

PT200Z AI SSD系列

国产高性能 PCIe Gen 5 企业级 AI SSD 存储解决方案

PT200z是针对人工智能高负载工作负载研发的专业硬盘，PT200z支持PCIeGen5接口，采用pSLC介质方案，提供超大带宽、超低访问延迟、超长写入寿命的持久化存储空间。

深度适配多模态大模型、长上下文推理等热点AI应用场景。为 AI 训练推理突破存储墙提供了全新的解决思路。

产品特点

- 平均随机读写延时：45 / 10 μ s (QD=1)
- PCIe Gen5 x4，双端口 2x2 通道
- 符合 PCI Express SFF-8639 Module Specification Revision 5.0
- 顺序读取：最高 14,800 MB/s、
- 顺序写入：最高 9,500 MB/s
- 随机读取：最高 3,100K IOPS、
- 随机写入：最高 950K IOPS
- 典型功耗（随机读）：< 22W
- 典型功耗（随机写）：< 19W
- 空闲功耗：< 5 W
- MTBF：250 万小时
- DWPD：60（超高耐久）

128 命名空间
掉电数据保护
TCG Opal 2.0 支持
AES-XTS 256-bit 加密
端到端数据保护
Sanitize
NVMe-MI(Management Interface)
SMBus



产品		PT200Z	
容量	1,600 GB	3,200 GB	6,400 GB
顺序读取 (MB/s)	14,800	14,800	14,800
顺序写入 (MB/s)	9,500	9,500	9,000
随机读取 (IOPS)	3,100K	2,800K	2,300K
随机写入 (IOPS)	950K	950K	950K
产品PN	GPSU2241P20HP71T6000	GPSU2241P20HP73T2000	GPSU2241P20HP76T4000
功能特性			
功耗	Idle 功耗 < 5 W, Active 读取 22 W, Active 写入 19 W		
平均随机读写延时	45 / 10 μ s (QD=1)		
接口	PCIe Gen5 x4, 双端口 2x2 通道		
介质	3D TLC (pSLC 模式)		
寿命 DWPD	60 DWPD		
UBER	< 1 sector per 10 ¹⁸ bits read		
MTBF	250 万小时		
工作温度	0°C ~ 70°C		
存储温度	-40°C ~ 85°C		
质保	5 年		
外形尺寸	U.2 15 mm: 100.10(L) × 69.85(W) × 14.70(H) mm		
功能特性	128 个命名空间、128 个命名空间、双端口 (Dual Port)、预留 (Reservation)、元数据保护、热节流 (Thermal throttling)、断电保护 (Power Loss Protection)、硬件 AES-XTS 256 位加密、支持 SMBus		

AI90 训推加速引擎——软硬件协同优化的AI解决方案

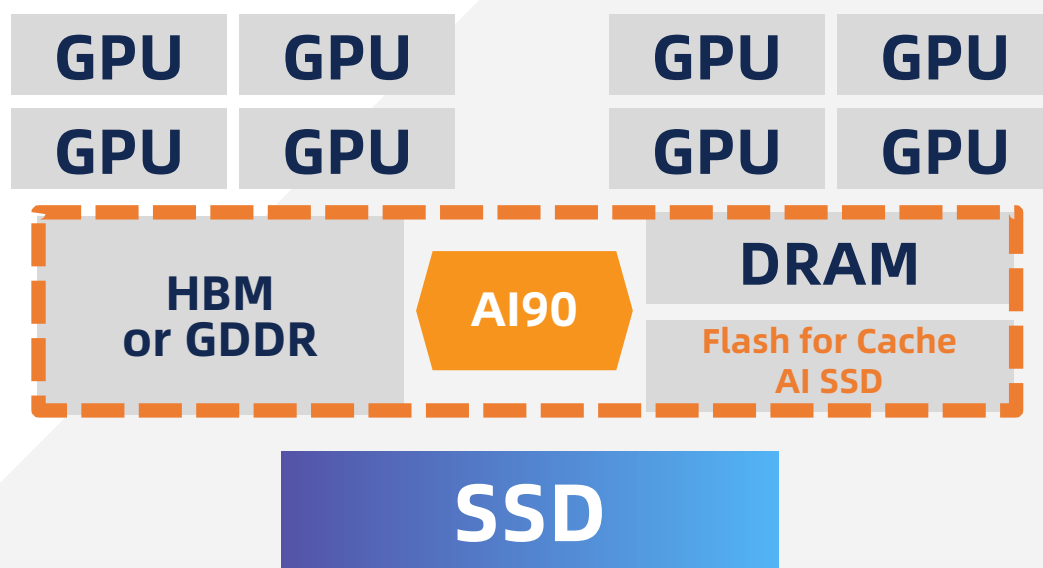
产品介绍

AI90是一套持续发展的软硬件协同优化方案，旨在通过释放GPU带宽潜力、强化KV Cache，有效提升GPU利用率，以合理的成本，释放多模态、长上下文推理的爆炸性需求。

产品特点

1. 软硬件协同——单纯软件方案不足以完全释放系统潜能，过于复杂的硬件设计也会提高不必要的初期投入，AI90以轻量化硬件辅助软件优化的方式，持续提供对训练/推理潜力的提升，助力用户充分发挥昂贵的计算资源。
2. 潜力巨大——加速效果随着模型复杂度，及上下文长度增加而显著增强。
3. 成本友好——对低端硬件同样能提供显著加速效果，且加速效果随GPU数量线性提升。

AI90 的架构图及效率提升



核心矛盾：算力增速远超存储发展

近20年来，GPU算力沿着摩尔定律指数级增长了近6万倍，而作为数据通道的显存容量和带宽始终近乎线性增长仅增加了十几倍，二者之间形成了巨大的鸿沟，从而导致“内存墙”困境。

算力提升
~60,000x
指数级爆发式增长

显存容量
仅16x
增长严重滞后

增长严重滞后
仅30x
数据通道拥堵

“算存分离”直接导致了“内存墙”的出现，使计算单元常因等待数据而处于空闲状态。在AI大模型任务中，主流GPU的有效算力利用率普遍徘徊在30%至60%之间。

传统推理方案 VS AI90加速方案 基于 Llama3 70B 模型测试

传统方案		AI90 加速方案	
硬件配置	8卡RTX 5090	8卡RTX 5090 + AI90模块	
推理吞吐量	~120 tk/s	~610 tk/s（提升5.1倍）	
首Token延迟	27.99s（64K上下文）	0.564s（降低50倍）	
显存利用率	30% - 40%	85% - 95%（提升2-3倍）	
多卡加速	无P2P	智能P2P互联	
服务能力	仅单租户	支持10+租户并发	

从消费级的RTX5090到最新的旗舰B300，无一例外，它们的标准配置都无法完全发挥其算力。实际算力利用率普遍在30-70%之间。这说明，即便是最顶级的硬件，也受制于内存瓶颈。这些标准配置都是厂商在成本和性能之间做出的商业妥协。

5090，H200， B300算力效率说明

RTX 5090

■ 标准配置：32GB GDDR7 / 1.79TB/s

⚡ 实际算力发挥：40% ~ 55%

🚧 核心瓶颈：容量和带宽均严重不足，无法承载大型模型的训练需求。

NVIDIA H200

■ 标准配置：141GB HBM3e / 4.8TB/s

⚡ 实际算力发挥：40% ~ 55%

🚧 核心瓶颈：带宽增长仍落后于算力，处理130B以上模型时，显存容量依然偏紧。

NVIDIA B300

■ 标准配置：288GB HBM3e / 8.0TB/s

⚡ 实际算力发挥：55% ~ 65%

🚧 核心瓶颈：虽为最新旗舰，但面对万亿级参数模型和超长上下文需求，带宽与容量依然存在瓶颈。

选择芯展速AI 90，就是选择在Token经济时代，以更低的成本、更高的效率，赢得竞争先机。



智能P2P互联

解锁消费级GPU多卡并行潜能，突破传统通信限制，实现集群线性扩展。

支持全系列3090/4090/5090及魔改。



动态扩展AI Cache

提供高带宽、低延迟、高耐久持久化存储，突破显存墙，实现对超长上下文的完美支持。
PSLC TB级虚拟显存扩展。



扩容与增速并重

获取千亿级参数模型本地运行的能力，同时大幅度提高token出产效率，大幅度降低token生产成本。