Open Scaling for Superior Tokens

打造领先Agent基础设施 加速产业AI转型

赵 帅 浪潮信息副总经理

Agentic AI时代 智能体深入千行百业

OpenClaw 破局问世
自主执行能力的智能体迎来规模化爆发拐点

2025年中国 AI Agent 企业级市场190亿人民币，2025-2028 CAGR 超110%

—— IDC, 2025

2026年将有40%企业应用集成 AI Agent（2025年不足5%）；2028年 33% 企业应用嵌入 Agentic AI

—— Gartner

大规模智能体产业化正在加速落地，激活全域产业生态

从单 Agent 到多 Agent 协同

实现跨领域协作，多 Agent 完成端到端复杂工作流，实现复杂场景的高保真模拟能力，多智能体系统将成为企业标配

具身智能与边缘智能体兴起

智能体进入机器人、汽车、制造等物理世界，轻量化 Agent 部署边缘设备，满足低延迟需求

智能体大规模应用成为行业趋势
各行各业均基于 AI Agent 推进行业转型

（2028年）
1700亿+
IDC：中国 AI Agent 企业级市场2028年总体规模

>40%
GitHub 2026：全球新增代码中超 40% 由AI辅助生成

1600万+
中国一人公司数量2025年同比增长 47%

制造业
智能工厂 · 氧化铝产线

工艺优化 Agent + 无人巡检 Agent

人工经验调参
↓
AI 实时优化工艺参数

+60%	-3.6%	90%
生成效率	能耗	自动化覆盖率

从“人盯设备”到“系统自驱动”

企业研发
AI“代码工厂” · 全栈研发体系

架构设计 Agent+ 代码生成 Agent +自动化测试 Agent

人工编码 + 周级交付
↓
AI 自主编码/审校/部署，小时级迭代

-70%	92%	85%+
开发周期	BUG 检出率	代码复用率

工程师指挥一支 Agent 编队

OPC 一人公司
超级个体 · AI 咨询工作室

多 Agent 虚拟团队 · 7x24h 自动运转

6人团队 + 80万 外包方案
↓
1人+AI 工具，数小时交付

<$500	实时	分钟级
启动成本	客户响应	图像审核

一个人，就是一家公司

AI Agent 实现社会生产效率跃升

海量Agent承载 需构建领先的CPU Agent底座

海量 Agent 承载

高密度并发

单机承载数百 Agent 实例
集群支撑万级并发运行

沙箱隔离

容器级隔离，为每个
Agent 提供独立安全环境

弹性调度

按负载自动扩缩容
算力资源利用率最大化

AI Coding 场景

- 需求拆解 Agent
- 代码生成 Agent
- 代码测试 Agent
- 文档生成 Agent

...

办公场景

- 会议纪要 Agent
- 邮件助理 Agent
- 文档撰写 Agent
- 日程规划 Agent

企业级 CPU Agent 运行底座

高主频多核 CPU，强劲并发处理 | 大容量内存
支撑海量实例 | 灵活扩展，随业务弹性生长

企业级 Agent 保障

高可用架构

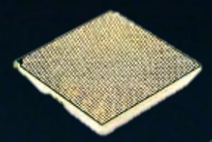
集群冗余、故障自愈
任务 7×24 不中断

安全合规

数据本地化处理
全链路权限管控与审计

“极致密度”MW机柜击破风冷极限 AIDC步入液冷原生

2kW 级加速芯片



- 更多计算核心
- 更高计算主频
- 更大片上显存

MW 级 AI 机柜

- 更大互联带宽 3.2T/s
- 更低互联延迟 ＜200ns
- 超低传输误码率 ＜10^{-15}
- 更高集成密度 500+芯

液冷原生 AIDC

- 更高效的散热能力
- 更低的生命周期运营成本
- 更高的算力性能输出
- 极致的Token/W压榨

OPEN Compute Project®

开放计算标准工作委员会 Open Compute Technology Committee

浪潮信息

面向群智协同 原生液冷CPU Agent主机重磅发布

以高密液冷底座支撑多元算力，释放海量Agent协作效能

开放算力模组

- 开源开放液冷 OCM 标准
- **x86、ARM多元算力兼容**

极致算力密度

- 最大支持 384*CPU，MW 级功耗设计
- 单机承载 **4万+ Agent**

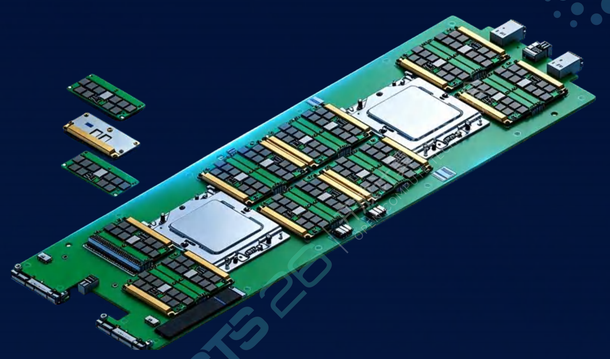
创新原生液冷设计

- 液冷原生设计，**部件级热维护**
- 一体化冷板设计，**100% 全域液冷**

原生液冷CPU Agent主机创新路径：开放、极致、重构

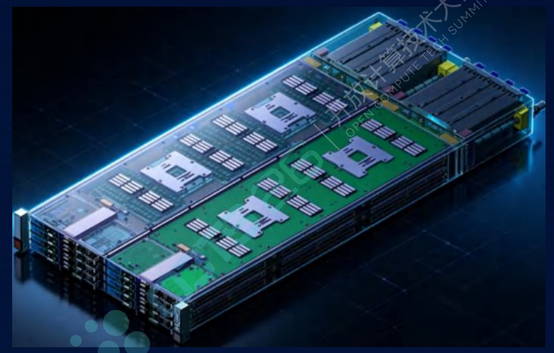
开放算力模组

（液冷 OCM2.0）



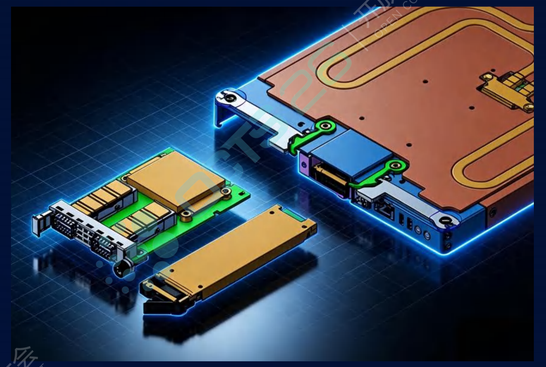
- 原生液冷标准化模组承载多元算力，实现**多形态 Agent 主机**快速衍生
- 液冷原生重构内存模组，多阶导热架构提高内存散热效率**30%+**

极致密度设计



- 一体化冷板基座，构建 ½ U 超薄算力节点，构建 **2U16 CPU** 高密通算系统
- **Cable-Less**无线缆设计，整机柜运维效率提升**150%+**

全域部件液冷重构



- SSD / NIC / 光模块接触导热式设计，全生命周期**带液热维护，业务零中断**
- 一体化冷板设计实现主板热源全域覆盖

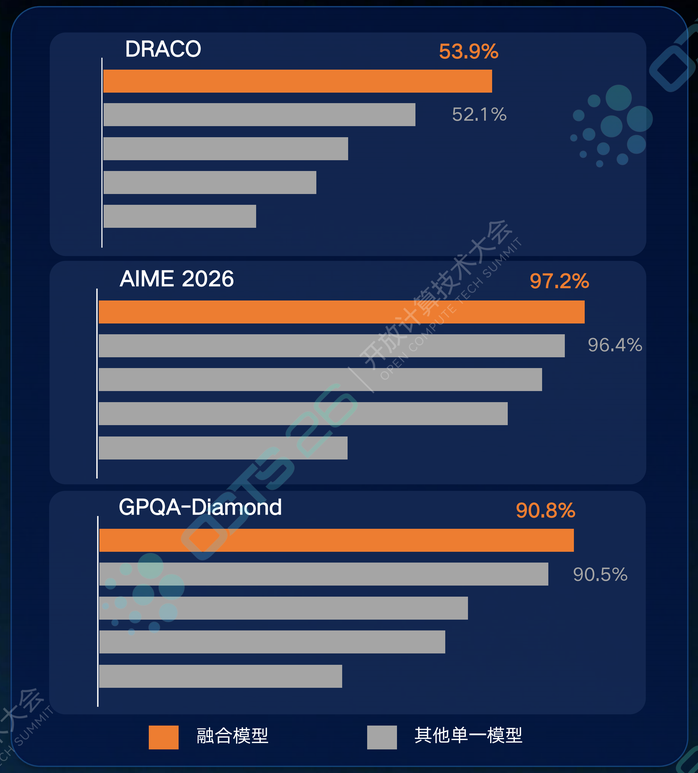
OPEN Compute Project

开放计算标准工作委员会 Open Compute Technology Committee

浪潮信息

多模融合突破单模上限 催生新一代Token Engine

浪潮信息实践表明，系统能力开始由模型协同决定，而不仅依赖单模型能力



多模融合超越单模上限

多模型协同推理的运行机制

Token Engine 能力要求

单机多模型共生
多模型同时驻留显存，免频繁加载切换

高吞吐 Token 服务
支撑数百 Agent 持续并发生成

超低时延响应
毫秒级完成 Agent 之间调用

长上下文稳定承载
支撑复杂 Coding Workflow 连续推理

显存统一共享
上下文统一管理，降低数据复制开销

高带宽互连
多卡协同推理，释放群智协同能力

多模融合对 Token Engine 提出更高要求

SD200：业界首个多模融合超节点

国内超节点系统 token 速度首次突破 5ms

8.9ms
2025.9 OCTS

4.77ms
2026.7 OCTS

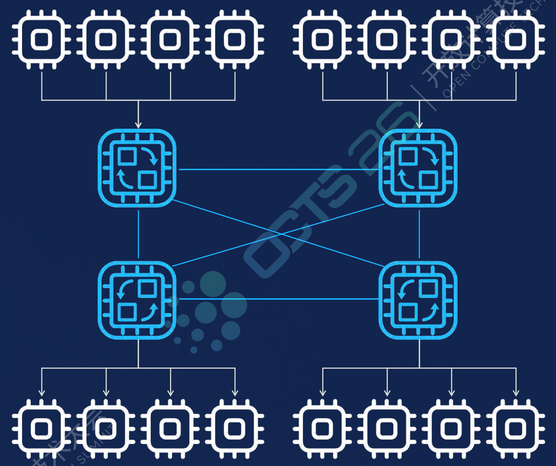
全面兼容国内开源模型，4T 显存单机万亿参数多模融合

DeepSeek	GLM	Kimi	LLaMA
Mimo	Minimax	Qwen	Step

基于 OCM（开放算力模组）与 OAM（开放加速模块）两大架构打造，
构建高性能交换单元打造3D Mesh高性能互连超扩展系统，支持64张本土AI芯片高密度算力扩展，满足大模型的低延迟推理需求，加快Token生成速度。

SD200-Enterprise: 面向企业智能体的 Token Engine

企业级
SD200-Enterprise



单机TB级统一显存，支撑Kimi K2.6等主流万亿参数开源模型本地部署，满足长文本、复杂推理及Agent自主协同等企业应用需求。

基于 Open Fabric Switch 构建16 GPU统一Scale-up互连域，实现原生内存语义通信与统一寻址，万亿参数模型推理TTFT提升40%。

开放架构兼容多元算力，适配主流AI开发生态与国内开源模型，降低企业应用迁移和开发适配成本，让超节点能力真正进入企业生产环境。

OPEN Compute Project® | 开放计算标准工作委员会 Open Compute Technology Committee

浪潮信息

双底座协同 多模融合驱动群体智能

高性能推理底座

多模型并行推理

大显存 / 多卡互联 / 高带宽通信
同时支撑多个候选模型并行运行

Token 服务高效供给

面向多 Agent 并发请求
提供稳定、高吞吐的 Token 生成能力

长上下文稳定承载

支撑长上下文输入、多轮任务调用和
复杂推理过程，保障模型持续稳定输出

候选模型 A
(推理强)

候选模型 B
(知识强)

候选模型 C
(代码强)

元脑企智
EPAI

Model Fusion（多模融合）

推理强 / 知识强 / 代码强
多模型按需协同调用

Agent 群体协同运行

Agent Swarm（群体智能）

融合输出
形成最优综合答案

SD200 超节点

大显存 · 高带宽 · 多卡互联，支撑多模型并行推理与高吞吐 Token 生成

液冷原生 CPU Agent 主机

高并发 · 强隔离 · 易扩展，承载 Agent 调度编排与工具服务运行

开放拓展打造领先Agent基础设施 加速全产业AI转型

加速产业 AI 转型

原生液冷 CPU 整机柜服务器
4万+ 并发 Agent 主机

群智协同

Agentic AI

多模融合

原生液冷 GPU 超节点
十万亿参数模型 Token 引擎

Open Scaling for Super Tokens

THANKS