China Internet

Hyperscalers Moving Into AI Chips: Custom Silicon Gains Traction, Compute Strategies Diverge

Highlights

- Chinese cloud and LLM companies are increasingly moving into chip development, with the strategic goals of reducing compute costs, securing supply, and improving model-hardware co-optimisation. The economic benefits ultimately show up in lower cost per token and better returns on AI capex.
- Maintain MARKET WEIGHT. Top BUYs: Alibaba and Z.AI.

Analysis

- **Custom AI silicon is gaining traction,** with players taking different routes to secure compute through proprietary chip development, strategic partnerships and multi-vendor procurement. Alibaba and Baidu are building more vertically integrated stacks, Tencent combines Zixiao with strategic exposure to Enflame, while Z.AI is moving upstream through greater control of domestic compute infrastructure and MiniMax remains more reliant on external multi-cloud capacity.
- **Alibaba is moving towards a more vertically-integrated AI compute stack,** spanning accelerators, networking, rack-scale systems and software. Alibaba unveiled the Zhenwu M890 in May 26 as a unified training-and-inference accelerator, with 144GB of high-bandwidth memory and 800GB/s of inter-chip bandwidth. The accompanying ICN Switch 1.0 enables a 64-accelerator full-bandwidth domain, while the Panjiu AL128 connects 128 M890 chips in a rack-scale supernode. T-Head plans to release the V900 in 3Q27, targeting roughly 3x M890 performance, and the J900 in 3Q28. We think the broader significance lies in Alibaba's ability to co-optimise chips, networking and cloud infrastructure instead of relying solely on standalone accelerator performance.
- **Zhenwu adoption is broadening, although external shipment and monetisation visibility remain limited.** T-Head has emerged as China's second-largest domestic AI accelerator supplier after Huawei. It had cumulatively shipped more than 560,000 Zhenwu chips as of Apr 26, of which more than 60% were serving external customers through Alibaba Cloud, to over 650 customers across more than 20 industries. In Mar 26, Eddie Wu disclosed that T-Head's annualised revenue exceeded Rmb10b. Alibaba has gradually opened SAIL, its AI accelerator software stack, supporting more than 260 training and inference frameworks including PyTorch, TensorFlow, vLLM and SGLang. Better compatibility should lower migration costs and support wider Zhenwu adoption, although SAIL remains under development and is not yet a proven compute unified device architecture substitute across all workloads.

Analyst(s)

Julia Pan
juliapan@uobkh.com
+86 21 5404 7225 ext 808

MARKET WEIGHT (Maintained)

Segmental Rating

Segment	Rating
E-commerce	MARKET WEIGHT (Maintained)
Online Game	OVERWEIGHT (Maintained)
OTA	OVERWEIGHT (Maintained)
AI	OVERWEIGHT (Maintained)
Online Property Vertical	MARKET WEIGHT (Maintained)
Local Life Service	UNDERWEIGHT (Maintained)
E-commerce	MARKET WEIGHT (Maintained)

Source: UOB Kay Hian

Sector Picks

Company	Ticker	Rec	Share Price (HK$)	Target Price (HK$)
Z.AI	2513 HK	BUY	769.50	1,400.00
Alibaba	9988 HK	BUY	108.80	188.00

Source: Bloomberg, UOB Kay Hian

Peer Comparison

Company	Ticker	Price @ 18 Sep 26 (lcy)	Target Price (lcy))	Upside To TP (%)	Market Cap (lcy m)	-----Calendarised PE----- 2026F (x)	2027F (x)	2028F (x)	Calendarised EV/EBITDA 2026F (x)	2027F (x)	2028F (x)	Calendarised EV/Sales 2026F (x)	2027F (x)	2028F (x)	---------- ROE ---------- 2026F (%)	2027F (%)	2028F (%)
China																	
Z AI Co Ltd	2513 HK	769.50	1,400.00	82%	375,443	n.a	n.a	n.a	n.a	n.a	n.a	15.5	5.2	2.2	n.a	n.a	n.a
Minimax Group Inc	100 HK	278.60	470.00	69%	97,367	n.a	n.a	n.a	n.a	n.a	n.a	19.8	7.7	3.5	n.a	n.a	n.a
Alibaba Group Holding Ltd	9988 HK	108.80	188.00	73%	2,165,475	17.2	12.4	9.4	11.8	8.9	6.8	1.3	1.1	1.0	8.2	11.1	13.2
Tencent Holdings Ltd	700 HK	426.00	608.00	43%	3,874,507	12.4	11.7	10.7	10.2	9.6	8.8	4.1	3.8	3.4	18.7	17.1	16.5
Baidu Inc	9888 HK	89.40	136.00	52%	244,380	15.8	12.9	11.0	10.1	10.0	8.3	0.8	0.8	0.7	4.2	4.8	5.4
Avg						15.1	12.4	10.4	10.7	9.5	8.0	8.3	3.7	2.2	10.3	11.0	11.7
US																	
Alphabet Inc	GOOG US	343.70	UR	n.a	4,226,148	18.0	21.5	17.5	18.2	14.0	10.9	9.5	7.5	6.2	59.9	42.6	32.8
Meta Platforms Inc	META US	682.30	UR	n.a	1,738,189	19.6	17.9	14.5	14.4	11.5	9.1	6.9	5.8	4.9	29.9	24.1	21.1
Microsoft Corp	MSFT US	497.80	UR	n.a	3,696,065	25.1	21.1	17.4	15.4	12.6	10.2	9.6	8.0	6.6	28.5	26.7	25.5
Avg						20.9	20.2	16.5	16.0	12.7	10.1	8.7	7.1	5.9	39.4	31.1	26.5

Source: Bloomberg, UOB Kay Hian

- **Baidu is scaling Kunlunxin from standalone accelerators towards larger-scale AI systems and broader model compatibility.** Kunlunxin has expanded support for domestic foundation models, including Kimi K3 and GLM-5.3-Flash, while the latest M100 is optimised for large-scale inference and the M300 is next on its roadmap. Baidu has also deployed the Tianchi 256 supernode and plans to introduce larger 512- and 1,024-card systems in early-27. We think wider model compatibility and larger supernodes should broaden Kunlunxin's addressable workloads across inference and increasingly large-scale AI compute.
- **Kunlunxin's proposed listing** could provide additional funding for its chip and system roadmap while establishing a standalone valuation for Baidu's AI-silicon asset. Kunlunxin confidentially submitted its HKEX A1 application in Jan 26 and subsequently started STAR Market IPO counselling in May 26. Investor engagement has also progressed ahead of the proposed listing, with a reported target valuation of US$50b (Rmb340b). Management reiterated during the Aug 26 earnings call that the listing process remains ongoing, with further updates to be provided when available.
- **Tencent is taking a more partnership-led approach to domestic compute alongside its proprietary Zixiao programme.** Tencent introduced its Zixiao AI inference chip in 2021, although disclosure on the latest proprietary roadmap and deployment scale remains limited. In practice, Tencent's domestic compute strategy is also closely linked with Enflame, with the two companies cooperating since 2019 on chip deployment, model optimisation and software adaptation. Tencent remained Enflame's largest individual shareholder with a 18 % post-IPO stake, while Tencent-linked sales accounted for 84% of Enflame's 2025 revenue. Enflame raised approximately Rmb6.12b in its Sep 26 STAR Market IPO, including funding for fifth- and sixth-generation AI chips and hardware-software development. We think the relationship provides Tencent with an additional domestic-compute option as Hunyuan, Yuanbao and agent workloads scale, without requiring full ownership of the underlying chip supplier.
- **Large hyperscalers are therefore moving towards silicon ownership,** while independent LLM companies are currently more focused on silicon optimisation. For companies such as Z.ai, MiniMax and DeepSeek, it is generally more economical today to optimise models, kernels and compilers for multiple GPU/NPU platforms rather than build a complete proprietary chip stack from scratch.

Custom-Silicon Strategies Vary Across China AI Players

Examples	Model	How it works	Primary use
Alibaba / T-Head	In-house fabless design	Develops architecture, IP, chip design and software stack internally, while outsourcing fabrication and packaging	Mainly internal Alibaba Cloud workloads
Baidu → Kunlunxin	Internal incubation → independent chip company	Originated as an internal chip team before being spun out and bringing in external capital/customers	Baidu workloads + external chip sales
Tencent	Proprietary ASIC + multi-chip ecosystem	Develops selected proprietary accelerators while continuing to use third-party GPUs/NPUs and domestic chips	Internal workloads and Tencent Cloud
ByteDance and potentially other large model companies	Co-design with semiconductor partners	Defines workloads, architecture and specifications internally, while relying on design partners for physical design, SerDes and tape-out	Large-scale internal inference

Source: Respective companies, UOB Kay Hian

- **Z.AI is moving further upstream into compute infrastructure** as additional capacity eases near-term compute constraints. The company now uses domestic chips as a primary inference resource and operates a production-grade inference system across more than 100,000 Chinese-made AI accelerators, which serves all GLM-5.3-Flash production inference. Infrastructure optimisation has improved end-to-end serving throughput by

around 3x, while Z.AI reported an 80% decline in unit token inference costs since the beginning of 2026. Z.AI has also reportedly approached domestic chip designers regarding a customised processor optimised for its models, although discussions remain preliminary. Alongside its 1GW AI data centre buildout, the latest US$5b financing could support close to another 100,000 AI accelerators, with around 40% allocated to training and 60% to inference, providing greater headroom for next-generation model training and commercial scaling. We think improving domestic compute availability and greater control over infrastructure should ease compute constraints and support inference economics, although greater vertical integration also raises capital intensity.

- **DeepSeek is reportedly exploring custom inference silicon as model usage scales.** The company has reportedly been developing an inference-focused AI chip for around a year and has engaged chip-design houses, foundries and memory suppliers, although the project remains at an early stage. We think the direction is consistent with broader industry efforts to improve inference economics through deeper model-hardware co-optimisation, but current disclosure remains insufficient to assess specifications, production timing or deployment scale.
- **MiniMax remains primarily reliant on externally procured multi-cloud compute** while broadening hardware portability across domestic accelerator platforms. Alibaba Cloud remains its clearest disclosed anchor provider, with MiniMax increasing its purchase caps to US$300m/US$400m/US$500m for 2026-28, respectively, or US$1.2b in aggregate, following faster-than-expected growth in training and inference demand. MiniMax has also adapted M3 to Huawei Cloud's Ascend-based CloudMatrix and multiple domestic accelerator platforms. We think this approach provides greater hardware flexibility without requiring MiniMax to own the underlying compute infrastructure, although it leaves the company more dependent on external capacity than vertically integrated peers.
- **OpenAI is moving towards custom silicon for scaled inference** while retaining a heterogeneous compute stack for broader workloads. OpenAI and Broadcom unveiled Jalapeño in Jun 26 as OpenAI's first custom Intelligence Processor for LLM inference, with deployment expected by end-26 and a broader 10GW custom-accelerator programme planned through 2029. Initial InferenceX benchmarks showed 1.5-1.9x higher peak throughput per watt and 1.7-3.6x lower end-to-end latency vs selected Nvidia GB200/GB300 comparison systems. We view Jalapeño primarily as an inference cost, power-efficiency and latency optimiser for predictable, high-volume workloads rather than a replacement for Nvidia and AMD GPUs across frontier training.
- **Anthropic is pursuing a diversified multi-architecture compute strategy** rather than relying on a single accelerator platform. AWS remains a major cloud and training partner, while Anthropic has also secured substantial capacity across Google tensor processing unit (TPU), Microsoft Azure, Nvidia and advanced micro devices (AMD) platforms. It has committed to up to 5GW of AWS capacity and 5GW of next-generation TPU capacity from 2027, alongside additional Nvidia and AMD deployments. We think the multi-vendor approach improves supply diversification and allows Anthropic to match workloads with the most suitable architecture, although supporting multiple accelerator stacks also requires deeper workload optimisation and software engineering.
- **Microsoft's Maia 200 provides a more direct proof point for the economics of model-silicon co-design.** Maia 200 is built on TSMC 3nm with 216GB of HBM3E and delivers more than 10 PFLOPS of FP4 performance within a 750W envelope. Microsoft said in its Jul 26 earnings call that Maia 200 was supporting both OpenAI and Microsoft AI models and delivering 30% better performance per dollar than the latest-generation hardware in its fleet. It also said co-designing its own models with Maia 200 delivered 40% better performance/watt for Microsoft AI models. We think this provides evidence that vertically integrating models, software and silicon can materially improve inference economics at sufficient scale.

- **Google is increasingly separating custom silicon by workload as training and inference requirements diverge.** Its eighth-generation TPU roadmap splits into TPU 8i for inference, serving and reinforcement learning, and TPU 8t for large-scale training. Google says TPU 8i delivers up to 80% better inference performance per dollar vs seventh-generation Ironwood, while TPU 8t delivers up to 2.7x better training performance per dollar, with both providing up to 2x better performance/watt. However, both products remain listed as "coming soon", so we would view these performance figures as forward-looking rather than evidence of broad commercial deployment today.
- **AWS demonstrates how anchor customers and software maturity can turn custom silicon into a broader commercial ecosystem.** Trainium3 is commercially available through EC2 Trn3 UltraServers, with AWS reporting up to 4.4x higher performance and 4x better performance/watt vs Trn2 UltraServers. Anthropic has committed to securing up to 5GW of Trainium capacity, while OpenAI has separately committed to approximately 2GW beginning in 2027 across Trainium3 and Trainium4. Amazon said in 2Q26 that its broader chips business, including Trainium, Graviton and Nitro, had exceeded a US$25b annual revenue run rate and was growing at a triple-digit percentage yoy. We think this suggests AWS's custom-silicon strategy is moving beyond an anchor-customer model towards a broader commercial ecosystem, although software portability and utilisation remain important to sustaining adoption.
- **We see custom silicon as increasingly complementary to, rather than an immediate replacement for, merchant GPUs.** Inference should see faster custom-silicon adoption given its more repeatable workloads and greater sensitivity to cost, power and latency, while frontier training is likely to remain more heterogeneous. Among the companies we cover, Alibaba has the broadest disclosed chip-to-system stack, Baidu is scaling Kunlunxin towards larger AI systems, Tencent retains a more partnership-led approach through Enflame, while Z.AI is increasing control over domestic compute infrastructure. We think the key factors to watch are deployment scale, software compatibility, utilisation and whether lower infrastructure costs translate into sustainable improvements in inference economics.

Valuation/Recommendation

- **Alibaba (9988 HK).** Maintain BUY with target price of HK$188.00. Its full-stack AI capabilities, spanning data centres, cloud infrastructure and models, provide greater control over compute capacity and costs, while its model portfolio covers both frontier performance and cost-efficient deployment. Ant Group's focus on trusted agent infrastructure could further broaden secure agent adoption across financial and enterprise use cases.
- **Z.AI (2513 HK).** Maintain BUY with target price of HK$1,400.00. Self-built data centres should provide greater control over costs and compute capacity while reducing reliance on external cloud providers, while the planned trillion-parameter model could help narrow the capability gap with global frontier models.

Sector Catalyst/Risk

- **Catalysts:** a) Broader commercial deployment of domestic AI accelerators, b) improving domestic compute availability, c) greater inference cost and power efficiency, d) stronger software compatibility and model-hardware co-optimisation, and e) expanding external adoption of proprietary AI-chip platforms.
- **Risks** include: a) slower-than-expected domestic chip performance improvement, b) software ecosystem and migration constraints, c) lower utilisation of proprietary infrastructure, d) elevated capex and funding requirements, and e) geopolitical and semiconductor supply-chain restrictions.