

行业点评

基于 GPT 4.0 模型的通信基础设施需求测算

超配

◆ 公司研究 · 公司快评

证券分析师： 马成龙 021-60933150
证券分析师： 付晓钦 0755-81982929
联系人： 袁文翀 021-60375411

◆ 通信

machenglong@guosen.com.cn
fuxq@guosen.com.cn
yuanwenchong@guosen.com.cn

◆ 投资评级：超配（维持评级）

执证编码：S0980518100002
执证编码：S0980520120003

事项：

近期关于 GPT 模型的更新和进展不断，2022 年 12 月 Open AI 推出的基于 GPT3.5 模型的 chatgpt 应用在发布 2 个月就拥有了 1 亿用户，短短 3 个多月后，2023 年 3 月 14 日 Open AI 便发布了多模态 GPT 4.0 大模型，相较于 GPT 3.5 模型，GPT 4.0 模型支持图文回答并实现了回答度的大幅提升。3 月 25 日，ChatGPT 再次功能升级实现了对插件的初始支持，帮助 ChatGPT 访问最新信息、运行计算或使用第三方服务。国内，百度发布了文心一言大模型，华为盘古大模型也将迎来发布，行业催化不断。应用层面的快速进展会直接带动以服务器-交换机-光模块为代表的硬件基础设施的需求提升。

国信通信观点：我们尝试从硬件基础设施角度对于当前以 Chatgpt 为代表的 AI 应用需求带来的硬件的弹性进行测算，并得到以下几个重要结论：（1）在 AI 集群数据中心网络架构中，服务器：交换机：光模块对应比例为 1：1.2：11；（2）短维度：从当下 GPT 模型一次性角度测算，服务器、交换机、光模块的市场弹性分别为 15%、5%、3%；（3）长维度：下游需求亿级别大规模响应角度测算，服务器、交换机、光模块的市场弹性分别为 65%、19%、10%。

投资建议：以 Chatgpt 为代表的 AI 应用正在得到快速发展，并进行着快速迭代，进而有望赋能多行业、多样化新型应用落地，提升行业效率。软端的快速发展离不开硬件基础设施的保障，因此 AI 数据中心内部的各类的硬件基础设施环节有望充分受益 AI 行业进步带来的需求提升：建议重点关注国内 ICT 领先企业【紫光股份】、【浪潮信息】、【锐捷网络】、【中兴通讯】、【菲菱科思】；光模块及光器件环节【中际旭创】、【天孚通信】、【新易盛】；IDC 温控企业【英维克】、【申菱环境】；IDC 电源端企业：【科华数据】、【科士达】；第三方 IDC【奥飞数据】；建议关注：【同飞股份】、【佳力图】。

评论：

◆ 基于英伟达 AI 网络架构硬件需求比例：服务器：交换机：光模块=1：1.2:11.4

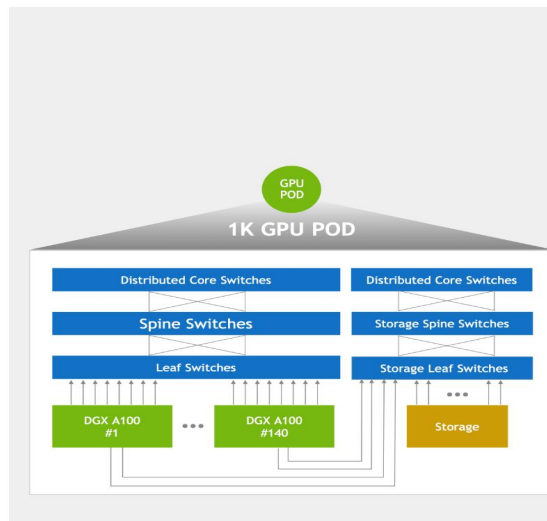
由于当前大部分 AIGC 模型都是基于英伟达方案来部署，我们从英伟达的 AI 集群模型架构进行拆解。

对于较大的 AI 数据中心集群，一般可多达几千台 AI 服务器的需求，在部署方面会拆分成一个个基本单元进行组件，英伟达对应的一个基本单元为 SuperPOD。

根据 SuperPOD 公开信息：一个标准的 SuperPOD 由 140 台 DGX A100 GPU 服务器、HDR InfiniBand 200G 网卡和 170 台 NVIDIA Quantum QM8790 交换机构建而成，其中交换机速率为 200G，每个端口数为 40 个。

图1: 英伟达 SuperPOD 基本结构

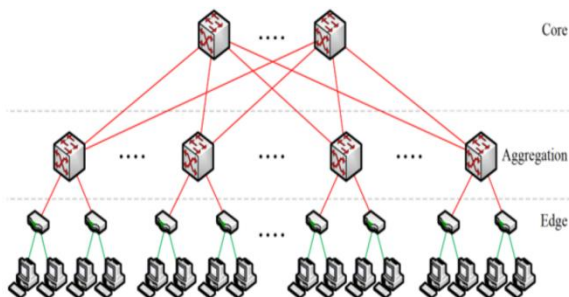
- ### DGX A100 SuperPOD
- A modular model
- 1K GPU SuperPOD Cluster**
- 140 DGX A100 nodes (1120 GPUs) in a GPU POD
 - 1st tier fast storage - DDN AI400x with Lustre
 - Mellanox HDR 200Gb/s InfiniBand - Full Fat-tree
 - Network optimized for AI and HPC
- DGX A100 Nodes**
- 2x AMD 7742 EPYC CPUs + 8x A100 GPUs
 - NVLINK 3.0 Fully Connected Switch
 - 8 Compute + 2 Storage HDR IB Ports
- A fast interconnect**
- Modular IB Fat-tree
 - Separate network for Compute vs Storage
 - Adaptive routing and SharpV2 support for offload



资料来源: 英伟达, 国信证券经济研究所整理

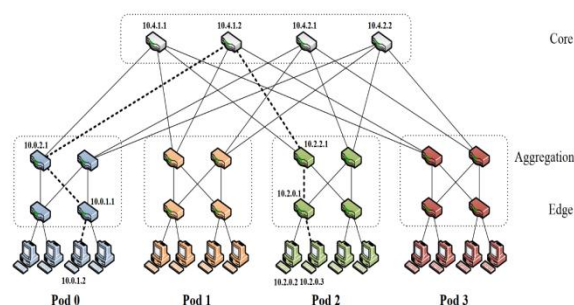
网络结构上, 英伟达采用 Infinband 技术 (“无限带宽” 技术, 简称 IB) 和 fat tree (胖树) 网络拓扑结构, 和传统的数据中心的区别在于, 在 IB fat tree 结构下, 使用的交换机数量更多, 且因为每个节点上行下行的端口数完全一致, 使得该网络是无收敛带宽的, **每个端口可采用同样速率的光模块**。

图2: 传统数据中心网络拓扑结构



资料来源: CSDN, 国信证券经济研究所整理

图3: Fat tree 网络拓扑结构



资料来源: CSDN, 国信证券经济研究所整理

光模块用量测算: 我们从线缆角度测算光模块需求, 一个 SuperPOD 170 个交换机, 每个交换机有 40 个端口, 最简单方式上下个 70 台服务器, 依次端口互联 (上下 1: 1 连接) 对应的线缆需求为 $40 \times 170 / 2 = 3400$ 根, 但是由于实际网络拓扑结构交换价不是该情况, 连接情况更加复杂且会分为三层结构, 因此线缆数需求有所提升, 我们假设上升至 **4000 根线缆需求**。

线缆的需求分为三种, 第一种用在机柜内部, 互联距离 5m 以内, 常用需求为铜缆, 不需要光模块; 第二类互联距离为 10m 以内, 可以采用 AOC (有源光纤) 连接, 也不需要光模块; 第三类, 带光模块的光纤, 单根需求为 2 个光模块。

考虑到 10m 以内的连接占据多数, 我们假设铜缆: AOC: 光模块光纤比例=4: 4: 2。

光模块需求=4000*0. 2*2=1600 个。

对于一个 SuperPod，服务器：交换机：光模块的用量比例=140：170：1600=1：1.2：11.4.

◆ 应用层面：单 GPT4.0 模型对于服务器需求用量测算

从用户使用角度来测算，我们对于服务器算力的测算受大模型参数，日活人数，每日每人提问等多因素影响。

在 ChatGPT 中，一个 token 通常指的是响应请求所需的最小文本单位，一般一个 30 词的提问大约对应 40 个 token，推理是 token 的算力调用是 2N。对应模型算力的需求我们分摊在一天 24h 的每一秒。

$$\text{服务器用量} = \frac{2N * \text{模型参数}}{\text{算力利用率} * \text{时间分摊} * \text{单台服务器算力}} \times \text{峰值需求倍数}$$

基于以下假设，我们可得到对应一个在 1 亿日活的应用需要的 AI 服务器的需求约为 1.5 万台。

表1：单 GPT4.0 服务器用量测算

假设	参数详情
模型 FLOPS 利用率	假设 20%（可提升）
峰值利用是平均算力	5 倍
单台 A100 服务器的算力	5 PTOPS
模型参数（GPT 4.0）	1 万亿+
用户月活（亿）	1 亿人次日活
假设每人每日提问（字）	1000 词
测算结果	
A100 服务器用量需求	约为 1.5 万台

资料来源：国信证券经济研究所整理

◆ 弹性测算一：现有采购部署层面弹性测算（基础投入维度）

角度 1 我们选择从现有完成一个类似 GPT4.0 入门级别要求的需求假设去测算硬件基础设施层面需求。

假设 1：结合现有各类公开数据，完成 ChatGPT4.0（训练+推理）需要至少 3 万张英伟达 A100 卡的算力投入，对应 3750 台 A100 的 DGX 服务器。

假设 2：全球假设国内和海外有潜在 20 家公司可能按照此规模进行测投入。

假设 3：网络结构比例按照单个 SuperPOD 方式部署，即服务器：交换机：光模块的用量比例=1：1.2：11.4。

假设 4：服务器价格参考英伟达价格，为 20 万美元；交换机结合 Mellanox 售价，假设单价为 2w 美金，光模块根据交换机速率，现在主流为 200G，假设售价为 250 美金。

结论：服务器、交换机、光模块的市场弹性分别为 15%、5%、3%。

表2: 一次性投入角度硬件需求角度测算

	全球服务器 (AI 服务器)	交换机	光模块
原有需求			
国内出货量 (万台)	1354 (AI 服务器 40)	1200	
全球出货量 (万台)	251 (AI 服务器 50)	53	2000
国内市场规模 (亿美元)	992 (AI 服务器 156)	376	
全球市场规模 (亿美元)	391 (AI 服务器 13)	170	85
单个 GPT4.0 模型 (一次性投入角度)			
单个需求量 (万台)	0.38	0.45	4.28
单价 (万美元)	20	2	0.025
增量市场 (假设 20 家同样规模投入)			
需求量 (万台)	7.5	9	85.5
投入规模 (亿美元)	150	18	2.14
弹性比例			
需求量增量 (全球)	1%	0.8%	4%
规模增量 (全球)	15%	5%	3%

资料来源:Lightcounting, IDC, 国信证券经济研究所整理

注: 原有需求按照 2021 年需求简单测算; 2019 年国内 AI 服务器出货量为 7.9w 台, 假设每年增速 30%。

注: 光模块需求出货为 100G 以上速率光模块出货总和大致估计

◆ 弹性测算二: 下游需求亿级别大规模响应 (远期应用规模起量角度)

角度 2: 基于下游应用呈现规模角度, 即按照单 GPT4.0 模型对于服务器需求用量测算。

假设 1: 单个应用的需求角度开看, 服务器潜在用量为 1.5 万台。

假设 2: 全球假设国内和海外有潜在 20 家公司可能形成同样类型规模应用。

假设 3: 网络结构比例按照单个 SuperPOD 方式部署, 即服务器: 交换机: 光模块的用量比例=1: 1.2: 11.4。

假设 4: 服务器价格参考英伟达价格, 为 20 万美元; 交换机结合 Mellanox 售价, 假设为 2.5-3w 美金, 光模块根据交换机速率, 现在主流为 200G, 假设售价为 250 美金。

结论: 服务器、交换机、光模块的市场弹性分别为 65%、19%、10%。

表3: 下游应用大规模使用角度硬件需求角度测算

	全球服务器 (AI 服务器)	交换机	光模块
原有需求			
国内出货量 (万台)	1354 (AI 服务器 40)	1200	
全球出货量 (万台)	251 (AI 服务器 50)	53	2000
国内市场规模 (亿美元)	992 (AI 服务器 156)	376	
全球市场规模 (亿美元)	391 (AI 服务器 13)	170	85
单个 GPT4.0 模型 (规模应用角度)			
单个需求量 (万台)	1.5	1.8	17.1
单价 (万美元)	20	2	0.025
增量市场 (假设 20 家同样规模投入)			
需求量 (万台)	30	36	342
投入规模 (亿美元)	600	72	8.55
弹性比例			
需求量增量 (全球)	2%	3.0%	17%
规模增量 (全球)	65%	19%	10%

资料来源:Lightcounting, IDC, 国信证券经济研究所整理

注: 原有需求按照 2021 年需求简单测算; 2019 年国内 AI 服务器出货量为 7.9w 台, 假设每年增速 30%。

注: 光模块需求出货为 100G 以上速率光模块出货总和大致估计

◆ 投资建议：

以 Chatgpt 为代表的 AI 应用正在得到快速发展，并进行着快速迭代，进而有望赋能多行业、多样化新型应用落地，提升行业效率。软端的快速发展离不开硬件基础设施的保障，因此 AI 数据中心内部的各类的硬件基础设施环节有望充分受益 AI 行业进步带来的需求提升：建议重点关注国内 ICT 领先企业【紫光股份】、【浪潮信息】、【锐捷网络】、【中兴通讯】、【菲菱科思】；光模块及光器件环节【中际旭创】、【天孚通信】、【新易盛】；IDC 温控企业【英维克】、【申菱环境】；IDC 电源端企业：【科华数据】、【科士达】；第三方 IDC【奥飞数据】；建议关注：【同飞股份】、【佳力图】。

◆ 风险提示：

Chatgpt 为代表的的应用落地不及预期，参数假设存在偏差，实际行业需求和投入力度不及预期。

相关研究报告：

《通信行业周报 2023 年第 10 期-运营商加强数字化投入，AI 迎来“iPhone 时刻”》——2023-03-27
《国信通信 - 云基建专题（二）-AI 加速发展下，云基建产业新趋势》——2023-03-23
《通信行业周报 2023 年第 9 期-“东数西算”加速建设，关注 AI 底层硬件需求》——2023-03-19
《国信通信-卫星专题二-卫星应用潜力充足，运营环节大有可为》——2023-03-13
《通信行业周报 2023 年第 8 期-关注数字经济主线，5G 卫星互联网标准构建启动》——2023-03-12

免责声明

分析师声明

作者保证报告所采用的数据均来自合规渠道；分析逻辑基于作者的职业理解，通过合理判断并得出结论，力求独立、客观、公正，结论不受任何第三方的授意或影响；作者在过去、现在或未来未就其研究报告所提供的具体建议或所表述的意见直接或间接收取任何报酬，特此声明。

国信证券投资评级

类别	级别	说明
股票 投资评级	买入	股价表现优于市场指数 20%以上
	增持	股价表现优于市场指数 10%-20%之间
	中性	股价表现介于市场指数 $\pm 10\%$ 之间
	卖出	股价表现弱于市场指数 10%以上
行业 投资评级	超配	行业指数表现优于市场指数 10%以上
	中性	行业指数表现介于市场指数 $\pm 10\%$ 之间
	低配	行业指数表现弱于市场指数 10%以上

重要声明

本报告由国信证券股份有限公司（已具备中国证监会许可的证券投资咨询业务资格）制作；报告版权归国信证券股份有限公司（以下简称“我公司”）所有。 ，本公司不会因接收人收到本报告而视其为客户。未经书面许可，任何机构和个人不得以任何形式使用、复制或传播。任何有关本报告的摘要或节选都不代表本报告正式完整的观点，一切须以我公司向客户发布的本报告完整版本为准。

本报告基于已公开的资料或信息撰写，但我公司不保证该资料及信息的完整性、准确性。本报告所载的信息、资料、建议及推测仅反映我公司于本报告公开发布当日的判断，在不同时期，我公司可能撰写并发布与本报告所载资料、建议及推测不一致的报告。我公司不保证本报告所含信息及资料处于最新状态；我公司可能随时补充、更新和修订有关信息及资料，投资者应当自行关注相关更新和修订内容。我公司或关联机构可能会持有本报告中所提到的公司所发行的证券并进行交易，还可能为这些公司提供或争取提供投资银行、财务顾问或金融产品等相关服务。本公司的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中意见或建议不一致的投资决策。

本报告仅供参考之用，不构成出售或购买证券或其他投资标的的要约或邀请。在任何情况下，本报告中的信息和意见均不构成对任何个人的投资建议。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。投资者应结合自己的投资目标和财务状况自行判断是否采用本报告所载内容和信息并自行承担风险，我公司及雇员对投资者使用本报告及其内容而造成的一切后果不承担任何法律责任。

证券投资咨询业务的说明

本公司具备中国证监会核准的证券投资咨询业务资格。证券投资咨询，是指从事证券投资咨询业务的机构及其投资咨询人员以下列形式为证券投资人或者客户提供证券投资分析、预测或者建议等直接或者间接有偿咨询服务的活动：接受投资人或者客户委托，提供证券投资咨询服务；举办有关证券投资咨询的讲座、报告会、分析会等；在报刊上发表证券投资咨询的文章、评论、报告，以及通过电台、电视台等公众传播媒体提供证券投资咨询服务；通过电话、传真、电脑网络等电信设备系统，提供证券投资咨询服务；中国证监会认定的其他形式。

发布证券研究报告是证券投资咨询业务的一种基本形式，指证券公司、证券投资咨询机构对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向客户发布的行为。

国信证券经济研究所

深圳

深圳市福田区福华一路 125 号国信金融大厦 36 层

邮编：518046 总机：0755-82130833

上海

上海浦东民生路 1199 弄证大五道口广场 1 号楼 12 层

邮编：200135

北京

北京西城区金融大街兴盛街 6 号国信证券 9 层

邮编：100032