南京证券 NANJING SECURITIES

行业评级：推荐

AIGC对光通信行业的影响

摘要：

- 本轮AI浪潮由ChatGPT掀起，并引发中外科技企业展开对大语言模型及生成式AI的追逐和对算力的军备竞赛，无论是传统龙头企业还是新进入者都在积极推出自家大模型，加大算力设备投资。展望2024年，多模态大模型与应用生态圈将成为AIGC的关键词，训练与推理需求的超预期增长将驱动算力硬件设备加速建设。
- AI服务器及相关硬件直接受益于算力需求增长，数据通信需求提升驱动光模块容量、数量双增长。单体服务器方面，为满足下游算力需求，芯片龙头企业加速产品迭代抢占市场份额。各家旗舰算力产品中，算力芯片异构、显存带宽扩容驱动服务器数据吞吐需求提升，光模块传输速率从400G向800G升级势在必行。算力集群组网方面，无阻塞网络架构、算力网络GPU全互联、计算单元规模提升，使得组网所需要的光模块、交换机在数量上显著提升。
- 投资建议：我们对光模块24年增长保持乐观，我们预计24年100G以上数通光模块行业规模增速在45%左右。相较于200/400G光模块驱动的行业周期，是由单个或少数云厂商资本开支驱动。而此次800G光模块是所有厂商参与测试，是服务器技术升级的关键节点。23年800G光模块是处于验证阶段，只有少量光模块企业有出货。24年伴随着英伟达H100的交付，800G光模块出货量将快速释放，标的估值得到消化。并伴随AI应用落地，推理算力需求预期提升，驱动光模块行业依然处于上行周期。重点关注：中际旭创、新易盛、光迅科技、天孚通信。
- 风险提升：AIGC应用推广不及预期；全球头部云服务厂商资本开支不及预期；800G光模块推广不及预期；行业竞争加剧，毛利率大幅不及预期。

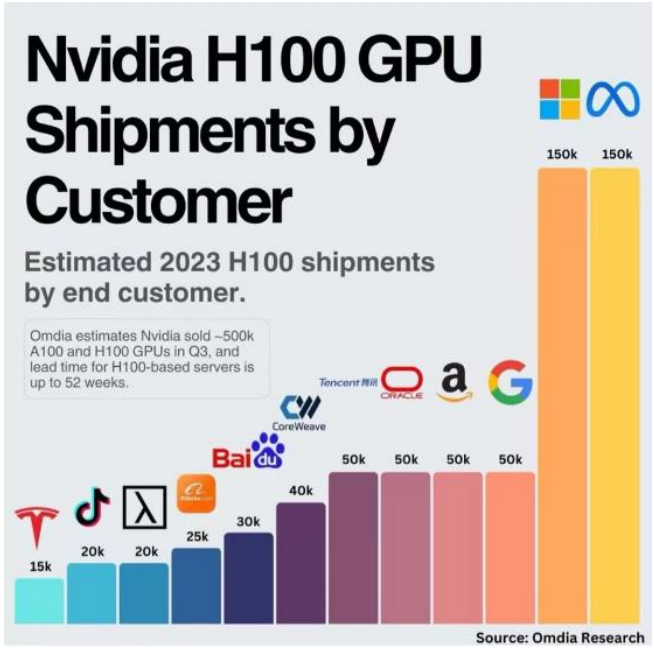
目录

- 1.0 行业背景：大模型驱动算力需求
  - 1.1 多模态大模型驱动训练算力需求增长
  - 1.2 大模型应用落地驱动推理算力需求增长
- 2.0 算力芯片与服务器发展趋势
  - 2.1 服务器性能迭代加速驱动数据吞吐带宽提升
  - 2.2 服务器组网结构升级驱动光通信设备零部件需求数量提升
- 3.0 光模块行业
  - 行业特征与技术发展趋势
  - 市场规模测算与投资建议
- 4.0 风险提示

1.0 行业背景

- 本轮AI浪潮由ChatGPT掀起，并引发中外科技企业展开对大语言模型及生成式AI的追逐和对算力的军备竞赛。无论是传统互联网企业还是新进入者都在加大算力投资，积极推出大模型。
- 我们认为多模态大模型与应用生态圈将成为2024年生成式AI的关键词，生成式AI依然处于行业发展的初期，多模态大模型训练与推理需求双驱动加速算力基础设施建设。

英伟达H100买家

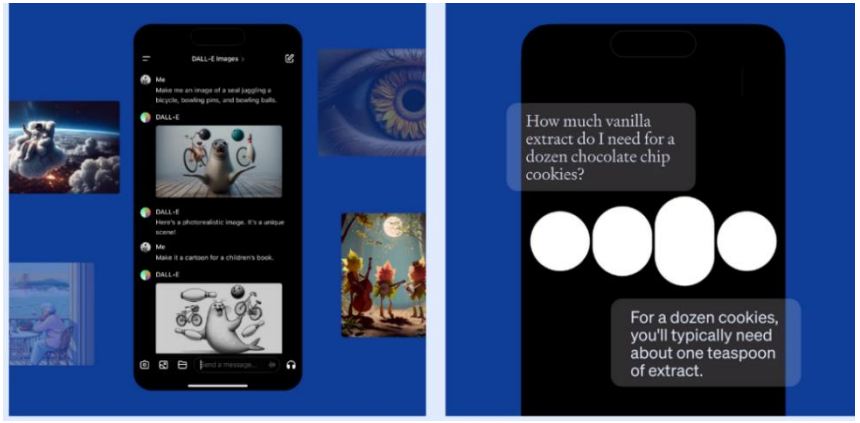


资料 ：财联社、南京证券研究所

1.1 多模态大模型，AI感知能力更加丰富

- 11月6日，OpenAI公布GPT-4 Turbo，平台提供多模态功能，包括视觉、图像创造（DALL.E3）和文本转语音（TTS）。
- 12月7号，谷歌发布AI多模态模型Gemini 1.0。Gemini 1.0根据不同尺寸进行了优化，分别是Ultra、Pro和Nano。Gemini Ultra适用于高度复杂的任务，Gemini Pro是通用版，而Gemini Nano则被用于各种客户端设备。Gemini 1.0可以同时理解并识别本文、图形、音频等。
- 伴随着大模型处理数据的类型持续扩展，多模态大模型（LMMS）成为主流，这使得未来大模型参数与训练集规模将持续扩大。当前多模态系统主要处理文本、图片和语音。伴随行业发展，大模型将可以处理视频、音乐、3D等内容。这也意味着大模型训练算力需求的增长将超预期。

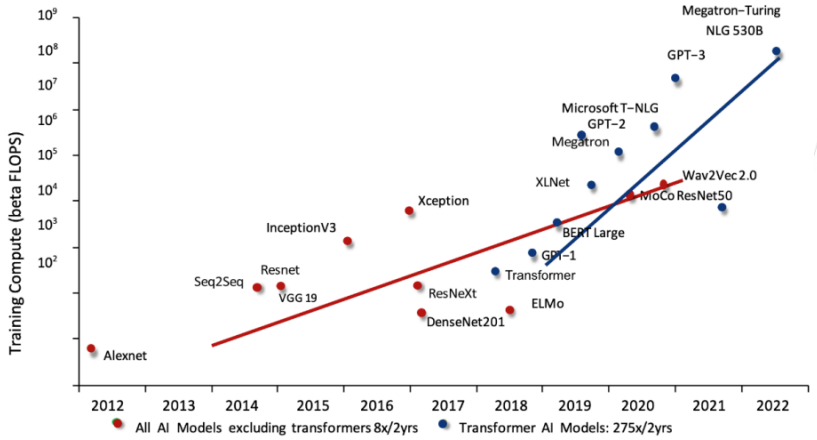
ChatGPT 4



谷歌Gemini 1.0



大模型训练对算力需求



资料 ：OpenAI官网、谷歌官网、 研究、南京证券研究所

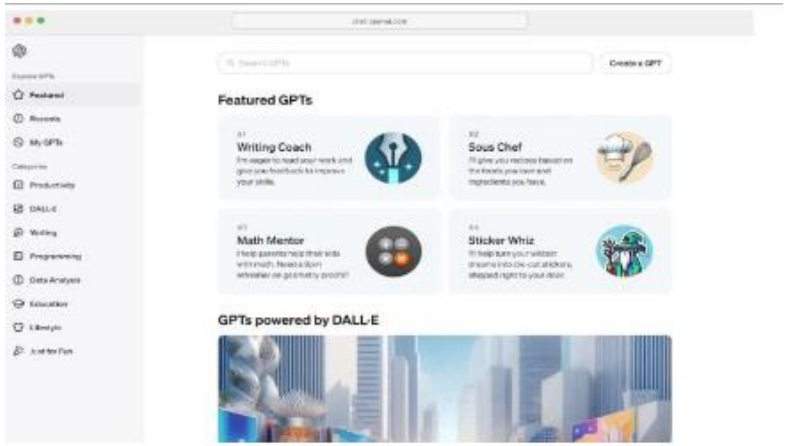
1.2 应用生态圈趋于成熟，推理算力需求显著增长

- OpenAI发布了GPTs，让用户们无需代码，结合自己的需求、外部知识和能力创造自定义版本的GhatGPT，满足了用户的定制化需求。此外，OpenAI在11月底上线GPT store，让开发者们能够分享、发布自己创建GPTs。GPTs和GPT Store的上线满足了客户的特色化需求，有望加速GPT应用生态建设，进一步提高用户黏性。
- GPTs降低了制作大模型应用门槛，用户无需编程基础，用自然语言就能做出专属GPTS，从而加速大模型向个人用户和垂直行业渗透。GPT应用量的增长，也意味着“杀手级”应用出现的可能性提升。大模型应用所带来的推理算力需求将会超出预期。

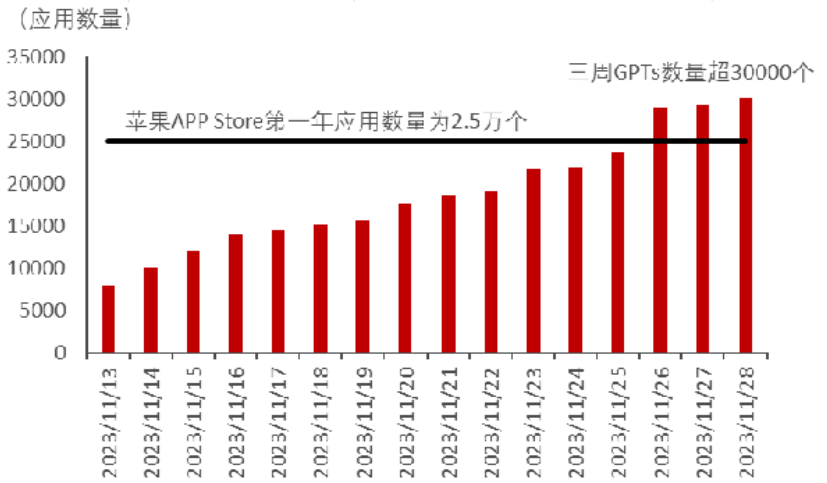
GPTs



GPT store



GPTs应用数量



资料　　：万联证券研究所、南京证券研究所

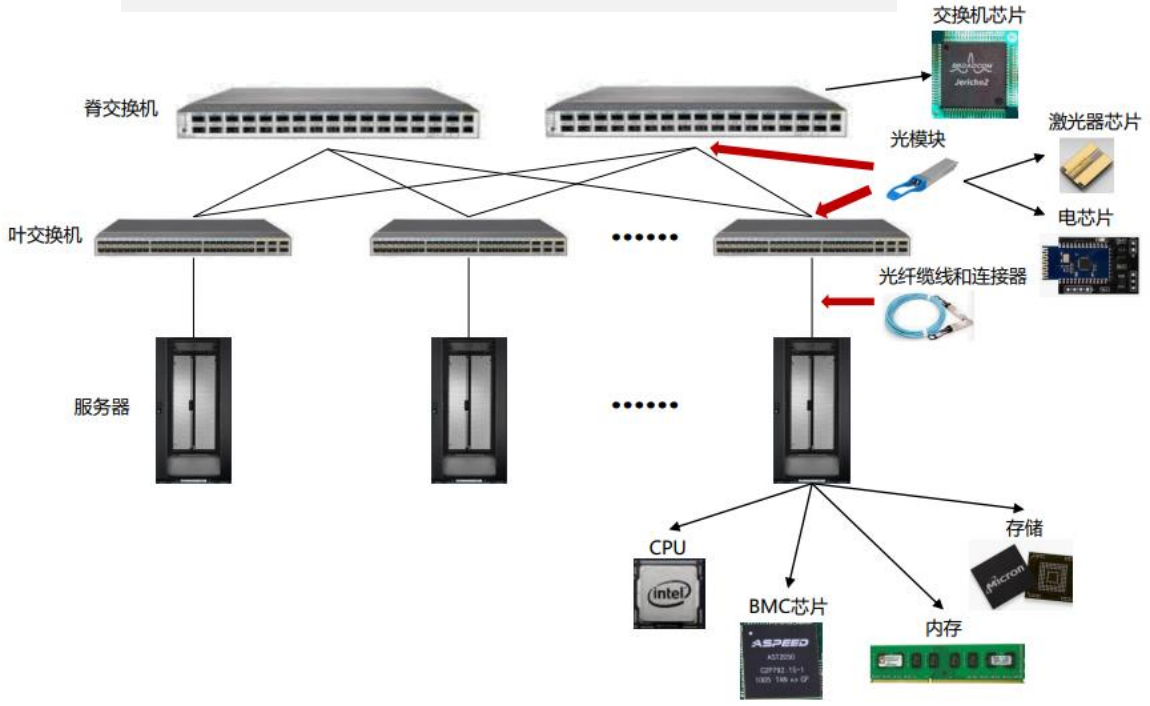
2.0 算力芯片与服务器发展趋势

- 大模型算力需求支撑AI服务器产业上行周期。算力芯片技术迭代加速，传统芯片厂商（英伟达、AMD、谷歌）以及新进入者（Amazon、Meta、微软）都在持续推出产品去抢占市场份额。服务器性能加速迭代对数据通信能力提出了更高的要求。
  - 技术迭代单体服务器数据吞吐需求提升。
  - 服务器之间组网结构变化驱动光模块、交换机需求提升。

自研AI芯片时间轴

时间	厂商	AI 芯片
2016 年 5 月	Nvidia	P100
2016 年 6 月	Google	TPU v1
2017 年 5 月	Nvidia	V100
2017 年 5 月	Google	TPU v2
2018 年 5 月	Google	TPU v3
2018 年 11 月	Amazon	Inferentia
2020 年 5 月	Nvidia	A100
2020 年 12 月	Amazon	Trainium
2021 年 5 月	Google	TPU v4
2022 年 3 月	Nvidia	H100
2023 年 4 月	Amazon	Inferentia 2
2023 年 5 月	Meta	MTIA
2023 年 8 月	Google	TPU v5
2023 年 11 月	Nvidia	H200
2023 年 11 月	Microsoft	Maia100

数据中心产业供应链

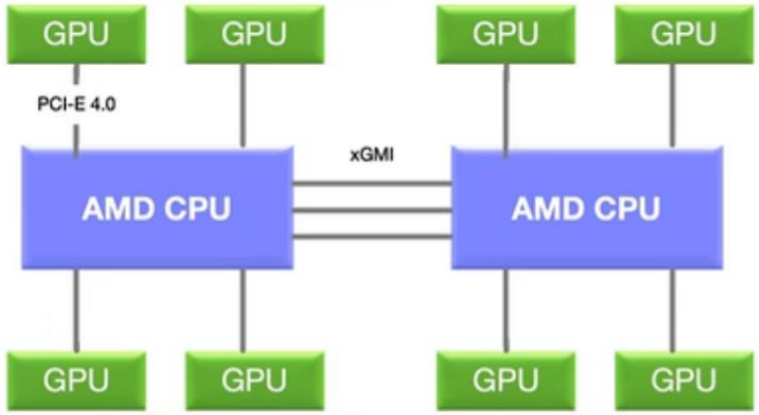


资料 ：长江证券研究所、申万宏源研究、南京证券研究所

2.1 传统标准GPU服务器内部架构

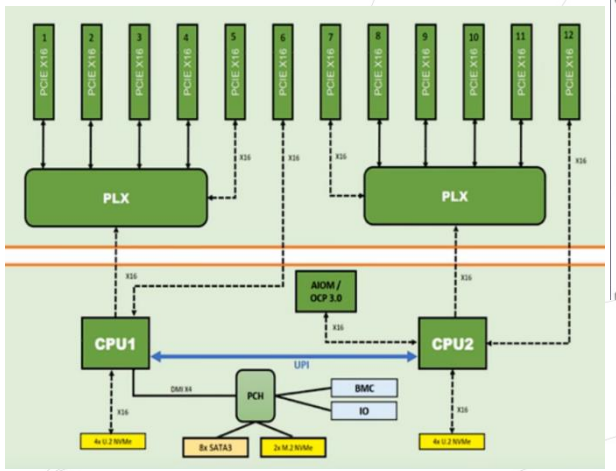
- PCIE直通：CPU PCIE line限制了服务器内显卡的数量：双路AMD milan中间通过三条xGMI链路，整机可提供128*2-32*3=160条PCIE line，8张CPU占用了16*8=128条PCIE通道，还有32条供其他网卡等部件使用。
  - 优势：成本较PLX方式的低；单个GPU独享x16，性能更优。
  - 劣势：4U只能支持8张双宽GPU卡；跨GPU无法直接点对点通信。
- PLX拓展：通过PLX芯片，将一个x16拓展成4-5个x16。
  - 优势：拓展性提升，4U支持10张双宽GPU卡；同PLX下GPU可点对点通信。
  - 劣势：成本较高；GPU上行总带宽收敛，性能有限制。

AMD Milan GPU服务器的内部拓扑





Intel icelake GPU服务器拓扑





资料 ：It server技术分享、南京证券研究所

2.1 NVLink技术、NVSwitch芯片提升芯片互联带宽打破PCIe限制

- Nvlink 4代的总带宽可达到900GB/s,为PCIe 5.0总线带宽的7倍，对比下一代PCIe 6.0的256GB/s也有显著优势。
- NVSwitch是英伟达的节点交换架构，通过连接多个NVLink，在单节点内和节点间实现多GPU的拓展。第三代NVSwitch能在带点服务器节点中支持8-16个完全链接的GPU，支持以900GB/s的速度互联每个GPU。
- 英伟达H100使用第四代 NVLink 和第三代 NVSwitch ，具有八个 NVIDIA H100 Tensor Core GPU 的系统具有 3.6TB/s的二等分带宽和450GB/s 的缩减操作带宽。与上一代相比，这两个数字分别增加了1.5倍和3倍。英伟达摆脱了PCIe在原始带宽、延迟以及缓存一直性方面的限制。通过NVLink和NVSwitch实现多个GPU大规模集群的传输，实现更快和更可拓展的计算系统。

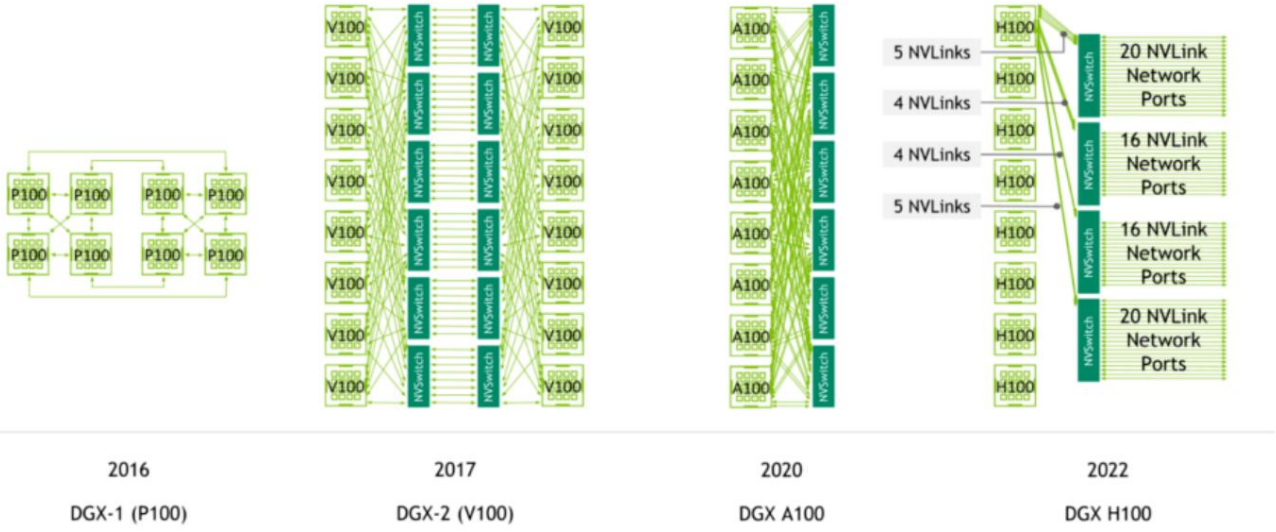
NVLink和NVSwitch性能指标

NVLink链路	第二代	第三代	第四代
NVLink 总带宽	300GB/s	600GB/s	900GB/s
每个 GPU 最大链路数	6	12	18
NVIDIA 架构支持	NVIDIA Volta™ 架构	NVIDIA Ampere 架构	NVIDIA Hopper™ 架构

NVSwitch芯片	第一代	第二代	第三代
直连或节点中 GPU 数量	最多 8 个	最多 8 个	最多 8 个
NVSwitch GPU 之间带宽	300GB/s	600GB/s	900GB/s
聚合总带宽	2.4TB/s	4.8TB/s	7.2TB/s
NVIDIA 架构支持	NVIDIA Volta 架构	NVIDIA Ampere 架构	NVIDIA Hopper 架构

资料 ：NVIDIA、南京证券研究所

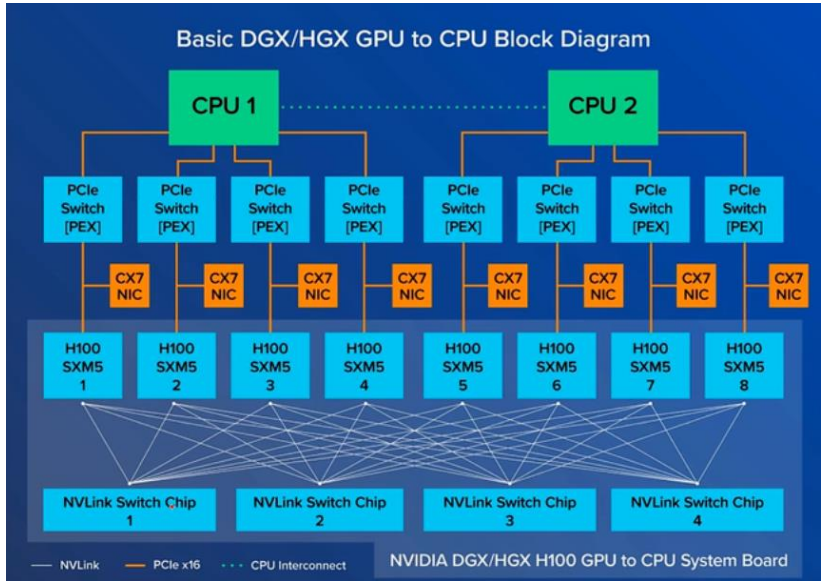
NVLink和NVSwitch协同



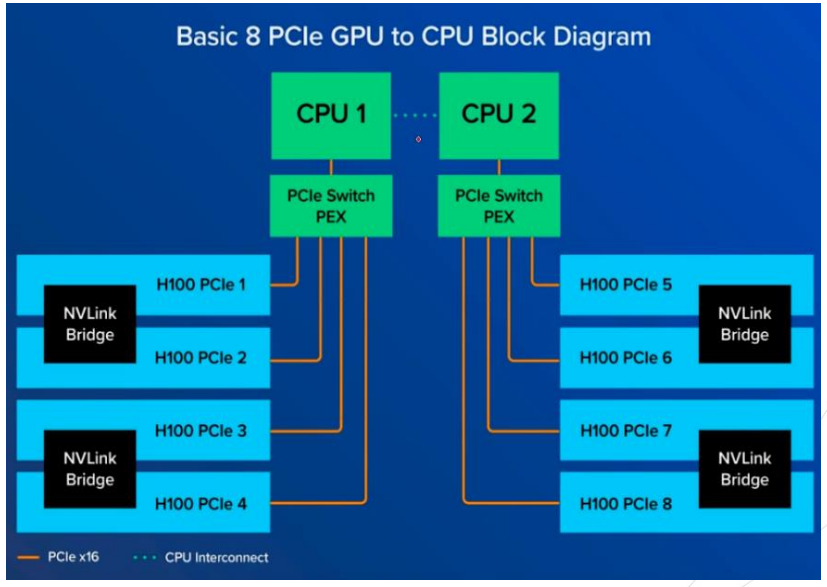
2.1 英伟达DGX服务器实现内部GPU全互联

- 以GTC SPRING 2022 发布的DGX H100为例，H100 分为SXM和PCIe两个版本。主流大模型依赖多卡协同，GPU之间的带宽要求较高，Nvlink全互联的GPU更加适合大模型应用场景。
  - SXM版本中8张H100芯片通过4张NVLink Switch芯片实现全互联。GPU之间带宽高达900GB/s（双向链路25GB/s*2*18条=900GB/s），相较于PCIE 5.0x16双向带宽128GB/s，互联速度得到快速提升。
  - PCIe版本中4张H100芯片通过PCLe Switch PEX4：1连接到CPU，2张H100芯片通过Nvlink Bridge互联。PCIE机型更加灵活，GPU卡的数量以及PCIE的拓扑可以进行调整。

NVIDIA DGX/HGX GPU和CPU互联架构



H100 PCIe GPU



H100 SXM vs PCIe

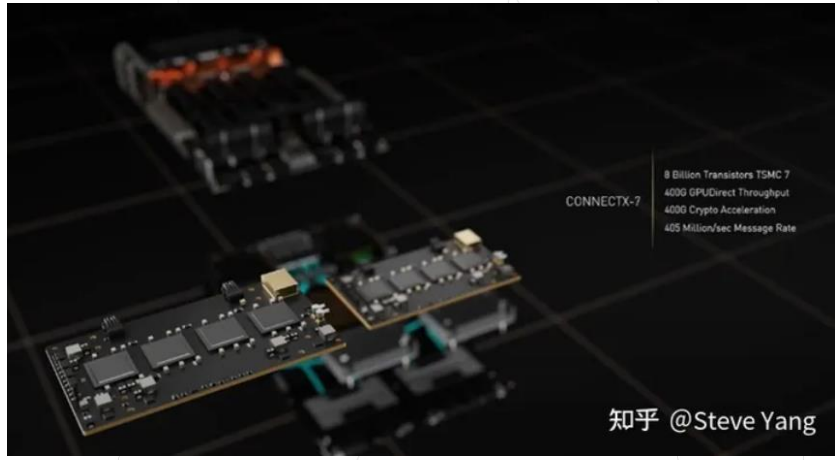
外形规格	H100 SXM	H100 PCIe
FP64	30 teraFLOPS	24 teraFLOPS
FP64 Tensor Core	60 teraFLOPS	48 teraFLOPS
FP32	60 teraFLOPS	48 teraFLOPS
TF32 Tensor Core	1000 teraFLOPS*	800 teraFLOPS*
BFLOAT16 Tensor Core	2000 teraFLOPS*	1600 teraFLOPS*
FP16 Tensor Core	2000 teraFLOPS*	1600 teraFLOPS*
FP8 Tensor Core	4000 teraFLOPS*	3200 teraFLOPS*
INT8 Tensor Core	4000 TOPS*	3200 TOPS*
GPU 显存	80GB	80GB
GPU 显存带宽	3TB/s	2TB/s
解码器	7 NVDEC 7 JPEG	7 NVDEC 7 JPEG
最大热设计功耗 (TDP)	700W	350W
多实例 GPU	最大为 7 MIG @ 10GB	
外形规格	SXM	PCIe，双插槽风冷式
互连	NVLink：900GB/s PCIe 5.0：128GB/s	NVLink：600GB/s PCIe 5.0：128GB/s
服务器选项	NVIDIA HGX™ H100 合作伙伴和配备 4 块或 8 块 GPU 的 NVIDIA 认证系统™ 配备 8 块 GPU 的 NVIDIA DGX™ H100	配备 1 至 8 个 GPU 的 NVIDIA 认证系统及合作伙伴服务器

资料 ：It server技术分享、NVIDIA、南京证券研究所

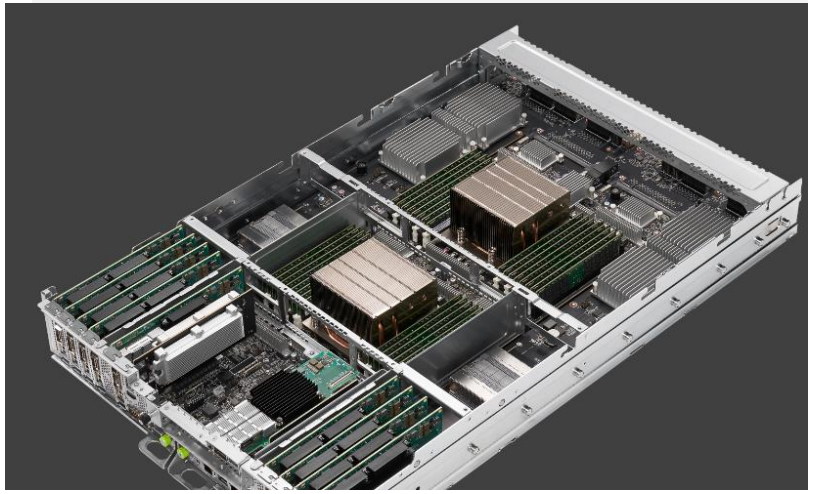
2.1 英伟达DGX服务器计算网络设计优化

- DGX H100在服务器内部通信构架中可以看出计算网络的重要性。相较于DGX A100，DGX H100正在放弃传统的PCIe网卡，转而使用“Cedar”的模块进行GPU与GPU之间的互联。DGX H100通过2个Cedar7模组，每个模组包含4个ConnextX-7 ICs，通过Densilink电缆连接到服务器的对外接口。虽然8张GPU配备了8张400G的网卡ICs，但接口合成成了4个800G接口。
- 优势：1.使用Cedar模块可以提升服务器空间效率，同时有助于服务器内部空气流通。2.Cedar模块直接通过电缆对外连接，有助于降低PCB设计复杂度与成本。

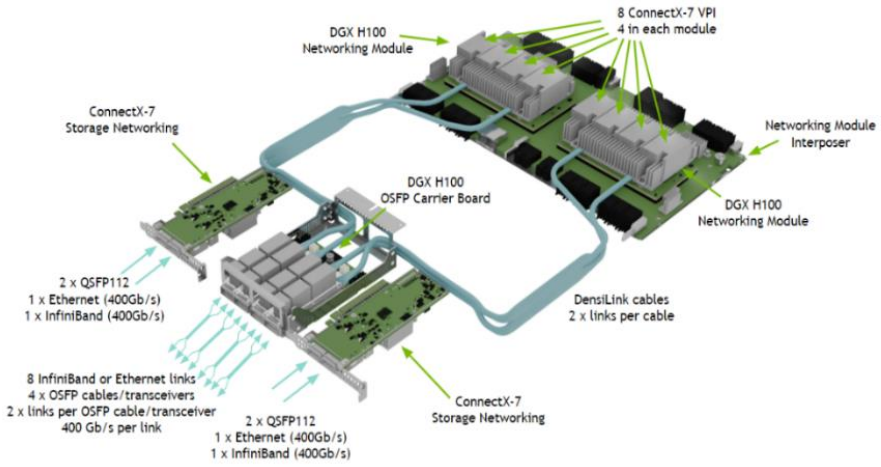
DGX H100 Cedar7



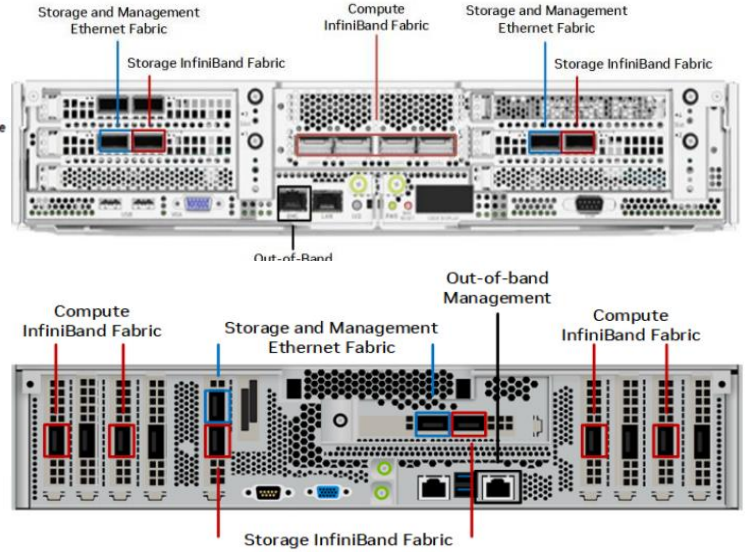
DGX A100



DGX H100 计算与存储网络



DGX H100/A100 接口



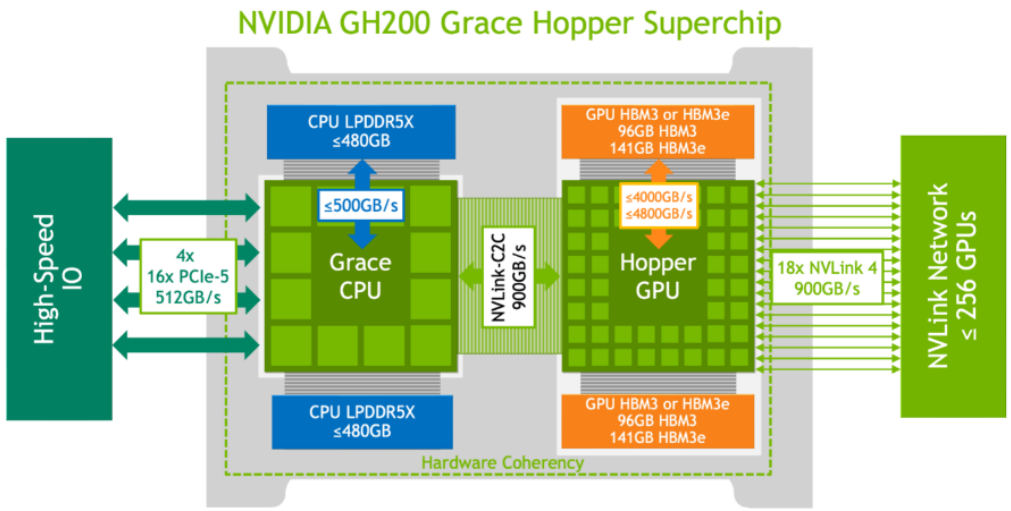
资料 ：NVIDIA、知乎Steve Yang、南京证券研究所

2.1 英伟达DGX服务器CPU与GPU实现互联

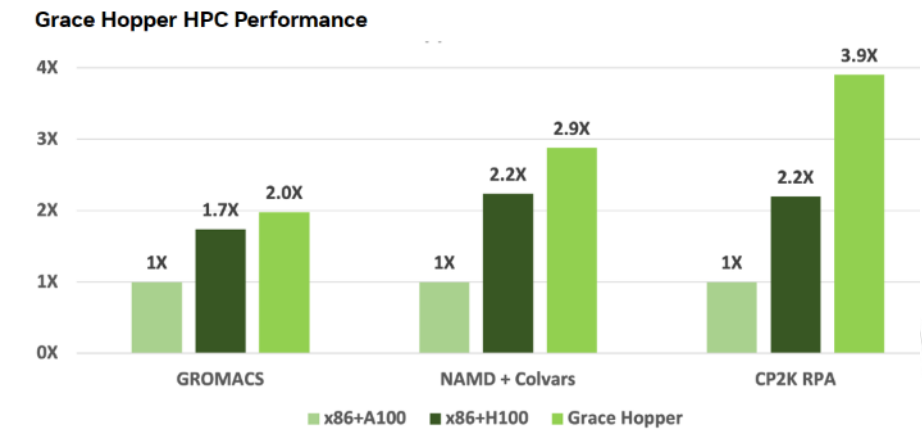
- Hopper架构中，CPU和GPU从原本的PCIe连接转变为了Nvlink传输，是英伟达的第一个异构加速平台，适用于HPC和AI工作负载。
- NVIDIA NVLink-C2C是一种内存连贯、高带宽和低延迟超级芯片互联。它允许定制芯片可以便利地连接到NVIDIA芯片。解决AI和HPC问题需要高容量和高带宽内存（HBM）。此项互联技术可以提供统一、缓存一致性的内存地址空间，后者将系统和HBM GPU 显存结合在一起，能够优化可编程性。CPU和GPU之间这种一致性、高宽带的连接是解决复杂AI和HPC问题的关键。

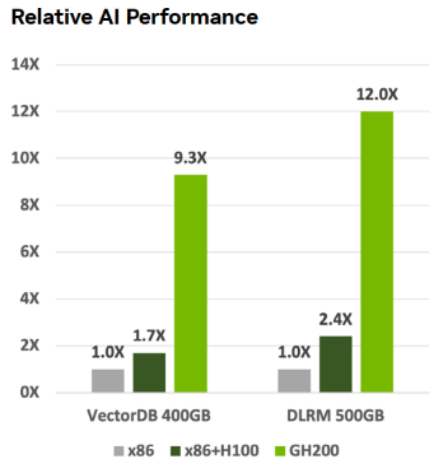
NVIDIA Grace Hopper 超级芯片逻辑概述

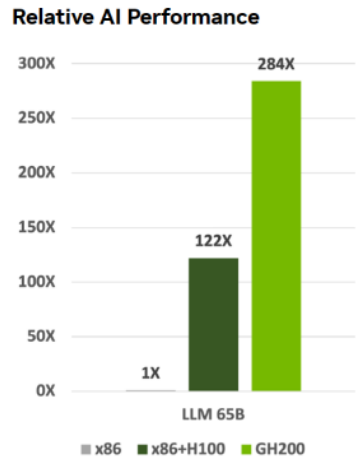




Grace Hopper 超级芯片性能表现





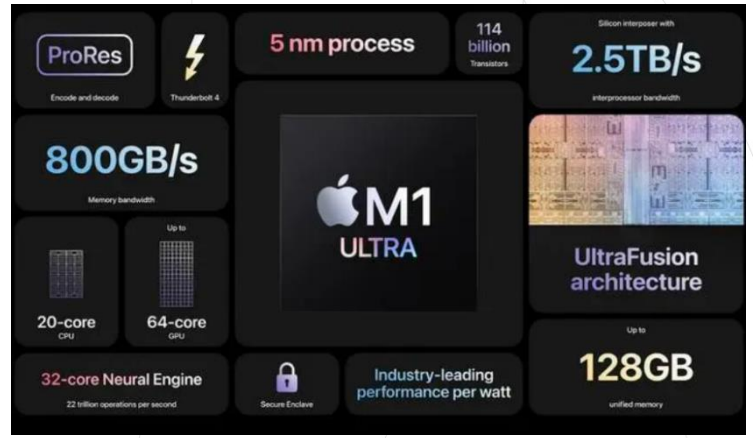


资料　　：NVIDIA、南京证券研究所

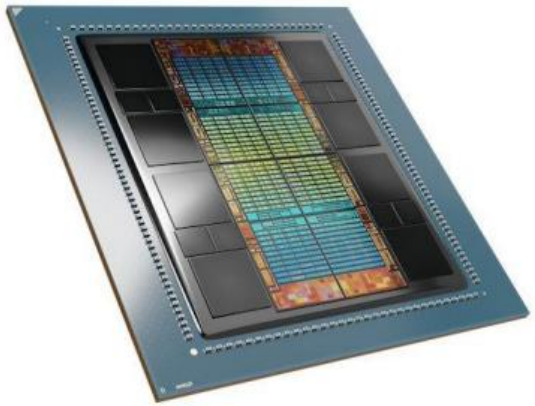
2.1 Chiplet和异构芯片成为行业发展的趋势

- Chiplet：AMD早在2011年发布了APU已经是CPU+GPU架构。在2017年，AMD发布的论文《Design and Analysis of an APU for Exascale Computing》中讨论包含CPU、GPU和HBM内存堆栈的APU芯片设计。在2023年6月推出了AMD首个CPU+GPU的数据中心产品。
- UltraFusion：M1 Ultra采用台积电5nm工艺，由两块M1 Max芯片拼接而成，这样的“组合”使得其晶体管数量达到1140亿颗。使得各项硬件指标翻倍：20 核 CPU 包含16个高性能核心、4个高能效核心；GPU核心数量增至64核；32核神经网络引擎能够带来每秒22万亿次的运算能力；统一内存最高规格达到 128GB；内存带宽提升至 800GB/s。这种多芯片（Multi-die）配置主要靠UltraFusion多晶粒架构，使用2.5D先进封装技术。
- 异构芯片互联的趋势，逐步摆脱了传统SoC的束缚，芯片面积提升不一定意味着成本大幅提升与良品率大幅下降。

M1 Ultra



AMD MI300X （8个GPU Chiplets）



英伟达和AMD主要GPU相关产品参数对比

	英伟达		AMD		
产品名称	A100 PCIe | SXM	H100 PCIe | SXM | NVL	MI250X	MI300A	MI300X
发布时间	2020.6	2022.3	2021.11	2023.1	2023.6
峰值算力（TFLOPS）	FP16: 312 | 624 FP32: 19.5 FP64: 19.5	FP8: 3,026 | 3,958 | 7,916 FP16: 1,513 | 1,979 | 3,958 FP32: 51 | 67 | 134 FP64: 51 | 64 | 134	FP16: 383 FP32/64: 47.9 FP32/64 Matrix: 95.7	暂无	暂无
工艺制程及芯片面积	7nm, 826mm^2	4nm, 814 mm^2	6nm, 724 mm^2	5nm, 1017 mm^2	5nm, 1017 mm^2
晶体管数量（亿）	540	800	582	1460	1530
内存容量	80 GB HBM2e	80 | 80 | 188 GB HBM3	128 GB HBM2e	128 GB HBM3	192 GB HBM3
内存带宽	1.9 | 2TB/s	2 | 3.35 TB/s | 7.8TB/s	3.2 TB/s	3.2TB/s	5.2TB/s
Interconnect	600 GB/s NVLink for 2 GPUs 64 GB/s PCIe Gen4	600 | 900 | 600 GB/s NVLink 125 GB/s PCIe Gen5	100GB/s	约 800GB/s	896GB/s
热设计功耗 TDP（W）	300 | 400	300-350 | 700 | 2x350-400	500	600	暂无

资料 ：AMD、NVIDIA、APPLE、 研究、南京证券研究所

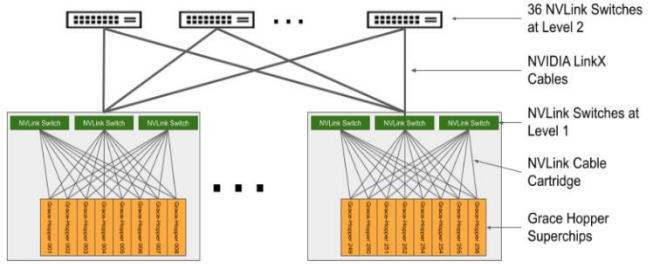
2.1 NVIDIA DGX GH200-256张Grace hopper超级芯片实现互联

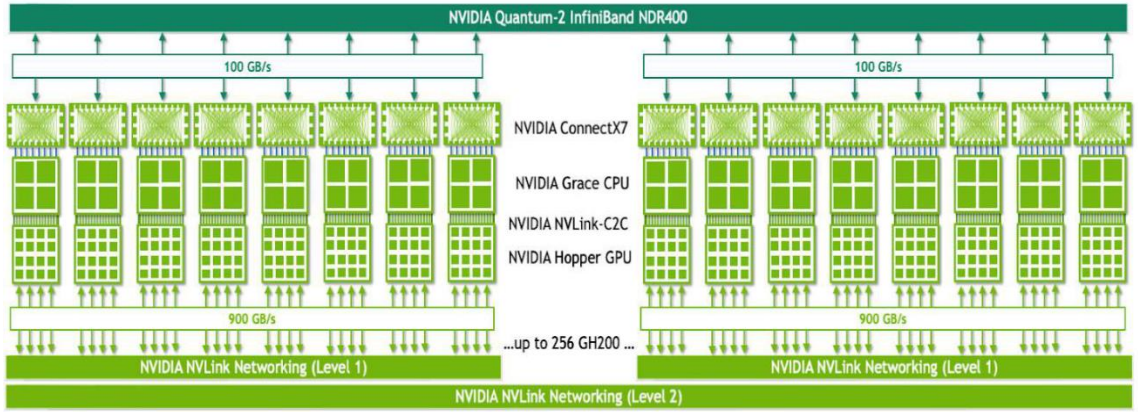
- 2023年5月28日，英伟达宣布推出NVIDIA DGX GH200，这是首款100TB级别的GPU内存系统。每8个Grace Hopper为一组，每组对应3台L1 Nvlink SW，256个节点分别为32组，需要32*3=96台 L1 NVLink Switches。96台L1 Nvswitch需要36台L2 Nvlink Switches。
- IB网络作为较为成熟的构架，每台Grace Hopper Superchip节点都配置了1个NVIDIA ConnectX-7单口400Gb/s网卡，和1个NVIDIA BlueField-4 NIC双口200Gb/s，实现IB全互联的同时，具备后续拓展能力。

NVIDIA DGX GH200







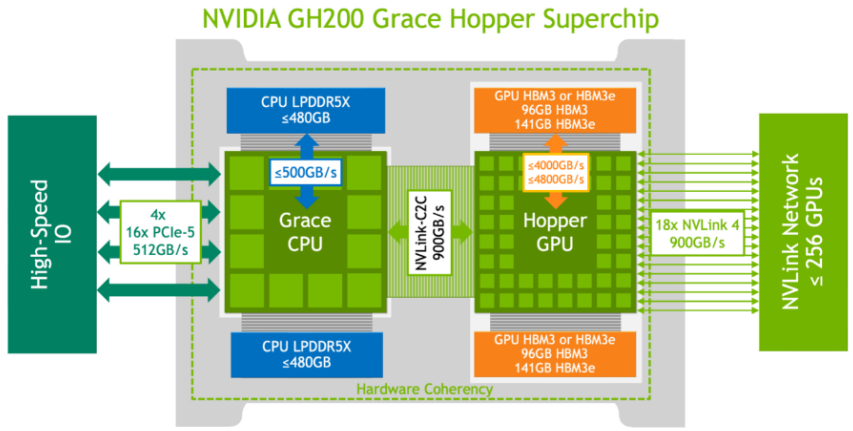


资料 ： It server技术分享、NVIDIA、南京证券研究所

2.1 NVIDIA DGX GH200-256张Grace Hopper超级芯片实现互联

- 从面板数据来看，显存144TB=256*（480GB内存（CPU Lpddr）+96GB显存（HBM）），之前从来没有把CPU内存当作GPU显存来算的。带宽230TB/s=256*900GM/s/1024。L1 NVSwitch对外提供3.2TB/s的双向带宽连接到L2 NVSwitch，256个GPU All-to-All总带宽为57.6TB/s，实际负载情况下900GM/s的带宽其实用不满。
- 从内存拓展、网络带宽配置和软件适配的角度，Grace Hopper架构中内存、网络延时问题，软件设计、客户使用生态问题都面临较大挑战。
- 面对ChatGPT的出圈，英伟达延续了22年发布的Grace Hopper去构建计算集群，该芯片设计场景是推荐系统。而大模型的特点就是内存占用大，每次训练与推理都要倒腾大量的数据，除了数据集这种冷数据外都是热数据。Grace Hopper超级芯片，CUP内存带宽处于面对热数据不够，面对冷数据浪费的尴尬情景。大模型涉及的热数据要尽可能多的放到HBM中，而不是单个GPU-CPU组合的HBM和LPDDR5X，所以提升显存带宽成为了必选项。

Grace Hopper 超级芯片



Grace&Hopper 性能参数

Grace CPU	Feature
CPU core count	72 Arm Neoverse V2 cores
L1 cache	64KB i-cache + 64KB d-cache
L2 cache	1MB per core
L3 cache	117MB
LPDDR5X size	Up to 480GB
Memory bandwidth	Up to 512GB/s
PCIe links	Up to 4x PCIe x16 (Gen5)

Hopper H100 GPU	Feature
FP64	34 teraFLOPS
FP64 Tensor Core	67 teraFLOPS
FP32	67 teraFLOPS
TF32 Tensor Core	989 teraFLOPS* | 494 teraFLOPS
BFLOAT16 Tensor Core	1,979 teraFLOPS* | 990 teraFLOPS
FP16 Tensor Core	1,979 teraFLOPS* | 990 teraFLOPS
FP8 Tensor Core	3,958 teraFLOPS* | 1,979 teraFLOPS
INT8 Tensor Core	3,958 TOPS* | 1,979 TOPS
High-bandwidth memory (HBM) size	Up to 96GB | 144GB HBM3e
Memory bandwidth	Up to 4TB/s | Up to 4.9TB/s HBM3e
NVIDIA NVLink-C2C CPU-to-GPU bandwidth	900 GB/s bidirectional
Module thermal design power (TDP)	Programmable from 450W to 1000W (CPU + GPU + memory)
Form factor	Superchip module
Thermal solution	Air cooled or liquid cooled

资料 ： It server技术分享、NVIDIA、南京证券研究所

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：https://d.book118.com/527032116015006030