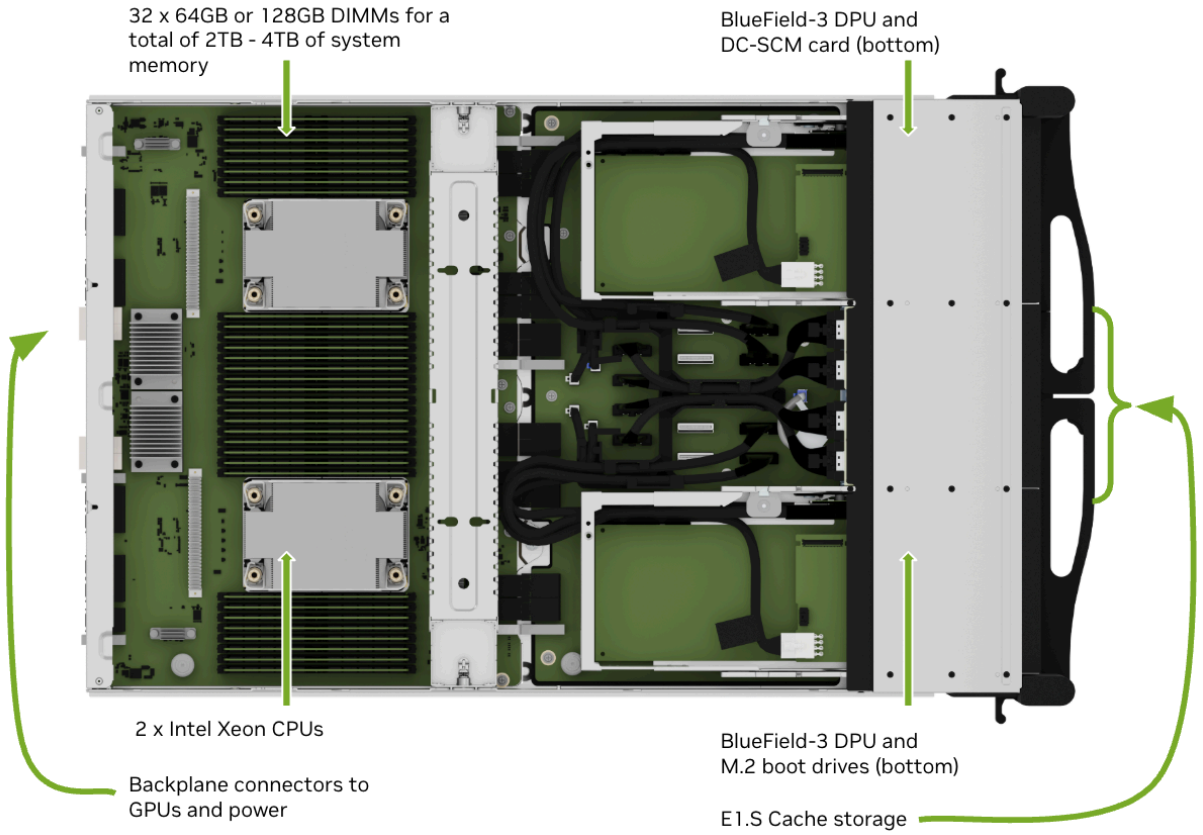


Figure 9: DGX B300 system motherboard, top view

The motherboard holds two Intel Xeon Platinum 6776P CPUs, and the system is configured with two terabytes of RAM, which can be upgraded by the customer to four terabytes if the customer's workload requires it. The rest of the components in the system are in a fixed configuration that has been optimized to run artificial intelligence workloads while providing ample IO and storage.

The GPU tray containing eight Blackwell Ultra GPUs can slide in and out of the system for ease of serviceability. The GPU tray design follows the 8-way architecture of previous DGX systems, connecting all of the GPUs using 5th generation NVIDIA NVLink switches for maximum performance while delivering 14.4 TB/s of aggregate bandwidth, doubling the bandwidth capabilities compared to Hopper generation systems.

As mentioned earlier, the NVIDIA ConnectX-8 network interface cards are capable of communicating at 800Gbps. In many architectures, these NICs would likely be connected to the IO subsystem that uses the latest PCIe Gen5 capable CPUs, limiting the output to the network at 480Gbps. In DGX B300, the ConnectX-8 interfaces are connected directly to the GPUs using PCIe Gen6, which doubles the bandwidth and fully supports the 800 Gbps throughput in NVIDIA's latest NICs.

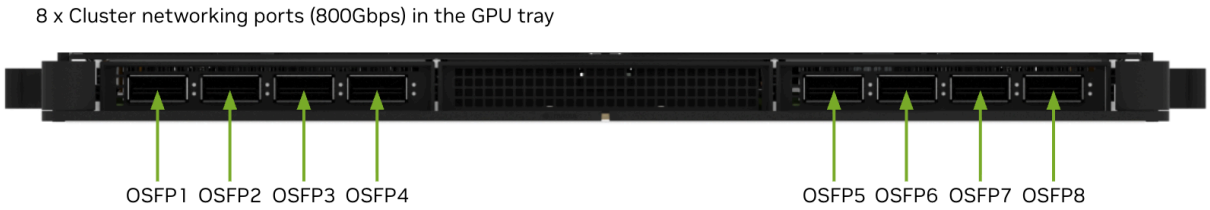


Figure 10: NVIDIA ConnectX-8 network interfaces in the lower section of the GPU tray

An important contribution of the ConnectX-8 cards is that they not only communicate the GPUs to the other nodes over the network, but they also provide the connectivity to the other components in the system. The NICs can be configured in either NVIDIA InfiniBand mode or in Ethernet mode, so they can be integrated with whichever network infrastructure on which the DGX cluster is built.

Conclusion

The NVIDIA DGX B300 system represents a significant advancement in infrastructure designed for enterprise AI. It addresses the challenges of scaling AI technologies, managing energy consumption, and reducing complexity for next-generation AI models in a traditional, air-cooled system. It features eight NVIDIA Blackwell Ultra GPUs within a 10 Rack Unit (RU) chassis, delivering exceptional performance for both complex training and large-scale inference. Key architectural innovations include a 5th generation NVLink fabric providing 14.4 TB/s of aggregate bandwidth and a novel implementation of PCIe Gen6 that directly connects high-speed network interfaces to the GPUs, ensuring maximum data throughput and minimizing I/O bottlenecks in scaled-out cluster environments.

DGX B300 also offers flexible deployment options, supporting traditional AC power via integrated Power Supply Units (PSUs) or highly efficient DC power through external busbars when integrated into NVIDIA MGX-compliant racks. Having two options for power configurations, combined with enhanced front serviceability for components like BlueField-3 network cards and M.2 NVMe boot drives, underscores a design philosophy focused on operational efficiency, reduced maintenance downtime, and adaptability to diverse data center infrastructures. The system's ability to deliver extreme compute density while ensuring all peripheral systems match this speed provides customers with a robust and adaptable solution for their high-performance AI workloads.

For more information, please review the NVIDIA DGX B300 system user guide.

Notice

This document is provided for information purposes only and shall not be regarded as a warranty of a certain functionality, condition, or quality of a product. NVIDIA Corporation (“NVIDIA”) makes no representations or warranties, expressed or implied, as to the accuracy or completeness of the information contained in this document and assumes no responsibility for any errors contained herein. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This document is not a commitment to develop, release, or deliver any Material (defined below), code, or functionality.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and any other changes to this document, at any time without notice.

Customers should obtain the latest relevant information before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer (“Terms of Sale”). NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this document. No contractual obligations are formed either directly or indirectly by this document.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer (“Terms of Sale”). NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this document. No contractual obligations are formed either directly or indirectly by this document.

NVIDIA products are not designed, authorized, or warranted to be suitable for use in medical, military, aircraft, space, or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death, or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer’s own risk.

NVIDIA makes no representation or warranty that products based on this document will be suitable for any specified use. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is the customer's sole responsibility to evaluate and determine the applicability of any information contained in this document, ensure the product is suitable and fit for the application planned by the customer, and perform the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer’s product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this document. NVIDIA accepts no liability related to any default, damage, costs, or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this document or (ii) customer product designs.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this document. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA.

Reproduction of information in this document is permissible only if approved in advance by NVIDIA in writing, reproduced without alteration and in full compliance with all applicable export laws and regulations, and accompanied by all associated conditions, limitations, and notices.

THIS DOCUMENT AND ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, “MATERIALS”) ARE BEING PROVIDED “AS IS.” NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL NVIDIA BE LIABLE FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF ANY USE OF THIS DOCUMENT, EVEN IF NVIDIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA’s aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the Terms of Sale for the product.

Trademarks

NVIDIA, the NVIDIA logo, are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

Arm

Arm, AMBA, and Arm Powered are registered trademarks of Arm Limited. Cortex, MPCore, and Mali are trademarks of Arm Limited. All other brands or product names are the property of their respective holders. "Arm" is used to represent Arm Holdings plc; its operating company Arm Limited; and the regional subsidiaries Arm Inc.; Arm KK; Arm Korea Limited.; Arm Taiwan Limited; Arm France SAS; Arm Consulting (Shanghai) Co. Ltd.; Arm Germany GmbH; Arm Embedded Technologies Pvt. Ltd.; Arm Norway, AS, and Arm Sweden AB.

Copyright

© 2025 NVIDIA Corporation. All rights reserved.

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