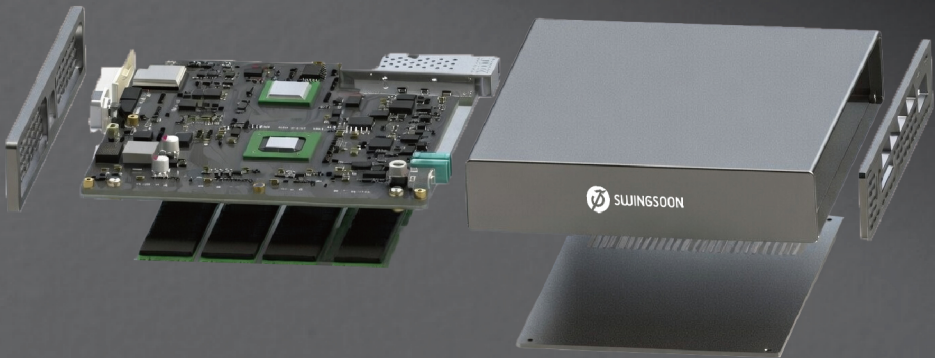




游隼系列 GP SPARK

General Direct Data Stack
G-DDS，通用直连数据栈

产品概述

GP Spark 是绿算技术专为高性能 AI 桌面计算打造的极速外置存储设备，核心定位为 NVIDIA DGX Spark 等 AI 开发主机的专属极速存储伴侣。它在仅 0.58 升（150mm x 150mm x 26mm）的迷你机身中，通过硬件卸载引擎将 RDMA 和 NVMe-oF 协议栈集成于芯片，支持 100GbE 网络和 GPUDirect RDMA 技术，实现数据直通 GPU 内存，提供 20μs 超低延迟、12GB/s 吞吐量 and 高达 270 万 IOPS 的极致性能，同时整机功耗控制在 100W 以内，铝镁合金机身被动散热、运行静音。它内置 4 个 M.2 NVMe SSD 插槽，可灵活扩展存储资源，彻底解决 AI 训练与推理中的“数据喂不饱算力”瓶颈——无论是加载 200B 参数的大模型，还是进行 70B 模型微调，都能实现权重加载与推理计算的流水线零等待，让个人开发者在桌面上就能拥有接近云端集群的存储体验，真正实现从数据准备到模型训练推理的桌面级 AI 闭环。

核心特性

极速扩展，容量无忧

- 单机支持4块NVMeSSD，最大单盘容量16TB，支持容量升级；
- 灵活搭配硬盘，满足不同存储需求。

无缝对接，即插即用

- 与Spark内核级无缝兼容，无需驱动，支持NVMe-OF、RDMA
- 专为DGX Spark优化，识别迅速、稳定可靠、支持GDS（GPU Direct storage）

桌面级性能王者

- 最大吞吐12GB/S
- 存取延迟<20μs
- 整机功耗<100W，能效比出众

便携设计，轻松部署

- 桌面级尺寸，不占空间
- 即插即用，快速上线，告别复杂配置

技术规格

项目	参数
最大支持	4块NVMe SSD
SSD规格	M.2 2280 22110
传输带宽	双向200GbE
存取延迟	<20μs
整机功耗	<100W
互联方式	支持RDMA高速网络、100GbE
协议	NVMe-OF、GDS(GPU Direct Storage)
兼容系统	NVIDIA DGx Spark、DGx Station、图形工作站、算力服务器
尺寸规格	150*150*26mm

核心优势

对比维度	游隼 GP spark	USB硬盘	NAS存储	雷电5
延迟	20μs	5-10ms	2-5ms	100μs
双向带宽	200Gbps	5.0Gbps	5.0Gbps	120Gbps
Spark兼容性	原生支持	需驱动	需NFS挂载	不支持
多机共享	200Gbps集群	不支持	支持(慢)	不支持
主机CPU资源占用	否	是	是	是

应用场景

场景一:AI大模型+RAG知识库构建

- 解决原生DGX Spark存储不足(仅1TB/4TB)的痛点
- 快速扩展存储空间，支持海量知识库数据缓存与训练

场景二:高速数据备份与迁移

- 通过RDMA高速网络互联，实现服务器与DGX Spark之间的极速数据传输
- 支持多设备集群快速同步，提升团队协作效率

场景三:KV Cache 外部化加速

- 提供高达200Gb带宽连接，实现KV Cache高速卸载
- 释放DGX Spark算力，提升整体训练与推理效率