仅供内部参考，请勿外传

从大模型演进到 Token 工厂，AI 产业链共筑基础设施升级

ODX 2026 首届算博会跟踪报告

招商电子团队近日参加了在北京举办的 ODX 2026 大会，AI 上下游产业链各公司专家做了非常详细的主题演讲，大会对当前 AI 产业发展现状深刻剖析，各论坛板块探讨火热，我们整理了本次大会部分演讲的核心观点，供投资者参考。

- **Token 经济与大模型：终端应用与模型范式是硬件迭代的第一性驱动力，硬件升级要落到有效 Token 产出。**生成式 AI 走向 Agentic AI，单 Agent 走向多 Agent 协同等，这些变化依次重写了算力类型、访存模式、通信拓扑、存储层级、CPU/GPU 配比和功耗曲线。GPU 仍是算力核心，但新增价值量越来越多地流向 CPU 调度、HBM/DRAM/SSD 多级内存、Scale Up 网络、KV Cache 服务、800V 供电、液冷以及整机柜系统工程。行业评价标准更关注 Token/s、Token/W、单用户 Token/s、TTFT、资源利用率及单 Token 成本等。
- **超节点与散热：液冷与 800V、高速互连和机柜结构协同设计，整机价值要以 Token/s 等衡量。**MoE、PD 分离、长上下文和训推一体要求更大的 Scale Up 域，同时单芯片功耗持续提升，使传统 8 卡服务器、风冷机柜及按部件独立设计的模式同时失效。超节点因此演进为以机柜为最小交付单元，将 CPU/GPU/XPU、内存和存储池、交换芯片、Cable Tray/正交背板、供电母线、液冷管路及管理单元统一设计。关于超节点架构，行业正朝着由模块化、三总线盲插和全液冷提升单柜密度及可维护性；更长期则与 800V HVDC、NPO/CPO、全光互连及算电协同共同重构机柜几何形态。整机价值最终要以有效 GPU 卡时、Token/s、Token/W 及可用率衡量，而非单纯比较理论峰值算力。**1）超聚变：**从 2017 年三总线液冷整机柜出发，已完成核心部件自研和百万级 XPU 规模商用，并把下一代目标推向 672 GPU、1MW 供电/制冷和全光互连；**2）百度：**一条线做计算节点模块化，另一条线推进天池 1.0→3.0 整柜演进，并已积累万卡级国产集群交付与运维数据；**3）新华三：**以 UXM 标准化解决异构 XPU 整机碎片化，同时探索正压与负压两条液冷路线。UXM 把类 PCIe、BFX 和 OAM 三类模组纳入共同体系，统一 54V 供电、机械、液冷和高速接口，并允许不同 Scale Up 协议、正交或分布式整机架构并行。面对单芯片 2kW 并向 5kW 演进、交换芯片功耗超过 3.5kW，公司同步推进 800V Power Shelf、VPD 和液冷方案。
- **网络与互连：互连复杂度持续提升行业光铜并进，业界对 NPO 重视度有所提高。**当前互连行业呈现多轨并行、双轨并进的发展方向，Scale up、Scale out 与 Scale across 同步扩容，以实现更高算力密度，互连的方式也从传统的 GPU-GPU 通信扩展为 GPU-CPU-内存-SSD 的全系统互连。**1）互连方案上，**依旧遵循“能用铜就用铜，必须用光再用光”的原则，产业趋势呈现出光铜并进的格局。Scale up 作为智算互联中带宽压力最大的互连架构，目前含有铜、光两种技术路径，但随着互连规模的扩大，单机柜面临物理承载极限，因此需要跨机柜互连，进一步催生光互连需求。铜互连方面，随着传输速率的持续提升，为降低损耗同时延长铜缆生命周期，从传统的 Cable Cartridge 铜线缆方案进一步延伸出 NPC 以及最终的 CPC 方案；光互连方面，我们观察到业界普遍对 NPO 的重视程度提高，尽管其并非最终技术演进路线，但却是在兼顾技术落地进度、性能、功耗以及可维护性下的阶段性最优解。**2）具体应**

推荐（维持）

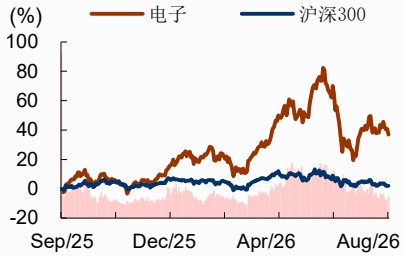
TMT 及中小盘/电子

行业规模

		占比%
股票家数（只）	533	10.3
总市值（十亿元）	21828.2	19.4
流通市值（十亿元）	15573.8	15.8

行业指数

%	1m	6m	12m
绝对表现	12.1	14.7	37.6
相对表现	13.3	15.9	33.4



资料来源：公司数据、招商证券

相关报告

1、《国产超节点跟踪报告—WAIC2026 超节点加速推出，关注国产算力未来放量节奏》2026-07-24

2、《英伟达 COMPUTEX 2026 跟踪报告—Rubin 全面量产，Vera 专为 Agent 设计性能超 x86》2026-06-02

3、《MemoryS 2026 闪存大会跟踪报告：行业缺货或将延续至 27 年，关注未来存储技术创新重构》2026-03-29

鄢凡 S1090511060002
yanfan@cmschina.com.cn
程鑫 S1090523070013
chengxin2@cmschina.com.cn
谌薇 S1090524070008
shenwei3@cmschina.com.cn
涂锟山 S1090525040004
tukunshan@cmschina.com.cn

仅供内部参考，请勿外传

用上，**立讯技术**遵循光铜并进的发展思路，在 224G 速率下具备完整的互联解决方案，覆盖从 GPU 到交换芯片、从板内到板间、从柜内到柜间的完整端到端链路，包括高速连接器、线缆组件、NPO/CPO 连光板组件，同时配套散热、电源方案等。同时立讯已在开展 448G 速率完整方案预研，CPC 方案实物样品已经产出，完全满足 PAM6 要求，PAM4 还会迭代新版本，预计今年内可以推出；**腾讯与阿里**在光互连领域联合发起 UPO 标准制定，产业链上下游大量合作伙伴共同参与，共三种形态适配不同应用场景，与主芯片集成可采用多种集成方案，UPO 标准采用 LGA Socket 连接方案，实测指标可满足 224G 应用需求，也可平滑演进至 448G。当前 UPO 标准制定速度加快，预计年底推出 Type1 XD Socket 产品，明年启动 448G UPO 标准制定。

- **存储：KV Cache 带来存储容量与传输两大难题，超节点形态或从连接算力走向连接存力。**KV Cache 逐步成为推理场景新瓶颈，随着上下文长度、并发数量提升，KV Cache 占用容量会快速扩张，给推理系统带来压力。针对 KV Cache，除了解决存储容量，还需要解决传输速度、数据复用两大问题，当缓存利用率超过 95%时，系统会从计算密集型转变为 I/O 密集型，KV Cache 传输瓶颈会凸显。因此，如何合理有效的管理 KV Cache 对于推理工作负载至关重要，内存层级架构的精细分工与紧密互连重要性凸显。当前内存层级主要包含 L1-GPU HBM→L2-CPU DRAM→L-本地 SSD→L3.5-ICMS 池→L4 云存储，针对 L1 与 L2 之间的传输 GAP，**1）腾讯：**提出 L1.5 解决方案，在现有超节点架构中接入独立内存池，直接在 Scale up Switch 接入专用 DDR 内存池，可为单颗 GPU 提供约 500GB 访问容量以及 400-800GB/s 访问带宽，可显著提高 batch size，带动 TPS 提升，大幅提高模型性能。腾讯推出的 Memory Switch Fabric 项目顺利开展，支持 DRAM、NAND、HBF，9 月启动软件层开发，明年会有 Memory Connect 芯片，将原 Scale up 网络从算力互连演进为内存互连。**2）三星：**针对 L1.5 层级，三星推出 zNAND-O，采用 SLC 架构，与 NPU 合封，推出适配端侧与服务器端架构，极大程度提高整体运行性能，预计 2028 年起深度赋能 AI 创新。
- **电源：800VDC 为兆瓦级机柜升级的物理必然，算电协同与源网荷储为行业持续演进趋势。**GPU 单芯片功耗由数百瓦走向 3.6-5kW、单柜功率由数十 kW 走向 300-600kW 乃至 1MW 后，供电系统需要同时完成四项重构：一是以 800V HVDC 降低长距离输电电流和铜损；二是把 DC/DC、VRM 持续推近负载，解决芯片“最后一英寸”的数千安培供电；三是通过超级电容、BBU/储能、分层控制消化 GPU 毫秒级动态功率；四是把供电与算力调度、液冷、电网和微电网统一设计。当前主线仍是先以 800V→50/54V 兼容既有服务器生态，再逐步演进至 12V/6V 中间母线、800V 进板、VPD/IVR；更长期则由 SST 推动中压直连和原生直流。电源由服务器配套部件升级为贯穿“源-网-荷-储-芯”的机架级能源平台。**1）麦格米特：**把 Sidecar 定义为可扩展、可管理的机架级能源平台。公司将数据中心电源分为 Shelf 内分布式、机架内集中式和机架级能源平台三类；**2）晶丰明源：**聚焦“最后一英寸”，以 6V 母线、TLVR、Co-Pack LDMOS 和 VPD 解决 4000A 级 Vcore 供电。
- **投资建议：**本次大会中行业内知名公司对自身所处领域以及未来技术发展做了深度解析与前瞻判断，产业链上下游公司共同推动技术升级，建议关注 ODCC AI 生态：服务器整机、高速互连、电源、机柜电源线、液冷、算力芯片和交换芯片、存储、配套芯片、滑轨、机箱等细分领域公司。
- **风险提示：宏观经济及政策风险；存储价格下滑风险；AI 资本开支不及预期的风险；行业竞争加剧的风险。**

仅供内部参考，请勿外传

表 1：ODCC 联盟相关标的

细分领域	标的
服务器整机	华为（未上市）、超聚变（拟上市）、紫光股份等
高速互连	立讯精密、汇聚科技、华丰科技、胜蓝股份、庆虹（未上市）等
电源	立讯精密、中宝科技、欧陆通、LEAD WEALTH（比亚迪电子）、麦格米特、中达电通、通达创智、纵贯线、瑞可达、肯铭达、东创精艺等
机柜电源线	立讯精密、瑞可达、肯铭达、纵贯线等
液冷	立讯精密、英维克、立敏达（领益智造）、LEAD WEALTH（比亚迪电子）、诺通（未上市）、泰诚精密（未上市）等
算力芯片和交换机芯片	华为、海光信息、壁仞科技、博通、中兴通讯等
存储	长鑫科技、长存控股（IPO 中）、海力士、美光、三星、闪迪、铠侠、西部数据、希捷、江波龙、大普微、佰维存储、澜起科技、忆恒创源等
配套芯片	杰华特、英诺赛科、圣邦股份、澜起科技、晶丰明源、芯朋微、紫光同创、新港海岸（未上市）、曦智科技等
滑轨	元拾、台一盈拓（未上市）等
机箱	时代博川（未上市）、CABINET（未上市）等

资料来源：招商证券整理（注：紫光股份、华丰科技、瑞可达、中兴通讯为通信组覆盖，同有科技为计算机组覆盖，麦格米特为电新组覆盖）

仅供内部参考，请勿外传

正文目录

一、ODX 大会部分现场展览图 6
二、主论坛 9
1、华为：星河 AI 数据中心网络，算力满载，业务永续 9
2、三星：Beyond the memory wall, Rearchitecting the data path for agentic AI 10
3、博通：AI at Scale：以太网作为通用互联底座 12
4、超聚变：架构跃迁，超节点创新之路 13
5、阳光电源：从源到芯，算电协同：基于 SST 的 AIDC 供电解决方案新变革 14
6、中兴通讯：“Token 经济”时代，全栈协同引领 TCO 最优解 15
7、新华三：以标准化共建开放超节点技术生态 17
8、晶丰明源：面向 AI 高密度算力集群，“最后一英寸”电源解决方案的演进 18
9、庆虹电子：AI 超节点高速互连方案的实践与思考 19
三、服务器分论坛 21
1、安费诺：AI 算力底座的互连与供电：超节点、NPO 与 HVDC 21
2、天翼云：面向 Agentic AI 应用的超节点硬件架构创新 22
3、锐捷网络：800V 高压直流机柜供电演进 24
4、百度：Agent 时代的计算节点模块化设计 25
四、超大规模集群分论坛 28
1、百度云：百度智能云 AI 全栈体系赋能超大规模集群建设 28
五、大模型算力分论坛 30
1、海光信息：大模型推理时代，国产算力与开放生态的双轮答卷 30
2、壁仞科技：异构并行分发架构下大模 Token 负载协同优化 31
六、百度智能云基础设施分论坛 33
1、百度云：Token 工业化时代：超节点架构重构与智算底座演进 33
七、算电协同分论坛 35
1、麦格米特：面向兆瓦级 AI 机柜的 800 VDC Sidecar 供电架构 35
八、AI 工厂分论坛 36
1、小红书：万卡 AI 集群稳定与智能运维实践 36
九、网络分论坛 38
1、立讯技术：AI Rack 高速互连方案介绍及 448G 高速铜互连可行性探讨 38
2、腾讯/阿里：UPO（Ultra Performance Optics）定义智算光互连新范式 40
3、腾讯：IO-NET：Scale Up 内存池构建内存互联超节点 42

图表目录

图 1：ODCC AI 工厂合作生态图 6

仅供内部参考，请勿外传

图 2：机柜示意图 6
图 3：液冷部分相关供应商展示 7
图 4：UQD/MQD 部分供应商展示 7
图 5：电源 Bus Bar 部分供应商展示 7
图 6：机柜电源线部分供应商展示 8
图 7：电源相关供应商展示 8

仅供内部参考，请勿外传

一、ODX 大会部分现场展览图

ODCC AI 工厂合作生态：2026 开放数据中心大会暨首届算博会（2026 ODX）重点聚焦计算、网络、存储、设施、液冷、算电、边缘等热点算力领域，基于大模型训练和推理等应用场景进行研讨。其中 ODCC 特设的 AI 工厂生态系统围绕机架与托盘标准化设计、整机柜集成方案、液冷散热技术、高速互联网络架构等核心模块进行展示，以开源设计为起点，致力于加速中国 AI 基础设施从规划走向规模部署。AI 工厂联合 NVIDIA、美团、小红书、新华三、远图、立讯、澜起科技、天数智芯等成员，覆盖互联网、芯片、服务器、网络设备、电源供应、液冷方案等全产业链关键环节，推动构建开放、协同、互信的产业生态。

图 1：ODCC AI 工厂合作生态图



资料来源：ODCC、招商证券

立讯技术覆盖高速互连、热管理、电源等领域完整产品布局：立讯技术作为全球领先的 ICT 核心零组件解决方案商，集中展示覆盖高速互连、热管理、电源管理等关键环节的全栈产品能力，为下一代 AI 基础设施提供高效、可靠、可扩展的底层支撑。高速互连方面，立讯技术已形成从 GPU 节点、交换机节点到机柜内及跨机柜互连的完整产品布局，并围绕 Chip-to-I/O、Chip-to-BP 等关键链路构建高速铜互连能力，同时，DAC、ACC、AEC 及 AOC 等高速互连产品持续演进，为 AI 集群提供端到端的高速连接支撑。

图 2：机柜示意图



资料来源：立讯技术、招商证券



热管理方面，立讯技术构建覆盖冷板、CDU、Manifold 以及快速流体连接器等产品的液冷技术组合，并进一步延伸至液冷双卡、四卡模组等核心计算单元。“芯片侧”到“机柜侧”的协同设计，提升了高热密度场景下的散热效率、温控稳定性和

仅供内部参考，请勿外传

系统可维护性。

图 3：液冷部分相关供应商展示



资料来源：立讯技术、招商证券

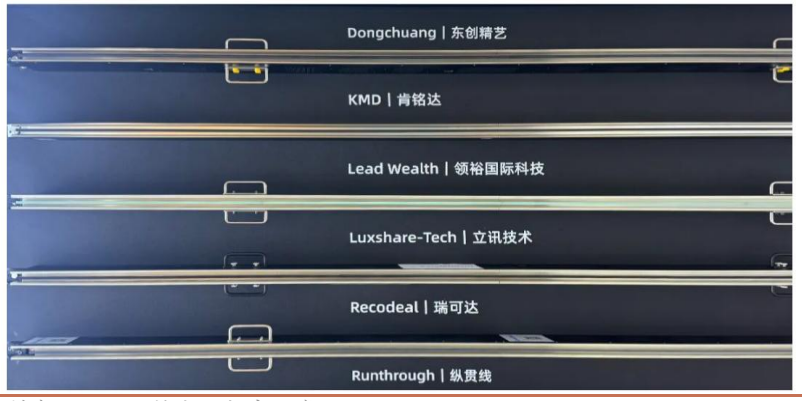
图 4：UQD/MQD 部分供应商展示



资料来源：立讯技术、招商证券

电源管理方面，立讯技术 Power Shelf 产品已覆盖 33kW、90kW、110kW 等不同功率等级，并延伸至 Rack Power Whip 等机柜级供电组件，以模块化产品组合适配不同 AI 机柜架构，在提升供电能力的同时兼顾效率、可靠性与后续扩展。

图 5：电源 Bus Bar 部分供应商展示



资料来源：立讯技术、招商证券

仅供内部参考，请勿外传

图 6：机柜电源线部分供应商展示



资料来源：立讯技术、招商证券

图 7：电源相关供应商展示



资料来源：立讯技术、招商证券

仅供内部参考，请勿外传

二、主论坛

1、华为：星河 AI 数据中心网络，算力满载，业务永续

演讲嘉宾：张白，数据通信产品线数据中心网络领域总裁

会议要点

AI 是第四次工业革命核心，大模型持续成熟，头部企业探索 AGI，多数企业落地 AI 推理业务；AI 推理成本快速下降，受杰文斯悖论影响，Token 应用扩大，算力需求持续走高。

国家层面对算网做前瞻性布局，算网定位为智能经济五大基础设施核心，需协同算、数、网、安、电要素；数据中心已经从传统应用中心、多云中心，进入算力中心时代。

算力中心网络待攻克三大问题：构建超 10 万卡算力集群、降低大规模集群算力损耗、保障集群稳定运行。华为从大规模组网、高算效互联、高可用机制三个方向解决上述难题。

1、大规模组网：

行业从万卡走向十万卡集群，传统 51.2T 交换机三层组网，存在成本高、转发跳数多、负载不均衡问题，行业转向两层扁平化架构。华为自研 100T 交换机，采用两层多平面架构，相较三层组网，交换机、光模块数量减少 70%，建网成本降低 40%。

网络带宽向 800G、1.6T 演进，传统可插拔光模块遭遇功耗墙、信号完整性墙、端口密度墙三重物理天花板。CPO 尚未大规模商用，NPO 兼顾功耗、散热、可维护性、成熟度，成为主流落地方案。华为推出自研 51.2T NPO 交换机，搭载 3.2T OE 光引擎，整机功耗降低 20%。

2、高算效互联：

1）AI 训练场景：稠密模型向 MoE 稀疏大模型演进，通信开销、通信复杂度大幅增长，传统逐流负载均衡不再适配。逐包负载均衡可与上层通信模型解耦，实现各场景网络吞吐最优。华为自研 NPLB 逐包网络负载均衡算法，可将集群训练因网络拥塞带来的性能损耗降低 2%-7%。应用逐包负载均衡后会出现包流混合调度，容易引发拥塞抖动，华为通过 ART 流混合流量动态调度技术，实现网络零抖动。

2）AI 推理场景：长序列、多轮对话带来 KV Cache 带宽需求暴涨近 8 倍，现有架构 8 张算力卡仅共享两条 200G 带宽，带宽缺口超 80%，产生计算-存储瓶颈，TPS 受限。

存算合一成为方向：存储直连参数面网络、GPU 直通，KV 带宽提升 8 倍；公司实验室测试 TPS 可提升 50%-4.8 倍，但该架构下计算、存储流量混跑，互相干扰损伤业务性能。公司对此进行优化，OpenCM 算网协同主动侧控制，P99 长尾时延降低 20%；网存联动 IO 智能调度，存储 IO 性能提升 30%，推理 TPS 再提升约 20%。

仅供内部参考，请勿外传

3、高可用机制：

从设备、链路、光模块三个维度构建全层级高可用，保障训推业务连续不中断。1）设备层面：推理业务中断阈值从训练的 28 秒收窄至 7 秒，公司发布闪启 2.0 技术，设备重启时间由 10 秒压缩至 5 秒，保障训推业务无中断。2）链路层面：原有链路故障可毫秒级切换；新增 LLR 链路级重传，针对万卡集群光模块频繁闪断问题，实现链路误码、故障微秒级重传，减少算力损失。3）光模块层面：自研星联光模块搭配短距光回损定位技术，1 分钟内排查光纤松动、脏污故障，提前消除集群调试阶段隐患。多重机制保障 AI Fabric 集群训练业务不中断。

4、行业展望：

华为联合产业输出多项标准、技术与测试规范，2025-2026 年入选 Gartner、Forrester 领导者象限。2019 年 ODCC 峰会首次提出 AI Fabric 构想，七年时间完成技术从萌芽到快速迭代。呼吁产业同仁深耕 AI Fabric，共建 AI Fabric 智能算网新底座，推动行业智能化。

2、三星：Beyond the memory wall, Rearchitecting the data path for agentic AI

演讲嘉宾：玄在雄，三星电子副总裁

会议纪要

（1）人工智能演进阶段

人工智能正在进入全新演进阶段。第一阶段为生成式人工智能，主要完成文本、图像训练；当前正在迈向代理式人工智能，系统按照规划、执行、反馈的循环完成运行。该演进正在从根本上重塑人工智能基础设施，记忆与灵感成为驱动变革的关键要素。实体 AI 工作流运行逻辑为：模型完成规划，多个智能体并行执行，评估输出结果，再次开展规划，循环往复，直至输出高质量结果。

（2）代理式人工智能的性能要求

代理式人工智能正将性能提升到一个全新的高度。在训练方面，Frontier AI 模型使用的数据量几乎每年翻三倍。计算资源方面，增长速度更是惊人——过去两年间，推理环节的增长尤为显著，其 token 使用量增长了近 100 倍。而这仅仅是由局部上下文窗口和思维链驱动的。这不仅仅是对性能提出更高要求，实际上还创造出了行业前所未有的性能规模。

（3）代理式人工智能数据处理规模

2026 年，中国每日 token 消耗量为 140 万亿。这是中国国家图书馆 AI 算力构建的案例，该机构藏书超过 4000 万册，相当于大约 6 万亿个 token。在中国，AI 仅用一天时间就能处理超过 20 倍于中国国家图书馆全部馆藏的数据量，这几乎是人类有史以来所写所有书籍总量的六倍。

风险在于 GTA 的规模：KV cache 增长极快，而仅 KV cache 就需要大量重复利用。与用户数据不同，KV cache 的存留时间非常短暂，却占据了巨大的存储容

仅供内部参考，请勿外传

量。这就是为什么这里需要专用存储的原因。

SSD 需要具备什么特性？高容量和耐用性。在大多数情况下，如果数据寿命为一天或更短，则使用 SLC；如果持续数天或更久，则使用 QLC。

2、PM 1763:

下面介绍三星为 KV Cache 共享存储设计的 SSD。三星去年推出了一款 PCIe Gen 6 SSD，并承诺相比上一代实现性能大幅提升，相关产品已经达成预期。PM 1763 实现了全球顶尖的性能表现。无论是顺序读写还是随机读写，所有参数都表现卓越。内部构造：采用 V9 TLC 闪存芯片，搭配 4 纳米制程的控制器，通过 PCIe Gen 6 接口，并配备了液冷系统。此外，我们计划在今年第四季度推出这款产品，敬请期待。

3、E3.S 256TB QLC SSD:

这是全球最小的 256 TB SSD 也是一款 256 TB 的硬盘，体积极其紧凑，却对数据中心极具价值。内部结构：搭载了 V9 QL C 闪存芯片，采用 5 纳米控制器，容量范围从 16 TB 到 56 TB 不等。我们在性能和可靠性方面实现了全面提升，堪称完美无缺。凭借这两款 SSD，我们将同时推动性能和容量的双重提升。

4、zNAND-O:

zNAND-O 基于 V10 BV-NAND 架构，凭借 TSV 技术，其性能和能效比传统 NAND 高出 7 倍。zNAND-O 控制器通过 TSV 与核心裸片相连。该产品在基底芯片上堆叠了多个核心芯片，并完全集成于单一系统之中。该产品通过 UCIe 标准接口实现芯片间连接，兼具高性能与低功耗特性。zNAND-O 支持通过增加堆叠层数实现纵向扩展，并支持横向扩展，满足 TB 级模型规模需求。2028 年起，zNAND-O 将深度赋能 AI 创新。

要点总结：核心芯片采用 SLC 架构，使用 TSV 硅通孔技术，将 NPU 与 zNAND-O 集成封装；芯片之间依靠 UCIe 标准接口互连，同时支持纵向扩展、横向扩展。

三星提出适配 zNAND-O 的全新设备端架构。三星全新 MPU 优化内存架构，将语音、KV Cache 保留在 DRAM 内部，规模大、计算密集的模型存放至 zNAND-O。凭借该套成本优化 AI 内存架构，zNAND-O 可以为设备端 AI 承载万亿级参数模型。

过去十年间，三星发布了一系列技术规范和白皮书，这些资料不仅推动了技术发展，还为中国数据中心行业注入了动力。不仅如此，三星还为业界带来了技术洞见，并在中国建立半导体晶圆厂方面迈出了重要一步。特别是在西安，三星投资建设了一座 120 万平方米的 NAND 闪存晶圆厂，另外两处 NAND 生产基地也已投入运营。三星已准备好参与任何形式的技术合作，这将推动中国人工智能产业的发展。

仅供内部参考，请勿外传

3、博通：AI at Scale：以太网作为通用互联底座

演讲嘉宾：Mohan Kalkunte，博通架构与技术副总裁

会议要点

1、行业变化：

AI 已经从集群计算走向分布式计算，网络本身成为计算架构的组成部分，以太网伴随 AI 负载持续迭代。以太网是通用化互联架构，三层内涵：适配任意规模场景（机架内、跨机架、跨数据中心、跨地域）；对接任意 XPU/GPU；承载训练、推理、分布式推理各类业务。

行业重大变化：以太网从传统 scale-out 场景，向 scale-up、scale-out、scale-across 三大领域扩张；scale-up 过去长期被私有互联主导，基于以太网的 scale-up 生态正在形成。

相关规范与组织：ESUN 1.0 规范正式发布，配套 SUE-T 协议；OCI 光计算互联组织成立，用于突破单机架铜缆传输距离限制；同时 UEC、MRC、字节、Meta 的 RoCE 等多条传输协议路线并行发展。

硬件迭代：交换芯片 Tomahawk6 带宽自 Tomahawk5（51.2T）提升至 102.4T，带宽与端口基数翻倍；端点速率从 400G 演进至 800G，未来向 1.6T 演进；scale-up 域不再局限单数据中心，支持多数据中心互联。

架构判断：现有问题已经不再是“以太网能否成为通用网络架构”，而是“以太网到底可以扩展到什么程度”。以太网在架构层面不存在硬性上限，可以持续承接不断演进的 AI 工作负载。

2、AI 工作负载演进及三大特征：

AI 负载从传统训练，拓展至强化学习、推理、分布式推理，具备三大特征：1）通信量大幅增长，张量并行、流水线并行带来大量实体间通信需求；2）流量异构性增强，除集合通信外，新增大量内存事务、存储事务、控制信令流量；3）业务部署物理空间不断扩张，由单机架走向多机架、大型集群，网络成为计算架构的一部分。

3、三大场景技术解析：

1）Scale up（高带宽低延迟紧耦合计算域）：是过去一年发生最大变革的领域。Scale-up 场景下 XPU 之间通过以太网执行类似内存 Load/Store 操作，需要对以太网做专项增强；协议栈分为 XPU 端点传输层、域内互联架构层。底层以标准以太网为基础，UEC 定义链路层增强：链路级重传、基于信用的流量控制 CBFC、优化报头格式。

行业存在多条自研 scale-up 传输协议路线，但行业共识：不同传输协议需要一套统一底层互联架构，这也是 ESUN 组织成立初衷，ESUN 1.0 让不同传输协议可跑在同一套通用架构。

规模演进目标：scale-up 域向 128、256、512 乃至 1K-2K 颗 XPU/GPU 发展。受铜缆距离限制，超大 scale-up 域无法容纳在单个物理机架内。OCI 组织目标

仅供内部参考，请勿外传

融合铜互联优势、拓展传输距离，实现 scale-up 域跨机架部署，重塑 AI 系统架构形态。

2）Scale out（多个 scale up 组成更大集群）：行业从 51.2T 向 102.4T 迭代，博通芯片从 Tomahawk 3 迭代到 Tomahawk 6，带宽大约提升 34 倍，端口基数相对上一代 51.2T 芯片实现翻倍。Tomahawk6 芯片可搭建 128K XPU 扁平化两级拓扑，对比三级拓扑，整机功耗、时延、光模块成本降低 40%，性能可靠性同步提升。

MRC（Multipath Reliable Connection）传输协议推出，解决传统 RoCEv2 RC 单路径、重传效率低、拥塞调优难的痛点，核心特性：1）自适应路由+数据包喷射，充分利用对分带宽，配套乱序数据放置机制；2）选择性重传，丢包仅重传对应数据包，而非整段序列；3）交换机侧数据包裁剪，拥塞时通知发送方。

3）scale-across（跨数据中心长距离互联）：算力超出单数据中心能力时使用，scale-across 互联多个数据中心，支撑百万级 GPU 规模。博通 Jericho 4 芯片面向该场景开发，高容量、高端口基数，支持长距离互联，集成深缓冲区吸收流量突发，长距离场景优先做延迟转发，规避丢包。

4、全协议栈创新总结：

1）应用层：集合通信、MoE 混合专家模型、解耦计算等创新；2）传输层：UET、SUE-T、MRC、Meta RoCE 等多条路线；3）网络侧：自适应路由、报文裁剪、BTS 拥塞信号、遥测技术；4）链路层：链路级重传、CBFC 信用流量控制、帧格式优化；5）交换侧：102.4T 交换机、高低延迟交换机、深缓冲区交换机；6）端点侧：端点速率 400G→800G；LPO、CPO、NPO 等光互联技术并行演进。

以太网依托开放标准，实现一套统一开放架构，适配任意 XPU、任意规模、任意 AI 业务负载。

4、超聚变：架构跃迁，超节点创新之路

演讲嘉宾：孙恪，超聚变数字技术股份有限公司智算中心领域副总经理

会议纪要

十年整柜与液冷演进：

公司自 2017 年开始布局三总线盲插液冷整机柜，用电、水、网三个维度同时解决高密服务器的供电、散热和互联问题。

通算时代风冷机柜约 15—20kW，公司液冷整柜做到 66kW、单柜 128 颗 CPU；2022—2023 年智算阶段，液冷整柜提高至 102kW，可部署 8 台 8 模组机型。

超节点阶段风冷不再适用，液冷整柜功率继续向 200kW、500kW 和 1MW 演进；升级不只涉及电源，还要求冷板、管路、连接器和高速线缆同步迭代。

“0—3”：三类核心部件自研

液冷侧覆盖冷板材料、加工工艺、快接头、二次管路和 TIM 导热材料，目的在于

仅供内部参考，请勿外传

使整个散热链路可控，而非只采购标准冷板。

电源侧由板级 DC-DC 延伸到整柜 Power Shelf，把效率、功率密度和可靠性作为系统能力。

高速侧与产业伙伴联合开发 Cable Tree 和连接器，速率由 25G 演进至 112G、224G。

公司认为超节点整机厂的差异化不只是集成 GPU，而在上述三类功能部件及系统工程能力；同时通过硬件架构兼容不同交换芯片和 Scale Up 协议，给客户保留选型空间。

“1—N”：规模复制与业务数据

超节点已从首个客户扩展到互联网、金融、运营商和科研院所，并建立全球供应与服务体系。进入新行业、区域并非简单复制，需要针对机房条件、运维和法规重新验证。

公司披露年交付超过 200 万 xPU、液冷节点年产能超过 30 万台；按单节点 4—8 颗 xPU 估算，已由万级样机进入数十万节点的规模商用。

规模带来三项反哺：更大采购量可推动液冷接头、冷板、电源等部件降本和良率提升；即便万分之一故障率也会形成大量故障样本，倒逼定位与修复体系；约 200 万设备运行数据可反向优化下一代架构。

5、阳光电源：从源到芯，算电协同：基于 SST 的 AIDC 供电解决方案新变革

演讲嘉宾：张迪，AIDC 电源事业部总裁

会议要点

1、阳光电源产业布局与 AIDC 领域进军思路：

企业基础：1997 年成立，深交所上市；光伏逆变器全球第一，储能系统全球前三；业务覆盖光、风、储、电、氢五大板块，2025 年营收 892 亿元。

技术与研发实力：2025 年研发投入 42 亿元；8000+研发工程师，硕士占比超 70%；参与 140 余项行业标准制定，专利 11000 余件，全球 7 大研发中心。

制造交付：拥有波兰、印度、泰国、合肥四大生产基地；电力电子设备累计交付 1000GW，电力电子产能 450GW，储能产能 100GWh，具备全球交付能力。

切入 AIDC 底层逻辑：依托电力电子核心技术，业务从发电侧、电网侧延伸至大型用电负载场景（AIDC、智慧矿山微电网）；掌握微电网、构网技术、电力电子变换、热管理、场景化应用等底层能力。

行业大背景：行业由传统 IDC 向 AIDC 演进，算力基地向大型集中基地转移，催生算电协同需求，公司具备大型基地工程实施与运维经验。

2、算力发展相关政策、行业趋势与现存痛点：

仅供内部参考，请勿外传

政策层面，算电协同写入十五五规划，AI 与能源双向赋能是未来五年重要课题。行业发展趋势方面，单机柜走向 1MW 成为确定性方向；制冷完成风冷向液冷迭代，供电即将由交流向原生直流演进。

AIDC 电力侧三大现存痛点：电力需求缺口大，算力建设快于电力配套；并网周期长，与数据中心建设节奏错配；AI 带来负载剧烈波动，对供电系统提出挑战。

现实挑战：电力短缺、项目交付速度要求提升，算-电、电-冷交织加深，前期仿真验证成为必备手段，规避投运故障。

公司研究思路：采用“以终为始”，先研判终局架构可行性，再倒推技术演进路径。

3、公司解决方案、产品能力与生态共建：

1）三类 AIDC 与电网融合典型场景及挑战：①离网起步、延迟并网场景：前期依靠微电网，后期并网；需要解决构网、黑启动、新能源配比，保障 AI 负载稳定。②电网配额受限场景：本地电网条件有限，项目具备自发电、限制峰值功率可优先并网；储能与算力深度融合，实现对负载、电网双向友好。③弱电网场景：部分新能源富集区域电网系统强度偏弱，大算力集群负载剧烈波动，存在断电风险，需要储能与关键电源融合保障系统稳定。

2）SST 固态变压器 EnerNeo 产品方案：供电架构变革方面，传统变压器技术迭代缓慢，AIDC 单机柜可达数百千瓦至 1MW，传统供电架构占用机房空间占比过高，亟需革新。SST 固态变压器把大量设备集成，适配行业从交流向直流演进的大趋势，实现中压直连高频化，解决占地难题。

核心产品参数：首款面向 AIDC 的 SST 固态变压器 EnerNeo，功率密度 312kW/㎡，相比传统供电系统占地面积减半；中压到 800V 低压全链路效率 98.5%，较传统方案提升 2 个百分点以上。

架构适配：针对 IDC 场景做专项优化，可覆盖分布式（可按照超节点单元分阶段上架算力）、2N（A/B 路替换）、4:3 等主流 AIDC 供电架构。

配套能力：公司输出完整整套解决方案，兼顾负载平滑、备电、交流负载、短时储能缓冲等过渡阶段需求，实现开箱即可在数据中心部署使用。

数字分身模型：同步发布硬件对应的数字分身，可接入仿真系统，完成电网架构、动态负载、全链路能量流验证，保障场站交付效果。

产品路标：已规划完整 SST 产品序列，后续推出不同电压等级、不同颗粒度产品，推动行业中压直流直连演进；SST 还可作为微电网能量路由器，落地算电协同。

6、中兴通讯：“Token 经济”时代，全栈协同引领 TCO 最优解

演讲嘉宾：陆威，中兴通讯股份有限公司智算产品规划总工

会议纪要

仅供内部参考，请勿外传

Token 经济的三项指标:

训练时代主要看算力规模；推理时代首先看系统每秒可生成 Token 数，决定总服务能力。

第二项是每瓦 Token 数，反映电能转化效率和长期运营成本；第三项是单用户每秒获得 Token 数，直接影响应用时延和用户体验。

三项指标共同指向 TCO：既不能只追求单卡峰值，也不能只看采购价格，而要看部署、供电、网络、运维和软件优化后的持续产出。

能力一：多芯协同

供应链和业务需求决定单一 GPU 难以覆盖全部场景，平台需兼容多款 CPU、GPU、DPU、网卡和大容量交换芯片。

传统方式先确定通用网络再集成芯片，可能无法释放芯片性能。中兴先通过软件仿真模型在不同芯片上的算力、带宽和通信需求，再定制互联拓扑，实现“芯片—模型—网络”协同。

多芯平台既用于供应安全，也可按训练、推理及成本目标选择不同组合，目标是在相同硬件预算下实现更高有效 Token 产出。

能力二：OEX 正交无背板架构

OEX 中垂直板卡为交换板、水平板卡为计算托盘，GPU 与交换芯片直接正交对插，缩短 112G/224G 信号传输距离，并实现电、水、网络三线无线缆盲插。

架构此前已在中高端路由器规模商用，并于 2025 中国算力大会获评重大突破成果，说明其并非仅停留在实验室阶段。

三项收益：无线缆释放风道和板卡空间，单柜支持 32/64/128 卡；短链路降低损耗并提高信号完整性；通过增加计算和交换部件继续扩展规模。

基于 OEX 可构建预制算力集装箱，在工厂完成硬件、软件和监控调试，现场连线后快速开通。计算托盘盲插、统一监控电/水/网，可降低部署和物理维护复杂度。

能力三：连接增强

MoE、PD 分离和集群扩张显著增加 GPU 通信。商用阶段强调成熟与可靠：小规模超节点在柜内使用电交换和电互联；十万卡场景由多个万卡资源池通过光网络连接。

前沿方向是把 OEX 与 dOCS 分布式光交换结合。dOCS 可按业务动态改变物理拓扑，减少光电转换环节和光模块数量，并实现百纳秒级低时延。

能力四：持续演进及全栈产品

整机层面推进算、存、网、供电和制冷协同；集群层面推进 Scale Up 与 Scale Out 融合，用更少交换机和光模块获得更高带宽；软件层面开展算电协同和算网协同调度。

仅供内部参考，请勿外传

产品分为芯片层、OEX 硬件基础设施层、软件基础设施层和推理服务层，软件覆盖推理引擎、KV Cache、算力调度、服务网关和服务调度。

公司称该开放架构可承接未来 6—10 年的芯片和软件升级，长期通过全栈协同持续改善 Token TCO。

7、新华三：以标准化共建开放超节点技术生态

演讲嘉宾：刘新民，新华三集团副总裁、技术委员会副主席

会议要点

1、行业现状与现实挑战：

算力需求高速增长，主流大模型性能约每 3.5 个月翻倍，国内两年新增 1400 个万亿级参数大模型，AI 驱动数据中心建设提速，基础设施承压。

行业痛点：1）算力碎片化：新建、扩容集群硬件容易不一致，业务迭代、硬件升级进一步加剧异构碎片化，服务器、CPU 型号多元。2）集群外部大互联技术成熟，但超节点节点内部互联呈现百花齐放的发展现状，同样存在不少问题。3）工程层面：过去只需要解决单设备内部问题，规模化时代，缺少标准化会抬高全行业成本，制约产业发展速度。

2、超节点标准化系统工程思路：

倡导行业共同攻克技术难题，沉淀软硬件货架式共享能力，核心方向包含三点：1）多元算力协同：多方协同项目不可避免出现碎片化，国内高端 CPU 成本偏高，需要标准化实现协同。2）互联走向开放：打破封闭“全家桶”方案，底层组件标准化，既可以使用“全家桶”整套方案，也可选用标准化组件，扩大技术选择空间、降低成本。3）供电、散热配套约束：单节点功率持续抬升，但受空间、冷却、模组体积限制，硬件无法无限做大，需要在有限空间实现高密度硬件设计。

各类模组标准出台：通过推进 UXM 标准化工作，完善模组标准层、整机柜技术底座；依托模块化、集成化，实现快速迭代、低成本落地。UXM 体系以统一供电、统一接口、统一高速互联实现开放，底层标准化组件之上，厂商仍可打造自有整套方案。分层架构支持分阶段迭代，便于行业形成统一技术语言、达成共识。

3、互联协议与物理连接：

互联协议层面，集群外部网络（节点外部的网络）成熟，但超节点节点内部网络尚处于发展阶段。机柜内横向 Scale-out 网络面临带宽、时延极限挑战，多个组织推进公开标准，各家解决底层问题一致。

新华三提出覆盖物理层、链路层、网络层、传输层的开放协议架构，协议开放，可交由第三方芯片厂商实现，部分接入芯片可封入 GPU 内部，部分交换协议芯片外置，构建 Scale-up 开放网络，同时支持定制化开发。

物理连接演进：连接形态由设备间连接走向芯片到芯片；传输距离从厘米级向毫米、微米级演进。到达性能边界后电信号难以满足需求，以光代电成为高速高可靠连接重要方向。

仅供内部参考，请勿外传

NPO 接口标准：NPO 技术本身并不完美，但从工程可行性看是现阶段最合适方案，可靠可用、承前启后；国内其他光互联方案落地挑战较大，NPO 具备现实可行性与演进潜力，已被行业接纳。新华三积极参与 NPO 多场景标准制定，推动配套工艺、光引擎规范建设。

4、互联架构路线：

正交架构、分布式架构均在落地，各有适用场景，更多体现工程实现能力，不存在绝对技术优劣。在 224G、448G 速率分界点，两种架构性能表现存在差异；未来工艺提升，二者性能差距存在不确定性。公司坚持标准化层面正交架构、分布式架构多路线并行，新工艺突破会重塑技术格局，公司内部有多条硬件平台，同步配套光轨标准层。

5、供电架构：

芯片功耗快速抬升，部分芯片功率突破 2000W。芯片功耗动态波动，对供电稳定性提出二次设计要求。整机功率从 200kW 向 500kW、兆瓦级演进。除高压直流供电外，供电垂直化成为重要方向，多年前公司和宇节合作过小批量产品，当前供电距离从数厘米压缩至毫米级，仍有大量技术挖掘空间

6、散热技术：

过去设备内部散热压力相对有限，行业仅需持续提升风扇转速、风量、风道、散热片设计。当前单芯片功耗突破 2000W，未来可达 5000W；交换芯片功耗突破 3500W，散热压力显著增大，散热体系层级下探至毫米、微米级，涉及多学科交叉，公司正在为高功耗交换芯片匹配散热方案。

当前行业主流为正压液冷（正压散热）路线，液冷部件、方案标准化是重点。国内已有负压液冷样机，运维层面具备潜在优势；极高功率密度场景效果尚待验证，现有功率条件下前景较好，负压降低对液体本身要求，有望提升散热效率。

新华三在多条技术方向开展研究，期待产业广泛研讨，协同创新共建多元开放 AI 生态。

8、晶丰明源：面向 AI 高密度算力集群，“最后一英寸”电源解决方案的演进

演讲嘉宾：张博一，晶丰明源高性能计算事业部中国区销售总经理

会议要点

1、功耗趋势：

2024 年全球数据中心耗电约 415TWh，接近英国全年用电；演讲保守预测 2030 年达到 945TWh、接近日本全年用电，激进口径可能达到 2024 年的 3 倍。

中国数据中心 2025 年耗电约 196TWh，2030 年可能达到 500—700TWh；全球占比可能由 2024 年的约 25%升至 2030 年的约 50%。上述均为演讲引用/测算口

仅供内部参考，请勿外传

径。

通算机柜 2021 年约 7kW，AI 机柜则可能由 2023 年的 40kW 提高至 2027 年的 600kW；按该口径，四年增长约 15 倍。

以 NVIDIA 芯片为参考，TDP 由 2020 年约 400W 提高至 2024 年约 1,000W，行业已出现 3.6kW 规划，并讨论 4—5kW 产品。按 0.8V 核心电压估算，Vcore 电流可能由约 450A 升至 4,050A。

2、四条供电技术路径：

母线降压：由 12V 向 6V 演进，使低压 MOS 获得更低 RDS(on)，并缩小电感等外围器件，有利于末级效率、动态响应和模块小型化；该路线可与其他方案叠加。

空间架构：由分立 SPS 转向模块化 VRM，布局从水平供电 LPD、背面供电，进一步演进至垂直供电 VPD 和 GPU 下方内嵌供电。演讲称 PDN 阻抗可改善一个数量级，VPD 电流密度有望达到 3A/mm²。

TLVR：通过跨导电感改善瞬态响应，减少外围电容，并将响应由毫秒/微秒级推进到纳秒级。

高频化：提高开关频率可缩小无源器件、提高功率密度，但会增加开关损耗，因此需与 6V 等低压母线配合。

3、公司产品与技术平台：

BPS 为 2008 年成立、2019 年科创板上市的 Fabless 公司，聚焦 Vcore 多相控制器和 DrMOS，不直接生产整机电源模组，而与模组厂商共同服务 GPU 和服务器客户。

LDMOS 合封方案将上管、下管和续流管以不同 Die 实现，可分别优化工艺；单颗 Die 可独立迭代，并通过金属层和基板封装在性能、良率、成本间灵活调整。

新一代 DrMOS 已量产，同时支持 6V 和 12V 母线，开关频率最高 2MHz，并集成电流检测和保护功能。

实测中，新器件自 15A 负载起较老产品具有效率优势，25A 重载效率优于海外竞品；50A 连续 5 分钟测试温升更低，强调高电流场景的热可靠性。

面向 VPD，器件可在同一平台从两相扩展至四相，并支持不同尺寸、长条外形和更薄封装；控制器兼容行业主流协议。量产产品覆盖当前需求，下一代产品面向未来 2—3 年的 GPU 升级。

9、庆虹电子：AI 超节点高速互连方案的实践与思考

演讲嘉宾：陈衍冲，庆虹电子市场部总监

会议要点

1、架构与产业变化：

国内头部厂商已推出磐久、大禹、银河、天池等多种超节点，不同产品在背板、

仅供内部参考，请勿外传

正交、飞线和光互联上各有侧重，互联部件不再是整机定型后的标准采购件。

芯片迭代由过去 3—5 年一代缩短为一年小版本、两年大版本，互联厂商需要同步支持不同芯片数量、封装和拓扑，研发与交期明显收紧。

主流形态包括背板式单面横插、无背板正交及无中板正交。随着信号由 112G、224G 向 448G 演进，PCB 链路裕量下降，行业由板级互联转向 Cable Tray、飞线等线缆化方案，长期是光铜共存。

2、五类互联硬件与 NPC 形态:

节点及整柜互联大致分为线缆背板、NPC 近芯片线缆、OTB 板外线缆、CPC 共封装铜互联和板级连接器。

NPC/CPC 必须前置参与芯片和封装布局;板级及客户定制产品则需与终端客户、ODM 协同，单一连接器厂商无法独立完成系统优化。

国产芯片在先进工艺和封装空间上约束更强。公司提供 4×4 规格、单缆 16 对差分通道的小型化方案，以及 8×8 规格、单缆 64 对差分通道的高带宽方案，在宽度、深度和出线空间之间权衡。

3、ARK 224G 及 448G 预研:

自研方舟 ARK 224G 高速背板接口单通道密度为 2.2×2.7mm，目标是在高密度下保留足够带宽。

224G 设计需缩短公母端子接触长度，同时解决冲压强度和制造良率；采用更低 DK、更低损耗材料，并缩小连接器屏蔽腔，避免磁场能量堆积造成 S 参数谐振。

面向 448G 的初步仿真已把原 224G 方案约 86GHz 的谐振点推迟到 110GHz，但尚需完成详细设计、制样和实测，不能等同于量产产品。

围绕 ARK 规划前置 I/O、OTB、板对线、线对线及连接器对连接器等端到端无源方案，同时储备低损耗、耐弯折的 224G/448G Raw Cable。

4、Cable Tray 量产与质量控制:

单柜 Cable Tray 可能包含数千至上万个接口，必须使用 XY 方向浮动实现对位自纠偏，并以 2—4mm Z 向浮动吸收装配落差和插拔力，降低连接器损伤。

设计验证包括极限插拔角度、偏位工况、高温老化、混合布线和长期可靠性。不同客户可能因质量风险放弃理论性能更高的方案，因此工程一致性比单次实验室结果更重要。

量产要求每根线缆、每条通道 100%执行无源 S 参数测试，关键尺寸进行 100% AOI，焊接工序检查尺寸和焊点强度；装配后再做全通道线序和 TDR 阻抗检测，排查断线、压伤和阻抗不连续。

公司后续将继续优化 ERL、ERPS、SCC 等指标的仿真与频域加权方法，联合上游开发异形铜材和 Raw Cable，并突破注塑、冲压等制造工艺。

仅供内部参考，请勿外传

三、服务器分论坛

1、安费诺：AI 算力底座的互连与供电：超节点、NPO 与 HVDC

演讲嘉宾：Roy Wu，技术总监

会议要点

本次分享内容覆盖超节点对应的同机互联部分，以及 NPO、HVDC 当中与互联强相关的内容。

1、超节点：

已实现大规模量产的超节点大多采用 Cable Tray 方案。GB200、GB300，以及未来将要量产的 VR200 均采用 Cable Tray 架构实现 GPU 集群互联。除英伟达外，北美多数厂商超节点以同机互联作为基础互联协议。国内各类超节点方案大多采用同机互联实现 scaleup 网络扩展。

架构差异：NV、AMD 超节点 GPU 与 CPU 部署在同一计算节点，无需 PCIe；国内已落地超节点 GPU、CPU 分属独立节点，带来 PCIe 互联机会；GPU 与交换单元之间 scale 网络通道数量大、架构复杂。国内两大主流架构为 Cable Tray 线缆背板、正交飞线架构，行业尚未形成单一主流，多路线并行。

安费诺对应产品方案：1）cable cartridge 模组化设计，解决传统线对板线缆可维护性差痛点，大量采用大范围浮动设计，适配整机对接的高要求。2）飞线架构两套方案：芯片侧飞接背板侧，节点之间依靠飞线完成互联；芯片侧飞接 IO/光模块侧，实现跨机柜 scale out 扩展。

速率与密度演进：国内当前量产速率 112G，后续向 224G、448G 演进；集群规模扩张，互联密度要求持续提升，面积密度目标由 5.76 向 4、2.4 演进；安费诺推出 3.2T、3.6T、6.4T、7.2T 带宽等级 NPC、CPC 共封装互联方案。

技术进度与链路实测：224G 北美已大批量量产；448G 安费诺已完成初步设计并产出实物样品，覆盖背板连接器、线缆对接、芯片侧飞线连接器完整链路；90GHz 条件下 SDD21 参数 35dB，电源串扰 55dB，具备 PAM6 下 448G 应用潜力，若用 PAM4 还需进一步优化。

2、NPO:

Cable cartridge、飞线架构逐步逼近传输物理极限，铜基互联随服务器规模扩张可传输距离缩短；光互联与电互联并行是技术趋势，铜、光互联在超节点不同场景均有落地选项。

行业进展：OFC、海内外 AI 大会多次讨论 CPO/NPO；OCA、ODCC、OIF 等标准化组织正在制定 NPO 相关标准，当前是规范 NPO 接口的关键窗口期。

安费诺产品方案：1）自研 NPC、CPC 方案已获北美大量客户使用，可提供兼容的 NPO 配套方案。2）配合 ODCC-UPO 工作组开发 NPO 方案，包含 LGA socket 插座形态，覆盖 6.4T-7.2T 场景，224G70GHz 条件下仿真性能优异，目标支持 448G，PAM6 可满足需求，PAM4 仍需优化。

仅供内部参考，请勿外传

形态路线观点：NPO 技术路线、外形规格多，socket 形态更利于标准化，规避两片式连接器跨厂商对接风险，同时公司也在开发两片式 Mirror Mezz 方案。

3、HVDC:

行业趋势：单卡功耗、超节点规模持续提升，HVDC 成为必然趋势；海内外 HVDC 架构形态不同，但核心互联环节大体一致：AC 取电→ACPDU 输出 800V 直流→Sidecar/Power Rack→FDA 配电单元→分发至各计算节点。

核心风险：连接器电流等级大幅提升，可靠性至关重要，需验证接触电阻漂移、热插拔电弧、震动、插拔松动等长期可靠性风险。

设计理念：行业厂商指标激进程度不一，安费诺采用保守设计，优先保障工业系统安全，重视长期稳定性、插拔寿命、安规要求。

产品布局：安费诺提供全链路电流分配完整解决方案，AC 侧覆盖 32 200 安培，DC 侧覆盖 power shelf 等环节；产品包含 PSU 输入输出连接器，800V 器件，节点、交换、计算设备取电连接器，覆盖 110 220 安培，适配客户差异化架构。

4. 高速互联与 800V HVDC 供电领域观点与产业需求:

技术布局：安费诺同时跟进超节点、NPO 光电两条路线，HVDC 海外推进节奏靠前，公司具备完整解决方案；作为本土化程度很高的美资企业，希望依托 ODCC 和国内伙伴共建高速、大电流互联生态。

可靠性重要性：高速互联领域 AI 训练场景下，高速互联误码会造成训练任务整体重启；供电领域 HVDC 多级转换、多环节分配，供电故障会对 AI 业务造成重大影响，高速、大电流新产品需要大量实验验证可靠性。

落地节奏判断：224G、448G 高速连接器国内落地尚需时间，448G 国内落地周期更长。北美或更快落地，336V-448V 方案未来 1-2 年处于研发与部署阶段。

产品价值：国内设备能力存在一定滞后，安费诺高密度高速产品适配未来场景；448G 在阻抗、串扰层面仍存在性能缺口；下一代高密度产品可改善国内 224G 场景下的 ISI 码间干扰问题。

2、天翼云：面向 Agentic AI 应用的超节点硬件架构创新

演讲嘉宾：白秀杨，中国电信天翼云研发专家

会议要点

1、Agentic AI 带来的硬件变化:

Agent 应用正从单一 Agent 走向任务规划、任务协调、结果确认等多个 Agent 协同；当输出不符合预期时，系统还会重新规划和执行，因此底层资源需求比传统单轮推理更复杂。

第一类瓶颈是 KV Cache。长上下文和并发量提升会快速占用 HBM，压缩 Batch Size 并降低吞吐；RAG、多轮对话又包含大量重复前缀，既要解决容量，也要解决高速传输和跨请求复用。

仅供内部参考，请勿外传

原纪要记录：70B 模型单 Token KV Cache 约 0.5MB、13B 模型约 0.78MB、采用 GQA 的 70B 模型约 0.31MB；并记录上下文达到 1.28K 时，A100/4090 显存可能被 KV Cache 基本占满。后两项数字与常见口径存在疑点，建议结合演讲 PPT 复核，但不影响“长上下文使 KV Cache 成为瓶颈”的方向判断。

缓存利用率超过 95%后，系统可能从计算密集型转为 I/O 密集型，KV Cache 搬运而非 GPU 计算成为吞吐制约。

第二类变化是 CPU 负载。传统业务中 CPU 主要做预处理和简单调度，CPU/GPU 常见配比为 1:4 或 1:8；多 Agent 需要工具调用、资源协调、结果同步，部分业务中 CPU 处理占总时延可达 90%以上，配比可能转向 1:2 甚至 1:1。

2、紫金超节点架构：

核心设计原则为开放生态、资源解耦、高速互联和极致能效。统一硬件平台兼容多款国产 AI 芯片，可按业务性能和成本选择最适合的芯片组合。

整机柜模块化设计打破固定 CPU/GPU 配比：Agent 业务可增加 CPU 模块，传统推理可增加 GPU 模块，并通过整柜交付简化部署和运维。

高速互联兼容多种 Scale Up 协议物理层及不同国产 GPU 技术路线，支持 224G 速率。部分国产 GPU 单卡功耗已经超过 1.6kW，因此采用直流入柜、高效率双输入电源模块和冷板液冷，并在机柜、节点两级部署漏液监控。

机柜统一监控 GPU 状态、互联链路及供电单元。集群包含 XPU 算力节点、CPU 通算节点、内存池节点，以及面向 L3/L4 不同层级 SSD 的混合存储节点；Scale Up 网络同时承载 GPU 互联与 KV Cache 流量，前端网络负责业务流量。

整体思路是将算力、通算、内存和存储由资源孤岛转为统一资源池，按任务动态调度，提升多 Agent 场景的资源利用率。

3、HPKV 多级 KV Cache 服务：

HPKV 将 KV Cache 从 GPU HBM 扩展至 CPU 内存和高性能存储，扩大预取与缓存复用范围。

长文本场景中，将无法容纳于显存的 KV Cache 卸载至 KV 存储，业务全局复用后 TTFT 最高可下降 80%。

智能客服、AI 助手、在线教育等多轮对话场景包含大量重复系统提示词，HPKV 可使缓存命中率提升至 90%以上，减少重复计算和 Token 成本。

云 API 等高并发场景中，缓存下沉可解除显存对并发量的限制；演讲实测 64 并发吞吐提升 2 倍，部分场景每秒吞吐最高提升 10 倍。

4、后续方向：

提升互联密度，并配套 800V HVDC 与全液冷，解决更高单卡功耗和机柜密度。

探索 Scale Up、Scale Light 等网络的同平面带宽共享，缓解不同业务下带宽利用不均。

从 CPU/GPU 灵活配比进一步走向内存和存储全面池化，解决多机 KV Cache 共

仅供内部参考，请勿外传

享。

Agent 场景不能只看 MFU，需同时关注整数、浮点和通算资源的综合利用率，推动通算与智算融合。

3、锐捷网络：800V 高压直流机柜供电演进

演讲嘉宾：李武，锐捷网络电源专家

会议要点

1、AI 算力功耗爆发背景下，50V 架构的局限性：

AI 芯片功耗快速抬升：从早期几十 W，提升至 H100 的 700W、B200 的 1kW，未来 GPU 单芯片功耗可达 5kW；单机柜最多 144 颗 GPU，单芯片功耗突破 5kW，机柜整机功率可达 600kW 1MW。

传统 50V 架构四大制约：1）配电容量触顶，功率上限约 300kW；2）PSU 功率密度存在瓶颈，实现 1MW 供电需占用 30U 设备空间，空间成本高；3）母线损耗随功率平方级激增，600kW 机柜 50V 架构电流约 12000A，切换 800V 后电流下降 15 倍，回路损耗降至原有 1/225；4）散热能力触顶，传统架构 6 次电能变换，变换损耗全部转为热量，风冷无法匹配。

2、800V HVDC 成为物理必然及产业共识：

两种实现方案：800V 直接架构、正负 400V 架构，本质都是 800V；正负 400V 为三线架构，存在中线不平衡风险。800V 架构带来多方面系统收益，端到端变换效率提升 3-5 个百分点；铜材使用量减少 45%；机房电能变换环节由 6 级缩减至 3 级。

产业共识：英伟达 Rubin 平台采用 800V HVDC；OCP 推进对应标准，80 余家 MGX 生态伙伴共建生态，头部厂商均选择 800V 路线。

3、数据中心三代供电架构演进以及 50V、12V、6V 三条母线迭代路径：

三代供电架构：1）第一代 12V 分布式供电：整机柜功率上限约 30kW；2）第二代 50V 架构：整机柜功率上限约 250kW；3）第三代 800V，分两个演进阶段：①第一阶段：800V 转 50V 集中供电。800V 经 Power Shelf 降压，沿用原有 50V 铜排与服务器板载方案，生态兼容性好；占用 8-10U 机柜空间，整机柜功率可达 700kW。英伟达采用 1U 90kW 架构，由 6 台 15kW PSU 组成，同时支持风冷、液冷；国内厂商方案差异化，普遍选择液冷，搭配 ATS 器件，功率覆盖 50kW、70kW、100kW。

②第二阶段：800V 直入服务器节点。取消 50V 变换，高压转低压在节点内部完成，供电路径大幅压缩；英伟达 Kyber 机架采用该方案；预计未来两年商业推广，整机柜功率上限可达 1MW。

800V 落地后三条中间母线技术路线：1）800V 转 50V，再转 12V/6V 给 XPU 供电。三级变换，整体效率约 85%；复用成熟 50V 体系，成熟度最高，有望率先落地。2）800V 直接转 12V 给 XPU 供电。母线电流大、12V VR 转换效率偏低，

仅供内部参考，请勿外传

大功率机柜适用性有限，大概率不会成为主流。3）800V 先转 6V，再由 6V VR 供电。效率提升 1 2 个百分点；短板为单节点大功率工况下，6V 母线传输电流可达数千安培。

4、800V 进柜、进板的核心挑战：

以英伟达 Kyber 机架为标杆，整机功率 600kW-1MW， 800V 直入节点带来多重难题：EMC、高功率密度 PDB 设计、器件生态不成熟、800V 节点安规标准缺失。

对应解决路径：1）EMC 治理：源头抑制干扰，RC 吸收电路把电压变化率控制在 300V 每纳秒以内；采用三点分离接地与 EMI 滤波。2）PDB 功率密度：采用多颗全砖电源模块并联；锐捷推出 800V 输入 50V 输出 PDB，输出功率 24kW，峰值效率 97.5%，整机尺寸 350×180×34 毫米，功率密度水平较高。3）器件生态：高压连接器、熔断器、第三代半导体、热插拔控制器等产业链产品加速成熟。4）安规设计：分级选材，选用 CTI＞600 的 FR-5 板材，严格管控爬电距离、电气间隙；大电流采用多路并联、线传变针传。计算板节点内部，多相模组扩容，引入 IVR；高动态工况做 EDPp 专项处理；采用垂直供电、供电堆叠方案。

垂直供电演进逻辑：VRM 不断贴近负载，缩短功率传输路径，降低损耗、提升功率密度。横向供电 I^2R 损耗可达 8%-10%；VRM 置于 PCB 背面，系统净效率提升约 10 个百分点；第三代 VRM 内嵌 PCB 内部，功率密度提升至 3A/mm²；终极形态为 IVR，适配未来 5kW GPU。

EDPp（GPU 动态脉冲功率）治理：EDPp 幅值可达 GPU TDP 的 120%-175%，管控不当会造成宕机、电网跳闸。1）800V 架构下，储能电容、超级电容迁移至 800V 母线或 sidecar；2）采用芯片-板-PSU-机房侧分级协同抑制；3）分频率治理：低频扰动在机房供电侧处理，高频扰动在机柜内部处理；4）当前缺少统一标准，需要产业链协同制定边界规范。

5、行业展望：

演进分为三维度：标准上 OCP 标准落地，国内同步对标推进；器件上第三代半导体、连接器、超级电容持续迭代；架构上 800V 转 50V 成为规模商用主流，12V、6V 中间母线持续验证。

产业时间节奏：2026 年 800V 转 50V 进入出货阶段；中间母线技术为未来两年演进方向；VPD、IVR 为量产终极方案。

核心结论：800V 是物理必然和产业共识；800V 转 50V 是未来 2-3 年工程落地主线；PDB、EDPp、IVR 是未来数年产业链重点研究方向。

4、百度：Agent 时代的计算节点模块化设计

演讲嘉宾：盛永，百度技术专家

会议要点

1、超节点行业现状：

仅供内部参考，请勿外传

（1）超节点的重要性

AI 时代到来，超节点作为 AI 基础设施的核心产品，正在推动数据中心发生深刻技术变革。

（2）超节点面临技术挑战

超节点实现规模化批量交付之后，会遇到多方面技术挑战，主要归纳为三点：

第一，规模跃迁，超节点规模化上线交付，会带动数据中心基础设施发生变革与演进。

第二，计算节点层面产品方案多元化。GPU 领域，除英伟达持续迭代产品之外，国内昆仑芯等多款 GPU 产品同步发展，算力供给呈现多元化，产品需要实现多方案兼容。CPU 领域，除英特尔、AMD 主流平台，还有海光等国产算力平台持续迭代。CPU 与 GPU 各类组合，每一类组合都需要全新的系统研发设计。

第三，不同硬件组合会带来研发资源投入、开发周期方面的挑战。从业务需求看，新兴算力迭代背景下，客户对新产品导入速度要求提高，传统计算服务器开发周期较长，现阶段要求产品在一年之内完成上线。

（3）百度的解决方案

将计算节点服务器拆分为硬件、结构、供电、散热、网络、管理六大模组，采用模块化设计。

模组可以像积木一样自由组合堆叠，快速完成产品开发迭代，不再需要针对每一套硬件组合做完整系统级验证。研发工作从整机系统级别下沉至模组级别，能够大幅压缩开发周期，减少资源投入。模块化设计同时需要兼顾不同客户业务场景需求，I/O 层面保留充足扩展弹性。从供应链角度，各模组均支持国产化替代方案，保障多路径供应能力。在 AI 集群高密度算力场景中，如果设备可维护性较差，故障会造成算力损失，设备可运维性也是重要设计考量。

2、百度天池超节点系统：

（1）系统架构

系统包含计算节点、GPU 节点、电源板、风扇模组、液冷模组、高密连接器等 13 项关键核心部件。

（2）分模组设计思路

第一，硬件模组，核心理念是实现多元化兼容，支持平滑向下一代产品迭代。

主板模组实现多平台兼容，不同平台主板物理尺寸统一，支持主板模块化迭代。GPU 模组兼容主 GPU 方案，昆仑芯等各类国产 GPU，在统一结构下完成模组兼容。背板、转接卡、交换板等其余部件实现通用化，搭配不同模组灵活组合，以此加快平台研发迭代速度。

第二，结构模组，重点关注极致 I/O 扩展能力、弹性网络、极简可运维性。

机箱采用标准 21 英寸 1SU 设计。机箱前窗划分为五个功能区域：第一区域布置管理网卡，支持普通网卡、双宽智能网卡等多类硬件，适配不同业务场景；第二区域设置 10 个 RDMA 网卡位；中间区域支持 SSD 设备；右侧布置系统盘。传统服务器设计为追求非热插拔，一般将系统盘布置在机箱内部。面向 AI 集群，运维要求提升，百度将系统盘布置在前窗，在有限空间内实现 I/O 能力最大化扩

仅供内部参考，请勿外传

展。机箱后窗全部采用热插拔设计，包含液冷冷板快接头、高密连接器、供电卡扣接口，以此提升整机可维护性。

第三，供电模组。超节点场景不再适用传统分布式供电方案，线缆密度、传输路径损耗会大幅上升，无法满足超节点产品发展。百度采用集中供电方案，节点后侧通过卡扣接口对接母排，实现快速运维接口，降低供电管理复杂度，简化线缆管理，实现统一电源管理。节点内部设置电源分配板，向各个模组分配电力，依靠高密度供电方案释放单机柜算力密度，避免供电成为算力瓶颈。

第四，散热模组。当前部分芯片功耗已经超过 1.2kW，传统风冷无法满足散热需求，行业普遍采用冷板式液冷散热方案。CPU、GPU 等高功耗核心芯片采用冷板液冷散热，其余低功耗部件保留传统风冷方案，整机功耗显著下降，散热能力得到提升。液冷快接头采用浮动盲插结构，支持热运维快速操作。冷板散热方案可将 CPU 温度降低 25℃以上，GPU 温度降低 20℃，充分释放 GPU 算力，避免因高温降频限制算力输出。

第五，管理模组，核心目标是保障设备交付之后模组对应的算力不会因为单机故障长时间无法对外提供服务。

超节点属于高密度算力产品，传统故障报修处理时效无法满足业务要求。管理模组设计核心要求为实时感知、自动响应，支撑快速故障处置，保障算力尽快交付业务使用，算力直接对应业务成本。整机沿用传统 BMC 带外管理架构，重点强化实时智能监控。冷板式液冷大规模部署场景下，液冷漏液监控与故障管理是核心设计要点。百度拥有长达十年冷板液冷线上部署运维经验，相关运维经验落地到超节点产品，实现液冷故障精准预测与维修。同时实现告警精准性管控，依靠准确告警提升维修效率。所有接口遵循行业标准接口，支持不同产品快速兼容。

总结：百度已经推出 AI 超节点相关产品，当前行业多家企业都在开发超节点高密度算力产品，百度呼吁行业协同推进模块化标准制定，减少多家企业并行开发带来的资源重复投入，助力 AI 行业发展。

仅供内部参考，请勿外传

四、超大规模集群分论坛

1、百度云：百度智能云 AI 全栈体系赋能超大规模集群建设

演讲嘉宾：李丁，百度云 百度混合云解决方案专家

会议纪要

1、超大集群的基础设施约束：

单卡设计功耗已接近 1kW。十万卡仅 GPU 功耗就达到百 MW，叠加 CPU、内存、网络和供电配套后可能达到 200—300MW，需直接从 220kV 主干电网取电并配套建设变电站。

GPU 训练负载不同于通用服务器，TP/DP 并行时仍有较高空载基线，业务功耗波动可达约 30%；映射到机房可能形成几十至上百 kW 的瞬时变化，因此不能只按设备标称功耗设计配电。

十万卡集群约需 6 栋、每栋约 2 万平方米且承载 50MW 的建筑。单栋楼宇、园区取电和跨楼网络必须从项目早期统一规划，普通园区很难直接满足条件。

RDMA 当前较优通信距离约 150m，超出后衰减和时延压力增加；百度正尝试将有效距离扩展到 200—300m，以适应楼内及跨楼互联。

风冷机柜约 30kW 后效率、噪声和风量均接近上限，万卡/十万卡项目需前置规划风液混冷或液冷。随着单柜功率由 4.4kW、20kW 继续上升，变配电和制冷设施占用的建筑空间将明显超过 IT 设备本身。

2、百度 AI 全栈产品：

P800 板卡服务器是百度全国产集群的主要硬件，3 万卡级集群均使用同型号，减少异构带来的运维和性能波动。

超节点可在单柜部署 32 卡或 64 卡，相比 4 台独立 8 卡服务器，GPU 距离更短、互联效率更高，更适合十万卡集群和 Token 产出优化。

百度希望在千卡、万卡项目早期参与楼宇、机柜及网络拓扑设计。虽然不是传统弱电集成商，但具备大规模智算集群系统集成和运行经验。

自研百舸平台负责集群纳管、训练/推理任务下发、资产登记和故障分域，可同时兼容百度自研、NVIDIA、寒武纪、海光等多种硬件。

开源模型和框架更多追求普遍可用，云厂商可针对客户模型和业务做训推优化；同一套硬件的 Token 产出差异由软件栈、并行策略、资源调度和故障恢复共同决定。

3、集群运营与案例：

对政府、城投和大型企业自建集群，百度可通过多云管理平台提供运维、计费和业务分发，解决“建成后不会运营”问题。

10,240 卡稳定性测试连续运行 84 小时，故障率 0.1%；期间共记录 40 张故障卡，平均恢复时间 20 分钟。核心在于将故障现象、历史统计和处置策略固化至平台，

仅供内部参考，请勿外传

实现自动自愈。

宁夏中卫 3.2 万卡全国产集群全部采用 P800，CPU 为海光四代 7490 系列，总算力约 16,000P、PUE 接近 1.2，并配套并行存储，主要承载训练任务。

集群规模扩大通常会使训练有效率由千卡时接近 99%下降至三千卡时约 70%—80%；百度称其万卡训练有效率仍可维持 80%以上，并经过第三方评测。

仅供内部参考，请勿外传

五、大模型算力分论坛

1、海光信息：大模型推理时代，国产算力与开放生态的双轮答卷

演讲嘉宾：杜夏威，总裁助理兼智能计算产品部总经理

会议要点

1、产业现状与软硬协同必由之路：

行业进入效率优先的迭代周期，芯片迭代以年为单位，模型迭代速度快，软硬件迭代周期存在天然鸿沟，弥合该鸿沟是 Token 经济发展的关键。

产业共识：软硬件协同优化是必由之路，是缩小与国际先进水平性能差距的关键。1）软件侧：需要完成从上层模型算法、推理引擎到底层算子、汇编层的全栈软件栈工程化优化。2）硬件侧：芯片定型后改动适配软件迭代难度高，采用 shift left 左移设计理念，强化与模型侧联动；同时拓展超节点、万卡/十万卡大规模集群等系统层面工作，实现与软件侧衔接。

2、产业四大关键挑战：

1）内存显存挑战（内存墙）：受容量、访存带宽约束凸显；PD 分离架构不同场景下 P、D 部件价值不同；倡议产业理性研判，避免一窝蜂式技术路线选择，防止衍生新产业问题。

2）通信墙：MoE 大模型对全局通信带宽需求提升；不能只追求纸面性能，要关注性价比，最终落脚到单 Token 成本合理化，才可以实现商业化落地。

2）MFU 算力利用率：不仅取决于芯片设计，还和大模型稀疏化（MoE）强相关；需要构建覆盖算力、访存、通信、IO 的完整适配系统。

4）能耗约束：单机柜功耗、土地、电力、机架（LPS）成为算力中心硬性约束；需要全局一体化设计，算电协同、绿电、储能成为大算力产业必备课题。

3、海光信息芯片与软件栈产品落地：

硬件芯片矩阵：覆盖边缘轻量设备、高性价比 PCIe 设备（双卡/四卡桥接）；主力产品基于 HSL 高速互连域，推出 8 卡、16 卡、40 卡超节点。下半年超节点需求上涨，驱动因素：①大模型参数量提升，需要更大 HSL 域、显存支撑推理与 PD 分离切分；②训推一体诉求，波峰跑业务、波谷回收算力做训练，超节点是 Token 工厂基础单元。

三层软件栈：1）DTK 底层基础软件栈：支撑编译层，对接各类框架算子，实现极致性能优化。2）DAS 中层 AI 软件栈：算子、推理引擎整合，开箱即用，逐步开源，构建开发者生态飞轮。3）DAP 顶层平台层：覆盖对话、Agent 智能体、DAP-Token，实现 Token 监控、统计、运营，打通算力-Token-业务。

4、算力交付模式与生态：

仅供内部参考，请勿外传

全域协同算力交付：1）大规模公有云侧：超算/算力互联网云化交付，用户便捷获取 Token，免去重资产运维。2）私域部署：兼顾数据边界，释放自有数据价值，迭代垂类模型、垂类 Agent。云上+私有化协同，实现从底层算力到上层 Token 由“可用”迈向“好用”。

生态共建：开放镜像仓库、模型仓库、开发者开源社区、Agent 与 Token 管控工具；依靠用户反馈驱动产品迭代，联合大模型、基础设施伙伴共同把算力转化为可交付业务能力。

2、壁仞科技：异构并行分发架构下大模 Token 负载协同优化

演讲嘉宾：高瑞阳，研发高级经理

会议要点

1、PD 异构设计背景：

上一代国产芯片 HBM2E 带宽存在瓶颈，单设备 460G/s，四层堆叠仅 1.8T；而旗舰大模型 Decode 阶段带宽需求达到 2T-10T，无法满足 SLA 要求，但该类芯片 Prefill 算力充足。

PD 异构方案：存量 HBM2E 国产芯片作为 Prefill（P）节点，发挥算力优势；下一代国产芯片或英伟达 GPU 作为 Decode（D）节点。复用存量 AIDC 硬件资源，支撑前沿大模型服务。

底层通信基础：依托 RDMA；规避 NCCL 仅适配英伟达 GPU 的限制；采用网卡直连通信，完成不同架构 GPU 内存映射，高性能、可跨服务器扩展，支撑异构推理。

2、三类 KV Cache 传输实现策略：

1）P2P 直连方案：P 节点生成 KV Cache 直接传给 D 节点做解码；路径短易优化；缺点 P、D 节点预先绑定，无法解耦，其他节点感知不到缓存。

2）共享存储方案：P 把 KV Cache 写入 store 集群（CPU-DRAM/SSD），全系统广播，D 节点查询读取；优势 P、D 完全解耦，由 store 统一管理。

3）生产环境主流-混合路径：P 节点一方面直接把 Cache 下发 D 节点；同时写入本地 store；新会话请求到来，其他 PD 节点可从 store 读取历史 KV Cache，兼顾时延与解耦能力。

3、存储资源需求测算：

以 1000 用户、每用户 20 个 session、单 session 上下文 32K token，经过 GLM5 模型 KV Cache 压缩后，单 token 缓存 90.5KB，整体需要 55.2TB 存储，无法全部放在 HBM 显存。多级缓存合理分层：HBM 承担热缓存，存储高频访问 session 对应的 cache，容量规模约 500GB-800GB；分布式 CPU DRAM 承担 11TB-13TB；SSD 承担 43TB-52TB。

仅供内部参考，请勿外传

4、国产芯片作为 Prefill 节点的商业可行性：

单机 P 节点收入计算公式：单机月收入=单机时长×单机吞吐×Token 定价×利用率。虽然 Agent 任务可以持续运行，但用户请求存在波峰波谷，夜间存在业务波谷，集群无法始终跑满单机吞吐，因此必须引入利用率参数。

单月总时长换算为 30×24，等价于 2.592 百万秒。Token 定价参考智谱官方对外报价：cache 命中的 Token，对应实际 Prefill 计算，每百万 Token 定价 8 元；cache 未命中 Token，每百万 Token 定价 2 元。基于上述参数，可以得到单机月收入计算表达式。

公式包含两个核心变量。第一个变量为 cache 命中率 hit rate，多轮对话场景中，如果 cache 已经在系统内完成计算，就可以直接复用缓存，不需要重新执行计算。该部分 Token 单价更低，同时不占用单机吞吐资源，仅完成检索读取即可，对业务收益影响极大。第二个变量为利用率。

代入单卡 500Gps、单机 4000Gps 的吞吐参数进行测算，得到单机月收入 8 万元。该测算结果属于保守基线，公式中多个变量都会对最终收益产生影响。

进一步分析，基线单机月收入为 8 万元。1）开展性能优化，若单卡吞吐由 500Gps 提升至 1000Gps，根据公式，单机收入将实现翻倍，因此性能优化具备极高业务价值。2）利用率：商业落地过程中提升利用率，收益呈正比例变化，利用率由 0.5 提升至 1.0，单机月收入可以提升至 33 万元。3）cache 命中率，该变量操作空间最大，与系统技术实现深度相关。可以通过系统层面多项手段提升 cache 命中率，例如多轮对话场景下对长 Decode 做处理、对 Prefill 结果做存储等。若 cache 命中率从 0.8 提升至 0.9，单机月收入可以从 33 万元提升至 53 万元。对比基线 8 万元，收益提升幅度十分显著。

以上就是上一代国产芯片承担 Prefill 任务能够算清经济账的底层原因，通过各个环节精细化优化，输出高质量、有效 Token。

仅供内部参考，请勿外传

六、百度智能云基础设施分论坛

1、百度云：Token 工业化时代：超节点架构重构与智算底座演进

演讲嘉宾：武正辉，百度智能云系统部资深工程师

会议要点

1、超节点为何需要重构：

大模型参数量由线性转为指数增长，单机 8 卡已难以支撑 MoE、PD 分离、长上下文和多种并行策略，集群规模由千卡向万卡、十万卡扩展。

传统 CPU+GPU 逐步变为 CPU+GPU+XPU/ASIC 的异构融合架构；硬件需要在同一系统内完成多类算力协同，而不是为每种组合重新定制服务器。

超节点的直接价值是提高单位空间卡数、缩短 GPU 距离并增加 Scale Up 带宽，但随之产生供电、散热、信号完整性和维护挑战。

2、四类关键瓶颈：

散热与功耗：天池第一代单柜约 33kW，后续提高到 70kW、140—300kW 乃至 500kW 以上，54V 供电难以长期支撑，需转向 800V；高密 GPU 则要求液冷和更精细的功耗稳态控制。

互联：速率由 112G 向 224G、448G 演进。机柜内铜互联仍具优势，长距离逐步由光承担；CPU、GPU、ASIC 等异构部件也需统一通信。

模块化：固定配置无法匹配多样模型，应让电源、计算、网络和互联单元按需组合、独立迭代。

带宽：Scale Up 带宽要由十余 TB 级提高至百 TB 级，才能减少 GPU 孤岛和通信等待，并支持更大模型及 KV 共享。

3、天池 1.0—3.0 路线：

天池 1.0 于 2024 年启动研发、2025 年 7 月上线并已规模落地；2.0 已进入部分客户试点和批量部署准备；3.0 为下一代规划。

单柜卡数由 32 卡、64 卡走向 128 卡和 256 卡；单柜功耗由 33kW、约 70kW 提高至规划的 140—300kW。

Scale Up 域由第一代 256 卡、第二代 512 卡，进一步规划至 1,024 卡及以上。

互联方面，1.0 使用 Cable Tray，2.0 继续提升其传输与可靠性；3.0 将结合 Cable Tray、正交背板和光互联/CPO/NPO，面向 448G 解决串扰、损耗及系统集成问题。

供电方面，前期采用 54V，3.0 在 128/256 卡阶段可能使用 54V 与 800V 共存、即 800V 入柜后转 54V；生态成熟后再走向 800V 直入节点。

仅供内部参考，请勿外传

4、整柜与计算节点设计：

天池整柜由电源、计算、网络、机柜本体、Cable Tray 和 RMC 管理单元六大模块组成，彼此解耦；单一模块故障可独立更换，减少对其余计算资源的影响。

节点内部继续拆分 CPU、存储、交换和 GPU 模块。天池 3.0 计划取消部分柜内线缆，实现更高程度的无线缆连接，并推动 CPU、GPU 及其他部件全液冷。

2.0 较 1.0 显著提高节点 I/O 带宽；机柜维持 46U、21 英寸规格，并进一步支持硬件热插拔和快速维护。

5、Scale Up 网络、内存池及 Token 工厂：

64 卡规模可使用 Cable Tray 或正交架构；128 卡以上逐步引入光。单柜或双柜 256 卡通过多类交换/高速互联芯片扩展，之后以 NPO 实现单层 1,024 卡全互联，服务 10 万亿参数及更大模型。

针对 PD 分离和长上下文显存不足，天池利用 Scale Up 域做 KV Cache 池化，并预研 CXL Memory Pool：通过 Fabric Switch 连接 Memory 器件，使内存与显存共享，扩展有效缓存容量。

Token 工厂第一阶段以铜互联天池超节点为基础，第二阶段随 NPO 成熟扩大光互联域，目标形成数百 TB 级内部带宽。

软件侧结合模型、推理和资源调度优化，提高同一套硬件的 Token 产出和性价比；架构按训推一体设计，可在业务需求变化时切换训练与推理。

3.0 计划依次实现单柜 64 卡、光铜混合 256 卡、512 卡，最终以 NPO 扩至 1,024 卡。关键待攻克项是光方案能否达到 Cable Tray 接近的低功耗、高可靠，以及高密机柜的供电和散热。

仅供内部参考，请勿外传

七、算电协同分论坛

1、麦格米特：面向兆瓦级 AI 机柜的 800 VDC Sidecar 供电架构

演讲嘉宾：施恒亮，麦格米特 AIDC 电源产品经理

会议要点

演讲关注的五项问题：

功率扩展：随着 GPU 和机柜功率提升，以模块方式增加供电能力，而不是整体替换。

高压安全：800V 系统需要完备的隔离、保护和状态监控。

动态负载：电源需适配 GPU 快速阶跃及高峰值功率。

能量管理：协调超级电容、电池和电网，降低负载波动对电网的冲击，并提供掉电穿越。

系统控制：将上述功能纳入统一实时控制和监控平台。

1、从分布式电源到机架能源平台：

数据中心电源可分为 IT Shelf 内部分布式、IT 机架内集中式，以及独立机架级能源平台三类。

Sidecar 把前端功率变换和能源管理从 IT 柜中解耦，可应对毫秒级负载阶跃，并在电网故障时延长供电时间。

GPU 及服务器迭代速度快，而供电基础设施寿命更长。解耦后，IT 侧可更换新 GPU，供电侧则按功率需求增加模块、储能或控制功能，无需同步重建整柜。

2、800VDC Sidecar 产品形态：

过渡方案在 IT 机柜内部配置 90kW Power Shelf，将 PFC 和 AC 前端放入 Sidecar，IT 柜只承担 DC/DC 转换，从而释放 IT 空间。

800V 进行机柜间传输，到 IT 柜附近再转为 50V，缩短低压大电流路径，减少铜缆用量和配电损耗；多个 Power 模块可并联，按实际上架功率扩容。

在 800V 机柜铜排成熟前，Sidecar 内部 110kW Power Shelf 可直接连接 DC/DC 电源架，一台 Sidecar 供 2—3 台 IT 机架。

兆瓦级机房成熟后，形态将转为一台 Sidecar 对应一台 IT 机架；最终由中高压直流远距离传输，经 SST 直接输出 800V 进入 IT Shelf，再由 PDB 向 GPU 供电，尽量减少中间变换级数。

仅供内部参考，请勿外传

八、AI 工厂分论坛

1、小红书：万卡 AI 集群稳定与智能运维实践

演讲嘉宾：张艺龄，小红书网络产品专家

会议要点

1、稳定性指标与口径：

单点故障会使其他 Rank 在 Collective 阶段等待，最终造成任务 Hang、Fail 或超时。一个包含 1,000 余张卡的 Pod 若停顿 10 分钟，约损失 170 GPU 卡时。

结果指标包括有效训练率和无效 GPU 卡时。有效训练率=有效 GPU 卡时/已分配 GPU 卡时；只有 Global Step、Step Time 及性能处于合理基线的时段才计为“有效”。

无效卡时按受影响 GPU 数量 × 无效持续时间累计；过程指标则分别记录故障发现、响应和恢复时长，便于定位损失发生在监控、决策还是维修环节。

2、五步稳定性闭环：

统一数据底座：把 GPU、网络、存储和作业作为业务对象，汇总 Metrics、Logs、Events 及业务/网络拓扑，再将原始数据收敛为标准事件。

判断 GPU 是否真正有效：同时观察 GPU 利用率、功耗等资源信号，Global Step/Step Time 等业务进展，以及 MFU、I/O 等待、TTFT 等性能指标，避免把降频或慢节点误算为有效卡时。

从面到点定位：先看 Pod、网络平面和作业拓扑，再下钻至时间段、端口、模块、线缆或单卡，通过增量和峰值识别异常对象。

固化证据：对 XID、GPU、网络和业务异常做持久化、去重和标准化；原始证据纳入 SLO，供人工及 AI 复盘，而不是只保留最终结论。

受控重建：事件上报后先隔离节点/故障域并保留现场，再检查 Checkpoint，选择回滚或重建策略；通过 Global Time 和时延指标确认业务真正恢复。

3、三类故障案例：

NVLink 瞬时 Inactive：驱动出现 XID 149/145，作业中断或重启。通过不同 Pod 和层级灰度，开启 NVLink 恢复及链路负载均衡，降低瞬态故障影响。

GPU ECC：累积不可纠正 ECC 直接隔离下线；Row Remap Pending 可在维护窗口 Reset。维修后需压力测试，通过后才重新回池。

RDMA：800G CX-8 网卡通过双 400G 链路上联两个网络平面，Bond 整体失效时，结合两台 Leaf 报错排除上联交换端口，再通过替换模块、线缆、直通组件交叉定位。

4、多 Agent 智能运维：

仅供内部参考，请勿外传

主 Agent 解析事件上下文并派生子 Agent，并行查询 Metrics、Logs 和 Events；部分子 Agent 按预定义意图执行带内、带外 GPU 及 RDMA 取证脚本，最终生成证据包。

工作台输出五类工件：事件卡、证据包、按置信度排序的根因假设、处置/回退动作、验收报告；每个完成 RCA 的 Case 会自动进入知识库和规则迭代。

安全边界包括默认只读、事实与推断分离、按风险等级选择自动执行或人工审批，以及对回退和止损动作实施权限隔离。

自动取证覆盖率接近 100%，人工约 30 分钟的工作压缩至约 3 分钟；P1 故障响应自动化率 98%，线路/端口/卡自动隔离累计执行 2,000 余次。

演讲披露：机器宕机等显性故障场景有效训练率约 99%，降频和慢节点等隐性故障约 94%；硬件维护平均少于 4 小时，P90 少于 12 小时。

仅供内部参考，请勿外传

九、网络分论坛

1、立讯技术：AI Rack 高速互连方案介绍及 448G 高速铜互连可行性探讨

演讲嘉宾：金龙，技术产品经理

会议要点

1、当前互联方案的发展趋势：

当前整个互联行业呈现多轨并行、双轨并进的发展大方向。无论是 Scale Up、Scale Out，还是 Scale Across，本质都是实现系统扩容，以此获取更高的算力密度。

不同的技术路线，对应的互联方案选型也存在差异。Scale Up 为单机柜内部纵向扩容，传输距离通常在 10 米以内。在 224G 速率时代，铜互联依然可以很好地满足该场景需求。从传统 DAC 无源铜缆，再到 AEC 有源铜缆，铜互联的可达距离也在持续延伸。该场景是铜介质的主力应用场景，遵循行业“能用铜则用铜”的选型原则。

Scale Out 属于横向扩展场景，传输距离达到 10 米以上。该场景中，可插拔光模块凭借部署灵活、支持长距离互联的特性，成为方案主力。Scale Across 面向更远距离互联场景，主要依靠光技术实现。

业界很多观点提及“光进铜退”，但从产业实际发展来看，更贴切的趋势是光铜并进。业界各类基于铜介质的 PHY 接口技术，也在持续延展铜互联的可达距离。铜介质凭借短距离、高密度、低成本的核心优势，在未来互联体系当中，依然会占据重要的一席之地。

英伟达各代互联目标演进趋势：无论是 Rubin、Rubin Ultra，还是 Feynman 架构，单机柜内部 Scale Up 场景，均以铜互联作为主力。随着速率持续迭代，铜互联也从 112G、224G，持续向 448G 演进。从技术路标可以看出，不管是节点内部还是机柜内部，铜互联都是主力，也印证了行业“能用铜则用铜，必须用光再用光”的发展思路。

2、立讯技术 224G 完整互联解决方案：

目前主要有两类主流架构，体现立讯技术光铜并进、端到端的产品布局。

第一类是全光互联前向架构，采用 NPO、CPO 组合方案，主要解决跨节点长距离连接，通过 NPO 跳线搭配 Shuffle box，构建高带宽、低损耗的光信号互联通道。

第二类为前向光+后向铜的混合架构，更贴合当前业务落地形态。机柜内部 Scale Up 互联，采用 Cable Cartridge 全铜无源线缆模组，完成 GPU 之间的短距互联；对外横向扩展则采用光互联方式，实现更高的部署密度。

可以看到，现在的方案覆盖从 GPU 到交换芯片、从板内到板间、从机柜到机柜的完整端到端链路。过去的模式一般是先完成系统整体设计，再去挑选适配的互联零部件。但现在，互联方案本身的瓶颈，会在一定程度上影响整体硬件架构的

仅供内部参考，请勿外传

实现。因此在项目设计最早期，就需要定义各类互联部件，零部件的性能上限，直接决定硬件系统的性能天花板。这也是立讯提出端到端解决方案能力的背景。客户选用立讯技术产品时，从连接器到线缆，再到整套系统布局，我们提供完整产品矩阵，帮助客户降低系统集成复杂度，实现一次对接完成全链路打通。

立讯技术部件级完整产品布局：覆盖高速连接器、线缆组件、CPO/NPO 连光板组件，同时配套机柜节点内部散热、电源方案，实现全场景端到端覆盖。从底层零部件开始，帮助客户在性能与成本之间取得平衡；同时针对硬件系统的性能天花板与短板，输出对应的产品方案。

3、立讯技术在 448G 技术上的研究现状：

随着速率不断提升，铜互联的有效传输距离持续被压缩。从 56G 到 112G，再到 224G，无源铜互联的有效距离持续缩短：56G 可以达到 3 米；224G 无源铜互联标准距离约 1 米；到 448G 时代，可能仅为 0.5 米，这是介质物理特性决定的客观趋势。

想要延续铜缆方案的生命周期，就需要更极致的设计，降低整条链路的信号损耗，提升链路整体性能。另一个核心挑战来自奈奎斯特频率，这是物理第一性原理约束。虽然可以通过提升调制阶数，例如从 PAM4 升级到 PAM6、PAM8，以此降低对奈奎斯特频率的要求，但会带来非常明显的影响：高阶调制对整个互联链路的信噪比要求会显著提高。

带宽压力虽然得到缓解，但时延、光电调制不匹配、额外器件带来的成本上升，都会成为新的瓶颈。除此之外，接口界面、substrate 封装基板部分带来的 stub 残桩问题，也会成为互联方案的重要制约因素。传统连接方案受 stub 影响，带宽大致处于 80-90GHz 区间，目前做到 80GHz 问题不大，但面向 PAM6、PAM4 这类更高标准，会面临巨大挑战。以上是 448G 互联面临的各类挑战。针对这些问题，立讯技术已经开展 448G 完整方案预研，2026 年已经完成多轮 448G 技术探索。

铜互联链路当中，线缆与连接器是两大核心组件。例如在 1 米线缆组件链路中，两端连接器带来的损耗约 3-4dB，绝大部分损耗集中在线缆本体，裸线的性能至关重要。公司裸线缆已经完成产品开发，测试带宽可以达到 100GHz 以上。在 110G 的测试条件下，整体带宽表现可达 110GHz 以上，可以很好适配 PAM4、PAM6。

想要构建低损耗链路，传统 NPC 近板铜互连方案会逐步向 CPC 共封装铜互连演进。也就是将距离芯片较远的连接器向芯片侧迁移。当下行业更多听到 NPC 方案；面向 448G 时代，行业将不得不考虑 CPC 共封装铜互连，把连接器和芯片 substrate 基板集成在一起，降低整条链路的信号损耗。

解决 CPC 链路之后，还有另外两大模块：第一是背板连接方案，立讯技术已经开展对应的产品规划。仿真条件下带宽大于 90GHz，实测带宽达到 100GHz 不存在明显障碍。

最后是可插拔 IO 部分，用于节点之间、设备之间互联。IO 层面，立讯针对 Lane-channel 方案，以及外部互联 DAC、AEC、AOC 都做了对应的产品规划。由此形成从母线、CPC、BP 背板连接器，再到 IO 的全场景互联方案。

下面展示 448G 裸线实测结果。立讯已经具备 27-30 AWG 规格的 448G cable

仅供内部参考，请勿外传

开发能力，仅从基础裸线缆层面，实现 448G 铜互联不存在瓶颈。

CPC 方案将连接器集成至芯片 substrate 基板，依托零短截 stub 连接技术，消除互配界面 stub 对带宽的负面影响，带宽可以满足 448G 业务需求。链路实测整体带宽大于 100GHz，实物样品已经产出。CPC 方案可以完全满足 PAM6 要求；面向 PAM4 还会迭代新版本，PAM6 场景下已经完全可用。

IO 部分的设计，我们创新采用非金属接触连接方案，规避金属接触 stub 带来的带宽影响，同时保留可插拔特性，接口密度与现有可插拔 IO 保持同一水平，带宽上限为 90GHz，支持 PAM6 场景；面向 PAM4 的迭代版本正在加速开发，预计今年内可以推出。

背板端连接器产品，延续立讯 Intrepid™系列设计风格，在互配可靠性、pin 针防护上做大量优化，规避公母头互配时的单 pin 损伤问题；整体采用全屏蔽设计，当前带宽约 90GHz，可以支持 PAM6；迭代版本持续向 PAM 4448G 目标优化。

前面为仿真测试结果，下面展示完整链路 demo 实测：实现 CPC→IO，通过 IO 线缆、DAC 再回到 IO→CPC 的完整闭环链路场景。

链路实测带宽可达 100GHz，单点、双点实测结果均验证 PAM6 448G 铜缆方案具备可行性。实际落地还需要结合客户业务场景、互联长度做方案取舍，但可以看到铜互联方案在支持 448G 上具备实现条件。

另外一套 demo 为背板链路，实现 CPC→BP、BP→BP 连接链路，实测带宽表现同样达标，完成 CPC→BP、CPC→IO 节点内全铜互联验证，具备 448G 互联支撑能力。

2、腾讯/阿里：UPO（Ultra Performance Optics）定义智算光互连新范式

演讲嘉宾：杨光、黄一元，腾讯网络硬件架构师/阿里

会议要点

1、Scale UP 互联趋势：

智算互联场景中，带宽压力最大的是超节点 Scale UP 互联架构。超节点 Scale UP 包含铜互联、光互联两条技术路线，机柜内部目前主要采用铜互联。当超节点规模持续扩张，扩展至 128 卡乃至 256 卡以上规模之后，单机柜会面临散热、供电、整机重量、cable tray 架构制造难度等多重上限约束，业务必然走向跨机柜互联，跨机柜场景需要采用光互联，光互联领域有多条技术路线并行推进。

国内已经实现产品落地的第一条路线为 LPO，U2 级别 LPO 已经实现一定规模部署。当速率提升至 224G 之后，LPO 会遇到应用层面的挑战，后续会做详细说明。第二条路线为 CPO，海外已经有部分部署案例，但 CPO 存在明显短板：封装难度大、良率偏低；整体生态封闭，光芯片、光引擎与电芯片深度绑定，不利于短期大规模落地。NPO 综合继承两类方案的优势：光引擎靠近芯片排布，光链路、电链路整体路径短，传输性能优异；通过标准化模组接口，实现光引擎与电芯片解耦，支持多款电芯片适配多家模组厂商产品，具备良好的产业化落地条件，因此预期 NPO 在未来会获得更多应用。

仅供内部参考，请勿外传

2、腾讯 NPO 领域实践：

腾讯较早识别 NPO 的应用前景，开展相关产品研发，自研 3.2T NPO 模组、配套 NPO 计算节点、102.4T NPO 交换机，依靠整套产品可以实现最大 512 卡 Scale Out。当前产品已经完成互联调通，BER 指标符合预期，预计不久之后将开展规模化部署。

产品开发过程中，NPO 计算节点、NPO 交换机均遇到现实技术问题。交换芯片单芯片带宽巨大，单颗芯片需要扇出大量模组，交换机硬件设计复杂度随之显著提升，主要面临三项挑战：

第一，NPO 模组安装一致性问题。单排模组数量多，需要保障压接可靠，保证各个位置压力均匀，这是设备调试阶段遇到的首要难题。

第二，机柜内部光纤理线挑战。单颗 NPO 模组引出 6 根光纤，32 个模组合计接近 200 根光纤。光纤带有保护套，线缆外径较粗，理线设计既要规避光纤应力损伤，同时兼顾安装便利性与后期可维护性。

第三，NPO 模组靠近芯片摆放，会挤占电源布局空间，同时带来光电耦合散热难题。以上问题已经在这一代产品完成解决。

NPO 交换机和传统可插拔光模块设备差异很大，需要模组与整机系统协同设计，无法依靠单一部件独立解决全部问题。面向下一代 200T 产品直驱光引擎，腾讯提出多项优化方向。NPO 模组层面，滤波器选型除性能指标之外，还需要重点关注安装可靠性、安装便捷性；探索下一代模组支持可插拔 FAU、内置光源，简化光纤布线，改善模组维护难度。整机层面，推进模块化引线设计、整机垂直供电技术，该代产品已经落地，下一代将应对更大规模带来的新增挑战；产品升级液冷方案，开展协同冷板设计。

3、224G LPO 技术：

112G 时代 LPO 已经实现批量商用，产品路线选择取决于整机架构设计以及芯片 service 的能力。224G 速率下，可插拔方案会面临巨大挑战。我们判断，224G 时代，如果 NPO 技术成熟，对比 LPO 会拥有更多性能与成本优势。基于三代产品研发经验，腾讯、阿里以及行业多家厂商，共同定义 3.2T、6.4T 下一代 NPO 模组规范。

4、6.4T UPO 标准：

该标准自去年年底启动，由阿里、腾讯联合发起，美团、快手等国内互联网企业，以及产业链大量上下游合作伙伴共同参与制定。

UPO 标准制定确立三项核心原则：第一，灵活性，定义三款不同产品形态，覆盖多样应用场景与系统需求；第二，高性能，标准原生支持 224G 速率，可平滑演进至下一代 448G；第三，可靠性与可维护性，这是 UPO 实现大规模量产、大规模上线的关键条件。

UPO 主要应用场景分为四类：交换机、GPU、MPU 存储扩展、网卡。不同场景对单卡密度、空间约束、执行条件的需求差异较大，单一模组形态无法适配全部场景，因此定义 UPO Type1 XD、UPO Type2 HD、UPO Type3 SD 三款形态。

仅供内部参考，请勿外传

UPO Type1 XD 密度最高，性能上限最高，但设计实现挑战较大，适合系统空间紧张、性能要求高的场景，需要配套高密度封装光器件实现。

UPO Type2 HD 为前期重点落地的形态，在密度、性能、工程实现难度上最均衡，适配现有各类硅光方案，系统适配范围最广。

UPO Type3 SD 外形尺寸更大，便于实现内置光源，适配传统 2D 硅光方案、对整机空间约束不苛刻的场景，可以采用该模组简化部署，同时降低成本。

UPO 模组与主芯片集成具备多种实现方案。On-PCB 为经典方案，UPO 模组与主芯片通过高速主板完成互联。Over-Substrate 属于较新集成方式，可以获得更优通道性能。On-Substrate、On-CoWoP 属于互联密度、集成度更高的方案。UPO 标准定义非焊接连接方式，可以适配以上多种主芯片集成方案。

性能层面，UPO 标准采用 LGA Socket 连接方案，该方案具备两项特征：连接高度低，可以实现很短的连接距离；仅设置两处接触点，采用非焊接结构，插损、回损、串扰指标表现优异。实测指标可以满足 224G 应用需求，经过评估可以平滑演进至 448G。

可靠性与可维护性方面，UPO 标准充分吸收腾讯 3.2T NPO 模组与系统的设计验证经验，做了多项优化。XY 方向增加定位销，提升对位精度；Z 轴推荐独立压块加背板固定方案，保障多模组连续压接的接触可靠性；采用无焊接连接器，规避焊接带来的污染风险；模组装配至 Socket 无需额外辅助工具。

UPO 标准内部正在评估、验证可分离 FAU 方案，实现 NPO 尾纤与模组本体分离，便于系统适配，同时支持光纤、光接头单独更换维修。

下面介绍 UPO 技术路标。2026 年 NPO 受到行业广泛关注，多家标准组织、厂商推进相关标准。本套 UPO 标准推进速度较快，V1.0 版本已于 8 月提交至 ODCC；8 月底 Type2 HD Socket 完成送样；6.4T UPO 模组与 6.4T UPO 系统预计 10 月可以开展联调。标准制定、关键连接器组件研发验证、模组系统验证多项工作并行推进。今年年底目标推出 Type1 XD Socket 产品；明年启动 448G UPO 标准制定。

3、腾讯：IO-NET：Scale Up 内存池构建内存互联超节点

演讲嘉宾：张潍丰，腾讯星脉网络高级架构师

会议要点

1、问题与系统方案：

现有层级中，L1 HBM 速度高但容量小，L2 存储容量大但带宽低，两者之间存在明显 GAP。Fabric Memory 作为 L1.5 层，主要服务稀疏 KV 和流式 KV 访问。

在原 GPU 超节点之外增加独立内存池，通过 Scale Up Switch 接入专用 DDR；交换侧配置 Memory Connect/协议转换芯片，以统一字节读写语义连接不同 DRAM DIMM。

演讲口径为单 GPU 增加约 500GB 访问容量和约 400—500GB“访问带宽”；原文未标注是否为 GB/s，建议回看 PPT 确认。

仅供内部参考，请勿外传

2、性能收益及适用边界：

上下文大于 32K 时，建模显示 Batch Size 中位数提升至原来的 5.3 倍，单卡 TPS 中位数提升至 2.8 倍；长上下文越大，释放 HBM 后可容纳的并发越多。

上下文大于 30K 时 Batch Size 可提高 4—5 倍；256K—1M 区间可超过 5 倍，TPS 约提高至 2.7 倍。

收益受三类条件约束：业务 SLO、上下文长度、模型注意力结构。方案更适合 DeepSeek、GLM、混元等稀疏注意力/长上下文模型，而非所有推理任务。

SLO 为 30—100ms 且上下文超过 32K 时，建模性能可提高 2 倍以上；30ms SLO 长上下文适合 Agent 任务，强化学习在 1M 上下文、70—100ms 条件下吞吐可提高 3—5 倍。

1M 上下文模拟中，GLM 5.2 中位性能提高 9 倍，DeepSeek V4 Pro、V3.2 及 Flash 约提高 3 倍，均为演讲建模口径。

3、成本与样机数据：

不同模型每 GPU 需额外配置约 300—600GB DRAM：DeepSeek V4 Pro 约 300—400GB，GLM 5.2 和混元约 300—600GB。

额外内存使单 GPU TCO 提高 20%—30%，因此必须通过 TPS 提升覆盖硬件成本，不能只比较峰值性能。

公司已在多款平台完成 GLM、DeepSeek V3 样机测试。GLM 在 32K—64K 上下文下 TPS 提升 5.15 倍；依靠高带宽缓存落盘，每 Token 耗时 TPOT 增加不足 10%。

4、PD 分离软件栈：

Prefill 阶段生成的主 KV 直接写入 Fabric Memory，Decode 实例的 HBM 只保留索引 KV；Decode 每生成一个 Token，再从内存池拉取所需数据。

新增两类通信：P 节点向内存池写 KV，D 节点逐 Token 读取 KV。软件需增加 Fabric Memory 控制面，负责实例预留、注册、分配和释放。

通信层新增重写、聚集和分散等算子，统一称为 FMIO（Fabric Memory I/O），要求网络、内存和推理引擎联合优化。

5、下一代产品形态与路线图：

2025 年的重点是通过 Scale Up Switch 连接 GPU、CPU 等算力；2026—2027 年进一步把 Scale Up Switch 升级为 Memory Switch，向上接入随机读写 DDR 及顺序读写 NAND/HBF 等 L1.5 介质。

形态上可由左右两个 64 卡 GPU 机柜加中间交换/内存柜，构成 128 卡、带独立内存池的超节点；Memory Switch Fabric 一侧连接 CPU/GPU/TPU/CPS，另一侧连接 DRAM 和 SSD。

项目计划 2026 年 9 月启动 Memory 软件层产品化、10 月开展业务验证，2027 年推进 Memory Connect 芯片、IO-NET 及 IO DIE 相关产品。

仅供内部参考，请勿外传

参考报告：

1、《国产超节点跟踪报告—WAIC2026 超节点加速推出，关注国产算力未来放量节奏》2026-07-24

2、《英伟达 COMPUTEX 2026 跟踪报告—Rubin 全面量产，Vera 专为 Agent 设计性能超 x86》2026-06-02

3、《MemoryS 2026 闪存大会跟踪报告：行业缺货或将延续至 27 年，关注未来存储技术创新重构》2026-03-29

仅供内部参考，请勿外传

分析师承诺

负责本研究报告的每一位证券分析师，在此申明，本报告清晰、准确地反映了分析师本人的研究观点。本人薪酬的任何部分过去不曾与、现在不与，未来也将不会与本报告中的具体推荐或观点直接或间接相关。

评级说明

报告中所涉及的投资评级采用相对评级体系，基于报告发布日后 6-12 个月内公司股价（或行业指数）相对同期当地市场基准指数的市场表现预期。其中，A 股市场以沪深 300 指数为基准；香港市场以恒生指数为基准；美国市场以标普 500 指数为基准。具体标准如下：

股票评级

强烈推荐：预期公司股价涨幅超越基准指数 20%以上

增持：预期公司股价涨幅超越基准指数 5-20%之间

中性：预期公司股价变动幅度相对基准指数介于±5%之间

减持：预期公司股价表现弱于基准指数 5%以上

行业评级

推荐：行业基本面向好，预期行业指数超越基准指数

中性：行业基本面稳定，预期行业指数跟随基准指数

回避：行业基本面转弱，预期行业指数弱于基准指数

重要声明

本报告由招商证券股份有限公司（以下简称“本公司”）编制。本公司具有中国证监会许可的证券投资咨询业务资格。本报告基于合法取得的信息，但本公司对这些信息的准确性和完整性不作任何保证。本报告所包含的分析基于各种假设，不同假设可能导致分析结果出现重大不同。报告中的内容和意见仅供参考，并不构成对所述证券买卖的出价，在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议。除法律或规则规定必须承担的责任外，本公司及其雇员不对使用本报告及其内容所引发的任何直接或间接损失负任何责任。本公司或关联机构可能会持有报告中所提到的公司所发行的证券头寸并进行交易，还可能为这些公司提供或争取提供投资银行业务服务。客户应当考虑到本公司可能存在可能影响本报告客观性的利益冲突。

本报告版权归本公司所有。本公司保留所有权利。未经本公司事先书面许可，任何机构和个人均不得以任何形式翻版、复制、引用或转载，否则，本公司将保留随时追究其法律责任的权利。