

Cloud AI

Booming cloud AI trend in 2024-25F

Key message

1. We expect total global server shipments to grow 5% YoY in 2024F to 11.8mn units, with non-AI server sales growth flat, while global server shipments will grow 10% in 2025F to 13mn units with non-training AI server growth of 6% YoY.
2. Blackwell production shipments will start in 2Q24F, ramp up in 3Q24F, and customers should have GPUs in data centers in 4Q24F. We expect Nvidia to see a 10% sales weighting for Blackwell in 2024F and 70-80% in 2025F. CSPs, enterprise, consumer IoT and governments are clients for AI GPUs & the main sales growth drivers.
3. GB200 rack demand is from top CSPs and server brands, and we expect 40-45k racks to ship in 2025F, with the top-4 US CSP comprising 80-85% of demand. Hon Hai (2317 TT) and Quanta Computer (2382 TT) are key assemblers. Thermal solution, power supply, DAC cable, and high-end optical fiber connector sales will also benefit from this trend.

Event

We expect cloud AI to grow on CSPs' aggressive capex expansion and a growing GPU supply. Global server shipment will grow 5-10% YoY in 2024-25F with key drivers from booming AI server demand, benefiting related suppliers.

Server demand growth in 2024-25F. Global server shipments fell by 19% YoY to 11.3mn units in 2023. This weakness was related to budget cannibalization from CSPs, as high AI infrastructure cost took a big portion of their budget that was not put AI into consideration in early of 2023. However, the top four US CSPs have intensified their capex to build technical infrastructure and AI offerings, evidenced by rising 2024F capex guidance during recent CY1Q24 earnings calls. Consensus now forecasts the top four US CSPs' capex to increase by 38% YoY in 2024F, up from the 26% forecast before the 1Q24 earnings season, and expects a further increase of 9% in 2025F. With increasing penetration rate of new CPU platform to drive recovery of general server in 2H24-2025F, we expect total global server shipments to grow 5% YoY in 2024F to 11.8mn units, with non-AI server sales growth flat, while global server shipments will grow 10% in 2025F to 13mn units with non-training AI server growth of 6% YoY.

Nvidia's Blackwell GPUs in the pipeline to boost 2025F AI server sales. TSMC (2330 TT) has expanded CoWoS capacity from 13k wafers per month (kwpm) in 4Q23 to 40kwpm in 4Q24F and 55kwpm in 4Q25F expected. Total AI GPU shipments for Nvidia will rise to 4.0mn units in 2024F, and up to 4.7-4.8mn units in 2025F, based on a Hopper GPU allocation of 85% in 2024F and 25% in 2025F, and Blackwell GPU allocations of 10% in 2024F and 75% in 2025F. We think Blackwell GPUs should come on stream in September, with some volume in the market in 4Q24F. B100 and B200 GPUs, and the GB200, should launch in the market around the same time, with significant volume available in 2025F. Combined with AMD's (US) MI300X/325X/350, Intel's (US) Gaudi 2 and 3, and ASIC shipments (mainly from Google's (US) TPU and Amazon Web Services' (AWS; US) Trainium), we estimate total training AI GPU shipments will grow to 5.5mn units in 2024F and 7.9mn units in 2025F, for AI training server shipments of 715k units in 2024F and 1.25mn in 2025F. The total AI server shipment weighting will be even higher if inference servers are taken into consideration.

AI servers carry higher content value for thermal, power supply & rack assembly sectors. During a May 22 earnings call, Nvidia cited Blackwell production shipments will start in 2Q24F, ramp in 3Q24F, and customers should have the GPUs in data centers in 4Q24F. CSP, enterprise, consumer IoT and sovereign are clients for the firm's AI GPUs, and the main sales growth drivers. We expect GB200 rack demand among top CSPs and server brands to result in 40-45k racks shipping in 2025F. Hon Hai (2317 TT) group's FII (CN) is the major GB200 rack supplier to Microsoft (US) and Oracle (US), and Quanta Computer (2382 TT) is for AWS, Meta (US) and Google. We expect Hon Hai's rack assembly market share for GB200 servers to be approximately 40%, with Quanta holding around 30%. The remaining 30% will go to Wiyynn (6669 TT), Wistron (3231 TT), ZT Systems (US) and Super Micro Computer (US). The high ASP of GB200 server racks of US\$1.8mn-3.5mn for NVL36/ 72 racks should lift ODMs' AI server sales weighting to over 50% of total server revenue in 2025F. Among component sectors, the thermal sector will see the greatest content value rise on liquid cooling adoption. With cold plate modules, coolant distribution units (CDU), manifolds, racks, RDHx, fan door and chassis required, we expect the thermal content value to rise to US\$40-80k for each GB200 rack, versus an air cooling content value of US\$2-3k per rack. On the power supply side, with power consumption per GB200 chip reaching 2.7kW, a NVL72 rack could require power of 120kW, or even higher. We expect the content value for each NVL72 rack to be at least US\$18-20k, versus US\$6-8K for H100 designs. We are also optimistic about the increase in the number and specifications of NVLinks when GPUs migrate from the Hopper-series to Blackwell-series, benefiting DAC cable and high-end optical fiber connector suppliers.

Stocks for Action

We expect key beneficiaries of the cloud and edge AI growth wave to include Hon Hai, Quanta Computer, Wiyynn, Wistron, Lotes (3533 TT), Asia Vital Components (AVC; 3017 TT), Auras Technology (3324 TT), Kaori Heat Treatment (8996 TT), Delta Electronics (2308 TT), Chenbro Micom (8210 TT), and King Slide Works (2059 TT).

Risks

Weak demand; over-ordering of AI servers by CSPs; margin dilution for AI servers.

Key assumption chart
Figure 2: AI GPU & AI server shipment forecasts in 2024-25F

Nvidia's training GPU units (k units)	2024F	2025F
Total Nvidia GPU	4,005	4,760
H100/ A100 shipments	3,605	1,190
Blackwell shipments (B100/B200/GB200)	401	3,570
B100/B200	200	1,474
GB200	200	2,096
Training GPU weighting (%)	2024F	2025F
H100/ A100 shipments	90	25
Blackwell shipments (B100/B200/GB200)	10	75
B100/B200	50	41
GB200	50	59
Nvidia's AI training server shipments (k units)	2024F	2025F
H100 /A100 (8 GPU per server)	451	149
B100/B200 (8 GPU per server)	25	184
GB200 (4 GPU per server)	50	524
Nvidia's total AI training server	526	857
AI training GPU shipments (k units)	2024F	2025F
Nvidia (H/B-series)	4,005	4,760
AMD (MI300X/350)	420	768
Intel (Gaudi 2/3)	232	450
Subtotal	4,657	5,978
ASIC (Google TPU/AWS Trainium)	864	1,888
Total training GPU	5,521	7,866
AI training server shipments (k units)	2024F	2025F
Nvidia (H/B-series)	526	857
AMD (MI300X/350)	53	96
Intel (Gaudi 2/3)	29	56
Subtotal	607	1,009
ASIC (Google TPU/AWS Trainium)	108	236
Total AI training server shipments	715	1,245
Server shipments (k units)	2024F	2025F
AI training server	715	1,245
Others (General server + AI inferencing)	11,106	11,759
Total server	11,822	13,004
YoY (%)	2024F	2025F
AI training server	188	74
Others (General server + AI inferencing)	1	6
Total server	5	10
Weighting (%)	2024F	2025F
AI training server	6	10
Others (General server + AI inferencing)	94	90
Total server	100	100

Source: KGI Research estimates

Figure 3: Capex of top-four US CSPs fell 2% YoY in 2023, but will grow 38% YoY in 2024F

Capex (US\$m)	1Q22	2Q22	3Q22	4Q22	1Q23	2Q23	3Q23	4Q23	1Q24	2019	2020	2021	2022	2023	2024F	2025F
Meta (Facebook)	5,441	7,572	9,375	9,043	6,842	6,216	6,543	7,665	6,400	15,102	15,115	18,567	31,431	27,266	37,383	41,418
Amazon	14,951	15,724	16,378	11,268	14,207	11,455	12,479	14,588	14,925	16,861	35,044	55,396	58,321	48,133	62,228	67,213
Microsoft	5,340	6,871	6,283	6,274	6,607	8,943	9,917	9,735	10,952	13,546	17,592	23,216	24,768	35,202	50,363	57,358
Google	9,786	6,828	7,276	7,595	6,289	6,888	8,055	11,019	12,012	23,548	22,281	24,640	31,485	32,251	46,817	48,241
US hyperscale subtotal	35,518	36,995	39,312	34,180	33,945	33,502	36,994	43,007	44,289	69,057	90,032	121,819	146,005	142,852	196,791	214,229
YoY (%)	1Q22	2Q22	3Q22	4Q22	1Q23	2Q23	3Q23	4Q23	1Q24	2019	2020	2021	2022	2023	2024F	2025F
Meta (Facebook)	27.4	64.2	117.4	68.4	25.7	(17.9)	(30.2)	(15.2)	(6.5)	8.0	0.1	22.8	69.3	(13.3)	37.1	10.8
Amazon	23.7	10.1	4.0	(40.5)	(5.0)	(27.1)	(23.8)	29.5	5.1	25.6	107.8	58.1	5.3	(17.5)	29.3	8.0
Microsoft	4.9	6.5	8.1	7.0	23.7	30.2	57.8	55.2	65.8	(4.8)	29.9	32.0	6.7	42.1	43.1	13.9
Google	64.7	24.2	6.7	19.0	(35.7)	0.9	10.7	45.1	91.0	(6.3)	(5.4)	10.6	27.8	2.4	45.2	3.0
US Hyperscale subtotal	29.7	19.9	20.3	(6.5)	(4.4)	(9.4)	(5.9)	25.8	30.5	3.4	30.4	35.3	19.9	(2.2)	37.8	8.9
QoQ (%)	1Q22	2Q22	3Q22	4Q22	1Q23	2Q23	3Q23	4Q23	1Q24	2019	2020	2021	2022	2023	2024F	2025F
Meta (Facebook)	1.3	39.2	23.8	(3.5)	(24.3)	(9.1)	5.3	17.1	(16.5)							
Amazon	(21.0)	5.2	4.2	(31.2)	26.1	(19.4)	8.9	16.9	2.3							
Microsoft	(9.0)	28.7	(8.6)	(0.1)	5.3	35.4	10.9	(1.8)	12.5							
Google	53.3	(30.2)	6.6	4.4	(17.2)	9.5	16.9	36.8	9.0							
US Hyperscale subtotal	(2.8)	4.2	6.3	(13.1)	(0.7)	(1.3)	10.4	16.3	3.0							

Source: Company data; Bloomberg; KGI Research

Figure 4: Decelerating CSP capex growth in 2023, but market expects CSP capex to resume YoY growth in 2024F

Capex, US\$m	2019	2020	2021	2022	2023	2024F	2025F
Meta	15,102	15,115	18,567	31,431	27,266	37,383	41,418
Amazon	16,861	35,044	55,396	58,321	48,133	62,228	67,213
Microsoft	13,546	17,592	23,216	24,768	35,202	50,363	57,358
Google	23,548	22,281	24,640	31,485	32,251	46,817	48,241
Baidu	931	738	1,689	1,586	1,687	1,669	1,745
Alibaba	6,517	6,379	8,311	5,014	5,286	6,011	6,232
Tencent	3,927	5,719	4,808	4,611	4,371	7,100	6,659
Hyperscale subtotal	80,432	102,867	136,627	157,216	154,196	211,571	228,865
Apple	9,247	8,702	10,388	11,692	9,564	10,918	11,921
IBM	2,286	2,618	2,062	1,346	1,488	1,720	1,934
Oracle	1,591	1,833	3,118	6,678	6,935	9,636	9,965
Paypal	704	866	908	706	759	800	946
eBay	508	463	444	420	455	500	504
Salesforce	643	710	717	798	813	739	821
Netflix	253	498	525	408	349	428	465
Uber	588	616	298	252	238	304	338
Enterprise subtotal	15,820	16,306	18,460	22,300	20,601	25,046	26,894
Total	96,793	119,173	155,086	179,516	174,797	236,617	255,759
YoY growth, percent	2019	2020	2021	2022	2023	2024F	2025F
Meta	8.5	0.1	22.8	69.3	(13.3)	37.1	10.8
Amazon	25.6	107.8	58.1	5.3	(17.5)	29.3	8.0
Microsoft	6.0	29.9	32.0	6.7	42.1	43.1	13.9
Google	(6.3)	(5.4)	10.6	27.8	2.4	45.2	3.0
Baidu	(29.9)	(20.7)	129.1	(6.1)	6.3	(1.1)	4.6
Alibaba	(11.9)	(2.1)	30.3	(39.7)	5.4	13.7	3.7
Tencent	17.0	45.6	(15.9)	(4.1)	(5.2)	62.4	(6.2)
Hyperscale subtotal	4.0	27.9	32.8	15.1	(1.9)	37.2	8.2
Apple	(26.7)	(5.9)	19.4	12.6	(18.2)	14.2	9.2
IBM	(32.7)	14.5	(21.2)	(34.7)	10.5	15.6	12.4
Oracle	8.4	15.2	70.1	114.2	3.8	38.9	3.4
Paypal	(14.5)	23.0	4.8	(22.2)	7.6	5.4	18.3
eBay	(22.0)	(8.9)	(4.1)	(5.3)	8.2	9.9	0.8
Salesforce	8.1	10.4	1.0	11.3	1.8	(9.1)	11.1
Netflix	45.5	96.8	5.4	(22.3)	(14.5)	22.9	8.7
Uber	5.4	4.8	(51.6)	(15.4)	(5.6)	27.7	11.2
Enterprise subtotal	(22.0)	3.1	13.2	20.8	(7.6)	21.6	7.4
Total	(1.3)	23.1	30.1	15.8	(2.6)	35.4	8.1

Source: Company data; Bloomberg; KGI Research

Figure 5: Top four US CSP capex outlook – Positive growth, with better-than-expected guidance

Company	Time	Actual & Guidance
Microsoft	1Q24	<ul style="list-style-type: none"> Capex grew 80% YoY and 22% QoQ to \$14bn to support cloud demand, inclusive of the need to scale AI infrastructure Expect capex to grow materially QoQ, driven by cloud and AI infrastructure investments and seasonality
	2Q24F	<ul style="list-style-type: none"> BBG consensus: US\$13.1bn (+20% QoQ)
	2024F	<ul style="list-style-type: none"> Guide capex to grow YoY in FY25F (year end June), to meet the growing demand signal for cloud and AI products. BBG consensus: US\$50.32bn (+43% YoY)
Google	1Q24	<ul style="list-style-type: none"> 1Q24 capex rose 91% YoY to US\$12bn, beating consensus by 17%, driven overwhelmingly by investment in technical infrastructure with the largest component for servers followed by data centers, especially for AI
	2024F	<ul style="list-style-type: none"> Guide capex in 2Q24-4Q24 to be roughly at or above the Q1 level. That's it, 2024F capex will grow over 50% YoY to above US\$48bn, beating consensus forecast by 6% (US\$45bn, up 40% YoY) Capex will be largely used for technical infrastructure in 2024, while investment in offices will be flat YoY, accounting for less than 10% of total capex in 2024
Meta	1Q24	<ul style="list-style-type: none"> 1Q24 capex was US\$6.4bn, down 7% YoY and 17% QoQ, below consensus by 9%, driven by investments in servers, data centers, and network infrastructure.
	2024F	<ul style="list-style-type: none"> Revise up 2024 capex guidance by 12%, up from US\$30-37bn to US\$35-40 (up 33% YoY at midpoint), beating consensus forecast by 9% as Meta continues to accelerate infra investments to support AI roadmap
	2025F	<ul style="list-style-type: none"> Expect capex to grow YoY in 2025F as the company invests aggressively to support AI research and product development efforts BBG consensus: \$40.9bn (+11% YoY)
amazon	1Q24	<ul style="list-style-type: none"> Capex was US\$14.9bn, up 5% YoY and 2% QoQ, in line with consensus
	2024F	<ul style="list-style-type: none"> Expect capex to meaningfully increase YoY in 2024, driven by infrastructure to support AWS's reaccelerating growth including high demand for gen AI Expect 1Q24 capex to be the lowest quarter of 2024F, implying full-year capex will increase YoY to over \$59.6bn (over 23% YoY), ahead of consensus Consensus forecast 2024F capex of \$62.05bn (+29% YoY)

Source: Company data, Bloomberg, KGI Research

Figure 6: Top four US CSP cloud business outlook

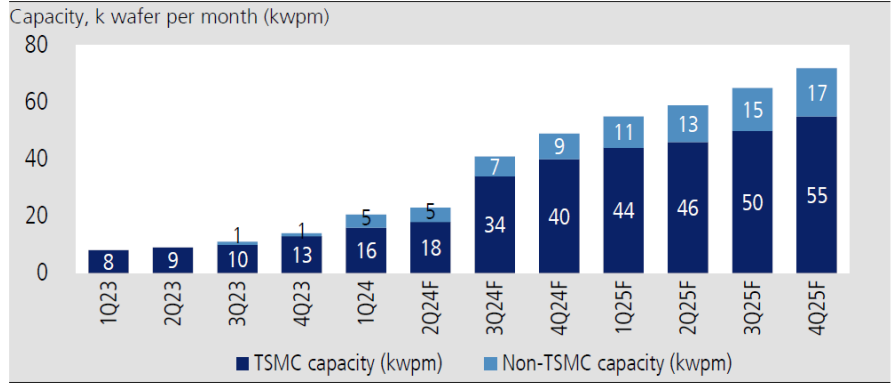
Company	Time	Actual & Guidance
Microsoft	1Q24	<ul style="list-style-type: none"> Intelligent Cloud revenue was \$26.7bn, up 21% YoY and 3% QoQ, beating both the company guidance (\$26-26.3bn) and the consensus (\$26.2bn) Azure and other cloud service revenue growth of 31% YoY in cc, beating both the consensus and the company's guidance of 28%
	2Q24F	<ul style="list-style-type: none"> Guides Azure sales to grow 30-31% YoY, ahead of consensus of 28% YoY
	FY24F	<ul style="list-style-type: none"> Consensus expects Intelligent Cloud sales to grow 19% YoY in FY2024 Consensus expects Azure sales to grow 28% YoY in cc in FY2024
Google	1Q24	<ul style="list-style-type: none"> Google Cloud sales rose 28% YoY and 4% QoQ to \$9.5bn, ahead consensus by 2%, driven by demand for GCP infrastructure and solutions
	2024F	<ul style="list-style-type: none"> Consensus expects Google Cloud sales to grow 25% YoY in 2024F
amazon	1Q24	<ul style="list-style-type: none"> AWS revenue grew 17% YoY and 3% QoQ to US\$25bn in 1Q24, beating consensus by 4% Witnessed growth in both Gen AI and non-Gen AI workloads across a diverse group of customers and across different industries, since companies are migrating more workloads to the cloud, while signing up for longer deals, making bigger commitments.
	2024F	<ul style="list-style-type: none"> The company continues to see the impact of cost optimization diminishing, causing customers turning their attention to newer initiatives and re-accelerating existing Consensus AWS sales growth of 15% YoY to \$104bn

Source: Company data, Bloomberg, KGI Research

Figure 7: AI training server weighting of total server shipments to rise from 2% in 2023 to 10% in 2025F

Server shipments (k units)	2022	2023	2024F	2025F
AI training server	124	249	715	1,245
Others (General server + AI inferencing)	13,703	11,010	11,106	11,759
Total server	13,827	11,259	11,822	13,004
YoY (%)	2022	2023F	2024F	2025F
AI training server		101	188	74
Others (General server + AI inferencing)		(20)	1	6
Total server	7	(19)	5	10
Weighting (%)	2022	2023F	2024F	2025F
AI training server	1	2	6	10
Others (General server + AI inferencing)	99	98	94	90
Total server	100	100	100	100

Source: Gartner; KGI Research estimates

Figure 8: CoWoS capacity expansion to support AI server demand increase

Figure 9: Nvidia's AI training server shipments will grow to 526k units in 2024F and 857k units in 2025F

Nvidia's training GPU units (k units)	2024F	2025F
Total Nvidia GPU	4,005	4,760
H100/ A100 shipments	3,605	1,190
Blackwell shipments (B100/B200/GB200)	401	3,570
B100/B200	200	1,474
GB200	200	2,096
Training GPU weighting (%)	2024F	2025F
H100/ A100 shipments	90	25
Blackwell shipments (B100/B200/GB200)	10	75
B100/B200	50	41
GB200	50	59
Nvidia's AI training server shipments (k units)	2024F	2025F
H100 /A100 (8 GPU per server)	451	149
B100/B200 (8 GPU per server)	25	184
GB200 (4 GPU per server)	50	524
Nvidia's total AI training server	526	857

Source: KGI Research estimates

Figure 10: AI server shipments to rise in 2024-25F; we revise up 2024-25F demand

AI training GPU shipments (k units)	2024F	2025F
Nvidia (H/B-series)	4,005	4,760
AMD (MI300X/350)	420	768
Intel (Gaudi 2/3)	232	450
Subtotal	4,657	5,978
ASIC (Google TPU/AWS Trainium)	864	1,888
Total training GPU	5,521	7,866
AI training server shipments (k units)	2024F	2025F
Nvidia (H/B-series)	526	857
AMD (MI300X/350)	53	96
Intel (Gaudi 2/3)	29	56
Subtotal	607	1,009
ASIC (Google TPU/AWS Trainium)	108	236
Total AI training server shipments	715	1,245

Source: KGI Research estimates

Figure 11: AI server matrix between major CSP, enterprises & ODMs in 2024F; all Taiwan ODMs are key assemblers (H100 GPU)

			2024F AI server shipments allocation (%) - based on L10							
AI server client	GPU solution		Quanta	Hon Hai	ZT / Inventec	Wiwynn	Wistron	Gigabyte	Supermicro	Dell / Lenovo / Inspur / Others
CSP	Microsoft	Nvidia / AMD	35	30	30	5				
	Google	Nvidia / TPU	50	15	35					
	AWS	Nvidia / Trainium	30	15	35	20				
	Meta	Nvidia / AMD	70	12		18				
	BBAT	Nvidia / Habana		5	10					85
	Oracle	Nvidia / AMD		90						10
Enterprise	Major enterprise	Nvidia		30	10		30		20	10
	Nvidia DGX	Nvidia					100			
	Tesla	Nvidia / Dojo					40		60	
Channels	Coreweave	Nvidia						10	75	15
	Other channels	Nvidia						35	50	15
Total server sales in 2024F (NT\$bn)			728	1,222	260	345	448	170	670	
Server sales weighting (%)			50	19	45	100	45	60	100	
AI server sales weighting (%)										
- of server sales			50	40	18	35	41	80		
- of total sales			25	7	8	35	18	48		

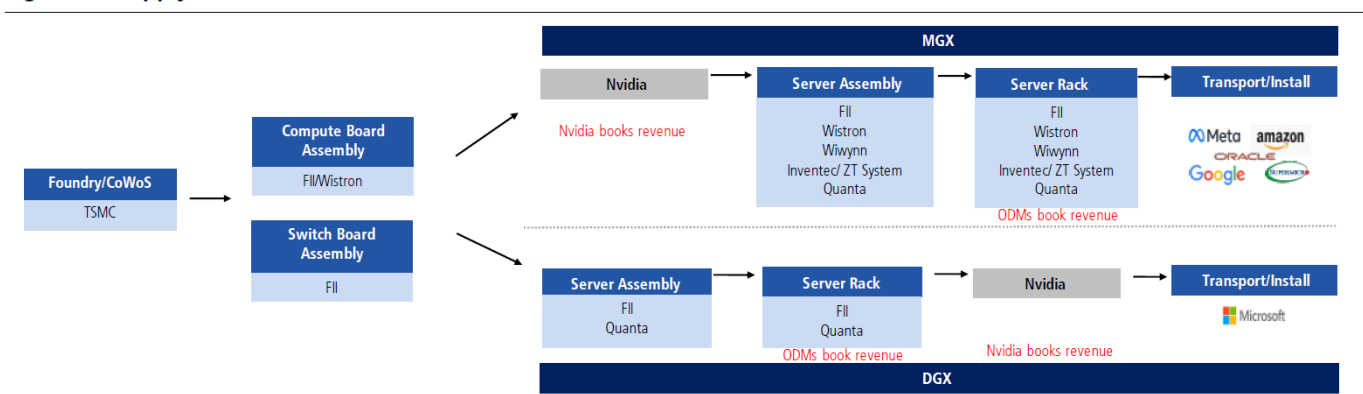
Source: KGI Research estimates

Figure 12: AI server sales to account for over 50% of total server sales for most ODMs in 2025F

Products	Ticker	Company	Server sales (NT\$bn)			Weighting of total sales (%)			YoY (%)		
			2023	2024F	2025F	2023	2024F	2025F	2023	2024F	2025F
ODM	2317 TT	Hon Hai	1,018	1,222	1,589	17	19	20	(35)	20	30
	2382 TT	Quanta	384	728	1,311	35	50	63	11	89	80
	3231 TT	Wistron	307	448	595	35	45	51	(13)	46	33
	2356 TT	Inventec	208	260	315	40	45	47	(7)	25	21
	6669 TT	Wiwynn	242	345	462	100	100	100	(17)	43	34
Brand	2357 TT	Asustek	16	45	70	3	8	11	60	181	56
	2376 TT	Gigabyte	53	170	187	39	60	60	162	222	10

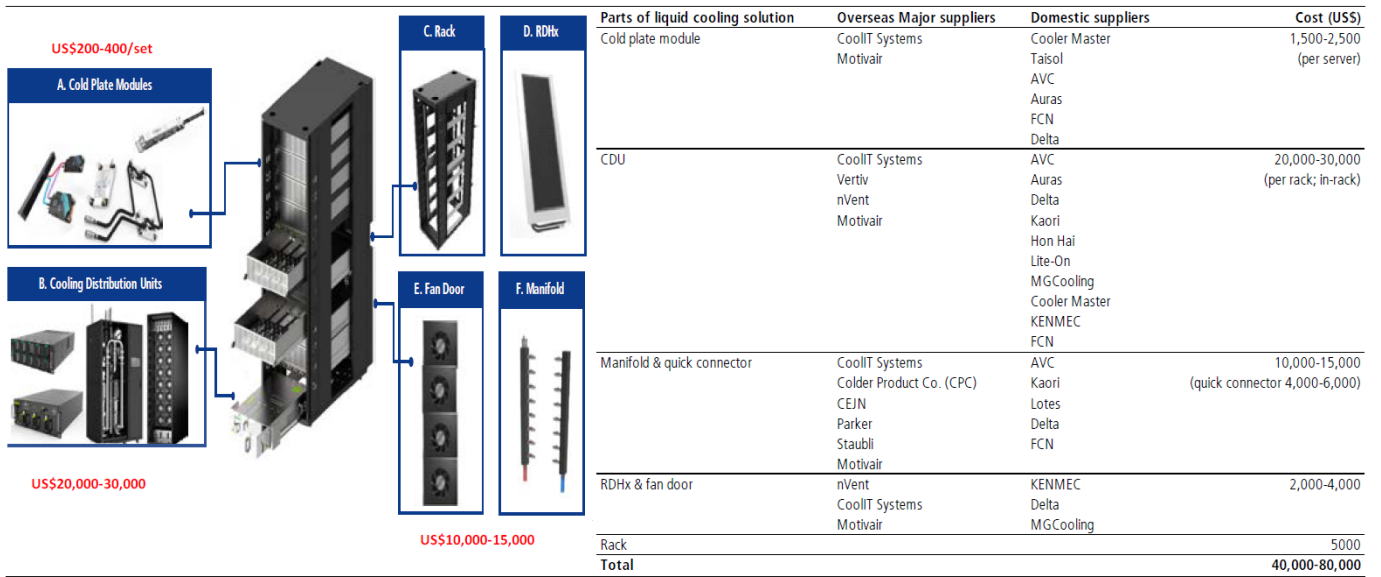
	Ticker	Company	AI server sales (NT\$bn)			Weighting of server sales (%)			YoY (%)		
			2023	2024F	2025F	2023	2024F	2025F	2023	2024F	2025F
ODM	2317 TT	Hon Hai	305.5	488.9	873.8	30	40	55		60	79
	2382 TT	Quanta	76.9	364.2	917.7	20	50	70		374	152
	3231 TT	Wistron	71.9	181.8	327.2	23	41	55		153	80
	2356 TT	Inventec	14.2	46.9	94.5	7	18	30		231	102
	6669 TT	Wiwynn	48.4	120.8	277.3	20	35	60		150	129
Brand	2357 TT	Asustek	5.6	36.0	56.0	35	80	80		543	56
	2376 TT	Gigabyte	32.1	135.8	149.3	61	80	80		323	10

Source: company data; KGI Research estimates

Figure 13: Supply chain flow chart for GB200


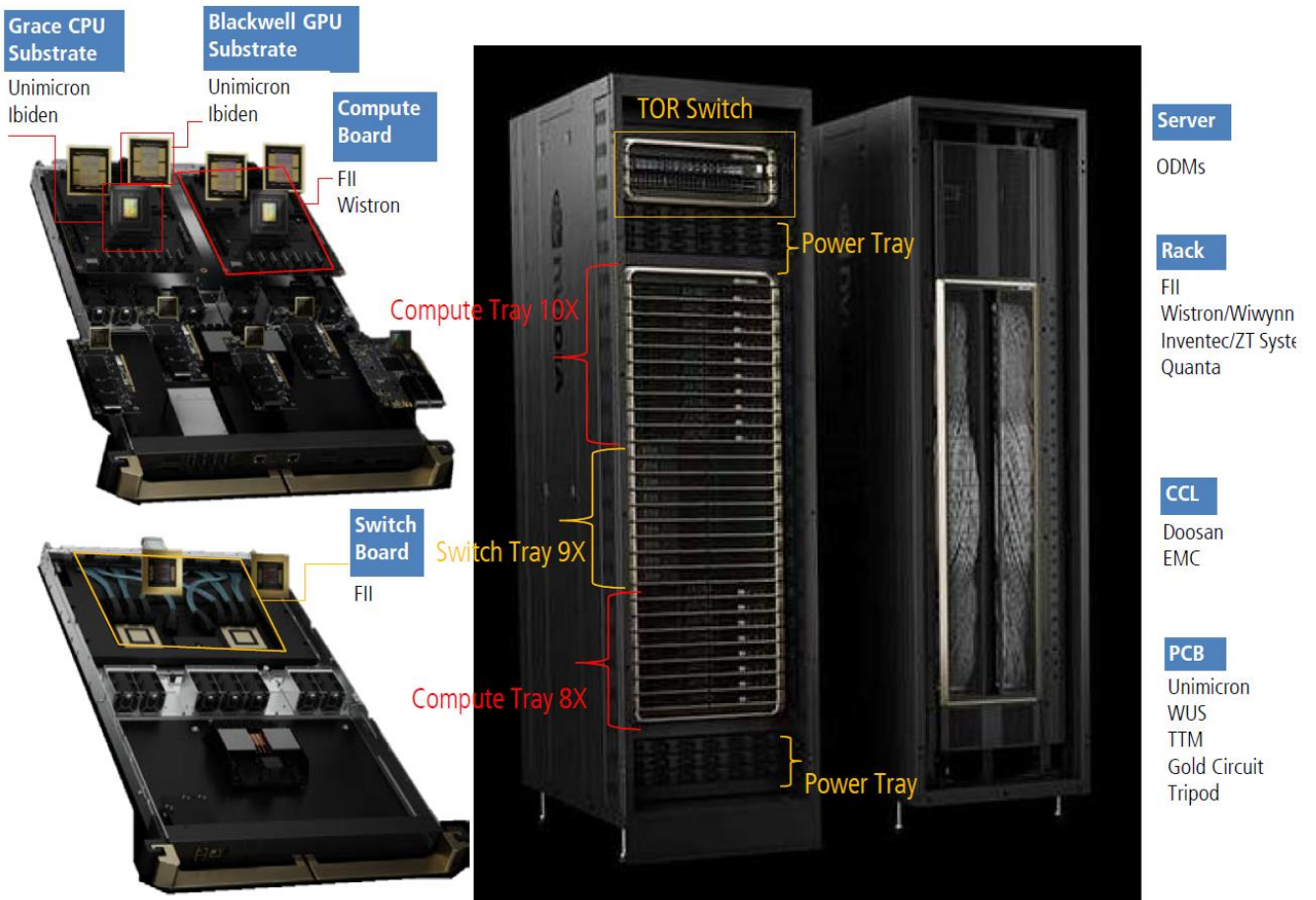
Source: KGI Research

Figure 14: Thermal solution transition to liquid cooling to create much higher content value for thermal plays



Source: Auras; KGI Research

Figure 15: GB200 NVL72 breakdown & supply chain



Source: Nvidia; KGI Research

