

英伟达 H20 恢复销售事件点评

Q2 国产云维持高增速, H20 出货预计带动国内云厂 Capex 上升

◆ 行业研究 · 行业快评

◆ 互联网 · 互联网 II

◆ 投资评级: 优于大市(维持)

证券分析师: 张伦可 0755-81982651 zhanglunke@guosen.com.cn
证券分析师: 刘子譚 liuzitan@guosen.com.cn

执证编码: S0980521120004
执证编码: S0980525060001

事项:

2023 年底, 美国拜登政府升级对华半导体出口管制后, 英伟达为维持中国市场, 推出面向中国市场的“阉割版” H20 芯片, 其性能较 H100 降低 28%, 但符合美国设定的出口合规门槛。

2025 年 4 月, 美国政府通知英伟达, 要求自即日起对华出口 H20 芯片需申请许可证, 后美国政府宣布 H20 实施无限期出口禁令。4 月 16 日英伟达正式公告禁令, 计提约 45 亿美元库存减值。

2025 年 7 月 14 日, 英伟达官网发布英伟达将恢复向中国销售 H20 芯片, 并宣布推出全新完全合规的中国特供版 GPU。英伟达正在重新提交销售 H20GPU 的申请, 美国政府已向英伟达保证将发放许可证, 公司希望尽快开始交付。

国信互联网观点: 1) AI 芯片算力层: 英伟达 H20 恢复销售将带动库存出清与国内资本开支增长, 后续特供 GPU 顺利则预计资本开支持续积极。英伟达 H20 作为技术妥协与商业平衡的产物依托 CUDA 生态壁垒仍是国内 AI 开发者的首要选择。根据英伟达一季报披露库存减记和采购义务损失为 45 亿美元, 对应 H20 库存收入口径预计达百亿美元量级, 恢复 H20 采购后预计各家云厂将抢购囤货、大幅带动资本开支增长。后续中国特供版 GPU 预计基于 BLACKWELL 架构, 如进展顺利国内资本开支将持续积极。2) 国内云厂商: 二季度云维持高增速, 随着客户方案落地下半年预计加速增长。H20 恢复供应将明显带动国内云厂资本开支上升, 利好云厂下游需求增长。二季度国内云厂商 AI 客户数小幅下降、但是场景渗透率持续提升, 预计国内云厂二季度维持高增速。跑通业务场景后, 中大客户调用量提升与私有化解决方案落地预计将带动云业务下半年加速增长。二季度国产卡买租占比提升, 但受限于良品率和交付能力仍无法满足算力增长需求, 各家云厂纷纷消耗库存卡。资本开支端国内云厂投资节奏显著受到 H20 断供的影响。二季度国内云厂商阶段性放缓整体 AI 服务器采购, 预计国内云厂资本开支二季度环比保持平稳。随着 H20 恢复供应以及英伟达 B 系列特供芯片开始供应, 预计将推动云厂需求增速出现季度性向上拐点。3) AI 模型层: 二季度模型层进一步向头部集中, 国内科技厂商加速追赶海外先进水平, 打造开源生态、推动工具建设与场景落地, 算力供应恢复将继续推动模型层迭代。二季度国内大模型发展势头强劲。阿里聚焦通用高性能开源模型, 4 月开源 Qwen3 登顶全球最强开源模型, 5 月推出众多模型与工具, 6 月发布 Qwen3-Embedding 系列模型、开源 3D 数字人项目, 打造完善 AI 生态与工具链。字节跳动发布豆包 1.5 深度思考模型、豆包视频生成模型等, 覆盖多创作场景。腾讯实现多领域模型创新与开源, 4 月上线“元宝”、开源图像生成插件, 混元 TurboS 在 Chatbot Arena 上排名攀升至全球前八, 提升应用体验。百度在基座模型与开发工具创新, 4 月发布文心大模型 4.5Turbo 和 X1Turbo, 6 月推出双数字人互动直播间、开源文心大模型 4.5 系列。

风险提示: AI 模型技术发展不及预期, AI 应用商业化进展不及预期, 大模型幻觉、伦理等安全性问题。

投资建议: H20 恢复供应解决了国内 AI GPU 短缺现状, 显著降低硬件投入与实际使用之间的时间与试错成本, 加速国内企业上云、利好云厂商下游需求增长, 推动大模型与应用落地。因此, 我们推荐海外芯片厂商龙头英伟达, 国内云厂商龙头阿里巴巴, 具备云业务与优质社交场景生态的腾讯控股以及拥有丰富端侧硬件生态的小米集团。

评论：

- ◆ AI 芯片算力层：英伟达 H20 恢复销售将带动库存出清与国内资本开支增长，后续特供 GPU 顺利则预计资本开支持续积极

英伟达 H20 作为技术妥协与商业平衡的产物依托 CUDA 生态壁垒仍是国内 AI 开发者的首要选择。H20 作为英伟达为中国市场定制的合规产品，其 FP16 算力仅为 H100 的 15%-30%，且禁用 NVLink 多卡互联功能。这种“性能阉割”虽削弱了其在训练场景的竞争力，却通过 96GB HBM3 显存和 4TB/s 带宽保留了推理优势。英伟达通过这种技术折中，在满足美国出口管制的同时，尽可能维持在中国 AI 推理市场的份额。

尽管 H20 性能受限，但其 CUDA 生态兼容性仍构成强大壁垒。中国主要云厂商和 AI 开发者仍然依赖英伟达软件工具链，这种路径依赖使得 H20 在模型训练、推理等场景仍具不可替代性。

根据英伟达一季报披露库存减记和采购义务损失为 45 亿美元，对应 H20 库存收入口径预计达百亿美元量级，恢复 H20 采购后预计各家云厂将抢购囤货、大幅带动资本开支增长。H20 芯片一季度因库存减记和采购义务产生 45 亿美元费用。虽然一季度 H20 产品在限制前销售 46 亿美元，但另有 25 亿美元订单因无法发货而损失（新加坡占第一季收入的近 20%）。预计二季度 H20 相关收入损失约 80 亿美元。

后续中国特供版 GPU 预计基于 BLACKWELL 架构，如进展顺利则国内四季度资本开支将持续积极。2025 年 7 月，英伟达曾推出基于 Blackwell 架构的全新特供卡 B30，根据 Wccfttech 报道性能为 H20 的 75%（约 222 TFLOPS FP8 算力），定价 6500-8000 美元（较 H20 低 40%），能效比提升 30%。但 B30 采用 GDDR7 显存，其显存带宽仅为 H20 HBM3 的 42.5%，在处理高精度计算时效率低下。此次提到的中国特供版 GPU 预计将改善 B30 部分“缺陷”、性能相比 H20 有所提升。

图1：H20 产品架构与性能对比

	HGX H20	L20 PCIe	L2 PCIe
GPU Architecture	NVIDIA Hopper	NVIDIA Ada Lovelace	NVIDIA Ada Lovelace
GPU Memory	96 GB HBM3	48 GB GDDR6 w/ ECC	24 GB GDDR6 w/ ECC
GPU Memory Bandwidth	4.0 TB/s	864 GB/s	300 GB/s
INT8 FP8 Tensor Core*	296 296 TFLOPS	239 239 TFLOPS	193 193 TFLOPS
BF16 FP16 Tensor Core*	148 148 TFLOPS	119.5 119.5 TFLOPS	96.5 96.5 TFLOPS
TF32 Tensor Core*	74 TFLOPS	59.8 TFLOPS	48.3 TFLOPS
FP32	44 TFLOPS	59.8 TFLOPS	24.1 TFLOPS
FP64	1 TFLOPS	N/A	N/A
RT Core	N/A	Yes	Yes
MIG	Up to 7 MIG	N/A	N/A
L2 Cache	60 MB	96 MB	36 MB
Media Engine	7 NVDEC 7 NVJPEG	3 NVENC (+AV1) 3 NVDEC 4 NVJPEG	2 NVENC (+AV1) 4 NVDEC 4 NVJPEG
Power	400 W	275W	TBD
Form Factor	8-way HGX	2-slot FHFL	1-slot LP
Interconnect	Pce Gen5 x16: 128 GB/s NVLink: 900GB/s	PCIe Gen4 x16: 64 GB/s	PCIe Gen4 x16: 64 GB/s
Availability	PS: Nov 2023 MP: Dec 2023	PS: Nov 2023 MP: Dec 2023	PS: Dec 2023 MP: Jan 2024

资料来源：英伟达官网、国信证券经济研究所整理

◆ 国内云厂商：二季度云维持高增速，随着客户方案落地下半年预计加速增长。H20 恢复供应将明显带动国内云厂资本开支上升，利好云厂下游需求增长

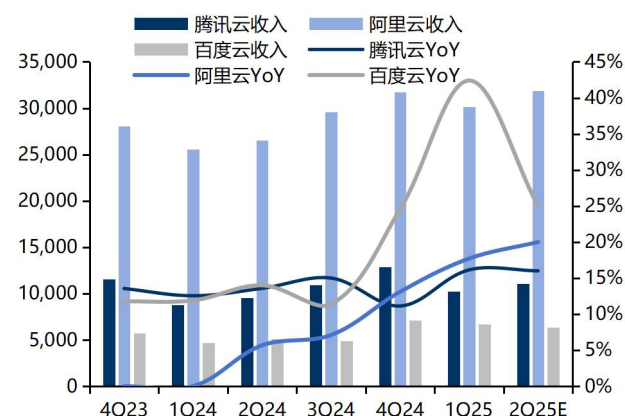
二季度伴随着 Deepseek 热度下滑、尝鲜用户减少，国内云厂商 AI 客户数小幅下降、但是场景渗透率持续提升。跑通业务场景后，中大客户调用量提升与私有化解决方案落地预计将带动云业务下半年加速增长。其中预计阿里云加速至同比+20%YoY（环比+2pcts）、腾讯云维持 16%YoY（同比+2pcts）、百度云预计环比增速至+25%YoY（同比+12pcts、环比-17pcts）。

二季度国产卡买租占比提升，但受限于良品率和交付能力仍无法满足算力增长需求，各家云厂纷纷消耗库存卡。国内卡虽然选择空间大，但生产的良品率和交付能力较弱，仍无法代替 NV 卡的使用。因此租赁放缓、非标训练卡持续采购，前期储备卡上线缓解部分压力。长此以往，大厂 AI 训练成本提升、新客户接入周期延长，这会对云厂商的 AI 云业务拓展产生阻碍，进而影响云业绩的增长速度。

资本开支端国内云厂投资节奏显著受到 H20 断供的影响。25 年云厂商预计将持续加大对 AI 的投入，但二季度由于 H20 断供各厂商面临无卡下单、无卡到货的情况，导致国内云厂商阶段性放缓整体 AI 服务器采购，预计国内云厂资本开支二季度环比保持平稳。预计阿里与腾讯 25Q2 Capex 预计与一季度整体持平。百度整体 Capex 稳健，预计同比持平。

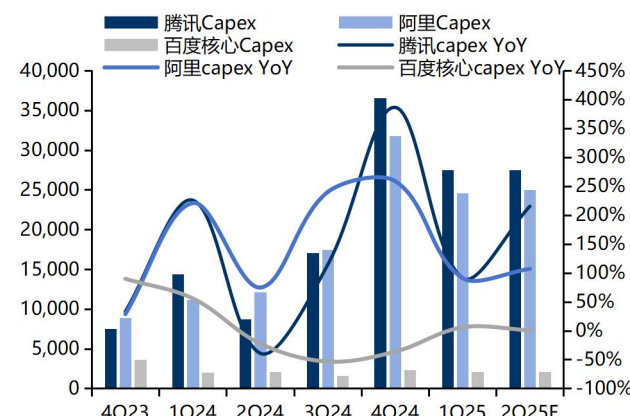
随着 H20 恢复供应以及英伟达 B 系列特供芯片开始供应，预计推动云厂需求增速出现季度性向上拐点。初步预计三季度各家将恢复原有 H20 采购计划、抢购囤货。同时根据 B 系列特供芯片性能与定价计划四季度采购方案的预算。

图2：国内云厂商云收入与同比变化预测（百万元/%）



资料来源：各公司财报、国信证券经济研究所整理

图3：国内云厂商资本开支与同比变化预测（百万元/%）



资料来源：各公司财报、国信证券经济研究所整理

◆ AI 模型层：二季度模型层进一步向头部集中，国内科技厂商加速追赶海外先进水平，打造开源生态、推动工具建设与场景落地

阿里：聚焦通用高性能开源模型，打造完善 AI 生态与工具链

4 月，阿里开源新一代通义千问模型 Qwen3，性能全面超越 R1、OpenAI-o1 等全球顶尖模型，阿里云 AI 社区魔搭推出全新 MCP 广场。5 月推出一系列模型和工具，如夸克的“深度搜索”产品、通义实验室的开源大模型搜索引擎 ZeroSearch、QwenChat 的网页开发功能等。6 月发布 Qwen3-Embedding 系列模型，开

源 3D 数字人项目 MNN TaoAvatar。17 日，发布 Qwen3 升级版，适配苹果 MLX 架构。6 月底推出多模态模型 QwenVLo。

字节跳动：模型升级与功能拓展，覆盖多创作场景

4 月 17 日，火山引擎发布豆包 1.5 深度思考模型，同日开源多模态智能体 UI-TARS-1.5。18 日，AI 开发平台扣子空间启动内测。5 月 Seed 团队发布代码模型 Seed-Coder，在 HuggingFace 上推出 Dream0 框架，火山引擎推出豆包·语音播客模型，还开源了多模态基础模型 BAGEL 等。6 月发布 SeedEdit3.0 图像编辑模型。11 日，升级豆包 App “一句话 P 图” 功能，发布豆包·视频生成模型 Seedance1.0Pro，火山引擎发布 PromptPilot 平台及豆包大模型 1.6。17 日，豆包上线 “AI 播客” 功能。

腾讯：多领域模型创新与开源，提升应用体验

4 月微信上线 AI 助手 “元宝”，集成混元与 DeepSeek 双引擎，支持解析多种文件并提供智能摘要等功能。18 日，腾讯混元开源图像生成插件 InstantCharacter。5 月腾讯混元大模型技术能力持续提升，混元 TurboS 在 ChatbotArena 上排名攀升至全球前八，国内仅次于 DeepSeek，同时推出了混元 T1-Vision、混元 Voice 等模型，混元图像 2.0 实现毫秒级实时生图，混元 3D 在可控性与超高清生成能力上实现代际飞跃。6 月 27 日，开源基于 MoE 架构的混元-A13B 大语言模型。

百度：基座模型与开发工具创新，推动技术落地

4 月 25 日，在 Create2025 百度 AI 开发者大会上发布全新一代文心大模型 4.5Turbo 和 X1Turbo，性能提升的同时价格大幅下降。6 月 17 日，推出基于文心大模型 4.5Turbo 的双数字人互动直播间。6 月底，开源文心大模型 4.5 系列，涵盖多款模型并提供开发套件。

DeepSeek：持续迭代深度推理模型

5 月 28 日，在 HuggingFace 平台发布 DeepSeekR1-0528，显著提升思维深度与推理能力，整体表现已接近 o3 与 Gemini-2.5-Pro 等国际顶尖模型。

◆ 投资建议：

建议关注 AI 领域端侧产品和应用进展，关注 AI 大模型进展。

◆ 风险提示：

AI 模型技术发展不及预期，AI 应用商业化进展不及预期，大模型幻觉、伦理等安全性问题。

相关研究报告：

《从 Circle 上市，探索稳定币与交易所的合作模式》——2025-06-24

《人工智能周报(25 年第 24 周)-OpenAI 上线 o3-pro 模型, 字节跳动发布豆包大模型 1.6》——2025-06-19

《在线音乐行业深度解析：复杂投资环境下的优质长期现金流资产》——2025-06-18

《国内互联网云厂商业绩与大模型进展跟踪-预计 Deepseek 带动云收入加速、利润率提升，大模型追赶海外先进水平》——2025-05-14

《美股科技互联网 25Q1 财报总结：营收利润超指引强劲，指引资本开支持续提升》——2025-05-13

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

投资评级标准	类别	级别	说明
报告中投资建议所涉及的评级（如有）分为股票评级和行业评级（另有说明的除外）。评级标准为报告发布日后 6 到 12 个月内的相对市场表现，也即报告发布日后的 6 到 12 个月内公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅作为基准。A 股市场以沪深 300 指数（000300.SH）作为基准；新三板市场以三板成指（899001.CSI）为基准；香港市场以恒生指数（HSI.HI）作为基准；美国市场以标普 500 指数（SPX.GI）或纳斯达克指数（IXIC.GI）为基准。	股票 投资评级	优于大市	股价表现优于市场代表性指数 10%以上
		中性	股价表现介于市场代表性指数±10%之间
		弱于大市	股价表现弱于市场代表性指数 10%以上
		无评级	股价与市场代表性指数相比无明确观点
	行业 投资评级	优于大市	行业指数表现优于市场代表性指数 10%以上
		中性	行业指数表现介于市场代表性指数±10%之间
		弱于大市	行业指数表现弱于市场代表性指数 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。本报告仅供我公司客户使用，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