

真武M890

全场景训推一体AI芯片



产品简介

真武系列芯片是面向 AI 训练与推理的数据中心 AI 芯片，由平头哥自主设计研发，采用自研并行计算架构与片间互联技术，配合自研软件栈T-Head SAIL®，实现软硬结合，支持自研软件生态，同时兼容主流AI生态应用。

真武M890原生支持FP32到FP4等多种数据精度，覆盖从高精度训练超低精度推理的全场景需求，既能满足训练精度要求，为大模型训练提供高可靠算力，也能提升推理计算密度，减少内存带宽消耗，显著降低推理成本。

真武M890支持ICN Switch组网和芯片直连两种方式：

- Switch组网：通过ICN Switch可实现64卡无差异化全带宽互联，芯片之间延迟一致性能均衡、节点间通信延迟大幅降低，显著提升大规模并行计算的效率与稳定性；
- 直连方式：每颗芯片有8个独立的ICN端口、提供800 GB/s的总片间互联带宽，可用于对等设备和共享内存访问，有效降低 AI 训练和推理中的通信开销，高效实现算力扩展

产品规格

分类		指标	
存储		容量(GB)	144
		形态	HBM
互联	主机接口	带宽(GB/s)	128
		形态	PCIe 5.0 X 16
	片间互联	带宽(GB/s)	800
		形态	ICN Link
算力		精度	支持FP32/BF16/FP16/ FP8/FP4等 多精度算力
编解码能力	编码	H.264（路）	16
		H.265（路）	16
	解码	H.265（路）	64
		H.265（路）	64

产品优势

全自研

硬件端到端全链路自研；
并行计算架构与片间互联等自研核心模块全面升级

高易用

自研全栈软件栈，全面兼容主流 AI 生态；
提供统一编程接口，业务迁移零代码修改

高性能

内存容量144 GB、片间互联带宽800 GB/s、
支持FP8、FP4低精度数据格式、
支持64卡全带宽互联

灵活部署

支持多进程资源共享、支持多租户安全隔离，
敏捷部署提升资源利用率

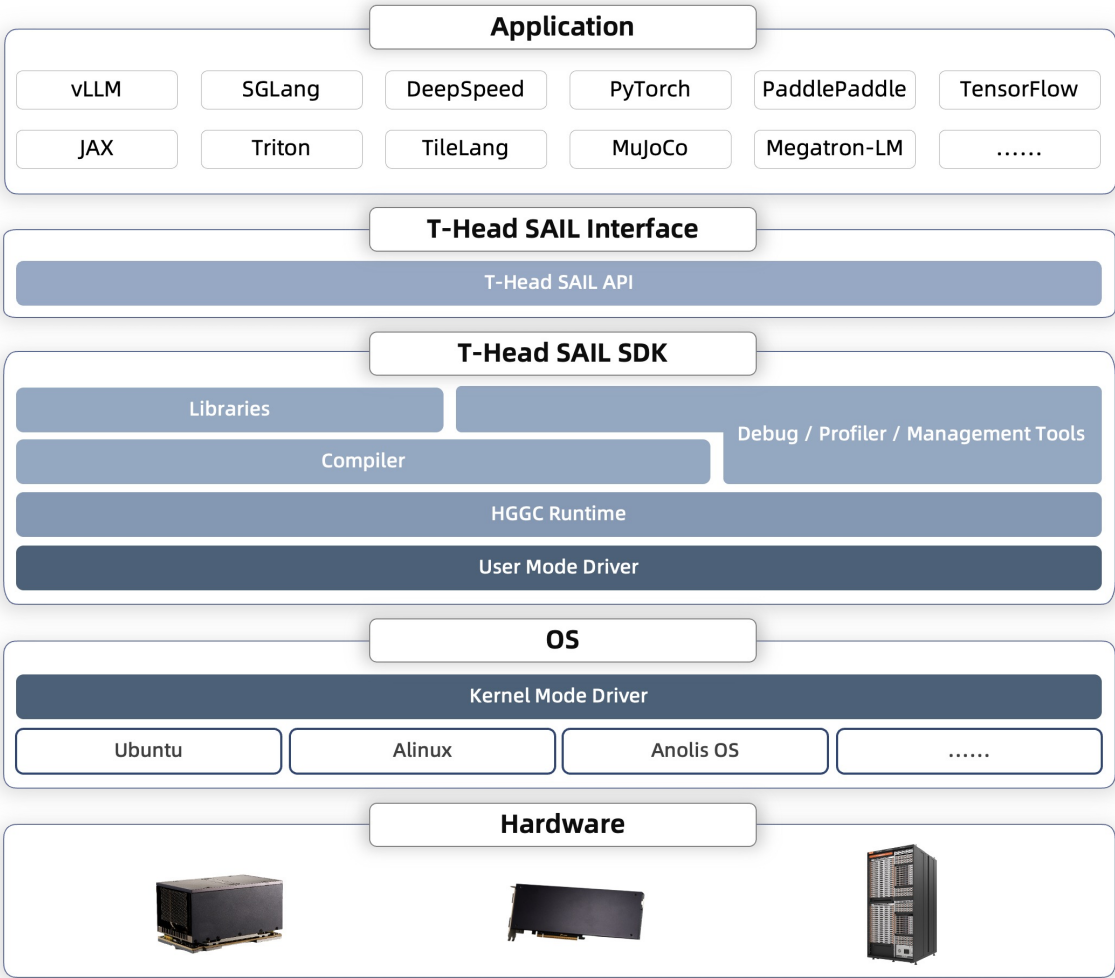
芯云模一体

芯片算力结合阿里完整的AI框架、平台、模型和应用，为客户提供一体化产品和服务

软件栈简介

T-Head SAIL[®]是平头哥自主研发的软件栈，拥有独立知识产权、完备的软件生态及工具链，向下对接底层硬件算力、向上承接上层应用需求，软硬协同、端到端支撑用户自主业务快速落地。软件栈由以下三层架构组成：

- 接口层：**提供统一的编程接口，全面兼容平头哥自研工具链及主流软件生态，屏蔽底层硬件差异，降低开发者接入门槛。
- SDK层：**具备完备的软件生态及开发工具链，赋能开发者与业务团队快速开展应用开发，实现从算法模型到硬件部署的高效流转。
- OS层：**通过KMD（内核驱动）与OS（操作系统）的深度结合，广泛兼容不同底层硬件平台与主流操作系统，通过系统级性能调优，确保全栈运行效率最大化。



应用场景

