The Chinese AI Infrastructure Boom: Introducing the SemiAnalysis China Datacenter Model

1,000+ facilities across 60+ operators mapped, built retail-first and flipped by AI, largest hyperscaler leases 1/5 national capacity, 100MW in 12 months, Eastern Data Western Compute

SEP 25 https://newsletter.semianalysis.com/p/the-chinese-ai-infrastructure-boom

View in browser

The Chinese AI Infrastructure Boom: Introducing the SemiAnalysis China Datacenter Model

1,000+ facilities across 60+ operators mapped, built retail-first and flipped by AI, largest hyperscaler leases 1/5 national capacity, 100MW in 12 months, Eastern Data Western Compute

EVERLYN, DYLAN PATEL, AND PATRICK SCHAABI

SEP 25 • PAID

READ IN APP

China sits at the frontier of the global model race. GLM 5.3 and Kimi K3 are the latest in a run of striking open-weights releases. ByteDance's Doubao serves 345M monthly users as China's ChatGPT, and Seedance is the State-Of-The-Art video generation model.

Every one of those models runs on a datacenter, and China has been building them at a pace that has gone largely unmeasured outside the country. The biggest tenant files no 10-K. Several of the largest landlords have never listed. Most of the primary sources are in Chinese. So the market settled on two lazy assumptions: China is big,

QQ:1413221435

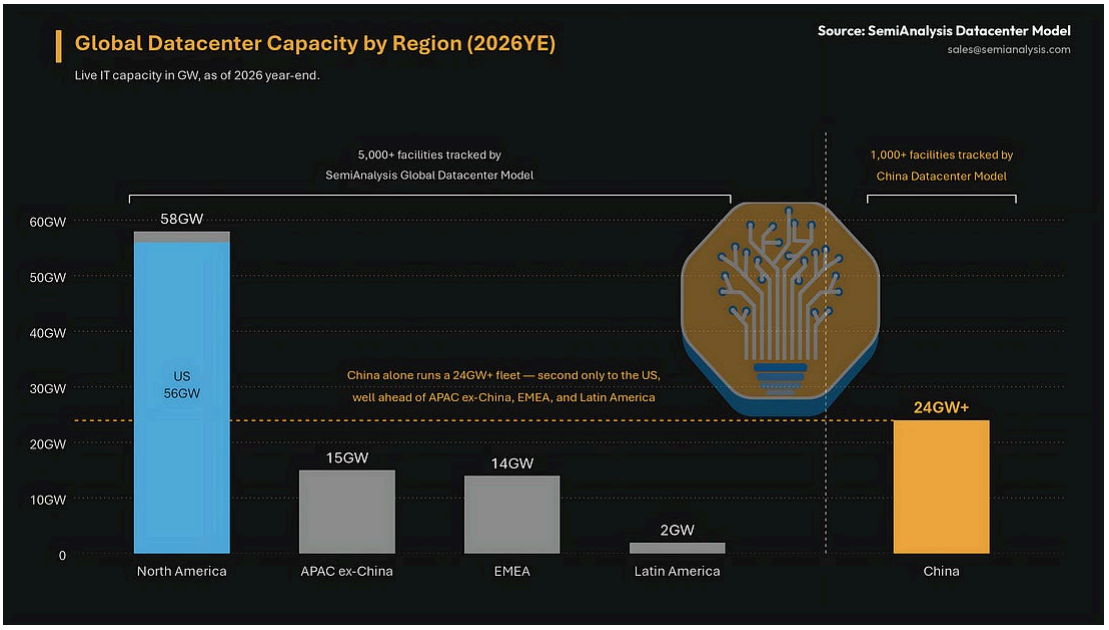
and China is empty. Published estimates of China's datacenter capacity differ by 15x, and reports keep citing high vacancy rates.

As with our flagship SemiAnalysis Datacenter Model, we use building-level data to show which of the most widely cited narratives hold, starting with how large the market is.

The global model tracks over 5,000 facilities across four regions. The US leads the world with 56GW of capacity as of 2026YE, followed by ~15GW for APAC ex-China, ~14GW for EMEA, and ~2GW for Latin America. Until now, it stopped at the border of China. Today we cross it.

Our tracking of 1,000+ datacenter facilities across over 60 players shows that China alone boasts a fleet of over 24GW. Bigger than EMEA. Bigger than the rest of Asia. This excludes ~20GW of dated pipeline and another ~30GW of announced projects.

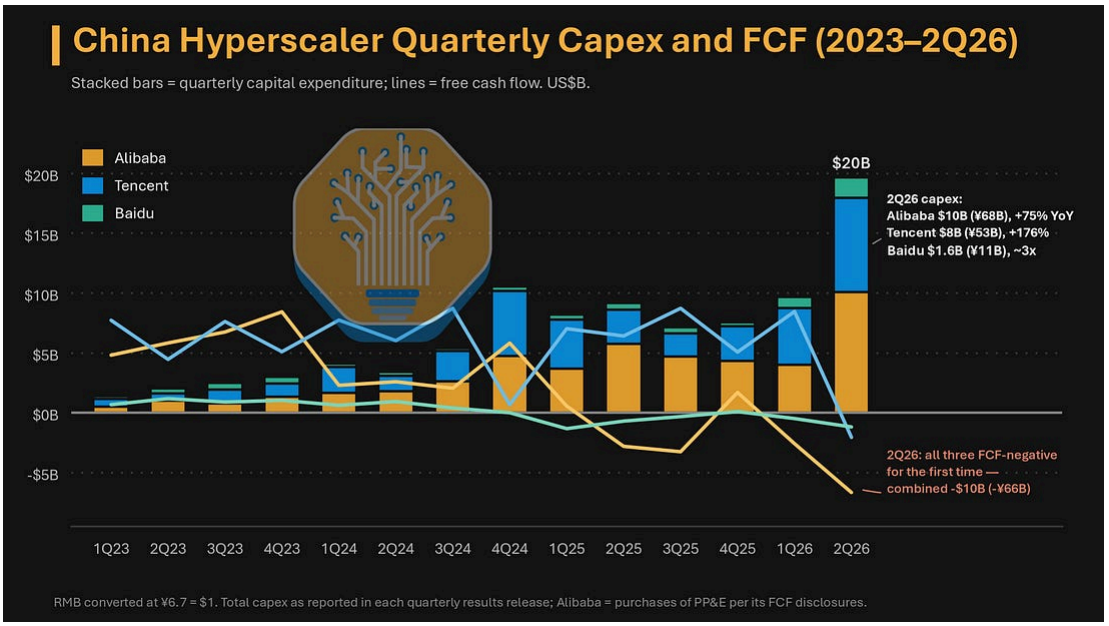
QQ:1413221435



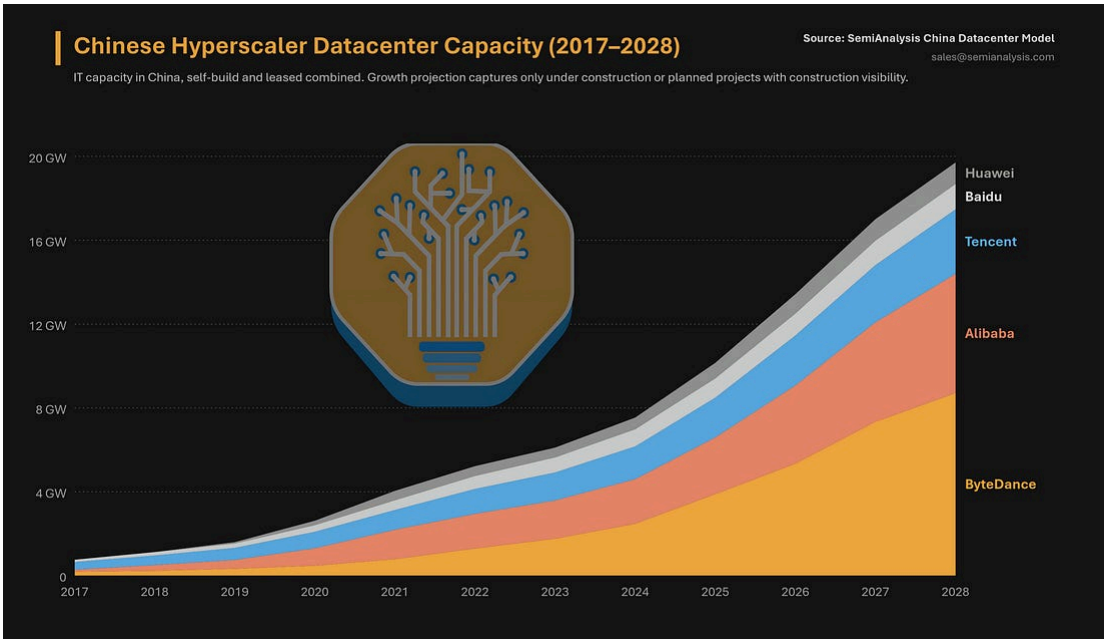
Source: SemiAnalysis Datacenter Model, SemiAnalysis China Datacenter Model

Growth is at an inflection point. In 2Q26, the combined capex of Alibaba, Tencent, and Baidu ("BAT") reached $20B, more than doubling YoY, and for the first time on record, all three posted negative free cash flow. This is the largest capex step-up in the sector's history.

QQ:1413221435



That total leaves out the largest spender of all, ByteDance, which remains private. According to our China Datacenter Model, ByteDance alone occupies roughly a fifth of delivered datacenter capacity in China, and it rents nearly all of it, making it the single most important customer for all wholesale colocation players in the country.



Source: [SemiAnalysis China Datacenter Model]

The listed players get the attention, but they are only the visible tip of the iceberg. This is also true for datacenter landlords. GDS and VNET, the only two Chinese datacenter landlords listed in the US, signed 1.3GW of wholesale orders in 1H26. But according to our China Datacenter Model, they captured barely a third of ByteDance and Alibaba orders in 2024–2026YTD.

QQ:1413221435

The buildout is a nationwide effort, and the state-owned carriers are also part of it. China's datacenter market was historically telecom-dominated. The state carriers held the majority of the market share in the 2010s, and still own a third of national capacity today. The national power grid companies invest aggressively too. Their combined capex accelerated from 2024, ending the 14th Five-Year Plan 24% over the original blueprint. The 15th plan (2026–2030) layers another 40% on top, to over $746B (¥5T).

Largely free of power constraints, labor shortages, and public protests, China routinely delivers 100MW datacenter facilities in under 12 months. Modular DCs are the standard playbook, which Tencent deployed its third-gen modular design in 2014. What is old news in China is only now being adopted at scale in America, as we detailed in The Wild Wild West of LEGO Datacenters.

The growth does not stop at China's border. Overseas leasing by Chinese hyperscalers is set to double from 2026 to 2029 and approach ~4GW of leased capacity in our SemiAnalysis Datacenter Model. That number still understates their offshore compute, because it excludes the hundreds of thousands of GPUs they rent from Western clouds, which we covered at length in How Oracle Is Winning the AI Compute Market.

However, the market also faces some dark clouds. Vacancy rates are high, and datacenter developers compete heavily on price. Chip supply is also constrained, due to export restrictions. But as explained below, neither has stopped AI datacenters from being built and filled at a remarkable speed.

In this report we cover, in order:

- Introduction to China's datacenter industry: the world's second-largest market, built retail-first, and why that history produced overbuild and high vacancy alongside an AI capacity shortage.
- The westward buildout: how the "Eastern Data, Western Compute" policy and hyperscale AI demand are moving the industry inland, at construction speeds and unit costs the West cannot match.
- Behind the paywall: we go tenant by tenant through ByteDance, Alibaba, Tencent, Baidu, and Huawei, from their AI and datacenter strategy to leasing and self-build footprints.

QQ:1413221435
QQ:1413221435

Introducing the SemiAnalysis China Datacenter Model

This report is a preview of the **SemiAnalysis China Datacenter Model**. It is the China counterpart to our global **SemiAnalysis Datacenter Model**, built to the same standard, bottom-up and building by building.

Contact our sales team at sales@semianalysis.com for more information.

What you'll find inside:

- Tracking over 1,000 facilities across 60+ operators
- Quarterly delivery curves through 2032 for every facility, with construction status and timelines
- Tenant identification per building, supplemented by tenant allocations per customer-mix for leading operators
- Hyperscaler self-build vs leased splits and per-tenant capacity paths
- A lease signing tracker and a lease delivery tracker, on one reconciled basis
- Eastern Data, Western Compute ("EDWC") hub and cluster analytics, load-growth and power-draw forecasts
- An interactive full map view showing where the capacity is concentrated

Consider it the opening chapter. As with our global Datacenter Model, we will keep adding layers (AI capacity, power, suppliers, supply/demand analysis) with every update.

Let's dive in.

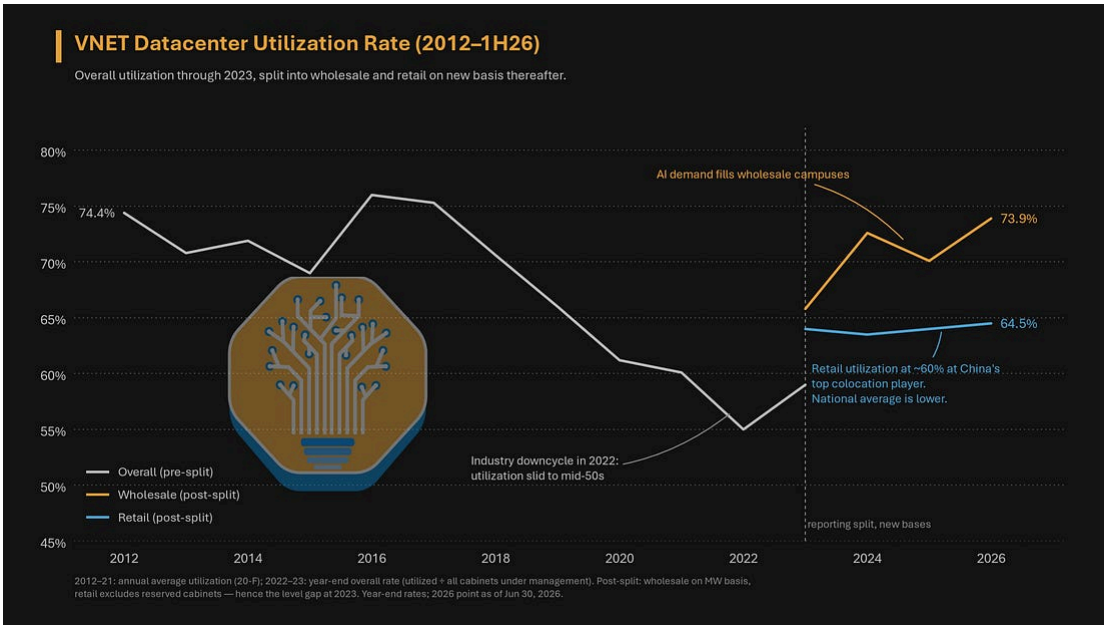
China's datacenter industry: built retail-first, flipped by AI

One of America's key advantages in the AI infrastructure race is that its datacenter industry grew up hyperscale. Multi-hundred-MW campuses have existed for years, purpose-built for a handful of enormous tenants. China's industry was shaped by telecom carriers and thousands of small retail deployments. Understanding that history explains what's strange about the market today, from overall high vacancy

QQ:1413221435
QQ:1413221435

coexisting with an AI capacity shortage, to why the carriers still own a third of everything.

The following chart previews the whole story. Between 2010 and 2022, a leading colocation player in China watched its utilization rate shrink from the mid-70% range to the mid-50% range. Then, from 2023 onwards, its segment disclosure reveals a two-speed market. Wholesale buildings are filled fast on AI demand, back above 70% utilization, while the legacy retail racks sit near 60%.



Source: VNET disclosure, SemiAnalysis

Era 1 — Carrier hosting (pre-2015): datacenters as telecom real estate

Unlike the rest of the world, where hyperscalers and colos dominate, China’s datacenter market was historically telecom-dominated. The three state carriers (China Mobile, China Telecom, China Unicom) held a combined 60–70% market share, with GDS and VNET, the largest third parties, below 5% each. The carriers enjoyed structural advantages in licensing and networking and benefited from captive demand. A large share of Chinese datacenter demand comes from government, SOEs and regulated industries, which default to state-owned suppliers.

For its first two decades, a Chinese datacenter was essentially telecom real estate, featuring low-density retail racks (~90% of facilities ran below 2kW) rented to websites, online gaming and CDN customers. This was often offered as a bundled product with bandwidth, served by small urban switching centers around Beijing, Shanghai and the Pearl River Delta. Third-party operators existed largely as resellers

QQ:1413221435
QQ:1413221435

of carrier space and bandwidth. The hyperscalers were barely a factor, as Alibaba Cloud was launched in 2009 and Tencent Cloud only opened to the public in 2013.

This era matters today because its stock is precisely the capacity now sitting vacant. The legacy retail racks no longer match what an H20 or Ascend server demands, and it’s impractical to retrofit given partial occupancy and high cost. The headline utilization number of ~50% is largely a monument to Era 1 (and Era 3, as we’ll see).

Era 2 — The cloud land grab (2015–2021): growth at any price

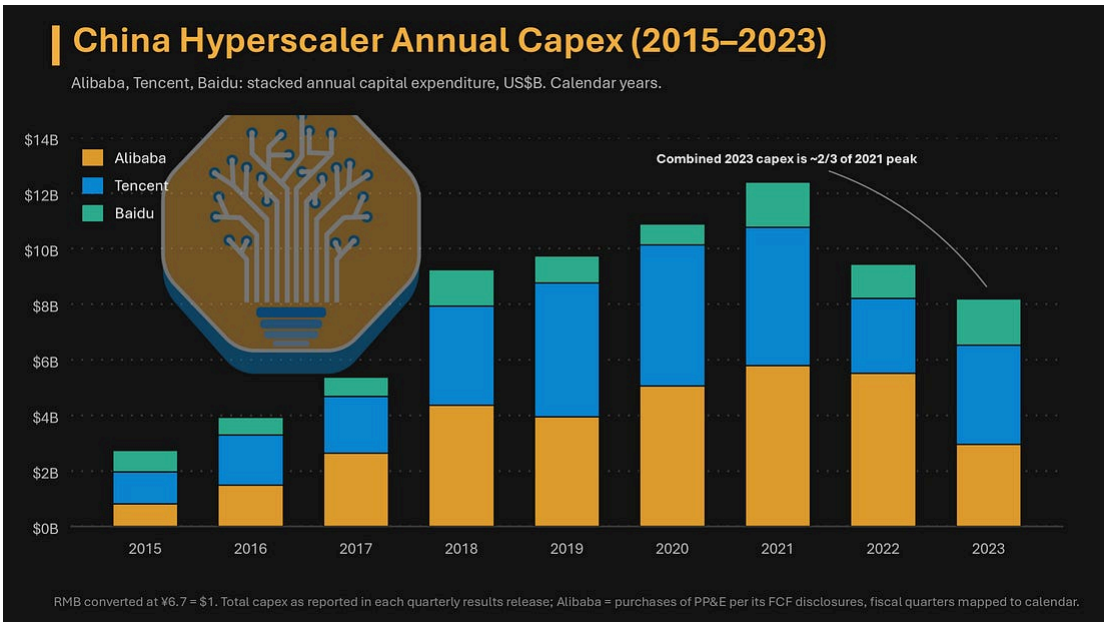
The cloud era began around 2015. The two leading players, Alibaba and Tencent, concluded that it was a winner-take-all market and opened a capex race. Alibaba Cloud revenue compounded at a ~110% CAGR across 2015–2019, at the cost of a brutal price war and EBITA losses. Its core products were marked down more than 50% after 17 consecutive price cuts during a 12-month window. Tencent responded in kind, driving combined hyperscaler capex to a 2021 peak.

Third-party colocation platforms (GDS, VNET, Chindata) scaled alongside them. Retail developers, meanwhile, kept building small urban sites on the assumption that cloud demand would spill over indefinitely.

Era 3 — Digestion (2022–2023): the music stops

The wild expansion stopped abruptly in 2022. Policy headwinds, macro slowdown, and share gains by state-owned clouds pushed Alibaba Cloud growth from +50% in 2021 to +23% in 2022, then to effectively zero. Tencent management framed it as a repositioning “from pursuing revenue growth to healthy growth” and saw a similar slowdown. Tencent cut capex budgets by nearly half in the middle of 2022. Alibaba did so in 2023.

QQ:1413221435
QQ:1413221435



Source: Company filings, SemiAnalysis

QQ:1413221435

Almost simultaneously, in February 2022, the National Development and Reform Commission (“NDRC”) launched Eastern Data, Western Compute guidelines and named ten computing clusters within eight national hubs. These regions, especially the western ones, enjoy fast track permitting and cheap green power.

The result was an oversupply despite a slowdown in delivery. Pricing was hit most. Power-exclusive market rates were cut in half from ~$80/kW/month. On the August 2026 earnings call, GDS management estimated another 18 months to fully digest the price reset, even though it has been going on for a few years.

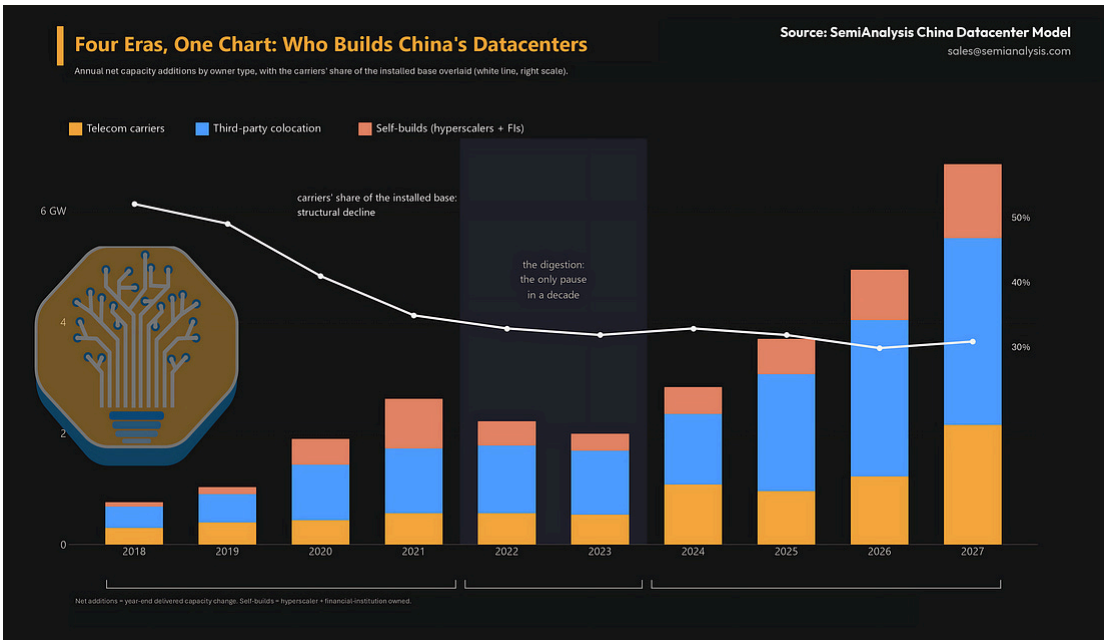
Era 4 — The AI supercycle (2024–): full steam ahead

DeepSeek and the hyperscalers’ own GenAI ambitions woke the market up over the course of 2024-2025. Combined BATB (ByteDance/Alibaba/Tencent/Baidu) capex went from roughly $35B in 2024 to over $50B in 2025 and is on a trajectory to $100B for 2026.

Wholesale orders returned at a scale never seen before, and the western hubs finally found their tenant. GDS and VNET together signed roughly 1.3GW of wholesale capacity in 1H26, with less than 10MW of retail orders. The AI era is wholesale. The legacy retail stock is not participating.

But watching only these two listed names badly understates total demand, as they only capture ~1/3 of ByteDance and Alibaba orders in 2024–2026YTD per our China Datacenter Model.

QQ:1413221435



Source: [SemiAnalysis China Datacenter Model]

To recap, the four-stage history explains China's datacenter market today: (1) a retail-focused legacy stock around Tier-1 cities, running 30-50% utilization, often physically unsuitable for AI densities; (2) a wholesale AI-driven buildout happening somewhere else entirely; and (3) a carrier share in structural decline, as hyperscaler self-build and third-party wholesale take over the growth.

SemiAnalysis is a reader-supported publication.
To receive new posts and support our work,
consider becoming a subscriber.

Subscribed

Eastern Data, Western Compute: the pipeline is heading west

The EDWC policy designates eight national computing hubs, implemented as ten clusters, spanning both the demand centers (Beijing-Tianjin-Hebei, Yangtze River Delta, Greater Bay Area, Chengdu-Chongqing) and the resource-rich west (Inner Mongolia, Ningxia, Gansu, Guizhou).

QQ:1413221435
QQ:1413221435



Source: NDRC, SemiAnalysis

QQ:1413221435

This is the industry's top-level design blueprint, and unlike most "guidance" documents, it actually changed how the industry develops.

The policy itself has been covered extensively. What is missing, and what matters, is the transmission mechanism.

In China, the first permit a datacenter developer applies for after securing land is an energy-consumption review filed with the local Development and Reform Commission. That quota is the binding cap on how large a facility can be. Any project consuming over 10,000 tons of coal equivalent (>10MW) cannot legally begin construction without passing this review.

The review regime dates to 2010 and became a rationing tool over time. When EDWC arrived in 2022, the quota system became the most effective mechanism because the same offices that issue quotas are the ones implementing the EDWC policy, so quotas effectively stopped construction flowing to non-hub locations.

Guangdong is the cleanest example. The province stopped issuing datacenter energy quotas in 2021-2022 and reopened applications only for the EDWC hub region in Shaoguan. Similar quota control governs the other eastern metros. Meanwhile, EDWC regions hand out energy quotas, land approvals and tax incentives generously.

Despite the mechanism, for the first two years, migration was slow because the demand of the day (cloud, gaming, e-commerce) required low latency to population, so hyperscalers kept fighting for scarce quota near Tier-1 cities.

QQ:1413221435

A little-noticed episode illustrates how the quota system played out in practice. In mid-2023, Guangdong's Energy Bureau inspected Tencent's two self-built campuses in Qingyuan and found each running at roughly 13–14x its approved scale. The projects were labeled "approved small, built big", and ordered rectification within six months. The local grid company was instructed to hold power supply to the approved quota. Even for Tencent, incremental builds now belong to the province's designated hub node.



Source: Guangdong Energy Bureau

Today, the campuses keep operating, but as our China Datacenter Model shows, Tencent's subsequent capacity went to Shaoguan. The reason is not simply regulatory pressure.

Two other things changed the picture.

First, telecom-led fiber investments have significantly reduced latency. Shaoguan again is the model case. Since 2022, the local government has invested ~$600M (~¥4B) in network infrastructure, building 13 ultra-high-speed 400G all-optical routes covering the Greater Bay Area, achieving 1.3ms one-way latency to Guangzhou, 1.66ms to Shenzhen, and under 3ms to Hong Kong. A direct Shenzhen-Guangzhou-Shaoguan hollow-core fiber corridor has been planned to cut latency a further ~30% from 2027 onwards. At those numbers, a Shaoguan rack is functionally a Shenzhen rack for all but the most latency-critical workloads.

Second, AI demand did the rest. Training workloads don't care about latency, but they care about power prices.

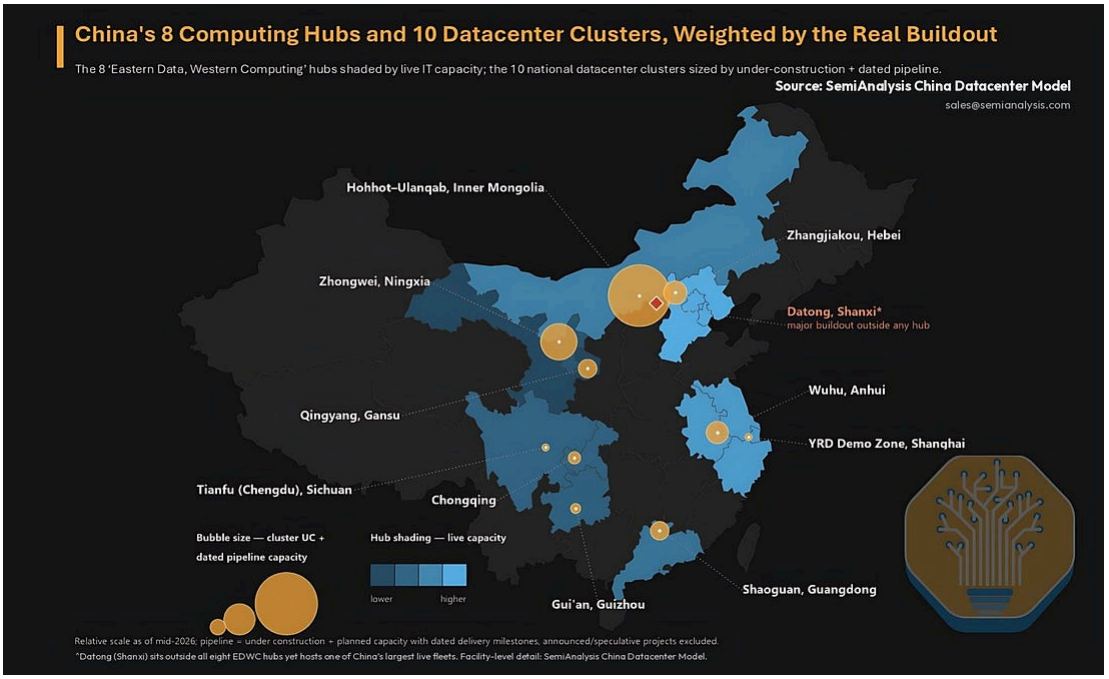
Inner Mongolia is the best example here. The region has China's largest renewable fleet and, critically, western Inner Mongolia is served by the Mengxi grid, China's only provincial grid outside State Grid and Southern Grid, with its own multilateral trading mechanism and independent pricing.

QQ:1413221435
QQ:1413221435

As a result, Inner Mongolia has emerged as China's Johor, the default destination for hyperscale AI capacity, with generous quota allocations, land measured in square kilometers, and, most importantly, power prices at roughly half the Tier-1 city level.

The Johor parallel runs deeper than land and power. On economics, Chinese rents are far lower, but so is the capex, while Johor rents are higher, but so is the cost to build. Our data shows that the two markets' yield-on-cost ranges overlap, with the best Inner Mongolia campuses underwriting to the return of Johor's international developers. However, the trends diverge. Inner Mongolia construction costs keep falling as the prefab approach and domestic substitution compound, while Johor developers face capex inflation and overrun risk in a supply chain running hot. Chinese operators building in Johor enjoy geographical arbitrage leveraging the Chinese supply chain and earn higher yields.

Zoom out from Inner Mongolia, though, and the ten clusters tell very different stories. It is easiest to see when the map is weighted by what is actually live and what is committed, rather than by what the policy designated, as the map below shows.



Source: [SemiAnalysis China Datacenter Model]

A recurring media take holds that EDWC is a boondoggle of empty western shells. The simplest rebuttal is to look at where the hyperscalers put their own money. As our tracking shows, ByteDance currently operates three self-build clusters: Datong in Shanxi (its earliest), Wuhu in Anhui (an EDWC cluster, serving east-China inference) and a twin-site Inner Mongolia program (Horinger and Ulanqab), with two Ningxia entities registered in Zhongwei (another EDWC region) in July 2026 pointing to a fourth.

QQ:1413221435

Outside of Inner Mongolia, development across the clusters is sharply divergent, but much of that divergence is by design. Gui'an was positioned from the start as a storage hub, as cold data doesn't mind the mountains. The most famous anchor tenant is Apple's iCloud China datacenter, operated with Guizhou-Cloud. Less reported is that Huawei and Tencent both have self-build clusters in this region.

Chengdu and Chongqing are different stories. They are pitched as national compute hubs but lost the AI-demand contest to Inner Mongolia on power price and grid flexibility. EDWC succeeded in concentrating the buildout inside the map, and the market then picked the winners within it.

Shanxi is the opposite. It is the largest datacenter region outside the EDWC hub system entirely (the red diamond on the map above), but attracted close to 2GW commitment from ByteDance, Baidu, and JD across self-build sites and colocation leases. The reasons are abundant cheap power (Shanxi is China's coal province), proximity to Beijing, and first mover advantage. Hyperscalers went there before EDWC existed and kept building out, proof that the quota regime channels the buildout, but electrons and geography still make the site selection.

QQ:1413221435

SemiAnalysis is a reader-supported publication.
To receive new posts and support my work,
consider becoming a free or paid subscriber.

Subscribed

Speed and cost: a different sport

The US buildout is power-gated with interconnection queues, transformer lead times, behind-the-meter workarounds, which we have covered at length here. China's is chip-gated: generation is abundant, state-owned utilities treat datacenter load as a development win, permitting is a policy funnel rather than a bottleneck, and construction labor is plentiful. With no physical friction on the facility side, the build cycle is driven purely by internet platform demand.

On speed, the standard delivery time for a 100MW facility has compressed from ~18 months in the cloud era to ~12 months in the AI era, with state-of-the-art projects finishing faster still.

QQ:1413221435

For reference, our US modular deep dive The Wild Wild West of LEGO Datacenters found the best-case full-modular buildout in the US still takes 12–18 months of construction, sitting behind 12–13 months of permitting that cannot overlap. For China, the permitting period before construction is typically 3–6 months, and faster in EDWC regions.

Several structural factors contribute to the overall faster speed.

First, speculative shells ahead of orders. Because hyperscalers now run T+6-month delivery requirements in their tenders, colocation developers pre-build civil works at their own risk, making the tenant competition effectively a race among half-finished shells. When GDS signed a 152MW AI mega-order in early 2025, the customer required delivery within six months and got it.

Second, building greenfield sites in western EDWC hubs changed the construction method. Away from Tier-1 land constraints, developers are moving from conventional multi-story concrete to low-rise, long-span steel halls. The steel is made in a factory while foundations are dug on site, then trucked in and lifted into place by crane. Concrete is used only in the ground, so there is no waiting for it to harden before the next floor goes up. VNET’s Ulanqab footage shows a hall go from bare foundations to a finished building within one summer-to-winter season.

QQ:1413221435

QQ:1413221435



Source: VNET videos, SemiAnalysis

Tencent’s T-Block takes the same logic to its extreme. The entire datacenter arrives as factory-built modules, with the “building” shrinking to a leveled site and a simple steel shell, sometimes no building at all.

QQ:1413221435
QQ:1413221435



Source: Tencent Video, SemiAnalysis

Third, prefabrication turns a serial schedule into a parallel one. The bulk of MEP work moves off-site into factories, so electrical and mechanical fit-out no longer waits for the building. The headline example is Alibaba's CUBE 5.0 "100-day datacenter", comprising 30 days of factory production and site prep, 50 days of on-site installation, and 20 days of commissioning, measured from a finished shell to the day servers arrive. This roughly halves M&E and commissioning stage compared with market standard 6-month, and brings ground-break to commissioning down to roughly 7–9 months.

Every factor above points in the same direction, which is more factory, less site. Follow that logic to its endpoint, and you would expect Tencent-style full-modular container campuses to be the national standard by now. They aren't, for two very Chinese reasons.

First, the labor math is different. Modularization's biggest payoff in the West is dodging scarce, expensive skilled trades. Chinese site labor is affordable and

QQ:1413221435

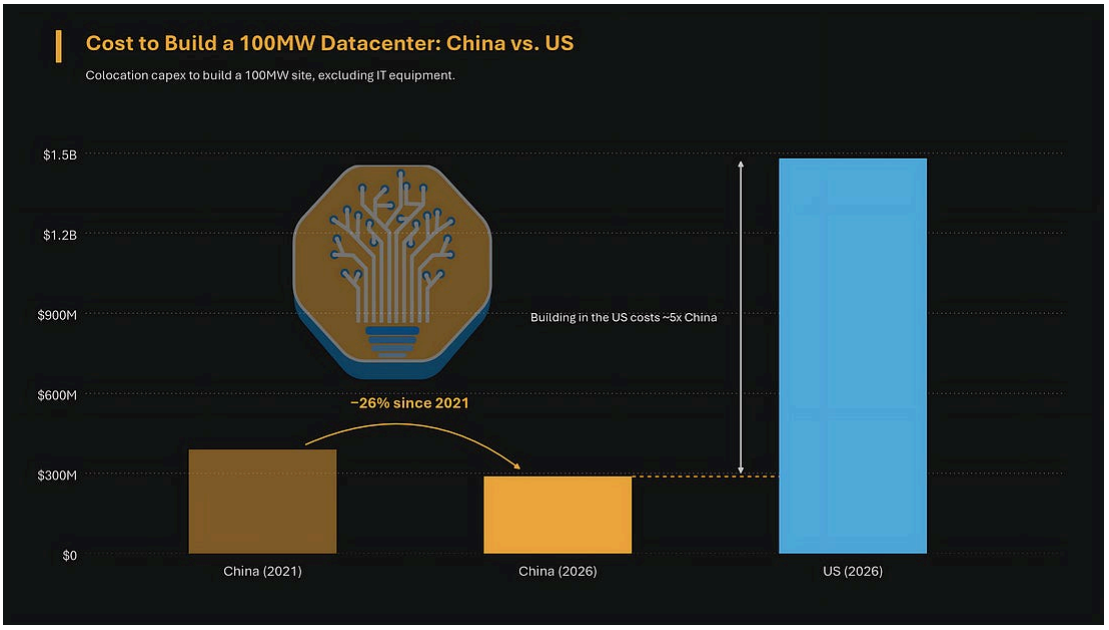
abundant, so a traditional steel-and-prefab build already hits aggressive schedules. The marginal speed gain from going full-container is small.

Second, the financing dynamics play a role. Chinese colocation developers pledge their datacenters to banks as loan collateral. A proper building with a 20-year design life appraises well, but a yard of containers appraises poorly and carries lower residual value. When your business model is “borrow against the asset, lease it to ByteDance,” you build a real building.

Today, full-modular builds remain limited to hyperscaler self-builds and Southeast Asia projects, but overall construction speed is as fast as needed.

On cost, capex keeps falling even as speed rises. Continued supply-chain compression and rising domestic content have pushed shell-and-fit-out costs to a fraction of comparable US wholesale costs, as we show in our model. Alibaba credits its modular approach with a further >10% cost reduction compared to its earlier version on top of the schedule gains.

QQ:1413221435



Source: SemiAnalysis China Datacenter Model

SemiAnalysis is a reader-supported publication.
To receive new posts and support my work,
consider becoming a free or paid subscriber.

Subscribed

QQ:1413221435

The demand: five hyperscalers, multiple playbooks

The demand side of China's AI buildout is concentrated in a handful of names: ByteDance, Alibaba, Tencent, Baidu, and Huawei.

Like their American counterparts, they span social media, e-commerce, search, gaming and short video. However, the consumer internet that built these companies has run out of new users. Today, China's internet penetration already sits at ~80% and grows only ~1.5% a year.

The growth engines of the past are all in single digits.

- Social media: Tencent's core asset, WeChat, grew MAU just 2% YoY in 2025.
- Search engine: Baidu's 2025 revenue fell 3% with online marketing down double digits.
- E-commerce: Alibaba's customer management revenue fell 7% YoY in the quarter ended June 2026.
- Short-video: Kuaishou's 2025 DAU grew just 2.7% YoY.

Outside of consumer apps, the same titans lead the cloud industry, with a structural handicap. The market is dominated by Alibaba, Huawei and Tencent alongside the three telecom carriers. But unlike the US hyperscalers, which generate roughly a third or more of their cloud revenue outside the US, Chinese cloud revenue is largely domestic. Alibaba Cloud, the most global of the group, still earns only ~20% of cloud revenue overseas, and the carrier clouds are effectively 100% domestic. In terms of "addressable GDP," the core market of the US tech titans is simply much larger.

Against that backdrop, AI has become the growth engine, which drives the capex inflection point in 2Q26 we mentioned earlier.

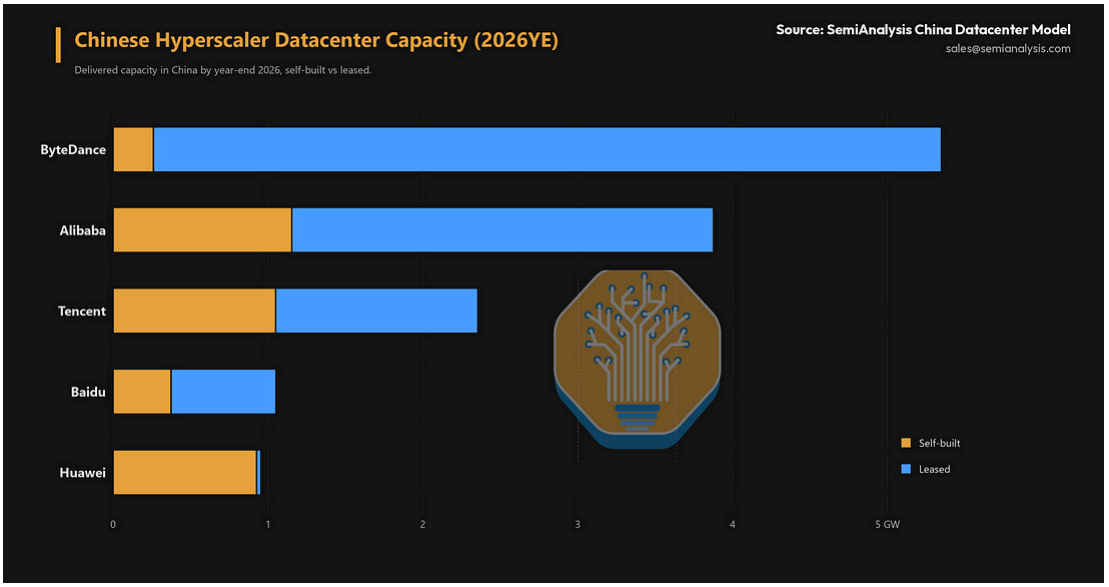
Beyond funding their own model R&D, the big platforms have built an investment web across China's GenAI labs. Alibaba invested in five LLM unicorns. Tencent backed six and wrote the largest external ticket in DeepSeek's Series A.

The labs, in turn, train and serve on the hyperscalers' clouds (in the case of the Alibaba/Moonshot relationship) or have their models served inside the hyperscalers' products (in the case of the Tencent/DeepSeek relationship). Either way, the

QQ:1413221435
QQ:1413221435

hyperscaler datacenter expansion is the compute foundation for the entire domestic AI ecosystem. Hence the deep dives that follow.

We provide a snapshot below showing where the hyperscalers stand:

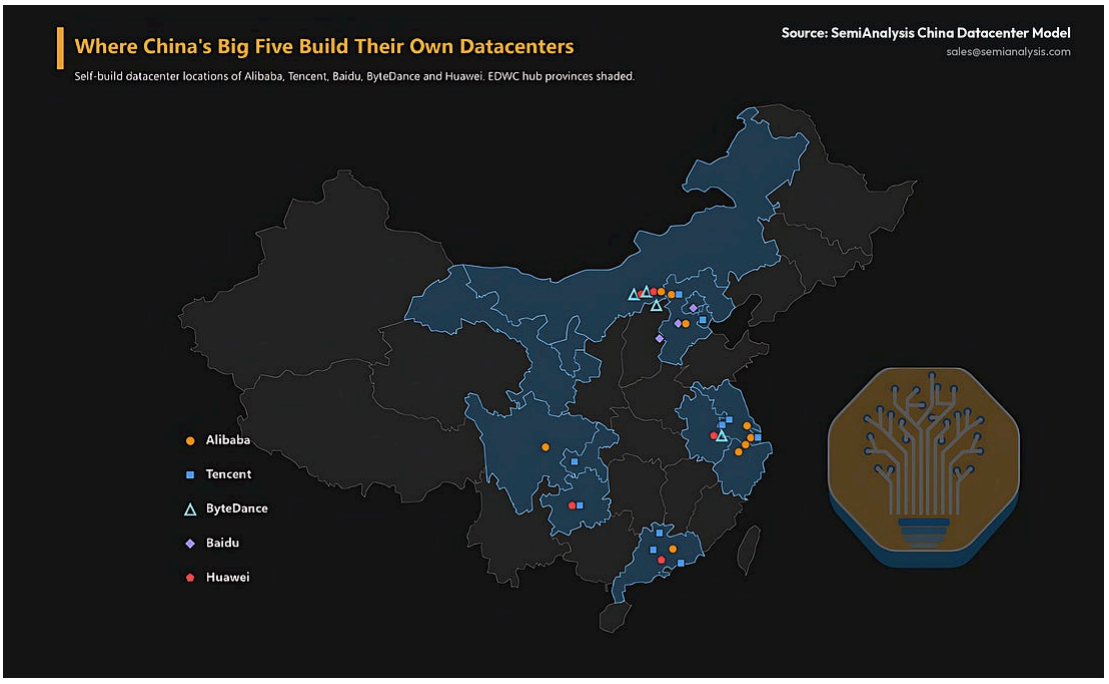


Source: [SemiAnalysis China Datacenter Model]

The leasing: Today, ByteDance and Alibaba each issue annual leasing demand at gigawatt scale and both are actively expanding overseas. ByteDance's is the largest and leases majority of its footprint, so its tender calendar effectively sets the order book for every wholesale operator in the country. Alibaba, as we show below, splits the demand across self-build, BOT and wholesale. Tencent is the most disciplined. It spent the 2022–23 downcycle closing small leased metro sites and consolidating into its own campuses, before leasing picked up again in the AI era. Baidu tenders at a fraction of the others' scale.

The self-builds: In the 2010s, Chinese hyperscalers began self-building datacenters as their businesses reached sufficient scale, and the strategies have diverged since. Alibaba, Tencent, and Baidu self-build their strategic campuses and lease the rest. Huawei self-builds almost everything because of its business model. ByteDance, the primary tenant, is now ramping its own campuses fast. Our China Datacenter Model tracks all self-builds, and the map below shows the current landscape.

QQ:1413221435
QQ:1413221435



Source: SemiAnalysis China Datacenter Model

QQ:1413221435

Between pure leasing and pure self-build, a hybrid structure called BOT also emerged. The real anatomy is a two-layer ownership split. The hyperscaler owns the land and the civil shell, while the colocation partner invests in, owns and operates the M&E layer and recovers its capital through 8–10-year service contracts. In this way, the hyperscaler keeps control and speed, the partner's balance sheet carries the capex, and the IRR is contracted rather than merchant. All three of BAT run some version of this playbook. Alibaba partners with AtHub, Tencent with Kehua Data, and Baidu with Aofei. Alibaba takes the mixing furthest, running self-build, BOT, and wholesale leases side by side within the same cluster, as we discuss below.

SemiAnalysis is a reader-supported publication.
To receive new posts and support my work,
consider becoming asubscriber.

Subscribed

ByteDance: TikTok parent, China's largest tenant, among the world's largest GPU users

Every wholesale datacenter conversation in China eventually arrives at the same question: what does ByteDance want? ByteDance is the largest single AI infrastructure buyer in China and, because most of it is leased, it is also the demand engine of the entire third-party colocation industry.

QQ:1413221435

The demand is real and funded by the world's most profitable consumer-app cash machine, Douyin / TikTok.

ByteDance's flagship LLM product is Doubao ("bean bun"). It is China's ChatGPT, not the strongest model on benchmarks, but a friendly, cheap-to-serve consumer product serving hundreds of millions of users. As of March 2026, Doubao was China's most-used AI assistant at 345M MAU, more than Alibaba Qwen (166M) and DeepSeek (127M) combined.

ByteDance also produced the best video models of the past year. Seedance 2.0 topped both text-to-video and image-to-video arenas outright in March 2026. Unlike nearly every other Chinese lab, ByteDance does not open-source its video models. When you are ahead and have enough computing power, you do not open source.

On the B2B side, ByteDance built the AI coding tool Trae and agent platform Coze, with less success. This August, the teams were folded into the Doubao organization. Doubao is now formally designated the company's "main-trunk AI business."

Volcano Engine is ByteDance's external cloud, growing rapidly and in direct competition with Alibaba Cloud. Volcano is also pushing hard into private deployment, where the high-ARPU production workloads in finance, government and industry have historically defaulted to Alibaba, Huawei, Tencent and Baidu.

ByteDance has yet to release a large open-weights model with frontier capabilities, but is reportedly in early stage training a model with as many as 10T total parameters, far exceeding Qwen3.8-Max (2.4T) and Kimi K3 (2.8T).

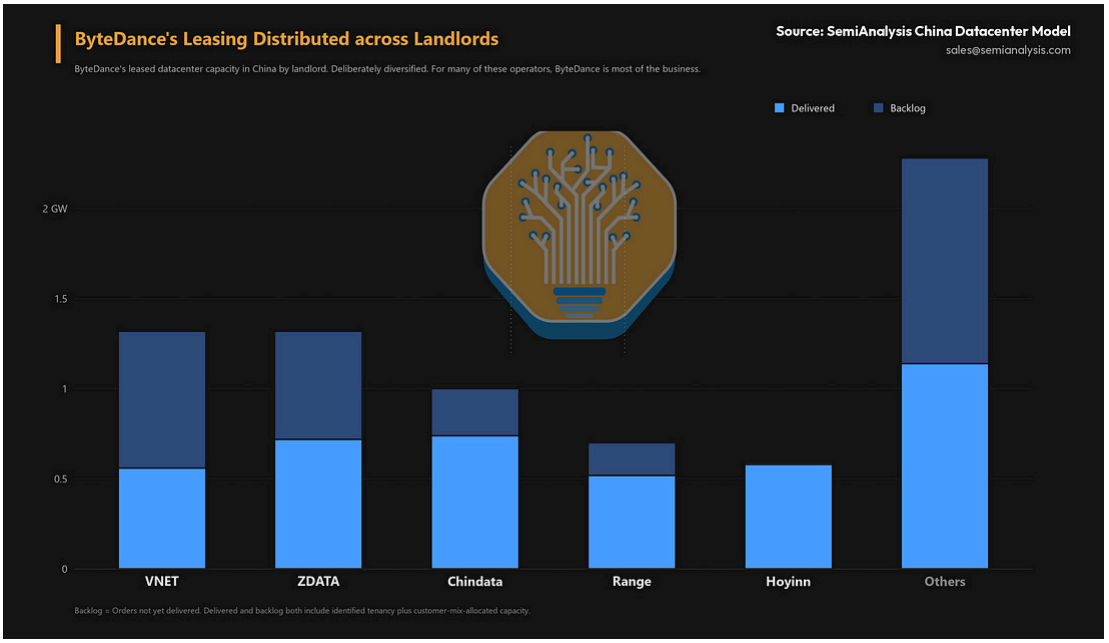
With all of the above business lines layered on top of Douyin's recommendation and ad inference and TikTok's global workloads, total training and inference demand comfortably outruns anything ByteDance can build itself.

The capex trajectory keeps getting revised upward. The 2026 AI infrastructure budget started at ~$24B (~¥160B), with roughly half for chips, and was raised at least 25% to over $30B (¥200B) by May. The most aggressive number in the press is $70B, a ceiling under discussion, not yet a budget.

Behind the capex numbers, ByteDance's domestic datacenter footprint was built almost entirely on wholesale colocation, leasing over 4GW of delivered capacity today across Chindata, ZDATA, VNET, Range, Hoyinn and a long tail, making it the anchor tenant of the entire Chinese wholesale market. Its supplier relationships run long and deep, but diversified across operators to manage supply risk, pricing, and

QQ:1413221435
QQ:1413221435

regional coverage. Because the orders are so large, suppliers develop existential dependence on ByteDance, often at more than 80% of revenue.



Source: [SemiAnalysis China Datacenter Model]

The earliest example is Chindata, ByteDance's most visible landlord. ByteDance still accounted for ~90% of its revenue into early 2026, as disclosed in a filing largely unreported. The company was once US-listed and taken private by Bain Capital in 2023. Its China business was sold to a local consortium in 2026.

ZDATA is the hidden winner. The company has never listed, funded by USD funds including ByteDance's own early investors, with little external data. Our China Datacenter Model shows its operating and under-construction scale now exceeds Chindata's, making it one of ByteDance's largest suppliers.

VNET is the latecomer, pivoting from retail to wholesale on ByteDance's demand at three nodes (Taicang, Huailai and Ulanqab). It booked 862MW of new wholesale orders in 1H26 and crossed 1GW of wholesale capacity in service for the first time, overwhelmingly ByteDance, including a record ~500MW single order phasing across 2026–28.

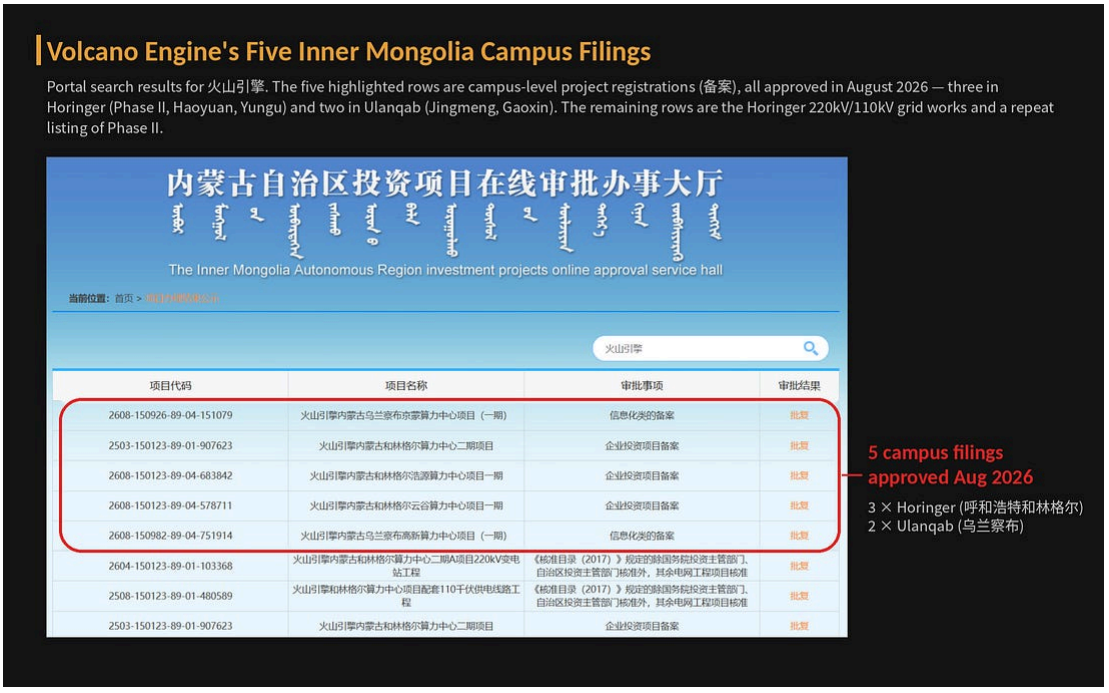
After reaching a certain scale, ByteDance began shifting to owned facilities through Volcano Engine, its cloud and AI infrastructure division. The move is intended to secure long-term capacity for AI training and inference while lowering unit costs and reducing dependence on third-party providers.

ByteDance currently operates three self-built clusters: Shanxi (~500MW planned), Wuhu in Anhui (~500MW planned), and the Inner Mongolia twin-node at Ulanqab

QQ:1413221435
QQ:1413221435

and Horinger (~1GW planned each, with upside to 2-3GW). The fourth cluster is taking shape in Ningxia: per our tracking of corporate registrations, ByteDance stood up two Zhongwei subsidiaries in July 2026, the same register-first, announce-later pattern that preceded Inner Mongolia.

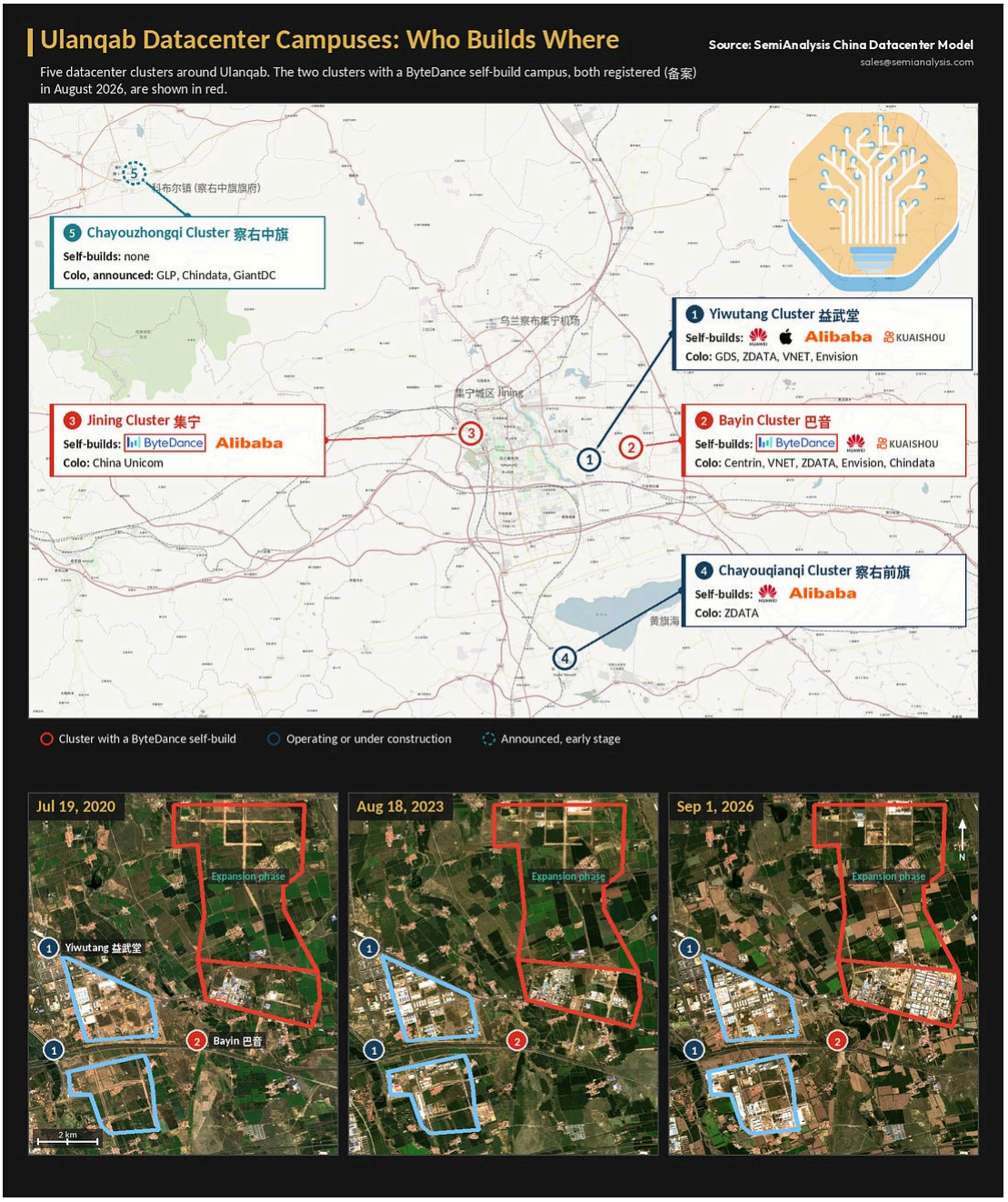
The Inner Mongolia cluster is the most recently built and fastest-growing one. In August 2026, the official government portal disclosed five datacenter project registration approvals for Volcano Engine (ByteDance cloud business), three in Horinger and two in Ulanqab. One is an expansion phase of a campus currently under construction, and the other four are new sites. The approved construction window spans from 2025 to 2029.



Source: The Inner Mongolia Autonomous Region Investment Projects Online Approval Service Hall

We show the core datacenter clusters in Ulanqab and major self-builds and colocation sites within each one in the chart below. The two new ByteDance campuses are located in Bayin and Jining clusters, respectively. We track them all in our China Datacenter Model.

QQ:1413221435
QQ:1413221435



Source: SemiAnalysis

Alibaba: the only true hyperscaler playbook in China

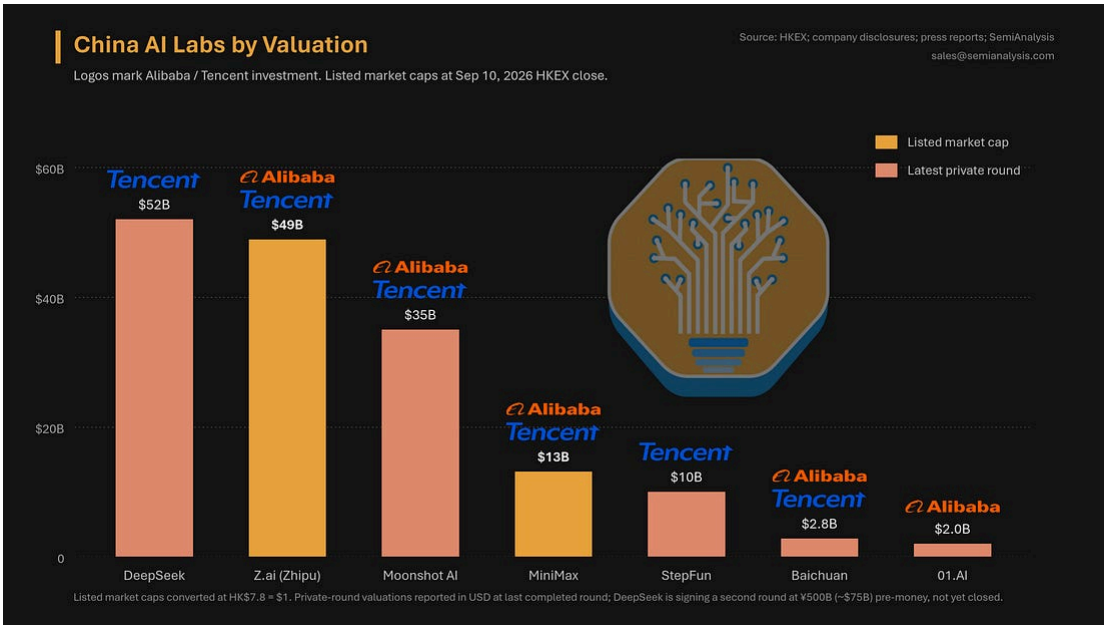
Alibaba is the largest cloud player in China, and the closest thing the country has to an AWS. The playbook rhymes, featuring a dominant IaaS base being upsold into AI, with management explicitly framing MaaS as the value-add layer on top of cloud. Geography does not. Overseas is only ~20% of Alibaba Cloud revenue, but analysts estimate international markets contribute ~30–40% of AWS revenue. Alibaba runs an AWS playbook inside a single addressable market.

The IaaS-to-MaaS transition is showing up in the P&L. AI-related product revenue is now ~35% of Alibaba Cloud's external revenue and guided to approach $10B

QQ:1413221435

annualized. In March, the group set up a dedicated Alibaba Token Hub business group, led by the group CEO himself, to align models, MaaS and applications.

The other half of the strategy is the investment web. Alibaba has invested in five of China's LLM unicorns (Moonshot, MiniMax, Z.AI, Baichuan and 01.AI), and the checks are structured to keep the models inside the Alibaba Cloud ecosystem rather than competing with it. The Moonshot deal is the template: ~$800M for ~36% of preferred equity, of which nearly half was paid in Alibaba Cloud credits. After MiniMax's Hong Kong IPO, Alibaba remains its largest external shareholder at 12%. Every unicorn training run is, one way or another, Alibaba Cloud demand.



Source: HKEX, company disclosures, press reports, SemiAnalysis

Alibaba has emerged as the leader in open-weights models with the Qwen series, with over 3B cumulative downloads and more than 300,000 derivative models. As an e-commerce giant, Alibaba has embedded Qwen into existing business lines as shopping agents and search bots while selling token inference. It is the only Chinese player running a true hyperscaler playbook with frontier models, a dominant cloud, and a captive e-commerce distribution machine, all pulling on the same compute pool.

On the hardware side, Alibaba's T-Head has moved past the pilot stage that stalls most in-house silicon programs and is now serving Alibaba Cloud workloads. Our data shows Alibaba and the carriers are jointly building several datacenters fully deployed with T-Head silicon.

In April 2026, China Telecom and Alibaba Cloud lit a T-Head Zhenwu 10,000-card cluster inside the telecom's Shaoguan campus, an EDWC hub node, with expansion

QQ:1413221435
QQ:1413221435

planned toward 100,000 cards.

The division of labor is as follows. Telecom brings the building, the power and the government-and-SOE sales channel. Alibaba brings everything that computes, from T-Head silicon to the cloud platform and the models on top, sold as full-stack IaaS-to-MaaS.

The carrier co-build is only one example of Alibaba’s flexible datacenter playbook. Our model shows even within its self-build clusters, Alibaba combines third-party build-to-suit leasing with true self-builds to reduce the capex burden.

This is well exemplified by the flagship Zhangbei complex. As we show in the map below, Alibaba mixes the two business models side by side across five campuses. Some are self-builds, while others are halls leased from third-party operators including AtHub, GDS and Yunlian, who put up the shells and power while Alibaba fills them with servers.



QQ:1413221435

Source: SemiAnalysis China Datacenter Model

Tencent: the disciplined one, with a 2026 step-change

Tencent has the most powerful distribution asset in the AI era and moves the slowest because of it. WeChat is China's largest social platform with over 1.4B MAU. With a surface that large, even a minor change ripples across a billion-plus users.

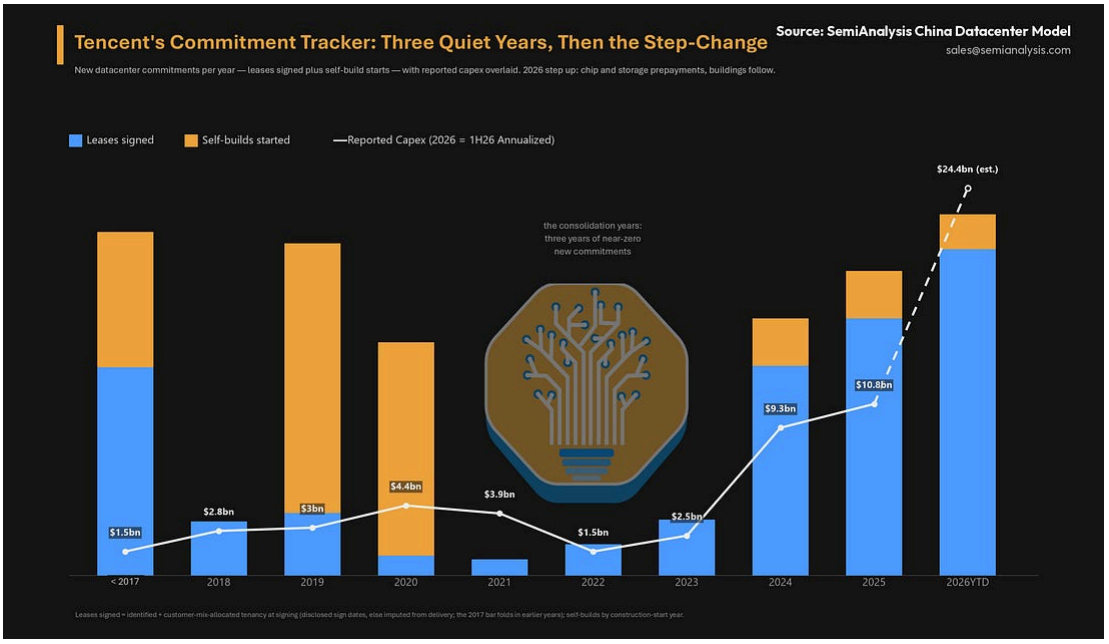
Management admitted at the January 2026 annual meeting that the company was late on AI by "9 months to a year". The catch-up started with a reorganization. By July 2026 all model work had been consolidated into a single foundation-model department under the Chief Scientist.

QQ:1413221435

The model has come up the curve. Hunyuan Hy3 went GA on July 6, 2026 with a 295B-total/21B-active MoE and 256K context, which now powers Tencent's AI chatbot (Yuanbao), coding agent (CodeBuddy), and the social platforms.

Still, the strategy is to own the interface, not the model. Tencent runs its own models alongside DeepSeek, Kimi and GLM across its AI products.

Tencent's datacenter development mirrors the business strategy.



Source: SemiAnalysis China Datacenter Model

QQ:1413221435

The company spent the 2022–23 downcycle consolidating, closing small leased metro sites and concentrating into self-built mega-campuses across Qingyuan,

Huailai and Yizheng. Then, from 2024, capex and commitments came roaring back together.

Capex in 1H26 already exceeds all of FY2025 and free cash flow swung negative, but capacity is growing steadily rather than spiking per our model tracking. That gap is the strongest evidence that incremental dollars are going into chip and storage pre-payments. Most of the incremental capacity now comes from wholesale leases rather than its own shells, a reversal of the self-build-heavy wave as the cash position tightens.

Tencent stayed east while everyone else went west. It has the smallest western footprint compared with ByteDance and Alibaba, consistent with a latency-sensitive workload mix (WeChat, games, video). One exception is a partnership with China Unicom in Zhongwei, Ningxia. The first hall was lit up in April 2025.

This is still the early stage of Tencent's all-in. Unlike announcements and benchmark scores, datacenter capacity is a hard, physical indicator of real computing power, which is why we will keep tracking Tencent's datacenter moves as the clearest test of how committed the catch-up really is.

QQ:1413221435

Baidu: the first mover falling behind, now a silicon play

Baidu invented China's AI story but is now its smallest spender.

The company was China's earliest mover in AI, from LLMs to autonomous driving. Its flagship LLM ERNIE was launched in 2019, and it was the first Chinese tech major to ship a ChatGPT-style product in March 2023. While it led the Chinese market for a few months, ByteDance's Doubao outpaced it and took the consumer crown.

Historically closed-source, Baidu open-sourced ERNIE 4.5 in June 2025, then shipped ERNIE 5.0 (2.4T para) in Jan 2026, and ERNIE 5.1 in May 2026 at decent headline performance metrics.

Like Alibaba, Baidu has integrated ERNIE Bot across search, maps, and coding assistants, but this does not rescue the core business. Online marketing revenue was down ~20% YoY in 2Q26. The exception is private deployment. ERNIE sells well to governments and SOEs.

QQ:1413221435

The capex line tells the same story. Baidu is the smallest spender of the four internet hyperscalers. The 1H26 capex is roughly what Alibaba spends in a month.

But Baidu runs the deepest vertical stack. Kunlunxin chips trained ERNIE 5.1 in Baidu's own datacenters (Alibaba's T-Head Zhenwu, by contrast, is used mostly for inference). Kunlunxin also sells outward, especially for state and financial procurement. It is being spun out for an A+H listing with the upper end of its mooted valuation above the parent's.

The datacenter map echoes this. Headline capex was up 201% YoY in 2Q26, a 45% core-capex-to-revenue ratio, the highest intensity of any Chinese platform. But our China Datacenter Model shows the smallest infrastructure footprint of the five, with third-party tendering running at a tenth of ByteDance's. No gigawatt campuses in the western hubs (Baidu's first Inner Mongolia compute order only arrived in 2025), and no overseas compute.

QQ:1413221435

The gap between the capex line and the buildings is mostly chips and storage, but Baidu isn't stocking up to chase a frontier training run. It is increasingly running a silicon business.

This is visible in the P&L. In 2Q26, GPU cloud revenue grew 283% YoY, accelerating from 184% in 1Q. Management spent the earnings call talking about returns, not capacity. GPU cloud, they note, carries a better margin profile than traditional CPU cloud, "underpinned by the cost advantage from our self-developed chips."

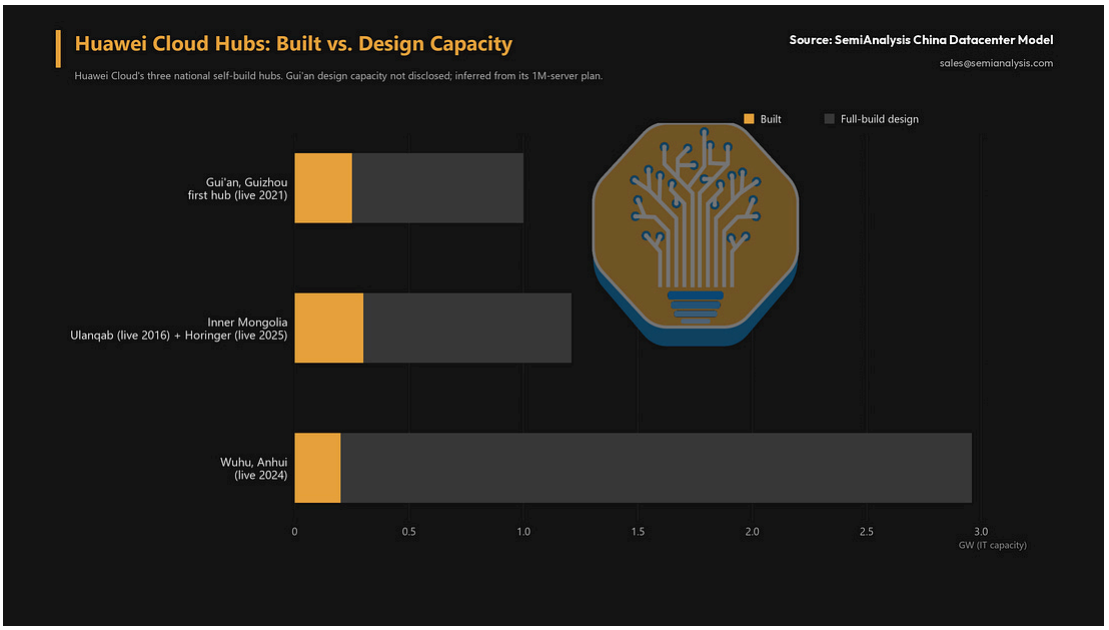
Huawei: the silicon supplier that also operates datacenters

Huawei is in this report for a different reason than the other four. It is simultaneously the second-smallest hyperscaler and the supplier that determines everyone else's ceiling.

As a datacenter operator, Huawei builds almost everything itself, leading to the highest self-build ratio of any player we track, over 90%.

QQ:1413221435

The footprint is anchored by three cloud hubs, each designed for a million servers or more.



Source: [SemiAnalysis China Datacenter Model]

Launched in 2021, Gui'an, Guizhou is the largest Huawei Cloud campus in operation, designed to hold 1M servers.

The second hub sits in Inner Mongolia across two sites. Ulanqab is Huawei's oldest self-build, dating back to 2016. The adjacent Horinger campus broke ground in 2023 and is designed to become Huawei's largest footprint in North China with 1,000 acres of land, 50 planned server halls, six dedicated 220kV substations, and over 1M servers at full build.

The Wuhu, Anhui campus opened in 2024 to complete the three-hub national layout. It is the most ambitious of the three, designed to host 3M servers on 1,000 acres of land.

Our model shows a clear pattern across all these campuses. The design capacity runs far ahead of deployment, with the operational-to-planned ratio around 1:6.

The right way to think about Huawei Cloud is not as a cloud business at all, but as a distribution channel for Ascend compute. It lets customers who cannot buy or operate CloudMatrix hardware rent it by the hour or by the token, converting Huawei's silicon advantage into recurring revenue. Huawei Cloud was less than 5% of group revenue in 2025, with a shrinking market share. The customer base is mainly government, finance and SOEs rather than internet-native workloads. Policy is pulling in the same direction. Since November 2025, new state-funded datacenters have been required to use domestic AI chips.

Hence the paradox: Huawei has the second-smallest hyperscaler footprint of the five and the largest exposure to everyone else's. Every gigawatt that state AIDC programs

QQ:1413221435
QQ:1413221435

energize from here forward is a bid for Huawei’s supply.

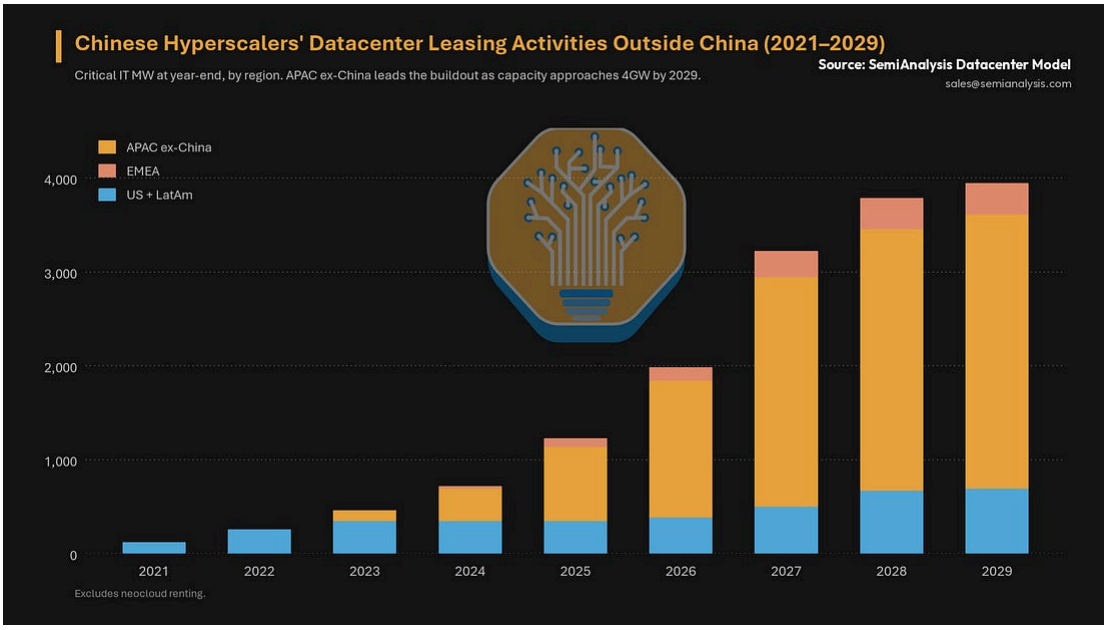
SemiAnalysis is a reader-supported publication.
To receive new posts and support my work,
consider becoming a free or paid subscriber.

Subscribed

The buildout doesn’t stop at the border: overseas capacity reaching ~4GW by 2029

In addition to the Chinese construction surge, Chinese tech titans are carrying out large-scale investments all around the world.

QQ:1413221435



Source: SemiAnalysis Datacenter Model

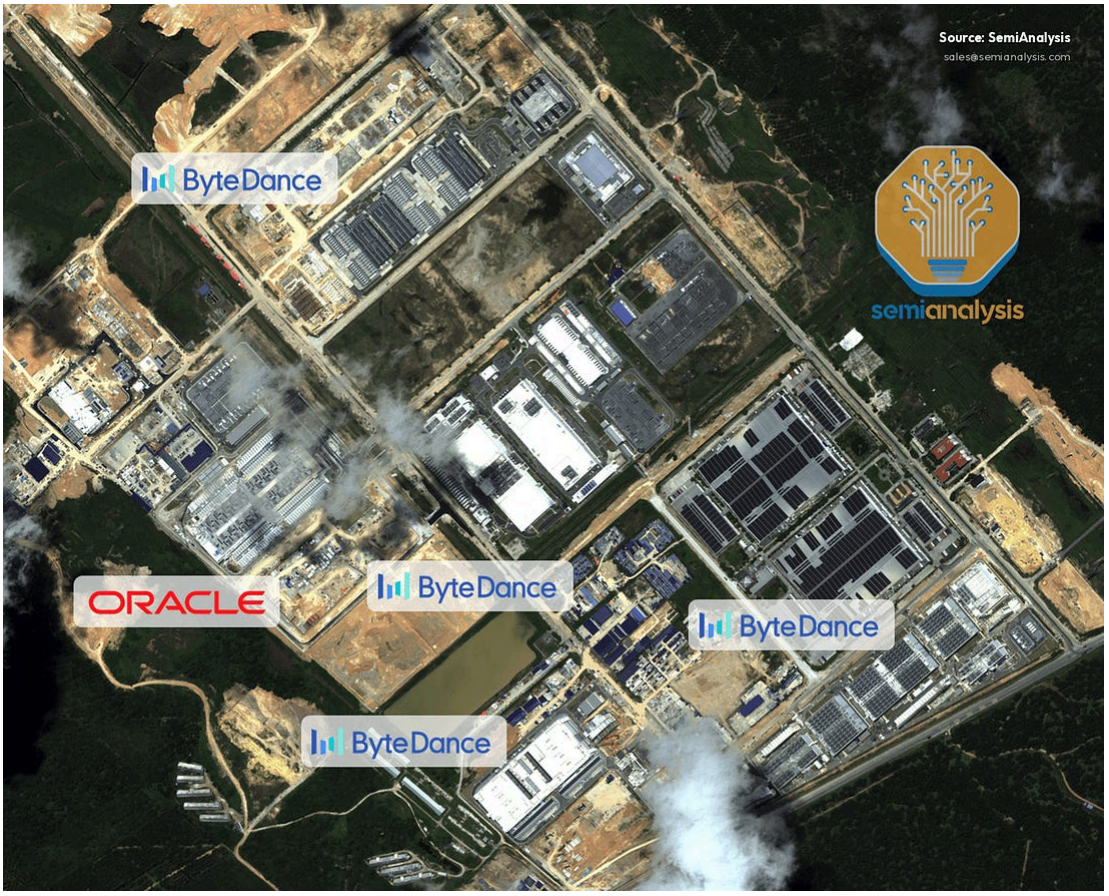
ByteDance is the clear leader. TikTok’s effort to process user data locally led it to contract significant capacity outside China in 2020–24. While the firm doesn’t self-build outside China, we track its major leases and neocloud rentals.

In the US, TikTok expanded aggressively in the 2020-2023 period, concentrated in Northern Virginia and Portland, but has added little of its own capacity since. In Europe, TikTok opened datacenters in Ireland and Norway in 2023, with a third site in Finland breaking ground in 2025. The most recent move is in Brazil, starting at ~200MW.

QQ:1413221435

Unlike the slow ramp up in the US and Europe, ASEAN is the current bright spot where ByteDance has accelerated sharply since 2023, as we covered in our ASEAN work and in How Oracle Is Winning the AI Compute Market.

Export controls cap what silicon can run onshore, so ByteDance's Nvidia fleet lives offshore in facilities built by developers who brought Chinese construction speed with them. Core landlords include DayOne, Bridge DC, and a wave of privately funded developers. Capacity is concentrated in the Singapore-Johor-Batam hub.



Source: SemiAnalysis Datacenter Model

These operators build at a relentless pace, delivering 100MW+ in under a year. Some established Western wholesale developers routinely take twice as long.



QQ:1413221435

Source: SemiAnalysis Datacenter Model

Alibaba and Tencent have also been active in overseas expansion, although at a smaller scale.

Alibaba is the more committed of the two. It currently runs 46 availability zones across 18 international regions, versus 60 availability zones and 13 regions in China. The pace of new region and availability zone launches has visibly accelerated over the past two years.

The expansion is funded in hard currency. In August 2026, Alibaba closed a $10B placement of new shares, which is the company's first new-share placement since its 2019 Hong Kong listing and the largest follow-on offering in Hong Kong history.

The filing is explicit about where the money goes. All to full-stack AI, with roughly 60% earmarked for expanding global computing infrastructure and the remainder for hyperscale AI datacenters and cloud upgrades.

This echoes our July call that Alibaba is looking for ~500MW capacity in the US by the end of 2027.

The demand driver is partly Model-as-a-Service business, with a twofold growth target: to increase AI-related services from 35% share of cloud revenue to over 50%, and to expand overseas revenue from ~20% of the cloud business to 25–30% within three years.

The other driver is access to advanced chips. As we covered extensively in GPU Remote Access and More, renting GPUs to Chinese hyperscalers is at present a legal and sizable business.

Tencent's expansion is smaller and started by serving the overseas push of Chinese gaming and internet companies. But the pivot is similar, expanding to three new overseas regions (Saudi Arabia, Japan, Malaysia) and adding 8 international availability zones since 2025, driven by the international rollout of its model-serving gateway and agent products, with Southeast Asia the clear focus.

International business is currently ~10% of Tencent cloud revenue, but growing north of 40% year over year. The stated ambition is over 30% by 2030.

QQ:1413221435

The bottom line

QQ:1413221435

Almost nothing in this report is visible from headline numbers. High vacancy sits alongside an AI capacity shortage. A record capex cycle runs on falling facility costs. The market's largest tenant discloses almost nothing yet anchors the entire wholesale industry. Each puzzle resolves only when the market is mapped building by building, and that is what the **SemiAnalysis China Datacenter Model** is: a bottom-up map of what is being built and leased, tenant by tenant, quarter by quarter.

As with their Western peers, the majority of capex will be driven by servers. But those servers need buildings, and datacenter construction is the earliest hard signal of where and when they will land. Because the model tracks under-construction and dated pipeline projects, the committed portion of the buildout is already in the data.

As we highlighted at the beginning, China's datacenter industry is entering an era of unprecedented expansion. The combined capacity of the country's leading hyperscale operators, Alibaba, Tencent, Huawei, ByteDance, and Baidu, is on track to more than triple from the start of the AI cycle: from roughly 6GW at end-2023 to ~21GW by 2030, with annual additions accelerating from ~1GW a year in the cloud era to 3-4GW a year in 2026–2027, before another ~6GW of their own speculative campus backlog breaks ground.

The demand behind that pipeline is structural. A protected domestic market, state-directed infrastructure, and an open-source flywheel that keeps Chinese models globally relevant despite export controls.

Subscribers get the map today and continued tracking of China's AI buildout from here.

QQ:1413221435

Subscribed

Like Comment Restack

© 2026 Dylan Patel
548 Market Street PMB 72296, San Francisco, CA 94104
Unsubscribe

Start writing

QQ:1413221435