Hong Kong Zhongke Hangxing Technology Co., Limited

COMPANY PROFILE

ZK-Storage WS5000 • All-Flash Accelerated Storage for AI

Make every GPU earn its keep · Disaggregated storage · Self-controlled

Item	Detail
Entity	Hong Kong Zhongke Hangxing Technology Co., Limited (香港中科航星科技有限公司)
R&D & manufacturing	Shenzhen Zhongke Hangxing Technology Co., Ltd.
Contact	Lisa CHEN (CEO)
Flagship product	ZK-Storage WS5000 (WS-HBMM5000) all-flash accelerated storage
Stage	Mature — product finalized and in mass production; independently validated
R&D heritage	~10 years of R&D, ~RMB 1000M cumulative investment
Core performance	300 GB/s bandwidth · 50M IOPS · 20 µs latency
Independent test	Beijing Information Science and Technology University on Huawei Ascend Atlas 910B; 7-metric median latency reduction 90.9%
Manufacturing	Luxshare Precision foundry, ~1,000 units/month; 2 demo units in stock for immediate PoC

This profile is an external communication document. Product specifications and third-party test figures are drawn from an independent test report; forward-looking figures are modeled and detailed in the accompanying Business Plan and Feasibility Study. Hong Kong registration details will be published upon completion.

PART 01

Company Overview

From precision electronics manufacturing to AI storage infrastructure

1 Company Overview

ZK-Storage is the brand of **Hong Kong Zhongke Hangxing Technology Co., Limited** (香港中科航星科技有限公司), 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.

Guided by "**innovation-driven, quality-first**," the group focuses on **AI compute infrastructure**. With **disaggregated storage** as its core architecture, it 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.

The flagship line, **ZK-Storage**, centers on **ZK-Storage WS5000 (WS-HBMM5000)**, developed over about **10 years** with about **RMB 1000M** of cumulative investment. It is now **finalized and in mass production**, has passed independent third-party validation by a national university, and is backed by a volume-manufacturing partnership — a complete engineering loop from demo-unit testing to volume supply.

1.1 Maturity at a glance

~10 yrs
Sustained R&D
Technology, product and manufacturing risk materially retired

~RMB 1000M
Cumulative R&D
Group's self-funded history

1,000/mo
Production capacity
Luxshare Precision foundry, within a month of order

2 units
Demo stock
Ready for immediate PoC, shortening validation

From concept to maturity: four certainties

Technology — Beijing Information Science and Technology University completed an independent test on the Huawei Ascend Atlas 910B platform; **Product** — WS5000 is finalized and in mass production; **Manufacturing** — a pre-production agreement with Luxshare Precision, ~1,000 units/month; **Ecosystem** — AMD and xFusion platform adaptation tests are in progress (subject to final reports).

1.2 Milestones and development stages

Stage	Milestone
Technology	~10 years of disaggregation and high-speed interconnect; ~RMB 1000M invested
Productization	WS5000 (WS-HBMM5000) finalized; hardware and software stack mature
Validation	Independent test by Beijing Information Science and Technology University; leading the NFS baseline on all 7 metrics
Manufacturing	Pre-production agreement with Luxshare Precision; ~1,000 units/month; 2 demo units in stock
Ecosystem	AMD and xFusion platform adaptation in testing; deep tuning for domestic accelerators

PART 02

Core Product & Technology

ZK-Storage WS5000 all-flash storage · Disaggregated architecture · Self-controlled

2 Flagship Product: ZK-Storage WS5000

ZK-Storage WS5000 is a high-performance all-flash accelerated storage appliance for AI training and inference. Through a disaggregated architecture and an end-to-end high-speed data path, it frees GPU clusters from waiting on data — markedly raising effective compute utilization and slashing total cost of ownership, with no changes to the upper-layer framework.

2.1 Core specifications

300 GB/s **Aggregate bandwidth** Line-rate data path	**50M** **Random IOPS** Friendly to high-concurrency small files	**20 µs** **Access latency** Microsecond response	**90%+** **GPU coverage** Broad mainstream accelerator support
48-72 h **Fast deployment** Turnkey; live within a day	**-40%** **Total cost** vs. mainstream 3-year TCO	**-60%** **Scale-out cost** Elastic, on-demand expansion	**2-3×** **GPU utilization** High-switch / long-context cases

2.2 Product portfolio

Product / Service	Form	Customer	Core value
ZK-Storage WS5000 appliance	Hardware	New AI clusters	High-bandwidth all-flash, turnkey
ZK-Storage storage software	Subscription	Existing-hardware customers	Disaggregation, continuous updates
Brownfield retrofit	Solution + service	Existing data centers	Speed-up without downtime
Accelerated storage service	Capacity / compute	SMB / cloud	On-demand, low barrier

Why storage?

In the LLM era, simply stacking more GPUs yields rapidly diminishing returns; the real bottleneck is on the data-supply side — model loading, checkpoint I/O and KV-cache scheduling. ZK-Storage upgrades storage from a supporting act into a compute amplifier, cutting KV-cache-related cost by about **74%** in testing.

3 Technology & Architecture

The core technology route is **disaggregation**: storage media are decoupled from compute nodes and pooled into an independently scalable all-flash pool, linked to the GPU compute pool over a high-speed lossless fabric. Compute and capacity scale independently and elastically, with pooled, efficiently shared resources.

- **NVMe-oF over RDMA/RoCE**: carry the NVMe protocol over remote direct memory access, bypassing redundant copies to approach local-disk performance.
- **GPUDirect**: data moves directly between storage and GPU memory, shortening the path and cutting CPU and latency overhead.
- **All-flash EBOF**: a controller-less, high-density flash pool whose bandwidth and IOPS scale near-linearly with capacity, at lower power.
- **KV-cache scheduling**: offload and reuse KV cache for long-context / high-switch inference, markedly lifting effective GPU utilization.
- **Self-controlled adaptation**: deeply tuned for Huawei Ascend and domestic accelerators, with 90%+ mainstream GPU coverage.

PART 03

Independent Third-Party Validation

Beijing Information Science and Technology University · Huawei Ascend Atlas 910B · leading on all 7 metrics

4 Independent Third-Party Validation

To objectively verify product performance, the group commissioned **Beijing Information Science and Technology University** to run an independent test on the **Huawei Ascend Atlas 910B** platform, against an **NFS network storage (NFS over TCP, 10GbE, ~1.25 GB/s)** baseline; the ZK-Storage side used a **NVMe-oF over RDMA/RoCE (2×200GbE, ~50 GB/s line rate)** high-speed data path. The test covered inference loading, training weight I/O and token throughput across **7 metrics**, with a median latency reduction of **90.9%** (lower is better) — leading the baseline on every metric.

4.1 Inference: model load and service speedup

Model	ZK-Storage load	NFS load	Load speedup	Latency cut	Service speedup
DeepSeek-32B	6.62 s	563.85 s	85.17×	98.83%	6.17×
DeepSeek-70B	35.38 s	1284.66 s	36.31×	97.25%	9.33×

4.2 Training: weights and checkpoint I/O

Test	ZK-Storage	NFS baseline	Speedup	Latency cut
Model load	12.72 s	140.23 s	11.02×	90.93%
Model save	31.16 s	165.87 s	5.32×	81.21%
Checkpoint load	10.55 s	131.37 s	12.45×	91.97%
Checkpoint save	81.94 s	451.14 s	5.51×	81.84%

4.3 Inference token throughput (= effective GPU utilization)

Switch frequency	ZK-Storage util.	NFS util.	Relative gain
10/day	99.8%	80.4%	+24.1%
20/day	99.5%	60.8%	+63.6%
40/day	**99.1%**	**21.7%**	**+356.9%**

Conclusion

In Beijing Information Science and Technology University's independent test, ZK-Storage WS5000 reached up to **~85× peak model-load speedup**, **5–12× training I/O speedup** and up to **+357% token efficiency**; the 7-metric median latency reduction was **90.9%** — providing reproducible, verifiable third-party endorsement of product performance.

PART 04

Ecosystem • Value • Future

Alliances + unit economics + global roadmap

5 Alliances and Industrial Ecosystem

- **Manufacturing — Luxshare Precision**: a pre-production agreement leverages its precision manufacturing and supply chain to deliver about 1,000 units within a month of order.
- **Validation — Beijing Information Science and Technology University**: completed an independent test on the Huawei Ascend platform, providing national-university endorsement of performance.
- **Adaptation — AMD / xFusion (in testing)**: platform adaptation tests with AMD and xFusion are in progress (work in progress; subject to final reports).
- **Self-control — Ascend ecosystem**: deep tuning for Huawei Ascend and domestic accelerators, aligned with the drive for self-controlled infrastructure.

6 Customer Value and Unit Economics

The group is anchored on "creating quantifiable value for customers." Against mainstream high-performance storage, **ZK-Storage** offers a clear three-year total-cost-of-ownership (TCO) advantage, and amplifies a customer's compute ROI by raising effective GPU utilization.

Option	3-year TCO (US$M)	vs. baseline
Industry baseline (high-end)	241	100%
ZK-Storage WS5000	144.6	-40%
3-year saving	96.4	—

Unit economics (per-system, modeled)

Per-system blended revenue is about **RMB 3940K** (hardware + software + O&M + compute service), at about a **49%** blended gross margin; combined with a **2.5×** GPU-utilization uplift, the customer's effective compute ROI is further amplified.

7 Applications and Market Network

The group advances across four market layers — **domestic greenfield, domestic brownfield retrofit, overseas greenfield, and overseas brownfield retrofit** — first proving scalable, repeatable delivery in domestic greenfield and retrofit scenarios, then extending overseas.

- **LLM training clusters**: accelerate model loading and checkpoint I/O to shorten training iterations.
- **LLM inference serving**: long-context and high-frequency multi-model switching — markedly higher effective GPU utilization.
- **AI centers / domestic stack**: disaggregation plus domestic adaptation for sovereign, self-controlled infrastructure.
- **Brownfield retrofit**: no GPU swap, no downtime — revive idle compute assets in place.

8 Future Blueprint and Roadmap

The group will keep advancing product iteration and market ramp with engineering pragmatism: near term, scale domestic benchmark customers and retrofit replication; mid term, complete multi-platform adaptation and round out the portfolio; long term, build a global AI storage infrastructure network.

Year	2026	2027	2028	2029	2030
Revenue (RMB 100M)	0.6	3.3	8.9	19.3	37.0

Note on figures

The 2026–2030 revenue above is a **modeled projection** (see the accompanying Business Plan), reflecting scaling potential given a finalized, mass-produced product, completed third-party validation and a manufacturing partnership; the 2030 figure is about **RMB 3700M**. Actual results are subject to final disclosure.

9 Contact

Item	Detail
Entity	Hong Kong Zhongke Hangxing Technology Co., Limited (香港中科航星科技有限公司)
R&D & manufacturing	Shenzhen Zhongke Hangxing Technology Co., Ltd.
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)
Phone	+86 138 2372 8880
Email	13823728880@139.com
Focus	ZK-Storage all-flash accelerated storage · AI compute infrastructure
Partnership	PoC units / joint validation / volume delivery / AI-center co-build

We welcome AI centers, model teams and industry partners to engage and scale ZK-Storage in AI compute infrastructure together.