Datasheet

# NVIDIA DGX GH200

Massive memory supercomputing for emerging AI.

As AI use cases continue to get more elaborate, the size and complexity of AI models are skyrocketing. While most organizations need to process many AI workloads in parallel, a segment of users have massive memory requirements for a single workload that exceed the bounds of a GPU or even a large, multi-GPU system. These users need a new approach that can extend the memory and processing power of hundreds of GPUs and CPUs with no performance bottleneck as scale increases, while preserving a single-GPU programming model for simplicity.

For these organizations, which include cloud service providers (CSPs), hyperscalers, large research organizations, and other leading-edge businesses pushing the boundaries of AI, NVIDIA DGX™ GH200 provides a new blueprint for giant model AI development. Using the NVIDIA Grace Hopper™ Superchip with integrated 4th generation NVIDIA® NVLink® technology that provides linear scalability and a massive, shared-memory space across all GPUs, this new class of AI supercomputer provides the capabilities needed to develop the world's largest graph neural networks, recommenders, simulation models, and generative AI applications.

As part of the DGX platform, DGX GH200 is more than hardware—it's a complete software and hardware solution designed and delivered by NVIDIA, with an end-to-end, turnkey experience; white-glove services that eliminate complexity, speed deployments, and simplify operation; and excellent power efficiency for a large memory supercomputer. DGX GH200 introduces a new epoch in AI, raising the state of the art for model size and complexity to a level that's never before been possible.

## Giant Memory for Giant Models

Unlike existing AI supercomputers that are designed to support workloads that fit within the memory of a single system, NVIDIA DGX GH200 is the only AI supercomputer that offers a shared memory space of up to 144TB across 256 Grace Hopper Superchips, providing developers with nearly 500X more fast-access memory to build massive models. DGX GH200 is the first supercomputer to pair Grace Hopper Superchips with the NVIDIA NVLink Switch System, which allows up to 256 GPUs to be united as one data-center-size GPU. This architecture provides 48X more bandwidth than the previous generation, delivering the power of a massive AI supercomputer with the simplicity of programming a single GPU.

## Key Features

**NVIDIA DGX GH200**

- 256 NVIDIA Grace Hopper Superchips, all interconnected with NVIDIA NVLink
- Massive, shared-GPU-memory space of 144 terabytes (TB)
- 900 gigabytes per second (GB/s) GPU-to-GPU bandwidth
- 1 exaFLOPS of FP8 AI performance
- NVIDIA Base Command™ and NVIDIA AI Enterprise software
- White-glove implementation experience

## Super Power-Efficient Computing

As the complexity of AI models has increased, the technology to develop and deploy them has become more resource intensive. However, using the NVIDIA Grace Hopper architecture, DGX GH200 achieves excellent power efficiency. Each NVIDIA Grace Hopper Superchip is both a CPU and GPU in one unit, connected with superfast NVIDIA NVLink-C2C. The Grace™ CPU uses LPDDR5X memory, which consumes one-eighth the power of traditional DDR5 system memory while providing 50 percent more bandwidth than eight-channel DDR5. And being on the same package, the Grace CPU and Hopper™ GPU interconnect consumes 5X less power and provides 7X the bandwidth compared to the latest PCIe technology used in other systems.

## Integrated and Ready to Run

Designing, integrating, and operationalizing a hyperscale data center tuned for massive-memory application development can be complex and time consuming. With DGX GH200, NVIDIA is not just a technology provider but a trusted partner that helps ensure success. As a fully tested and integrated solution with software, compute, and networking, that includes white-glove services spanning installation and infrastructure management to expert advice on optimizing workloads, DGX GH200 lets teams hit the ground running.

| DGX GH200 Technical Specifications | |
|---|---|
| **CPU and GPU** | 256x NVIDIA Grace Hopper Superchips |
| **CPU Cores** | 18,432 Arm® Neoverse V2 Cores with SVE2 4X 128b |
| **GPU Memory** | 144TB |
| **Performance** | 1 exaFLOPS |
| **Networking** | 256x OSFP single-port NVIDIA ConnectX®-7 VPI with 400Gb/s InfiniBand<br>256x dual-port NVIDIA BlueField®-3 VPI with 200Gb/s InfiniBand and Ethernet<br>24x NVIDIA Quantum-2 QM9700 InfiniBand Switches<br>20x NVIDIA Spectrum™ SN2201 Ethernet Switches<br>22x NVIDIA Spectrum SN3700 Ethernet Switches |
| **NVIDIA NVLink Switch System** | 96x L1 NVIDIA NVLink Switches<br>36x L2 NVIDIA NVLink Switches |
| **Management Network** | Host baseboard management controller<br>(BMC) with RJ45 |
| **Software** | NVIDIA AI Enterprise (optimized AI software)<br>NVIDIA Base Command (orchestration, scheduling, and cluster management)<br>DGX OS / Ubuntu / Red Hat Enterprise Linux / Rocky (operating system) |
| **Support** | Comes with three-year business-standard hardware and software support |

## Ready to Get Started?

To learn more about DGX GH200, visit: **nvidia.com/dgx-gh200**



NVIDIA