腾讯程序员

英伟达GPU全系列硬核科普手册

一文读懂NVIDIA芯片的定位、规格与应用场景

作者：nickycao

核心逻辑只有三条：

- 显存 — 决定模型装不装得下
- 算力 — 决定算得快不快
- 互连带宽 — 决定多卡扩不扩得动

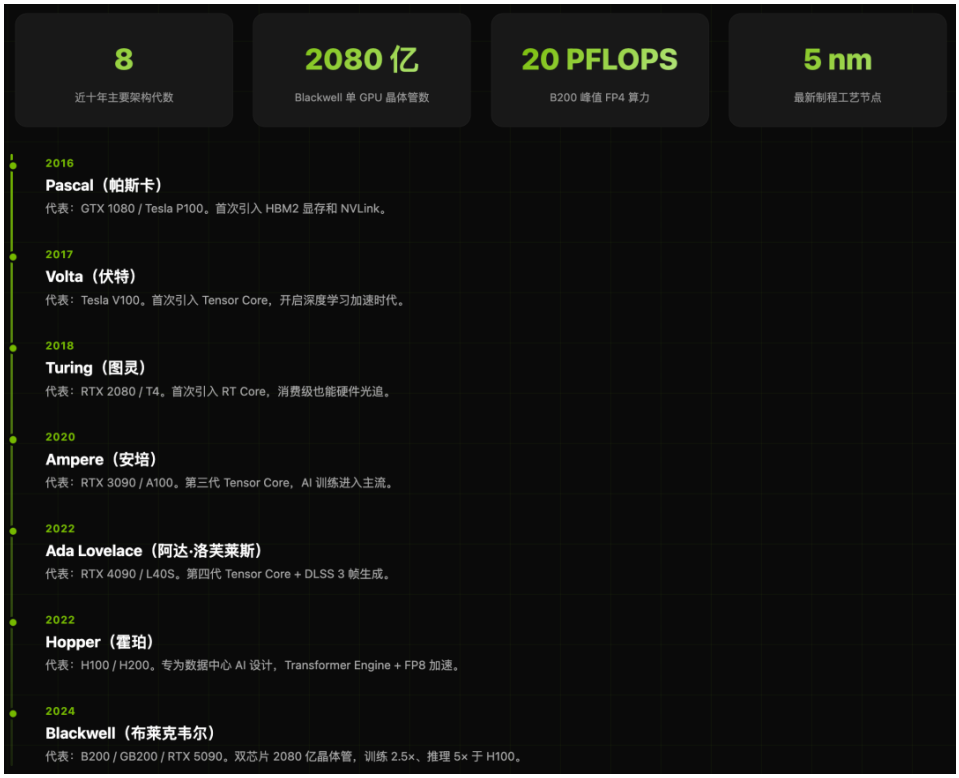
目 录

阅读导图与学习路径

这篇不是一口气必须读完的论文，而是一份按需查阅的 GPU 百科；先抓住产品线主线，再按你的角色和场景深入感兴趣的板块。

架构演进史

NVIDIA 每隔约 2 年推出新一代 GPU 架构，每一代都带来计算能力的质变。



GPU 命名规则解读

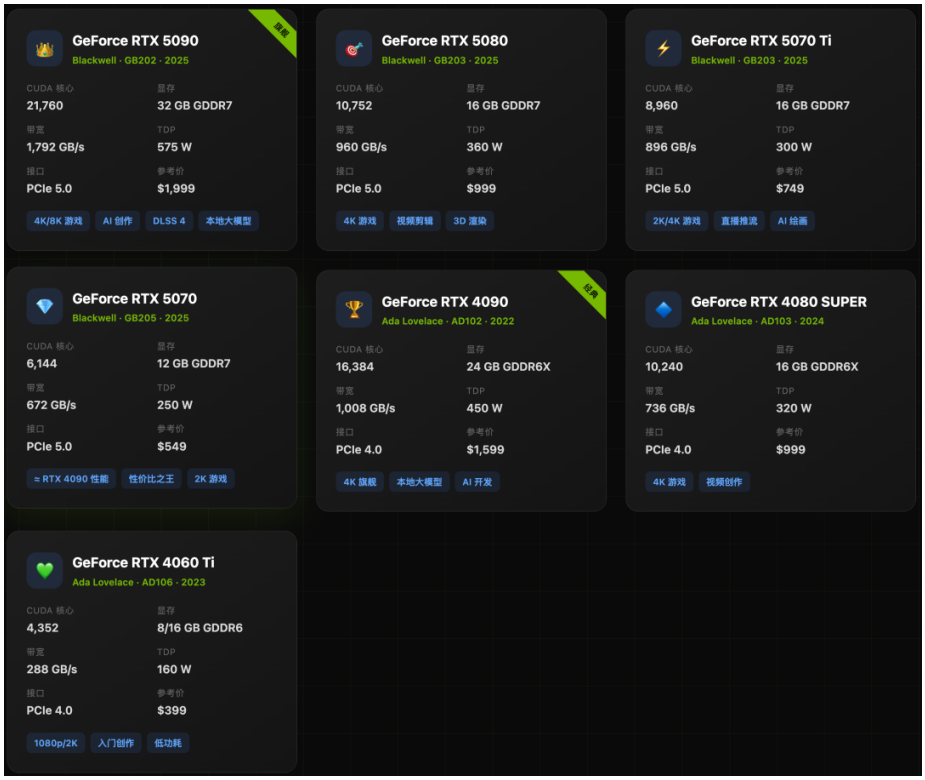
看似复杂的型号其实藏着清晰逻辑。学会「拆字」，一眼识别定位。

核心概念扫盲

搞懂这些关键术语，后面的参数不再陌生。

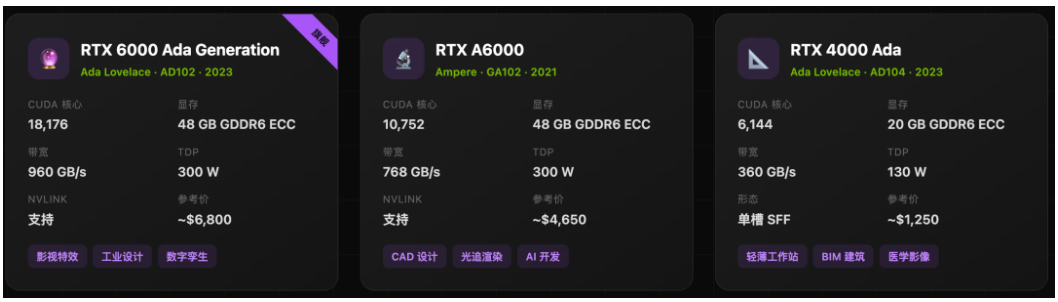
消费级 GPU（GeForce 系列）

面向游戏玩家与创作者。性价比高、驱动生态丰富，支持光追和 DLSS。



专业工作站 GPU

面向 CAD/CAE、影视特效、科学可视化。大显存 ECC、ISV 认证、长周期驱动。

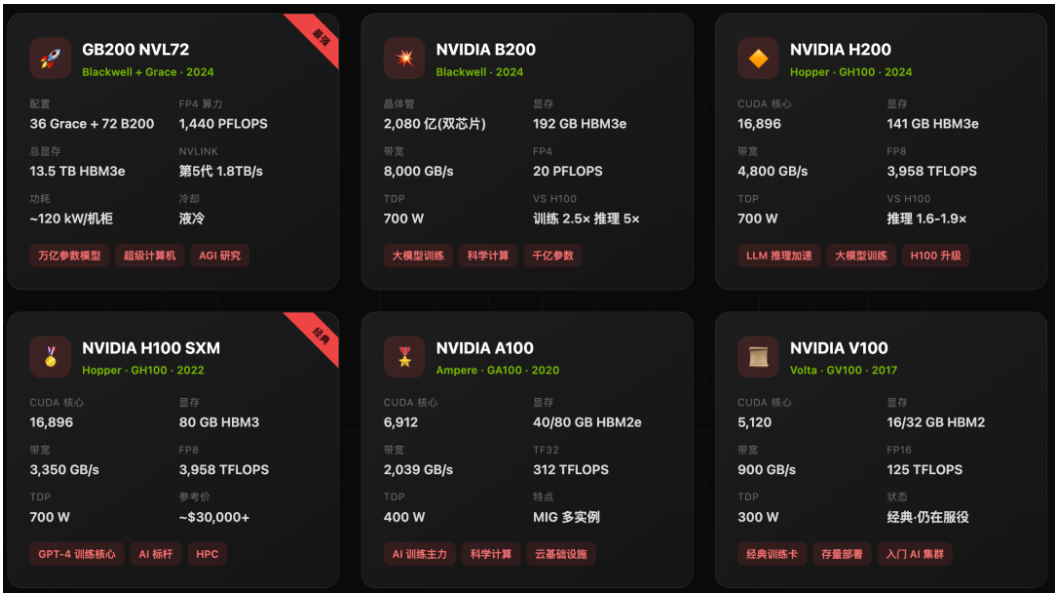


推理加速卡（数据中心）

专为 AI 推理优化，追求高吞吐、低延迟和低功耗。承担 ChatGPT 等服务的实际运算。

训练旗舰（数据中心）

AI 大模型训练核心。GPT-4、Llama 等大模型背后靠成千上万张这类卡协同工作。



中国特供版 **GPU**

受美国出口管制限制，NVIDIA 为中国市场定制的合规版本。核心规格有所调整，但仍是国内 AI 产业的主力芯片。

出口管制背景：

自 2022 年起，美国陆续限制高端 AI 芯片对华出口。NVIDIA 通过降低互连带宽（NVLink）、削减 CUDA 核心数或限制算力精度等方式，推出符合管制要求的「中国特供版」。这些芯片在中国 AI 大模型训练和推理中仍发挥关键作用。命名规律：

数据中心卡在原型号数字修改为 「800」（如 A100→A800、H100→H800），或推出全新编号（H20、L20）；

消费级在型号后加「D」（如 RTX 4090D）。

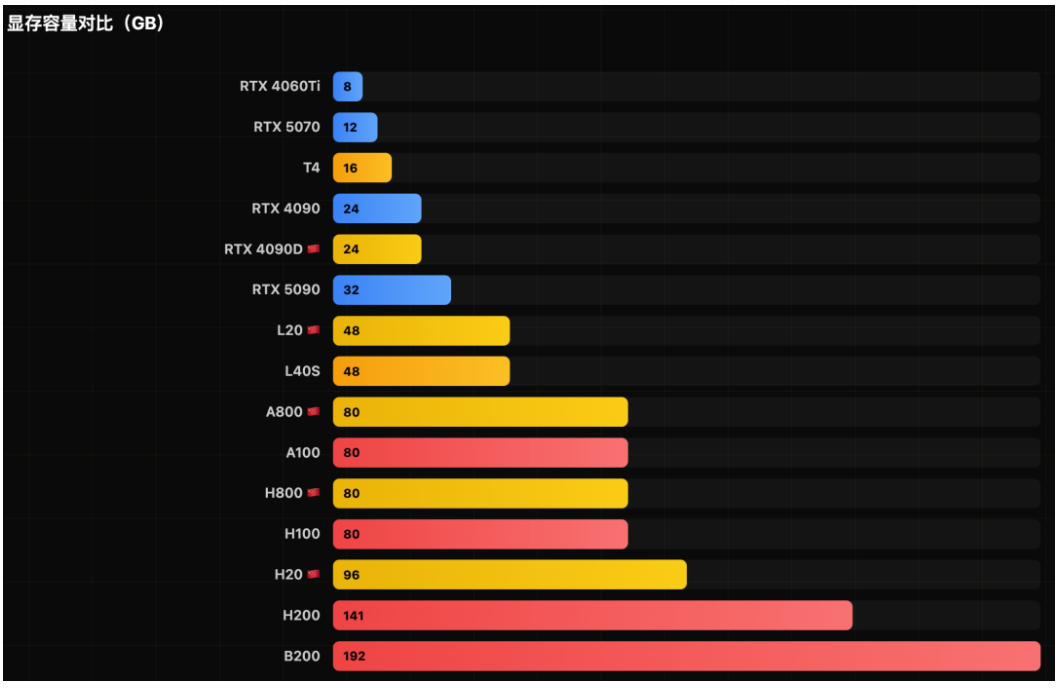
横向对比总表

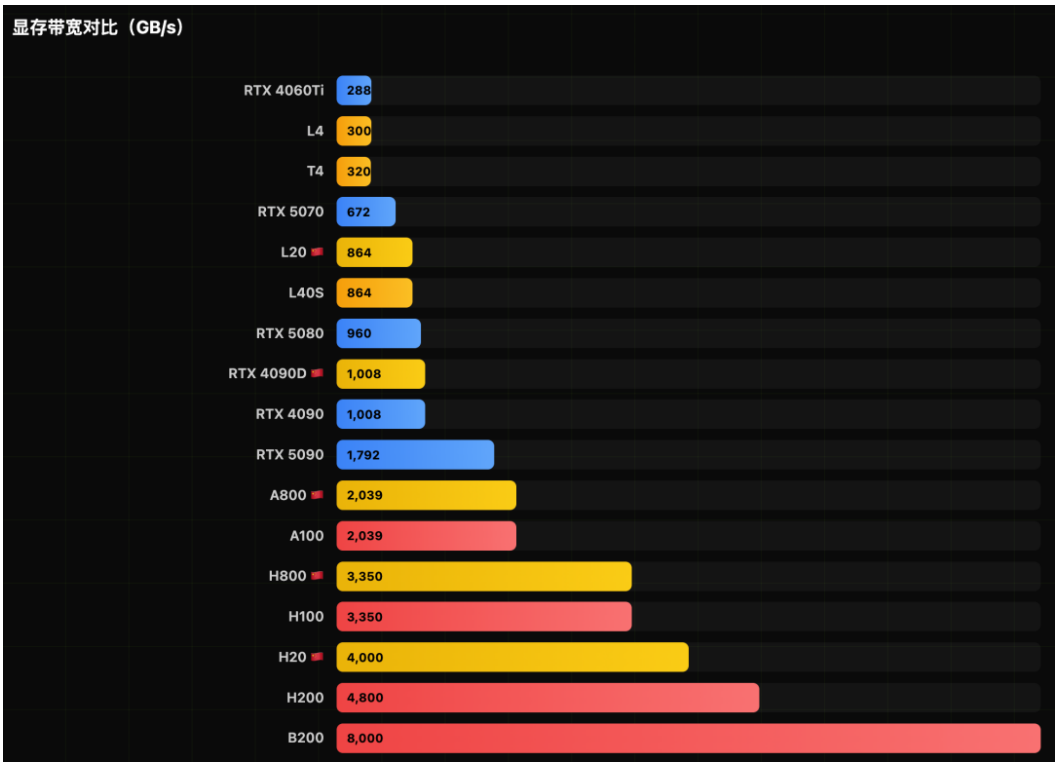
一张表快速对比关键参数差异。

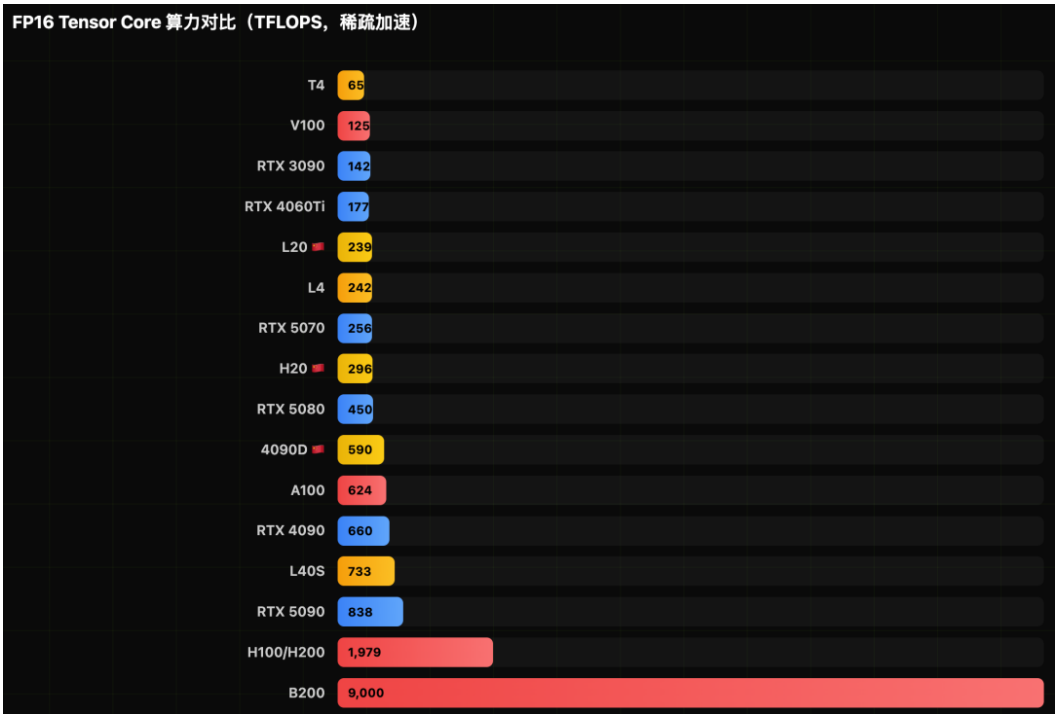
型号	类别	架构	CUDA	显存	带宽	TDP	定位
RTX 5090	消费	Blackwell	21,760	32GB GDDR7	1,792 GB/s	575W	旗舰游戏+AI
RTX 5080	消费	Blackwell	10,752	16GB GDDR7	960 GB/s	360W	高端游戏
RTX 5070	消费	Blackwell	6,144	12GB GDDR7	672 GB/s	250W	甜品≈4090
RTX 4090	消费	Ada Lovelace	16,384	24GB GDDR6X	1,008 GB/s	450W	上代旗舰
RTX 6000 Ada	工作站	Ada Lovelace	18,176	48GB GDDR6	960 GB/s	300W	专业渲染
RTX A6000	工作站	Ampere	10,752	48GB GDDR6	768 GB/s	300W	专业设计
L40S	推理	Ada Lovelace	18,176	48GB GDDR6	864 GB/s	350W	全能推理
L4	推理	Ada Lovelace	7,424	24GB GDDR6	300 GB/s	72W	低功耗推理
T4	推理	Turing	2,560	16GB GDDR6	320 GB/s	70W	经典推理
B200	训练	Blackwell	—	192GB HBM3e	8,000 GB/s	700W	顶级训练
H100	训练	Hopper	16,896	80GB HBM3	3,350 GB/s	700W	AI 标杆
H200	训练	Hopper	16,896	141GB HBM3e	4,800 GB/s	700W	H100 增强
A100	训练	Ampere	6,912	80GB HBM2e	2,039 GB/s	400W	经典主力
H20	中国版	Hopper	~9,984	96GB HBM3	4,000 GB/s	400W	中国市场主力
H800	中国版	Hopper	16,896	80GB HBM3	3,350 GB/s	700W	H100 合规版
A800	中国版	Ampere	6,912	80GB HBM2e	2,039 GB/s	300W	A100 合规版
L20	中国版	Ada Lovelace	11,776	48GB GDDR6	864 GB/s	275W	推理合规版
RTX 4090D	中国版	Ada Lovelace	14,592	24GB GDDR6X	1,008 GB/s	425W	4090 合规版

性能可视化

图形直观感受关键指标差异。

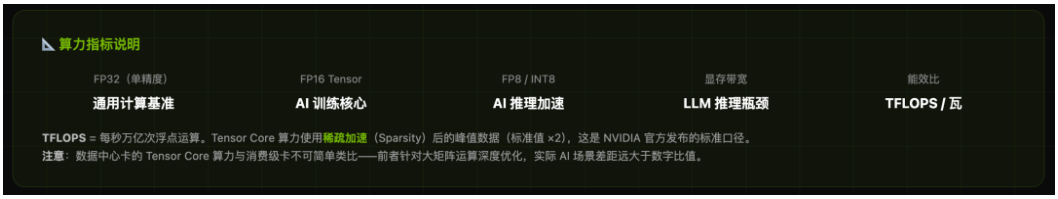






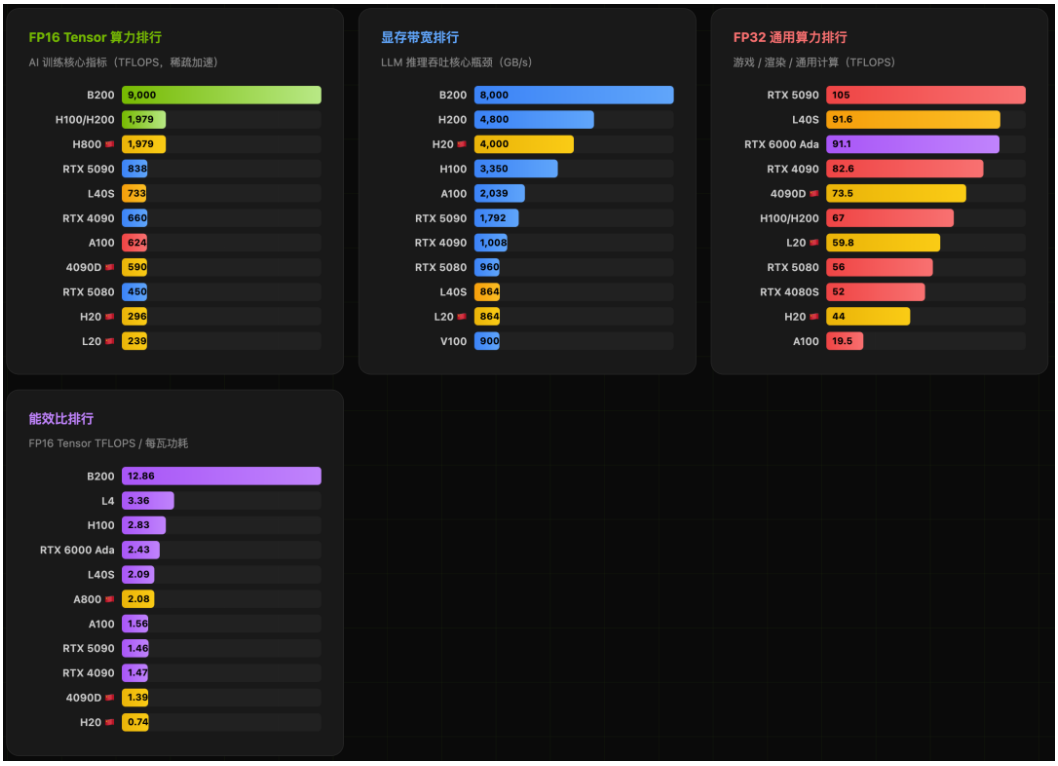
GPU 算力指标对比与排行

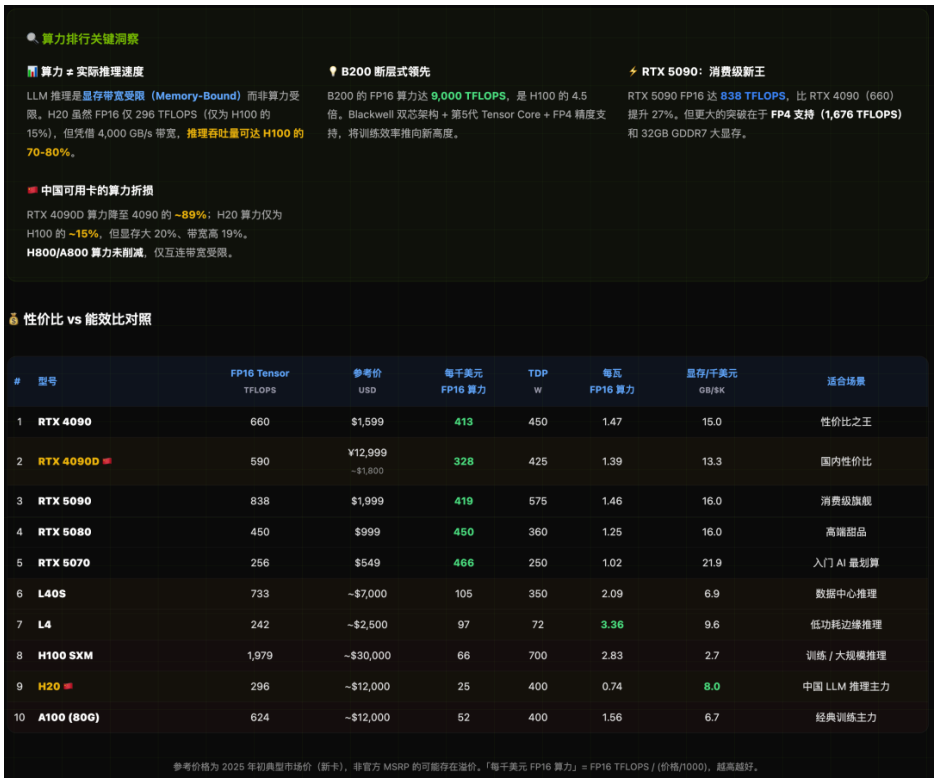
算力是衡量 GPU 实力的核心指标。这里将所有主流 GPU 的关键算力数据汇总排列，涵盖从消费级到数据中心的完整图谱。



排名按 FP16 Tensor Core 算力（稀疏加速峰值）降序排列。H800/A800 算力与原版相同（仅互连降速），故单独列出。B200 FP16 数据为 Blackwell 双芯架构峰值。

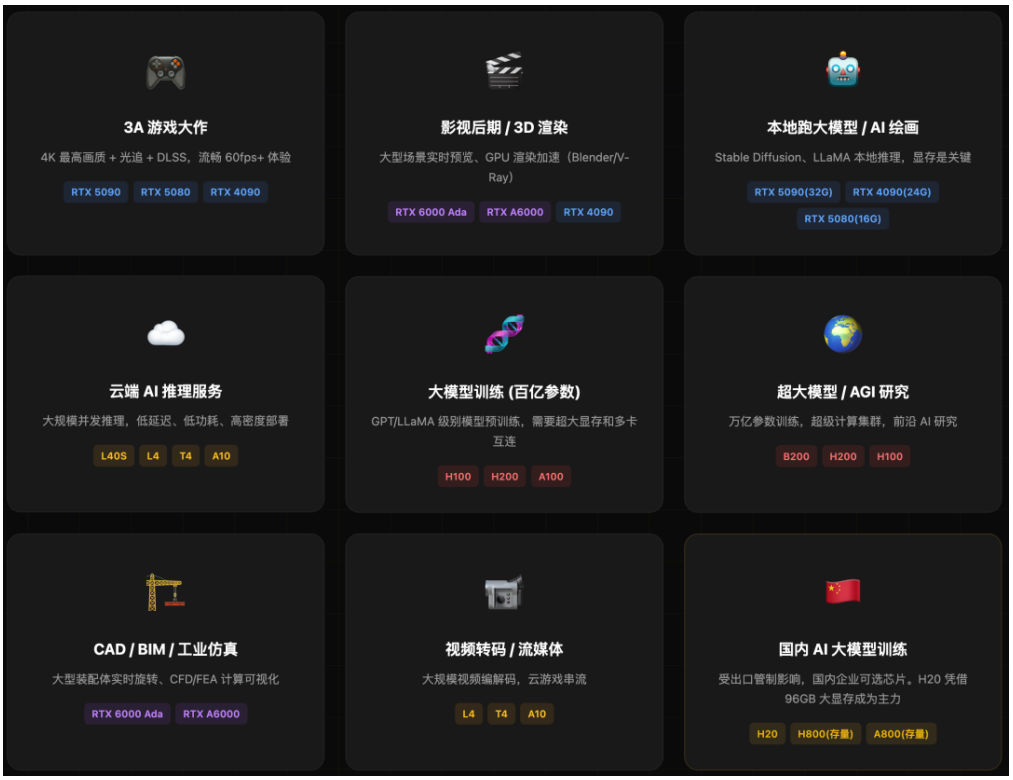
#	型号	类别	架构	FP32 TFLOPS	FP16 Tensor TFLOPS（稀疏）	FP8/INT8 TFLOPS/TOPS	显存 容量	带宽 GB/s	TDP W	能效比 FP16/W
1	B200 训练	Blackwell	2024	—	9,000	18,000 (FP4)	192 GB HBM3e	8,000	700	12.86
2	H200 训练	Hopper	2024	67	1,979	3,958 (FP8)	141 GB HBM3e	4,800	700	2.83
3	H100 SXM 训练	Hopper	2022	67	1,979	3,958 (FP8)	80 GB HBM3	3,350	700	2.83
4	L40S 推理	Ada Lovelace	2023	91.6	733	733 (FP8)	48 GB GDDR6	864	350	2.09
5	RTX 5090 消费	Blackwell	2025	105	838	1,676 (FP4)	32 GB GDDR7	1,792	575	1.46
6	RTX 6000 Ada 工作站	Ada Lovelace	2023	91.1	728	728 (FP8)	48 GB GDDR6	960	300	2.43
7	RTX 4090 消费	Ada Lovelace	2022	82.6	660	660 (FP8)	24 GB GDDR6X	1,008	450	1.47
8	A100 (80GB) 训练	Ampere	2020	19.5	624	624 (INT8)	80 GB HBM2e	2,039	400	1.56
9	RTX 5080 消费	Blackwell	2025	56	450	900 (FP4)	16 GB GDDR7	960	360	1.25
10	L4 推理	Ada Lovelace	2023	30.3	242	485 (INT8)	24 GB GDDR6	300	72	3.36
11	RTX 4090D 中国版	Ada Lovelace	2023	73.5	590	590 (FP8)	24 GB GDDR6X	1,008	425	1.39
12	RTX 5070 Ti 消费	Blackwell	2025	47	376	752 (FP4)	16 GB GDDR7	896	300	1.25
13	RTX A6000 工作站	Ampere	2021	38.7	310	310 (INT8)	48 GB GDDR6	768	300	1.03
14	H20 中国版	Hopper	2024	44	296	296 (FP8)	96 GB HBM3	4,000	400	0.74
15	RTX 5070 消费	Blackwell	2025	32	256	512 (FP4)	12 GB GDDR7	672	250	1.02
16	RTX 4080S 消费	Ada Lovelace	2024	52	418	418 (FP8)	16 GB GDDR6X	736	320	1.31
17	L20 中国版	Ada Lovelace	2024	59.8	239	239 (FP8)	48 GB GDDR6	864	275	0.87
18	H800 中国版	Hopper	2022	67	1,979	3,958 (FP8)	80 GB HBM3	3,350	700	2.83
19	A800 中国版	Ampere	2022	19.5	624	624 (INT8)	80 GB HBM2e	2,039	300	2.08
20	V100 (32GB) 训练	Volta	2017	15.7	125	—	32 GB HBM2	900	300	0.42
21	RTX 3090 消费	Ampere	2020	35.6	142	284 (INT8)	24 GB GDDR6X	936	350	0.41
22	RTX 4060 Ti 消费	Ada Lovelace	2023	22.1	177	177 (FP8)	8 GB GDDR6	288	160	1.11
23	T4 推理	Turing	2019	8.1	65	130 (INT8)	16 GB GDDR6	320	70	0.93
24	A10 推理	Ampere	2021	31.2	125	250 (INT8)	24 GB GDDR6	600	150	0.83





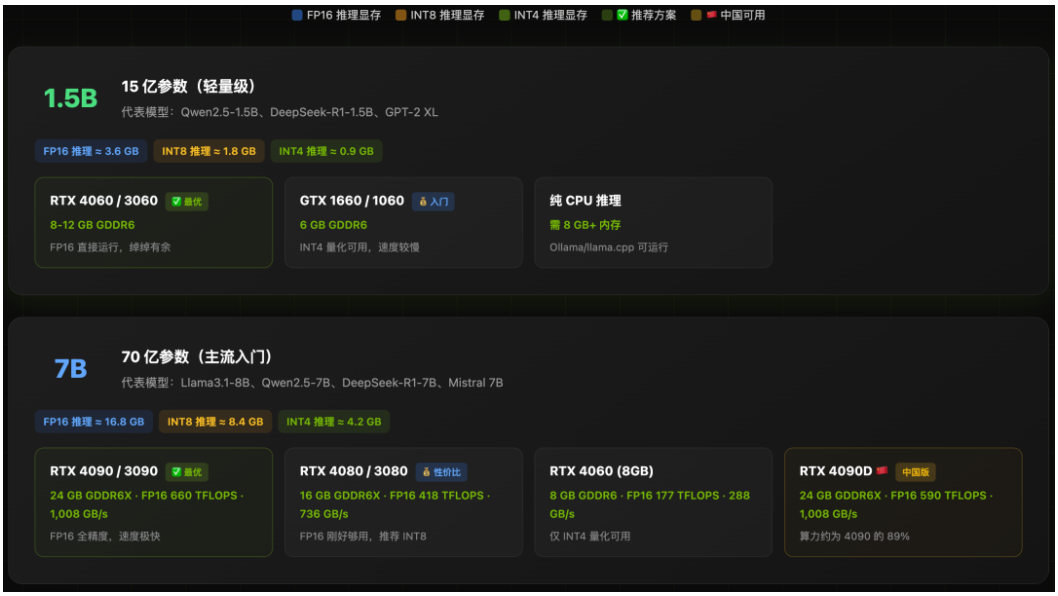
应用场景推荐

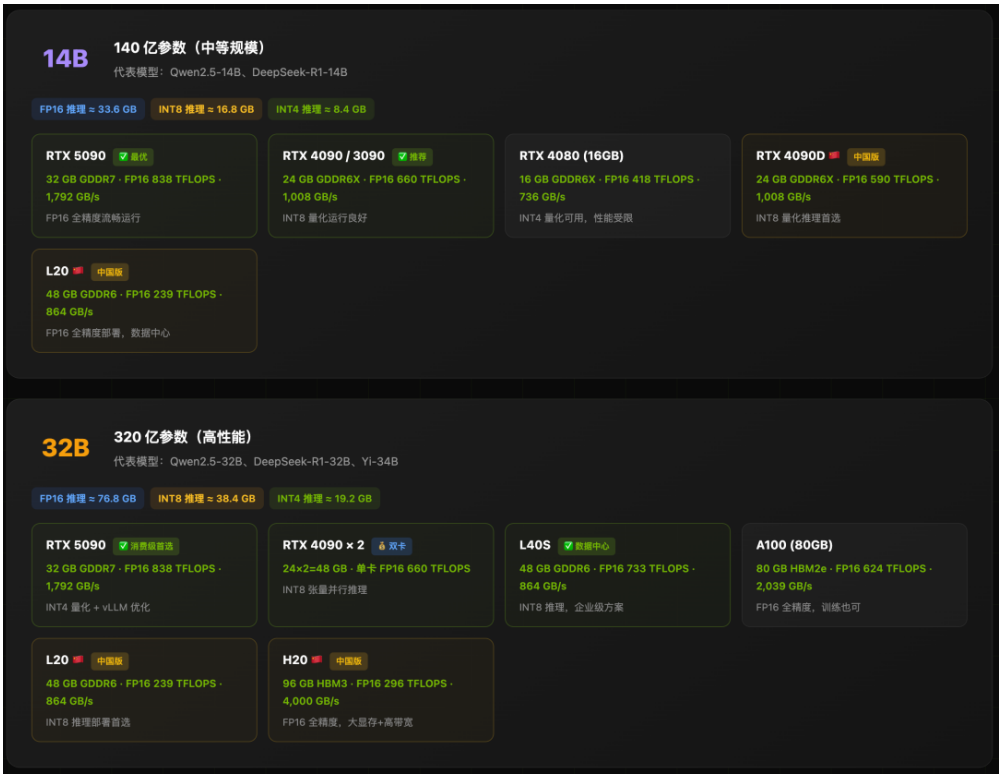
根据你的实际需求，找到最适合的 GPU。

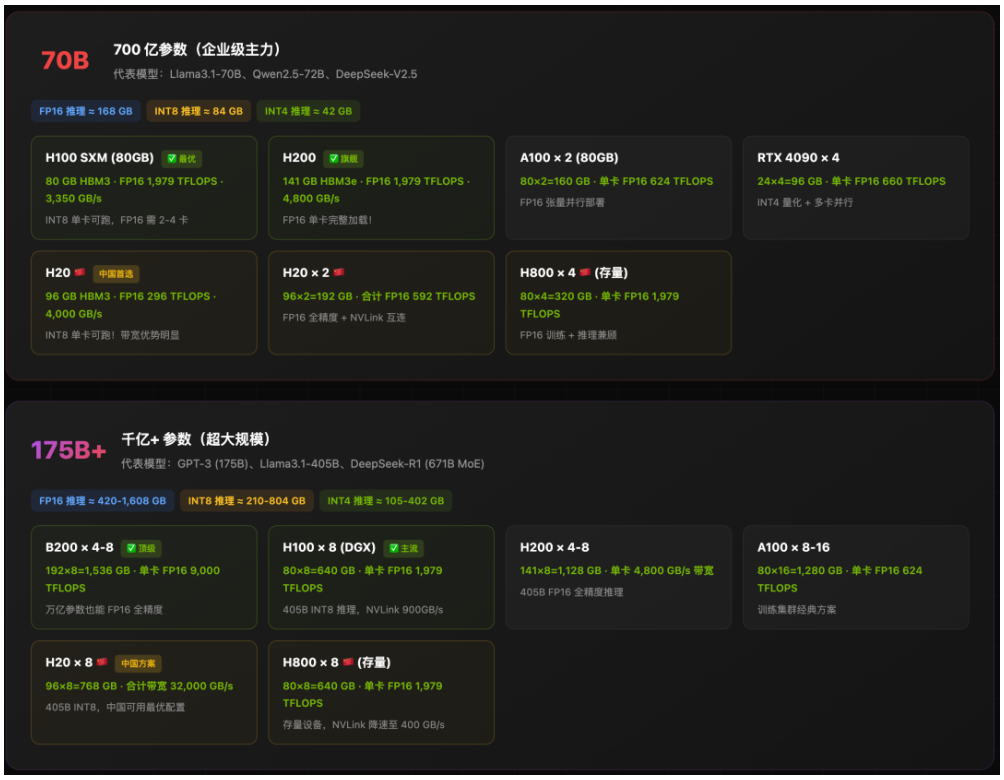


模型参数 × GPU 选择对照

不同参数规模的大模型（LLM），在推理和训练场景下对 GPU 显存的需求差异巨大。掌握估算方法，精准匹配硬件。


FP16 推理显存
INT8 推理显存
INT4 推理显存
推荐方案
中国可用
1.5B
15 亿参数（轻量级）
代表模型：Qwen2.5-1.5B、DeepSeek-R1-1.5B、GPT-2 XL
FP16 推理 ≈ 3.6 GB
INT8 推理 ≈ 1.8 GB
INT4 推理 ≈ 0.9 GB
RTX 4060 / 3060
最优
8-12 GB GDDR6
FP16 直接运行，绰绰有余
GTX 1660 / 1060
入门
6 GB GDDR6
INT4 量化可用，速度较慢
纯 CPU 推理
需 8 GB+ 内存
Ollama/llama.cpp 可运行
7B
70 亿参数（主流入门）
代表模型：Llama3.1-8B、Qwen2.5-7B、DeepSeek-R1-7B、Mistral 7B
FP16 推理 ≈ 16.8 GB
INT8 推理 ≈ 8.4 GB
INT4 推理 ≈ 4.2 GB
RTX 4090 / 3090
最优
24 GB GDDR6X · FP16 660 TFLOPS · 1,008 GB/s
FP16 全精度，速度极快
RTX 4080 / 3080
性价比
16 GB GDDR6X · FP16 418 TFLOPS · 736 GB/s
FP16 刚好够用，推荐 INT8
RTX 4060 (8GB)
8 GB GDDR6 · FP16 177 TFLOPS · 288 GB/s
仅 INT4 量化可用
RTX 4090D
中国版
24 GB GDDR6X · FP16 590 TFLOPS · 1,008 GB/s
算力约为 4090 的 89%


14B
140 亿参数（中等规模）
代表模型：Qwen2.5-14B、DeepSeek-R1-14B
FP16 推理 ≈ 33.6 GB
INT8 推理 ≈ 16.8 GB
INT4 推理 ≈ 8.4 GB
RTX 5090
最优
32 GB GDDR7 · FP16 838 TFLOPS · 1,792 GB/s
FP16 全精度流畅运行
RTX 4090 / 3090
推荐
24 GB GDDR6X · FP16 660 TFLOPS · 1,008 GB/s
INT8 量化运行良好
RTX 4080 (16GB)
16 GB GDDR6X · FP16 418 TFLOPS · 736 GB/s
INT4 量化可用，性能受限
RTX 4090D
中国版
24 GB GDDR6X · FP16 590 TFLOPS · 1,008 GB/s
INT8 量化推理首选
L20
中国版
48 GB GDDR6 · FP16 239 TFLOPS · 864 GB/s
FP16 全精度部署，数据中心
32B
320 亿参数（高性能）
代表模型：Qwen2.5-32B、DeepSeek-R1-32B、Yi-34B
FP16 推理 ≈ 76.8 GB
INT8 推理 ≈ 38.4 GB
INT4 推理 ≈ 19.2 GB
RTX 5090
消费级首选
32 GB GDDR7 · FP16 838 TFLOPS · 1,792 GB/s
INT4 量化 + vLLM 优化
RTX 4090 × 2
双卡
24×2=48 GB · 单卡 FP16 660 TFLOPS
INT8 张量并行推理
L40S
数据中心
48 GB GDDR6 · FP16 733 TFLOPS · 864 GB/s
INT8 推理，企业级方案
A100 (80GB)
80 GB HBM2e · FP16 624 TFLOPS · 2,039 GB/s
FP16 全精度，训练也可
L20
中国版
48 GB GDDR6 · FP16 239 TFLOPS · 864 GB/s
INT8 推理部署首选
H20
中国版
96 GB HBM3 · FP16 296 TFLOPS · 4,000 GB/s
FP16 全精度，大显存+高带宽



模型参数 × GPU 快速对照总表

模型规模	FP16 显存	INT8 显存	INT4 显存	消费级推荐	数据中心推荐	中国可用
1.5B	3.6 GB	1.8 GB	0.9 GB	RTX 4060 (8G)	—	任意消费级卡
7B	16.8 GB	8.4 GB	4.2 GB	RTX 4090 (24G)	L4 / T4	RTX 4090D
14B	33.6 GB	16.8 GB	8.4 GB	RTX 5090 (32G)	L40S (48G)	L20 / RTX 4090D
32B	76.8 GB	38.4 GB	19.2 GB	RTX 5090 + 量化	A100 / L40S	H20 / L20
70B	168 GB	84 GB	42 GB	RTX 4090×4 + INT4	H100 / H200	H20 (96G) 单卡INT8
175B	420 GB	210 GB	105 GB	不现实	H100×8 / H200×4	H20×8
405B	972 GB	486 GB	243 GB	不现实	H200×8 / B200×4	H20×8 INT8
671B (MoE)	~320 GB*	~160 GB*	~80 GB*	RTX 4090×4 INT4*	H100×4 / H200×2	H20×4 INT8*

* MoE（混合专家）架构仅激活部分参数，实际显存需求远小于等效密集模型。如 DeepSeek-R1 671B 实际活跃参数约 37B。

训练 vs 推理 — 显存需求差异

推理（Inference）

- 仅加载模型权重 + KV Cache
- 显存 ≈ **参数量 × 精度 × 1.2**
- INT4/INT8 量化可大幅压缩
- 单卡即可运行中小模型

全参训练（Full Fine-tuning）

- 需存储：权重 + 梯度 + 优化器状态 + 激活值
- 显存 ≈ **参数量 × 精度 × 4~6 倍**
- 如 7B FP16 训练需 ~84-100 GB
- 通常需要多卡 + DeepSpeed/FSDP

LoRA 微调（Parameter-Efficient）

- 冻结主体权重，仅训练少量适配器
- 显存 ≈ **推理显存 × 1.3~1.5**
- 如 7B LoRA 仅需 ~20 GB
- RTX 4090 单卡即可微调 7B 模型

H20 为什么是中国 AI 的「真香卡」？

H20 虽然算力被大幅削减（CUDA 核心仅 ~9,984，FP8 仅 296 TFLOPS），但拥有 **96 GB HBM3 + 4,000 GB/s 带宽**——这恰好是 LLM 推理的核心瓶颈。

- 推理场景：大模型推理是**显存带宽受限（Memory-Bound）**而非计算受限，H20 的超大显存和高带宽完美匹配
- 70B 模型 INT8 量化：仅需 ~84 GB，H20 **单卡即可运行**，而 H100 (80GB) 反而装不下
- 性价比：H20 价格约为 H100 的 1/3~1/2，但在推理吞吐量上仅低 20~30%
- NVLink 900 GB/s：多卡互连带宽未被削减，适合多卡并行大模型

选购决策指南

不知道选哪张卡？跟着这棵决策树走一遍。

 本文数据整理自 NVIDIA 官方资料及公开报道，参考价格为发布时建议零售价（MSRP），实际价格因市场供需而异。

 中国特供版信息基于公开报道整理，部分规格可能因出口管制政策调整而变化。

 仅供科普参考，不构成购买建议。

附录：快速参考

GPU选型对照表

使用场景	推荐系列	代表型号
游戏玩家	GeForce	RTX 4090 / 4080
内容创作	GeForce	RTX 4080 Super
专业设计	RTX Pro	RTX 6000 Ada
AI推理	L / T系列	L40S / H20 / T4
大模型训练	H / B系列	H100 / H200 / B200
国内合规	特供版	H20 / L20 / A800

GPU算力排行（FP16 Tensor Core）

排名	型号	特点
1	B200	双芯旗舰，性能最强
2	H200	超大141GB显存
3	H100	行业标准首选
4	A100	高性价比训练卡
5	H800/H20	中国合规版本
6	L40S	高效能推理卡
7	RTX 4090	消费级旗舰

- 数据来源：NVIDIA官方资料及公开报道
- 价格参考：发布时建议零售价（MSRP），实际价格因市场供需而异
- 中国特供版信息基于公开报道整理，部分规格可能因出口管制政策调整而变化
- 本文仅供科普参考，不构成购买建议

内容整理自腾讯程序员公众号

整理日期：2026年3月21日