China Data Centers: Highlights from China AI Computing Industry Ecological Development Conference

Timothy Zhao
+852-2978-2673
timothy.zhao@gs.com
Goldman Sachs (Asia) L.L.C.

Ronald Keung, CFA
+852-2978-0856
ronald.keung@gs.com
Goldman Sachs (Asia) L.L.C.

Eunice Liu
+852-2978-7472 eunice.liu@gs.com
Goldman Sachs (Asia) L.L.C.

We recently attended the China AI Computing Industry Ecological Development Conference, where representatives from a variety of industry players (e.g. data center operators, hyperscalers, token routing platforms) and independent research institutes (e.g. CAICT, KZ Consulting) delivered presentations. Key topics included national computing power network, GW-scale data center clusters, domestic chip ramp-up and limitations, token operations, etc.

GS takes: We think China's national computing power network may reshape the competitive landscape and geographic footprint of the China data center industry over the next 5 years; in other words, capital and technology are heavily shifting toward Western China hubs, while the value proposition of tier-1 city data centers is shifting toward ultra-low latency hot compute, edge nodes and AI inference. The emergence of GW-scale computing power clusters implies higher technology and resource barriers and provides differentiation among data center operators. Within our China data center coverage, we are Buy rated on VNET, GDS and Range Intelligent, Neutral rated on Shanghai Athub, and Sell rated on Beijing Sinnet.

We summarize key highlights from the conference below.

Computing power network

- **Computing power network as part of China's core infrastructure construction project in 2026-30:** China's "six major networks", including the water network, new-type power grid, computing power network, next-generation communication network, urban underground pipe network, and logistics network, will attract Rmb7tn investments in 2026 as per NDRC, and Bloomberg reported on Jun 9 that China is to spend c.Rmb2tn (US$300bn) over the next 5 years in data center investments (GS note).
- **Computing power network architecture:** 1) Computing-networking infrastructure layer that integrates heterogeneous compute with high-speed transmission via RDMA (Remote Direct Memory Access) and elastic networks; 2) Interconnection resource layer driven by a "1+M+N" node hierarchy i.e. 1 national public computing-networking service platform, M regional platforms, and N industry platforms; 3) Application service layer. Hence, wide-area data transmission and long-distance RDMA, and cross-domain orchestration platforms are key capabilities; currently, out of 1.1k+ computing nodes and 110+ scheduling platforms, only 10% possess cross-domain orchestration capabilities,

Goldman Sachs does and seeks to do business with companies covered in its research reports. As a result, investors should be aware that the firm may have a conflict of interest that could affect the objectivity of this report. Investors should consider this report as only a single factor in making their investment decision. For Reg AC certification and other important disclosures, see the Disclosure Appendix, or go to www.gs.com/research/hedge.html. Analysts employed by non-US affiliates are not registered/qualified as research analysts with FINRA in the U.S.

as per CAICT.

GW-scale data center clusters

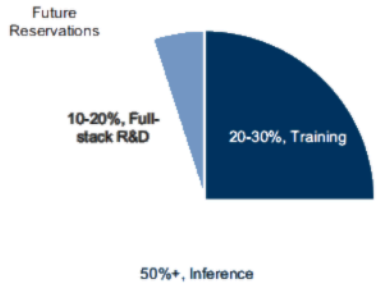
- **Foundation of a GW-scale data center clusters lies in network architecture** as increasing switch port counts from 65 to 144 ports require higher individual network capacity. While there have been a number of operational GW-scale clusters in operation in the US (e.g. Project Rainier), computing clusters composed of 100k+ chips remain scarce in China.
- **Breakthrough of 200MW data centers in China:** Range Intelligence (300442.SZ, Buy) has brought a 200MW, 100k chip-scale data center building operationally YTD, as part of its GW-scale computing cluster on a 500 mu (0.3 sqkm) land parcel, and sees a 400MW data center theoretically feasible to implement.
- **Computing power economics gap between domestic chips and imported chips remain big** (Exhibit 1). While capex per IT power for domestic chips is 40-50% lower than imported chips, computing power related efficiency metrics lag e.g. capex per computing power is 2-4x and computing power per IT power is 10-30% vs. imported chips.

Exhibit 1: Capex per 200MW IT power data center based on different IT infrastructure

Capex per 200MW IT power data center	Nvidia B300	Huawei CM384	Panjiu 2.0
Computing power (EFLOPS, FP16)	**250**	**80**	**40**
Total IT power (kW)	**187,500**	**208,800**	**234,600**
Servers/supernodes	12,500	261	782
IT power per unit	15	800	300
Total capex (Rmb bn)	**61.0**	**41.0**	**41.5**
Land	0.05	0.05	0.05
Land area (mu)	100	100	100
Land cost (Rmb k/mu)	500	500	500
Building construction	0.90	0.90	0.90
GFA (k sqm)	200	200	200
Cost per GFA (Rmb k/sqm)	4.5	4.5	4.5
Electrical equipment	3.8	4.8	5.4
Cost per IT power (Rmb k/kW)	20	23	23
Servers/supernodes	56.3	35.2	35.2
Unit cost (Rmb mn)	4.5	135.0	45.0
Capex per IT power (Rmb bn/GW)	**325**	**196**	**177**
Capex per computing power (Rmb bn/EFLOPS)	**0.24**	**0.51**	**1.04**
Computing power per IT power (EFLOPS/GW)	**1,333**	**383**	**171**

Source: Company data

Exhibit 2: Typical workload distribution for a GW-level campus

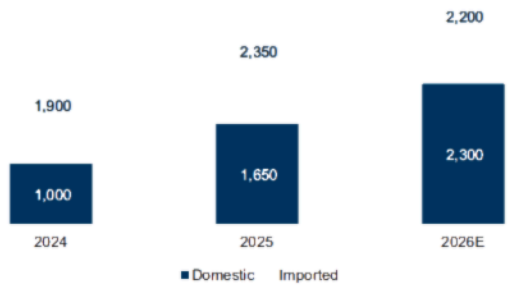


Source: Company data

Domestic chips

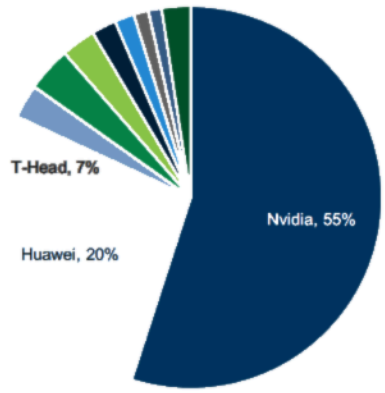
- **Imported computing power remains dominant among CSPs:** As per CAICT, 70-90% of outstanding computing power of major CSPs (each with operational computing power 2.2-4.5 PFLOPS) in China is based on imported chips.
- **Domestic AI accelerator chip ramp-up:** As per KZ Consulting, domestic AI accelerator chip shipments may exceed 50% market share in 2026 in China, with Huawei and T-Head leading the domestic market share. Internet companies and government entities appear to be the biggest customers of domestic chips, with internet companies accounting for 63%/23%/60%/30% of Huawei/T-Head/Cambricon/Hygon shipments in 2026, and government entities accounting for 15%/15%/35% of Huawei/Cambricon/Hygon shipments in 2026.
- **Gap between domestic and imported chips on token efficiency and profitability:** As per CAICT, domestic chips remain lagging behind foreign peers in performance, cost and ecosystem, which constrain token output efficiency (Exhibit 5) and efficiency (Exhibit 6).

Exhibit 3: Domestic AI accelerator chip shipments are expected to exceed 50% market share in 2026
China AI accelerator chip shipments: domestic vs. imported



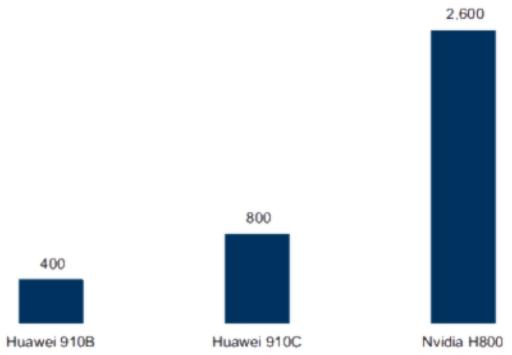
Source: KZ Consulting

Exhibit 4: Huawei and T-Head are expected to lead domestic AI accelerator chip shipments in 2026
China AI accelerator chip shipment share in 2026



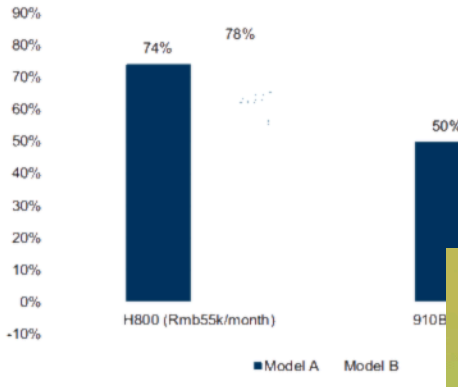
Source: KZ Consulting

Exhibit 5: Huawei 910B/910C token output volume is 1/6-1/3 of that of Nvidia H800 servers
Average daily token output per server for typical tasks (mn)



Source: CAICT

Exhibit 6: Profit margins of tokens/APIs based on Huawei 910B servers lag behind that on Nvidia H800
Illustration: Profit margin by IT infrastructure



Source: CAICT

Token operations

- **Token pricing divergence:** Against the backdrop of the "impossible triangle" between intelligence, cost and quality (i.e. low latency, high concurrency, high availability), the token pricing trend appears to diverge. This is also related to business model variations as competition is evolving from single model breakthroughs to ecosystems.
- **Token export:** 1) Shantou City is leveraging its "Inbound Data Processing" policy to build China's first city-level practice for token export, with advantages in its submarine cable connectivity and abundant offshore wind power resources; 2) Key applications currently involve AI toys and AI short-form videos, while cross-border digital employee may represent a future business model.