

NVIDIA HGX AI 超级计算机

先进的 AI 计算平台。



专为 AI 和 HPC 而打造

AI、复杂模拟和海量数据集需要多个具有极快互连速度的 GPU 和完全加速的软件堆栈。NVIDIA HGX™ AI 超级计算平台整合了 NVIDIA GPU、NVLink®、NVIDIA 网络以及全面优化的 AI 和 高性能计算 (HPC) 软件堆栈的全部功能，可提供最高的应用性能并加快获得见解的速度。

将 HGX 与 NVIDIA 网络相连

	NVIDIA Quantum-2 InfiniBand 平台： Quantum-2 交换机，ConnectX-7 适配器，BlueField-3 DPU	NVIDIA Spectrum-X 平台： Spectrum-4 交换机，BlueField-3 DPU，Spectrum-X 许可证	NVIDIA Spectrum 以太网平台： Spectrum 交换机，ConnectX 适配器，BlueField DPU
深度学习训练	最佳	更佳	适用
科学计算模拟	最佳	更佳	适用
数据分析	最佳	更佳	适用
深度学习推理	最佳	更佳	适用

NVIDIA HGX 规格

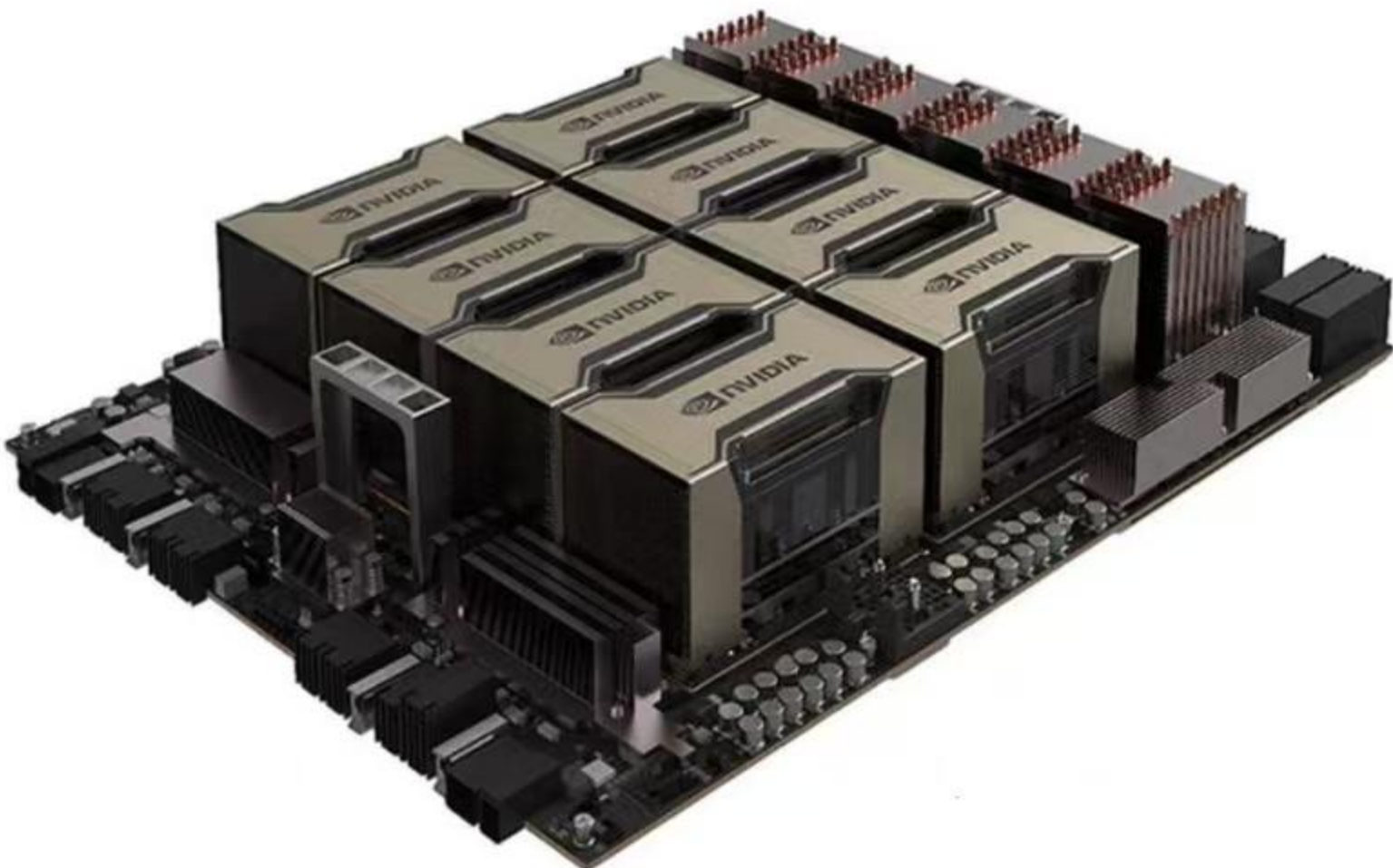
NVIDIA HGX 采用单一基板，配备 4 个 H200 或 H100 GPU 或是 8 个 H200、H100、B200 或 B100 GPU。这些强大的硬件和软件组合为打造强劲的 AI 超级计算平台奠定了基础。

	HGX B200	HGX B100
GPUs	HGX B200 8-GPU	HGX B100 8-GPU
外形规格	8x NVIDIA B200 SXM	8x NVIDIA B100 SXM
HPC 和 AI 计算 (FP64/TF32/FP16/FP8/FP4)*	320TF/18PF/36PF/72PF/144PF	240TF/14PF/28PF/56PF/112PF
显存	最高可达 1.5TB	最高可达 1.5TB
NVIDIA NVLink	第五代	第五代
NVIDIA NVSwitch™	第四代	第四代
NVSwitch 互联 GPU 间带宽	1.8TB/s	1.8TB/s
聚合总带宽	14.4TB/s	14.4TB/s

	HGX H200	
	4-GPU	8-GPU
GPUs	HGX H200 4-GPU	HGX H200 8-GPU
外形规格	4x NVIDIA H200 SXM	8x NVIDIA H200 SXM
HPC 和 AI 计算 (FP64/TF32/FP16/FP8/INT8)*	268TF/4PF/8PF/16PF/16 POPS	535TF/8PF/16PF/32PF/32 POPS
显存	最高可达 564GB	最高可达 1.1TB
NVLink	第四代	第四代
NVSwitch	不适用	第三代
NVSwitch 互联 GPU 间带宽	不适用	900GB/s
聚合总带宽	3.6TB/s	7.2TB/s

	HGX H100	
	4-GPU	8-GPU
GPUs	HGX H100 4-GPU	HGX H100 8-GPU
外形规格	4x NVIDIA H100 SXM	8x NVIDIA H100 SXM
HPC 和 AI 计算 (FP64/TF32/FP16/FP8/INT8)*	268TF/4PF/8PF/16PF/16 POPS	535TF/8PF/16PF/32PF/32 POPS
显存	最高可达 320GB	最高可达 640GB
NVLink	第四代	第四代
NVSwitch	不适用	第三代
NVLink Switch	不适用	不适用
NVSwitch 互联 GPU 间带宽	不适用	900GB/s
聚合总带宽	3.6TB/s	7.2TB/s

* 采用稀疏技术



目录

以前 Inspur A800|A100 PCIe服务器

- HGX H100|H800模组

以后 NVIDIA DGX H100