Not for redistribution without written consent of Morgan Stanley

Morgan Stanley | RESEARCH

INSIGHT

September 13, 2026 10:28 PM GMT

China's AI Path

From Rmb8.5tr Capex to 3x Compute

We forecast Rmb8.5tr total capex, 2026-30E, supporting IT capacity reaching 81GW by 2030E. We address the financing need and ROIC. We are OW AI infra (BABA, KC, VNET), AI labs (MiniMax, Zhipu) and GPU localization.

Key Takeaways

- We forecast total capex of Rmb8.5tr (US$1.3tr), 2026-30E, in China, equating to, on average, 17% of the US peer total.
- Based on our capex and domestic AI shipment forecasts, we expect IT power capacity to grow from 26GW in 2025 to 81GW in 2030E.
- We see potential offshore financing needed to support a total of Rmb2.0tr (US $294bn) offshore capex, 2026-30E.
- We continue to see 10-20% ROIC and further to 30% for Maas (3P APIU) on high-performing chips but lower to 9% for net generation domestic chip.

Tech Diffusion

A Morgan Stanley Research Key Theme of 2026

Capex - Rmb8.5tr over 5 years: We forecast total capex of Rmb8.5tn (US$1.3tn) over 2026-30E, including hyperscalers, neoclouds and telcos. We see China capex reaching ~17% of the US peer level, on average.

Compute - 81GW IT capacity by 2030E: We estimate IT power capacity in China will expand from 26GW in 2025 to 81GW in 2030, with hyperscalers the main contributor. We forecast Chinese hyperscalers adding 47GW of compute globally in 5 years, with net adds in 2026-28E, equating to ~22% of the US peer level.

Financing - Healthy hyperscaler balance sheets, but offshore financing might be needed: We see CSPs supporting AI development via four avenues: 1) by spending capex, 2) with prepayments to key supply chain partners, 3) by contract commitments for long-term visibility, and 4) by investing upstream and downstream. That said, we think offshore cash financing needs are the key consideration for CSP fundraising over the next five years to support Rmb2.0tr of offshore capex.

ROIC - Domestic AI chips also deliver decent return: We continue to see a 10-20% ROIC (2.5-3 years payback) for IaaS/MaaS (1P) and 30% (~2 years payback) for MaaS (3P), but lower at a 9% ROIC (3.6 years payback) with the upcoming generation of domestic chips. However, this could improve further as domestic localization continues to advance.

MORGAN STANLEY ASIA LIMITED+
Gary Yu
Equity Analyst
Gary.Yu@morganstanley.com +852 2848 6418

Tom Tang
Equity Analyst
Tom.Tang@morganstanley.com +852 3963 2810

Yang Liu
Equity Analyst
Yang.Liu2@morganstanley.com +852 2239 7311

MORGAN STANLEY TAIWAN LIMITED+
Charlie Chan
Equity Analyst
Charlie.Chan@morganstanley.com +886 2 7730 2765

MORGAN STANLEY ASIA LIMITED+
Rebecca Xu
Equity Analyst
Rebecca.Xu@morganstanley.com +852 2848 2749

Lydia Lin
Equity Analyst
Lydia.Lin@morganstanley.com +852 2239 7567

Joanne Lau
Research Associate
Joanne.CY.Lau@morganstanley.com +852 3963 2502

Eddy Wang, CFA
Equity Analyst
Eddy.Wang@morganstanley.com +852 2239 7039

Kathy Zhu
Research Associate
Kathy.Zhu@morganstanley.com +852 3963 2618

CHINA INTERNET AND OTHER SERVICES
Asia Pacific
Industry View **Attractive**

Key reports:

Alibaba Group Holding: Path to 13-20% ROIC (17 Aug 2026)
Kingsoft Cloud Holdings: Transition to AI Cloud Bearing Fruit; MaaS Brings Further Upside (6 Jul 2026)
China's Emerging Frontiers: China's AI Path: More Bang For The Buck (27 Apr 2026)
China's Emerging Frontiers: China's AI Path: Monetizing Surging Token Use via AI Cloud (16 Mar 2026)

Morgan Stanley does and seeks to do business with companies covered in Morgan Stanley Research. As a result, investors should be aware that the firm may have a conflict of interest that could affect the objectivity of Morgan Stanley Research. Investors should consider Morgan Stanley Research as only a single factor in making their investment decision.

For analyst certification and other important disclosures, refer to the Disclosure Section, located at the end of this report.

+= Analysts employed by non-U.S. affiliates are not registered/qualified as research analysts with FINRA, may not be associated persons of the member firm, and may not be subject to FINRA Rule 2241 restrictions on communications with a subject company, public appearances and trading securities held by a research analyst account.



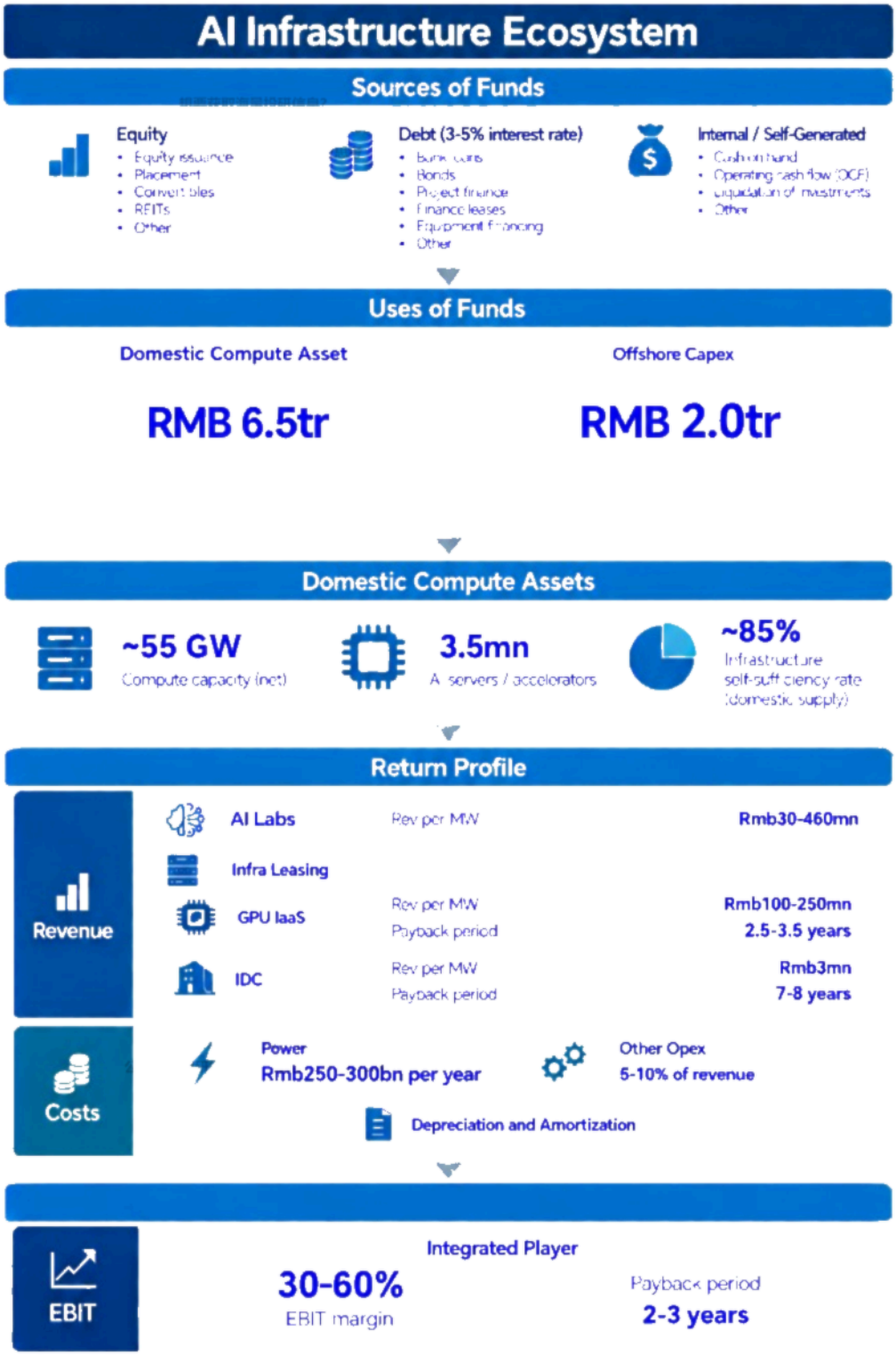
Key stock calls: We continue to prefer the China cloud supply chain backed by healthy financing and strong ROIC – we stay OW on key infrastructure providers Alibaba, Kingsoft Cloud and VNET; AI labs Zhipu and MiniMax; and GPU localization Cambricon, Illuvatar and Hygon.

China's Emerging Frontiers: China's AI Path: Owning the Full AI Stack via In-house Chips (11 Mar 2026)

China's Emerging Frontiers – Data Centers: Inflection and Disruption in Remote Areas (21 Jan 2026)

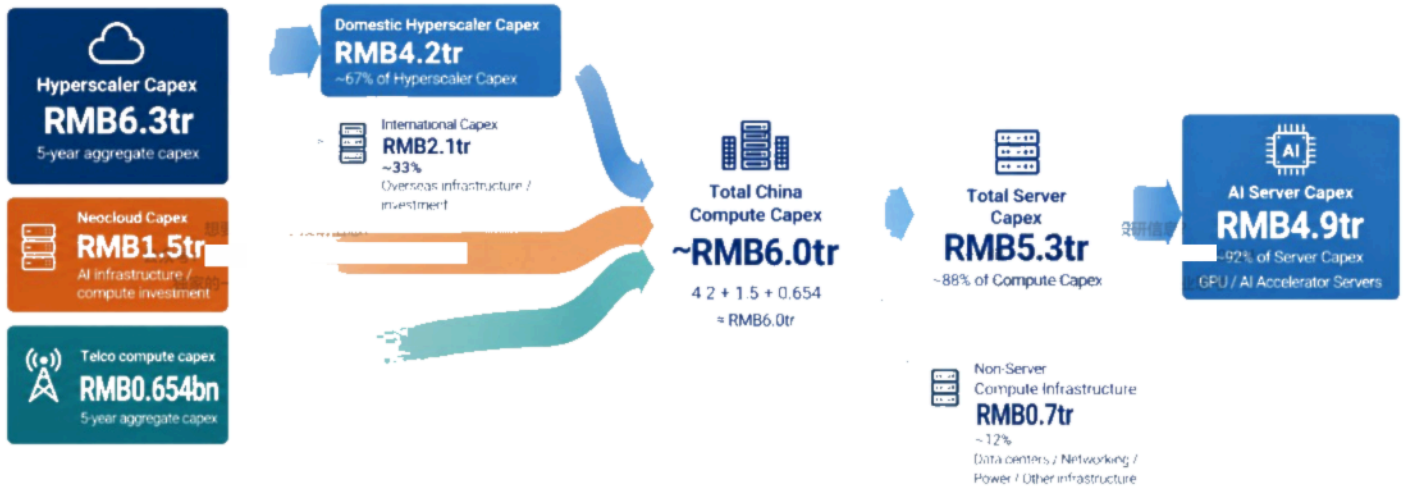
Key Charts

Exhibit 1: Our report in 1 chart: Fund Flow through the AI Infrastructure Ecosystem



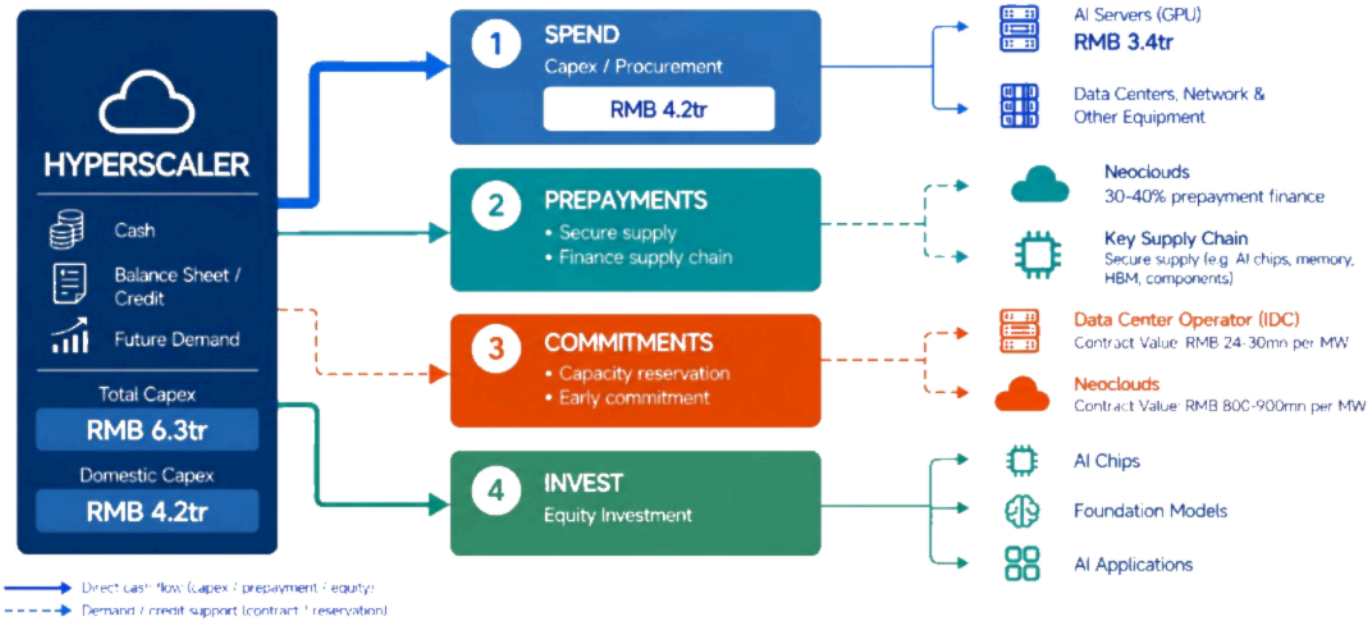
Source: Company data, Morgan Stanley Research estimates. Our ARR per MW is calculated based on a range of assumptions on TFS, token cost and training/inference mix, for which we have details in the ROIC section.

Exhibit 2: How is capex being spent over the next 5 years



Source Company data, Morgan Stanley Research

Exhibit 3: Hypsercalers' investment and business across major vendors



Source Company data, Morgan Stanley Research

Executive Summary

1. Capex: China AI spending is stepping up, with upside risk from hyperscalers

We forecast total China capex of Rmb1.3tr (US$193bn) in 2026E, growing further to Rmb2.0tr (US$291bn) in 2028E. The 5-year total spend of Rmb8.5tr (US$655bn) would equate to ~17% of US peers, we estimate.

First pillar - Core hyperscaler spending: Total **hyperscaler +key internet capex of Rmb995bn/US$142bn in 2026E, rising 23% YoY, to Rmb1.2tr/US$175bn in 2027E, and further to Rmb1.4tr/US$205bn in 2030E, 13-19% of US peers**. China hyperscaler growth of 121% in 2026E will outgrow US peers at 41% in 2026E, following a 2025 setback due to China's capacity constraint. We forecast 65-70% of the China capex to be spent domestically.

Second pillar - Neoclouds adding high-end servers: Beyond hyperscalers, we estimate **neocloud investment at a total of Rmb1.5tr (US$227bn) over the next 5 years,** as an important high-end compute addition to hyperscaler capex. We expect most of the spending to be on server procurement.

Third pillar - Telco operators: Telcos provide a third pillar of demand. We estimate roughly **Rmb81bn of computing capex in 2026E**, with computing capex growing at a 23% five-year CAGR, to Rmb179bn by 2030E, even as overall telco capex remains broadly stable.

2. Compute: Domestic chip supply should unlock the next leg of capacity growth

We believe China's AI compute expansion has so far been constrained less by demand than by availability of high-end compute. As domestic GPU/ASIC supply begins to ramp, we expect this bottleneck to ease, and we forecast **total compute capacity to increase from 26GW in 2025 to 81GW by 2030E.**

The IT power capacity expansion of 55GW is driven primarily by:

- **Hyperscalers:** +34GW from 2025 to 2030E;
- **Telcos**' self-built external capacity: +9GW;
- **Other players/AI labs:** +12GW in our base case, with upside potential.

Within hyperscalers, we expect Alibaba and Tencent to each add 6-9GW, and Baidu around 1.0GW. We believe that increasing use of domestic GPUs and proprietary ASICs will become an important enabler of capacity growth, particularly as Alibaba, Baidu, and potentially other platforms deepen vertical integration in silicon.

Globally, we expect hyperscalers to add 47GW of compute in 2026-30E; while the net addition of 24GW seen for 2026-28E is ~22% of US peers.

The associated infrastructure requirement is substantial. We forecast incremental AI power demand increasing from 5.6GW in 2026E to roughly 13GW in 2030E, while third-

party datacenter orders rise from 6.1GW to 11GW, assuming roughly 20% of capacity is self-built by hyperscalers and AI labs.

3. Financing: Funding needs differ across the ecosystem; Hyperscalers may need offshore financing

We see an important distinction between China and the US. China's hyperscalers generally have stronger bargaining power and greater vertical integration, while ecosystem investments remain more selective. As a result, domestic AI capex is funded primarily through operating cash flow, cash balances, leases and commercial prepayments rather than large-scale strategic cross-financing.

We expect domestic capex to be well funded. We view BAT well funded for onshore capex given strong OCF generation. However, this is not completely without risks - BABA may have greater cash drain if food delivery competition is re-ignited, and Baidu's search stabilization is also a key variable. We think Tencent is the most stable of the three.

The greater constraint is offshore funding, where capex, shareholder returns and offshore debt repayments compete for the same pool of cash. We see hyperscalers spending a total of Rmb2.0tr (mainly contributed by ByteDance) in 2026-30E; additional financing may be needed through 1) monetizing investment portfolio, 2) debt and 3) equity financing.

In contrast, we see materially greater financing need for neoclouds, but the funding sources are also diverse. Typical funding structures **include 30-40% financing leases at ~5% lease interest rate, bank loans at roughly 3-4%, customer prepayments and incremental debt/equity issuance**. These structures support rapid capacity expansion, but they also increase sensitivity to utilization, delivery schedules, refinancing costs and server demand.

4. ROIC: Returns improve materially as monetization moves up the AI stack

Our ROIC framework suggests that the key question is not simply how much capex is being spent, but what monetization model sits on top of that investment.

- **Self-owned GPU IaaS:** We estimate approximately 44% operating margin, 13% ROIC and ~3-year cash payback under our base case.
- **Rented/neocloud IaaS:** We estimate around 20% operating margin. Margins are lower, but there is effectively no upfront server capex, allowing immediate positive cash generation once the rental spread turns positive.
- **Self-built compute + self-built model:** We estimate 53% operating margin, 19% ROIC and 2.5-2.6-year payback.
- **Self-built compute + third-party API:** Assuming 10% revenue sharing, our base case reaches approximately 57% operating margin, 29% ROIC and 2.1-year payback.

We also see potential for domestic chips to materially increase computing capacity, although this might result in a lower ROIC in the near term, due mainly to lower performance.

- **P/D disaggregation** - With regard to adopting Huawei 950PR for prefill, we think

adopting PD disaggregation can result in marginally better ROIC, in the case where input tokens are higher (such as summarization, long document Q&A).

- **Full domestic infrastructure** - With next-generation domestic capacity coupled with HBM 3E, we see potential to run the full inferencing process; We estimate base case 31% operating margin, 9.1% ROIC and 3.6 years of payback, still lagging foreign servers.

Key Stock Calls

China /Cloud – Alibaba, Tencent, Kingsoft Cloud – We continue to like the cloud service providers in China; we believe strong demand for compute and supply-chain relief should further add to growth potential.

China datacenter – VNET – We like VNET for being the leader in rural area development in hubs, such as Ulanqab in Inner Mongolia.

China LLM - We see computing growth supporting further growth of China model players. OW on MiniMax and Zhipu.

China Semi Localizaiton - We remain optimistic on GPU localization including Cambricon, Illuvatar and Hygon.

China to spend Rmb8.5tr total capex in 5 years

We forecast total compute capex spending of Rmb1.3tn (US$193bn) in 2026 to expand to Rmb2.0tr (US$291bn) by 2030E. As a % of US hyperscaler + neoclouds, China's capex will rise from 18% in 2025 to 21% in 2026E.

By category, we forecast:

- Hyperscaler capex to grow from Rmb995bn (US$142bn) in 2026E to Rmb1.4tr in 2030E (US$205bn)
- Neocloud capex to grow from Rmb324bn (US$34bn) in 2026E to Rmb363bn (US$53bn) in 2030E
- Telco compute capex to grow from Rmb81bn in 2026E to Rmb179bn in 2030E

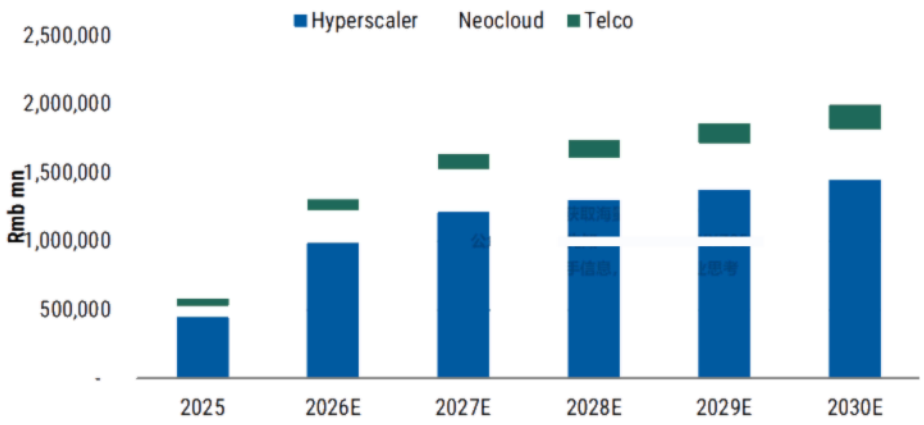
Domestic capex will make up 70-80% of the spending, aggregating to Rmb6.4tr over the next 5 years.

Exhibit 4: How is capex being spent in the next 5 years



Source: Company data, Morgan Stanley Research

Exhibit 5: Total China capex, by category



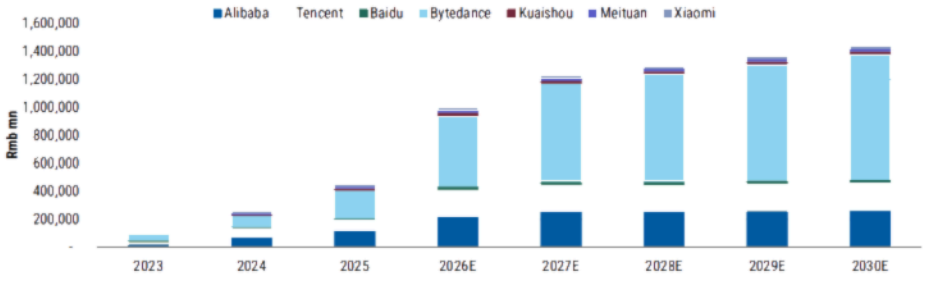
Source: Company data, Morgan Stanley Research estimates

Hyperscalers – Capex of Rmb6.3tr seen over 2026-30E

After the June-quarter results, we have updated our capex forecasts for major capex spenders in China. We forecast that total cloud capex will grow 121% YoY, to Rmb995bn (US$142bn) in 2026E, outgrowing the US peers (41% YoY) since 2023. We forecast it to further grow 23% YoY, to Rmb1.2tr (US$175bn) in 2027E, and to Rmb1.4tr (US$205bn) in 2030E. After the June-quarter result, we have raised:

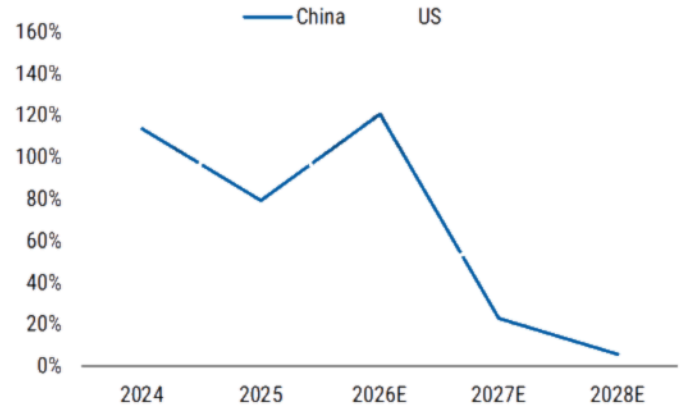
- Tencent's capex forecast to a Rmb193bn-Rmb200bn range. Despite the company's reported Rmb50bn capex in the second quarter, we do not expect the same quarterly run rate throughout the year, as there is a seasonality pattern within the capex spending.
- Alibaba's capex forecast to a range of Rmb217 - Rmb259bn, 2026-2030E on a calendarized basis. We expect Alibaba to spend more than Tencent, given a clearer visibility to its ROIC through AI GPU IaaS as well as MaaS. In the June quarter, Alibaba's MaaS ARR has surpassed Rmb10bn and is well on track to reach Rmb30bn by the end of the year. Management also indicated that the cloud payback period is 2.5 years and that ROIC could be above our previous estimate.
- Baidu's capex forecast to a Rmb20-24bn range rate after its result and its stated ambition to grow its compute capacity.

Exhibit 6: Total hyperscaler capex - 13-18% of US peers



Source: Company data, Morgan Stanley Research estimates

Exhibit 7: China vs US hyperscaler capex growth



Source: Company data, Morgan Stanley Research estimates. US hyperscalers include Amazon, Microsoft, Meta, Google and Spacex

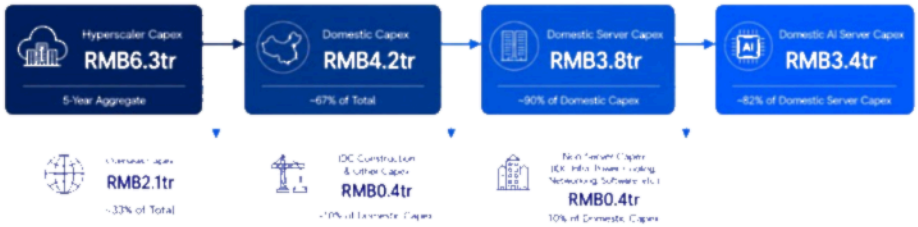
How is capex spent?

Comparing with the global hyperscalers, we expect a larger share of the capex to be spent directly on compute-related product, given other items, such as datacenter construction, are much cheaper. We assume the following:

- 30-35% overseas capex - this is mainly contributed by 20-40% of the top 3 hyperscalers' overseas expansion.
- 10-15% of domestic capex as non-server procurement - which will be going to datacenter rollout and other networking equipment
- 10% of server capex as CPU - we think this number could climb because of the rising importance of CPU in agentic workflows.

The remainder will be spent on GPU-related computing, which includes the entire server (including memory spending).

Exhibit 8: How we expect CSP capex to be spent over the next 5 years



Source: Company data, Morgan Stanley Research estimates

Neoclouds – Rmb1.5tr seen over 2026-30E

A challenging part of sizing total AI spending is that hyperscalers usually engage in off-balance-sheet financing through lease commitments to manage cash flow and source additional computing power outside of the domestic supply chain. Information on this tends not to be transparent, given that even if the lease commitment is disclosed, there is not a clear classification on what has been spent on what – i.e., whether it is computing equipment or just office buildings, etc.

Only Tencent has a clear line of showing such expenditure, which comes under **technical infrastructure operations costs**, which we can infer as the quarterly lease fee paid for computing-related infrastructure. Assuming that additional leases paid per quarter after 2023 are attributed to AI computing, we note that Tencent's lease fee increased by Rmb6bn from 4Q25 to 2Q26, from which we estimate a total lease commitment of Rmb24bn during the first half of the year.

Bottom-up sizing. We size neocloud spending based on the amount of high-end server acquisitions and an estimated cost per server of Rmb5-10bn, depending on the supply chain tightness. **We forecast a total neocloud investment of Rmb234bn (US$34bn) in 2026E to Rmb363bn (US$53bn) in 2030E.**

We also analyzed historical data for listed neocloud companies, including KC, and names such as Sharetronic Data Technology (300857.SZ, Not Covered) and Range Intelligent (300442.SZ; Not Covered), to size their total actual spending through the following lines:

- **Cash capex** - This is the most straightforward spending, from which we can also separate the computing-equipment-related capex on an annual basis by looking at the PPE breakdown or construction in progress.
- **Leases** - These companies also sign financing lease agreements with banks or financial service providers to lease back the computing equipment as a means to manage cash outflow and debt burden; this is partially reflected in the right-of-use assets on the statements.

Combining the 1H26 results of the above companies, the total capex, including lease arrangements, reached Rmb47bn. We think this number could be higher in the second half, given the continuously trending higher server procurement and leasing prices in China, although we also note that channel uncertainty is the key overhang to their procurement capability. Using Kingsoft Cloud as an example, we forecast total investment of Rmb15bn in 2026E, and the current completion rate is 40%.

How capex will be spent: For neoclouds, capex spending tends to be straightforward. We expect the majority of the cash to be used on server procurement (>95%), given how expensive it is to procure high-end servers in China at present.

Telcos to Spend Rmb653bn compute capex seen over 2026-30E

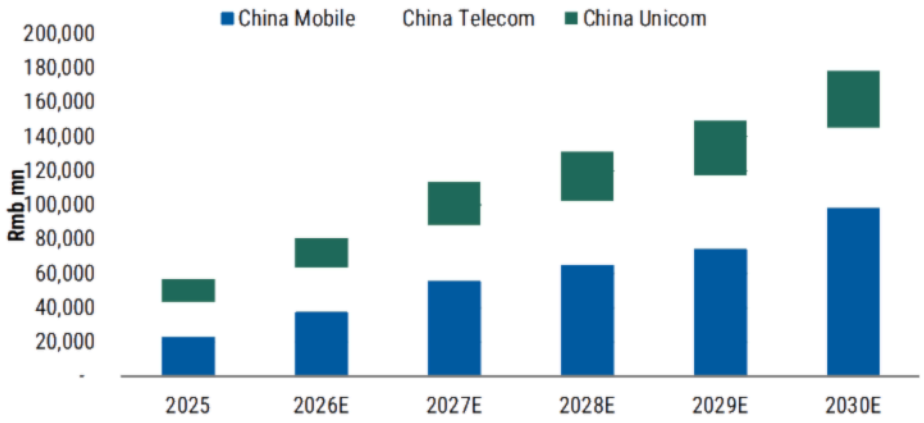
Chinese telcos will be the major investment platform for sovereign AI, in our view, given their responsibility to serve the government and SOE clients. Historically, the telcos have given the market clear guidance on their one-year forward spending, and they usually deliver this quite on track. As of 2026E, we expect total spending of ~Rmb81bn from the telcos, including:

- **China Mobile** - Rmb38bn on computing capex
- **China Telecom** - Rmb26bn on computing capex
- **China Unicom** - Rmb18bn on computing capex.

We continue to expect the Chinese telcos to reduce their total capex before meaningful 6G deployment (likely 2029E and onwards). Computing capex should become a larger part of the budget share and enjoy faster growth than traditional capex.

We forecast that Chinese telcos' total capex will remain stable until 2030, but that computing capex will expand from Rmb81bn in 2026E to Rmb179bn in 2030E, a 10% CAGR.

Exhibit 9: Total telco computing capex



Source: Company data, Morgan Stanley Research estimates

How capex will be spent

We expect the Chinese telcos, on average, to spend ~50% of their total computing capex on non-compute-related items, such as datacenter rollout and networking, while the remainder will be spent on computing power, with the majority (75%+) being smart computing or GPU/ASIC-related spending. In general, we think most of the compute equipment will go to domestic GPU/ASIC vendors, given the importance of national security.

Exhibit 10: How capex will be spent



Source: Company data, Morgan Stanley Research estimates

Domestic AI chip supporting expansion towards 81GW total power capacity

Backed by strong domestic chip ramp-up from 2H26, we expect China compute capacity to expand and hence have raised our forecasts. We now expect:

- Total IT capacity to expand from 26GW in 2026E to 81GW by the end of 2030.
- We forecast hyperscaler capacity to expand by 34GW – to 44GW, from 2025 to end-2030E.
- Datacenter third-party orders to expand from 6.1GW in 2026E to 11GW in 2030E.

Globally, we forecast total net adds of 47GW by hyperscalers in 5 years, with 24GW net adds in 2026-28E – 22% of the 108GW level for US peers.

We expect 81GW IT power capacity 2030E

Why is domestic chip capacity so important? Despite being the No.2 AI spender globally, China's compute capacity additions have been lagging those of the US and other regions in the past year, due mainly to the constraints on chip supply.

Based on our model, we estimate the total IT capacity in China was 26GW as of 2025. Net computing capacity addition in China was a respective 2.3GW and 3.5GW in 2024 and 2025.

We think this is set to accelerate with domestic chip capacity ramp-up in the next few years, giving China the necessary compute supply.

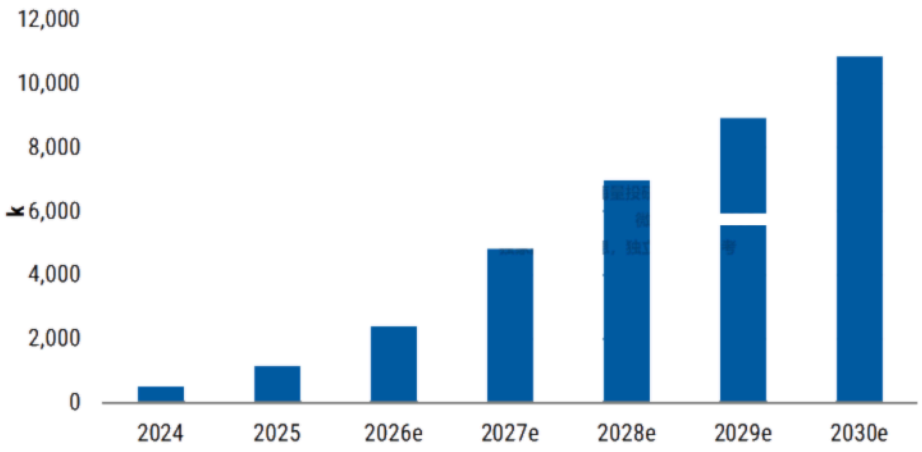
Working with our Semis team, we tried to bridge the capacity additions from domestic chipset to datacenter capacities with the following procedure:

1) Shipment and production

Our Semis team forecasts total AI chip shipments of 1.1mn in 2025, expanding to 2.4mn (+100%+YoY) in 2026E, 4.8mn in 2027E and further to 11mn by 2030E. This will result in a domestic chip TAM expanding from Rmb94bn in 2025 to Rmb646bn by the end of 2030E.

We expect domestic chips to account for 70-85% of the total server deployment throughout 2025-30E.

Exhibit 11: Domestic AI chip shipments



Source: Company data, Morgan Stanley Research (e) estimates

2) Cost of chip per MW

We estimate the total cost of chip procurement in China is US$50-70bn per GW for high-end training chipsets, while domestic chips mainly used for inference will be cheaper, at US$19bn per GW in 2026E, and US$22bn per GW in 2027E, on average. The price may vary depending on the chip vendor – with Huawei likely to be the most expensive, at US $42bn per GW, per our assessment.

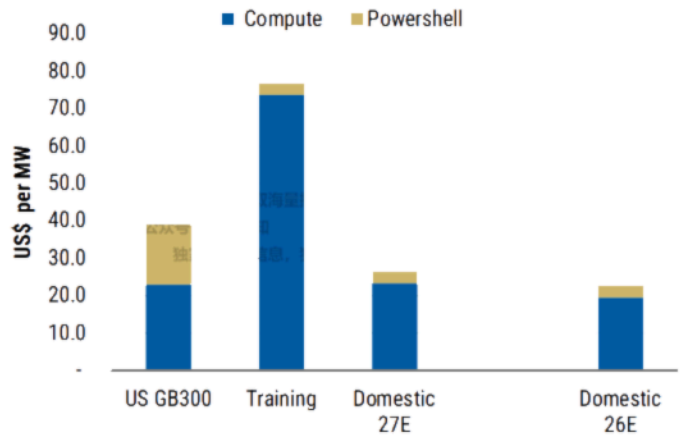
We forecast our cost per GW based on the ASP of a single GPU from our Semis team – assuming the standard 8x GPU per server – and we multiply that by factor of 3-4x, accounting for non-GPU items (e.g., mostly memory, CPU and networking components).

We think the increase in the per GW price in 2027E mainly stems from accounting for more superpod penetration and higher memory prices, which may require even more non-GPU content. For 2028E and onward, the cost of per GW deployment could even drop as domestic capacities increase.

We also note that our current ASP estimates from our Semis team on chip cost do not fully account for further memory price hikes. When such changes are reflected for the future, we might have to raise our overall capex forecast to accommodate such increases.

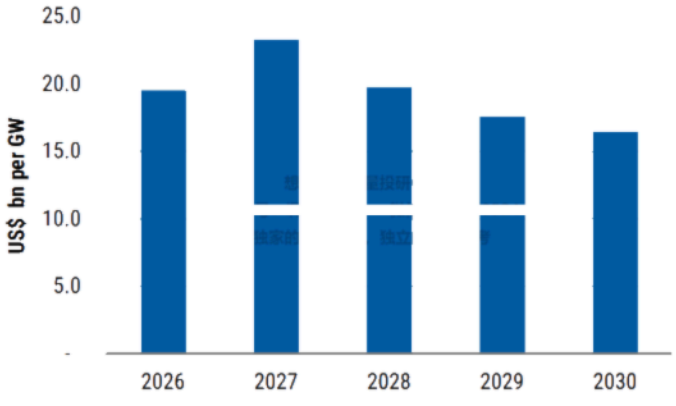
In the US, deploying one GW of GB300 costs roughly US$39bn, with 41% of the cost attributable to the datacenter power shell. However, in China, the cost of the power shell is much cheaper, accounting for just 10-15% of the deployment cost, but the cost of chip procurement will be much higher.

Exhibit 12: Cost of per GW capacity: US vs China



Source: Company data, Morgan Stanley Research estimates

Exhibit 13: Cost per GW of domestic chip over time



Source: Company data, Morgan Stanley Research estimates

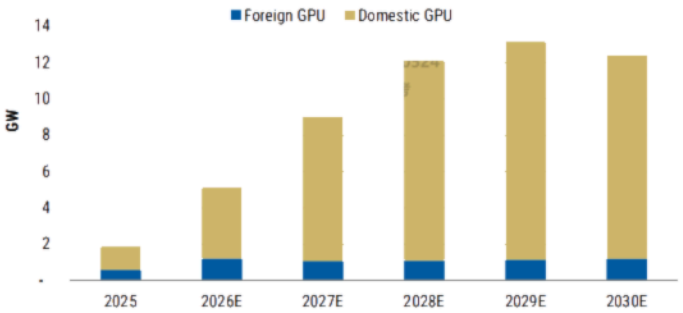
Combining the shipment forecasts from our China Semis team on domestic GPU capacity and the capacity additions from neoclouds and incremental CPU servers, we forecast that total incremental power demand in China will increase from 5.6GW in 2026E to 13GW in 2030E, mainly driven by the deployment of domestic chipsets.

We forecast that IT power in China will increase **from 26GW in 2025 to 81GW by 2030E**, driven mainly by:

- Top 4 hyperscalers' capacity increasing by 34GW in 2025-30E
- Telcos' self-built internally used racks increasing by 9GW in 2025-30E
- Others' increasing by 12GW in 2025-30E.

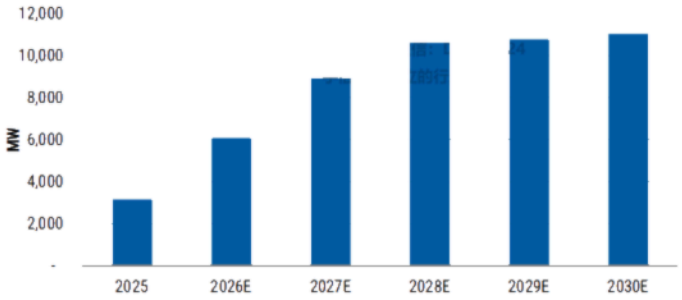
From a datacenter perspective, we expect third-party datacenter orders to increase from 6.1GW in 2026E to 11GW in 2030E, assuming 20% of capacity being self-built by the hyperscalers and AI labs. By region, we expect western regions (Inner Mongolia, Ningxia, etc.) to account for 70% of the incremental orders (5Y total of 33GW) while eastern regions account for the remaining 14GW.

Exhibit 14: China AI power demand to increase to 13GW in 2030E, driven mainly by domestic chipsets



Source: Company data, Morgan Stanley Research (E) estimates

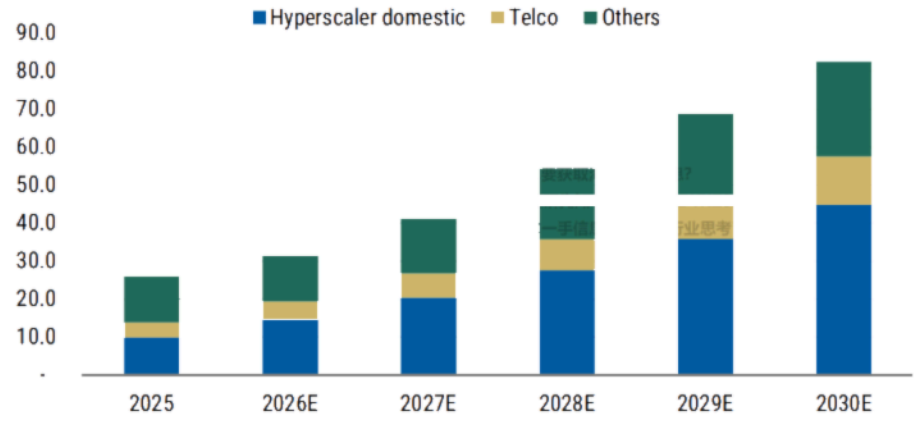
Exhibit 15: Third-party datacenter orders to increase to 11GW in 2030E



Source: Company data, Morgan Stanley Research (E) estimates

Exhibit 16:

We forecast IT power capacity to reach 81GW by end of 2030E



Source: Company data, Morgan Stanley Research (E) estimates

1) Hyperscalers

We forecast that the top 4 hyperscalers, combined, will increase their IT power capacity by 34GW in 2025-2030E, including 6-9GW net adds, each, from Alibaba and Tencent, and 1.0GW net adds from Baidu.

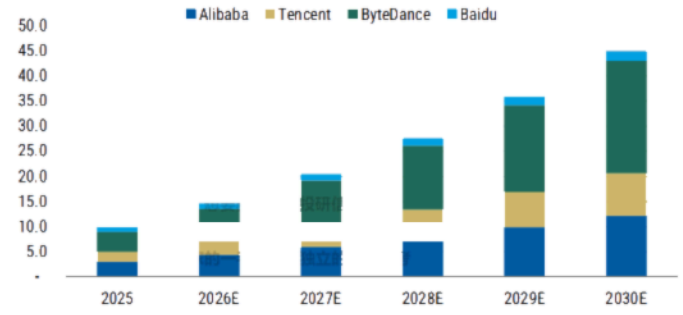
We translate our GW additions from our domestic capex forecast and our estimate of per GW deployment price above. We assume BABA to have a cheaper deployment price, mainly because cheaper T-head pricing and in-house chips should translate into a cost advantage.

We think this can be supported by a combination of domestic chipset procurement as well as self-built ASIC. Alibaba and Baidu are already leading the development with T-head and KLX, and deploying them for internal as well as external use.

For hyperscalers, we also estimate their foreign capacity additions over the next 5 years based on our offshore capex forecasts. We assume that they will spend ~80% of capex on server procurement and 15% on GPU, with the remainder deployed in owned datacenters on CPU servers. This should add a total of 2-3GW per year in terms of compute capacity.

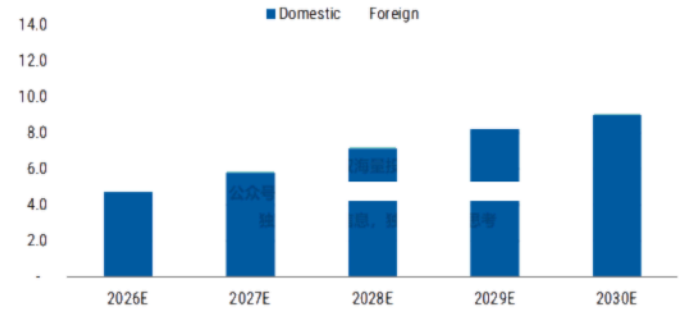
In total, we expect **the top 4 hyperscalers to add compute capacity of 47GW in 2026-30E globally, with 24GW in 2026-28E**, compared to US hyperscalers adding an estimated 108GW.

Exhibit 17: Total domestic IT capacity, by hyperscaler



Source: Company data, Morgan Stanley Research (E) estimates

Exhibit 18: Total net adds in compute capacities – 34GW domestic and 13GW foreign



Source: Company data, Morgan Stanley Research estimates

2) Telco cabinets

Our estimates of telco cabinets are probably much less than their headline number, given that part of datacenters are leased to the hyperscalers and AI labs, which we already account for in the other categories. We forecast a total net add of 9GW over the 5 years, mainly used for internal and SOE/government-related demand.

3) Other parties (retail, AI labs, SOE infra, and other internet)

Other players could have more meaningful implications for future compute capacity, driven mainly by: 1) individual model labs; 2) tier 2 internet companies, such as Xiaomi and Meituan; and 3) SOE players that also build and rent compute to AI labs. Currently, we forecast a net addition of 12GW over the 5 years, which is the remainder of hyperscaler and telco net adds. We expect SOE and AI labs as the main players to digest the remaining domestic AI compute chips.

We think the AI model companies can be more aggressive in acquiring computing in the future. Over the past months, we have seen multiple newsflows on AI labs intending to construct their own compute capacity and develop their ASIC, including:

- **DeepSeek:** Plans to build a 1GW self-owned AI data center in Inner Mongolia while still leasing third-party capacity (Bloomberg, July 2026), and is also developing its own inference-focused AI chip (report here).
- **Z.ai / Zhipu AI:** Reportedly completed a 1GW-class AI data center using domestic chips, with multiple 10,000+ chip clusters (Bloomberg, July 2026), and is exploring a custom ASIC for the GLM model family (report here).

We think it could be an increasingly important consideration for the AI labs to self build/ co-build compute capacity and design their own chipsets.

- **Data security:** CSPs are in competition with the AI labs and completely reliant on their compute capacity; thus, it may not be an ideal choice for the AI labs and may pose data security concerns.
- **More flexibility to optimize infra:** With self-owned compute capacity (whether built in-house or IDC), the AI labs have more flexibility to optimize the underlying infrastructure through avenues such as network construction. This could improve the overall token per second (TPS) metric of the compute capacity they own.

- **ASIC is also crucial for unit economics:** We have noted earlier that the adoption of the ASIC + model could create synergy in improving the overall token generation cost or the TPS on a per server basis through means such as improving model flop utilization (MFU).

However, we are also aware that from a financial perspective, it could also become more risky if such a strategy is implemented:

- **Upfront capex:** Datacenter construction is not expensive in China, but could still prove a burden for the AI labs when their demand becomes greater. It is also a less attractive investment for the AI labs vs spending it on compute to roll out more training and inferencing capabilities.
- **Datacenter vendors are also cautious:** Based on our discussions with datacenter vendors, they are also cautious on taking AI labs' orders, due mainly to the concern over their cash flow stability. From an IDC perspective, a 6- to 12-month delay in ramp-up could become a meaningful drag on their unlevered IRR, which management teams would not regard as favorable.

Financing need differs across the supply chain

Circularity case study

The market sometimes questions an increasing circularity in the AI ecosystem – wherein the cash cows invest in the neoclouds (high leverage) and AI labs (cash burning) – which are also the key customers of their products and services.

Our GVAT team, in an earlier report, noted that the AI space is becoming increasingly interwoven, **with suppliers funding customers, rising customer concentration, revenue sharing between customer and vendor, take-or-pay contracts, and vendor repurchase agreements**, accelerating datacenter build-out but making it difficult for investors to understand the true economics.

In China, we think the financing case is different. China has an ecosystem in which the hyperscalers probably have the most cash and better bargaining power against a more fragmented landscape both upstream (chip makers) and downstream (AI labs). At the same time, the hyperscalers are also incubating their own chipsets and models and running a full integration model. Circularity in China mainly exists between the hyperscalers and the AI labs, but to a lesser extent than in the US, given that the size of the financing is much smaller.

In addition, the US infrastructure providers are increasingly relying on the credit ratings of their customers for financing – in extreme scenarios, infrastructure providers can raise debt simply with a hyperscaler contract, and the spread will be based on the hyperscaler's credit rating for that particular project. In China, infrastructure providers can still access cheap project loans at a 3-4% rate with tangible assets; financiers are not looking for hyperscaler backing at the moment.

However, the importance of CSPs is still crucial within the entire ecosystem. We think the hyperscalers in China assume 4 major roles within the AI ecosystem:

- **Spend:** CSPs are the largest source of capex spending in China, as we noted in our earlier sections.
- **Prepayments:** CSPs are also a source of credit for the supply chain, especially the neoclouds, through: 1) prepayments to support part of the capex for those companies; and 2) locking in supplies on key components such as memory.
- **Contracts:** Hyperscalers are also providing lease contracts or long-term commitment to both neoclouds and datacenters, securing their long-term demand. From a datacenter point of view, we estimate that every 1GW additional compute demand should translate into a total contract value of Rmb24-30bn to the datacenter vendors, assuming an 8- to 10-year contract tenure. This is also the basis for datacenters to incur their capex spending.
- **Invest:** Hyperscalers are also active in expanding their investment across the AI ecosystem – both the AI chip and AI model companies.

In the next section, we will also explore the ways different parties finance their capex need

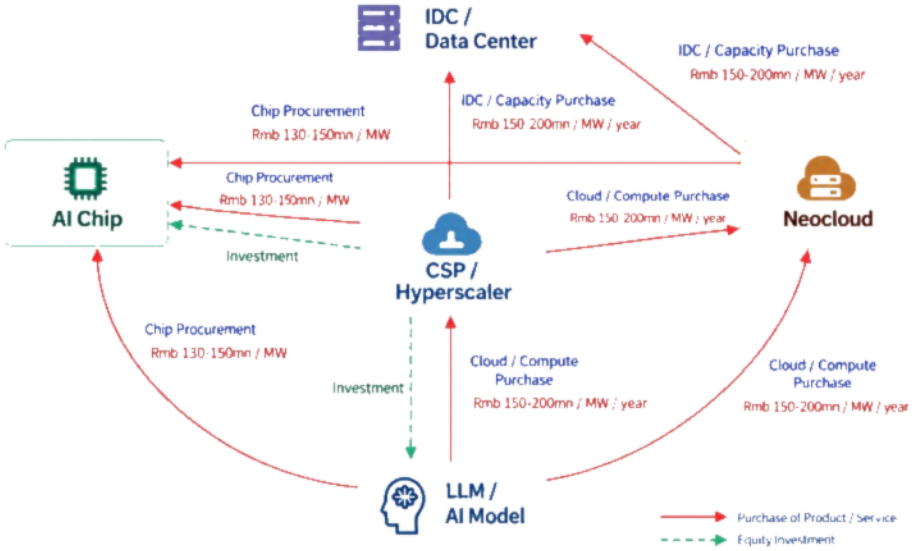
for the next 5 years.

Exhibit 19: Hypsercaler's investment and business across major vendors



Source: Company data, Morgan Stanley Research

Exhibit 20: In China, circularity is limited



Source: Company data, Morgan Stanley Research estimates

Hyperscalers – Cash and OCF are the primary onshore funding, but offshore funding is key to watch

We think the hyperscalers, themselves, are well funded to cover capex domestically, but offshore financing is the key complication, as we need to account for offshore capex, debt repayment, and offshore cash on hand.

We think there are few layers of financing for the hyperscalers:

- **OCF and cash on hand:** Both Tencent and Alibaba should have strong domestic OCF generation and cash on hand to support the capex in China and other new business investments.
- **Lease commitments:** Part of the capex is done via lease and lease back, which we believe is mostly accomplished through the neoclouds, which we discuss later in this report. These leases and capex form the primary investment to support domestic AI development.
- **Divestments:** Tencent, especially, has a strong portfolio of investment – whose liquidity can be easily accessed on the secondary market; such investments can serve as an important source of cash to repay debt and conduct overseas investment.
- **Reduced shareholder return:** Given that both Alibaba and Tencent are listed offshore, they need to deploy offshore cash to conduct share repurchase and dividend payments. We forecast that Alibaba and Tencent will spend a respective 30%/20% of their capex overseas. Such cash could be partly financed by a reduced shareholder return.
- **Debt and equity financing:** This is most straightforward financing but probably more costly to shareholders (equity financing, especially). However, when large lump-sum repayments are required, these may be the likely financing avenues.

However, at the same time, we believe that investors should be mindful that there could be some upside risk to the cash outflows, which is not incorporated in our projections:

- **Datacenter commitments** - Although datacenter rent is much cheaper in China vs. the US – with a Rmb200-250 per kw per month rental price – when lease size increases, it could have more meaningful implications. We estimate hyperscaler datacenter capacity to increase by 38GW over the next 5 years. Assuming a 20% self-built capacity, this would add a total leasing fee of Rmb91bn.
- **Acting as an investment platform** - The hyperscalers usually act as an investment platform to fund the development of the AI ecosystem, which could require financing for AI labs, chipsets, and other technology innovations.

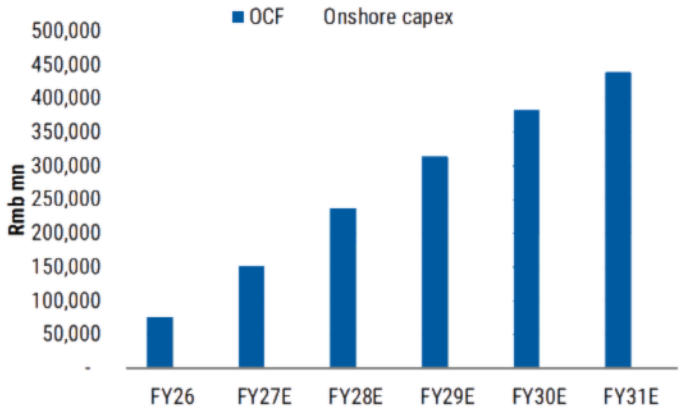
Domestic capex is well funded by OCF and cash balance.

We think domestic capex is well funded by the existing OCF and cash and cash equivalents, based on our forecasts for the hyperscalers. As shown in the charts below, all three players have the OCF to cover capex the majority of the time. For Tencent, we estimate onshore OCF by subtracting its international game's operating profit after tax – and it still appears well funded.

However, this is not completely without risk. Alibaba might further draw down its cash on

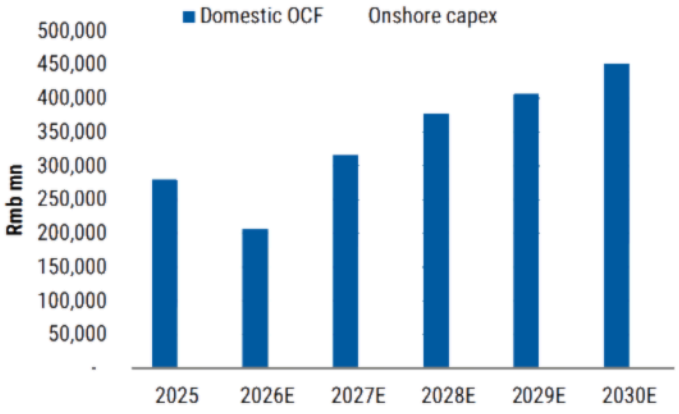
hand if the same food delivery subsidies happen again to compete with Meituan, and Baidu could also have negative free cash flow if search revenue does not stabilize in the future. Tencent appears to be the most stable of the three, given its solid operation and cashflow management.

Exhibit 21: Alibaba OCF vs onshore capex



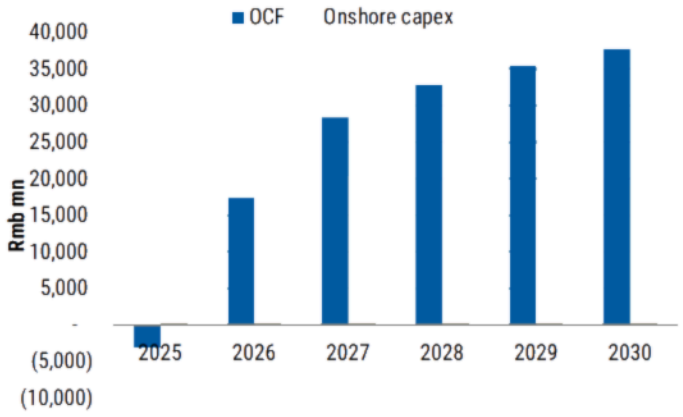
Source: Company data, Morgan Stanley Research estimates

Exhibit 22: Tencent OCF vs onshore capex



Source: Company data, Morgan Stanley Research estimates

Exhibit 23: Baidu OCF vs onshore capex



Source: Company data, Morgan Stanley Research estimates.

Offshore capex totaling Rmb2.0tr may need additional financing

Based on our estimates, we forecast 4 hyperscalers to spend a total of Rmb2.0tr capex for offshore expansion, including chip procurement and datacenter construction. At the same time, there are also some refinancing needs associated with previously borrowed funds. Offshore cash is a more limited resource, and BAT does not have a sizable offshore OCF to support the expansion. For Bytedance, the largest spender offshore, we do not know its exact cashflow, but its recent US$37.6bn loan also suggests a potential need to borrow for offshore expansion.

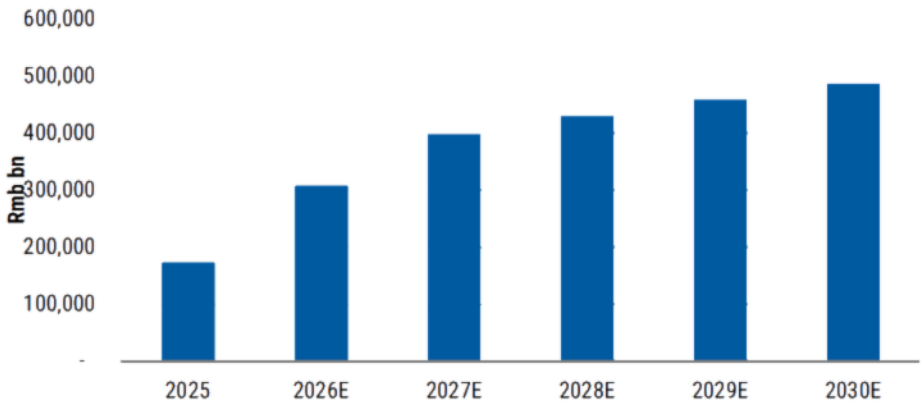
We think the hyperscalers can generally fund their offshore expansions in three ways:

1. **Monetizing investment portfolio** - All hyperscalers have investment portfolios that include offshore listed equities. Those tend to be easy to monetize and can

provide cash to support expansion. Tencent, in particular, has a large portfolio (US $72bn) with listed equities.

2. **Debt** - It is also a common practice for hyperscalers to raise offshore debt. For example, Tencent has raised US$2.45bn and Rmb15bn senior notes YTD; Baidu raised a total of Rmb14.4bn senior notes last year.
3. **Equity** - Hyperscalers can also raise secondary equity or through convertibles to fund large cashflows.

Exhibit 24: Hyperscaler offshore capex



Source: Company data, Morgan Stanley Research estimates

Neocloud and Datacenters – More Diverse Sources of Financing, But Also Carrying the Highest Debt Burden

Neocloud - Surging growth with stronger hyperscaler support

The neoclouds, given their size and capital limits, engage in more innovative financing to fund their capex spend. We use Kingsoft Cloud as an example to explore how neocloud companies can fund their capex, and the key cash outflows to monitor.

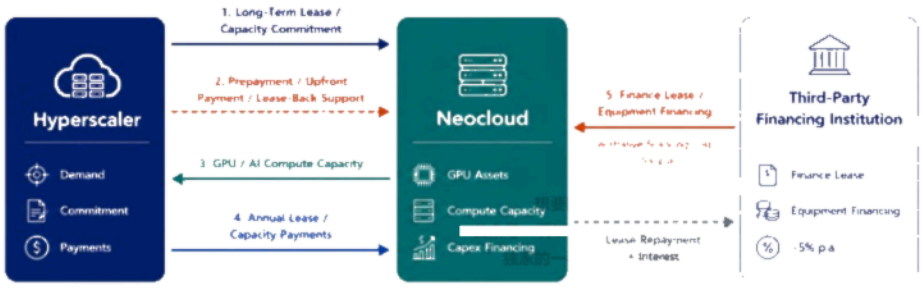
In general, we see the following means as the key financing inflows for KC:

1. **Straight loan:** For the cash part of the capex, KC can borrow up to a loan-to-cost (LTC) ratio of 90%, although such an extreme financing practice would be unlikely, given its burden on future cash flow and repayments. Bank loans and credit are usually quite cheap in China, with an interest rate of 3-4% per annum.
2. **Financing lease:** Per our KC model, we expect the company to finance 40% of the total investment per year through financing lease arrangements. The leases are usually 5-year terms (matching the depreciation of equipment), with an interest rate of ~5%.
3. **OCF with customer prepayment:** OCF is an interesting layer as it accounts for customer prepayment, where certain customers can pay the lifetime lease value in a lump sum (with some discount), or most commonly for one year forward. This could provide support to cashflow as was evident in the 2Q26 result, where KC's reported Rmb2bn OCF was supported by a big prepayment.
4. **Equity and debt:** We note that, when there is still a shortfall in cash, KC could have the option to engage in further equity or debt rasing.

From an industry perspective, we note that other neoclouds are likely taking more debt than KC through both bank loans and financing lease agreements to support expansion. we aggregate the key metrics of the larger neocloud players, and we note that they usually engage in 30-40% lease financing, and after lease payment should cover 20-30% of cash capex spending. The remainder is largely financed by debt, which could be a concern over long-term cashflows when there is lump-sum repayment.

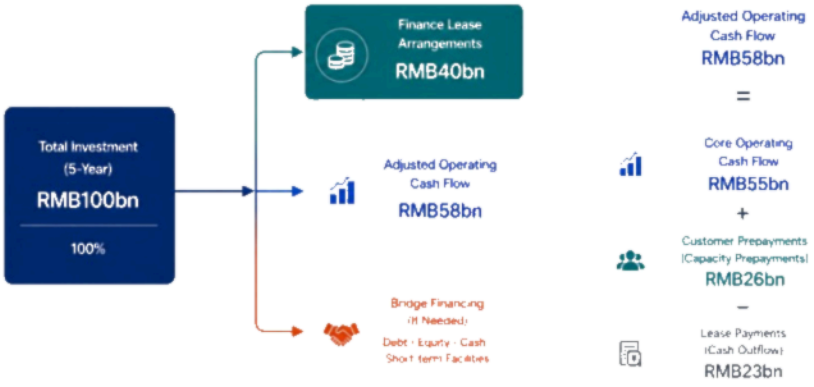
While KC's financing practices are more durable thanks to the support of a large customer, we think the neocloud industry, in general, still exhibits risk on debt levels, given how fast they have been expanding. We see key downside risks including: 1) slowdown in delivery, 2) rising interest cost, and 3) sharp drop in high-end server demand.

Exhibit 25: Financing is an important driver for neocloud business



Source: Company data, Morgan Stanley Research

Exhibit 26: KC's financing schedule - as an example, but other neoclouds will rely more heavily on debt + lease



Source: Company data, Morgan Stanley Research estimates

Datacenters – Higher Debt Burden, and Alternative Financing Is a Key Element

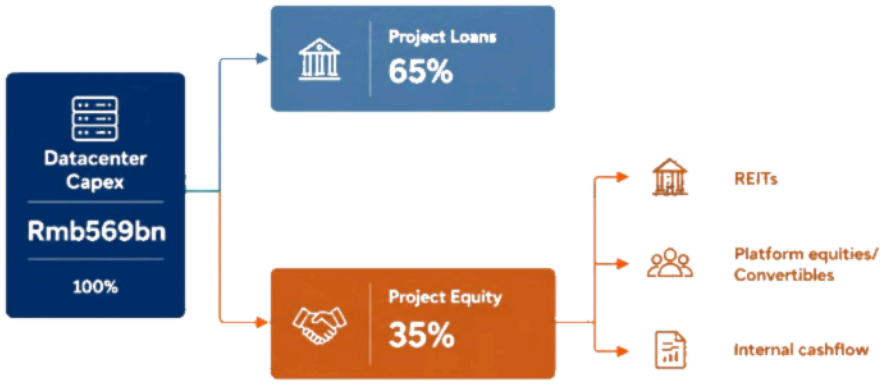
Based on our capacity addition forecast for third-party datacenters and our assumed 50% distribution to telcos, we forecast that carrier-neutral datacenter vendors need to invest a total capex of Rmb569bn over the next 5 years.

We see following potential financing instruments for datacenters:

- **Project loan** - This is the major financing instrument for the datacenter vendors. The IDC vendors could borrow at a 60-80% LTC for new datacenter projects, at a cost of 3-4%, assuming industry standard financing. No dilution and cheap cost are the major advantage here.
- **OCF -** Datacenters, in general, can generate a ~12% project yield on new construction, which will turn into OCF for the coming year. However, given the size of the investment (usually 5-6x OCF), OCF is not capable to finance a major part of the budget. We estimate 10-15% of the capex can be financed by OCF.
- **Alternative financing** - We note increasing popularity of instruments such as REITs (including pre-REIT C-REIT and private ABS) as financing instruments for datacenter vendors (e.g., VNET's plan of Rmb1-2bn per year cash recycling). The advantage is that this can be quick to convert to cash and at a premium multiple compared to the listco. We also see potential private credit and equity involvement at the project level, which could fund initial capex for construction.
- **Additional financing** - New convertibles and equity remain an option for IDC

vendors to fund near-term cash flow.

Exhibit 27: Datacenter financing mix



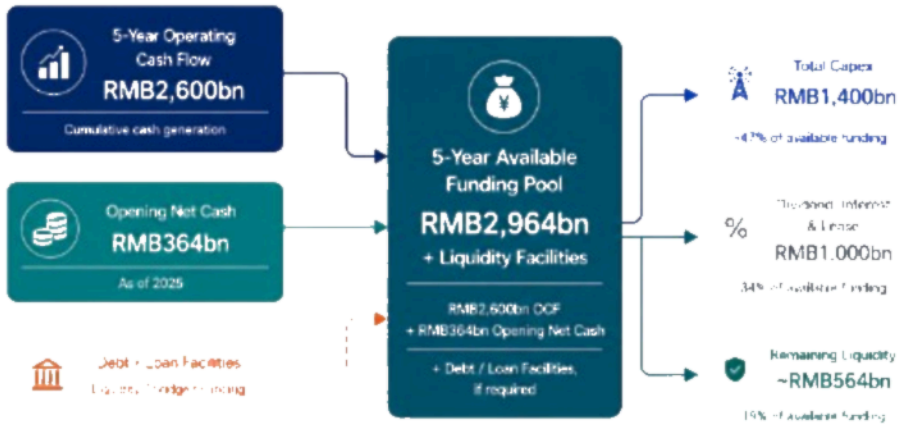
Source: Company data, Morgan Stanley Research estimates

Telcos

The telcos probably have the easiest-to-understand balance sheets and cashflow structures, although we see some downside risk on OCF forecasts – mostly stemming from cash collection issues, but the amount of OCF in the next 5 years still appears achievable. All three telcos' capex are well supported by their OCF, and their debt service levels also appear healthy. We do not see any substantial financing activities needed, given the current balance sheet strength.

We expect most of the telco capex to be self-funded with net cash, annual OCF and reduced traditional capex. There could be some additional debt financing, but this would most likely be manageable, in our view.

Exhibit 28: 5Y fund uses for China telcos



Source: Company data, Morgan Stanley Research estimates

Domestic chip production provides decent ROIC; Unlocking compute is the key

We've previously analyzed ROIC of AI investment under three different scenarios. Here, we introduce a 4th framework:

1. Self-owned GPU IaaS - 44% operating margin, 13% ROIC and 3-year payback
2. Rented IaaS - 20% operating margin, no upfront capex
3. Self-built MaaS - 53% operating margin, 19% ROIC and 2.6-year payback
4. Self-built with 3rd-party API - 57% operating margin, 29% ROIC and 2.1-year payback

As domestic chipset operations start to ramp up, we also analyze the ROIC of running inference workloads on the next generation of domestic chipsets with HBM3e.

- In a prefill/decode disaggregation (PD) scenario - We think ROIC could improve, depending on the type of workload and the compute resources that decode consumes in a mixed-compute scenario.
- Self-built with 3rd-party API on domestic chipset - We see next generation domestic chip potentially achieving 31% operating margin, 9.1% ROIC and 3.6 years payback, which could improve in the future as technology is further iterated.

Recapping Our ROIC Analysis

In our August 17, 2026, report, we assessed three different scenarios and the margin, ROIC, and payback period under each scenario:

1. Self-owned GPU IaaS: A potential path towards 44% OP margin, 13% ROIC, and three-year payback

In our base case, we assume monthly rental fees of Rmb250k/server/month (for long-term rental), or US$8.5 per GPU hour with 75% utilization (for short-term rental, which also yields Rmb250k per month) and Rmb8mn per server capex. Key costs flowing through are server depreciation, IDC depreciation, energy costs, and other opex (labor, maintenance, etc.), which generates 44% operating margin, 13% ROIC, and three-year cash payback.

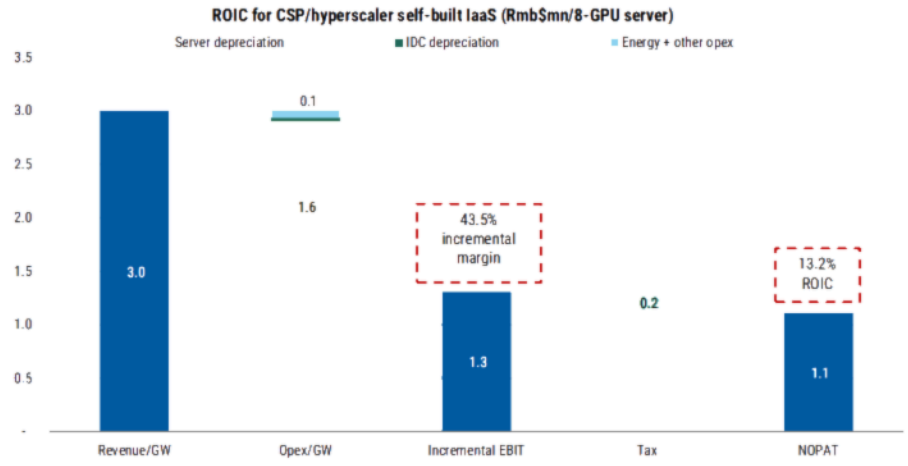
Our base case for China is below the US team's 67% incremental EBIT margin/31% ROIC for GB300 GPU IaaS, mainly reflecting much higher server costs (close to 3x), but partly offset by much lower IDC and energy costs. The direction of our analysis is similar to that of the US – the return curve is highly convex as hardware costs decline and monetization improves.

Exhibit 29: China – unit economics scenarios of self-owned GPU IaaS per server

ROIC for CSP / hyperscalers self-built IaaS (Rmb)					
Monthly server rental fee (Rmb)	350,000	300,000	250,000	200,000	150,000
Revenue	**4,200,000**	**3,600,000**	**3,000,000**	**2,400,000**	**1,800,000**
Server depreciation	2,400,000	2,000,000	1,600,000	1,200,000	800,000
IDC depreciation	26,667	26,667	26,667	26,667	26,667
Energy cost	67,392	67,392	67,392	67,392	67,392
Total opex	**2,494,059**	**2,094,059**	**1,694,059**	**1,294,059**	**894,059**
OP / EBIT	**1,705,941**	**1,505,941**	**1,305,941**	**1,105,941**	**905,941**
% margin	40.6%	41.8%	43.5%	46.1%	50.3%
Tax rate	15.0%	15.0%	15.0%	15.0%	15.0%
NOPAT	**1,450,050**	**1,280,050**	**1,110,050**	**940,050**	**770,050**
Server	12,000,000	10,000,000	8,000,000	6,000,000	4,000,000
IDC	400,000	400,000	400,000	400,000	400,000
Total capex	**12,400,000**	**10,400,000**	**8,400,000**	**6,400,000**	**4,400,000**
ROIC (NOPAT / Capex)	**11.7%**	**12.3%**	**13.2%**	**14.7%**	**17.5%**
EBITDA	4,132,608	3,532,608	2,932,608	2,332,608	1,732,608
Tax	(255,891)	(225,891)	(195,891)	(165,891)	(135,891)
Operating cash flow	3,876,717	3,306,717	2,736,717	2,166,717	1,596,717
Cash payback (yrs)	**3.20**	**3.15**	**3.07**	**2.95**	**2.76**

Source: Company data, Morgan Stanley Research estimates

Exhibit 30: CSP self-built IAAS generating 44% incremental margins and 13% ROIC



Source: Company data, Morgan Stanley Research estimates

2. Rented/neocloud IaaS: Lower margin, but no upfront capex and immediate cash generation

The second part of our framework considers a different way for Alibaba to address incremental AI computing demand: by renting the underlying servers from a neocloud/ third-party provider and re-renting that capacity to customers.

Because AliCloud does not own the server or the computing asset, there is effectively no upfront capex in this part of the framework.

The economics therefore become a spread business between the rental price charged to the end-customers (revenue) and the price paid to neocloud providers for compute (cost).

In our base case, Rmb250k/month customer pricing against Rmb200k/month computing rental cost generates ~20% operating margin.

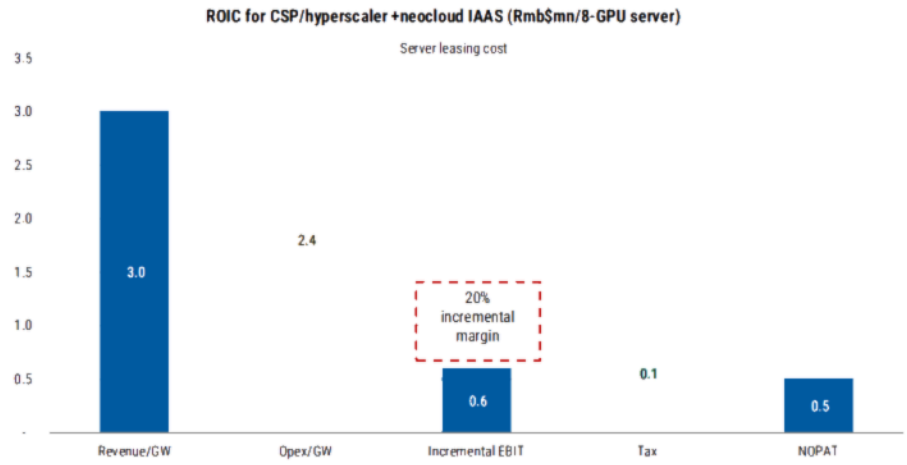
Rented capacity generates lower operating margins because the economics accrue in part to the neocloud provider. However, it offers immediate positive cash contribution once the rental spread turns positive, given no upfront server capex to recover.

Exhibit 31: Unit economics scenarios of rented/neocloud IaaS per server

ROIC for CSP / hyperscalers +neocloud IaaS (Rmb)					
Monthly server rental fee (Rmb)	200,000	225,000	250,000	275,000	300,000
Revenue	**2,400,000**	**2,700,000**	**3,000,000**	**3,300,000**	**3,600,000**
Monthly server rental cost	200,000	200,000	200,000	200,000	200,000
Server leasing cost	2,400,000	2,400,000	2,400,000	2,400,000	2,400,000
Total opex	**2,400,000**	**2,400,000**	**2,400,000**	**2,400,000**	**2,400,000**
OP / EBIT	**-**	**300,000**	**600,000**	**900,000**	**1,200,000**
% margin	0.0%	11.1%	20.0%	27.3%	33.3%
Tax rate	15.0%	15.0%	15.0%	15.0%	15.0%
NOPAT	**-**	**255,000**	**510,000**	**765,000**	**1,020,000**
Server	-	-	-	-	-
IDC	-	-	-	-	-
Total capex	**-**	**-**	**-**	**-**	**-**
ROIC (NOPAT / Capex)					
EBITDA	-	300,000	600,000	900,000	1,200,000
Tax	-	(45,000)	(90,000)	(135,000)	(180,000)
Operating cash flow	-	255,000	510,000	765,000	1,020,000
Cash payback (yrs)	-	-	-	-	-

Source: Company data, Morgan Stanley Research estimates

Exhibit 32: CSP rented/neocloud IaaS generates 20% EBIT margin with no upfront costs



Source: Company data, Morgan Stanley Research estimates

3. Self-built MaaS: the greatest sensitivity to inference mix, throughput and pricing

The third part of our framework moves one layer higher in the AI stack, with CSPs/model providers monetizing owned computing infrastructure and generating revenue through model APIs/MaaS.

The key drivers are: (1) token throughput per GPU, (2) inference vs. training workload mix, and (3) token pricing.

In our base case, we assume 4,000 tokens/sec/GPU, 50% inference in the mix, and Rmb9.5/mn blended token pricing.

These drivers compound – higher throughput and inference mix increase monetizable token volume, while higher token pricing increases revenue per token. Training generates no direct revenue but remains essential for model improvement.

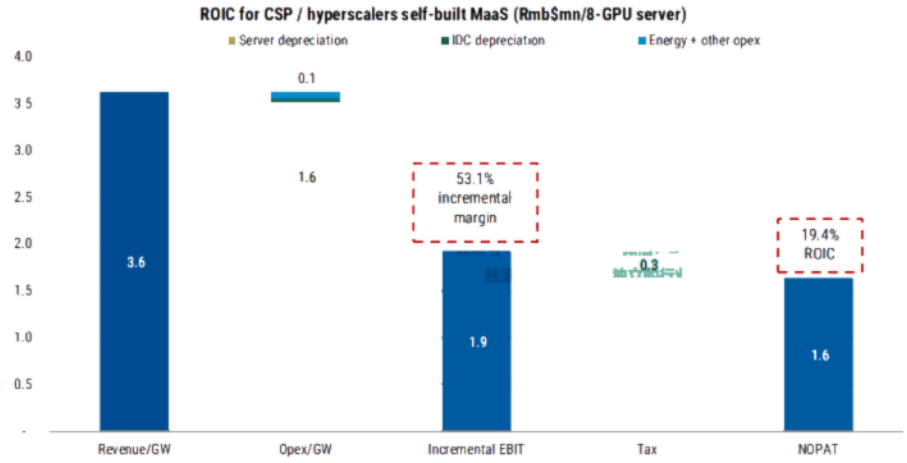
Our base case generates 76%+ gross margin, 53% operating margin, 19% ROIC and 2.5-year payback – the highest margin potential across the three parts of our framework. However, outcomes vary widely: training-heavy, low-throughput workloads can remain loss-making, while improving proportion of inference and model efficiency can drive significant operating leverage from the same investment base.

Exhibit 33: Unit economics scenarios of a self-built MaaS in China per server

ROIC for CSP / hyperscalers self-built MaaS (Rmb)					
Token throughput (token / second / GPU)	2,000	3,000	4,000	4,500	5,000
Inference vs training workload	30.0%	40.0%	50.0%	60.0%	70.0%
Utilization rate	75.0%	75.0%	75.0%	75.0%	75.0%
Revenue generating tokens (mn)	**113,530**	**227,059**	**378,432**	**510,883**	**662,256**
Input	4.00	5.50	7.00	8.50	10.00
Input (KV Cache)	0.40	0.55	0.70	0.85	1.00
Output	12.00	16.50	21.00	25.50	30.00
Blended (Rmb/mn)	**5.46**	**7.51**	**9.56**	**11.60**	**13.65**
Blended (US$/mn)	**0.81**	**1.11**	**1.42**	**1.72**	**2.02**
Revenue	**619,872**	**1,704,647**	**3,615,918**	**5,927,522**	**9,039,794**
Server depreciation	1,600,000	1,600,000	1,600,000	1,600,000	1,600,000
IDC depreciation	26,667	26,667	26,667	26,667	26,667
Energy cost	67,392	67,392	67,392	67,392	67,392
Total cost and opex	**1,694,059**	**1,694,059**	**1,694,059**	**1,694,059**	**1,694,059**
Inference cost	508,218	677,623	847,029	1,016,435	1,185,841
Gross profit	**111,654**	**1,027,023**	**2,768,888**	**4,911,087**	**7,853,953**
% margin	18.0%	60.2%	76.6%	82.9%	86.9%
Training cost	1,185,841	1,016,435	847,029	677,623	508,218
OP / EBIT	**(1,074,187)**	**10,588**	**1,921,859**	**4,233,464**	**7,345,736**
% margin	-173.3%	0.6%	53.1%	71.4%	81.3%
Tax rate	15.0%	15.0%	15.0%	15.0%	15.0%
NOPAT	**(913,059)**	**9,000**	**1,633,580**	**3,598,444**	**6,243,875**
Server	8,000,000	8,000,000	8,000,000	8,000,000	8,000,000
IDC	400,000	400,000	400,000	400,000	400,000
Total capex	**8,400,000**	**8,400,000**	**8,400,000**	**8,400,000**	**8,400,000**
ROIC (NOPAT / Capex)	**-10.9%**	**0.1%**	**19.4%**	**42.8%**	**74.3%**
EBITDA	552,480	1,637,255	3,548,526	5,860,130	8,972,402
Tax	161,128	(1,588)	(288,279)	(635,020)	(1,101,860)
Operating cash flow	713,608	1,635,667	3,260,247	5,225,111	7,870,542
Cash payback (yrs)	**11.77**	**5.14**	**2.58**	**1.61**	**1.07**

Source: Company data, Morgan Stanley Research estimates

Exhibit 34: In China, a CSP's self-built MaaS generates 53% EBIT margin and 19% ROIC



Source: Company data, Morgan Stanley Research estimates

4. A slight tweak - Introducing self-built infra with 3rd-party API

In a 3P API model, we assume a base case TPS of 4k per second, and there is no training cost in this scenario.

We introduce two more assumptions: 1) **a 10% revenue sharing fee to the third-party model provider,** and we assume a gross accounting (i.e., revenue fully booked and sharing cost is recognized in COGS); and 2) we assume a performance discount of **10-50% when running 3P model due to less infra optimization**. We note that the gross or net assumption may have an impact on margin, but should be neutral to our ROIC or payback period calculation.

As as result, this model could generate a **ROIC of 29% and cash payback of ~2 years under our base case assumption vs our base case in-house model at 19.7% ROIC and 2.58 years of cash payback**.

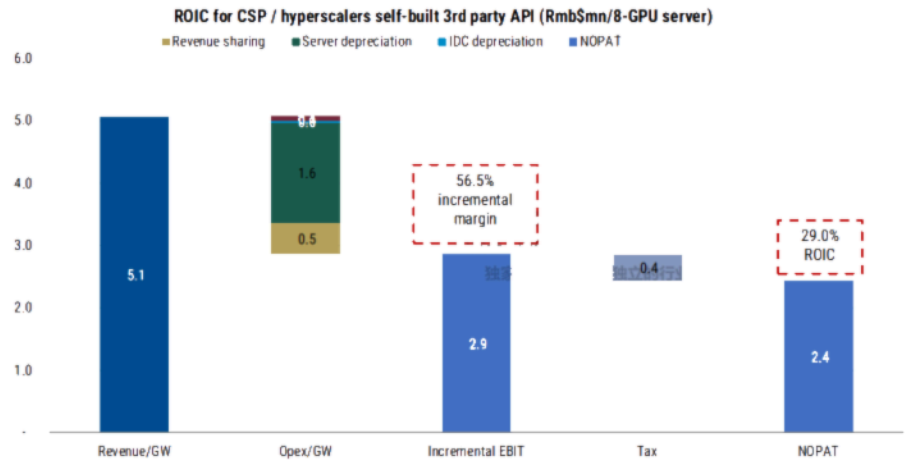
We think the key variable is mainly the performance discount, which is hard to measure. But based on our talk with the supply chain, the slower the model architecture improvement, the more advantage the cloud players would have given a longer time to improve the underlying infra.

Exhibit 35: Unit economics scenarios of self-built compute + 3rd-party API

ROIC for CSP / hyperscalers self-built 3rd party API (Rmb)					
1P base case throughput	4,000	4,000	4,000	4,000	4,000
Discount to 1P model	50%	40%	30%	20%	10%
Token throughput (token / second / GPU)	2,000	2,400	2,800	3,200	3,600
Inference vs training workload	100.0%	100.0%	100.0%	100.0%	100.0%
Utilization rate	75.0%	75.0%	75.0%	75.0%	75.0%
Revenue generating tokens (mn)	**378,432**	**454,118**	**529,805**	**605,491**	**681,178**
Input	4.00	5.50	7.00	8.50	10.00
Input (KV Cache)	0.40	0.55	0.70	0.85	1.00
Output	12.00	16.50	21.00	25.50	30.00
Blended (Rmb/mn)	**5.46**	**7.51**	**9.56**	**11.60**	**13.65**
Blended (US$/mn)	**0.81**	**1.11**	**1.42**	**1.72**	**2.02**
Revenue	**2,066,239**	**3,409,294**	**5,062,285**	**7,025,212**	**9,298,074**
Sharing ratio	10%	10%	10%	10%	10%
Revenue sharing	206,624	340,929	506,228	702,521	929,807
Server depreciation	1,600,000	1,600,000	1,600,000	1,600,000	1,600,000
IDC depreciation	26,667	26,667	26,667	26,667	26,667
Energy cost	68,000	68,000	68,000	68,000	68,000
Total cost and opex	**1,901,291**	**2,035,596**	**2,200,895**	**2,397,188**	**2,624,474**
Inference cost	1,901,291	2,035,596	2,200,895	2,397,188	2,624,474
Gross profit	**164,948**	**1,373,698**	**2,861,390**	**4,628,024**	**6,673,600**
% margin	8.0%	40.3%	56.5%	65.9%	71.8%
OP / EBIT	**164,948**	**1,373,698**	**2,861,390**	**4,628,024**	**6,673,600**
% margin	8.0%	40.3%	56.5%	65.9%	71.8%
Tax rate	15.0%	15.0%	15.0%	15.0%	15.0%
NOPAT	**140,206**	**1,167,643**	**2,432,181**	**3,933,820**	**5,672,560**
Server	8,000,000	8,000,000	8,000,000	8,000,000	8,000,000
IDC	400,000	400,000	400,000	400,000	400,000
Total capex	**8,400,000**	**8,400,000**	**8,400,000**	**8,400,000**	**8,400,000**
ROIC (NOPAT / Capex)	**1.7%**	**13.9%**	**29.0%**	**46.8%**	**67.5%**
EBITDA	1,791,615	3,000,364	4,488,056	6,254,690	8,300,267
Tax	(24,742)	(206,055)	(429,208)	(694,204)	(1,001,040)
Operating cash flow	1,766,873	2,794,310	4,058,848	5,560,487	7,299,227
Cash payback (yrs)	**4.75**	**3.01**	**2.07**	**1.51**	**1.15**

Source: Company data, Morgan Stanley Research estimates

Exhibit 36: In China, hosting third-party API generates 46.6% incremental margin and 34.1% ROIC on 30% revenue sharing



Source: Company data, Morgan Stanley Research estimates

Introducing ROIC Analysis on Domestic Chip Production

Why now? Although domestic chipset mass rollout started in 2H26, we note that the current version of the chipset is not equipped with HBM 3E, which means that the memory bandwidth may not be enough to run the decoding of large language models (output tokens). Rather, the currently more advanced domestic chipset is probably specialized in the prefill stage (input tokens) through PD disaggregation.

Based on our discussion with our Semiconductor team, we note that the next-generation domestic chipset will start to adopt HBM3E, increasing the attractiveness of such chips in running the full inference. There are already chip designers who are starting to test new chip design with better memory bandwidth. We think this could be a more meaningful advancement for the domestic supply chain as it removes the real bottleneck on the compute capacity addition.

Several papers also note that it is not always ideal to adopt PD disaggregation. It is usually a better choice when running tasks with a much higher input token length, a larger model or a stricter requirement on TPOT (time per output token).

Exhibit 37: PD disaggregation - How it works



Source: Morgan Stanley Research

We analyze the ROIC and impact when domestic AI chips are adopted in a scenario of PD disaggregation and also full domestic inference process (likely with the next generation of HBM3E-equipped chips):

1. PD disaggregation could temporarily bridge the gap.

Generally, PD disaggregation will improve the output throughput on the high-end server, and the % improvement in output throughput in a decode-only scenario is determined by how much computing resources prefill consumes. This is also correlated to the input/ output length of a particular query. The advantage of adopting PD disaggregation would be higher when the input mix is higher in a given query. Of course, in the real world, the context length and mix for an agentic task can differ a lot – and it's unlikely that a perfect mix could be achieved.

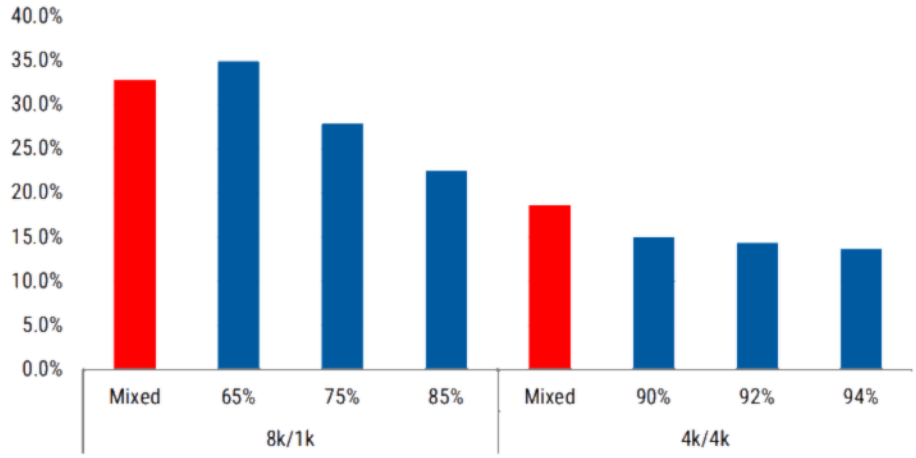
In our analysis, we also assume a prefill server equipped with Ascend 950PR and a model with 69B active parameters. We tested the difference in adopting PD disaggregation vs running on a single high-end server in two scenarios: 1) 8k/1k output/input, which is long input and short output, typically seen in tasks such as summarizing, long document query and coding assistance; and 2) 4k/4k, which is typically seen in high-value-added areas,

including reasoning and deep research.

We find that in a long input task, it is possible to achieve better ROIC when PD disaggregation is adopted. However, in a 4k/4k task scenario, we think it is harder, given that prefill would generally account for a lesser mix of computing resources.

We also highlight that our analysis is more hypothetical and does not take account of more complicated real-life considerations, such as the requirement on TTFT (time to first token), TPOT (time per output token), KV (key value) cache transmission efficiency, etc. We want to illustrate the basic principle underlying the PD disaggregation and potential effect on the token output.

Exhibit 38: Sensitivity on PD disaggregation ROIC in 8k/1k and 4k/4k scenarios



Source: Company data, Morgan Stanley Research estimates

2. The final stage - a full domestic infrastructure

We do not know the exact specs of the new generation of domestic compute, but we can test the ROIC under different efficiency assumptions.

Using our base case ROIC model, which assumes a 4k TPS per GPU on a third-party MaaS, we assume that the new generation compute is capable of achieving a 20-40% compute efficiency vs our base case compute. However, domestic AI servers are probably cheaper – at a procurement price of Rmb4mn per server. We estimate that **ROIC can be in the range of 0.3%-17.9% and a payback period of 2.7-5.2 years, depending on the final compute efficiency.**

Our base case assumes a 30% performance vs foreign servers, resulting in an incremental **EBIT margin of 31%, ROIC of 9.1% and a payback period of 3.6 years** under a self-built capacity running third-party API scenario.

What could hit the low end? Zhipu recently adopted full domestic infrastructure for its Flash version, for which, in our view, ROIC could be poor, as most of the current domestic chips can run only prefill or tend to be inefficient on decoding. This suggests a much higher performance discount, and could hit our bear case of a low to negative ROIC.

In the case of in-house chip design, we think incremental inference margin could improve further on higher performance as well as 15-20% cheaper server cost (no need to pay chip

margin). In that case, we think ROIC could improve to 20-25%. However, we also note that the in-house model may still rely on more expensive foreign servers for training.

Exhibit 39: Full domestic compute sensitivity analysis

ROIC for CSP / hyperscalers self-built 3rd party API (Rmb)	Foreign					
Performance vs foreign server		20%	25%	30%	35%	40%
Token throughput (token / second / GPU)	2,800	560	700	840	980	1,120
Inference vs training workload	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Utilization rate	75.0%	75.0%	75.0%	75.0%	75.0%	75.0%
Revenue generating tokens (mn)	**529,805**	**105,961**	**132,451**	**158,941**	**185,432**	**211,922**
Input	7.00	7.00	7.00	7.00	7.00	7.00
Input (KV Cache)	0.70	0.70	0.70	0.70	0.70	0.70
Output	21.00	21.00	21.00	21.00	21.00	21.00
Blended (Rmb/mn)	**9.56**	**9.56**	**9.56**	**9.56**	**9.56**	**9.56**
Blended (US$/mn)	**1.42**	**1.42**	**1.42**	**1.42**	**1.42**	**1.42**
Revenue	**5,062,285**	**1,012,457**	**1,265,571**	**1,518,685**	**1,771,800**	**2,024,914**
Sharing ratio		10%	10%	10%	10%	10%
Revenue sharing	506,228	101,246	126,557	151,869	177,180	202,491
Server depreciation	1,600,000	800,000	800,000	800,000	800,000	800,000
IDC depreciation	26,667	26,667	26,667	26,667	26,667	26,667
Energy cost	68,000	68,000	68,000	68,000	68,000	68,000
Total cost and opex	**2,200,895**	**995,912**	**1,021,224**	**1,046,535**	**1,071,847**	**1,097,158**
Inference cost	2,200,895	995,912	1,021,224	1,046,535	1,071,847	1,097,158
Gross profit	**2,861,390**	**16,545**	**244,347**	**472,150**	**699,953**	**927,756**
% margin	56.5%	1.6%	19.3%	31.1%	39.5%	45.8%
OP / EBIT	**2,861,390**	**16,545**	**244,347**	**472,150**	**699,953**	**927,756**
% margin	56.5%	1.6%	19.3%	31.1%	39.5%	45.8%
Tax rate	15.0%	15.0%	15.0%	15.0%	15.0%	15.0%
NOPAT	**2,432,181**	**14,063**	**207,695**	**401,328**	**594,960**	**788,593**
Server	8,000,000	4,000,000	4,000,000	4,000,000	4,000,000	4,000,000
IDC	400,000	400,000	400,000	400,000	400,000	400,000
Total capex	**8,400,000**	**4,400,000**	**4,400,000**	**4,400,000**	**4,400,000**	**4,400,000**
ROIC (NOPAT / Capex)	**29.0%**	**0.3%**	**4.7%**	**9.1%**	**13.5%**	**17.9%**
EBITDA	4,488,056	843,211	1,071,014	1,298,817	1,526,620	1,754,423
Tax	(429,208)	(2,482)	(36,652)	(70,823)	(104,993)	(139,163)
Operating cash flow	4,058,848	840,730	1,034,362	1,227,994	1,421,627	1,615,259
Cash payback (yrs)	**2.07**	**5.23**	**4.25**	**3.58**	**3.10**	**2.72**

Source: Company data, Morgan Stanley Research estimates

This report references U.S. Executive Order 14032 and/or entities or securities that are designated thereunder. U.S. persons may be prohibited from buying certain securities of entities named in this report. Readers are solely responsible for ensuring that their investment activities are carried out in compliance with applicable laws.

This report references export controls and/or entities that may be subject to export control restrictions. Readers are solely responsible for ensuring that their investment or trade activities are carried out in compliance with applicable laws.