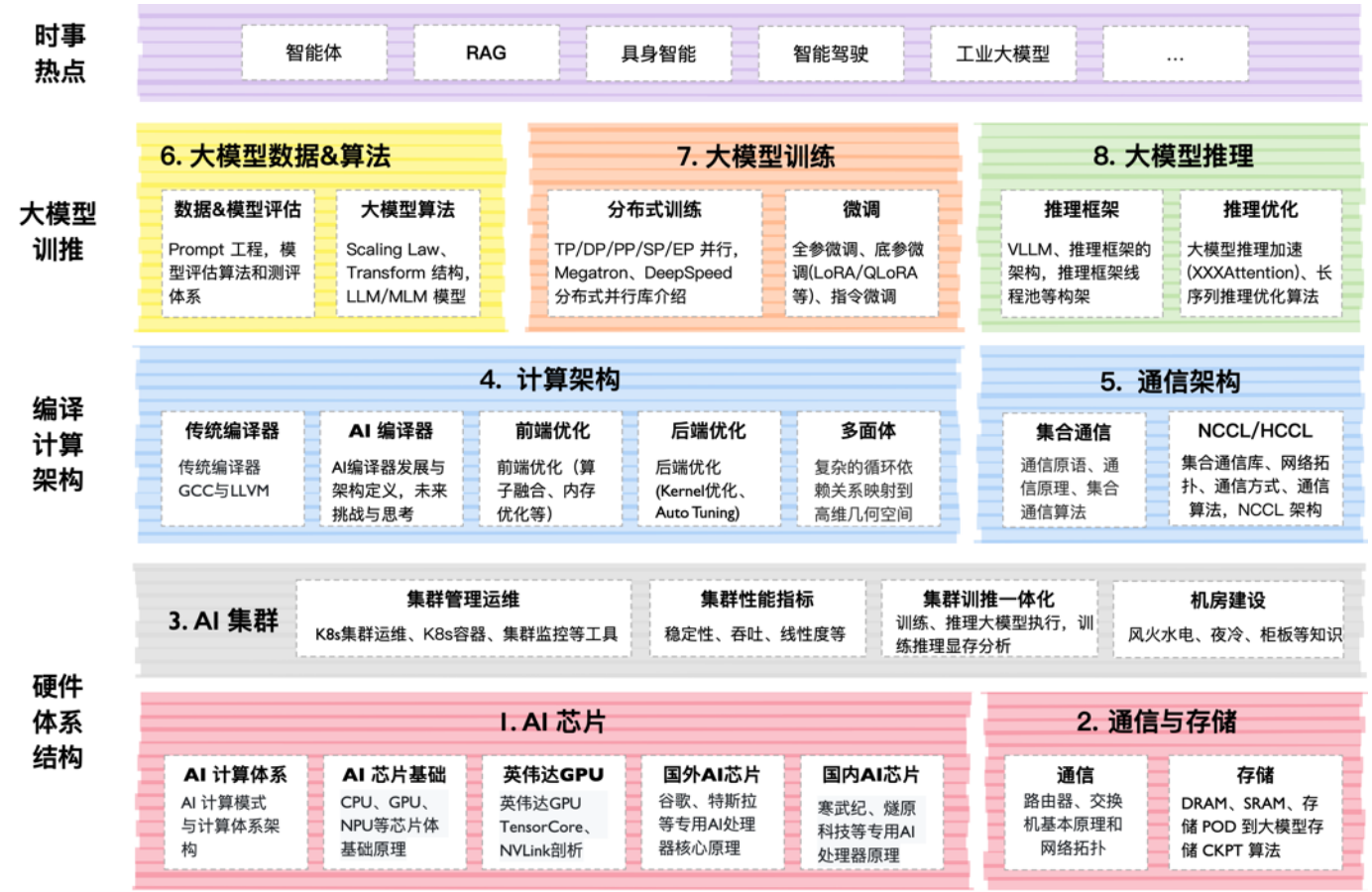
万卡AI集群建设挑战

ZOMI

Question?

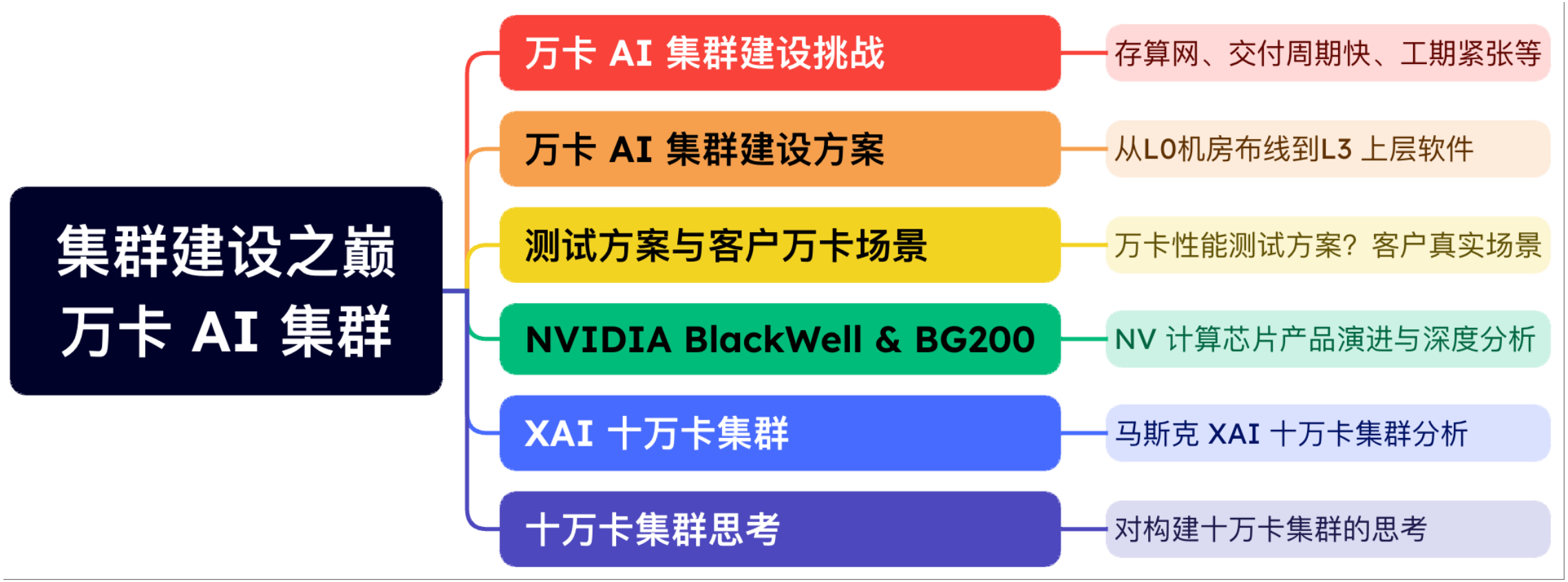
- 万卡集群，感觉很难，但不就是把 GPU/NPU 都放在一起堆料吗？

Content github.com/Infrasys-AI/AIInfra



GitHub github.com/Infrasys-AI/AIInfra book infrasys-ai.github.io

Content


集群建设之巅
万卡 AI 集群
万卡 AI 集群建设挑战
存算网、交付周期快、工期紧张等
万卡 AI 集群建设方案
从L0机房布线到L3 上层软件
测试方案与客户万卡场景
万卡性能测试方案？客户真实场景
NVIDIA BlackWell & BG200
NV 计算芯片产品演进与深度分析
XAI 十万卡集群
马斯克 XAI 十万卡集群分析
十万卡集群思考
对构建十万卡集群的思考

Content

1. 整体建设挑战
2. 能耗极限、网络瓶颈、计算效率、系统可靠性

01

整体建设挑战

万卡 AI 集群建设挑战

- **超万卡 GPU/NPU 大规模 AI 集群**
  - **大容量：**算力容量 >5000P，存储资源池 >200P，交换机 ~1000台
  - **多设备：**N 个机房 >1000 机柜（5000台设备）
  - **多线缆：**>10w 光模块布线，>8W根数据线缆，>20W个接头，~8W 焙纤端子
- **参与方多，实施强交叉，工期紧机房准备度低**
  - **参与万众多：**设计院、研发部、云管中心、设备供应商、第三方设备厂商、土建方
  - **基建重叠：**L1/L2 交叉实施 ~X 个月
- **能耗密度突破物理极限，电力供应与散热挑战高**
  - **单节点能耗高：**单卡能耗 350W，单节点 3000W，单柜 20kW
  - **多设备集中：**N 个机房 >1000 机柜（5000台设备），地理位置集中

万卡 AI 集群建设挑战

- **万卡互联的网络通信带宽与延迟瓶颈**
  - **多层互联：**交换机 ~1000台，框框、盒盒盒多层互联结构
  - **通信算法：**集合通信、点对点通信超规模后 Ring、Have Doubling 回环时间长
- **计算效率提升，万卡规模互联算力损耗难题**
  - **AI 节点：>2000** 节点，AI 卡数 ~16000
  - **存储容量：~800** ROCE 交换机
  - **大模型：PyTorch** + Megatron + Dense/MoE 从十以到千亿规模训练测试，万卡集群 K8S 调度
- **系统可靠性，故障率指数级上升**
- **国产化挑战：生态与性能的双重差距**
- **成本与运营：千亿级投资的商业风险**

星际之门

- OpenAI、软银等企业总投资5000亿美元，可配 40 万颗 NV AI 芯片

02

能耗密度破极限

电力供应与散热挑战：能耗密度突破物理极限

- **万卡建设能耗现状：**
  - **单节点能耗高：**单卡能耗 350W，单节点 3000W，单柜 20kW
  - **多设备集中：**N 个机房 >1000 机柜（5000台设备），地理位置集中

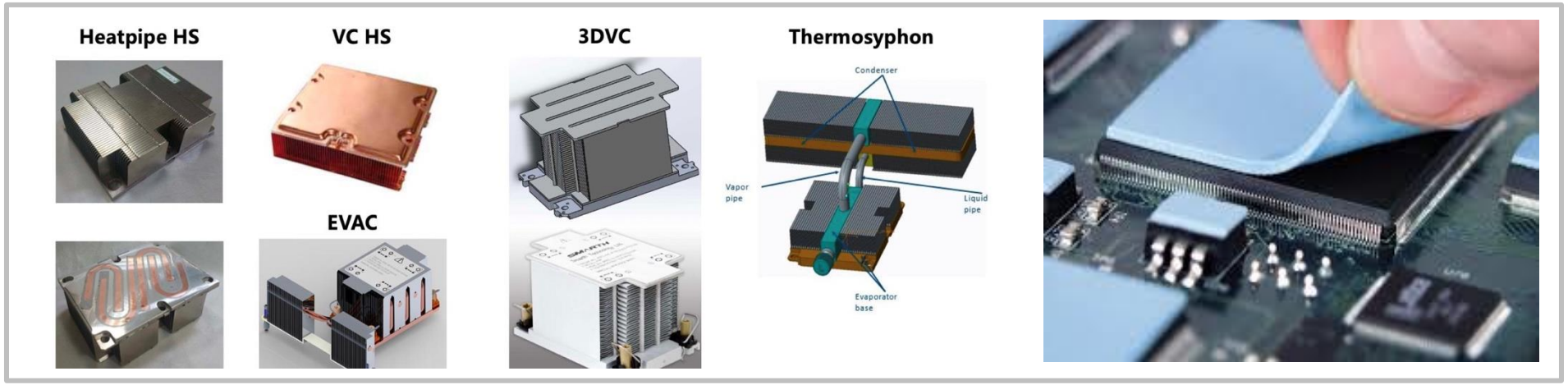
- **超高功耗需求：**
  - **高耗电：**10 万张 H100 GPU 集群 IT 功耗 >150MW，e.g. 相当于小型城市的峰值用电，年耗电量 ~1.59 太瓦时，电费成本超 1.2 亿美元/年。
  - **单机柜功率密度飙升：**NVIDIA 的 NVL72 机柜功耗达 120kW（a.k.a. 冗余后198kW），液冷成为强制选项

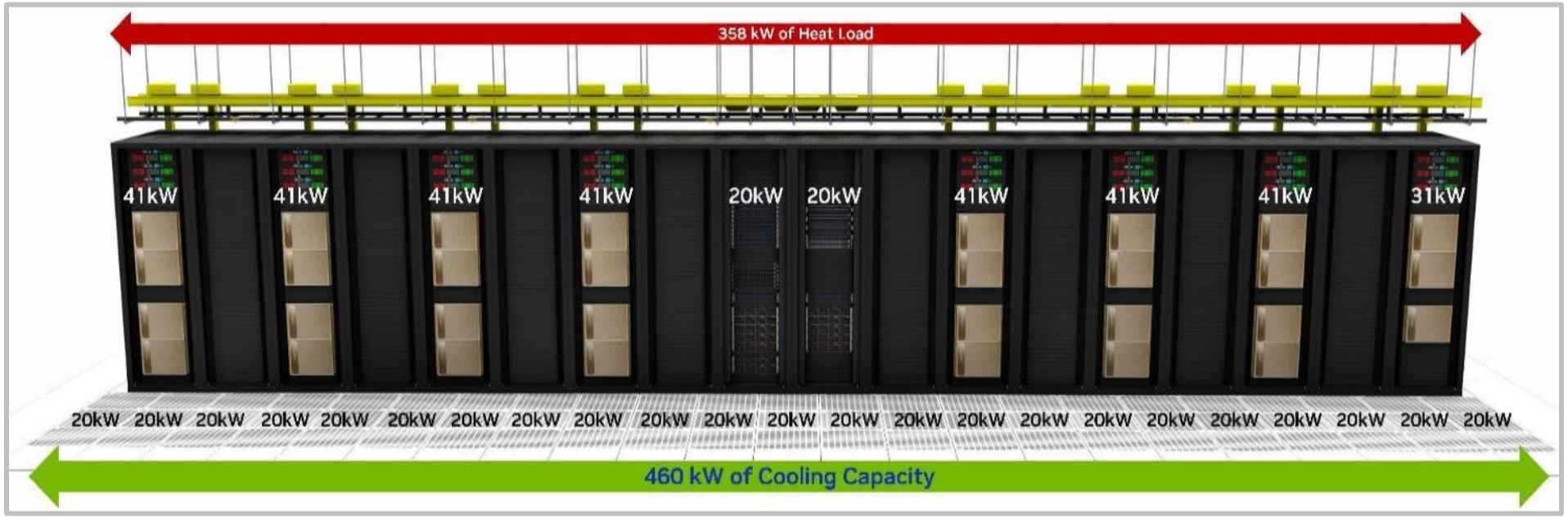
电力供应与散热挑战：能耗密度突破物理极限

- **供电架构重构：**
  - **供电方案：**传统 12V 供电链路损耗过高，48V 直连方案（e.g. OCP 标准）成为主流，但仍需解决高电流（>1000A/芯片）下的电压转换效率问题
  - **新架构：**Vicor 等厂商提出“垂直供电架构”（VPD），通过分比式电源（FPA）将电流倍增模块嵌入处理器下方，减少PCB传输损耗 95%

- **液冷协同设计**
  - **定制化：**高密度机柜需定制密封式液冷管道，防止冷却液渗漏导致电气短路
  - **协同优化：**液冷系统占数据中心总能耗 ~40%，需与配电网络协同优化

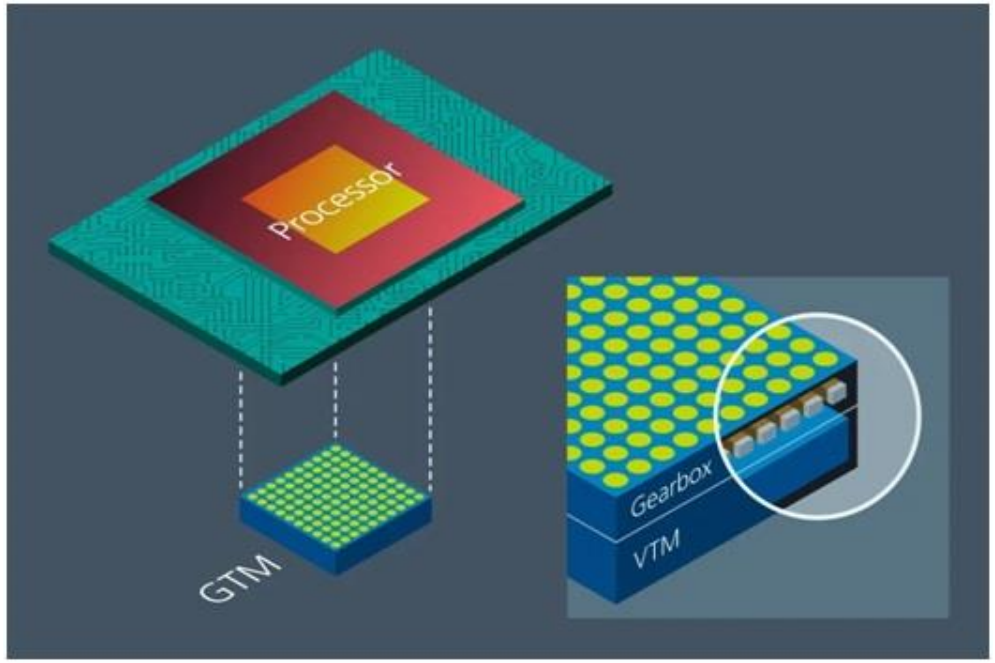
液冷、功耗、器件相互影响

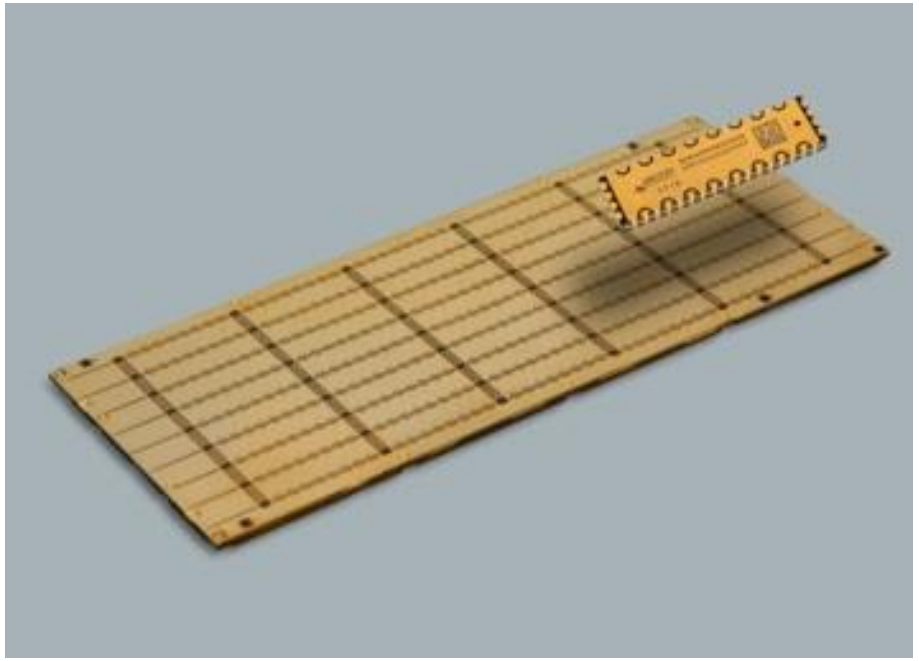

Heatpipe HS
VC HS
3DVC
Thermosyphon
EVAC
Condenser
Vapor pipe
Liquid pipe
Evaporator base


358 kW of Heat Load
41kW
41kW
41kW
41kW
20kW
20kW
41kW
41kW
41kW
31kW
20kW 20kW 20kW 20kW 20kW 20kW 20kW 20kW 20kW 20kW 20kW 20kW 20kW 20kW 20kW 20kW 20kW 20kW 20kW 20kW
460 kW of Cooling Capacity

新型的封装和电源提供方案

- 搭配电流倍增器 GTM 置于处理器下方，最大限度地提高电源传输性能。 Vicor 垂直电源传输VPD 包括更高 I/O 路由、板载内存等设计大大减少了外围供电和散热器件数量。
- SM-ChiP 封装将所有无源器件、磁性器件、MOSFET 和控制器集成到一个模块中，降低噪声改善散热性能。





02

网络通信

万卡互联的网络通信带宽与延迟瓶颈

- **万卡建设互联现状：**
  - **多层互联**：交换机 ~1000台，框框、盒盒盒多层互联结构，多层交换拓扑
  - **通信算法**：集合通信、点对点通信超规模后 Ring、 Have Doubling 回环时间长

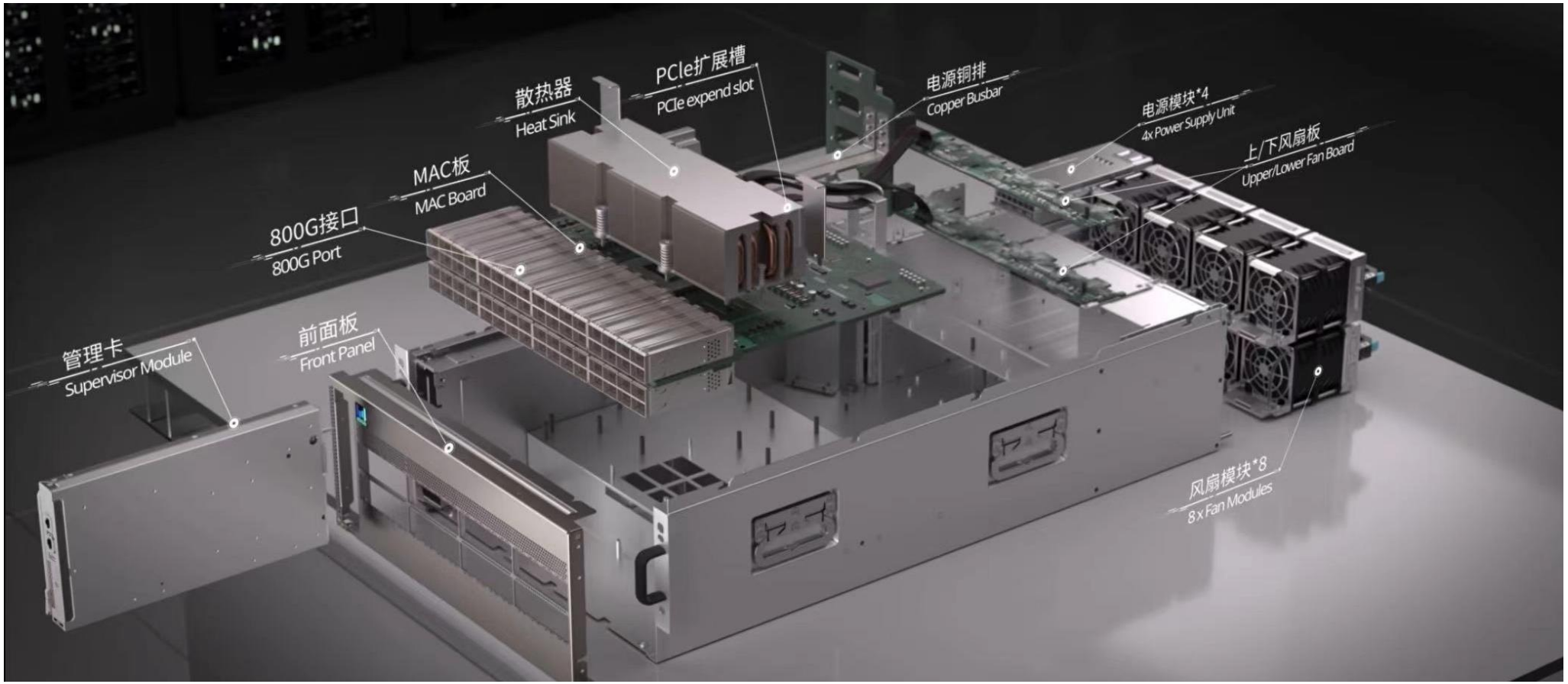
- **超大规模组网复杂度：**
  - **成本增加：**光模块成本激增，10 万 AI 集群需 ~10 万个光模块，长距单模式光模块成本是多模 10 倍
  - **InfiniBand等协议：**低延迟但扩展性差，超10万卡时需4层交换机，成本高昂

万卡互联的网络通信带宽与延迟瓶颈

- **万卡建设互联现状：**
  - **自研交换机：**字节采用三级交换机（e.g. Broadcom Tomahawk 4芯片），通过减少 ECMP 哈希冲突和动态拥塞控制（Swift +DCQCN 算法）提升吞吐量 ~30%
  - **引入以太网：**以太网交换机成本低，但需优化 All-Reduce 等集合通信效率，Meta 采用 7:1 超额订阅拓扑平衡带宽与成本
  - **NVLink 与网络协同：**TP 严重依赖 NVLink，但跨机柜 PP/DP 通信需融合 InfiniBand/RDMA，多网络协同易引发死锁，需要实现跨节点、跨柜与网络协同

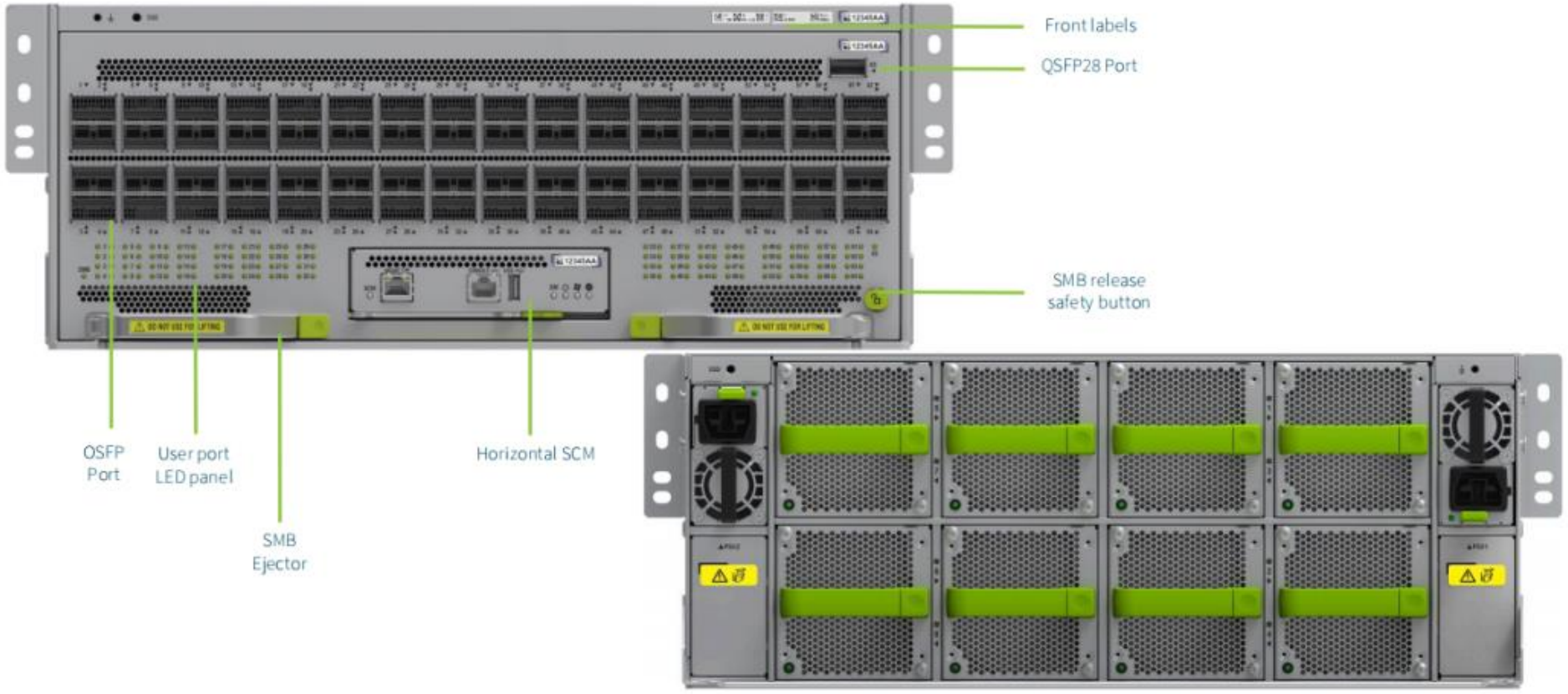
字节跳动自研交换机 B5020

- 4U 机架式交换机；64 x 800GbE 端口；51.2 Tbps 交换容量（全双工算法下达 102.4 Tbps）



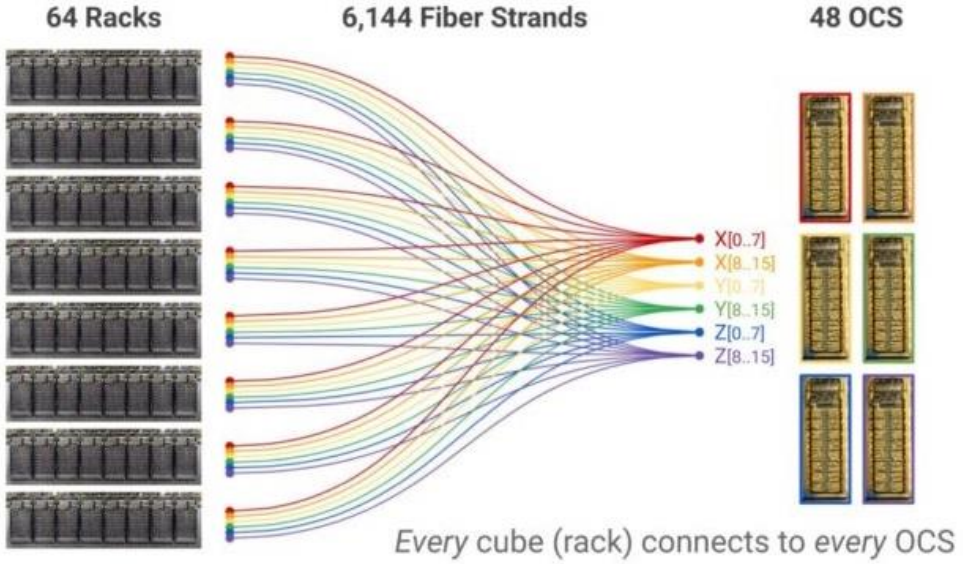
Mate 自研交换机 Meta 51.2T Ethernet Switch

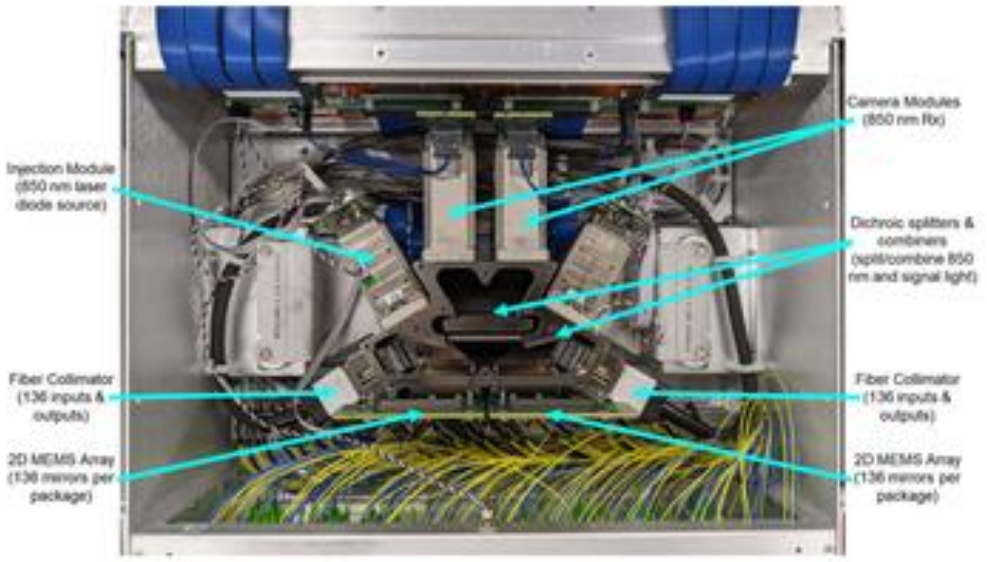
- 交换机单芯片带宽 51.2T，支持 800G 高速端口，满足 AIDC 对高带宽需求。
- 融合 CPO（共封装光学）技术，有助于降低能耗和成本，同时提高传输效率。



Google 自研 OCS 和 Tours 4D 拓扑

- **OCS (Optical Circuit Switch)：** MEMS 光路交换架构，通过动态重构光子连接实现纳秒级拓扑重组，支持分布式网络控制平面，突破传统电交换机的端口密度和能效比限制。
- **Torus 4D 拓扑：** 将三维 Clos 拓扑扩展为时间分片四维结构，通过 时空复用 TDM 在多跳层叠网络中实现物理链路统计复用，以硬件级调度优化多对多通信效率。





03

计算效率

计算效率提升，万卡规模互联算力损耗难题

- **万卡建设计算效率现状：**
  - AI 节点：>2000 节点，AI 卡数 ~16000
  - 存储容量：~800 ROCE 交换机
  - 大模型：PyTorch + Megatron + Dense/MoE 从十亿到千亿规模训练测试，万卡集群 K8S 调度

- **提升有效计算率 MFU：**
  - 千卡集群 MFU ~40-50%，但万卡因通信延迟和同步开销，特别是面向 MoE 稀疏模型 MFU 将至 35%
  - 混合专家模型（MoE）需All-to-All通信，加剧带宽压力；

计算效率提升，万卡规模互联算力损耗难题

- **分布式训练策略**
  - 字节跳动 MegaScale 通过 3D 并行优化和计算+通信重叠技术，在 1.2W 集群上 MFU 提升 ~55.2%
  - 摩尔线程“夸娥集群”采用自适应混合并行策略，支持显存池化管理，减少I/O瓶颈

DeepSeek 性能优化方案

- **GEMM FP8：**通过 FP8 通用矩阵乘法库，利用 Tensor Core 加速和 JIT 优化，同时通过 CUDA-core 二级累加技术缓解精度损失。
- **Dual Pipe：**双向流水线并行架构，将 Forward 与 Backward 计算过程重叠执行，并通过硬件级调度隐藏通信开销。

Method	Bubble	Parameter	Activation
1F1B	$(PP-1)(F+B)$	1×	PP
ZB1P	$(PP-1)(F+B-2W)$	1×	PP
DualPipe (Ours)	$(\frac{PP}{2}-1)(F\&B+B-3W)$	2×	$PP+1$

04

系统可靠性

系统可靠性

- **硬件故障常态化**
  - 10 万 GPU 集群日均故障 GPU 超 50 张，单卡故障可导致整个训练任务中断
  - 传统故障定位需 1-2 天，复杂故障达数十天，训练中断损失巨大

- **快速恢复机制**
  - 设计分钟级故障定位+断点续训方案，通过 Kubernetes 自动驱逐故障节点，秒级内切换备份节点
  - 通过全栈运行时打点技术，将故障恢复时间压缩至分钟级，提升训练有效率

05

国产化与成本运营

国产化挑战：生态与性能的双重差距

- **GPU供应受限：**
  - NV 高端芯片 H100/B200 对华禁运，国产 NPU 华为昇腾、寒武纪在互联带宽和显存容量上存在代差
  - 国产 AI 集群虽实现万卡组网，但 MFU 仍低于国际水平
- **软件生态割裂：**
  - CUDA 生态垄断大模型工具链，国产 NPU 需兼容层迁移，算子适配成本增加 5-10%

成本与运营：千亿级投资的商业风险

- **天量资本支出**
  - 10 万 H100 集群硬件成本 >40 亿美元，加上液冷和网络设备，总投资 >60 亿美元
  - 国产替代方案相同算力下成本并不低
- **利用率与空转风险**
  - 中国已建 250+ 智算中心，但多地出现算力空置，河南万卡集群因模型需求不足利用率仅 40%
  - 紫光集团于英涛呼吁避免盲目建设，需匹配实际产业需求

总结与思考

Question?

- 总结？干就完了！摸着石头过河，总结经验，下一个视频更精彩！

Thank you

把 AIInfra 带入每个开发者、每个家庭、每个组织，构建万物互联的智能世界

Bring AI Infra to every person, home and organization for a fully connected, intelligent world.

ZOMI

GitHub github.com/Infrasys-AI/AIInfra

Book infrasys-ai.github.io

引用与参考

1. http://www.ccw.com.cn/ai/2022-03-01/23727.html
2. https://cloud.tencent.cn/developer/article/1705673

PPT 开源在： https://github.com/Infrasys-AI/AIInfra