NV Blackwell 产品演进分析

NVIDIA ZOMI

NVIDIA GPU架构发展

架构名称	Fermi	Kepler	Maxwell	Pascal	Volta	Turing	Ampere	Hopper
中文名字	费米	开普勒	麦克斯韦	帕斯卡	伏特	图灵	安培	赫柏
发布时间	2010	2012	2014	2016	2017	2018	2020	2022
核心参数	16个SM，每个SM包含32个CUDA Cores，一共512 CUDA Cores	15个SMX，每个SMX包括192个FP32+64个FP64 CUDA Cores	16个SM，每个SM包括4个处理块，每个处理块包括32个CUDA Cores +8个LD/ST Unit + 8 SFU	GP100有60个SM，每个SM包括64个CUDA Cores，32个DP Cores	80个SM，每个SM包括32个FP64 +64 Int32+64 FP32+8个Tensor Cores	102核心92个SM，SM重新设计，每个SM包含64个Int32+64个FP32+8个Tensor Cores	108个SM，每个SM包含64个FP32+64个INT32+32个FP64+4个Tensor Cores	132个SM，每个SM包含128个FP32+64个INT32+64个FP64+4个Tensor Cores
特点&优势	首个完整GPU计算架构，支持与共享存储结合的Cache层次GPU架构，支持ECC GPU架构	游戏性能大幅提升，首次支持GPU Direct技术	每组SM单元从192个减少到每组128个，每个SMM单元拥有更多逻辑控制电路	NVLink第一代，双向互联带宽160 GB/s，P100拥有56个SM HBM	NVLink2.0，Tensor Cores第一代，支持AI运算	Tensor Core2.0，RT Core第一代	Tensor Core3.0，RT Core2.0，NVLink3.0，结构稀疏性矩阵MIG1.0	Tensor Core4.0，NVlink4.0，结构稀疏性矩阵MIG2.0
纳米制程	40/28nm 30亿晶体管	28nm 71亿晶体管	28nm 80亿晶体管	16nm 153亿晶体管	12nm 211亿晶体管	12nm 186亿晶体管	7nm 283亿晶体管	4nm 800亿晶体管
代表型号	Quadro 7000	K80 K40M	M5000 M4000 GTX 9XX系列	P100 P6000 TTX1080	V100 TiTan V	T4，2080TI RTX 5000	A100 A30系列	H100

NVIDIA GPU架构发展

- B200、B100、GB200、NVL72、NVL32、SuperPod、GH200、H200、H100、L20、SuperPod-576
- ConnectX-800G网卡、网络交换机

本节内容

1. 单 GPU 介绍
2. HGX 产品
3. NVL 产品

01

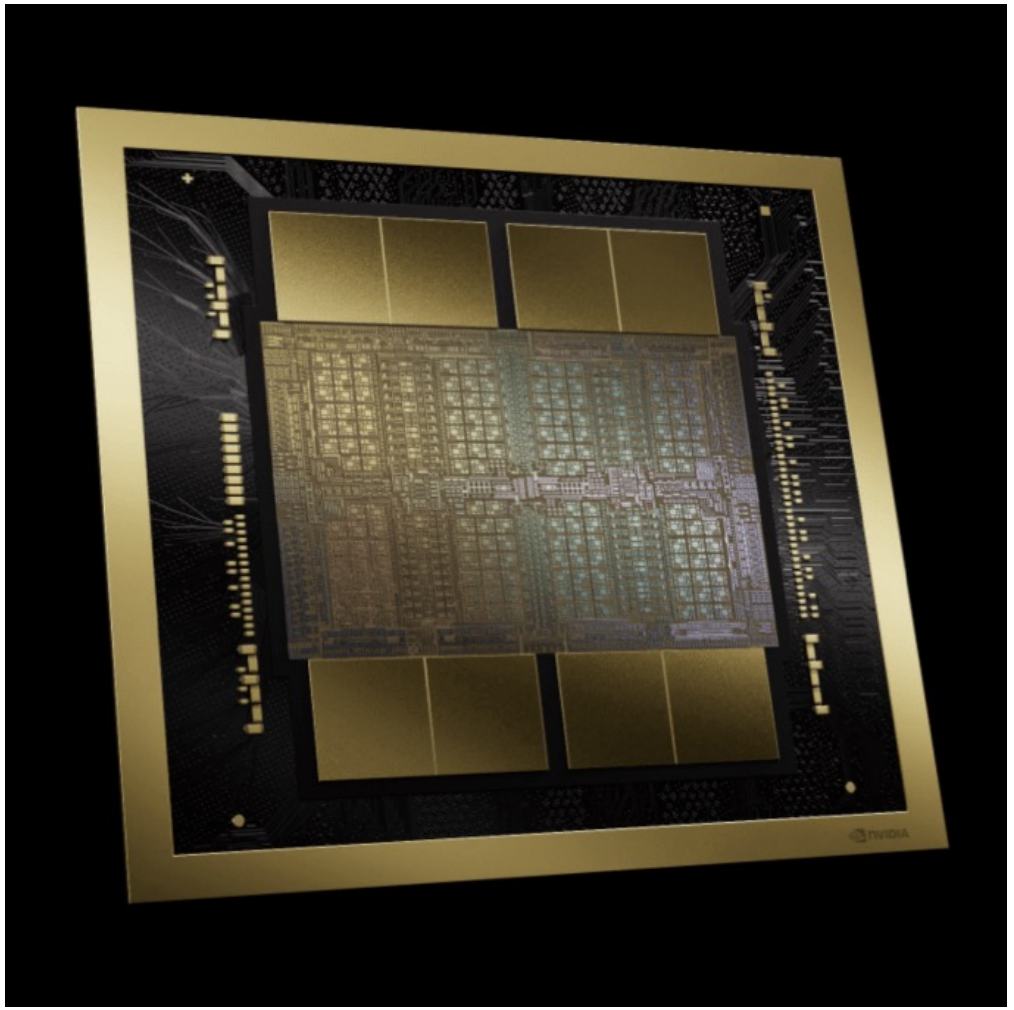
GPU产品介绍

单 GPU 产品介绍

产品	A100	H100	H200	GH200	B100	B200	Full B200	GB200
架构	Ampere	Hopper			Blackwell			
HBM 大小	80GB	80GB	141GB	96GB/141GB	180GB/192GB	180GB/192GB	192GB	384GB
HBM 带宽	2TB/s	3.35TB/s	4.8TB/s	4TB/s 4.9TB/s	8TB/s	8TB/s	8TB/s	16TB/s
FP16 (FLOPS)	312T	1P	1P	1P	1.75P	2.25P	2.5P	5P
INT8 (OPS)	624T	2P	2P	2P	3.5P	4.5P	5P	10P
FP8 (FLOPS)	N	2P	2P	2P	3.5P	4.5P	5P	10P
FP6 (FLOPS)	N	N	N	N	3.5P	4.5P	5P	10P
FP4 (FLOPS)	N	N	N	N	7P	9P	10P	20P
NVLink 带宽	600GB/s	900GB/s	900GB/s	900GB/s	1.8TB/s	1.8TB/s	1.8TB/s	3.6TB/s
Powers	400w	700w	700w	1000w	700w	1000w	1200w	2700w
Others	1 Die	1 Die	1 Die	1 Grace CPU + 1 H200 GPU	2 Die	2 Die	2 Die	1 Grace CPU + B200 GPU

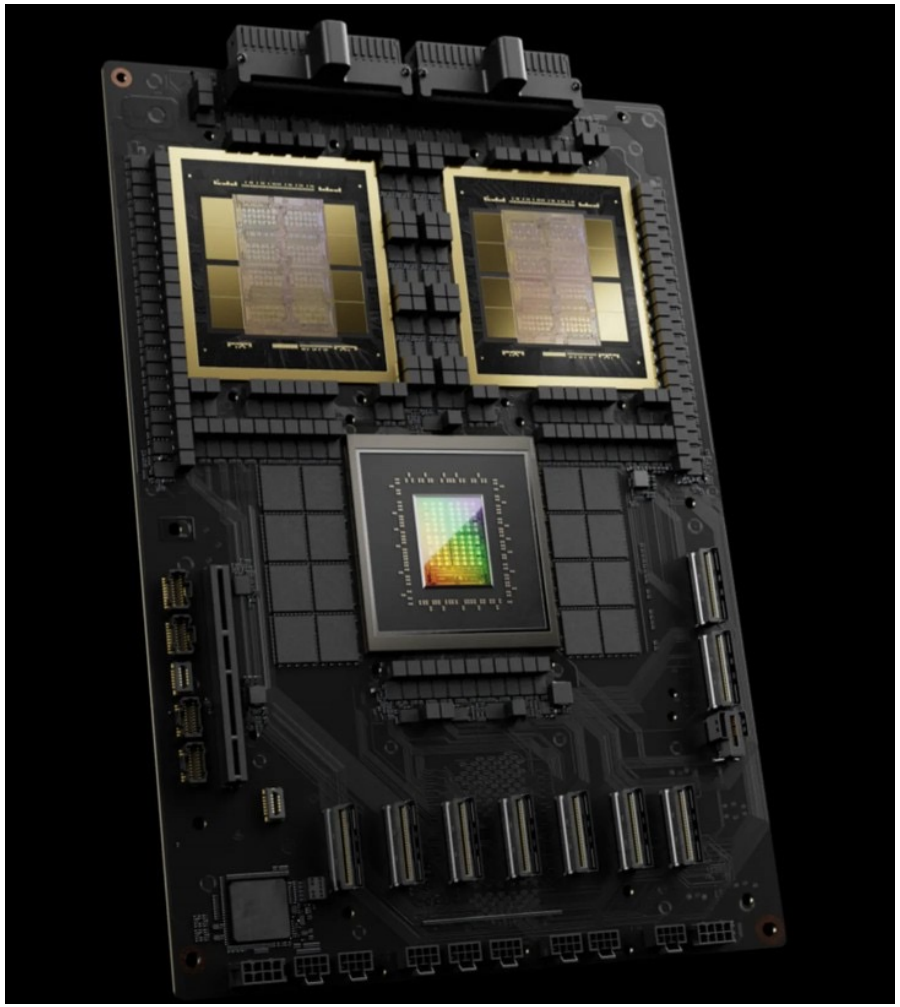
单 GPU 产品介绍

产品	B100	B200	Full B200	GB200
架构	Blackwell			
HBM 大小	180GB/192GB	180GB/192GB	192GB	384GB
HBM 带宽	8TB/s	8TB/s	8TB/s	16TB/s
FP16 (FLOPS)	1.75P	2.25P	2.5P	5P
INT8 (OPS)	3.5P	4.5P	5P	10P
FP8 (FLOPS)	3.5P	4.5P	5P	10P
FP6 (FLOPS)	3.5P	4.5P	5P	10P
FP4 (FLOPS)	7P	9P	10P	20P
NVLink 带宽	1.8TB/s	1.8TB/s	1.8TB/s	3.6TB/s
Powers	700w	1000w	1200w	2700w
Others	2 Die	2 Die	2 Die	1 Grace CPU + B200 GPU



单 GPU 产品介绍

产品	B100	B200	Full B200	GB200
架构	Blackwell			
HBM 大小	180GB/192GB	180GB/192GB	192GB	384GB
HBM 带宽	8TB/s	8TB/s	8TB/s	16TB/s
FP16 (FLOPS)	1.75P	2.25P	2.5P	5P
INT8 (OPS)	3.5P	4.5P	5P	10P
FP8 (FLOPS)	3.5P	4.5P	5P	10P
FP6 (FLOPS)	3.5P	4.5P	5P	10P
FP4 (FLOPS)	7P	9P	10P	20P
NVLink 带宽	1.8TB/s	1.8TB/s	1.8TB/s	3.6TB/s
Powers	700w	1000w	1200w	2700w
Others	2 Die	2 Die	2 Die	1 Grace CPU + B200 GPU



单 GPU 产品介绍

- A100 -> H100 FP16 算力增加 3 倍多，功耗只从 400w 增加到 700w

产品	A100	H100	H200	GH200	B100	B200	Full B200	GB200
架构	Ampere	Hopper			Blackwell			
HBM 大小	80GB	80GB	141GB	96GB/141GB	180GB/192GB	180GB/192GB	192GB	384GB
HBM 带宽	2TB/s	3.35TB/s	4.8TB/s	4TB/s 4.9TB/s	8TB/s	8TB/s	8TB/s	16TB/s
FP16 (FLOPS)	312T	1P	1P	1P	1.75P	2.25P	2.5P	5P
INT8 (OPS)	624T	2P	2P	2P	3.5P	4.5P	5P	10P
FP8 (FLOPS)	N	2P	2P	2P	3.5P	4.5P	5P	10P
FP6 (FLOPS)	N	N	N	N	3.5P	4.5P	5P	10P
FP4 (FLOPS)	N	N	N	N	7P	9P	10P	20P
NVLink 带宽	600GB/s	900GB/s	900GB/s	900GB/s	1.8TB/s	1.8TB/s	1.8TB/s	3.6TB/s
Powers	400w	700w	700w	1000w	700w	1000w	1200w	2700w
Others	1 Die	1 Die	1 Die	1 Grace CPU + 1 H200 GPU	2 Die	2 Die	2 Die	1 Grace CPU + B200 GPU

单 GPU 产品介绍

- H200 -> B200 FP16 算力增加 2 倍多，功耗只从 700w 增加到 1000w

产品	A100	H100	H200	GH200	B100	B200	Full B200	GB200
架构	Ampere	Hopper			Blackwell			
HBM 大小	80GB	80GB	141GB	96GB/141GB	180GB/192GB	180GB/192GB	192GB	384GB
HBM 带宽	2TB/s	3.35TB/s	4.8TB/s	4TB/s 4.9TB/s	8TB/s	8TB/s	8TB/s	16TB/s
FP16 (FLOPS)	312T	1P	1P	1P	1.75P	2.25P	2.5P	5P
INT8 (OPS)	624T	2P	2P	2P	3.5P	4.5P	5P	10P
FP8 (FLOPS)	N	2P	2P	2P	3.5P	4.5P	5P	10P
FP6 (FLOPS)	N	N	N	N	3.5P	4.5P	5P	10P
FP4 (FLOPS)	N	N	N	N	7P	9P	10P	20P
NVLink 带宽	600GB/s	900GB/s	900GB/s	900GB/s	1.8TB/s	1.8TB/s	1.8TB/s	3.6TB/s
Powers	400w	700w	700w	1000w	700w	1000w	1200w	2700w
Others	1 Die	1 Die	1 Die	1 Grace CPU + 1 H200 GPU	2 Die	2 Die	2 Die	1 Grace CPU + B200 GPU

单 GPU 产品介绍

- B200 FP16 算力是 A100 的 7 倍，而功耗是其 2.5 倍

产品	A100	H100	H200	GH200	B100	B200	Full B200	GB200
架构	Ampere	Hopper			Blackwell			
HBM 大小	80GB	80GB	141GB	96GB/141GB	180GB/192GB	180GB/192GB	192GB	384GB
HBM 带宽	2TB/s	3.35TB/s	4.8TB/s	4TB/s 4.9TB/s	8TB/s	8TB/s	8TB/s	16TB/s
FP16 (FLOPS)	312T	1P	1P	1P	1.75P	2.25P	2.5P	5P
INT8 (OPS)	624T	2P	2P	2P	3.5P	4.5P	5P	10P
FP8 (FLOPS)	N	2P	2P	2P	3.5P	4.5P	5P	10P
FP6 (FLOPS)	N	N	N	N	3.5P	4.5P	5P	10P
FP4 (FLOPS)	N	N	N	N	7P	9P	10P	20P
NVLink 带宽	600GB/s	900GB/s	900GB/s	900GB/s	1.8TB/s	1.8TB/s	1.8TB/s	3.6TB/s
Powers	400w	700w	700w	1000w	700w	1000w	1200w	2700w
Others	1 Die	1 Die	1 Die	1 Grace CPU + 1 H200 GPU	2 Die	2 Die	2 Die	1 Grace CPU + B200 GPU

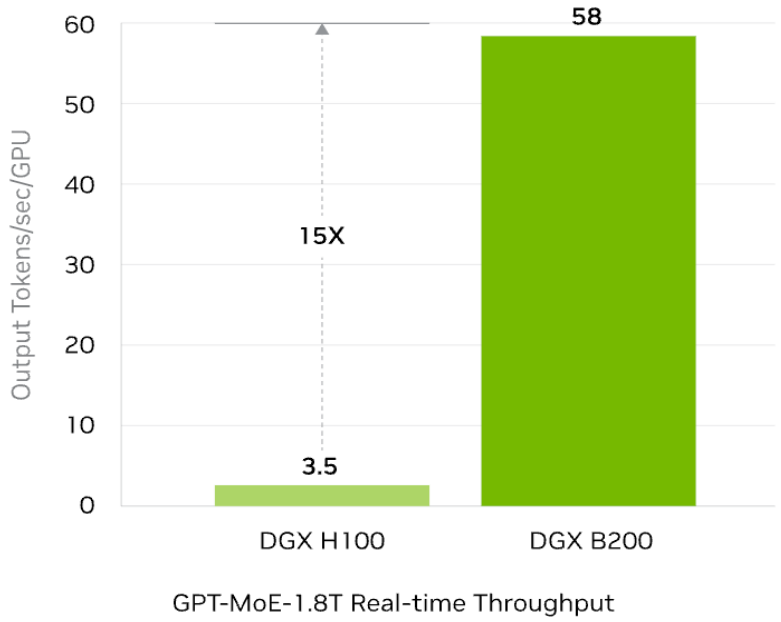
单 GPU 产品介绍

- Blackwell GPU 支持 FP4 精度，其算力为 FP8 的两倍

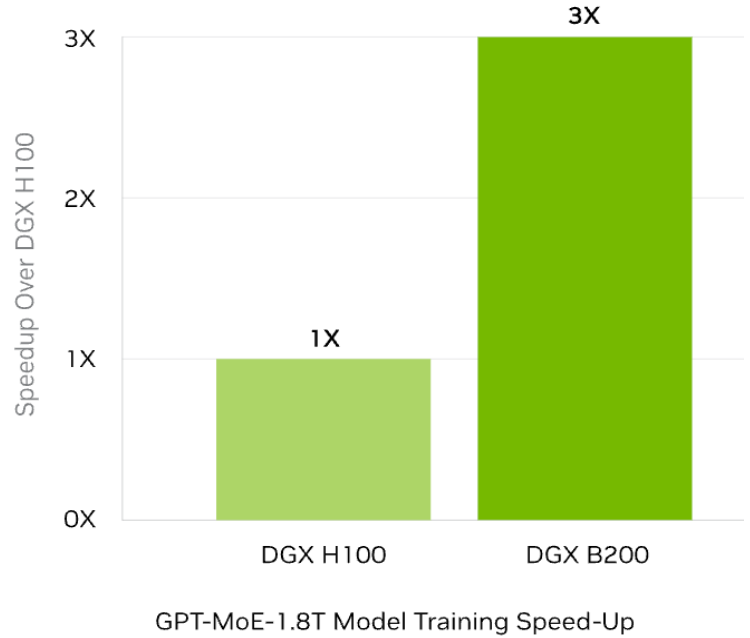
产品	A100	H100	H200	GH200	B100	B200	Full B200	GB200
架构	Ampere	Hopper			Blackwell			
HBM 大小	80GB	80GB	141GB	96GB/141GB	180GB/192GB	180GB/192GB	192GB	384GB
HBM 带宽	2TB/s	3.35TB/s	4.8TB/s	4TB/s 4.9TB/s	8TB/s	8TB/s	8TB/s	16TB/s
FP16 (FLOPS)	312T	1P	1P	1P	1.75P	2.25P	2.5P	5P
INT8 (OPS)	624T	2P	2P	2P	3.5P	4.5P	5P	10P
FP8 (FLOPS)	N	2P	2P	2P	3.5P	4.5P	5P	10P
FP6 (FLOPS)	N	N	N	N	3.5P	4.5P	5P	10P
FP4 (FLOPS)	N	N	N	N	7P	9P	10P	20P
NVLink 带宽	600GB/s	900GB/s	900GB/s	900GB/s	1.8TB/s	1.8TB/s	1.8TB/s	3.6TB/s
Powers	400w	700w	700w	1000w	700w	1000w	1200w	2700w
Others	1 Die	1 Die	1 Die	1 Grace CPU + 1 H200 GPU	2 Die	2 Die	2 Die	1 Grace CPU + B200 GPU

DGX B200 适用于推理！

- Real Time Large Language Model Inference



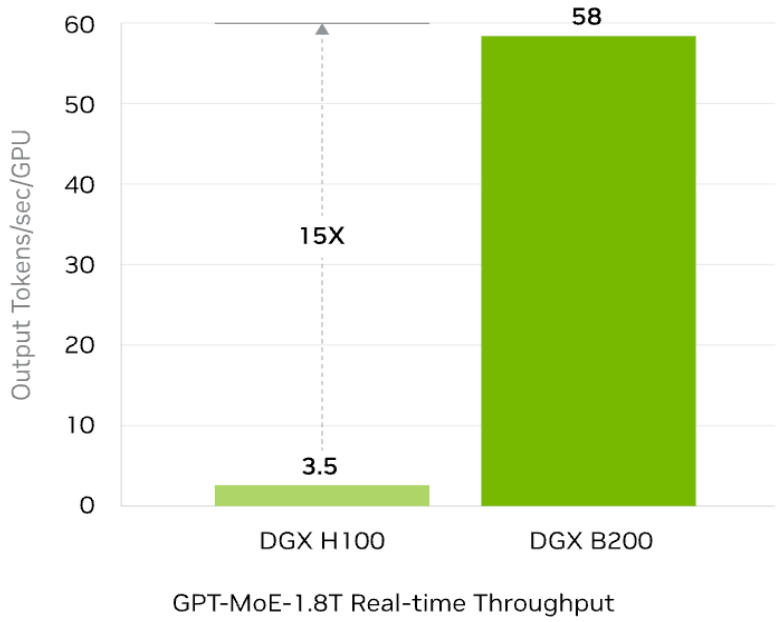
- Supercharged AI Training Performance



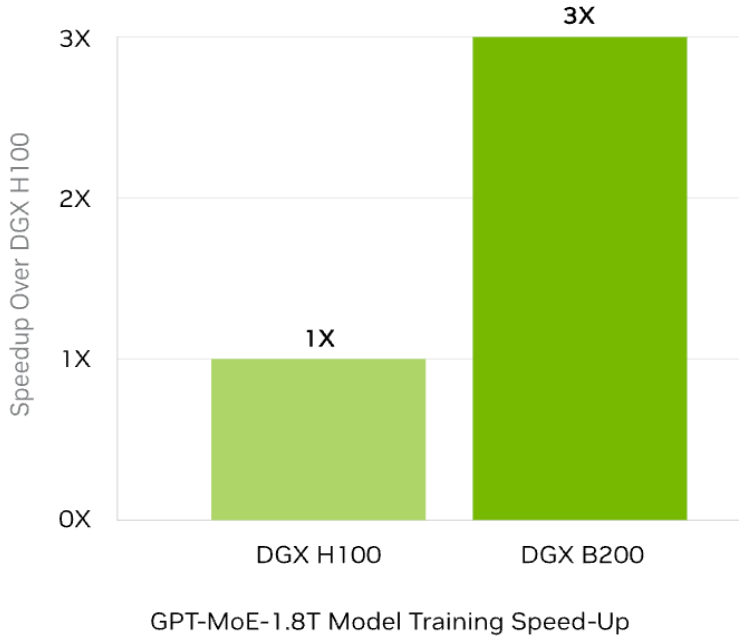
HGX H100 -> HGX B200 FP16/FP8 算力提升 2.25 倍，HBM 更大，显存带宽也达到 2.3 倍左右，NVLink 带宽加倍，假设 MFU 为 50%，那整体训练速度提升 3 倍符合预期。

DGX B200 不一定适用于推理！

- Real Time Large Language Model Inference



- Supercharged AI Training Performance



- 用 FP4 算力和 Hopper 架构的 FP8 算力比较

02

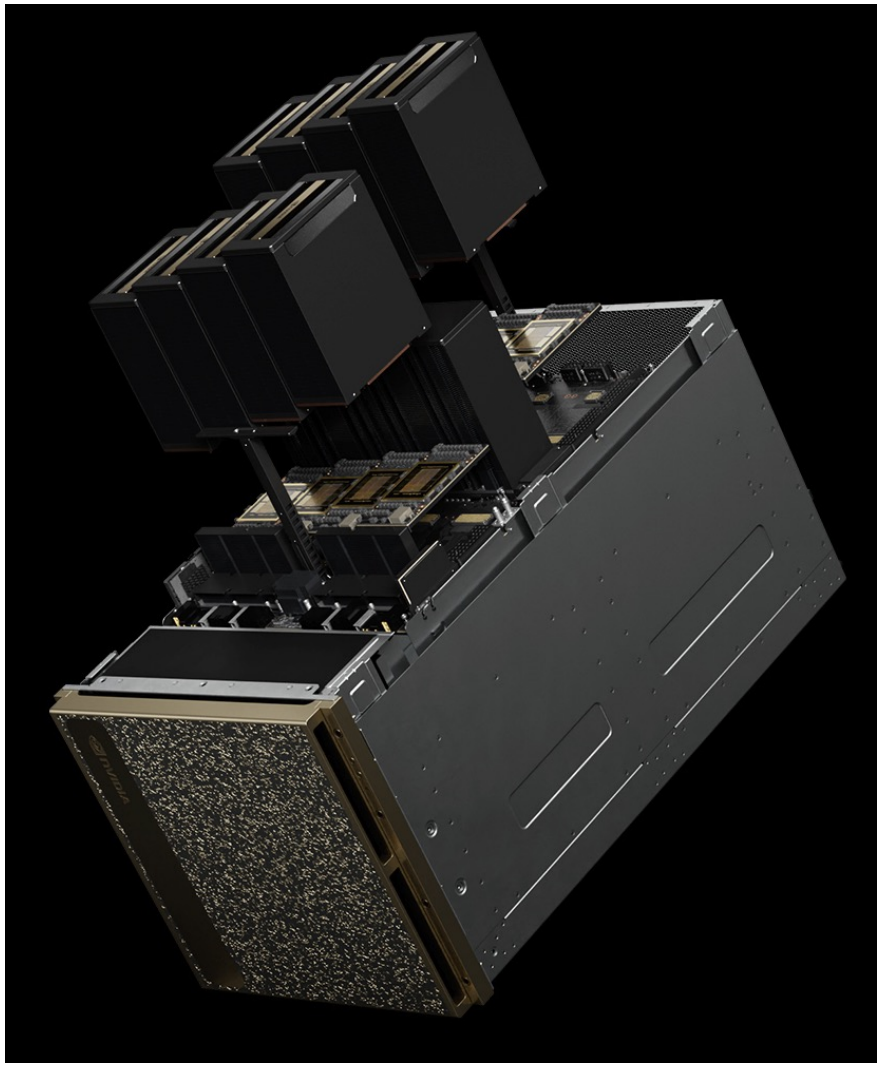
HGX 服务器

HGX 服务器介绍

产品	HGX A100	HGX H100	HGX H200	HGX B100	HGX B200
	8xA100 SXM	8xH100 SXM	8xH200 SXM	8xB100 SXM	8xB200 SXM
架构	Ampere	Hopper		Blackwell	
HBM 大小	640GB	1.1TB		1.44TB/1.5TB	
HBM 带宽	2TB/s	3.35TB/s	4.8TB/s	8TB/s	8TB/s
FP16 (FLOPS)	312T	1P	1P	1.75P	2.25P
INT8 (OPS)	624T	2P	2P	3.5P	4.5P
FP8 (FLOPS)	N	2P	2P	3.5P	4.5P
FP6 (FLOPS)	N	N	N	3.5P	4.5P
FP4 (FLOPS)	N	N	N	7P	9P
GPU-GPU 带宽	600 GB/s	900 GB/s		1.8 TB/s	
NVLink 带宽	4.8 TB/s	7.2 TB/s		14.4 T/s	
Ethernet 带宽	200 Gb/s	400 Gb/s + 20 0Gb/s		2 x 400 Gb/s	
IB 带宽	8 x 200 Gb/s	8 x 400Gb/s		8 x 400 Gb/s	
GPUs Power	3.2kw	5.6kw		5.6kw	8kw
总Power	6.5kw	10.2kw		10.2kw	14.3kw
网络产品	ConnectX-6 NIC	ConnectX-7 NIC		BlueField-3 DPU + ConnectX-7 NIC	

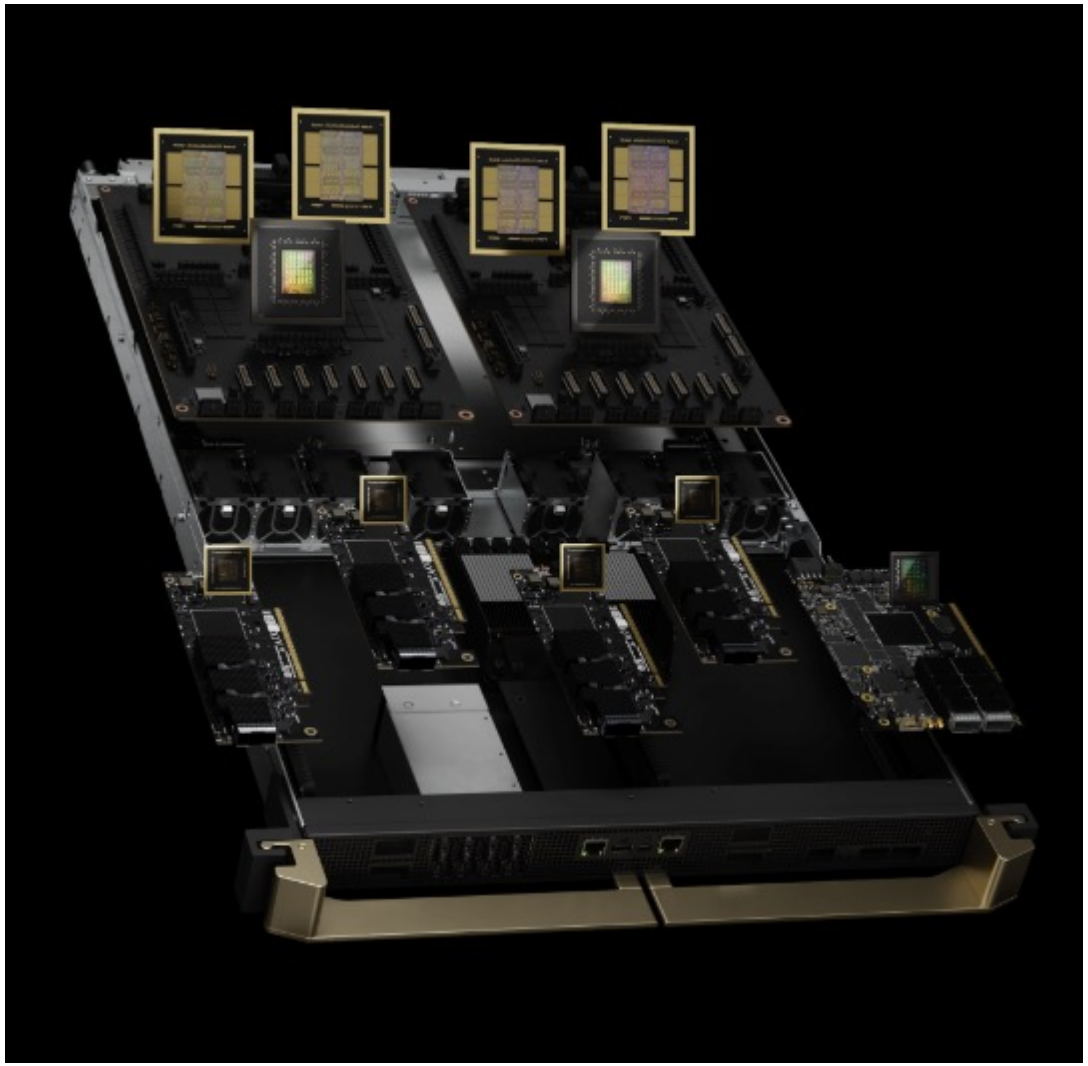
HGX 服务器介绍

产品	HGX B100	HGX B200
	8xB100 SXM	8xB200 SXM
架构	Blackwell	
HBM 大小	1.44TB/1.5TB	
HBM 带宽	8TB/s	8TB/s
FP16 (FLOPS)	1.75P	2.25P
INT8 (OPS)	3.5P	4.5P
FP8 (FLOPS)	3.5P	4.5P
FP6 (FLOPS)	3.5P	4.5P
FP4 (FLOPS)	7P	9P
GPU-GPU 带宽	1.8 TB/s	
NVLink 带宽	14.4 T/s	
Ethernet 带宽	2 x 400 Gb/s	
IB 带宽	8 x 400 Gb/s	
GPUs Power	5.6kw	8kw
总Power	10.2kw	14.3kw
网络产品	BlueField-3 DPU + ConnectX-7 NIC	



HGX 服务器介绍

产品	HGX B100	HGX B200
	8xB100 SXM	8xB200 SXM
架构	Blackwell	
HBM 大小	1.44TB/1.5TB	
HBM 带宽	8TB/s	8TB/s
FP16 (FLOPS)	1.75P	2.25P
INT8 (OPS)	3.5P	4.5P
FP8 (FLOPS)	3.5P	4.5P
FP6 (FLOPS)	3.5P	4.5P
FP4 (FLOPS)	7P	9P
GPU-GPU 带宽	1.8 TB/s	
NVLink 带宽	14.4 T/s	
Ethernet 带宽	2 x 400 Gb/s	
IB 带宽	8 x 400 Gb/s	
GPUs Power	5.6kw	8kw
总Power	10.2kw	14.3kw
网络产品	BlueField-3 DPU + ConnectX-7 NIC	



HGX 服务器介绍

- HGX A100 -> HGX H100 or HGX H200，FP16 算力增加到 3.3 倍，功耗不到原来 2 倍

产品	HGX A100	HGX H100	HGX H200	HGX B100	HGX B200
	8xA100 SXM	8xH100 SXM	8xH200 SXM	8xB100 SXM	8xB200 SXM
架构	Ampere	Hopper		Blackwell	
HBM 大小	640GB	1.1TB		1.44TB/1.5TB	
HBM 带宽	2TB/s	3.35TB/s	4.8TB/s	8TB/s	8TB/s
FP16 (FLOPS)	312T	1P	1P	1.75P	2.25P
INT8 (OPS)	624T	2P	2P	3.5P	4.5P
FP8 (FLOPS)	N	2P	2P	3.5P	4.5P
FP6 (FLOPS)	N	N	N	3.5P	4.5P
FP4 (FLOPS)	N	N	N	7P	9P
GPU-GPU 带宽	600 GB/s	900 GB/s		1.8 TB/s	
NVLink 带宽	4.8 TB/s	7.2 TB/s		14.4 T/s	
Ethernet 带宽	200 Gb/s	400 Gb/s + 20 0Gb/s		2 x 400 Gb/s	
IB 带宽	8 x 200 Gb/s	8 x 400Gb/s		8 x 400 Gb/s	
GPUs Power	3.2kw	5.6kw		5.6kw	8kw
总Power	6.5kw	10.2kw		10.2kw	14.3kw
网络产品	ConnectX-6 NIC	ConnectX-7 NIC		BlueField-3 DPU + ConnectX-7 NIC	

HGX 服务器介绍

- HGX H100 & HGX H200 -> HGX B100 & HGX B200，FP16 算力增加到 2 倍左右，功耗相当

产品	HGX A100	HGX H100	HGX H200	HGX B100	HGX B200
	8xA100 SXM	8xH100 SXM	8xH200 SXM	8xB100 SXM	8xB200 SXM
架构	Ampere	Hopper		Blackwell	
HBM 大小	640GB	1.1TB		1.44TB/1.5TB	
HBM 带宽	2TB/s	3.35TB/s	4.8TB/s	8TB/s	8TB/s
FP16 (FLOPS)	312T	1P	1P	1.75P	2.25P
INT8 (OPS)	624T	2P	2P	3.5P	4.5P
FP8 (FLOPS)	N	2P	2P	3.5P	4.5P
FP6 (FLOPS)	N	N	N	3.5P	4.5P
FP4 (FLOPS)	N	N	N	7P	9P
GPU-GPU 带宽	600 GB/s	900 GB/s		1.8 TB/s	
NVLink 带宽	4.8 TB/s	7.2 TB/s		14.4 T/s	
Ethernet 带宽	200 Gb/s	400 Gb/s + 20 0Gb/s		2 x 400 Gb/s	
IB 带宽	8 x 200 Gb/s	8 x 400Gb/s		8 x 400 Gb/s	
GPUs Power	3.2kw	5.6kw		5.6kw	8kw
总Power	6.5kw	10.2kw		10.2kw	14.3kw
网络产品	ConnectX-6 NIC	ConnectX-7 NIC		BlueField-3 DPU + ConnectX-7 NIC	

HGX 服务器介绍

- HGX B100 & HGX B200 网络基本没有升级，IB 仍然选择 8x400Gb/s

产品	HGX A100	HGX H100	HGX H200	HGX B100	HGX B200
	8xA100 SXM	8xH100 SXM	8xH200 SXM	8xB100 SXM	8xB200 SXM
架构	Ampere	Hopper		Blackwell	
HBM 大小	640GB	1.1TB		1.44TB/1.5TB	
HBM 带宽	2TB/s	3.35TB/s	4.8TB/s	8TB/s	8TB/s
FP16 (FLOPS)	312T	1P	1P	1.75P	2.25P
INT8 (OPS)	624T	2P	2P	3.5P	4.5P
FP8 (FLOPS)	N	2P	2P	3.5P	4.5P
FP6 (FLOPS)	N	N	N	3.5P	4.5P
FP4 (FLOPS)	N	N	N	7P	9P
GPU-GPU 带宽	600 GB/s	900 GB/s		1.8 TB/s	
NVLink 带宽	4.8 TB/s	7.2 TB/s		14.4 T/s	
Ethernet 带宽	200 Gb/s	400 Gb/s + 20 0Gb/s		2 x 400 Gb/s	
IB 带宽	8 x 200 Gb/s	8 x 400Gb/s		8 x 400 Gb/s	
GPUs Power	3.2kw	5.6kw		5.6kw	8kw
总Power	6.5kw	10.2kw		10.2kw	14.3kw
网络产品	ConnectX-6 NIC	ConnectX-7 NIC		BlueField-3 DPU + ConnectX-7 NIC	

03 NVL & SuperPod

NVL & SuperPod 介绍

产品	NVL 32	GH200 SuperPod	NVL72	GB200 SuperPod
	32 x GH200	256 x GH200	36 x GB200	288 x GB200
	32 x GPU	256 GPU	72 GPU	576 x GPU
架构	Hopper		Blackwell	
HBM 大小	32 x 144GB = 4.6T	256 x 96GB = 24.5T	32 x 384GB = 13.8T	288 x 384GB = 110T
LPDDR5X 大小	32 x 480GB = 15.4T	256 x 480GB = 123T	32 x 480GB = 17.3T	288 x 480GB = 138T
HBM 带宽	3.35TB/s	4.8TB/s	8TB/s	8TB/s
FP16 (FLOPS)	32P	256P	180P	1440P
INT8 (OPS)	64P	64P	360P	2880P
FP8 (FLOPS)	64P	64P	360P	2880P
FP6 (FLOPS)			360P	2880P
FP4 (FLOPS)			720P	5760P
GPU-GPU 带宽	0.9TB/s	0.9TB/s	1.8TB/s	1.8TB/s
NVSwitch	Gen3 64 Port/NVSwitch		Gen4 72 Port/NVSwitch	
NVLink 带宽	36 x 0.9T/s=32TB/s	256 x 0.9T/s=230TB/s	72 x 1.8T/s=130TB/s	576 x 1.8T/s=1PB/s
Ethernet 带宽	16 x 200Gb/s	256 x 200Gb/s	18 x 400Gb/s	576 x 400Gb/s
IB 带宽	32 x 400Gb/s	256 x 400Gb/s	72 x 800Gb/s	576 x 800Gb/s
GPUs Power	32 x 1kw=32kw	256 x 1kw=256kw	36 x 2.7kw=97.2kw	288 x 2.7kw=777.6kw
网络产品	ConnectX-7 NIC		ConnectX-8 NIC	

NVL & SuperPod 介绍

产品	NVL72	GB200 SuperPod
	36 x GB200	288 x GB200
	72 GPU	576 x GPU
架构	Blackwell	
HBM 大小	32 x 384GB = 13.8T	288 x 384GB = 110T
LPDDR5X 大小	32 x 480GB = 17.3T	288 x 480GB = 138T
HBM 带宽	8TB/s	8TB/s
FP16 (FLOPS)	180P	1440P
INT8 (OPS)	360P	2880P
FP8 (FLOPS)	360P	2880P
FP6 (FLOPS)	360P	2880P
FP4 (FLOPS)	720P	5760P
GPU-GPU 带宽	1.8TB/s	1.8TB/s
NVSwitch	Gen4 72 Port/NVSwitch	
NVLink 带宽	72 x 1.8T/s=130TB/s	576 x 1.8T/s=1PB/s
Ethernet 带宽	18 x 400Gb/s	576 x 400Gb/s
IB 带宽	72 x 800Gb/s	576 x 800Gb/s
GPUs Power	36 x 2.7kw=97.2kw	288 x 2.7kw=777.6kw
网络产品	ConnectX-8 NIC	

NVL & SuperPod 介绍

- NVL32 -> NVL72，GPU 数从 32 增加到 72，FP16 从 32P 到 180P (近 6 倍)

产品	NVL 32	GH200 SuperPod	NVL72	GB200 SuperPod
	32 x GH200	256 x GH200	36 x GB200	288 x GB200
	32 x GPU	256 GPU	72 GPU	576 x GPU
架构	Hopper		Blackwell	
HBM 大小	32 x 144GB = 4.6T	256 x 96GB = 24.5T	32 x 384GB = 13.8T	288 x 384GB = 110T
LPDDR5X 大小	32 x 480GB = 15.4T	256 x 480GB = 123T	32 x 480GB = 17.3T	288 x 480GB = 138T
HBM 带宽	3.35TB/s	4.8TB/s	8TB/s	8TB/s
FP16 (FLOPS)	32P	256P	180P	1440P
INT8 (OPS)	64P	64P	360P	2880P
FP8 (FLOPS)	64P	64P	360P	2880P
GPU-GPU 带宽	0.9TB/s	0.9TB/s	1.8TB/s	1.8TB/s
NVSwitch	Gen3 64 Port/NVSwitch		Gen4 72 Port/NVSwitch	
NVLink 带宽	36 x 0.9T/s=32TB/s	256 x 0.9T/s=230TB/s	72 x 1.8T/s=130TB/s	576 x 1.8T/s=1PB/s
Ethernet 带宽	16 x 200Gb/s	256 x 200Gb/s	18 x 400Gb/s	576 x 400Gb/s
IB 带宽	32 x 400Gb/s	256 x 400Gb/s	72 x 800Gb/s	576 x 800Gb/s
GPUs Power	32 x 1kw=32kw	256 x 1kw=256kw	36 x 2.7kw=97.2kw	288 x 2.7kw=777.6kw
网络产品	ConnectX-7 NIC		ConnectX-8 NIC	

NVL & SuperPod 介绍

- GH200 SuperPod -> GB200 SuperPod，GPU 数从 256 到 576，FP16 从 256P 到 1440P (近 6 倍)

产品	NVL 32	GH200 SuperPod	NVL72	GB200 SuperPod
	32 x GH200	256 x GH200	36 x GB200	288 x GB200
	32 x GPU	256 GPU	72 GPU	576 x GPU
架构	Hopper		Blackwell	
HBM 大小	32 x 144GB = 4.6T	256 x 96GB = 24.5T	32 x 384GB = 13.8T	288 x 384GB = 110T
LPDDR5X 大小	32 x 480GB = 15.4T	256 x 480GB = 123T	32 x 480GB = 17.3T	288 x 480GB = 138T
HBM 带宽	3.35TB/s	4.8TB/s	8TB/s	8TB/s
FP16 (FLOPS)	32P	256P	180P	1440P
INT8 (OPS)	64P	64P	360P	2880P
FP8 (FLOPS)	64P	64P	360P	2880P
GPU-GPU 带宽	0.9TB/s	0.9TB/s	1.8TB/s	1.8TB/s
NVSwitch	Gen3 64 Port/NVSwitch		Gen4 72 Port/NVSwitch	
NVLink 带宽	36 x 0.9T/s=32TB/s	256 x 0.9T/s=230TB/s	72 x 1.8T/s=130TB/s	576 x 1.8T/s=1PB/s
Ethernet 带宽	16 x 200Gb/s	256 x 200Gb/s	18 x 400Gb/s	576 x 400Gb/s
IB 带宽	32 x 400Gb/s	256 x 400Gb/s	72 x 800Gb/s	576 x 800Gb/s
GPUs Power	32 x 1kw=32kw	256 x 1kw=256kw	36 x 2.7kw=97.2kw	288 x 2.7kw=777.6kw
网络产品	ConnectX-7 NIC		ConnectX-8 NIC	

NVL & SuperPod 介绍

- NVL72 & GB200 SuperPod 采用 ConnectX-8 IB 网卡，带宽 800Gb/s

产品	NVL 32	GH200 SuperPod	NVL72	GB200 SuperPod
	32 x GH200	256 x GH200	36 x GB200	288 x GB200
	32 x GPU	256 GPU	72 GPU	576 x GPU
架构	Hopper		Blackwell	
HBM 大小	32 x 144GB = 4.6T	256 x 96GB = 24.5T	32 x 384GB = 13.8T	288 x 384GB = 110T
LPDDR5X 大小	32 x 480GB = 15.4T	256 x 480GB = 123T	32 x 480GB = 17.3T	288 x 480GB = 138T
HBM 带宽	3.35TB/s	4.8TB/s	8TB/s	8TB/s
FP16 (FLOPS)	32P	256P	180P	1440P
INT8 (OPS)	64P	64P	360P	2880P
FP8 (FLOPS)	64P	64P	360P	2880P
GPU-GPU 带宽	0.9TB/s	0.9TB/s	1.8TB/s	1.8TB/s
NVSwitch	Gen3 64 Port/NVSwitch		Gen4 72 Port/NVSwitch	
NVLink 带宽	36 x 0.9T/s=32TB/s	256 x 0.9T/s=230TB/s	72 x 1.8T/s=130TB/s	576 x 1.8T/s=1PB/s
Ethernet 带宽	16 x 200Gb/s	256 x 200Gb/s	18 x 400Gb/s	576 x 400Gb/s
IB 带宽	32 x 400Gb/s	256 x 400Gb/s	72 x 800Gb/s	576 x 800Gb/s
GPUs Power	32 x 1kw=32kw	256 x 1kw=256kw	36 x 2.7kw=97.2kw	288 x 2.7kw=777.6kw
网络产品	ConnectX-7 NIC		ConnectX-8 NIC	

04

总结与思考

NVIDIA GPU架构发展

- B200、B100、GB200、NVL72、NVL32、SuperPod、GH200、H200、H100、L20、SuperPod-576
- ConnectX-800G 网卡、网络交换机

Thank you

把AI系统带入每个开发者、每个家庭、每个组织，构建万物互联的智能世界

Bring AI System to every person, home and organization for a fully connected, intelligent world.

ZOMI

Course chenzomi12.github.io

GitHub https://github.com/chenzomi12/AIFoundation

Reference 参考&引用

1. https://www.fibermall.com/blog/nvidia-b100-b200-gh200-nvl72-superpod.htm