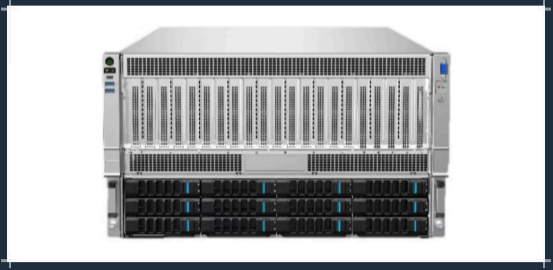
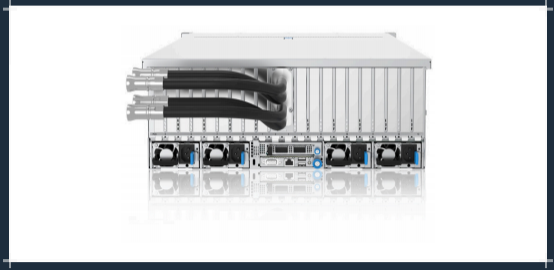
军航科工半导体
JH SEMICONDUCTOR

PRODUCT PORTFOLIO





JH SEMICONDUCTOR · AI COMPUTING SERVERS & INFRASTRUCTURE SOLUTIONS

AI服务器产品与算力基础设施解决方案

企业级服务器 · 液冷集群 · 企业级SSD与DDR5内存 · 私有化算力平台

军航科工 · PRODUCT PORTFOLIO · 2026

军航科工半导体（深圳）有限公司

4 / 8 / 10 GPU | PCIE 5.0 | 风冷 / 液冷

POSITIONING · 战略定位

军航科工的战略定位

面向企业级AI部署，交付从服务器、存储到算力平台的完整基础设施——聚焦可控、可扩展、可交付的AI算力基础设施，以AI服务器为核心，以企业级SSD与DDR5内存为差异化能力，向私有化算力集群和运营平台延伸。

01

算力服务器

+ 4 / 8 / 10卡GPU服务器
+ 风冷、液冷、边缘推理

02

企业级存储

+ PCIe 4.0 / 5.0 SSD
+ U.2 / U.3、E3.S、AIC

03

服务器内存

+ DDR5 RDIMM
+ 容量、频率与平台兼容验证

04

算力平台

+ 统一纳管、调度、监控
+ 多租户与私有化部署

军航价值主张 | 可控——私有化部署与供应链可控 · 可扩展——节点、存储、网络按需扩展 · 可交付——BOM、集成、测试、上线一体化

FULL STACK · 全景

产品与解决方案全景

硬件、数据平面、平台软件与交付服务形成完整技术栈。

应用层	**行业应用与AI工作负载**	大模型推理 | AIGC | 智能制造 | 金融风控 | 科研计算 | 企业智能体
平台层	**JH AI算力管理平台**	统一纳管 | 资源池化 | 任务调度 | 多租户 | 监控告警 | 计量计费
基础设施层	**AI服务器与集群**	4/8/10卡GPU服务器 | 风冷/液冷 | PC-Farm | 机柜级集群
基础设施层	**数据平面与基础组件**	企业级SSD | DDR5 RDIMM | 高速网络 | 冗余电源 | 液冷CDU
交付与服务	**项目交付与服务**	方案设计 · 集成测试 · 性能调优 · 现场部署 · 运维保障

市场布局及全栈解决方案

To C 个人

OPC大众创业万众创新

- 个人AI工作站全栈产品
- AI PC
- 音视频AI终端产品

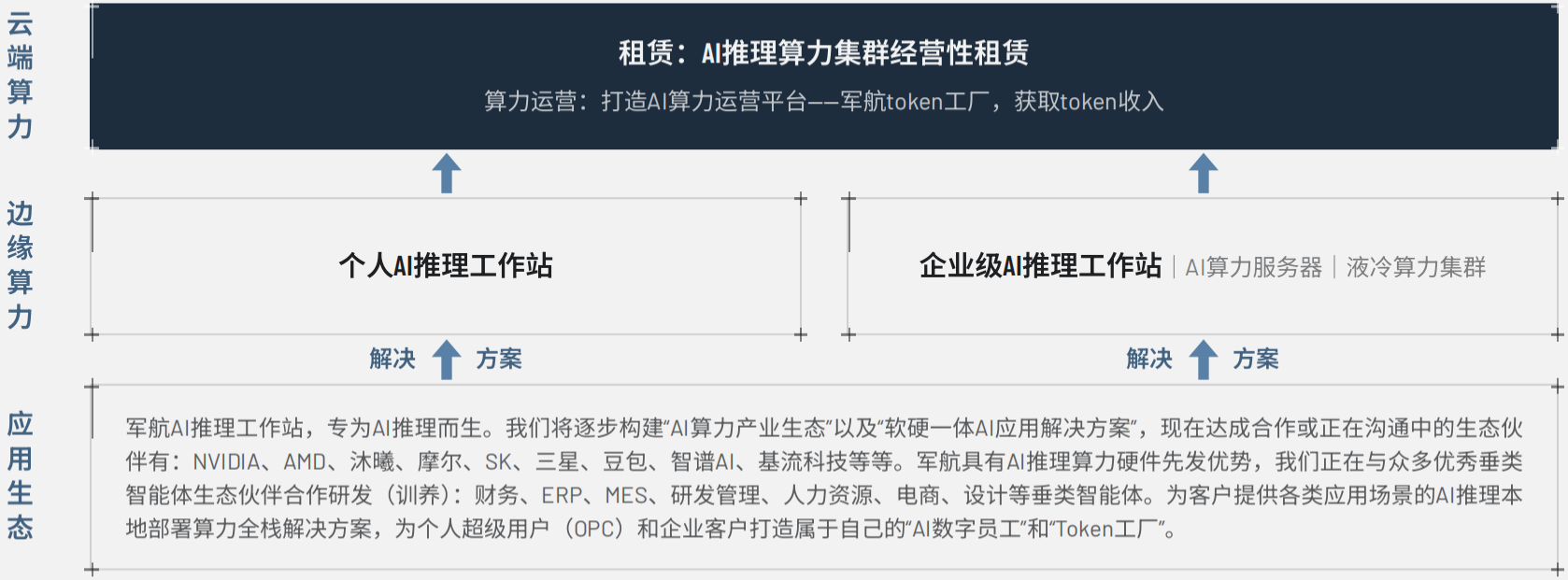
To B 企业级

企业AI应用全栈解决方案

- 企业级AI推理服务器
- 4 / 8卡AI算力服务器
- 算力集群

应用场景 | 智能家居、智慧办公、智慧出行、金融、政务、医疗、科研.....

商业模式


云端算力
租赁：AI推理算力集群经营性租赁
算力运营：打造AI算力运营平台——军航token工厂，获取token收入
边缘算力
个人AI推理工作站
企业级AI推理工作站 | AI算力服务器 | 液冷算力集群
解决 方案
解决 方案
应用生态
军航AI推理工作站，专为AI推理而生。我们将逐步构建"AI算力产业生态"以及"软硬一体AI应用解决方案"，现在达成合作或正在沟通中的生态伙伴有：NVIDIA、AMD、沐曦、摩尔、SK、三星、豆包、智谱AI、基流科技等等。军航具有AI推理算力硬件先发优势，我们正在与众多优秀垂类智能体生态伙伴合作研发（训养）：财务、ERP、MES、研发管理、人力资源、电商、设计等垂类智能体。为客户提供各类应用场景的AI推理本地部署算力全栈解决方案，为个人超级用户（OPC）和企业客户打造属于自己的"AI数字员工"和"Token工厂"。

业务布局及价值版图

2024

军航智算

聚焦推理算力设备，投入工作站、高性能PC的研发与生产。

2025

军航算云

打造算力运营平台，提供推理算力网络一体化服务。

2025

JH AI HUB

发布第一代桌面级AI推理工作站；AI算力服务器实现研产销。

2026

军航智维

开展算力运维业务，提供算力设备运营保障服务。

2026

领航AI平台

打造分布式算力调度平台（JH Pilot AI），提供低成本、高效率、高可控的算力资源服务。

价值终点 | Token工厂

业务闭环:研、配、采、产、测、交、运、保

以项目制交付模型管理服务器全生命周期。

研	配	采	产	测	交	运	保
需求分析	**方案配置**	**供应链组织**	**集成制造**	**验证调优**	**现场上线**	**平台运营**	**售后保障**
业务场景 模型与并发	GPU / CPU 存储 / 网络	核心器件 BOM优化	装配、线缆 固件烧录	Burn-in NCCL / RDMA	上架、组网 验收交付	监控、调度 资源优化	远程支持 备件与维修

项目输入

- 模型规模、并发与时延目标
- 机房功率、网络、散热条件
- 采购预算、交付周期、合规要求

交付输出

- 最终BOM与配置清单
- 性能、稳定性与兼容性测试报告
- 上架部署、验收文档与运维手册

核心竞争力

以存储能力和工程交付能力构建AI服务器差异化。

1

企业级存储与内存整合

- 企业级SSD选型、固件与耐久度匹配
- DDR5 RDIMM容量、频率与平台兼容验证
- 系统盘、缓存盘、数据盘分层配置

2

GPU服务器工程能力

- GPU拓扑、PCIe 5.0、供电与风道设计
- BIOS/BMC、驱动、CUDA与固件兼容
- Burn-in、ECC、掉卡与降速问题定位

3

集群与网络调优

- NCCL、GPUDirect RDMA与多节点测试
- RoCE / InfiniBand网络方案与压测
- 训练/推理吞吐、时延和资源利用率优化

4

供应链与全生命周期交付

- GPU、CPU、SSD、RDIMM、NIC协同采购
- 项目BOM成本控制与交付节奏管理
- 部署、运维、返修、备件与售后闭环

AI服务器产品矩阵

覆盖边缘推理、企业级推理、训练集群和高密度节点。

产品平台	形态	GPU扩展	散热	核心定位
8卡AI算力服务器 RTX 6000D平台	4U/5U机架式	8张双宽卡，单卡最高600W	独立风道风冷	高性价比企业级推理整机
10卡高密度推理服务器 Thor T5000平台	高密度机架式	最大10张双芯卡	高风量风冷	高密度低时延推理
8卡风冷AI服务器 Intel / AMD平台	4U/5U机架式	8张双宽/四宽卡	CPU/GPU独立风道	通用训练与推理
8-10卡液冷AI服务器	4U/5U机架式	8-10张三宽卡	CPU/GPU冷板液冷	高功率密度与能效优先
PC-Farm 4U4H	4U机架式	4个独立PC节点	节点独立风冷	分布式推理、云桌面、渲染
PC-Farm 6U6H	6U机架式	6个独立PC节点	节点独立风冷	高密度分布式节点集群

产品名称为军航方案命名；平台图片与配置用于方案示意，最终以项目规格书为准。

AI SERVER

核心产品

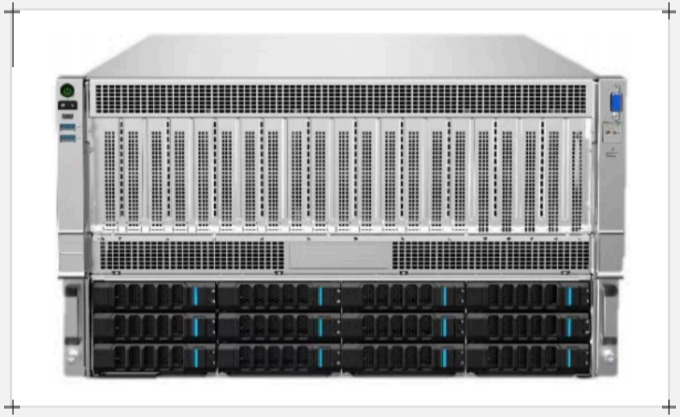
从单机到集群，覆盖风冷、液冷、边缘推理与分布式节点

风冷 | 液冷 | PCIE 5.0 | GPU高密度 | 集群交付

JH 8-GPU SERVER

8卡AI算力服务器 | RTX 6000D平台

面向大模型推理、AIGC与视觉计算的高性价比整机方案。



8×双宽GPU　单卡最高600W　PCIE 5.0

定位 | 最优性价比企业级推理整机

处理器与扩展

- 支持2颗Intel至强四代/五代或AMD EPYC 9004/9005
- 11个PCIe 5.0插槽，支持高速网络适配器

存储与I/O

- 前置12个LFF盘位
- SATA/SAS×8＋SATA/SAS/NVMe×4
- VGA、BMC、Type-C串口

GPU与散热

- 支持8张双宽GPU
- 单卡功耗最高600W
- CPU/GPU独立风道设计

典型场景

- 大语言模型推理
- AIGC与多模态生成
- 视觉计算、数字人、渲染

平台图片及配置用于方案示意，最终以项目BOM、交付规格书和实际测试结果为准。

10卡高密度推理服务器 | Thor T5000平台

高密度双路机架式推理方案，强调GPU密度与低时延。

最大10张双芯T5000 | 4×CRPS电源 | PCIE 5.0×16

定位 | 高性能、高密度、双路推理整机

处理器与存储

- 支持2颗Intel至强四代/五代
- 前置12个3.5/2.5英寸SAS/SATA盘位
- 选配支持4个U.2

GPU密度

- 最大支持10张双芯Thor T5000 GPU
- 10个PCIe 5.0 x16双槽位
- 1个PCIe 5.0 x16单槽位

电源与管理

- 4个标准CRPS电源
- 支持2+2、3+1冗余模式
- 模块化I/O：网口、USB、IPMI、VGA、COM、UID

典型场景

- 在线大模型推理
- 高并发API服务
- 企业级智能体与知识库

平台图片及配置用于方案示意，最终以项目BOM、交付规格书和实际测试结果为准。

8卡风冷AI服务器 | Intel / AMD平台

通用型训练与推理服务器，兼顾扩展能力、稳定性与成本——通用风冷平台，适合标准机房快速部署，降低液冷改造门槛。

1	双路处理器	Intel至强四代/五代或AMD EPYC 9004/9005
2	GPU扩展	支持8张双宽/四宽涡轮卡，单卡最高600W
3	存储与I/O	12个LFF盘位；12个PCIe 5.0插槽
4	热设计	CPU/GPU独立风道，高风量服务器风扇

平台图片及配置用于方案示意，最终以项目BOM、交付规格书和实际测试结果为准。

JH LIQUID-COOLED

8-10卡液冷AI服务器

面向高功率密度、低PUE与规模化集群的冷板式液冷平台——CPU/GPU全覆盖冷板液冷，降低机房环境要求，提升高密度部署能力。

高密度GPU

- 支持8-10张三宽GPU
- 单卡功耗最高600W
- 适配高功率推理与训练负载

平台扩展

- 双路Intel至强或AMD EPYC
- 12个PCIe 5.0插槽
- 高速网络与存储扩展

液冷系统

- CPU/GPU冷板全覆盖
- 可对接CDU与机房水路
- 支持集群级供回水监控

目标价值

- 提高单机柜算力密度
- 降低风冷噪声与热负荷
- 提升长期能效和可维护性

平台图片及配置用于方案示意，最终以项目BOM、交付规格书和实际测试结果为准。

边缘AI算力服务器

AI算力服务器 – Intel / AMD液冷平台

- **顶级处理器支持：**支持2颗Intel至强四代/五代或AMD EPYC 9004/9005
- **前置12 LFF盘位：**SATA/SAS×8＋SATA/SAS/NVME×4，灵活选配
- **极致GPU扩展能力：**支持8–10张三宽卡，单卡功耗高达600W，驾驭顶级算力
- **智能挂耳设计：**支持VGA、BMC接口、Type-C串口
- **先进冷板式液冷：**CPU/GPU全覆盖液冷散热，效率更高，降低机房环境要求
- **高速I/O与网络：**12个PCIe5.0插槽，带宽翻倍，支持最新高速网络适配器

定位｜最优能效比，推理整机方案

AI算力服务器 – Intel / AMD涡轮卡平台

- **顶级处理器支持：**支持2颗Intel至强四代/五代或AMD EPYC 9004/9005
- **前置12 LFF盘位：**SATA/SAS×8＋SATA/SAS/NVME×4，灵活选配
- **极致GPU扩展能力：**支持8张双宽/四宽涡轮卡，单卡功耗高达600W
- **智能挂耳设计：**支持VGA、BMC接口、Type-C串口
- **CPU/GPU独立风道设计：**GPU散热与CPU散热互相独立
- **高速I/O与网络：**12个PCIe5.0插槽，带宽翻倍，支持最新高速网络适配器

定位｜最优性价比，推理整机方案

PC-Farm分布式节点服务器

模块化、热插拔、独立上下电，适配分布式推理与云桌面场景。

6U6H

6个可插拔PC节点

- 标准6U机架式服务器架构，支持6个可插拔PC节点（每节点支持一张GPU卡）
- 机箱内部集成供电和管理背板，可为每个节点提供750W峰值供电功率
- 各节点支持热插拔、完全独立上下电，单节点故障或插拔维护不影响其他节点工作
- 每个节点配置2个可插拔8038高功率服务器风扇，提供温控SmartFAN功能
- 4路CRPS 1100W AC电源，支持冗余和热插拔功能
- 长×宽×高：715×440×265mm｜重量52.3KG｜工作温度5°C-35°C

4U4H

4个可插拔PC节点

- 标准4U机架式服务器架构，支持4个可插拔PC节点
- 机箱内部集成供电和管理背板，可为每个节点提供750W峰值供电功率
- 各节点支持热插拔、完全独立上下电，单节点故障或插拔维护不影响其他节点工作
- 每个节点配置2个可插拔8038高功率服务器风扇，提供温控SmartFAN功能
- 4路CRPS 1100W AC电源，支持冗余和热插拔功能
- 长×宽×高：715×440×176mm｜重量43.5KG｜工作温度5°C-35°C

平台图片及配置用于方案示意，最终以项目BOM、交付规格书和实际测试结果为准。

军航 Junhangclaw 个人超级工作站系列

入门款

JH-W50

芯片	国产芯片
算力	50 TOPS
适用	个人用户

增强款

JH-W120

芯片	算力芯片
算力	120 TOPS
适用	小团队

专业款

JH-W280

芯片	沐曦 国产芯片
算力	280 TOPS INT8
适用	政企信创

旗舰款

JH-W2000N

芯片	NVIDIA Thor
算力	2070 TOP FP4稀疏
适用	企业 / 部门 / 团队

军航JH-W2000工作站

接口（两型号相同）：USB4.0全功能20G×1、USB3.2 Type-A×3、USB2.0 Type-A×4；HDMI2.1×1、DP1.4a（HBR3, MST）×1、USB4.0×1

JH-W2000N

架构 / GPU	NVIDIA Blackwell GPU
CPU	14-core Arm Neoverse
CUDA核心	2560
Tensor核心	96颗第五代Tensor核心
内存	高达128GB LPDDR5x 统一内存
存储	1TB NVME；4个2.5英寸盘位扩展组件
散热	140mm冷排液冷
网络	1×2.5GbE；最大4×25GbE（QSFP28）
AI性能	2070 TFLOPS（FP4稀疏）
系统尺寸	199.3×199.3×199.3mm
电源规格	DC外置电源 300W / 20V

JH-W2000

架构 / GPU	NVIDIA Blackwell GPU
CPU	14-core Arm Neoverse
CUDA核心	2560
Tensor核心	96颗第五代Tensor核心
内存	高达128GB LPDDR5x 统一内存
存储	1TB NVME
散热	140mm冷排液冷
网络	1×2.5GbE；最大4×25GbE（QSFP28）
AI性能	2070 TFLOPS（FP4稀疏）
系统尺寸	199.3×199.3×100.3mm
电源规格	DC外置电源 300W / 20V

ENTERPRISE · 企业级方案

军航 Junhangclaw 企业级解决方案

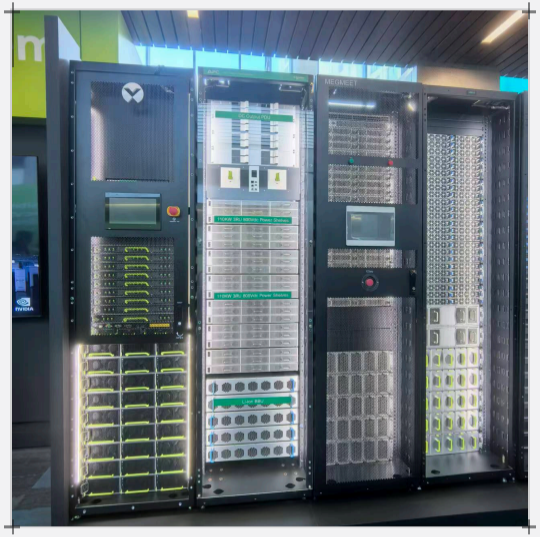
为AI应用提供从团队到企业的整体算力解决方案：私有化部署，按需选配，弹性扩展，Token工厂。

产品线	算力规模	适用
双卡工作站	500-600 TFLOPS	中小团队
四卡工作站	1000-1200 TFLOPS	中型企业
8卡算力服务器	5000+ TFLOPS	中大型企业
液冷柜式集群	20000+ TFLOPS	含CDU液冷

软件平台 | 算力调度软件 · 企业级管理平台 · 集群运维管理

机柜级液冷算力集群

服务器、网络、存储、供配电与液冷系统的一体化交付。



液冷整柜架构示意

计算节点

- 4/8/10卡GPU服务器
- 风冷或液冷

高速网络

- RoCE / InfiniBand
- GPUDirect RDMA

存储层

- 本地NVMe＋共享存储
- 冷热数据分层

基础设施

- PDU / BBU / CDU
- 水路与环境监控

平台软件

- 资源纳管、任务调度
- 监控告警与计量

机柜图片为架构示意；具体机柜功率、冷却方式、网络和节点数量按项目设计。

企业级存储与内存子系统

军航将服务器、SSD与RDIMM纳入同一兼容性和质量闭环。

企业级SSD

- PCIe 4.0 / 5.0，U.2 / U.3、E3.S、AIC
- 1.92TB、3.84TB、7.68TB、15.36TB、30.72TB
- PLP、端到端保护、QoS、Telemetry与安全擦除

服务器存储分层

系统盘	1.92 / 3.84TB
缓存盘	高写入耐久度
数据盘	7.68 - 30.72TB
日志盘	稳定QoS

DDR5 RDIMM

- 容量：64GB / 96GB / 128GB及项目定制
- 频率：4800 / 5600 / 6400 MT/s
- 面向主流双路服务器平台的SPD与BMC兼容验证

平台验证重点

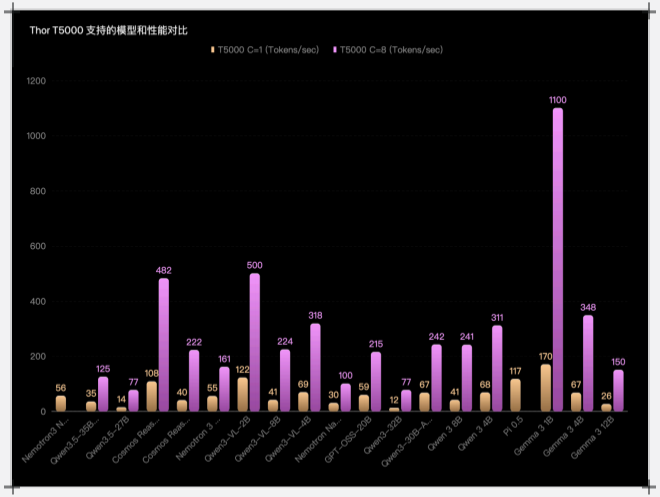
- BIOS识别｜BMC / iDRAC｜全容量显示
- 压力稳定性｜温度与功耗

容量、频率和功能按项目选型；最终参数以产品规格书和兼容性测试报告为准。

Thor T5000模型兼容与推理性能示例

覆盖主流开源模型系列，支持从小参数模型到MoE与大模型推理。

模型系列	细节
Google Gemma3	Gemma 3 270M、1B、4B、12B、27B
NVIDIA Nemotron	Nemotron3 Nano 4B ☀、Nemotron3 Nano 30B-A3B、Nemotron Nano 9B v2、Nemotron Nano 12B VL
NVIDIA Cosmos Reason	Cosmos Reason 1 7B、Cosmos Reason 2 2B、Cosmos Reason 2 8B ☀
Alibaba Qwen3.5	Qwen3.5 35B-A3B（MoE）☀、27B、9B ☀、4B、0.8B
OpenAI GPT OSS	GPT OSS 20B、GPT OSS 120B
Alibaba Qwen3	Qwen3 4B、8B、30B-A3B（MoE）、32B、Qwen3 VL 4B、VL 8B
Mistral AI Ministral 3	Ministral 3 3B/8B/14B Instruct、Ministral 3 3B/8B/14B Reasoning
Meta Llama 3	Llama 3.2 3B、Llama 3.1 8B、Llama 3.1 70B



性能受模型版本、量化精度、上下文长度、并发数、软件栈和测试环境影响；图表数据用于平台能力示例。

JH AI算力管理平台

统一纳管异构算力，实现资源池化、调度、监控与计量。

JH AI算力管理平台 总览 设备管理 算力池 任务监控 报表分析 管理员

24	18	128	76%
节点总数	在线节点	GPU总数	资源利用率

节点运行状态

节点	设备类型	GPU	状态	利用率	温度
JH-EDGE-01	边缘推理	8×GPU	在线	82%	62°C
JH-TRAIN-02	训练节点	8×GPU	在线	71%	59°C
JH-LC-03	液冷节点	10×GPU	维护	—	38°C

1 · 统一纳管

整合自营与第三方异构XPU资源

2 · 智能调度

按成本、性能和优先级分配资源

3 · 弹性计费

支持包年包月、按需与项目计费

4 · 全栈支持

覆盖训练、推理、监控和生命周期管理

平台界面为军航产品化示意，可根据客户现有云平台、调度系统和权限体系进行定制集成。

私有化算力平台交付架构

按业务规模从单机、节点池扩展至机柜级集群。

企业应用层

知识库 / Agent / AIGC

模型服务层

推理引擎 / API网关

算力平台层

调度 / 监控 / 多租户

GPU服务器池

4 / 8 / 10卡节点

高速网络

RoCE / IB / 以太网

存储与内存

NVMe SSD / RDIMM

闭环：业务请求 → 模型服务 → 资源调度 → 计算 / 网络 / 存储 → 结果返回

支持单机、私有集群、混合云与多地节点。

典型部署场景

金融	私有化模型与智能投研
制造	质检、知识库与工艺智能
政企	安全可控的本地AI平台
科研	模型验证与高性能计算
AI企业	推理API与Token工厂
海外	本地化交付与边缘节点

项目交付与服务保障

以可验收、可运维、可扩展为交付标准。

01

售前设计

- 需求澄清
- 容量规划
- 预算与BOM

02

集成验证

- 装配与烧机
- 驱动 / 固件
- 性能基线

03

现场交付

- 上架组网
- 平台部署
- 验收培训

04

运维保障

- 远程监控
- 问题定位
- 备件与返修

PRODUCT PORTFOLIO

JH SEMICONDUCTOR

让企业以更低成本获得可控、可扩展的AI算力基础设施

服务器 · 企业级SSD · DDR5内存 · 液冷集群 · 算力平台

军航科工半导体（深圳）有限公司

AI SERVER

STORAGE

PLATFORM