NVIDIA GPU Product Lines

Hopper to Blackwell

20251128 Allen Sun

🧠 NVIDIA Industry Terminology Hierarchy

Tier	Common Name	Examples	Description
Consumer GPU	GeForce RTX	RTX 5090 / RTX 5080	Gaming, content creation, AI development entry
Professional Workstation GPU	RTX Professional	RTX 6000 Ada / RTX PRO 6000	Professional graphics, engineering design, AI dev environments
Data Center GPU	NVIDIA Data Center GPU or "A/H/B Series GPU"	A100 / H100 / B200	AI training, inference, large model serving
GPU Module Platform	HGX	HGX H100 / HGX B200	Server GPU baseboard (SXM version)
Complete System / Supercomputing Platform	DGX / GB200 / NVL72	DGX H100 / GB200 NVL72	Complete AI supercomputing solutions

Power Distinction (TDP): SXM versions (HGX) typically have much higher power consumption than PCIe versions (e.g., H100 SXM is 700W, PCIe is 300-350W), B200 SXM: **~1000W** (40% higher than H100).

🧱 Common Industry Communication Terminology

Common Term	Actual Meaning	Corresponding Technical Level
RTX	Generally refers to workstation or consumer GPUs	RTX 5090 / RTX PRO 6000, etc.
PCIe GPU	Standard motherboard slot GPU	RTX series A100 PCIe H100 PCIe B200 PCIe
Data Center GPU	Generally refers to SXM / NVLink version GPUs	A100 / H100 / B200 2020 (A)Ampere 2022 (H)Hopper 2024 (B)Blackwell
HGX Platform	Server module with multiple SXM GPUs	HGX H100 HGX B200
DGX System	NVIDIA's complete server system (includes CPU, storage, networking)	DGX H100 DGX B200
NVL72 / GB200 Cluster	Ultra-large-scale AI system (multi-DGX structure)	36 CPU: Grace 72 GPU: Blackwell

📊 Real Industry Usage Examples (Engineer Colloquialisms)

Statement	Meaning
This server has two RTX 6000 Ada cards	Workstation / small AI inference machine, PCIe interface
We're training on H100 HGX	Using SXM + NVLink platform for distributed training
We're running 8-card A100	Common DGX A100 configuration
Client wants B200 PCIe	Wants inference/acceleration on standard server motherboard
That's a DGX H100 rack	Complete NVIDIA official system
GB200 NVL72 cluster	Refers to large-scale cluster architecture integrating Grace CPU + Blackwell GPU

🧭 Summary in One Statement:

- 🔹 "RTX" is the umbrella term for workstation and consumer products
- 🔹 "A/H/B Series (Data Center GPU)" is the umbrella term for server-grade GPUs
- 🔹 "HGX" is the server GPU baseboard
- 🔹 "DGX" is the complete server system
- 🔹 "SXM GPU" supports NVLink, "PCIe GPU" does not support NVLink
- 🔹 In Hopper and Blackwell generations, only SXM versions support high-speed NVLink interconnect

🔬 Core Specifications Quick Reference

Item	A100	H100	B200
Architecture	Ampere	Hopper	Blackwell
Process	7nm	4nm	4nm (TSMC)
Memory	40/80GB HBM2e	80GB HBM3	192GB HBM3e
FP8 Performance	-	~2000 TFLOPS	~4500 TFLOPS
NVLink	600GB/s	900GB/s	1800GB/s
Typical Price (USD)	~$15K	~$30K	~$40K (estimated)

🎯 How to Choose a GPU?

- **Training GPT-4 scale models** → GB200 NVL72
- **Training 7B-70B models** → HGX H100 (8-card)
- **LLM inference services** → H100/B200 PCIe
- **Fine-tuning / RAG applications** → A100 or RTX 6000 Ada
- **Development & testing** → RTX 4090 / 5090

🚀 Future Outlook

With a solid grasp of these core concepts and terminology hierarchies, navigating the upcoming Rubin architecture and subsequent technical roadmaps will become much more intuitive, allowing you to quickly understand NVIDIA's evolving blueprint in the AI computing landscape.