智算多芯——大模型时代的算力算法创新实践

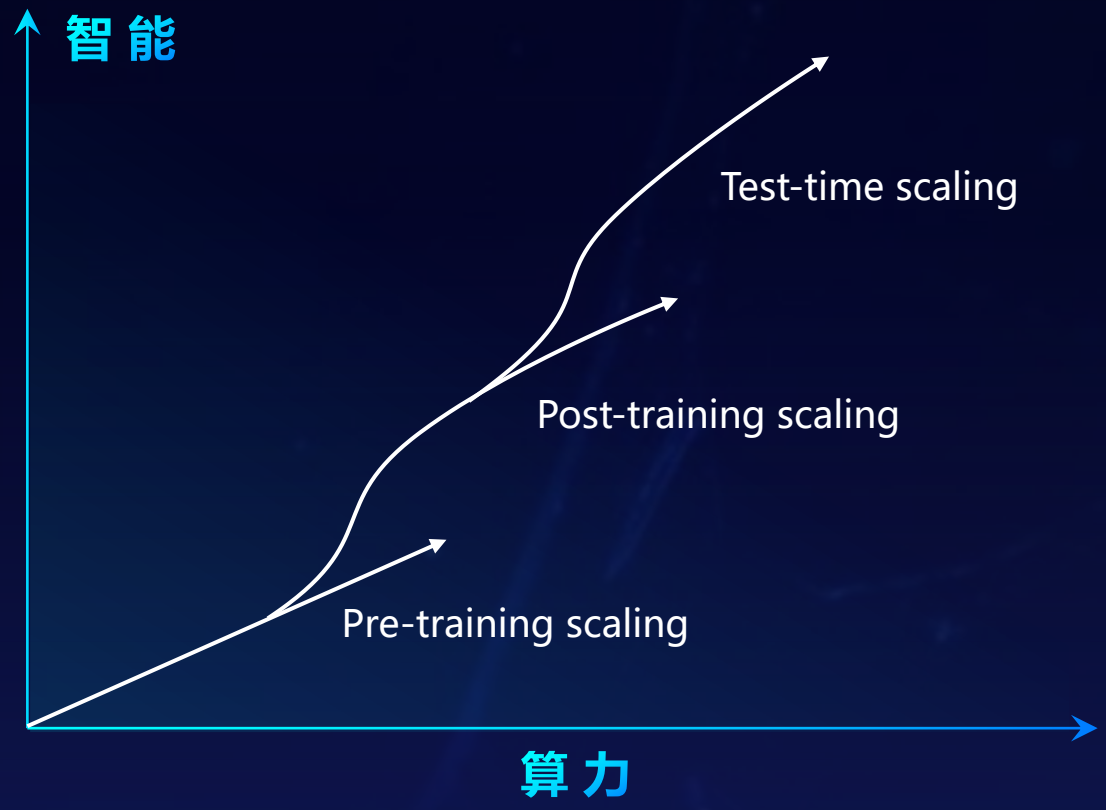
杜怀亮

大模型时代：Scaling Law驱动智能持续进化

从 pre-training 到 post-training 和 test-time scaling，scaling law 持续刷新模型能力上限

代表复杂思维的推理(Reasoning)能力的涌现，是构建智能体的基础

以 DeepSeek 为代表的模型开源极大降低了创新门槛，加速应用范式变革



智能体产业化的核心三要素



能力(Capability)

模型能力
决定了智能体的应用上限



速度(Speed)

交互速度
决定了智能体的商业价值



成本(Cost)

Token成本
决定了智能体的盈利能力

速度决定了智能体的商业价值

从人机交互到智能体交互的根本转变

- **人机交互**：满足20 tokens/s 的流畅阅读体验
- **智能体交互**：速度优势直接转化为商业价值

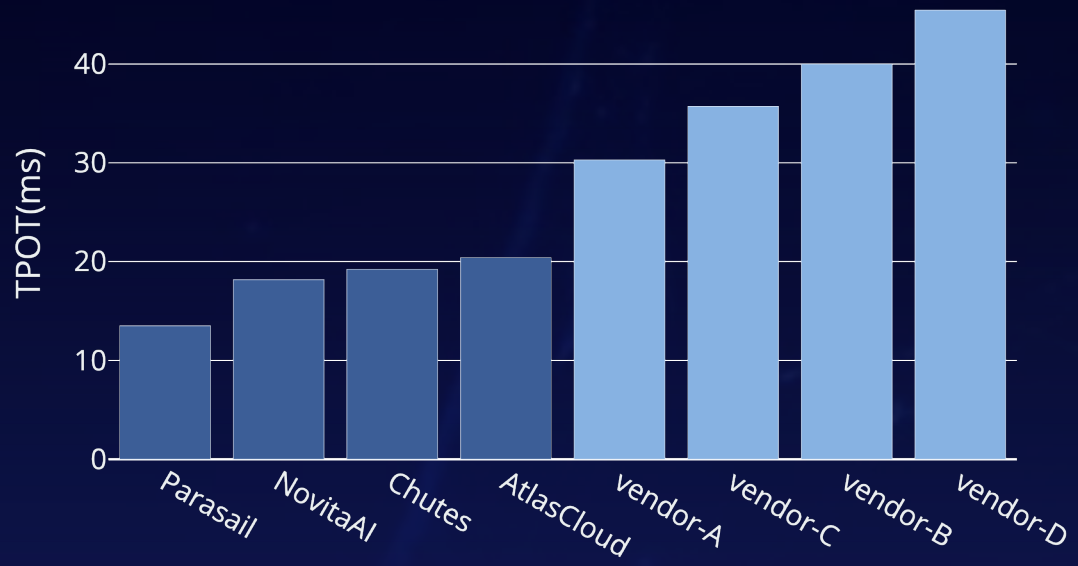
商业应用需要毫秒级响应

欺诈防控
~10 ms

具身智能
~50 ms

电子商务
~ 100 ms

全球大模型API服务商
DeepSeek Token生成速度（TPOT）



*数据来源：OpenRouter等开放API平台

元脑 SD200 超节点 AI 服务器

Token生成速度率先进入10ms 时代



8.9ms

元脑SD200 AI服务器 DeepSeek R1大模型
token生成速度(TPOT)

元脑 deepseek

基于创新研发的多主机低延迟内存语义通信架构，打造的以开放系统设计聚合64路本土AI芯片、可支持万亿参数大模型推理的超节点AI服务器

面向未来，创新AI计算架构，推动AI产业可持续发展

当前AI算力的可持续发展面临挑战

规模



电力



投入



发展AI专用计算架构：从规模导向转为效率导向，重新思考和设计AI计算系统

探索开发大模型芯片：算法硬件化的专用计算架构更适合推理场景，是未来发展方向

浪潮信息 HPC全栈解决方案

教科研数据中心全栈解决方案

应用场景

AI4S、HPC行业应用

气象海洋 | 生命科学 | 物理化学 | 材料科学 | 工业仿真 | 高能物理 | 石油勘探 | 动漫渲染 | 天文遥感

软件平台

高性能计算管理平台



集群管理 | 集群调度 | 作业提交 | 精细计费

Cluster Engine

基础设施管理平台



ISPIM

应用特征分析平台



TEYE

系统环境

操作系统 | 编译器 | 并行环境 | 数学库

硬件平台

CPU计算节点



通用服务器
NF5180/5280



高密服务器
i24、i48



大内存服务器
NF8260/8480、TS860

GPU计算节点



HGX机型
NF5488、NF5688



PCIe机型
NF5468

整机柜产品



整机柜服务器
ORS6000S

高速网络



InfiniBand网络

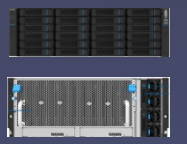


RoCE网络

并行文件存储



全闪
12/24盘位



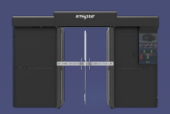
混闪
36/60盘位

基础设施

模块化数据中心



微模块数据中心（MDC）



液冷MDC



风液式解决方案

软件产品：AIStation

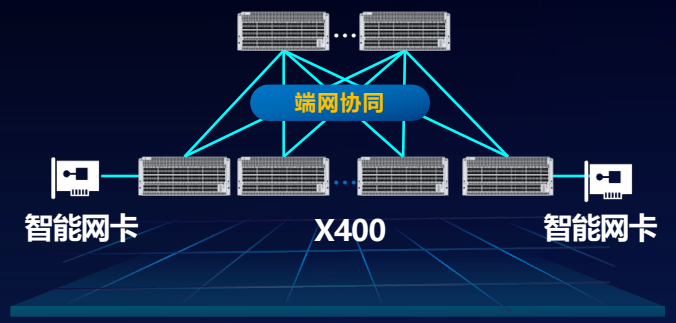
- **AIStation V5.2版本拟于4月30日发布**
  - 新增**大模型的分布式推理部署能力**，同步提供细粒度服务性能与质量监控。支持**DeepSeek全系列模型的单机和分布式部署**，快速完成DeepSeek 私有化部署
  - 支持全本土方案：**AIStation+HG3/4号+MXC550/C500/DCU+QLOS**
  - 增强数据处理、运营管理及用户角色管理能力：**增加图片数据标注，提供账户充值和消费管理能力，进一步增强用户权限的管理细粒度**
- **目标客户和客户价值：**
  - DeepSeek部署的客户：AIStation帮助客户在单机/多机上快速部署DeepSeek，保障服务稳定性
  - 泛AI行业客户：金融、能交、医疗等对本土化要求较高的客户，通过AIStation快速建立AI开发标准流程
- **产品策略：**
  - 加强本土化方案的推广，针对金融、能交、医疗等客户，联合硬件产品，整体方案交付客户
  - 持续挖掘利旧设备部署DeepSeek诉求的客户，推广AIStation分布式推理部署能力

发布元脑企智（EPAI）

- **元脑企智（EPAI）**
  - 丰富拓展对话、知识库功能，支持**多模态模型，进行基于复杂文本、图片、表格的理解与对话**
  - 更贴近用户业务数据，**支持主流数据库（mysql 、postgres、mariadb等）接入、分析与绘图**，帮助用户利用大模型**智能分析结构化数据**
  - 强化智能体能力，能够**自主规划智能体，支持基于客户业务流程快速构建与接入智能体服务**，帮助用户简便、快捷使用最新大模型
  - 新增数据处理、模型训练、模型微调、模型评估的**北向API**，方便外部系统的调用和集成
- **元脑企智+ISP 新增大模型应用支撑**
  - 面向政务、教育、金融、企业等客户，提供**公文处理、智能会议、智能客服、教学辅助、文献辅助阅读、金融研报生成、司法问询机器人**等智能体应用
  - 面向政府、央国企、金融等对大模型内容安全场景。提供**针对核心价值观、歧视性、违规违法、侵权、不准确的内容，提供输入和输出内容风险识别和处理能力**，满足《生成式人工智能服务安全基本要求》
- **产品策略：**
  - 针对新增的大模型应用，针对性在政务、教育、金融、企业客户推广，挖掘应用商机

端网协同突破以太网性能极限 充分释放算力潜能

X400 AI Fabric方案



万卡集群
支持**十万卡级**

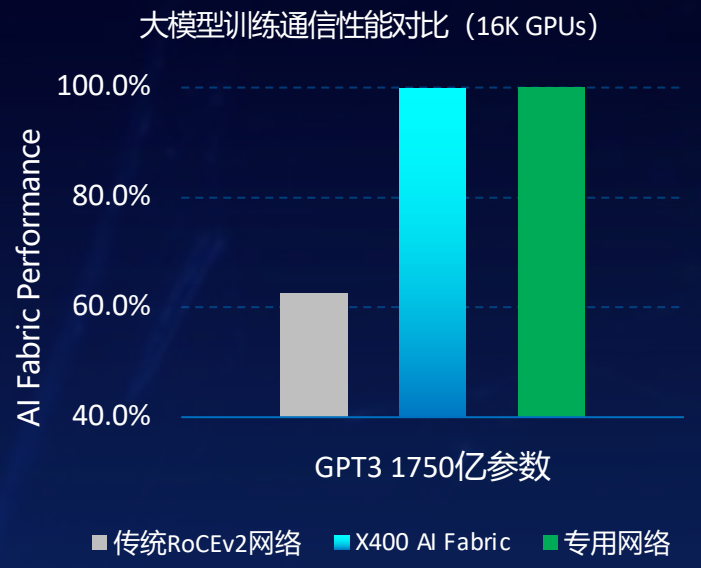
传输0丢包
端到端**400G**

有效带宽
高达**95%**

超级AI以太=X400+智能网卡



性能领先传统RoCE 1.6倍



传统数据中心向智算中心的演进

单机柜功率以10kW以下为主

受限于通用服务器功率及制冷技术限制

单机柜功率以20kW以上为主

受到单台服务器高达10kW功率、组网距离及布线成本限制

制冷形式以风冷为主

单柜制冷能力一般低于10kW，常规采用列间空调+封闭冷/热通道的形式

制冷形式向高密风冷/液冷转变

高密风冷支持单柜50kW+

液冷支持单柜100kW+

先机房后IT，缓慢建设

传统模式先按照机架功率确定机房规划，后进行IT设备选型招采

先IT后机房，快速上线

智算中心先确定IT设备选型，同步进行机房规划

机房

双排微模块





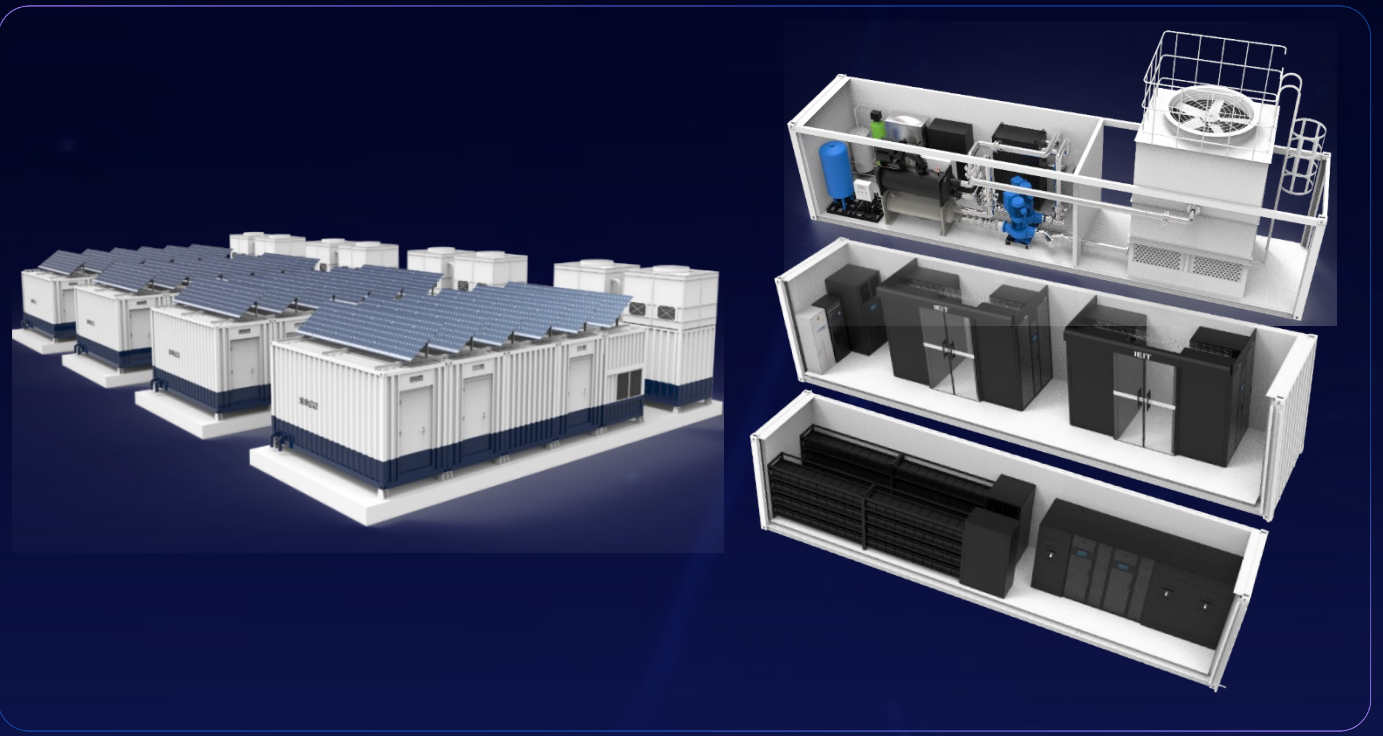
智算算力仓

智算中心特点及模式规划

- 智算中心需求总结：单机柜功率密度高50kW以上，紧抓时机因地制宜，设备价值高安全性要求高
- 智算中心核心观点：预制化+高密供电+安全液冷



高密模块化算力仓解决方案



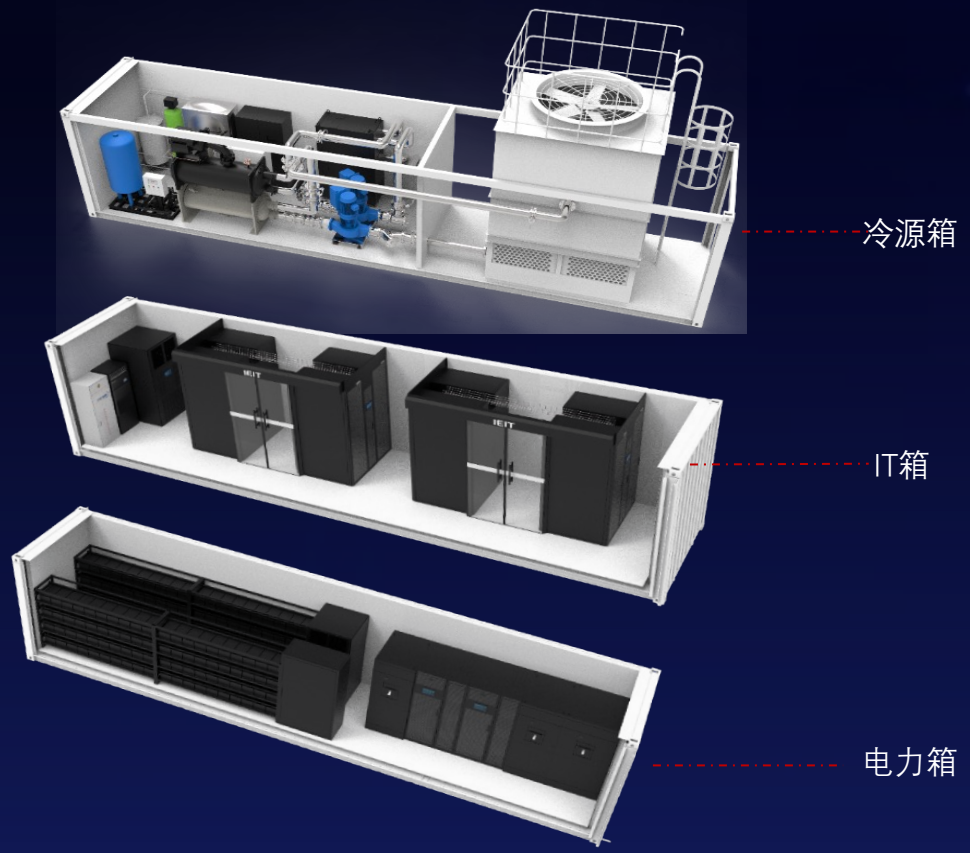
高密预制化数据中心解决方案

预制化液冷智算算力仓

预制化液冷智算算力仓

预制化液冷智算算力仓，以国际标准 40 尺集装箱为载体，由 IT 箱、电力箱、冷源箱组成， 为液冷整机柜提供高效基础设施支撑，包括 4 个 120kW液冷整机柜，满配算力柜+1 个网络柜，满足 510kW 功率密度部署，具备高度集成、安全可靠、风液同源、高密快速交付等特 点，适用于 AI 计算等多种计算场景 。

产品图片



产品优势

高度集成

- 采用符合国际标准 40 尺集装箱为载体，高效 集成液冷整机柜、网络、配电、CDU、空调、 集成冷源、监控、消防等系统
- 冷源机组内置式水力模块，节省占地

风液同源

- 风冷、液冷散热采用同一套冷源系统，实现风 液同源
- 先进 BA 控制系统，结合变频水泵和变频风机 等节能技术为 CDU 和列间空调精确供冷

安全可靠

- 隔热、防火、防水、防锈、防盐雾设计，配电、 空调系统冗余设计，保障机组可靠运行
- 采用智能动环监控系统，全面监测集装箱内的环境和设备参数

快速交付

- 满足 40 尺标箱内部署 510kW 设备功率密 度，支持 4 台 NVL72 满配部署 □ 快速交付，最大程度去工程化安装
- 客户只 需提前预置场地，2 周完成 4 箱部署

建设案例：10MW预制化算力中心



项目概况：

总建筑面积3980㎡，采用119台集装箱，规划419架IT机柜，机柜功率10~30kW，按照使用功能进行分区，划分为数据机房区、液冷测试区、外网测试区、配电区、监控室、展厅及其他辅助区域；建筑北侧配套集成冷水机组并预留冷机扩展区域空间。

该项目仅历时9个月即完成设计施工交付，即将投入使用。





建设案例： 10MW预制化算力中心

项目点：

定制化

- 匹配智算算力、通用算力、边缘算力应用场景
- 提供低密、中密、高密集群规划
- 支持风液混合

快速部署

- 土建与工厂预制并行，缩短60%建设周期
- 现场去工程化交付，提升**50%**交付效率
- 支持横向、纵向按需快速扩容

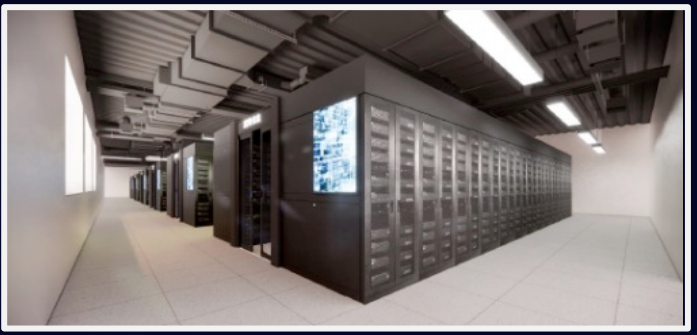
绿色高密

- 光伏+储能+液冷+自然冷源，PUE低至1.15以下，降低碳排放25%
- 风冷单柜功率40kW，液冷单柜功率100kW

集约高效

- 10kV简链路供电，冷源预集成，节省占地**40%**
- 系统智能调优，高效运行，能效提升**20%**







土建施工



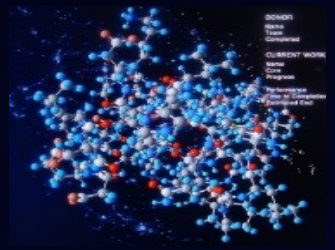
钢结构生产





应用创新推动科研发展

应用创新推动科研发展



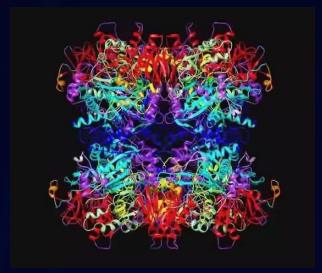
材料计算
复旦大学材料量化仿真应用



天文计算
华中科大引力波计算和数值模拟平台



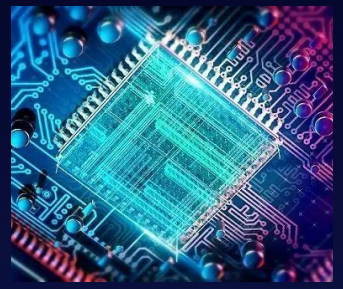
生命科学
浙江大学药物设计、分子摸拟应用



冷冻电镜
清华大学结构生物研究平台



粒子物理
华中师大格点量子色动力学研究



半导体设计
复旦大学EDA仿真应用



计算化学
南京大学量子化学、分子动力学与模拟应用



高通量材料
南科大高通量材料计算与分析平台


谢谢！