

# ZK-DPU WS7000 (WS-HBMM7000)

## STORAGE ACCELERATION TECHNOLOGY WHITEPAPER

### Disaggregated All-Flash Acceleration Storage for AI Training & Inference

Make every GPU count · Disaggregated storage · Self-controlled

Item	Details
Publisher	Hong Kong Zhongke Hangxing Technology Co., Limited
Model	ZK-DPU WS7000 (WS-HBMM7000) all-flash acceleration storage platform
Headline spec	70M IOPS · 300 GB/s (2.4 Tbps) · 20 $\mu$ s latency (vendor spec)
Form / interface	24-bay NVMe · active-active controllers · 6 $\times$ PCIe 5.0 · up to 12 $\times$ 200GbE
Reference class	System-level all-flash platform comparable to GP7000-class
Third-party evidence	Same-architecture WS5000 independently tested by Beijing Information Science and Technology University (Huawei Ascend Atlas 910B): peak 85.2 $\times$ speedup, median 90.9% reduction across 7 metrics

This whitepaper is an external technical communication. Figures marked "vendor spec" are vendor design/specification values for WS7000; measured data are from the same-architecture WS5000 independent test report; the Alibaba Cloud CIPU / T-Head comparison figures are taken from public sources and cited item by item. All metrics are subject to the final delivery documents and third-party reports.

## Chapter 1

# About Zhongke Hangxing

Heir to an innovation gene — from precision electronics to AI storage infrastructure

# 1 About Zhongke Hangxing

---

**Hong Kong Zhongke Hangxing Technology Co., Limited** (brand: **ZK-DPU**) is the group's outward-facing entity for international business and partnerships. Research, development and manufacturing are carried out together with its affiliate **Shenzhen Zhongke Hangxing Technology Co., Ltd.**, whose founding team has worked in electronics manufacturing since **1996** — nearly three decades — building full-chain capability across semiconductor R&D, intelligent-terminal production and systems solutions. With **disaggregated storage** as its core architecture, the group delivers high-bandwidth, low-latency, self-controlled all-flash accelerated storage for AI training and inference, solving the era's pain point of **GPUs that wait on data**.

## 1.1 Product family and R&D heritage

The flagship line is branded **ZK-DPU**: the already mass-produced **ZK-DPU WS5000 (WS-HBMM5000)** targets mainstream AI nodes; the **ZK-DPU WS7000 (WS-HBMM7000)** described in this whitepaper is a higher-density, higher-bandwidth, system-level all-flash storage platform for ultra-large AI clusters, specified to a GP7000-class level. The line has been developed over roughly **10 years** with cumulative investment of about **RMB 1 billion**, and has passed independent third-party validation by a national university.

**~10 yrs**

**Sustained R&D**

Disaggregated storage /  
fast interconnect

**~RMB 1B**

**Cumulative R&D**

**spend**

Company historical  
investment

**85.2×**

**Peak load speedup**

Same-architecture  
WS5000, third-party

**90.9%**

**Median reduction**

7 metrics vs NFS baseline

### **Ecosystem & manufacturing certainty**

**Independent validation** — Beijing Information Science and Technology University completed same-architecture WS5000 testing on Huawei Ascend Atlas 910B; **manufacturing** — a volume-production preliminary agreement with Luxshare Precision provides scale delivery capability; **self-controlled** — deeply adapted to domestic compute platforms such as Huawei Ascend, with data kept in-domain.

## Chapter 2

# WS7000 (WS-HBMM7000)

## Series — Product Overview

Disaggregated · all-flash · a system-level storage platform for AI training & inference

## 2 ZK-DPU WS7000 (WS-HBMM7000) Product

### Overview

**ZK-DPU WS7000 (WS-HBMM7000)** is a disaggregated all-flash acceleration storage platform for ultra-large AI training and inference. With active-active controllers, a single system delivers up to **70M IOPS**, **300 GB/s (2.4 Tbps)** aggregate throughput and **20  $\mu$ s**-class access latency (**vendor spec**). NVMe-oF / RDMA / RoCEv2 with GPUDirect Storage connect the all-flash pool straight to GPU memory, removing the "GPU waits on data" bottleneck at its root.

**70M**

**Random IOPS**

Vendor spec · system  
aggregate

**300 GB/s**

**Aggregate  
throughput**

2.4 Tbps bandwidth

**20  $\mu$ s**

**Access latency**

Vendor spec

**24-bay**

**NVMe all-flash**

Active-active · PCIe 5.0

#### 2.1 Seven product highlights

- **Highlight 1 — Extreme performance:** system-level 70M IOPS, 300 GB/s throughput and 20  $\mu$ s latency, meeting the extreme I/O of training checkpoints and inference KV-cache.
- **Highlight 2 — Disaggregated storage:** storage and compute scale independently; GPUs are no longer limited by local-disk capacity, avoiding the resource mismatch of buying compute just to get storage.
- **Highlight 3 — All-flash EBOF:** 24-bay NVMe U.2 all-flash, up to 250 TB per drive; the EBOF design scales capacity and performance linearly.
- **Highlight 4 — Protocol passthrough:** NVMe-oF / RDMA / RoCEv2 plus GPUDirect Storage let data bypass host CPU and extra copies, reaching GPU memory directly.

- **Highlight 5 — Active-active controllers:** dual controllers with automatic failover keep long training jobs and online inference continuous.
- **Highlight 6 — KV-cache scheduling:** tiered KV-cache offload and scheduling sharply raise effective GPU utilization for long-context and high-frequency multi-model switching.
- **Highlight 7 — Self-controlled:** 6×PCIe 5.0 expansion and up to 12×200GbE networking; deeply adapted to domestic platforms such as Huawei Ascend, data in-domain.

## Chapter 3

# Benchmark: WS7000 vs Alibaba Cloud CIPU

System-level storage platform vs cloud infrastructure offload card — a fair-caliber comparison

## 3 Benchmark: WS7000 vs Alibaba Cloud CIPU (T-Head ecosystem)

### Fair-caliber statement (please read first)

**ZK-DPU WS7000 (WS-HBMM7000)** is a **system-level disaggregated all-flash storage platform** (aggregating 24 NVMe bays); Alibaba **CIPU** is a **per-server cloud infrastructure offload card (smart-NIC DPU)**, and T-Head **Zhenyue 510** is an **enterprise SSD controller chip**. They sit at different system layers, so the IOPS / bandwidth below are representative values in each one's own caliber — **WS7000 is a system-level aggregate; CIPU / Zhenyue are single-card / single-chip**. Absolute numbers must be read with this caliber difference in mind; in real deployments the three are usually **complementary** (CIPU offloads host virtualization/networking, WS7000 provides the shared high-performance storage pool). This chapter illustrates **positioning differences and respective magnitudes**, not a same-caliber benchmark of one device.

### 3.1 Key metric comparison

Metric	ZK-DPU WS7000 (WS-HBMM7000) (system platform)	Alibaba CIPU 1.0 (card)	Alibaba CIPU 2.0 (card)	T-Head Zhenyue 510 (SSD ctrl chip)
Random IOPS	<b>70M</b> (vendor spec, aggregate)	3M (storage)	3.6M (EBS)	3.4M (chip)
Throughput / bandwidth	<b>300 GB/s</b> (2.4 Tbps, aggregate)	200 Gbps ( $\approx$ 25 GB/s, storage)	400 Gb/s ( $\approx$ 50 GB/s, network)	14 GB/s (chip)
Access latency	<b>20 <math>\mu</math>s</b> (vendor spec)	min 30 $\mu$ s (storage)	—	—
Energy efficiency	system all-flash · PUE-optimized	—	—	420K IOPS/Watt (chip)
Positioning	disaggregated all-flash storage platform	cloud infra offload card (smart-NIC)	cloud infra offload card (smart-NIC)	enterprise SSD controller chip
Scaling	6×PCIe 5.0, up to 12×200GbE, independent linear scale	within one server	within one server	within drive
Protocols	NVMe-oF / RDMA / RoCEv2 / GPUDirect	eRDMA / NVMe / virtio	eRDMA / NVMe / virtio	NVMe

Sources: WS7000 is vendor spec; CIPU 1.0 storage 200 Gbps / 3M IOPS / 30  $\mu$ s latency and RDMA 5.5  $\mu$ s network latency are from 阿里云官方技术资料 (与非网转述) ; CIPU 2.0 400 Gb/s bandwidth and 3.6M EBS IOPS, and Zhenyue 510 (3.4M IOPS / 14 GB/s / 420K IOPS-W<sup>-1</sup>) are from 平头哥公开资料 / 公开报道汇总 (博客园) . See the caliber statement above for how to read each value.

### 3.2 Positioning: offload card vs disaggregated storage platform

- **Different system layer:** CIPU is a "one card per server" infrastructure offload device that pushes virtualization, networking and storage access off the host CPU; WS7000 is a "one platform per cluster" storage system exposing an aggregated high-performance all-flash pool. They are inherently at different layers.
- **Different metric caliber:** CIPU/Zhenyue IOPS and bandwidth are **single-card / single-chip** capabilities; WS7000's 70M IOPS and 300 GB/s are **whole-system aggregates** (24 bays + multi-controller + multi-port), so the absolute values are not on the same scale.
- **Complementary, not substitutive:** in large deployments, offload cards like CIPU accelerate host-side networking/virtualization/storage clients, while storage platforms like WS7000 provide back-end shared storage and GPU passthrough — together they deliver both "host offload" and "storage pooling" benefits.
- **Substitution scenario:** when the need is "build a high-bandwidth, low-latency, independently scalable, self-controlled shared storage foundation for a GPU cluster", a system-level disaggregated platform like WS7000 is the more direct choice; a pure offload card cannot, on its own, provide storage pooling and cross-node scaling.

### 3.3 Third-party evidence (same-architecture WS5000)

WS7000 and the mass-produced WS5000 share the **same disaggregated architecture and NVMe-oF/RDMA stack**. WS5000's independent third-party test at Beijing Information Science and Technology University on Huawei Ascend Atlas 910B against an NFS baseline provides first-hand evidence for the architecture (**this is WS5000's measurement, not WS7000's own**):

Measured item (same-arch WS5000)	WS5000	NFS baseline	Speedup / reduction
DeepSeek-32B model load	6.62 s	563.85 s	<b>85.17×</b> (98.83% lower)
Training model load	12.72 s	140.23 s	11.02× (90.9% lower)
Effective token output (40 switches/day)	99.1% util	21.7% util	<b>+356.9%</b>

## Chapter 4

# Technical Architecture

NVMe-oF · RDMA / RoCEv2 · GPUDirect · all-flash EBOF · active-active controllers

## 4 Technical Architecture

---

**ZK-DPU WS7000 (WS-HBMM7000)** is built around disaggregated storage: a fast lossless network decouples the all-flash pool from GPU compute, and protocol passthrough minimizes the data path. Core elements:

### 4.1 Disaggregation and all-flash EBOF

An EBOF (Ethernet-Bunch-of-Flash) all-flash chassis serves 24 NVMe U.2 SSDs through the controllers; capacity and bandwidth scale linearly with bays/nodes, decoupled from GPU compute, eliminating the waste of buying compute just to add capacity.

### 4.2 NVMe-oF / RDMA / RoCEv2 lossless network

Up to 12×200GbE front-end carries NVMe-oF over RDMA (RoCEv2), delivering near-local-NVMe remote latency on lossless Ethernet and avoiding TCP/IP stack and memory-copy overheads.

### 4.3 GPUDirect Storage passthrough

GPUDirect Storage lets data DMA straight from the all-flash pool into GPU memory, bypassing the host CPU and bounce buffers and sharply cutting weight / checkpoint load latency.

## 4.4 Tiered KV-cache scheduling

For inference, tiered KV-cache offload and scheduling push long-context / session caches down to the all-flash pool, relieving GPU memory pressure and raising effective utilization under long context and high-frequency model switching.

## 4.5 Active-active controllers and PCIe 5.0 expansion

Active-active controllers provide automatic failover and load balancing; 6×PCIe 5.0 x16 slots (backward compatible with 4.0/3.0) support fast interconnect and future evolution.

## Chapter 5

# Technical Specifications

WS7000 (WS-HBMM7000) system-level all-flash acceleration storage platform  
· vendor-spec caliber

## 5 Technical Specifications

Specification	ZK-DPU WS7000 (WS-HBMM7000) (vendor spec)
Model	WS7000 (WS-HBMM7000)
Form factor	system-level disaggregated all-flash storage platform (active-active)
Random IOPS	up to 70M
Aggregate throughput	300 GB/s (2.4 Tbps)
Access latency	20 $\mu$ s-class
Drive bays	24 × NVMe U.2 SSD (single/dual-port)
Max drive capacity	250 TB
Expansion slots	6 × PCIe 5.0 x16 (backward compatible 4.0/3.0/2.0/1.0)
Front-end network	up to 12 × 200GbE
Storage protocols	NVMe-oF / RDMA / RoCEv2 / GPUDirect Storage
High availability	active-active controllers, automatic failover
Domestic adaptation	deeply adapted to platforms such as Huawei Ascend, data in-domain

### Caliber note

Values marked "vendor spec" are WS7000 (WS-HBMM7000)'s vendor design/specification figures (specified to a GP7000-class level), for positioning and selection reference; measured performance is subject to third-party reports and final delivery documents (see Chapter 6).

## Chapter 6

# Third-party Validation

Beijing Information Science and Technology University · Huawei Ascend Atlas  
910B · NFS baseline (same-architecture WS5000)

## 6 Third-party Validation (same-architecture WS5000)

### Data caliber

The data in this chapter are from [same-architecture WS5000 \(WS-HBMM5000\)](#) at Beijing Information Science and Technology University on Huawei Ascend Atlas 910B, against an NFS (10GbE) baseline, over NVMe-oF over RDMA/RoCE (2×200GbE, 50 GB/s line rate). WS7000 shares WS5000's architecture and stack, so these are cited as architectural evidence — **not WS7000's own measurement.**

### 6.1 Inference: model load and service speedup

Model	WS5000 load(s)	NFS load(s)	Load speedup	Service speedup
DeepSeek-32B	6.62	563.85	85.17×	6.17×
DeepSeek-70B	35.38	1284.66	36.31×	9.33×

### 6.2 Training: model and checkpoint I/O

Measured item	WS5000(s)	NFS(s)	Speedup	Reduction
Model load	12.72	140.23	11.02×	90.9%
Model save	31.16	165.87	5.32×	81.2%
Checkpoint load	10.55	131.37	12.45×	92.0%
Checkpoint save	81.94	451.14	5.51×	81.8%

## 6.3 Effective token output (high-frequency model switching)

Switches/day	WS5000 util	NFS util	Effective output gain
10	99.8%	80.4%	+24.1%
20	99.5%	60.8%	+63.6%
40	99.1%	21.7%	+356.9%

### Test conclusion

Across 7 metrics the **median reduction vs NFS is 90.9%**, peak load speedup **85.2×**, and effective token output improves up to **356.9%** under high-frequency switching — the performance of the same disaggregated + all-flash + protocol-passthrough architecture, independently endorsed by a national university.

## Chapter 7

# Application Scenarios

LLM training · inference serving · HPC · AI compute centers · finance · simulation

## 7 Application Scenarios

---

- **LLM training clusters:** accelerate model-weight and checkpoint I/O, shorten training iterations and resume-from-checkpoint, and cut idle waiting on expensive GPUs.
- **LLM inference serving:** for long context and high-frequency multi-model switching, tiered KV-cache scheduling sharply raises effective GPU utilization and concurrent throughput.
- **HPC:** high-bandwidth, low-latency shared storage for scientific computing, genome sequencing and CFD simulation.
- **AI compute centers / domestic stacks:** disaggregation plus deep domestic-GPU adaptation underpin self-controlled city- / industry-level AI infrastructure.
- **Finance & data compliance:** in-domain data, low latency and high concurrency for risk control, quant and real-time analytics under strict compliance.
- **Industrial simulation & autonomous driving:** high-throughput supply of massive training samples and simulation data to speed the data loop and model iteration.
- **Brownfield data-center retrofit:** speed up without replacing GPUs or stopping service, reactivating existing compute assets and raising per-node output.

## 8 Contact

Item	Details
Publisher	Hong Kong Zhongke Hangxing Technology Co., Limited
R&D / manufacturing affiliate	Shenzhen Zhongke Hangxing Technology Co., Ltd.
Brand / model	ZK-DPU · WS7000 (WS-HBMM7000) all-flash acceleration storage platform
Registered office	Hong Kong SAR, China — registered office to be published upon completion of registration
R&D base	Room 302, Building 3, Ship Front Plaza, Sea World, Nanshan District, Shenzhen, China
Contact	Lisa CHEN (CEO)
Tel	+86 138 2372 8880
Email	13823728880@139.com
Partnership	evaluation units / joint validation / volume delivery / ecosystem

We welcome compute centers, model teams and industry partners to engage with us and scale ZK-DPU WS7000 (WS-HBMM7000) across AI compute infrastructure.