IDC
FutureScape

AI infrastructure race: The compute, power and trust stack reshaping Asia/Pacific



Dr William Lee

Senior Research Director, Datacenter and Core Infrastructure Research, IDC

The trajectory and the framework

The model you license matters less than the stack beneath it

1B+
AI agents worldwide by 2029, more than half in Asia/Pacific

217B
agent actions a day

3.7T
tokens processed a day

>$68B
cost to deliver

Three layers are being decided at once

01 Compute is dividing Inference displaces training, and architectures diversify.

02 Power and placement gate the build Energy, cooling and sovereignty override price and performance.

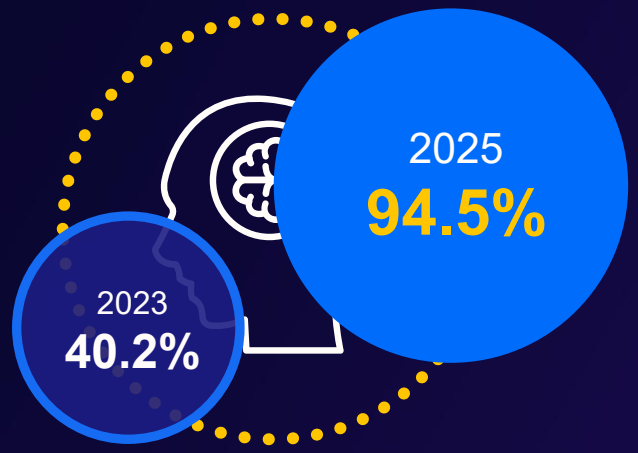
03 Trust becomes architecture A compliance obligation becomes a design requirement.

© 2026 IDC Distribution is forbidden without written consent of IDC. All rights reserved.
Source: IDC Agent Economics Adoption and Delivery Model, December 2025

Compute is dividing

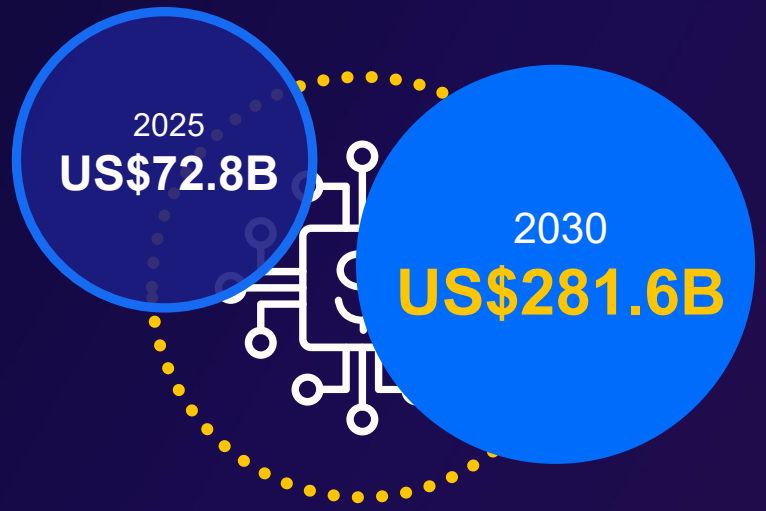
Inference takes over, spend quadruples and no one architecture wins

Inference goes mainstream



of Asia/Pacific organizations were using AI inferencing infrastructure; it is continuous and latency-bound.

Hardware spend quadruples



AI hardware spend in Asia/Pacific. Still the largest AI budget line; at 50.7% of total AI spend by 2030.

No single architecture wins

$52.7B vs $33.9B

Worldwide, 1Q26: non-x86 accelerated servers overtook x86 for the first time.

75.3%

of server revenue goes to AI-supporting systems by 2030, up from 54.5%.

Greater China: AMD ahead of NVIDIA; DeepSeek and Ascend mainstream.

© 2026 IDC

Distribution is forbidden without written consent of IDC. All rights reserved.

Source: AI-Optimized Cloud Infrastructure in Asia/Pacific: The Rise of AI-Native Providers, Sep 2025; IDC Worldwide AI and Generative AI Spending Guide, Aug 2026; IDC Quarterly AI Infrastructure Tracker, 1Q26 and Aug 2026 forecast

IDC PREDICTION

With hyperscaler premium AI infrastructure costs rising rapidly, by 2030, 75% of enterprises will shift advanced services to specialized clouds and private AI factories to mitigate costs.

75% of enterprises, by 2030

Always-on inference turns AI into a running cost, so economics push workloads off premium hyperscale capacity. Compute splits across hyperscalers, specialized clouds and private AI factories.



The signals so far

86%

of CIOs planned some workload repatriation, the highest rate yet

+1.0 pt

public cloud's share of Asia/Pacific compute spend barely moves: 24.6% to 25.6% in a year

© 2026 IDC

Distribution is forbidden without written consent of IDC. All rights reserved.

Source: IDC FutureScape 2027 – AI Factories Rewrite IT #6 (APJ); IDC Enterprise Infrastructure Pulse, 1Q26 and May 2026

Power and placement

Power and cooling cap AI capacity, and governments decide where it goes



82.9GW

of datacenter IT power capacity in Asia/Pacific by 2030

~16%

of all net new electricity demand globally through 2030 will come from datacenters.

Singapore

700MW

Jurong site set aside, but the second call awarded just 200MW. Datacenters use ~7% of power.

South Korea

3GW

of AI clusters planned outside Seoul.

Johor and Batam

212MW

went live in Johor in 1H2026 at 0.7% vacancy; Batam has a ~450MW power deal.

Liquid cooling, worldwide

49%

of datacenter owners and operators plan to add liquid cooling capability in the next two years, and 52% plan to increase rack density.

© 2026 IDC

Distribution is forbidden without written consent of IDC. All rights reserved.

Source: IDC Semiannual Datacenter Facilities Tracker, 2026; IDC Accelerated Racks, 2025; IDC Datacenter Facilities and Operations Survey, Dec 2025; Singapore EMA; Knight Frank; Cushman & Wakefield

IDC PREDICTION

By 2028, GPU, memory, networking, and power scarcity will force 70% of AI datacenter operators to shift from component-level buying to integrated, whole-system capacity planning.

70% of AI datacenter operators, by 2028

Scarcity in GPUs, memory, networking and power makes capacity a system decision, not a price-performance comparison of individual components.



The signals so far

+80%

DRAM revenue growth quarter on quarter in 1Q26: price +79%, volume +1%

15.2%

a year AI server revenue growth in Asia/Pacific against 4.5% for shipments, 2026–2030: fewer, costlier systems

© 2026 IDC

Distribution is forbidden without written consent of IDC. All rights reserved.

Source: IDC FutureScape 2027 – Rewiring Global IT Supply Chains #4 (APJ); IDC Memory Demand and Supply, Jun 2026; IDC AI Infrastructure Tracker, Aug 2026 (APJ)

Placement and sovereignty

Training stays in hubs; inference moves in-country

Decide jurisdiction before capacity: where tuning and serving must sit comes first, and only then how much to build or buy.



Train where the scale is

Shared cloud and global hubs

Frontier and foundation training stays where the fabric, accelerators and elasticity already sit.



Tune where the rules are

Sovereign and dedicated

Fine-tuning on regulated data needs sovereign capacity a customer controls, in a named jurisdiction.



Serve where the data is

In-country and at the edge

Inference follows latency and data gravity out to plants, networks and users.



Sovereignty in practice

79% of Asia/Pacific organizations running AI workloads have data residency requirements, and about 68% are using or planning sovereign AI within 12 months. That pulls inference and tuning in-country while heavy training stays in a few global hubs.

45.4% of AI-centric spend goes to dedicated capacity by 2030, growing 26.7% a year against 21.2% for shared

© 2026 IDC Distribution is forbidden without written consent of IDC. All rights reserved.

Source: IDC AI Infrastructure Tracker, Aug 2026 (APJ); IDC Survey, Sovereign AI Momentum, Mar 2026

Placement and sovereignty

IDC PREDICTION

With increasing regulatory nationalism, 50% of the A2000 will rely on AI factories designed to comply with nation-specific data standards, legal frameworks, and economic policies by 2030.

50% of A2000 companies, by 2030

Sovereignty decides where and how capacity is built, overriding price and performance. In Asia/Pacific, each market sets its own rules.



The signals so far

3X

resilience gap by 2029 between enterprises that adopt multicloud, portable or sovereign edge and those that don't

Placement first

Residency and jurisdiction have moved ahead of price and performance in the data-placement decision

© 2026 IDC Distribution is forbidden without written consent of IDC. All rights reserved.

Source: IDC FutureScape 2027 – AI Factories Rewrite IT #5 (APJ); IDC Cloud Pulse Survey 1Q26: APAC AI Infrastructure

Trust becomes architecture

Trust now outweighs model performance, and residency is rarely all-or-nothing

IDC PREDICTION

45%

With the EU AI Act and China's filing regime imposing incompatible assurance rules, 45% of A2000 companies will run jurisdiction-specific AI model registries and controls by 2029.

What organizations now weigh above model performance



Security



Data residency



Responsible AI



Open standards

Model performance now weighed after all four

>60%

of enterprises say changing privacy, security and AI rules already disrupt IT operations.

79% require data residency, at least for sensitive data



Strict rules for all data



Strict rules for sensitive data only



Flexible

© 2026 IDC

Distribution is forbidden without written consent of IDC. All rights reserved.

Source: IDC FutureScape 2027 predictions – Sovereignty & Regulations Reset IT's Rules #3 (APJ); IDC Asia/Pacific Emerging AI Technologies Survey 2025

IDC PREDICTION

With the growth of AI agents, 50% of APJ enterprises will restructure identity and data access under zero-trust architecture, requiring per-agent verification before autonomous action by 2028.

50% of enterprises, by 2028

If data can be faked, identity can't be the control point. Per-agent verification has to be designed in, not added later as a compliance check.



The signals so far

85%

of organizations in Asia/Pacific will face synthetic-identity phishing by 2028

75%

of security operations center alerts will be triaged by AI agents by 2028, worldwide

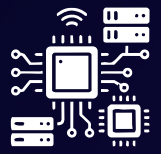
© 2026 IDC

Distribution is forbidden without written consent of IDC. All rights reserved.

Source: IDC FutureScape 2027 – Agentic Edge Continuum #2 (APJ), IDC FutureScape 2026 – Security and Trust #1 and #5 (APeJ)

Bringing it together

Compute, power and trust are one decision



Compute

The chips you pick set your power density.



Power and placement

That density, and the jurisdiction, decide which facilities can host you.



Sovereignty and trust

Each market's rules decide which trust controls you must meet.

Decisions made in 2026–27 follow you into the 2030s

Chips and servers	**Chip generation**	**Chip generation**	**Chip generation**
Power contracts, sites and sovereignty setups	**One commitment, a decade or more**		

Call 1
Where your inference runs, jurisdiction first

Call 2
Which platforms you standardize on

Call 3
Which partners you sign with

© 2026 IDC Distribution is forbidden without written consent of IDC. All rights reserved.

Source: IDC analysis, 2026. Timeline is illustrative; chip refresh cycles vary by platform.

Outlook and what to carry back

Placement will drive the next wave of growth

What could speed it up

- Power allocations turning into usable capacity on time
- Open liquid cooling standards
- Sovereign cloud regions moving pilots to production

What could slow it down

- Grid delays and limited renewable energy
- Chip shortages and rising power bills
- A different rulebook in every country

Who benefits

Service providers. Hyperscalers are only 20.9% of AI server revenue in 2030; digital service providers grow 34.2% a year and telcos 37.4%.

Organizations that plan compute, power and trust together will scale faster.

Early signals to watch



New capacity for high-density AI racks



Whether DeepSeek-style models spread beyond India and Greater China



Whether sovereign AI deals move from pilots to multiyear contracts

© 2026 IDC

Distribution is forbidden without written consent of IDC. All rights reserved.

IDC Research, Inc. (IDC) is the technology intelligence layer of the AI economy. A global leader in research and data for more than 60 years, IDC's expert analysts, proprietary datasets, and rigorous methodologies are trusted by business and IT leaders to guide critical business strategies and IT investments. Today, that intelligence is built into the tools and workflows where work gets done with IDC Quanta, making work sharper, teams faster, and businesses harder to beat.

Visit www.idc.com/ap, subscribe to the IDC blog, or follow IDC on X @IDCAP and LinkedIn Asia Pacific for the latest industry insights.

Terms of Use: Except as otherwise noted, the information enclosed is the intellectual property of IDC, Copyright 2026. Reproduction is forbidden unless authorized; contact permissions@idc.com before use. All rights reserved.


IDC