升技云计算 ABIT CLOUD

国产化服务器与算力解决方案提供商

高校AI

国产化 • 自主可控 • 开箱即用

高校人工智能算力解决方案 Single Machine & Cluster

海光 C86 3号 / 4号 处理器 × 华为昇腾计算卡 × AI 单机 • 智算集群



海光 4 号双路 4U8 卡 AI 服务器 · H4G2-408

单机 • AI 训练服务器

校级 • 智算集群

全栈国产化

www.abitcloud.cn

· 面向高校科研与教学的 AI 基础设施 · 2026.11 · V1.0

方案总览

从课题级单机到校级智算集群，一套国产化体系全覆盖

01

单机算力 • 课题级

H4G2-408 4U 8卡 AI 服务器

· 双路海光 4 号 74XX 处理器

· 最高 8 块双宽 AI 计算卡

· 24 DIMM DDR5 · 最高 3TB 内存

· 预装 AI 环境 · 开箱即用

面向课题组 / 实验室，一台设备即可开展大模型训练与推理

02

校级集群 • 院系级

GPU 节点 + CPU 节点 + 并行存储

· H4G2-408 GPU 计算节点池

· H4R2-212 通用计算节点池

· 200Gbps RoCE 高速互联

· Slurm 调度 · Web 门户管理

多院系共享、用户隔离、资源统计记账，校级算力平台

03

全栈国产 • 自主可控

从芯片到软件全链路国产化

· 海光 C86 x86 架构国产 CPU

· 华为昇腾国产 AI 计算卡

· 支持 openEuler / 麒麟 OS

· 国密硬件级安全可信

满足信创合规要求，科研数据自主可控，符合教育行业采购导向

425 TFLOPS FP16	112 GB HBM	8 卡 / 节点	3 TB DDR5	200 Gbps
昇腾 Atlas 350 单卡算力	单卡显存 · 1.4TB/s 带宽	双宽 AI 计算卡扩展	单节点内存容量	集群 RoCE 互联带宽

计算基座 • 海光 C86 处理器（3号 / 4号）

国产 x86 架构高性能处理器 —— 无需改代码，现有科研软件直接运行

旗舰性能 • 7000 系列

集成 16–32 核心，面向数据中心高负载　128 路 PCIe 通道 · 8 通道内存控制器

单颗 CPU 最大支持 2TB 内存　大容量三级缓存 · 高内存带宽与 I/O 带宽

生态兼容 • 零成本迁移

主流 x86 指令集，科研软件生态成熟　现有 AI / HPC 代码无需移植即可运行

PyTorch、MPI、Slurm 等直接可用　应用迁移成本低，避免重复开发

安全可信 • 信创合规

内置国密算法硬件加速引擎　基于专用硬件的安全启动

完整新一代可信计算方案　满足教育行业国产化采购要求

C86-3G • 3 号平台

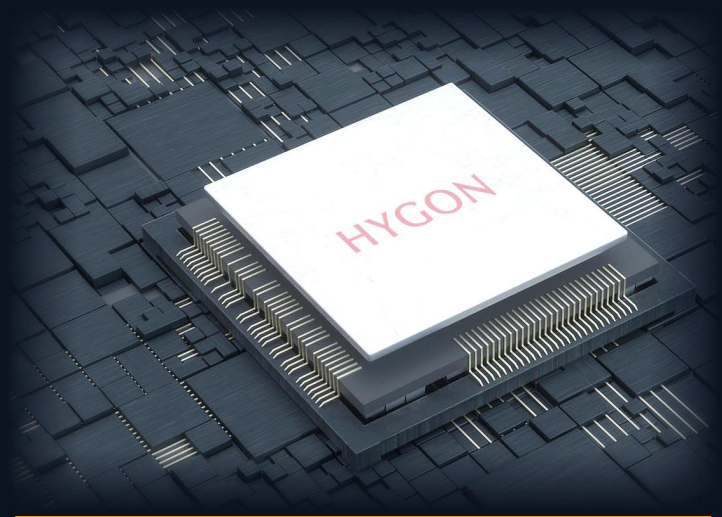
H3R2-212 2U 通用服务器 · 7000 系列 7 系处理器

面向通用科研计算、数据预处理与 AI 推理服务

C86-4G • 4 号平台

H4G2-408 / H4R2-212 · 74XX 处理器

内存升级 DDR5-4800 · PCIe 5.0 高速扩展，面向 AI 大算力



国产 x86 架构

PCIe 5.0 扩展

DDR5-4800

国密算法引擎

海光 C86 处理器产品视觉 · 图源：海光官网 hygon.cn

AI 算力核心 • 华为昇腾计算卡

训练 + 推理双卡型选择，全部官方规格 —— 单机与集群均以此为算力单元

Atlas 350 加速卡

训练 / 高算力

AI 算力	425 TFLOPS FP16 / BF16
低精度算力	1561 TFLOPS mxFP4 • 804 TOPS INT8
显存	112 GB HBM • 带宽 1.4TB/s
接口	PCIe 5.0 x16 • 双向 128GB/s
卡间互联	灵衢 UB 端口互联 318–424GB/s
功耗 / 形态	≤600W • 双宽全长被动散热

适用 H4G2-408 训练节点 · 建议 4 卡 / 节点

Atlas 300I Duo 推理卡

推理 / 高能效

AI 算力	280 TOPS INT8 / 140 TFLOPS FP16
显存	96 GB LPDDR4X • 408GB/s
视频分析	256 路 1080P 实时解码分析
接口	PCIe 4.0 x16
能效	1.86 TOPS/W 业界领先
功耗 / 形态	150W • 单槽全高全长

适用推理节点 · 单节点可部署 8–12 卡

软硬协同：H4G2-408 提供 12 个 PCIe 插槽 / 8 个双宽卡位 + 3200W 冗余电源，双宽训练卡与单槽推理卡均可适配

训练用 Atlas 350（4 卡/节点）· 推理用 300I Duo（8 卡起）· 两型可混插，按预算与场景配比

华为昇腾 · 软件生态与高校选型

从底层硬件到上层框架的全栈使能 —— 按场景选卡，按预算组网

应用层

大模型训练 / 推理 · CV / NLP / 多模态科研
主流开源大模型昇腾适配版本可直接部署

框架层

PyTorch（torch_npu）· MindSpore · PaddlePaddle
DeepSpeed / Accelerate 分布式训练框架

使能层

CANN 异构计算架构 · 算子库与自动调优
MindStudio 开发套件 · 模型迁移工具

硬件层

Atlas 350 训练加速卡 · Atlas 300I Duo 推理卡
PCIe 标准形态，兼容海光 C86 服务器平台

高校场景	推荐配置
教学实验 / 推理服务	**Atlas 300I Duo • 2U 推理节点 • 8-12 卡** 单卡 96GB 大显存 · 150W 低功耗
课题组训练	**Atlas 350 • H4G2-408 • 4 卡 / 节点** 单节点 1.7 PFLOPS · 开箱即训
院系大模型训练	**Atlas 350 集群 • 多 GPU 节点 • 200G 组网** HCCL 多卡互联 · 线性扩展
训推混合负载	**双卡型混插 • 训练 + 推理分层部署** 白天推理 · 夜间训练调度复用

生态提示：昇腾生态兼容 PyTorch / MindSpore 主流框架，现有代码迁移成本低；300I Duo 与 350 可按预算分批采购、混部调度

单机方案 • H4G2-408 海光 4 号 4U 8卡 AI 服务器

大模型训练 · 高密度 AI 计算 · 课题组级算力主力机型

一台设备 = 一个 AI 实验室

双路海光 4 号 74XX + 昇腾计算卡 + 3TB 内存
单机即可开展 7B~70B 级模型训练与推理

2 颗
海光 74XX 处理器

24 DIMM
DDR5-4800 · 最高 3TB

8 卡位
双宽 AI 计算卡

12 个
PCIe 扩展插槽

3200 W
CRPS 冗余电源

4 块
NVMe U.2 高速盘位

交付预装昇腾 CANN + PyTorch / MindSpore 环境

配置项	产品规格
处理器	支持两颗海光 74XX 处理器（C86-4G）
内存	24 DIMM · DDR5-4800 MT/s · 最大可扩展至 3TB
硬盘	前置最高 12 个 3.5"/2.5" SAS/SATA · 4 个 NVMe U.2
内置存储	2 个 M.2 SSD（系统盘 / 缓存盘）
GPU 扩展	最大 12 个 PCIe 插槽 · 支持 8 块双宽 AI 加速卡
接口	1 × RJ45 管理口 · 2 × VGA · 4 × USB
电源	3+1 或 2+2 冗余 · 2700W/3200W CRPS 高效电源
系统管理	集成 BMC · 支持 IPMI 2.0 / Redfish / SNMP

单机方案 • 2U 通用计算服务器

推理服务 · 数据处理 · CPU 并行计算 · 集群通用节点

H3R2-212

海光 3 号 7 系 • 2U

· 海光 3 号 7000 系列 7 系处理器平台

· 2U 机架式通用服务器 · 标准机柜部署

· 高性价比 CPU 算力：并行计算、仿真、生信

· 可部署昇腾 300I Duo 推理卡，兼作 AI 推理节点

H4R2-212

海光 4 号 7 系 / 5 系 • 2U

· 海光 4 号 7 系（高性能）与 5 系（主流）可选

· DDR5 内存平台 · 新一代 I/O 扩展能力

· 集群 CPU 计算节点 / 登录节点首选机型

· 按院系负载分层配置，控制整体造价



H 系列通用服务器 • 2U 12 盘位机架机型

AI 推理服务　数据预处理　CPU 并行计算

单机组合建议：H4G2-408（AI 训练）+ H4R2-212（推理 / 预处理）

两台起步即可支撑院系级 AI 科研，后续无缝扩展为集群节点

软件环境 • 开箱即用

从驱动、框架到调度系统全部预装调优，交付即可提交训练任务

AI 框架与应用

PyTorch（torch_npu 昇腾适配）· MindSpore
DeepSpeed 分布式训练 · Transformers 大模型生态

昇腾使能层

CANN 异构计算架构 · MindStudio 开发套件
算子自动调优 · 混合精度训练加速

并行与调度

Slurm 队列调度 · 用户 / 队列 / 权限管理
OpenMPI / HCCL 集合通信 · 分布式数据并行

系统与运维

openEuler / 麒麟 / Ubuntu 可选
Docker 容器 · Conda 环境 · SSH / XRDP / VNC 远程

交付即用 • 五步交付流程

硬件上架通电 → 系统与驱动安装 → AI 框架与算子调优 → 示例任务验证 → 交付培训

交付含：服务器使用手册、教学视频、首次联网调试、环境参数调优报告

集群方案 • 校级智算集群架构

GPU + CPU 异构计算 · 高速无损网络 · 并行存储 · 统一调度

师生用户 • Web 门户 / SSH 提交

作业提交 · 资源查看 · 数据管理 · 计费统计

登录节点 ×2

编译 / 提交 / 开发环境

管理 / 调度节点

Slurm 控制器 · BMC 集中监控

运维服务

监控告警 · 记账 · 一键恢复

200Gbps RoCE 无损网络 • Spine-Leaf 架构 • RDMA 远程直接内存访问

GPU 计算节点池 • H4G2-408 × N

昇腾训练卡 / 推理卡 · 大模型训练与推理

CPU 计算节点池 • H4R2-212 / H3R2-212 × N

数据预处理 · CPU 并行 · 传统 HPC 负载

NVMe 并行存储集群

分布式文件系统 · TB 级聚合带宽 · 数据集与模型集中管理

按需弹性扩展

GPU 节点即插即扩 · 集群规模线性增长

多院系共享：按队列隔离资源 · 用户级配额 · 作业记账统计 · 对接校园网统一认证

组网方案 · 高速计算网络说明

为什么组网 · 怎么组网 —— 分布式训练的效率由网络决定

梯度同步 AllReduce

多卡分布式训练中，各节点每一步都要同步梯度；网络慢一步，千卡同慢，GPU 算力再强也会闲置等待

数据高速供给

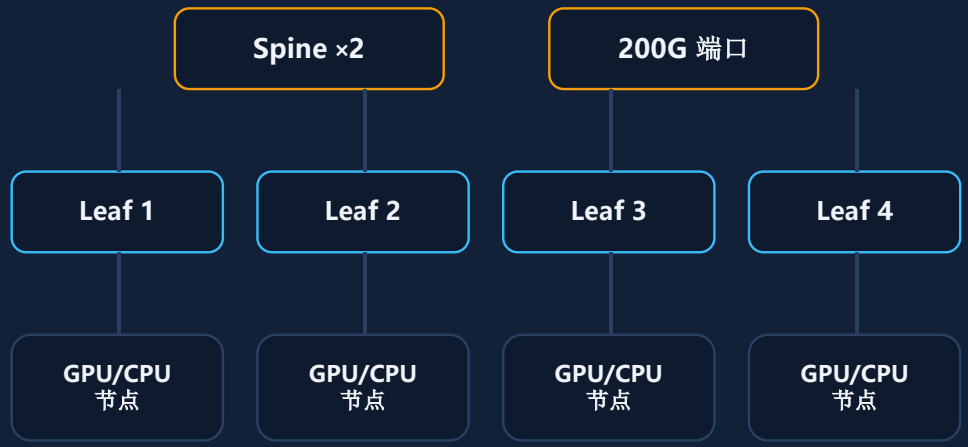
TB 级训练数据集与模型权重需要从并行存储并发读取，带宽不足时 GPU 喂数据跟不上

通信时延决定扩展效率

节点间延迟每降低 1 微秒，集群线性度显著提升；RDMA 绕过 CPU 直接读写内存，时延以微秒计

无损三要素： PFC 优先级流控 · ECN 拥塞通知 · 大缓存交换机

Spine-Leaf 两级无阻塞架构（示意）



每 GPU 节点 2×200G 上联 · 1:1 无收敛 · RoCEv2 RDMA

GPU 节点规模	Leaf 交换机	Spine 交换机	收敛比
2 - 8 节点	2 台	2 台	1:1
8 - 32 节点	4 - 8 台	4 台	1:1
32 节点以上	8+ 台	8 台	1:1 / 超无收敛

组网方案 • 典型规模参考

从两台起步到校级平台，按预算与需求平滑演进

起步型 • 课题级

≈ 3.4 PFLOPS FP16 峰值

GPU 节点

H4G2-408 × 2（每节点 4 卡 Atlas 350）

CPU / 推理

H4R2-212 × 1（300I Duo 推理卡）

存储

节点本地 NVMe + 万兆 NAS

网络

万兆以太接入，节点直连

8 张训练卡起步，两台设备即可支撑课题组级大模型训练与推理，预算门槛最低

标准型 • 院系级

≈ 27.2 PFLOPS FP16 峰值

GPU 节点

H4G2-408 × 8（32 张 Atlas 350）

CPU / 推理

H4R2-212 × 4 · 登录 / 预处理 / 推理

存储

NVMe 并行存储集群（百 TB 级）

网络

100-200G RoCE · Spine-Leaf

多课题组共享，Slurm 队列隔离，白天推理夜间训练，算力利用率最大化

旗舰型 • 校级

100+ PFLOPS FP16 峰值

GPU 节点

H4G2-408 × 32+（128 张以上）

CPU / 推理

CPU 节点池 + 独立推理服务池

存储

多套并行存储 + 冷热分层

网络

200G RoCE 全互联 · 多集群

校级智算平台，对接校园统一认证与计费，支撑全校 AI 教学科研与学科建设

演进路径：起步型直接并入标准型集群 —— 节点即插即扩，Slurm 自动纳管，无需推倒重建，保护每一分投资

服务与保障

从选型测试到质保运维，全周期支持高校科研教学

免费算力测试

提供测试样机与远程环境，跑通课题代码、验证精度与性能后再采购

预装交付

驱动、框架、调度系统全部预装调优，通电联网即可开跑，附使用手册与教学视频

三年整机质保

硬件三年质保 · 质保期内重大故障上门处理，往返运费全部承担

软件技术支持

环境升级、报错排查、小众软件部署指导，售后期内不限次数

使用培训

首次联网调试 + 服务器使用教学，帮助师生快速上手 Linux 集群

招投标支持

支持各省政采平台流程，配合完成方案书、参数、演示等投标环节

交付承诺：验收后 7 日内因硬件故障导致无法计算，免费更换或退货　质保期内提供远程技术支持（软件编译与升级、系统故障解除等）

升技云计算 ABIT CLOUD

以国产算力
育未来智能

高校人工智能算力解决方案 · 海光 C86 × 华为昇腾 · 单机 / 集群

官网：**www.abitcloud.cn**

产品线：AI 服务器 · 国产平台服务器 · 工作站 · 存储 · 边缘计算 · 配件

2026.11 · V1.0