

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

01 阅读导图与学习路径

02 架构演进史

03 GPU命名规则解读

04 核心概念扫盲

05 消费级GPU (GeForce系列)

06 专业工作站GPU

07 推理加速卡 (数据中心)

08 训练旗舰 (数据中心)

09 中国特供版GPU

10 横向对比总表

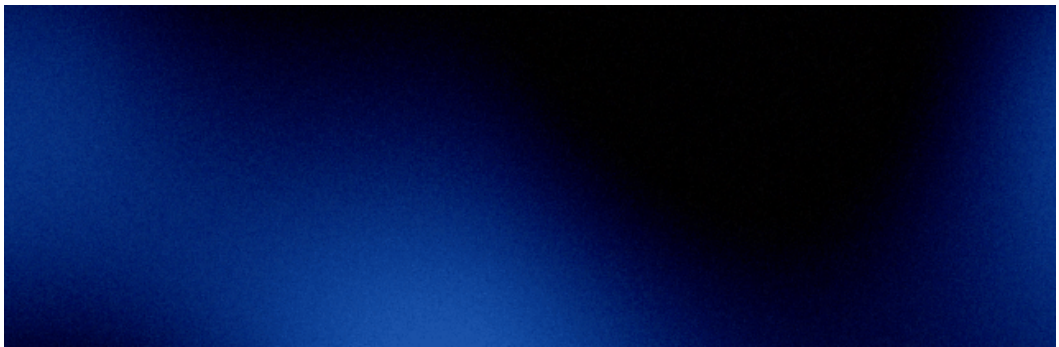
11 性能可视化

12 GPU算力指标对比

13 应用场景推荐

14 模型参数×GPU选择对照

15 选购决策指南



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。



这篇文章适合谁看？ 新增

Who Is This For?

如果你不是来查芯片拆解评测，而是想搞明白**英伟达这么多 GPU 到底谁是谁、该选哪张**——无论你是游戏玩家、设计师、AI 工程师还是采购决策者，这篇都能帮你把产品线理清、把选型逻辑讲透。

这篇到底想讲明白什么？ 核心

Three Core Questions

核心回答三件事：第一，英伟达的 GPU 分哪些产品线，各自面向谁；第二，同一产品线里，不同型号的核心差异在哪（显存、算力、功耗、互连）；第三，面对具体场景（游戏/设计/推理/训练/国内合规），该怎么选。

先看什么，后看什么？ 建议

Recommended Reading Order

如果你只想快速选型，建议按 **快速摘要 → 应用场景 → 选购指南** 三步搞定；如果你想系统了解，按导航顺序从架构演进读到最后。前一种像先看结论再翻数据，后一种像从头到尾看完整报告。

📌 打个比方：像去餐厅——赶时间直接看“必点推荐”，想深入了解就翻完整菜单和食材说明。

主线 1：产品线定位 核心

Product Lines

全篇最重要的第一件事：**搞清楚五大产品线各自面向谁**。消费级 GeForce 面向游戏和创作；工作站 RTX Pro 面向专业设计；推理卡 L/T 面向数据中心部署；训练卡 B/H/A 面向大模型研发；中国特供版面向国内合规场景。认准定位，才不会看花眼。

主线 2：硬件规格 关键

Key Specs

看 GPU 规格时只需抓住四个关键：**显存容量**（决定能装多大模型）、**算力 TFLOPS**（决定计算速度）、**显存带宽**（决定数据搬运速度）、**互连带宽**（决定多卡协作效率）。这四个数字能让你在对比表里快速做出判断。

主线 3：场景选型 关键

Scenario Matching

规格参数再好看，最终要落到你的**实际场景**上：游戏看核心性能和 DLSS；设计看显存和 ISV 认证；推理看吞吐和功耗比；训练看显存、算力和多卡互连；国内场景还要加一条合规性。文末的应用场景和选购决策树就是帮你做这步匹配的。

如果你是游戏玩家 / 创作者
For Gamers & Creators

重点看 **消费级 + 性能图表 + 选购指南**。关注 CUDA 核心数、光追性能、DLSS 版本和显存大小。RTX 50 系新增多帧生成，提升巨大。

如果你是 AI 工程师 / 研究员
For AI Engineers & Researchers

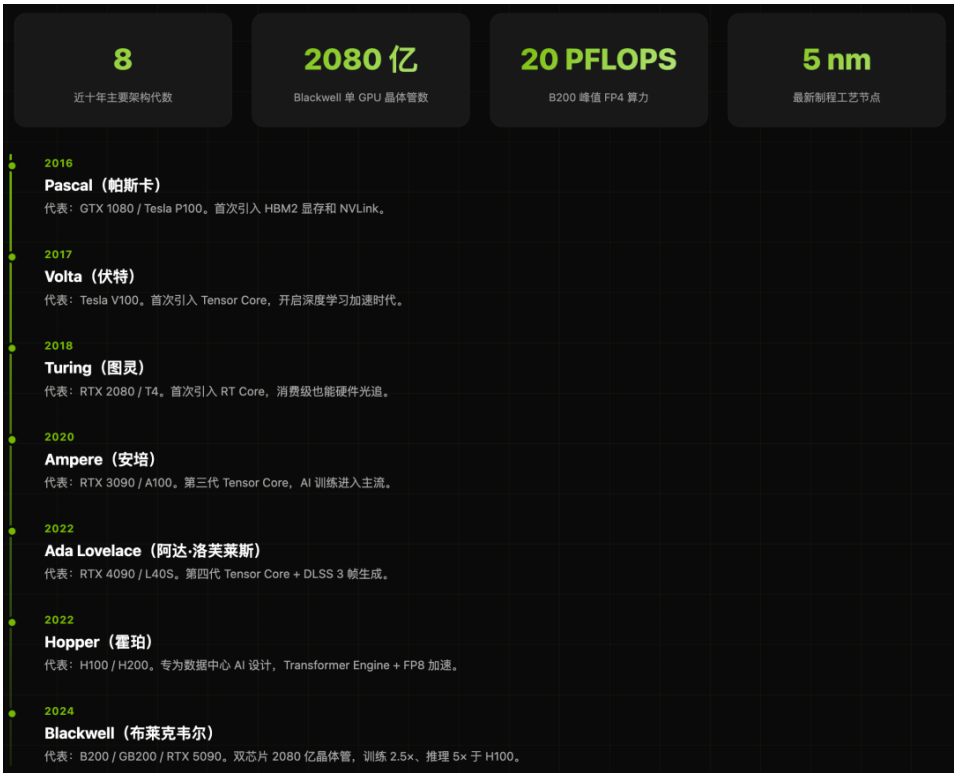
先快速过一遍导图，再重点看 **训练卡 + 算力排行 + 模型选 GPU**。阅读时追问三个问题：**显存够不够装下模型？算力够不够训完？多卡互连有没有瓶颈？**

如果你是国内企业 IT / 采购
For CN Enterprise Buyers

直奔 **中国特供版 + 应用场景 + 选购指南**。优先关注 H20、L20、RTX 4090D 的合规可用性和实际算力，以及存量 A800/H800 设备的利用策略。

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。



CUDA 核心

GPU 基本计算单元，核心越多并行能力越强。消费级旗舰可达 2 万+。



Tensor Core

专为矩阵运算设计的加速单元，极大加速 AI 训练和推理，是区分游戏卡和 AI 卡的关键。



显存 (VRAM)

GPU 高速内存。显存越大，能加载的模型越大。AI 大模型常需 40-192 GB。类型：GDDR6/7、HBM。



显存带宽

数据搬运速度。HBM 带宽远超 GDDR，决定 AI 推理延迟和吞吐。



NVLink

NVIDIA 多卡高速互连技术，速度远超 PCIe，大模型训练必备。



精度 (FP32/16/8)

训练用 FP32/FP16 混合精度；推理可用 FP8/INT8 量化换取高吞吐。



TDP 功耗

最大功耗。消费级 150-575W，数据中心 400-700W。直接影响散热和电费。



DLSS

AI 画面升级技术。低分辨率渲染 + AI 补帧到高分辨率，帧数暴涨画质几乎无损。

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。



GeForce RTX 5090

Blackwell · GB202 · 2025

CUDA 核心
21,760

显存
32 GB GDDR7

带宽
1,792 GB/s

TDP
575 W

接口
PCIe 5.0

参考价
\$1,999

4K/8K 游戏 AI 创作 DLSS 4 本地大模型



GeForce RTX 5080

Blackwell · GB203 · 2025

CUDA 核心
10,752

显存
16 GB GDDR7

带宽
960 GB/s

TDP
360 W

接口
PCIe 5.0

参考价
\$999

4K 游戏 视频剪辑 3D 渲染



GeForce RTX 5070 Ti

Blackwell · GB203 · 2025

CUDA 核心
8,960

显存
16 GB GDDR7

带宽
896 GB/s

TDP
300 W

接口
PCIe 5.0

参考价
\$749

2K/4K 游戏 直播推流 AI 绘画



GeForce RTX 5070

Blackwell · GB205 · 2025

CUDA 核心
6,144

显存
12 GB GDDR7

带宽
672 GB/s

TDP
250 W

接口
PCIe 5.0

参考价
\$549

≈ RTX 4090 性能 性价比之王 2K 游戏



GeForce RTX 4090

Ada Lovelace · AD102 · 2022

CUDA 核心
16,384

显存
24 GB GDDR6X

带宽
1,008 GB/s

TDP
450 W

接口
PCIe 4.0

参考价
\$1,599

4K 旗舰 本地大模型 AI 开发



GeForce RTX 4080 SUPER

Ada Lovelace · AD103 · 2024

CUDA 核心
10,240

显存
16 GB GDDR6X

带宽
736 GB/s

TDP
320 W

接口
PCIe 4.0

参考价
\$999

4K 游戏 视频创作



GeForce RTX 4060 Ti

Ada Lovelace · AD106 · 2023

CUDA 核心
4,352

显存
8/16 GB GDDR6

带宽
288 GB/s

TDP
160 W

接口
PCIe 4.0

参考价
\$399

1080p/2K 入门创作 低功耗

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。



RTX 6000 Ada Generation

Ada Lovelace · AD102 · 2023

CUDA 核心
18,176

显存
48 GB GDDR6 ECC

带宽
960 GB/s

TDP
300 W

NVLink
支持

参考价
~\$6,800

影视特效 工业设计 数字孪生



RTX A6000

Ampere · GA102 · 2021

CUDA 核心
10,752

显存
48 GB GDDR6 ECC

带宽
768 GB/s

TDP
300 W

NVLink
支持

参考价
~\$4,650

CAD 设计 光追渲染 AI 开发



RTX 4000 Ada

Ada Lovelace · AD104 · 2023

CUDA 核心
6,144

显存
20 GB GDDR6 ECC

带宽
360 GB/s

TDP
130 W

形态
单槽 SFF

参考价
~\$1,250

轻薄工作站 BIM 建筑 医学影像

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。

NVIDIA L40S

Ada Lovelace · AD102 · 2023

CUDA 核心

18,176

显存

48 GB GDDR6

带宽

864 GB/s

FP8

733 TFLOPS

TDP

350 W

参考价

~\$7,000

LLM 推理

AI 视频

图形+AI

轻训练

NVIDIA L4

Ada Lovelace · AD104 · 2023

CUDA 核心

7,424

显存

24 GB GDDR6

带宽

300 GB/s

INT8

485 TOPS

TDP

72 W

形态

低功耗半高卡

边缘推理

视频分析

云游戏

低功耗

NVIDIA T4

Turing · TU104 · 2019

CUDA 核心

2,560

显存

16 GB GDDR6

带宽

320 GB/s

INT8

130 TOPS

TDP

70 W

特点

被动散热-无需外接电源

经典推理卡

云推理

视频转码

NVIDIA A10

Ampere · GA102 · 2021

CUDA 核心

9,216

显存

24 GB GDDR6

带宽

600 GB/s

TDP

150 W

形态

单槽低功耗

特点

图形+AI 兼顾

云桌面

AI 推理

虚拟化

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。

GB200 NVL72

Blackwell + Grace · 2024

配置

36 Grace + 72 B200

FP4 算力

1,440 PFLOPS

总显存

13.5 TB HBM3e

NVLINK

第5代 1.8TB/s

功耗

~120 kW/机柜

冷却

液冷

万亿参数模型

超级计算机

AGI 研究

NVIDIA B200

Blackwell · 2024

晶体管

2,080 亿(双芯片)

显存

192 GB HBM3e

带宽

8,000 GB/s

FP4

20 PFLOPS

TDP

700 W

VS H100

训练 2.5x 推理 5x

大模型训练

科学计算

千亿参数

NVIDIA H200

Hopper · GH100 · 2024

CUDA 核心

16,896

显存

141 GB HBM3e

带宽

4,800 GB/s

FP8

3,958 TFLOPS

TDP

700 W

VS H100

推理 1.6-1.9x

LLM 推理加速

大模型训练

H100 升级

NVIDIA H100 SXM

Hopper · GH100 · 2022

CUDA 核心

16,896

显存

80 GB HBM3

带宽

3,350 GB/s

FP8

3,958 TFLOPS

TDP

700 W

参考价

~\$30,000+

GPT-4 训练核心

AI 标杆

HPC

NVIDIA A100

Ampere · GA100 · 2020

CUDA 核心

6,912

显存

40/80 GB HBM2e

带宽

2,039 GB/s

TF32

312 TFLOPS

TDP

400 W

特点

MIG 多实例

AI 训练主力

科学计算

云基础设施

NVIDIA V100

Volta · GV100 · 2017

CUDA 核心

5,120

显存

16/32 GB HBM2

带宽

900 GB/s

FP16

125 TFLOPS

TDP

300 W

状态

经典-仍在服役

经典训练卡

存量部署

入门 AI 集群

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）。

数据中心级（训练/推理）

NVIDIA H20

Hopper · GH100 · 2024 · 中国特供

CUDA 核心

~9,984 (vs H100 16,896)

显存

96 GB HBM3

带宽

4,000 GB/s

FP8

296 TFLOPS

NVLink

900 GB/s

TDP

400 W

LLM 推理首选

大显存优势

国内主力卡

MIG 支持

NVIDIA H800

Hopper · GH100 · 2022 · 中国特供

CUDA 核心

16,896 (同 H100)

显存

80 GB HBM3

带宽

3,350 GB/s

NVLink

400 GB/s (vs H100 900)

FP64

1 TFLOPS (大幅削减)

TDP

700 W

AI 训练

NVLink 降速版

已禁售(2023.10)

NVIDIA A800

Ampere · GA100 · 2022 · 中国特供

CUDA 核心

6,912 (同 A100)

显存

80 GB HBM2e

带宽

2,039 GB/s

NVLink

400 GB/s (vs A100 600)

TDP

300 W (PCIe)

差异

仅 NVLink 降速

AI 训练

A100 合规版

已禁售(2023.10)

NVIDIA L20

Ada Lovelace · AD102 · 2024 · 中国特供

CUDA 核心

11,776

显存

48 GB GDDR6 ECC

带宽

864 GB/s

INT8

239 TOPS

TDP

275 W

形态

双槽全高全长 PCIe

AI 推理

视频分析

L40S 合规替代

NVIDIA B20 (规划中)

Blackwell · 2025+ · 中国特供

定位

B200 中国合规版

显存

待定 (预计缩减)

互连

受限 (合规要求)

状态

计划 2025 Q2+ 出货

合作

浪潮等经销商

对标

H20 升级版

下一代中国市场

Blackwell 架构

待发布

消费级中国特供版

GeForce RTX 4090D

Ada Lovelace · AD102 · 2023 · 中国特供

CUDA 核心

14,592 (vs 4090 16,384)

显存

24 GB GDDR6X

带宽

1,008 GB/s

TDP

425 W (vs 4090 450W)

参考价

¥12,999

差异

CUDA -11%, 性能 -5~10%

国内旗舰游戏卡

本地大模型

AI 开发

GeForce RTX 5090D (预期)

Blackwell · GB202 · 2025 · 中国特供

CUDA 核心

21,760 (预计同 5090)

显存

32 GB GDDR7

带宽

1,792 GB/s

TDP

575 W

差异

预计仅限制 AI 算力

据传游戏性能不变

国行旗舰

游戏性能=5090

待上市

中国特供版 vs 原版核心差异一览

原版	中国版	主要削减项	性能影响	状态
A100	A800	NVLink 600→400 GB/s	多卡通信降速，单卡不受影响	已禁售
H100	H800	NVLink 900→400 GB/s, FP64 阉割	多卡训练效率下降，高精度计算受限	已禁售
H100	H20	CUDA -41%, 算力 -80%	算力大幅降低，但显存增至 96GB	管制收紧中
L40S	L20	CUDA 18176→11776, INT8 削减	推理吞吐降低约 35%	在售
RTX 4090	RTX 4090D	CUDA 16384→14592	游戏 -5~10%, AI 性能 -10%	在售
B200	B20	预计大幅削减互连与算力	具体待发布	规划中

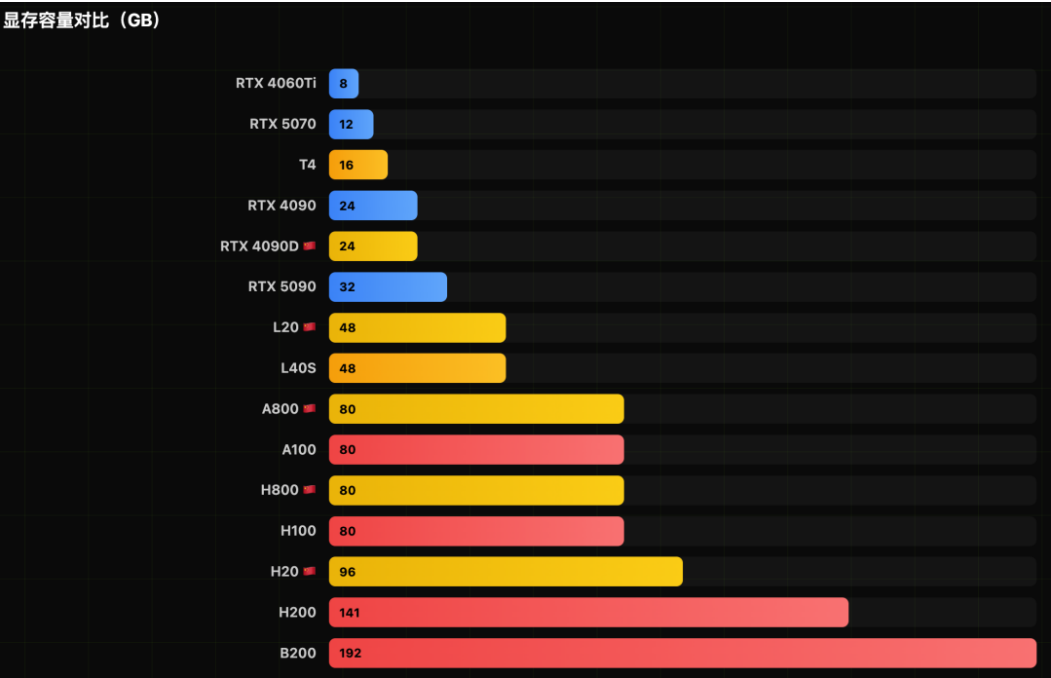
横向对比总表

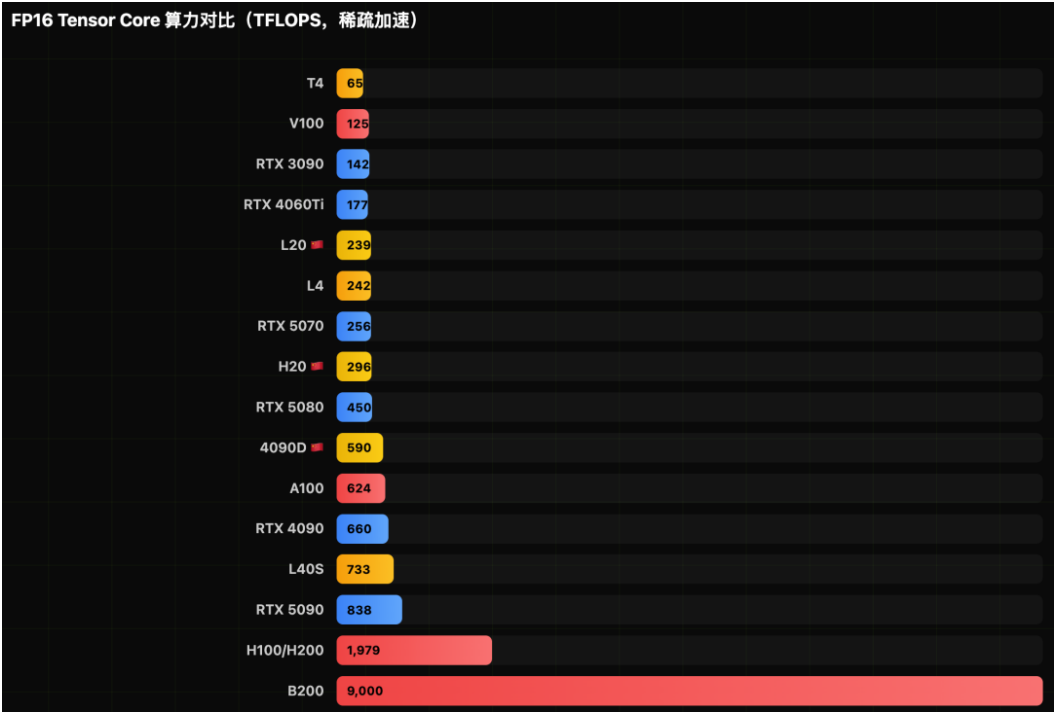
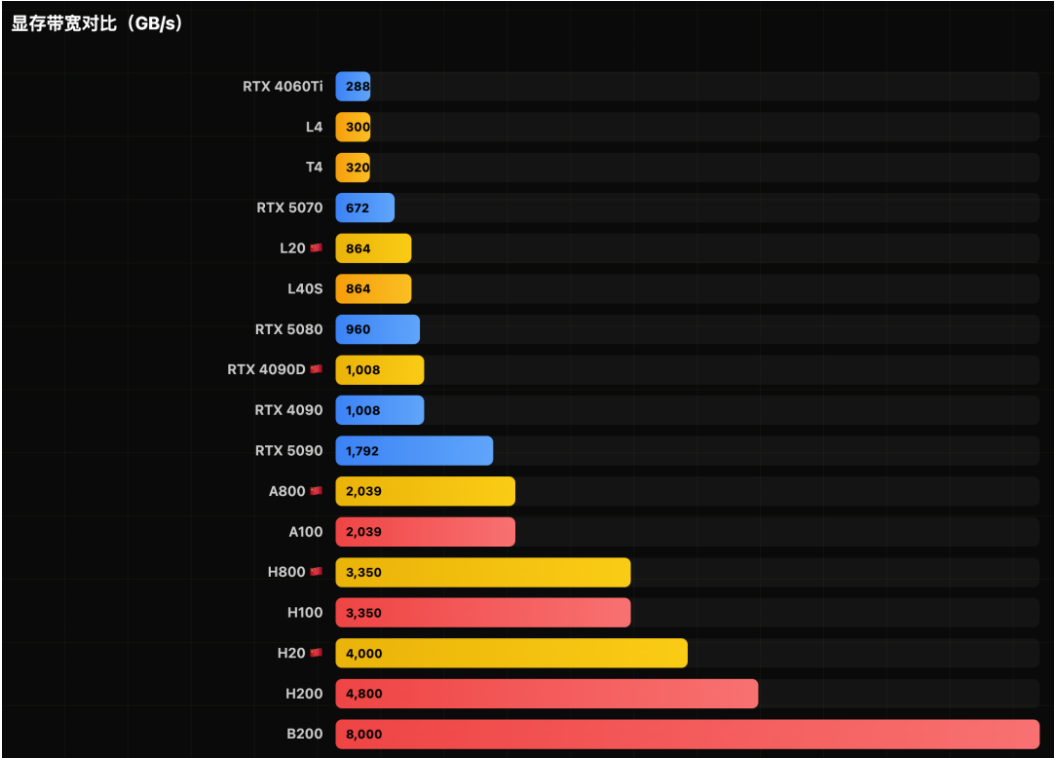
一张表快速对比关键参数差异。

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	中国版	Hopper	~9,984	96GB HBM3	4,000 GB/s	400W	中国市场主力
H800 中国版	中国版	Hopper	16,896	80GB HBM3	3,350 GB/s	700W	H100 合规版
A800 中国版	中国版	Ampere	6,912	80GB HBM2e	2,039 GB/s	300W	A100 合规版
L20 中国版	中国版	Ada Lovelace	11,776	48GB GDDR6	864 GB/s	275W	推理合规版
RTX 4090D 中国版	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。

算力指标说明

FP32 (单精度)

通用计算基准

FP16 Tensor

AI 训练核心

FP8 / INT8

AI 推理加速

显存带宽

LLM 推理瓶颈

能效比

TFLOPS / 瓦

TFLOPS = 每秒万亿次浮点运算。Tensor Core 算力使用稀疏加速 (Sparsity) 后的峰值数据 (标准值 x2)，这是 NVIDIA 官方发布的标准口径。

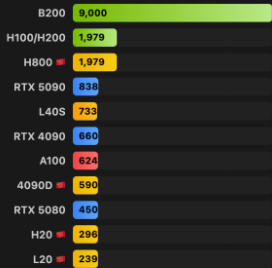
注意：数据中心卡的 Tensor Core 算力与消费级卡不可简单类比——前者针对大矩阵运算深度优化，实际 AI 场景差距远大于数字比值。

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	年份	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 训练		Ada Lovelace	2023	91.6	733	733 (FP8)	48 GB GDDR6	864	350	2.09
5	RTX 5090 训练		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 训练		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 训练		Blackwell	2025	56	450	900 (FP4)	16 GB GDDR7	960	360	1.25
10	L4 训练		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 训练		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 训练		Blackwell	2025	32	256	512 (FP4)	12 GB GDDR7	672	250	1.02
16	RTX 4080S 训练		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 训练		Ampere	2020	35.6	142	284 (INT8)	24 GB GDDR6X	936	350	0.41
22	RTX 4060 Ti 训练		Ada Lovelace	2023	22.1	177	177 (FP8)	8 GB GDDR6	288	160	1.11
23	T4 训练		Turing	2019	8.1	65	130 (INT8)	16 GB GDDR6	320	70	0.93
24	A10 训练		Ampere	2021	31.2	125	250 (INT8)	24 GB GDDR6	600	150	0.83

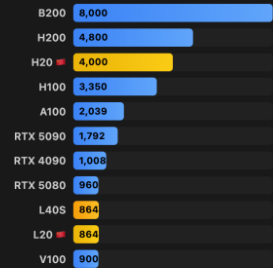
FP16 Tensor 算力排行

AI 训练核心指标（TFLOPS，稀疏加速）



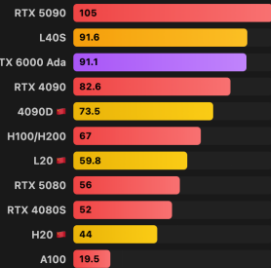
显存带宽排行

LLM 推理吞吐核心瓶颈（GB/s）



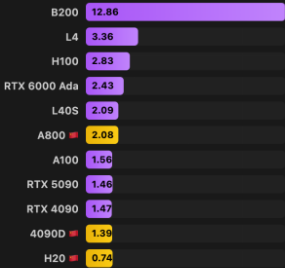
FP32 通用算力排行

游戏 / 渲染 / 通用计算（TFLOPS）



能效比排行

FP16 Tensor TFLOPS / 每瓦功耗



🔍 算力排行关键洞察

📊 算力 ≠ 实际推理速度

LLM 推理是**显存带宽受限 (Memory-Bound)** 而非算力受限。H20 虽然 FP16 仅 296 TFLOPS (仅为 H100 的 15%)，但凭借 4,000 GB/s 带宽，**推理吞吐量可达 H100 的 70~80%**。

📍 中国可用卡的算力折损

RTX 4090D 算力降至 4090 的 **~89%**；H20 算力仅为 H100 的 **~15%**，但显存大 20%，带宽高 19%。
H800/A800 算力未削减，仅互连带宽受限。

📈 B200 断层式领先

B200 的 FP16 算力达 **9,000 TFLOPS**，是 H100 的 4.5 倍。Blackwell 双芯架构 + 第6代 Tensor Core + FP4 精度支持，将训练效率推向新高度。

⚡ RTX 5090：消费级新王

RTX 5090 FP16 达 **838 TFLOPS**，比 RTX 4090 (660) 提升 27%。但更大的突破在于 **FP4 支持 (1,676 TFLOPS)** 和 32GB GDDR7 大显存。

🏆 性价比 vs 能效比对照

#	型号	FP16 Tensor TFLOPS	参考价 USD	每千美元 FP16 算力	TDP W	每瓦 FP16 算力	显存/千美元 GB/\$K	适合场景
1	RTX 4090	660	\$1,599	413	450	1.47	15.0	性价比之王
2	RTX 4090D	590	¥12,999 ~\$1,800	328	425	1.39	13.3	国内性价比
3	RTX 5090	838	\$1,999	419	575	1.46	16.0	消费级旗舰
4	RTX 5080	450	\$999	450	360	1.25	16.0	高端甜品
5	RTX 5070	256	\$549	466	250	1.02	21.9	入门 AI 最划算
6	L40S	733	~\$7,000	105	350	2.09	6.9	数据中心推理
7	L4	242	~\$2,500	97	72	3.36	9.6	低功耗边缘推理
8	H100 SXM	1,979	~\$30,000	66	700	2.83	2.7	训练 / 大规模推理
9	H20	296	~\$12,000	25	400	0.74	8.0	中国 LLM 推理主力
10	A100 (80G)	624	~\$12,000	52	400	1.56	6.7	经典训练主力

参考价格为 2025 年初典型市场价 (新卡)，非官方 MSRP 约可能存在溢价。「每千美元 FP16 算力」= FP16 TFLOPS / (价格/1000)，越高越好。

🔥 按场景选算力：该看哪个指标？			
应用场景	核心指标	为什么？	推荐 GPU
🎮 游戏光追	FP32 + RT Core	光栅化和光追依赖通用浮点算力	RTX 5090 > 4090 > 5080
🧠 LLM 推理	显存带宽 + 显存容量	推理是 Memory-Bound，带宽决定吞吐	H20 🚩 > H200 > H100
🔥 大模型训练	FP16/BF16 Tensor + NVLink	矩阵运算主导，多卡互连是关键	B200 > H100 > A100
🎨 AI 绘画/视频	FP16 Tensor + 显存容量	Stable Diffusion 等依赖半精度矩阵运算	RTX 5090 > 4090 > L40S
☁️ 云端批量推理	INT8/FP8 + 能效比	追求每瓦吞吐最大化，降低电费成本	L4 > L40S > T4
🏗️ CAD/渲染	FP32 + 显存容量 + ECC	大场景需要大显存和精确计算	RTX 6000 Ada > A6000

应用场景推荐

根据你的实际需求，找到最适合的 GPU。



3A 游戏大作

4K 最高画质 + 光追 + DLSS, 流畅 60fps+ 体验

RTX 5090 RTX 5080 RTX 4090



影视后期 / 3D 渲染

大型场景实时预览、GPU 渲染加速 (Blender/V-Ray)

RTX 6000 Ada RTX A6000 RTX 4090



本地跑大模型 / AI 绘画

Stable Diffusion, LLaMA 本地推理, 显存是关键

RTX 5090(32G) RTX 4090(24G)

RTX 5080(16G)



云端 AI 推理服务

大规模并发推理, 低延迟、低功耗、高密度部署

L40S L4 T4 A10



大模型训练 (百亿参数)

GPT/LLaMA 级别模型预训练, 需要超大显存和多卡互连

H100 H200 A100



超大模型 / AGI 研究

万亿参数训练, 超级计算集群, 前沿 AI 研究

B200 H200 H100



CAD / BIM / 工业仿真

大型装配体实时旋转、CFD/FEA 计算可视化

RTX 6000 Ada RTX A6000



视频转码 / 流媒体

大规模视频编解码, 云游戏串流

L4 T4 A10



国内 AI 大模型训练

受出口管制影响, 国内企业可选芯片。H20 凭借 96GB 大显存成为主力

H20 H800(存量) A800(存量)



国内 AI 推理部署

国内云厂商推理场景, L20 低功耗大显存适合部署

L20 H20 RTX 4090D



教育企业 AI 应用

智能批改、自适应学习、AI 助教、智能题库生成、虚拟教学场景。需平衡推理吞吐与部署成本

RTX 4090D L20 H20 H800(存量)



高校 / 科研院所

深度学习科研、NLP/CV 前沿实验、多模态大模型微调、学生实训集群。需兼顾科研灵活性与显存容量

H20 A800(存量) RTX 4090D L20



高校教学实训平台

AI/机器学习课程实训、多用户共享 GPU 集群、Jupyter 云平台。需高并发分时复用, 单卡成本敏感

T4 A10 RTX 4090D RTX 4060 Ti



科研大模型预训练

高校/实验室自研大语言模型或基础模型, 百亿至千亿参数预训练。需超大显存 + 高速互连集群

H20 H800(存量) A800(存量)

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。

显存需求估算公式

显存 ≈ 参数量 × 精度系数 × 安全系数(1.2)

安全系数 1.2 用于预留 KV Cache、激活值、框架开销等。训练场景需额外 3-4x 存储梯度和优化器状态。

FP32 (全精度)

4 字节/参数

FP16 / BF16 (半精度)

2 字节/参数

INT8 (8位量化)

1 字节/参数

INT4 (4位量化)

0.5 字节/参数

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

8-12 GB GDDR6

FP16 直接运行, 绰绰有余

GTX 1660 / 1060

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

24 GB GDDR6X · FP16 660 TFLOPS · 1,008 GB/s

FP16 全精度, 速度极快

RTX 4080 / 3080

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

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

32 GB GDDR7 · FP16 838 TFLOPS · 1,792 GB/s

FP16 全精度流畅运行

RTX 4090 / 3090

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

24 GB GDDR6X · FP16 590 TFLOPS · 1,008 GB/s

INT8 量化推理首选

L20

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

32 GB GDDR7 · FP16 838 TFLOPS · 1,792 GB/s

INT4 量化 + vLLM 优化

RTX 4090 × 2

24×2=48 GB · 单卡 FP16 660 TFLOPS

INT8 张量并行推理

L40S

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

48 GB GDDR6 · FP16 239 TFLOPS · 864 GB/s

INT8 推理部署首选

H20

96 GB HBM3 · FP16 296 TFLOPS · 4,000 GB/s

FP16 全精度, 大显存+高带宽

70B

700 亿参数（企业级主力）

代表模型：Llama3.1-70B、Qwen2.5-72B、DeepSeek-V2.5

FP16 推理 ≈ 168 GB

INT8 推理 ≈ 84 GB

INT4 推理 ≈ 42 GB

H100 SXM (80GB)

✔ 最优

80 GB HBM3 · FP16 1,979 TFLOPS · 3,350 GB/s

INT8 单卡可跑，FP16 需 2~4 卡

H200

✔ 顶级

141 GB HBM3e · FP16 1,979 TFLOPS · 4,800 GB/s

FP16 单卡完整加载！

A100 × 2 (80GB)

80×2=160 GB · 单卡 FP16 624 TFLOPS

FP16 张量并行部署

RTX 4090 × 4

24×4=96 GB · 单卡 FP16 660 TFLOPS

INT4 量化 + 多卡并行

H20

中国特选

96 GB HBM3 · FP16 296 TFLOPS · 4,000 GB/s

INT8 单卡可跑！带宽优势明显

H200 × 2

96×2=192 GB · 合计 FP16 592 TFLOPS

FP16 全精度 + NVLink 互连

H800 × 4 (存量)

80×4=320 GB · 单卡 FP16 1,979 TFLOPS

FP16 训练 + 推理兼跑

175B+

千亿+ 参数（超大规模）

代表模型：GPT-3 (175B)、Llama3.1-405B、DeepSeek-R1 (671B MoE)

FP16 推理 ≈ 420~1,608 GB

INT8 推理 ≈ 210-804 GB

INT4 推理 ≈ 105-402 GB

B200 × 4-8

✔ 顶级

192×8=1,536 GB · 单卡 FP16 9,000 TFLOPS

万亿参数也能 FP16 全精度

H100 × 8 (DGX)

✔ 主流

80×8=640 GB · 单卡 FP16 1,979 TFLOPS

405B INT8 推理，NVLink 900GB/s

H200 × 4-8

141×8=1,128 GB · 单卡 4,800 GB/s 带宽

405B FP16 全精度推理

A100 × 8-16

80×16=1,280 GB · 单卡 FP16 624 TFLOPS

训练集群经典方案

H20 × 8

中国方案

96×8=768 GB · 合计带宽 32,000 GB/s

405B INT8，中国可用最优配置

H800 × 8 (存量)

80×8=640 GB · 单卡 FP16 1,979 TFLOPS

存量设备，NVLink 降至 400 GB/s

■ 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/FP8DP

💡 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