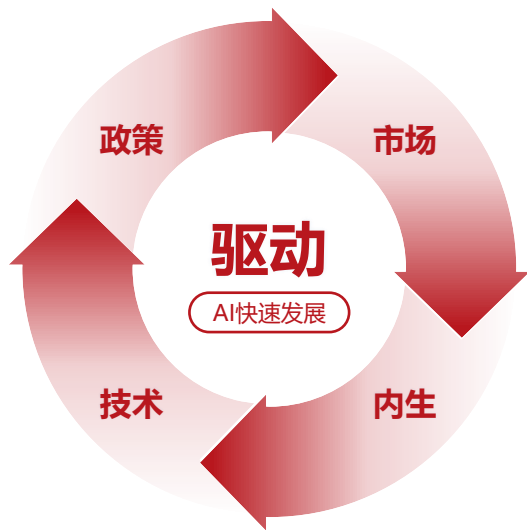


算力领航 打造敏捷化Token工厂

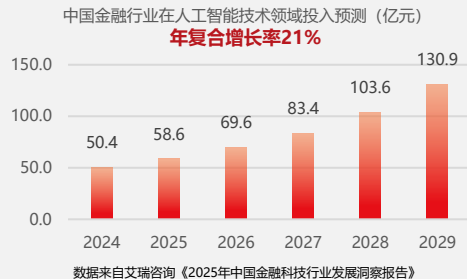
2026/08

精耕务实 为时代赋智慧

金融行业人工智能发展趋势

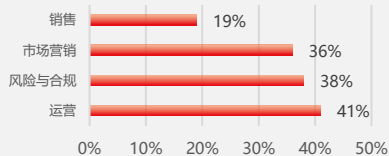


金融机构在人工智能领域的投入不断增长



人工智能在金融行业的应用领域

贵公司目前在以下哪些领域使用了AI 解决方案? (中国)



金融人工智能应用发展特点

由于金融客户目前主要是内部使用AI应用,以推理场景为主不涉及预训练,所以算力总体规模相对有限



人工智能给金融机构带来的价值

AI 如何提升了贵公司业务运营 (中国)

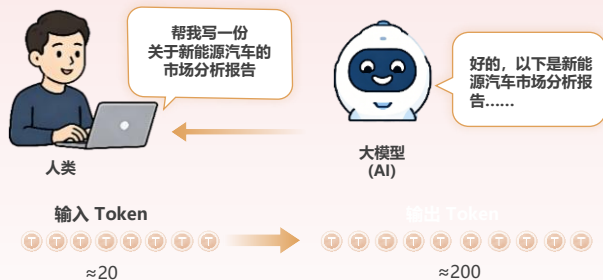
提高运营效率	68%
改善客户体验	44%
降低总持有成本	24%
生成更准确的模型	16%
创造竞争优势	13%
开辟新商机	7%

信息来自 英伟达《2024年金融服务业中国AI现状与趋势》

Agent 应用导致Token消耗大幅增加

传统模式: 单轮提问: 简单、高效、可控

人类“偷懒”提问, AI 一次性回答



Token 消耗特点:

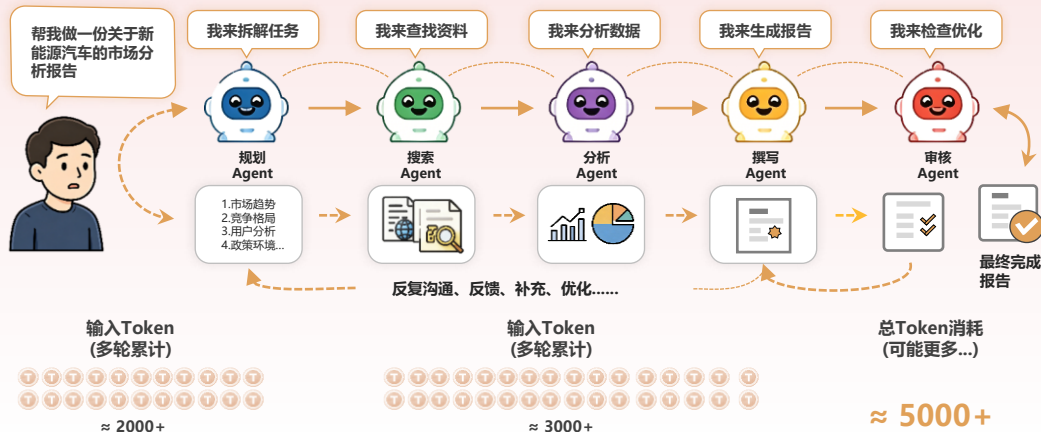
- 一次输入, 一次输出
- Token 线性、可预测
- 成本低, 效率高

总 Token 消耗:

≈ 220

Agent 时代: 多轮协作: 复杂、放大、难以控制

多个Agent分工协作, 反复交流, 逐步完成任务



Token放大的三大原因

① 交互链放大

任务被拆解, Agent之间多轮对话, 链条越长Token越多



② 上下文膨胀

每一轮都要带上历史信息, 上下文不断累积, 输入Token越来越大



③ 分支探索

多种方案尝试、反复验证、多路径并行导致Token组合式增长



本质变化



原来: 人类大脑思考(隐式)
现在: 系统展开思考(显式)
→ 将思考成本转化为Token成本



智算平台面临的挑战



多角色使用复杂 从员工到智能体

权限、场景、工具繁多
协同调用复杂



成本与性能失衡 从 FLOPS 到 Token

模型推理成本高
资源利用率低



安全合规要求高 从账号到 Token 权限

数据敏感、合规严格
需多层审计与控制



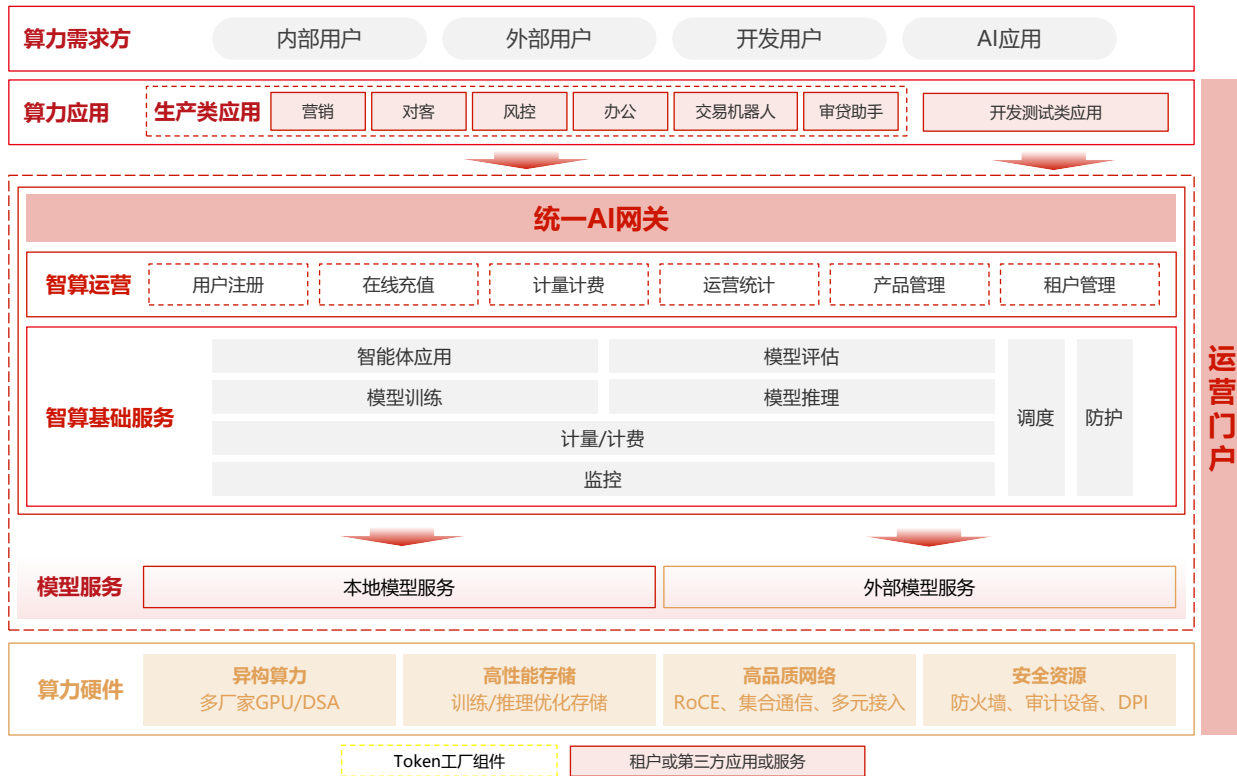
系统碎片化严重 从工具到平台

异构GPU、多模型、多工具孤岛
缺乏统一管理治理



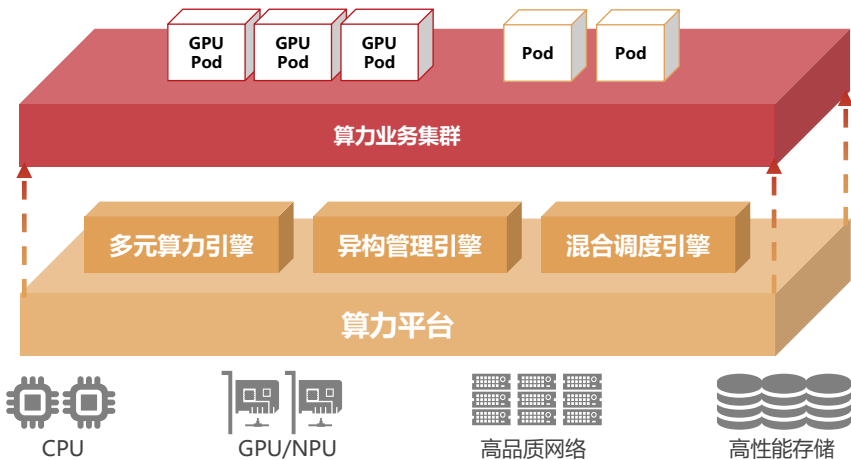
需求：安全、稳定、可控、可扩展的企业级AI 运营平台

图灵Token工厂架构



异构资源统一纳管及vGPU调度

多元异构算力统一管理，协同配合，打造灵活高效的算力平台，满足多样化算力任务需求



- 支持大规模集群算力统一调度
- 支持多元异构CPU和GPU的统一管理和调度



- CPU已验证支持X86\ARM架构
- GPU验证支持多家厂商**30+**款加速卡款型



- 自研GPU通用适配框架，一款加速卡适配周期缩短至**1周**内



- 支持NVIDIA卡物理切分、MIG、vGPU，算力和显存可按**1%**和**MB**的细粒度切分

多维数据统计



Token 精细运营

输入输出 Token、请求速率、
响应时延可量化分析



任务与服务调度

推理服务、训练任务、
开发环境运行状态清晰可视



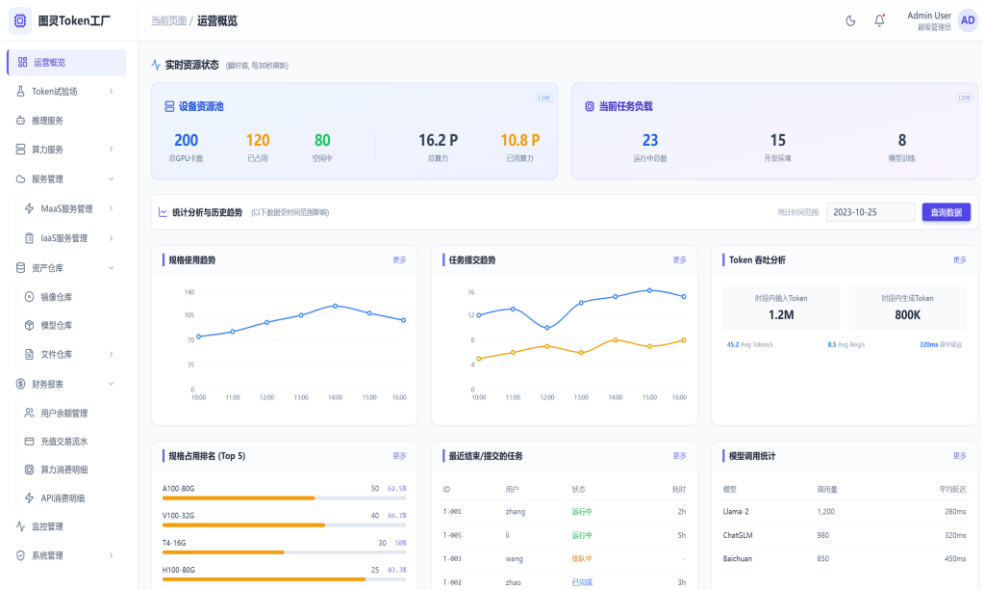
统一运营总览

设备资源池、任务负载、趋势看板一屏掌握



资产与费用治理

模型仓库、镜像仓库、文件仓库及消费明细统一管理



可视运营

全景看板, 实时洞察, 直观掌控

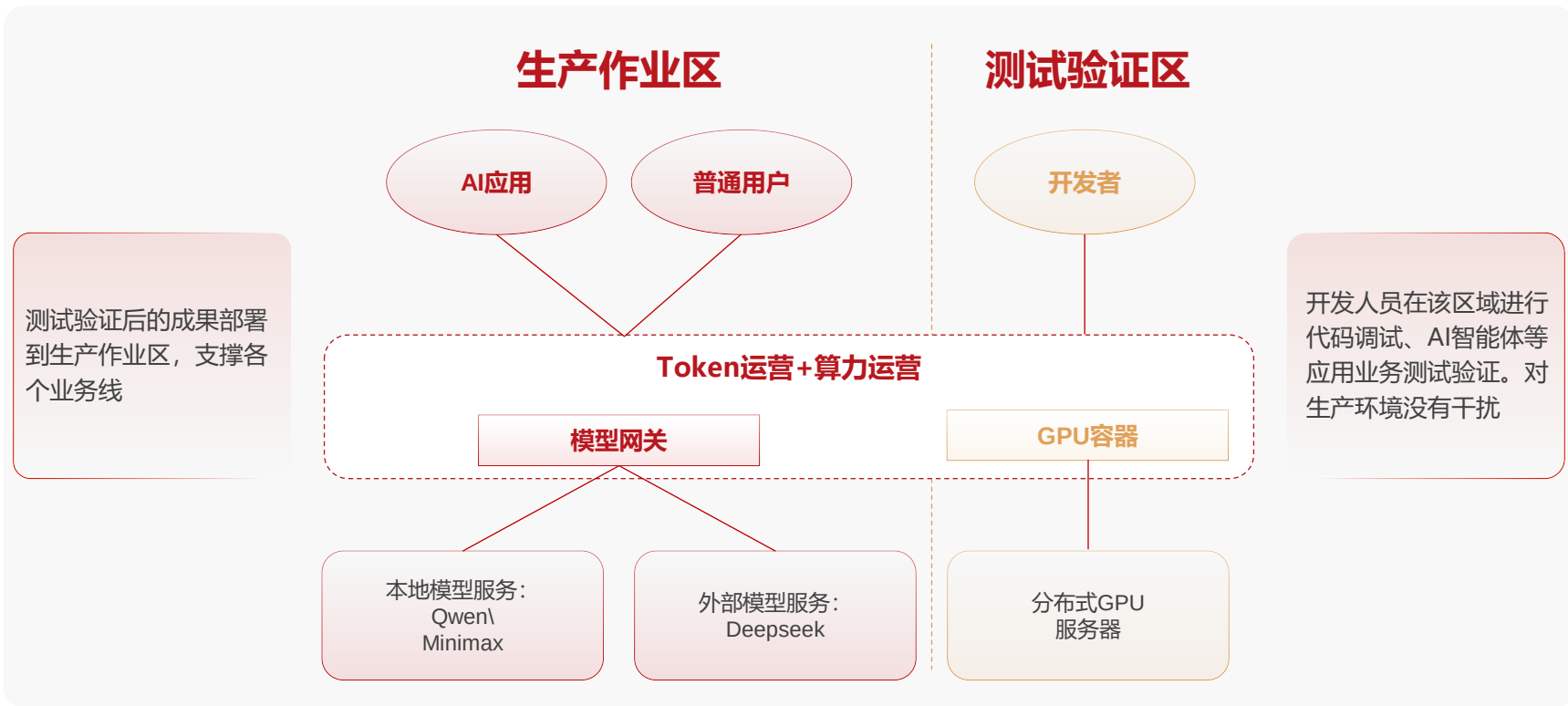
精细运营

多维度量化分析, 支撑精细决策

降本增效

优化资源利用, 降低成本, 提升效能

建设模式参考



硬件决定下限 调优提升上限



核心调优方向

模型层

模型权重压缩、量化，降低显存占用，提升推理速度

框架层

框架选型、参数调整（例：Prefix Cache）、流程优化（例：异步策略）、前沿特性适配，以提高显存利用及训/推效率

算子库 集合通信库

算子开发、算子融合、流程优化、前沿特性适配，以提高带宽利用、训/推效率。

智算中心全域资源监控

100+类监控对象，300+项监控指标数据，监控智算中心全域资源状态，提升从算力、应用系统到基础资源的全面运维感知力



从AI辅助到AI原生，金融智算筑基



通过安全可靠测评的芯片名单

通过安全可靠测评的芯片名单以及对应加速卡名称

附表一：人工智能训练推理芯片（同一等级按产品名称首字笔画为序排列）

序号	产品名称	送测单位	安全可靠等级
1	昇腾310	深圳市海思半导体有限公司	I级
2	昇腾910	深圳市海思半导体有限公司	I级
3	真武M530	平头哥（上海）半导体技术有限公司	I级
4	真武M890	平头哥（上海）半导体技术有限公司	I级
5	壁砺™ 166	上海壁仞科技股份有限公司	I级
6	DCU-3G	海光信息技术股份有限公司	I级
7	KCC-V100X 芯片	上海天数智芯半导体股份有限公司	I级
8	MXC600	沐曦集成电路（上海）股份有限公司	I级
9	PH100	摩尔线程智能科技（北京）股份有限公司	I级

安全可靠测评的芯片名单信息来自中国信息安全测评中心官网
https://www.itsec.gov.cn/aqkkcp/cpgg/202605/t20260526_259396.html

大模型需要的算力集群 需要参数网络



PD分离效果更好

2-5倍TPS性能提升
50%TTFT降低

模型并行与数据并行分离是提升效率的关键。
在相同硬件下，可实现吞吐量提升2-5倍，
首次令牌生成时间降低50%以上，
有效解决资源争抢和延迟抖动问题。



参数量持续增长

Kimi K3达到2800B
Qwen3.8达到2400B
DeepSeek V4 PRO1600B

国产大模型参数量飞速增长，如DeepSeek-V3.2
达6710亿，Kimi K2.5突破万亿。模型规模的急
剧扩大直接带来了算力需求的指数级增长。



高并发超长上下文

单台服务器显存明显不足

单卡算力已无法满足需求。
必须采用大规模GPU集群进行分布式部署，
大参数模型的超长上下文和多并发的访问需求。

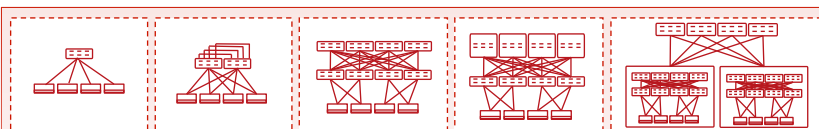
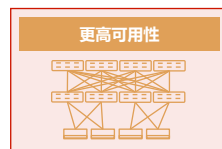


核心洞察：

单机算力已无法满足需求。必须采用大规模GPU集群进行分布式部署，以提供足够的算力支持，确保模型高效稳定运行。
PD分离技术提升了单机利用率上限，但模型规模的爆发式增长使得单机性能成为瓶颈，推动行业向集群化算力基础设施演进。

高效智算网络

智算底座需要高效的智算网络



百卡 → **十万卡**

新华三可提供高密800G/400G框/盒交换机，低转发时延大幅领先

S9820-8C-G 32*400G S9825-8C-G 64*400G

S9855-24B16DH-G 24*200G+16*400G

+

S12500CR S12500G-AF 48*100G

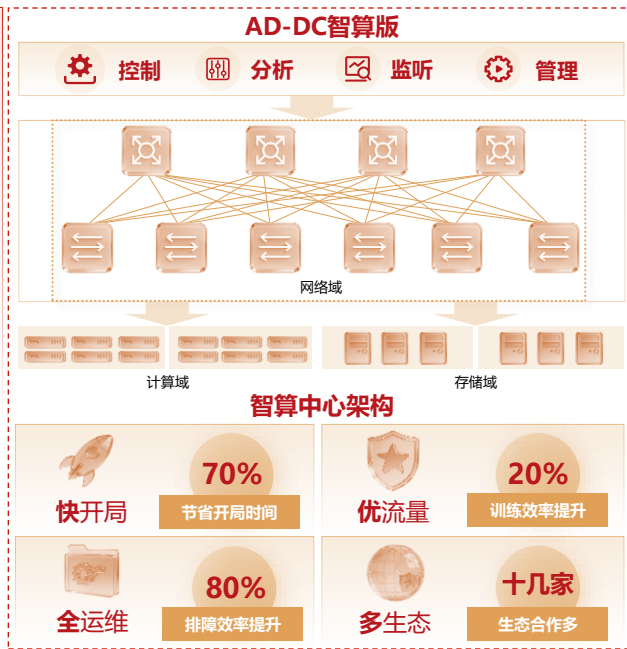
36*400G 36*200G

性能领先

- 国芯25.6T最高性能智算产品，预留800G/1.6T端口升级能力
- 交换芯片/CPU/内存等全部国产化
- 支持全面智算网络功能，对标商用芯片最高水准
- 兼容冷板液冷设计、支持LPO模块

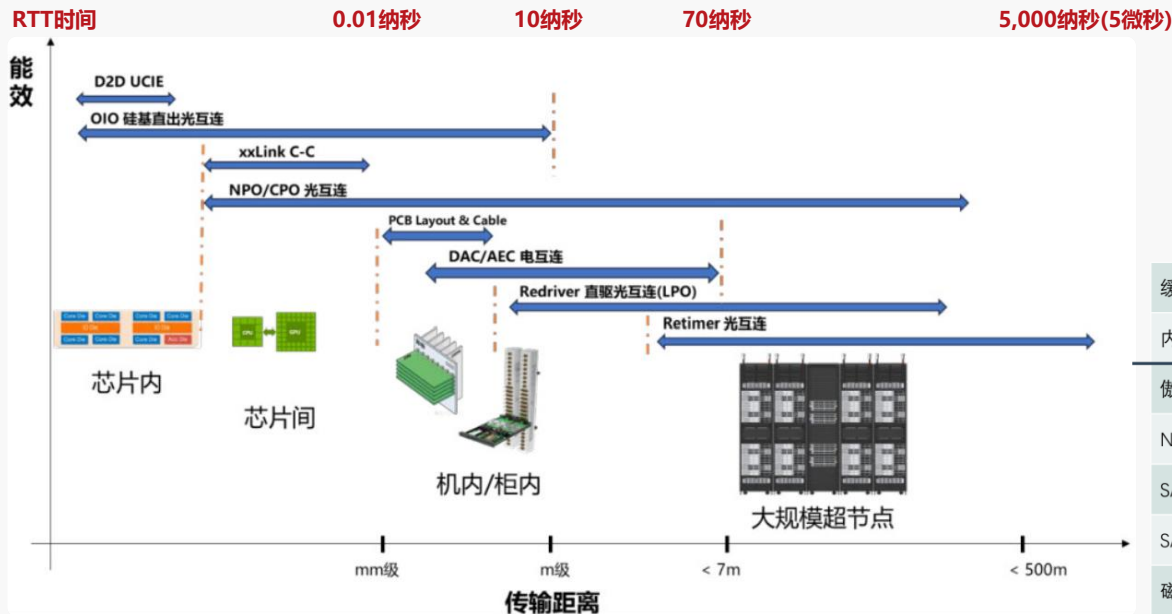
开放兼容

- 支持对接国产化网卡
- 兼容多品牌国产GPU卡
- 支持多种CPU智算平台



超节点是越大越好吗？

信号传输距离与能效，对互联技术路线之间的核心制约关系



主流存储介质访问延时对照表

缓存: L2/L3	易失存储	2-2x ns
内存/显存		< 100 ns
傲腾SSD	非易失存储	3.5-5 μ s
NVMe 卡/盘		10-30 μ s
SAS/SATA SSD		3x-1xx μ s
SAS/SATA磁盘		x ms
磁带		X min

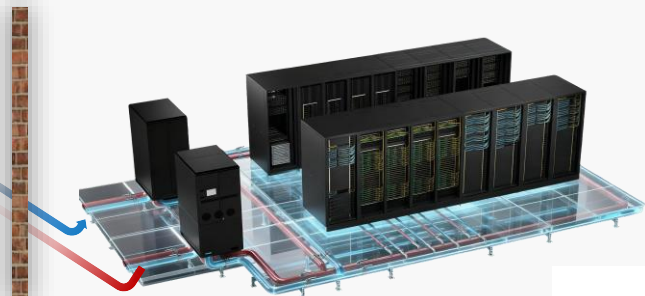
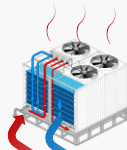
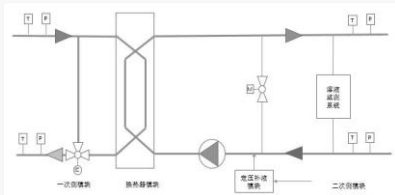
注：光走1米需要5纳秒的时间/RTT时延10纳秒；光走200米要用1000纳秒（1微秒）/RTT时延2000纳秒（2微秒）

超节点机房建设方案

- 冷却液分配单元（CDU）是液冷系统的核心组件，负责**均匀分配并调节冷却液的流量与温度，确保系统稳定高效运行**。它通过与泵、板换散热器等部件协同工作，并具备过滤杂质功能，防止系统堵塞与损坏。
- 风液式CDU适用于边缘计算节点、小型数据中心或无法进行液冷改造等场景，**无需依赖冷冻水系统**，通过风冷完成最终散热，具有**部署灵活、结构紧凑**的特点。随着液冷技术向边缘侧扩展，风液式CDU作为“冷却孤岛”的关键设备，应用日益广泛。

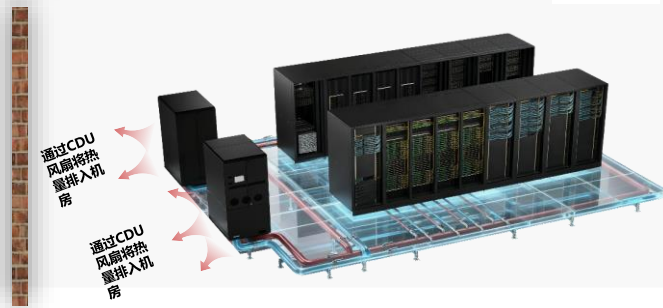
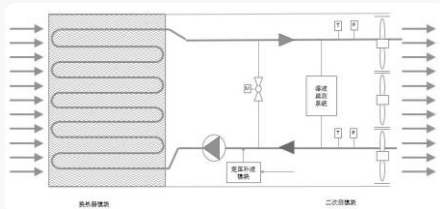
液液式换热CDU

加热后的二次侧冷却液进入CDU内，
经一次侧循环将热量带出机房
真正意义上的液冷散热



气液式换热CDU

加热后的二次侧冷却液进入CDU内经
风冷降温，然后将热量排到机房内
免去了一次侧的施工周期及设备成本问题



新华三金融Token工厂：打造六位一体全栈能力体系

AI Infra全栈深度整合，算网安存管维一体化交付；运维管理和业务调度全栈感知，软硬件协同优化性能提升

金融应用

营销

对客

风控

办公

交易机器人

信贷助手

.....

AI业务服务平台

智能体应用

模型训练

模型评估

模型推理

数据工程服务

AI 基础服务

数字资产管理

管控运营运维平台

资源管理平台

大数据平台

Token
工厂

基础设施

多元异构算力

超节点

GPU
服务器

通用
服务器

高性能存储

下一代存储

X2000
存储

高品质RoCE网络

商芯

国芯

框框/框盒/盒盒

端到端安全

智能体安全

数据安全
模型安全

运

运营、运维管理：多租户、全域状态运维监控，全面运维感知，全域故障生命周期管理

管

系统管理和调度：多元算力整合，实现算力资源统一调度与运营管理，网关分发实现算力均衡

算

高密算力底座升级：算力底座超节点全面升级，多元架构，快速适配主流大模型

网

网络通信性能优化：Scale-Up和Scale-Out双层互联架构，算网协同，集合通信优化提升通信性能，异构兼容

存

KV Cache效率提升：打造全新重构的分布式存储；超节点KV cache卸载到外部存储，对KV Cache框架和存储的性能优化

安

智能体安全：AI流量识别和AI安全网关对流量进行分析和策略管控

算力 x 联接 释放金融人工智能无限潜能

—— 感谢观看 ——

新华三集团
www.h3c.com