RefleXio Intelligence
Investor Brief · Issue No. 01

The AI Chip War

NVIDIA, Huawei and the Geopolitics of Compute

An investor brief on margins, infrastructure, sovereignty and the future of the AI stack.

refleXio

INVESTOR BRIEF

Issue No. 01 · April/ May 2026 · v3.0

The AI Chip War

NVIDIA, Huawei and the geopolitics of compute

An investor brief on margins, infrastructure, sovereignty and the future of the AI stack.

Publication date	April/ May 2026
Publisher	RefleXio Intelligence
Series	RefleXio Intelligence Briefs, Issue No. 01
Version	3.0, which added the methodological appendix, the oversupply scenario (S5), expanded Europe analysis, quantified dashboard thresholds and HBM supply-chain detail
Prepared by	RefleXio Intelligence Editorial Desk

Independent intelligence on AI, compute and the geopolitics of the stack.

Contents

Key numbers at a glance

Nine figures that frame the rest of the brief. Each is sourced and repeated, with context, in the relevant section.

$193.7B NVIDIA FY26 Data Center revenue (+68% YoY)	**$78.0B ±2%** Q1 FY27 guidance, no China DC compute	**$4.5B** H20 charge absorbed in Q1 FY26
41% Chinese domestic share of China's 2025 AI accelerators	**~750k** Huawei Ascend 950PR 2026 shipment ambition	**2.2M** NVIDIA cards still shipped to China in 2025
415 → 945 TWh Global DC electricity demand 2024 → 2030 (IEA)	**45% / 25%** US vs China share of 2024 DC electricity	**280×** Collapse in inference cost Nov-22 → Oct-24 (Stanford AI Index)

As a reading note, fiscal 2026 for NVIDIA ended 25 January 2026. Q1 FY27 guidance is forward-looking. Chinese shipment figures are market-share estimates from IDC, reported via Reuters. The 280× inference-cost collapse refers to the price of running a GPT-3.5-equivalent model on published commercial endpoints between November 2022 and October 2024 (Stanford AI Index 2025).

1. Executive summary

The AI chip war between the United States and China can no longer be read as a linear race for FLOPS. It is a contest between **national compute ecosystems**, where silicon, memory, networking, energy, software, industrial capacity, regulation and adoption all move together. Brookings has recently framed the race as one played across multiple dimensions of compute, models, adoption, integration and deployment.[7]

NVIDIA still leads the technological frontier. Blackwell pushes FP4 with micro-tensor scaling, and the company argues this step doubles throughput and the size of models that fit in memory while preserving accuracy.[2] NVIDIA has also defended, with technical evidence, that formats such as NVFP4 can reach downstream accuracy comparable to BF16 on some training runs, provided recipe and calibration are right.[3] At the FY26 close, NVIDIA unveiled the Rubin platform, a set of six new chips which the company claims deliver up to a 10× reduction in inference token cost versus Blackwell.[13]

However, NVIDIA's advantage can no longer be valued in isolation. According to IDC data reviewed by Reuters, Chinese domestic accelerator makers captured roughly **41%** of China's AI accelerator server market in 2025; NVIDIA kept the lead at **55%** (approximately 2.2 million cards), while Huawei led the domestic bloc with around **812,000** AI chips shipped.[4] In March 2026, Reuters also reported that Huawei's new **Ascend 950PR** was finding traction with ByteDance and Alibaba, with a 2026 shipment ambition of roughly **750,000** units.[5] In April 2026, the picture sharpened further, as DeepSeek's forthcoming V4 model is being optimized to run on Huawei Ascend silicon, with Chinese tech giants reportedly placing large advance orders, which signals that the software–hardware layer is aligning with the domestic stack.12

NVIDIA closed fiscal year 2026 (ended 25 January 2026) with record revenue of **$215.9 billion** (Data Center **$193.7 billion**, +68% year over year) and guided Q1 FY27 revenue at **$78.0 billion ±2%**, assuming no Data Center compute revenue from China.[13] That guidance stance, paired with the May 2025 **$4.5 billion** H20 charge, frames the current China position not as a temporary bump but as an effective foreclosure.[10, 9] The Council on Foreign Relations and CSET both stress that, despite this, Huawei remains a full generation behind on memory subsystem capability, which serves as a reminder that Chinese autonomy is growing from a genuine base, but not a leading one.15, 16

> **One-page takeaways**
>
> • What is happening. Compute is decoupling into two partially isolated national stacks, not just two chip vendors.
>
> • Why it matters. The asset at stake is the standard (software, systems and developer base), not just quarterly revenue.

> • Where the risk sits. NVIDIA's own guidance now treats the Chinese data center market as effectively closed.
> • Where value is created. Memory, advanced packaging, networking, utilities and data-center infrastructure re-rate alongside silicon.
> • Where the margin migrates. Away from pure silicon design toward HBM, advanced packaging, networking and DC infrastructure (see Figure 2).

The bottom line for investors is that the real risk is no longer only whether NVIDIA sells more or fewer high-margin GPUs. It is that China converts a peak-performance disadvantage into a **sovereignty, availability, deployment-scale and regulatory-preference** advantage, which compresses the CUDA ecosystem's value locally and, over time, exports that stack to non-aligned markets in Southeast Asia, the Gulf and parts of Africa.[18]

2. Central thesis

Thesis · Anti-thesis · What the market is underpricing · What to watch

Thesis. The AI chip war is a war for control of the entire compute stack, not for the best individual accelerator. Winners will own the combination of silicon, memory, networking, software and energy, not just the top-line FLOPS.

Anti-thesis. NVIDIA's software moat (CUDA) and system integration (NVLink, NVL72, reference architectures) are so deep that no competing stack can replicate them inside the investable horizon.

What the market is underpricing. The probability of a durable two-stack world in which the Chinese compute ecosystem becomes self-sufficient enough to serve its domestic market and export a sovereign alternative to non-aligned buyers.

What investors should watch next includes NVIDIA's guidance treatment of China; Huawei 950PR and 950DT production and customer adoption; DeepSeek V4 and Alibaba model–hardware alignment; HBM and CoWoS supply; time-to-power at new AI campuses; BIS and Chinese procurement actions.

The correct investment question is therefore not *"is the B200 better than the Ascend 950PR?"* but these five:

1. Which complete system offers the best total cost of ownership per useful token, per trained model and per national deployment?
2. Which country can scale energy, data centers, advanced packaging, memory and networking faster?
3. Which stack receives software and model optimization first?
4. Which regulatory regime accelerates its own domestic adoption the most and constrains its rival the most?
5. How does all of the above redistribute future margins, multiples and market share across semiconductors, memory, networking, energy and infrastructure?

3. Why chip-vs-chip analysis fails

The most common analytical error is to take a marketing or benchmark number and turn it into an investment thesis. In AI that approach fails for several reasons.

First, throughput numbers depend on the **precision used**. Blackwell can look spectacular in FP4, but the economic relevance of FP4 depends on whether the workload, the model and the training or inference process actually tolerate that quantization without material degradation.[2, 3]

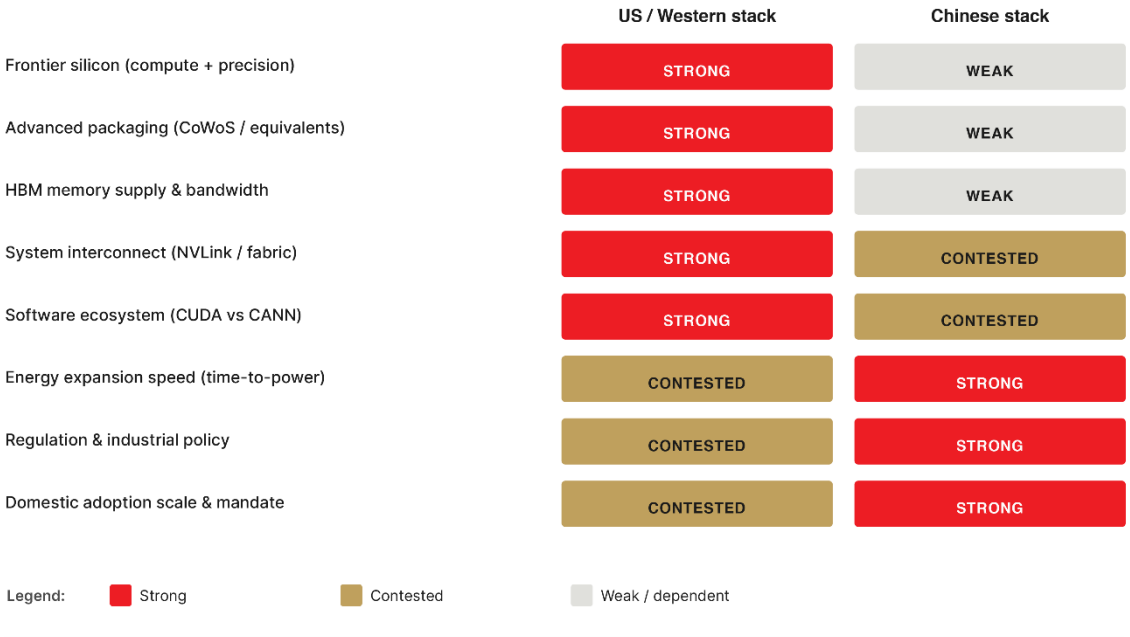
Second, the economics of compute are no longer explained by the performance of a single GPU. They are explained by the **rack**, the **cluster**, the **network** and the **software**. The gap between "I have a better chip" and "I have a better system" is wide, and NVIDIA knows it, as Blackwell is increasingly sold as a rack-scale platform rather than as a discrete component, and Rubin pushes that logic further.[2, 13]

Third, the Chinese market is no longer playing only to maximize performance. It is playing to maximize **autonomy**. Reuters has reported that much of the domestic chip surge was accompanied by public infrastructure spending and implicit "buy Chinese" guidance.[4] CSET's own tracking of more than 180 Chinese AI chip products across 30+ Chinese organizations shows that Huawei's Ascend line, despite being entity-listed since 2020, is now the most widely deployed Chinese accelerator and is comparable in some dimensions to earlier NVIDIA generations.[15]

Fourth, energy is emerging as a binding constraint. The IEA estimates that data centers consumed around 1.5% of global electricity in 2024 (roughly 415 TWh), with the United States at **45%**, China at **25%** and Europe at **15%** of the global data-center electricity footprint.[6] Brookings, however, warns of a potential *"electron gap,"* meaning China may enjoy a strategic edge in energy expansion and in how quickly new capacity can be wired to AI deployments, even without leading today's total consumption.[1]

FIGURE 1 — Compute stack: US / Western vs Chinese positioning by layer (2026)

Qualitative scoring across eight stack layers. Positions are relative and directional, not a static benchmark.



Source: RefleXio Intelligence synthesis (2026), based on CSET, Brookings, IEA and NVIDIA filings.

Figure 1. Compute stack, US vs China, layer by layer. Source, RefleXio Intelligence (2026), based on CSET, Brookings, IEA and NVIDIA filings.

4. What really matters in the NVIDIA–Huawei comparison

4.1 Compute and precision

Blackwell does not just raise peak FLOPS; it attempts to turn low precision into a stable commercial advantage. NVIDIA argues that its Transformer Engine with micro-tensor scaling makes FP4 usable while preserving accuracy, and that the precision step itself contributes to doubling performance and effective capacity for certain models.[2]

The investor distinction worth making is between peak advertised performance, sustained performance on real workloads, precision acceptable for training, precision acceptable for inference, and the engineering cost required to extract the promised performance. A big mistake would be to assume that every FP4 benchmark equals an immediately monetizable advantage. NVIDIA's own technical writing admits that low precision requires specific recipes, higher-precision layers in some cases and careful calibration to preserve downstream behavior.[3] Rubin, announced at FY26 close, claims a further 10× reduction in inference token cost versus Blackwell on SemiAnalysis InferenceX benchmarks, an external validation point worth tracking as Rubin ramps.[13, 17]

> **What investors should watch**
>
> • MLPerf Inference and Training results, convergence times and training stability across reduced-precision formats.
>
> • Accuracy retention after quantization, particularly on mixture-of-experts and long-context workloads.
>
> • Portability signals across stacks. How easily models move between NVIDIA and non-NVIDIA targets, especially the CANN ↔ CUDA compatibility layer announced with DeepSeek V4.[12]

4.2 Memory and bandwidth

In advanced AI, memory is no longer a compute peripheral; it is the central bottleneck. An accelerator's value depends on how much model it can hold, at what bandwidth, and with what latency it can feed the critical kernels.

Public debate oversimplifies this. A chip with weaker compute can get remarkably close if it compensates with the right memory configuration, efficient partitioning, distributed inference and a well-integrated system architecture. Comparing B200 or GB200 with an Ascend accelerator on FLOPS alone is misleading. But direction of travel matters, because SemiAnalysis estimates that China's largest DRAM national champion, CXMT, is on track to produce roughly 40,000 HBM stacks in 2025 and 2.2 million in 2026, while global HBM shipments will be around 23.7 billion Gb in 2025 rising to 30 billion in 2026, which leaves China with a ~70× shortfall in HBM bits even

after the ramp.[17] That gap, not peak FLOPS, is the real bottleneck on Huawei's ability to compete at scale.[15, 16]

> **What investors should watch**
>
> - HBM evolution by generation, capacity per GPU, effective bandwidth and lead times.
> - HBM3 / HBM3E / HBM4 availability and supply-side constraints.
> - CXMT ramp, HBM2/3 yields, and any signal that Huawei and CXMT move from stockpile dependency to domestic flow.
> - Substitution signals toward cheaper inference-optimized memory stacks.

OSAT, substrates and the HBM4 transition

The memory bottleneck is not only about who fabricates the DRAM. Three adjacent constraints determine how quickly HBM translates into deployed compute, and all three are tightening through 2026. The first is advanced packaging capacity. TSMC's CoWoS-L (the variant used for Blackwell and Rubin) is the single hardest bottleneck in the chain, as SemiAnalysis and issuer commentary consistently identify it as the throughput ceiling on NVIDIA deliveries.[17] The second is the OSAT layer outside Taiwan, where ASE, Amkor and SPIL are scaling capacity for HBM-adjacent assembly, but geographic concentration remains high and substrate supply (particularly ABF substrates used in large-die AI packages) has been in multi-quarter deficit. The third is glass substrates, the emerging alternative that Intel, Samsung and TSMC are all developing; large-scale deployment is not expected before 2027–2028.

For HBM4, the transition from HBM3E introduces a further complication, in that the base die moves to a logic-foundry process (TSMC N5/N3), which means every HBM4 stack now consumes not just DRAM wafer capacity but also leading-edge logic wafer capacity. That tightens the coupling between the memory supply chain and the same advanced nodes used for GPUs themselves, a second-order effect that compresses the pool of frontier wafer starts available across the industry.

Indicative lead times, AI supply chain components (Q2 2026):

Component	Indicative lead time	Direction 2025 → 2026	Primary suppliers
HBM3E (12-Hi stacks)	22–32 weeks	Extending	SK hynix, Micron, Samsung
HBM4 (samples / early volume)	36–52 weeks	New, first shipments 2H26	SK hynix lead; Micron/Samsung Q4 2026+
CoWoS-L advanced packaging	26–38 weeks	Extending	TSMC (sole at scale)
ABF substrates (large-die)	20–30 weeks	Stable / tight	Ibiden, Unimicron, Shinko

Component	Indicative lead time	Direction 2025 → 2026	Primary suppliers
Glass substrates (pilot)	Pilot / 2027+	Ramping	Intel, Samsung, TSMC (dev)
High-voltage transformers (DC-grade)	104–208 weeks	Extending sharply	Global, multi-year backlog

Source, RefleXio synthesis (2026) from SemiAnalysis, issuer commentary and industry reporting. Lead times are indicative and vary by buyer, priority and volume. The transformer row is included because a GPU without grid-delivered power is an idle GPU, and in 2026 it has become the most binding non-silicon constraint (see Section 4.5).

4.3 Interconnect and system performance

Interconnect is where superficial comparisons break. NVLink, NVSwitch and the NVL72 system architecture have become a structural advantage for NVIDIA. The economic value is not only in the chip, but in making thousands of GPUs work as a single coherent system. Rubin extends this advantage with tighter integration between the CPU (Vera), GPU (Rubin) and networking (BlueField-4).[13]

Jensen Huang's strategic concern in the Dwarkesh Patel interview is easier to understand from this angle. If China gets its models and clusters to run increasingly well on Huawei hardware, the battle stops being about node lag and becomes a battle of **complete architectures**.[8] Huawei's announced CloudMatrix 384 and Atlas 950 SuperPoD (the latter linking 8,192 Ascend chips in late 2026) are the first concrete evidence that China is building at that architectural level and not just at the chip level.[18]

> **What investors should watch**
>
> - Huawei cluster announcements (CloudMatrix 384, Atlas 950) and proprietary-fabric deployments.
> - Progress in silicon photonics and multi-node MLPerf benchmarks.
> - Distributed inference performance and penetration into national and provincial data centers.

4.4 Software moat and migration cost

This is NVIDIA's central moat. CUDA, the libraries, optimized frameworks, TensorRT, NeMo and the sheer familiarity of the developer base reduce adoption cost and raise effective performance. NVIDIA's own filings reveal how much the company sees the loss of the Chinese market as an ecosystem threat, given that the company has stated that the effective closure of the Chinese

market has helped its competitors build larger developer and customer ecosystems to challenge it globally.[9]

An investor focused only on quarterly margins can underprice the future cost of losing a developer community. If Chinese software matures enough, CUDA's economic advantage may compress first in China and then in markets that align with technological-sovereignty strategies. Two 2026 data points sharpen this risk. First, DeepSeek V4 will run on Huawei Ascend with a CANN-Next heterogeneous compute architecture that advertises CUDA compatibility, with training still on NVIDIA but inference on Ascend;12 and second, weekly token consumption from Chinese open models on the OpenRouter platform surpassed US models in February 2026, and the gap has widened since.[18]

> **What investors should watch**
>
> - Framework support, compiler quality and PyTorch / JAX portability.
> - CUDA-equivalent libraries (CANN, CANN-Next) and ease of kernel migration across stacks.
> - Adoption by Chinese labs and native optimization of leading models for domestic hardware. DeepSeek V4 on Ascend is the benchmark case.

4.5 Energy and grid readiness

Two opposite oversimplifications should be avoided here. The first is to claim that China already "vastly exceeds" the United States in AI energy. The IEA does not support that framing, as the United States represented **45%** of global data-center electricity consumption in 2024, against **25%** for China, with global data-center demand rising from **415 TWh in 2024 to ~945 TWh by 2030** in the IEA base case.[6, 14]

The second oversimplification is to conclude that China therefore has no advantage. Brookings has argued that China may enjoy a structural advantage in energy for the next phase of the race, citing faster capacity expansion, fewer bottlenecks for new facilities and greater state ability to pair new electrical infrastructure with industrial strategy. Brookings notes that US data-center electricity demand could more than double to **426 TWh by 2030**, or roughly 9% of total US electricity demand, while Chinese data center demand could reach **277 TWh by 2030**.[1]

For investors, the conclusion is simple. What matters is not only who consumes more electricity today, but who can **connect the next useful gigawatt to the next useful cluster** faster.

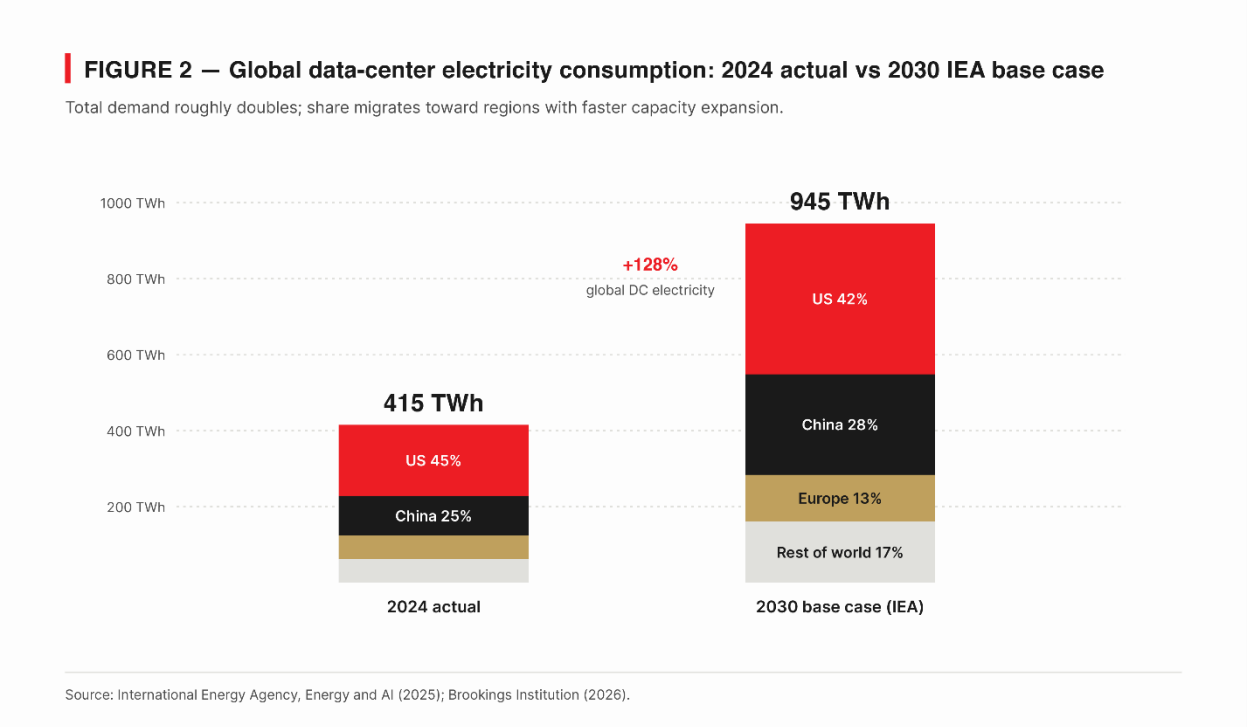


Figure 2. Global data-center electricity consumption, 2024 levels and 2030 trajectory. Source, IEA (2025); Brookings (2026); RefleXio Intelligence.

> **What investors should watch**
>
> - Time-to-power for new AI campuses and substation approvals.
> - Transmission capacity and marginal industrial electricity prices.
> - Cooling and firm-capacity availability for data-center-grade loads.
> - Turbine OEM backlogs (GE Vernova, Siemens Energy) as a leading indicator of firm-capacity scarcity.

4.6 Industrial capacity and supply chain

There is no AI leadership without supply-chain leadership. Advanced packaging, HBM, wafers, substrates, servers, cabling, optics, cooling and rack integration form a chain where a single bottleneck can stall the whole system.

RefleXio
www.reflexio.es
press@reflexio.es

RefleXio Intelligence

Independent research on compute, infrastructure and sovereign AI