长江证券 CHANGJIANG SECURITIES

| 证券研究报告 |

行业研究 | 深度报告 | 电子元件

AI 行业十问十答系列之一：AI 服务器为算力载体，需求弹性空间几何

请阅读最后评级说明和重要声明

报告要点

AI 作为未来十年一遇的产业机遇，当前正处于 AI 的“iPhone 时刻”，该系列报告旨在通过问答的形式来梳理行业投资脉络，跟踪行业发展的趋势。本篇报告作为我们 AI 行业十问十答系列第一篇，更多的是对 AI 服务器行业的投资逻辑思考以及行业的需求弹性空间、市场格局等多方面的梳理。AI 服务器作为算力的载体，未来充分受益于算力时代的算力爆发式增长，多个环节都存在投资机会，详情可参见正文。

分析师及联系人

杨洋

SAC：S0490517070012

蔡少东

SAC：S0490522090001

丨证券研究报告丨

2023-05-20

行业研究丨深度报告

投资评级 看好丨上调

电子元件

AI 行业十问十答系列之一：AI 服务器为算力载体，需求弹性空间几何

AI 行业的“iPhone 时刻”对于服务器意味着什么

我们认为，在 AIGC 时代作为算力载体的服务器将迎来 AI 化的加速升级，正如在移动互联网时代对移动终端硬件算力升级带来的智能手机对功能机的取代，未来 AI 服务器在整体服务市场占比有望快速提升。复盘智能手机对功能机的替代过程不难发现，全球智能手机在 2007 年出货超过 1 亿部，渗透率超过 10%，在 2013 年出货量超过十亿部，渗透率超过 50%，值得注意的是，在渗透率越过 10%节点后，智能手机渗透斜率出现了加速趋势。当前全球 AI 服务器的渗透率约在 1%，参考智能手机以及新能源车发展历程，越过 1%临界点后，渗透率超过 10%的时间周期在 4-5 年左右，因此未来几年 AI 服务器行业出货量是非常值得期待的。

AI 服务器需求弹性测算

AI 服务器的需求包括训练端以及推理端，当前市场更多关注大模型前期训练消耗的算力需求，未来伴随大模型应用的全面推广，远期在推理端的算力及服务器需求或超过训练端。

训练端：若以 GPT-3 模型为例（175B 参数量，3000 亿 tokens），在假设训练周期为一个月，使用 A100 GPU 算力利用率在 50%的情况下，大概需要约 100 台服务器，当然若单次训练周期较短以及实际算力利用率较低，实际的服务器需求量或高于此处测算值。

推理端：若以 GPT-3 模型为例（175B 参数量），假设单日访问人次为 0.5 亿次（2023 年 4 月份 ChatGPT 官网访问量为 17.6 亿次）、单次提问 tokens 为 1000、使用 T4 服务器的情况下，推理端需要的服务器数量约 390 台，超过此前训练端服务器需求量。考虑到未来用户数量仍在高速增长中，未来推理服务器需求量或将继续保持高增长。**详细测算过程参见正文。**

如何跟踪（AI）服务器行业

在跟踪（AI）服务器行业边际变化或者基本面时，我们认为应该具备全球产业链视角。AI 行业日新月异，市场充斥着大量资讯与产业变化，这些会被过滤后反映在高频的上市公司股价波动。另一方面，从投资角度来看，基本面的变化依旧是研究的抓手，除了第三方机构季度或月度的行业出货量及市场规模数据，部分上市公司有着高频的月度数据，以及季度法说会对下一季度的展望也是判断行业景气度的关键参考因素。当然考虑到服务器行业下游主要为云计算/互联网客户，国内外的云服务商资本开支数据依旧是行业冷暖的重要信号。

AI 行业投资策略

AI 大时代将催生诸多 AI+应用，破坏式革新赋能各行各业，对海量文字、图片、视频等 AIGC 信息的存-算-传-显，意味着全球将进入算力需求爆发式增长阶段，带来的是云-网-端-边的重大机遇，科技硬件无疑是 AIGC“起高楼”的基石，迎来长期重大发展机遇。从产业链受益先后角度，我们依次推荐算力载体服务器链、边缘 AI 产业链、算力芯片辅助链；在选股思路上侧重“含 AI 量”、“大客户”、“低估值”与“竞争力”等因素；投资节奏层面，可密切关注国内外大模型及应用更新动态，跟踪云厂商及互联网企业的资本开资情况。AI 产业链公司估值与业绩梳理可参考后文表格。

风险提示

1、AI 服务器出货不及预期；

2、大模型发展不及预期。

行业内重点公司推荐

公司代码	公司名称	投资评级
601138	工业富联	买入

市场表现对比图(近 12 个月)

资料来源：Wind

相关研究

- 《2013 年-2015 年电子行情复盘》2023-05-14
- 《“屏地风雷”系列报告之十三：走向虚拟世界的 LED 显示》2023-04-15
- 《新显示技术，高成长赛道——电子纸行业深度报告》2023-02-18

更多研报请访问
长江研究小程序

目录

图表目录

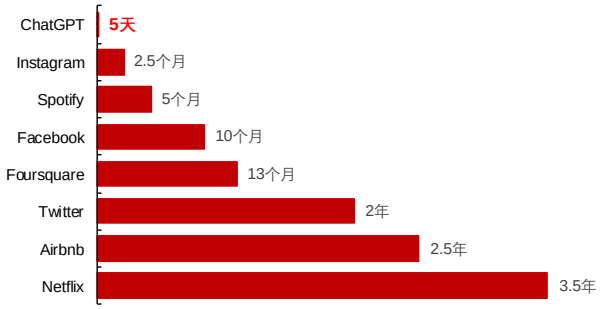
AI 行业十问十答系列之一

AI 作为未来十年一遇的产业机遇，当前正处于 AI 的“iPhone 时刻”，该系列报告旨在通过问答的形式来梳理行业投资脉络，跟踪行业发展的趋势。本篇报告作为我们 AI 行业十问十答系列第一篇，更多的是对 AI 服务器行业的投资逻辑思考以及行业的需求弹性空间、市场格局等多方面的梳理。AI 服务器作为算力的载体，未来充分受益于算力时代的算力爆发式增长，多个环节都存在投资机会，详情可参见正文。

Q：AI 行业的“iPhone 时刻”对于服务器意味着什么

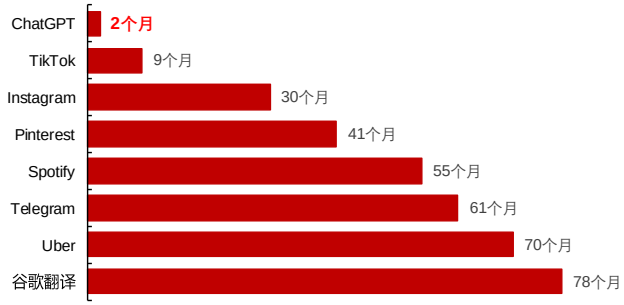
回顾曾经的第一代搭载多点触控技术的 iPhone 推出时，iPhone 对功能机及人机交互方式进行了颠覆式变革，对智能手机行业发展产生深远的影响，其产品迅速获得大批消费者的认可及购买。OpenAI ChatGPT 的推出对 AI 行业的发展、人机交互及社会发展都将是颠覆式影响，其交互可通过自然语言进行并较为精准理解、解答用户问题，正如英伟达 CEO 黄仁勋在公开场合提及 AI 行业进入到“iPhone 时刻”，ChatGPT 的注册用户、月活数迅速突破百万级别、亿级别，远超其他互联网应用。

图 1：ChatGPT 注册用户突破百万用时 5 天



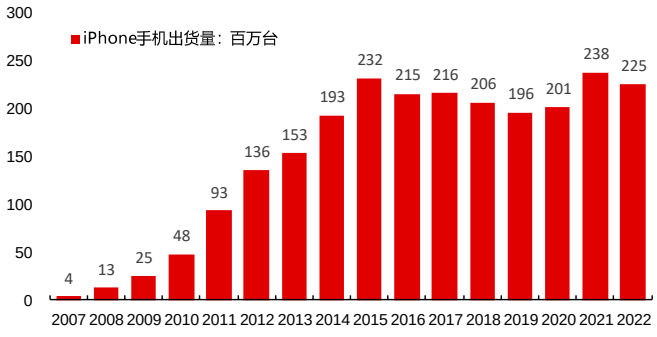
资料来源：财讯网，长江证券研究所

图 2：ChatGPT 突破 1 亿月活用时 2 个月



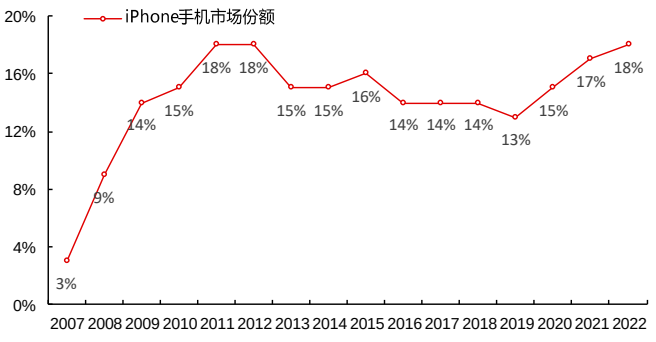
资料来源：财讯网，长江证券研究所

图 3：iPhone 手机历年出货量（2007-2022）



资料来源：Counterpoint，长江证券研究所

图 4：iPhone 手机历年市场份额（2007-2022）



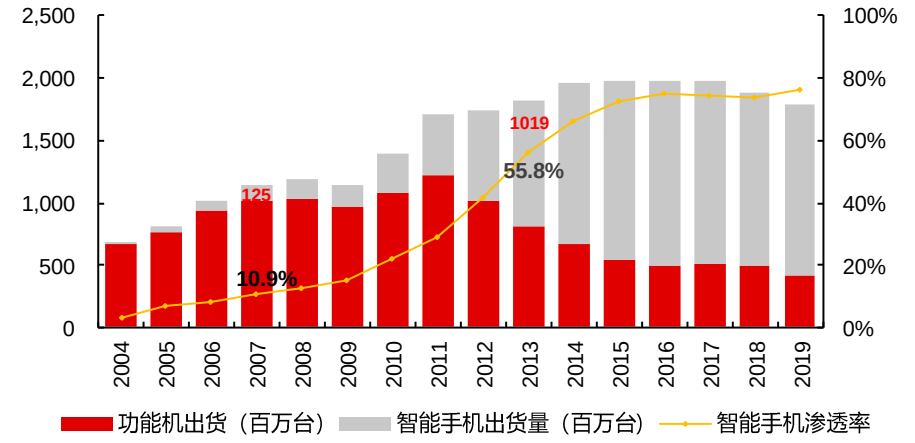
资料来源：Counterpoint，长江证券研究所

ChatGPT 的智能化核心来源于前期大量语料及参数的大模型的预训练，ChatGPT-3 拥有 1750 亿参数，无论是训练端或推理端，都需要算法及算力的支持，即大模型核心三要素：数据、算法、算力。我们认为，在 AIGC 时代作为算力载体的服务器将迎来 AI 化

的加速升级，正如在移动互联网时代对移动终端硬件算力升级带来的智能手机对功能机的取代，未来 AI 服务器在整体服务市场占比有望快速提升。

复盘智能手机对功能机的替代过程不难发现，全球智能手机在 2007 年出货超过 1 亿部，渗透率超过 10%，在 2013 年出货量超过十亿部，渗透率超过 50%，值得注意的是，在渗透率越过 10%节点后，智能手机渗透斜率出现了加速趋势。当前全球 AI 服务器的渗透率约在 1%，参考智能手机以及新能源车发展历程，越过 1%临界点后，渗透率超过 10%的时间周期在 4-5 年左右，因此未来几年 AI 服务器行业出货量是非常值得期待的。

图 5：2004-2019 年全球功能机与智能机出货量及智能手机渗透率

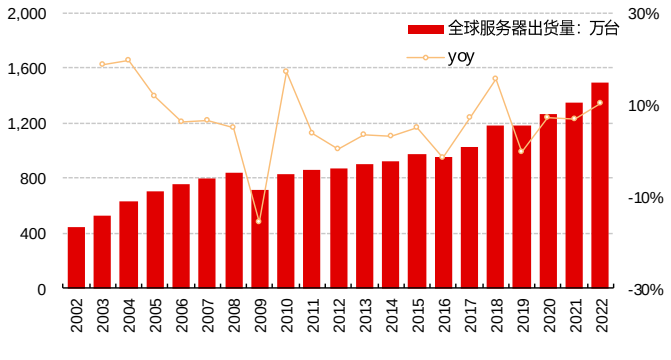


资料来源：智研咨询，长江证券研究所

Q：服务器市场的规模多大

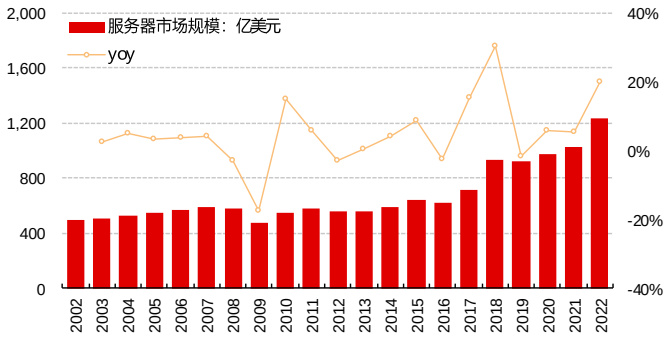
全球服务器出货量在近几年来整体保持较为稳定的增长趋势，2022 年全球服务器出货量约 1,495 万台，同比增长 10.39%，相较于量增幅度而言，2022 年全球服务器市场的规模增速要更高，2022 年全球服务器市场规模约 1,230 亿美元，同比增长 20%，我们判断主要系高 ASP 的 AI 服务器出货拉动。根据 TrendForce 数据显示，2022 年搭载 GPGPU 的 AI 服务器出货量占整体服务器约 1%。

图 6：2002-2022 年全球服务器出货量及增速



资料来源：彭博，长江证券研究所

图 7：2002-2022 年全球服器市场规模及增速

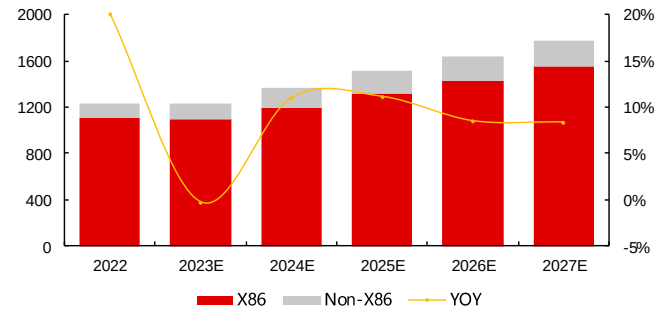


资料来源：彭博，长江证券研究所

根据 IDC 预测数据，2023 年全球服务器市场规模同比几乎持平，2024 年及以后服务器市场将保持 8-11%区间增速，预计到 2027 年市场规模达到 1,780 亿美元。此前

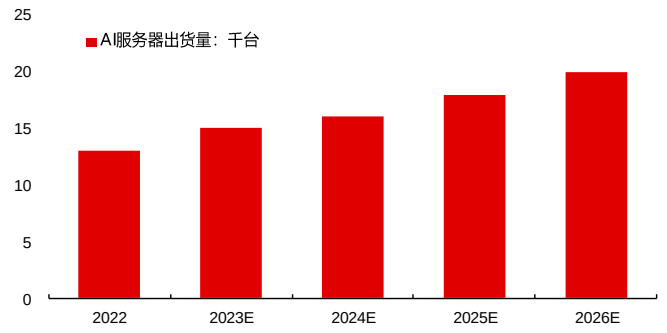
TrendForce 预计全球 AI 服务器出货量 2022-2026 年复合增速有望达到 10.8%，考虑到 AIGC 行业的飞速发展，实际增速或更高。

图 8：2022-2027E 全球服务器市场规模预测（亿美元）



资料来源：IDC，长江证券研究所

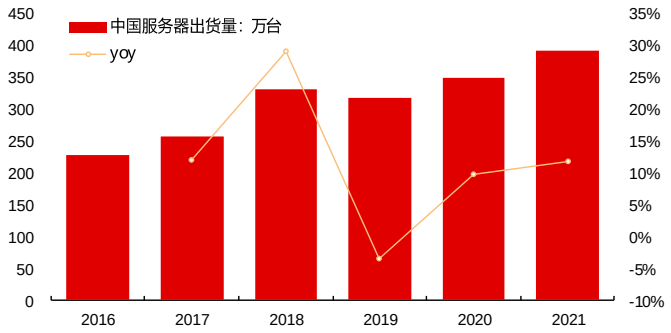
图 9：2022-2026E 全球 AI 服务器出货量预测



资料来源：TrendForce，长江证券研究所

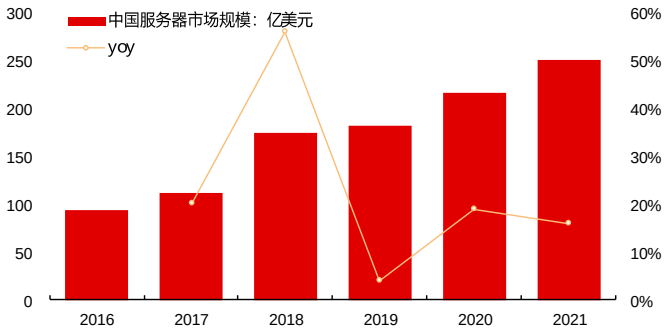
2021 年国内服务器市场出货量为 391.1 万台，同比增长 11.7%，对应市场规模为 250.9 亿美元，同比增长 15.9%。

图 10：中国服务器市场出货量及增速



资料来源：华经情报网，长江证券研究所

图 11：中国服务器市场规模及增速



资料来源：华经情报网，长江证券研究所

Q：普通服务器和 AI 服务器的区别

普通服务器与 AI 服务器的显性区别在于计算能力的不同，AI 服务器相比普通服务器在图形处理以及高性能计算方面表现更为突出，适合应用在对于算力要求更高的深度学习领域。AI 服务器多采用异构形式，如 CPU+GPU、CPU+TPU、CPU+其他的加速卡等，单台 AI 服务器 GPU 卡的用量多为 4 张以上，普通 GPU 服务器则多为单卡或双卡，由于 AI 服务器增加了 GPU 的使用数量，相配套的在带宽、散热、内存、存储等方面也会有相应升级。

表 1：浪潮信息通用服务器与 AI 服务器技术规格对比

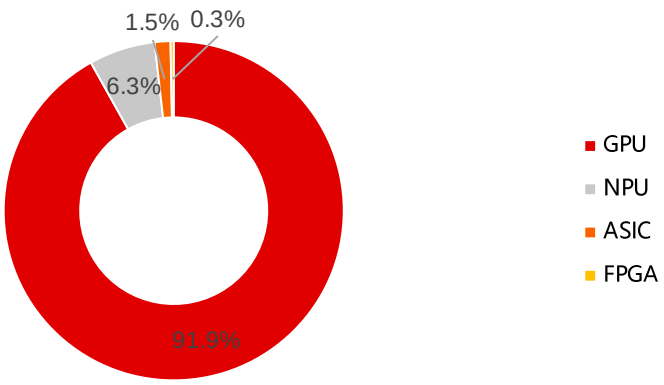
规格	NF5688M6（AI 服务器）	NF5180A6（通用服务器）
GPU 计算模块	1* HGX 8GPU（Ampere 架构）	
处理器	2 颗第三代 Intel® Xeon® 可扩展处理器(Ice Lake)，TDP 270W，支持 3 条 UPI 互联	支持 1 到 2 个 AMD®EPYCTM Milan 系列处理器： 最多支持 64 核（主频 2.45GHz） 最高主频 3.7GHz（8 核） 单 CPU 256MB L3 缓存和 18 GT/s XGMI2 互连链路 最大热设计功率 280W

内存	支持 32 条 DDR4 RDIMM/LRDIMM 内存，速率最高支持 3200MT/s	最大支持 32 根内存. 每个处理器支持 8 个内存通道，每个通道最大支持 2 个内存插槽. 内存最大速度可达 3200MT/s (1DPC).支持 RDIMM 与 LRDIMM 内存. 内存保护支持 ECC，内存镜像，内存等级保护
存储	8 块 2.5 英寸 NVMe SSD or SATA/SAS SSD 16 块 2.5 英寸 SATA/SAS SSD	前置 4*2.5”+4*3.5” SATA/SAS/NVME 支持热插拔 10*2.5” SATA/SAS/NVMe 支持热插拔 内置 可选 2 个 SATA M.2 或 1 个 PCIe M.2+1 个 SATA M.2
PCIe 扩展	10 个 PCIe 4.0 x16 插槽， 2 个 PCIe 4.0 x16 插槽 (PCIe 4.0 x8 速率)， 1 个 OCP3.0 插槽 6 个 PCIe 4.0 x16 插槽， 1 个 OCP3.0 插槽	最大支持 3 个 PCIe 插槽 支持 2 个单宽 GPU； 支持 1 个 OCP3.0 x16 热插拔网卡
网络	可选配 1 张 PCIe 4.0 x16 OCP 3.0 网卡，速率支持 10G/25G/100G	1 个可选 OCP3.0 模块支持 10Gb/s,25Gb/s,40Gb/s, 100Gb/s, 200Gb/s 网卡 2 个可选板载 1G 电口 支持标准 1Gb/10Gb/25Gb/40G/100Gb 网卡
前置 I/O	1 个 USB 3.0 端口，1 个 USB 2.0 端口，1 个 VGA 端口，1 个 RJ45 管理口	1 个前置 VGA，1 个前置 USB 3.0
后置 I/O	1 个 USB 3.0 端口，1 个 VGA 端口	1 个后置 VGA，2 个后置 USB3.0
远程管理	内置 BMC 远程管理模块，支持 Redfish/IPMI/SOL/KVM 等	集成 1 个独立的 1000Mbps 网络接口，专门用于 IPMI 的远程管理
操作系统	Red Hat Enterprise 7.8 64bit、CentOS 7.8、Ubuntu 18.04 或更高版本	Microsoft Windows Sever、Red Hat Enterprise Linux、SUSE Linux Enterprise Server、CentOS 等
散热	N+1 冗余热插拔风扇	8 个热插拔 N+1 冗余风扇
电源	6 块 3000W 80Plus 铂金电源，支持 3+3 冗余	支持 800W/1300W CRPS 标准 1+1 冗余电源
机箱尺寸	宽 447mm，高 263.9mm，深 850mm	含挂耳：W（宽）482mm；H（高）43.05mm；D（深）811.8 mm 不含挂耳：W（宽）438mm；H（高）43.05mm；D（深）780 mm
工作温度	10℃~35℃/50℉~95℉	5℃-45℃
满配重量	≤88kg	满配<31kg

资料来源：浪潮信息公司官网，长江证券研究所

从国内 2021 年 AI 服务器加速卡类别来看，目前主要是以 GPU 芯片作为数据中心加速卡，占比高达 9 成以上，GPU 在性能以及通用性层面更好，ASIC 则在功耗及效率方面占优，预计未来非 GPU 占比有望提升，2025 年有望超过 20%。

图 12：2021 年中国 AI 服务器加速卡类别



资料来源：华经产业研究院，长江证券研究所

英伟达作为全球算力“卖铲人”，全球 AI 服务器大量使用其 GPU，英伟达针对数据中心领域推出的 Tesla 系列产品，包括 P100、V100、T4、A100、H100 等系列，主导了 AI 训练及推理芯片市场。作为面向算力需求指数级上升的 HPC 和数据中心市场的 H100 芯片，其算力相较于 A100 倍数提升，可以大幅缩短深度学习模型的训练时间。在应用于大模型的 AI 训练时，H100 相较于 A100 快 9 倍，在用于大模型推理时，H100 的吞吐量达到了 A100 芯片的 30 倍。

表 2：英伟达 V100、A100 和 H100 芯片参数对比

	HI00	A100(80GB)	V100
FP32 CUDA Cores	16896	6912	5120
Tensor Cores	528	432	640
Boost Clock	~1.78GHz (Not Finalized)	1.41GHz	1.53GHz
Memory Clock	4.8GbpsHBM 3	3.2GbpsHBM2e	1.75GbpsHBM 2
Memory Bus Width	5120-bit	5120-bit	4096-bit
Memory Bandwidth	3TB/sec	2TB/sec	900GB/sec
VRAM	80GB	80GB	16GB/32GB
FP 32 Vector	60TFLOPS	19.5TFLOPS	15.7TFLOPS
FP 64 Vector	30TFLOPS	9.7TFLOPS (1/2FP 32 rate)	7.8TFLOPS (1/2FP 32 rate)
INT 8 Tensor	2000TOPS	624TOPS	N/A
FP 16 Tensor	1000TFLOPS	312TFLOPS	125TFLOPS
TF 32 Tensor	500TFLOPS	156TFLOPS	N/A
FP 64 Tensor	60TFLOPS	19.5TFLOPS	N/A
互联	NV Link 4 18 Links(900GB/sec)	NV Link 3 12 Links(600GB/sec)	NV Link 2 6 Links(300GB/sec)
GPU	GH100 (814mm2)	GA100 (826mm2)	GV100 (815mm2)
晶体管数量	80B	54.2B	21.1B
功耗	700W	400W	300W/350W
制造工艺	TSMC4N	TSMC7N	TSMC12nmFFN
接口	SXM 5	SXM 4	SXM 2/SXM 3
架构	Hopper	Ampere	Volta

资料来源：CSDN，长江证券研究所