See Generative AI's Impact on the AI Server Market to 2025

Presenter: Frank Kung/Senior Analyst

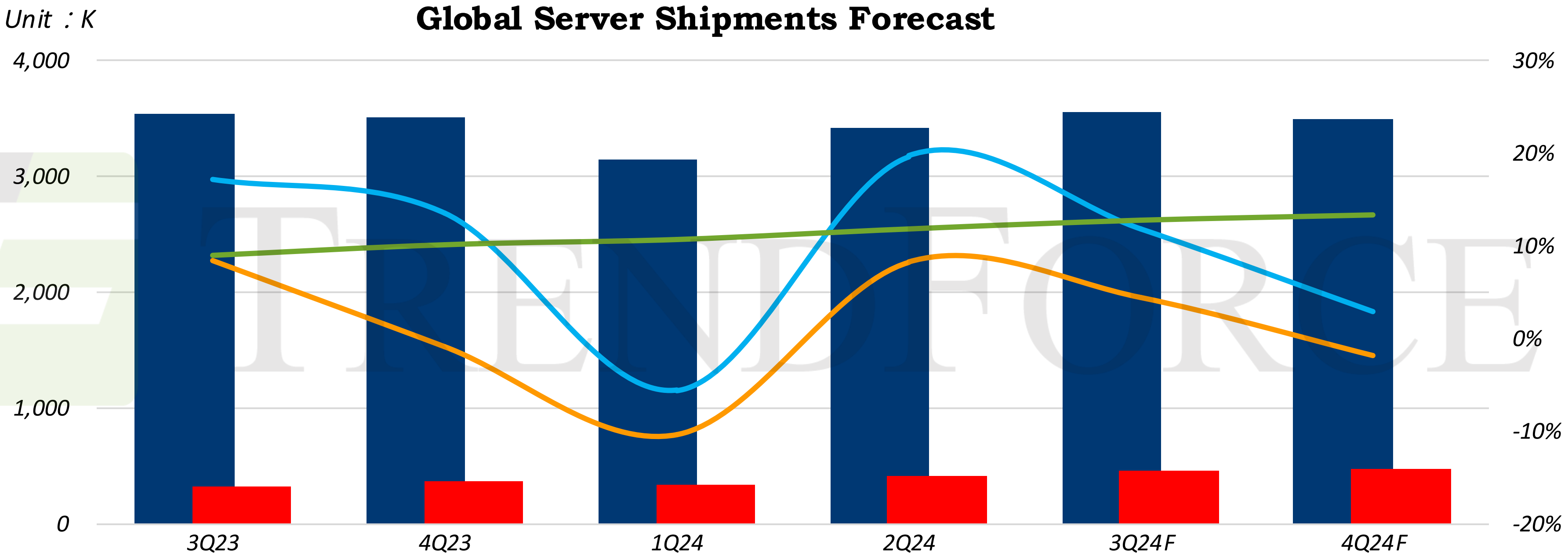
TRENDFORCE

FMS the Future of Memory and Storage

Projected Global Shipment and Ratio Changes on Servers and AI Servers

- *Global server shipments are expected to grow by only around **1.9%** in 2024, continuously being **squeezed out** by budgets for **AI servers**.*
- *It’s projected that AI servers will climb to about a **41.5%** YoY growth in 2024, to meet the strong demand of CSPs and OEMs generative AI training and inference application.*

Global Server Shipments Forecast



	3Q23	4Q23	1Q24	2Q24	3Q24F	4Q24F
Total Server Shipments	3,550	3,520	3,154	3,415	3,564	3,499
AI Server Shipments	317	360	340	407	455	468
Total Server QoQ	8.4%	-0.8%	-10.4%	8.3%	4.4%	-1.8%
AI Server QoQ	17.2%	13.6%	-5.6%	19.7%	11.8%	2.9%
AI Server M/S	8.9%	10.2%	10.8%	11.9%	12.8%	13.4%

Note: AI Servers include AI Training and AI Inference Servers.

Source: TrendForce, Aug., 2024

Major AI Chip Suppliers Include NVIDIA, AMD, Intel, and CSPs

TRENDFORCE

Supplier	AI Chip	Major AI Application	Major AI Chip Solution Name	Process	Memory
NVIDIA	GPU	AI Training	H100 、H20 、GH200 、H200(2Q24)	4nm	HBM3/3e
			B-series (B100/B200, GB200) (2H24) 、B Ultra(2025)	4nm	HBM3e
			A100	7nm	HBM2e
		AI Training/AI Inference	A30	7nm	HBM2e
			L40s/L20	5nm	GDDR6
		AI Inference	L4/L2	5nm	GDDR6
AMD XILINX	GPU	AI Training	MI200	6nm	HBM2e
			MI300/MI308/MI325(4Q24)	5nm	HBM3/3e
			MI350 (2025)	3nm(f)	HBM3e
		AI Inference	Radeon V	7nm	GDDR6
	FPGA	AI Inference	Versal	7nm	HBM2e
			Virtex	16nm	-
intel	GPU	AI Training	Max GPU	5nm	HBM2e
		AI Training	Gaudi 2/3	5~7nm	HBM2e
		AI Inference	Flex GPU	6nm	GDDR6
	FPGA	AI Inference	Altera Stratix	14nm	HBM2
Google	ASIC	AI Training/AI Inference	TPU v5/v6(f)	4(f)~5nm	HBM2e 、HBM3
aws	ASIC	AI Training/AI Inference	Trainium 、Inferentia	5~7nm	HBM2e/3
Others	ASIC	AI Training/AI Inference	➢ MSFT, Meta, etc. ➢ China players (Like as BAT, Huawei, etc.)	7~12nm	HBM2/2e/3

Source: TrendForce, Aug., 2024

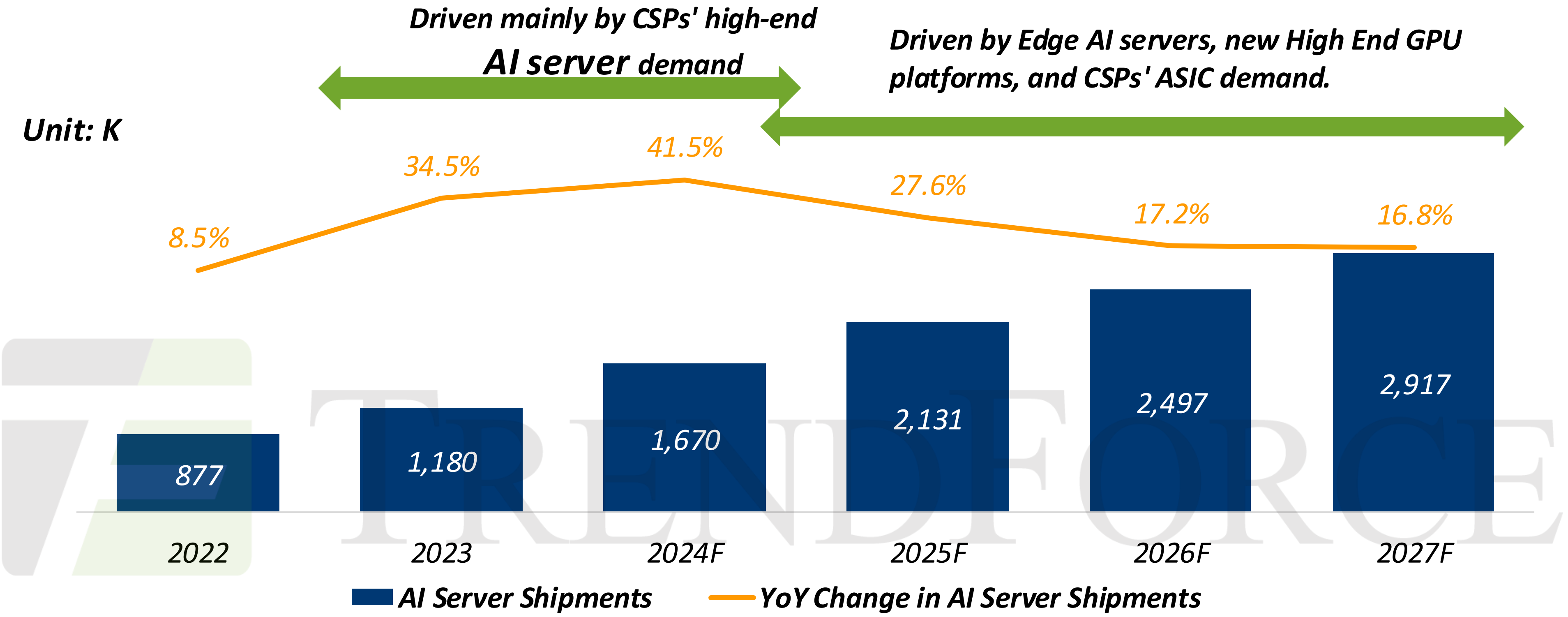
Projected Global Shipment on AI Chips and AI Servers, 2023-2025F

Major Suppliers Adopted with AI Chips *Unit: K*	*AI Application Category*	*Estimated AI Chip Shipment M/S*			*Estimated AI Server Shipment M/S*		
		2023	*2024E*	*2025F*	*2023*	*2024E*	*2025F*
NVIDIA	*AI Training (High-end)*	*26.2%*	*36.0%*	*40.4%*	*18.4%*	*31.3%*	*40.8%*
	AI Inference (low-end)	*21.7%*	*13.2%*	*11.0%*	*47.2%*	*32.3%*	*25.5%*
AMD XILINX	*AI Training (High-end)*	*2.6%*	*3.6%*	*3.7%*	*1.9%*	*3.0%*	*3.4%*
	AI Inference (low-end)	*7.4%*	*4.8%*	*4.0%*	*5.4%*	*5.2%*	*4.4%*
intel	*AI Training (High-end)*	*1.7%*	*1.7%*	*1.1%*	*1.3%*	*1.3%*	*0.9%*
	AI Inference (low-end)	*2.2%*	*1.4%*	*1.2%*	*1.8%*	*1.6%*	*1.6%*
Others(Google, AWS, etc.)	*AI Training (High-end)*	*15.3%*	*15.7%*	*15.4%*	*9.7%*	*11.4%*	*11.1%*
	AI Inference (low-end)	*23.0%*	*23.6%*	*23.1%*	*14.4%*	*13.9%*	*12.3%*
YoY	*-*	*59.7%*	*65%*	*43.8%*	*34.6%*	*41.5%*	*27.6%*
Overall Ratio of AI Servers		*-*	*-*	*-*	*8.8%*	*12.2%*	*14.8%*

Note: The primary configuration of the NVIDIA GB200 solution consists of 1 Grace CPU and 2 Blackwell GPU AI chips.

Source: TrendForce, Aug., 2024

Growth Forecast for AI Server Market, 2022-2027F



Note: Designed for AI training and inference, AI servers are equipped with acceleration chips such as GPU, FPGA, and ASIC.

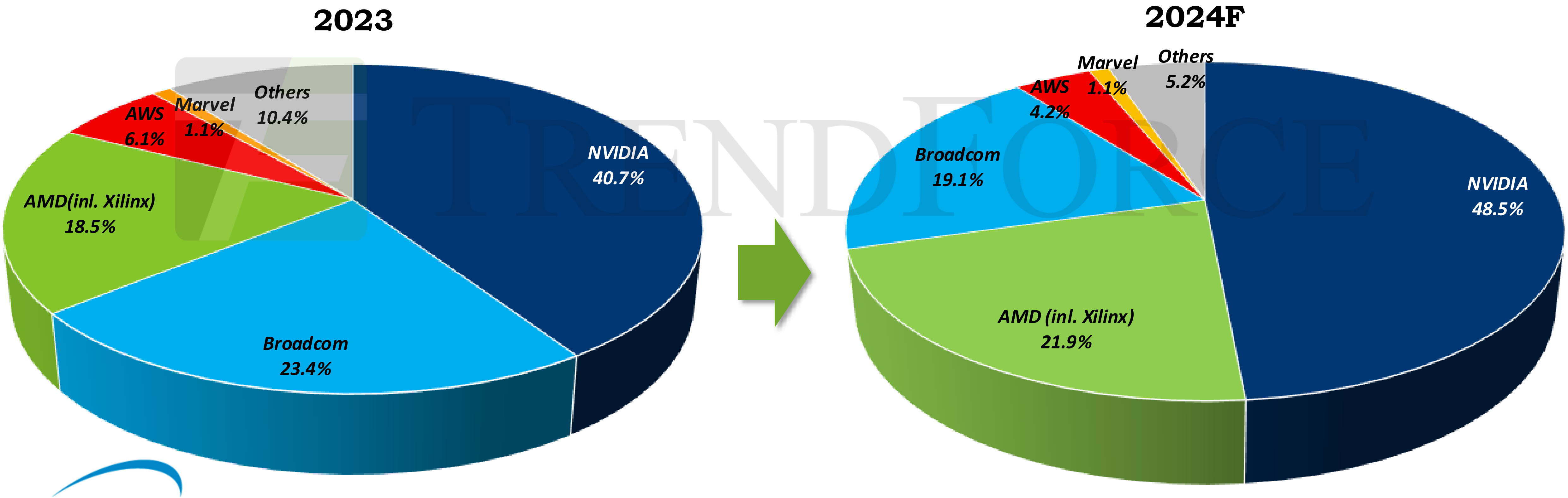
- *The market for AI servers will experience a surging growth during 2023-2024, with YoY growth rates for shipments averaging at around **38%**.*
- *Global shipments of AI servers are projected to increase at a **CAGR of 27.2%** during 2022-2027. By 2027, AI servers are forecasted to account for **around 19%** of the total annual server shipments.*

Source: TrendForce, Aug., 2024

NVIDIA and AMD Will Account for a Greater Portion of TSMC's CoWoS Production Capacity in 2024

- □ *TSMC's CoWoS production capacity is projected to reach over* ***300K*** *at the end of 2024.*
- □ *It is expected that TSMC's CoWoS production capacity goal will reach* ***550~600K*** *by 2025,* and the demand is expected to nearly double next year.

Distribution of TSMC CoWoS Demand among Major AI Chip Suppliers



Source: TrendForce, Aug., 2024

Development of AI Chips and Comparison of HBM Specifications between NVIDIA and AMD in 2023~2025F

Company	AI Chips	2022	2023				2024F				2025F			
			1Q23	2Q23	3Q23	4Q23	1Q24	2Q24	3Q24	4Q24	1Q25	2Q25	3Q25	4Q25
NVIDIA	H100	HBM3 8hi 80GB (16GB*5)												
	GH200 (CPU+GPU)						HBM3e 8hi 141GB (24GB*6)							
	H20						HBM3 8hi 96GB (16GB*6)							
	H200						HBM3e 8hi 141GB (24GB*6)							
	B100/B200								HBM3e 8hi 192GB (24GB*8)					
	GB200 (CPU+GPU)								HBM3e 8hi 192/384GB (24GB*8 /192GB*2)					
	Blackwell Ultra										HBM3e 12hi 288GB (36GB*8)			
AMD	MI200	HBM2e 8hi 128GB (16GB*8)												
	MI300X			HBM3 12hi 192GB (24GB*8)										
	MI300A (CPU+GPU)			HBM3 8hi 128GB (16GB*8)										
	MI325X									HBM3e 12hi 288GB (36GB*8)				
	MI350/MI375 (TBD)											HBM3e 12hi 288GB (36GB*8)		

Source: TrendForce, Aug., 2024

AI Server Supply Chain Will Promote Product Specification and Shipments for NVIDIA New Platform

Supply Chain	*Major Players*	*Forecast of Key Development Trends from 2024 to 2025*
Upstream Key Components	- ***CoWoS*** ：*TSMC, Intel etc.* - ***HBM*** ：*SK hynix, Samsung, Micron* - ***Power related*** ：*Delta, LiteOn, AVC, AURAS, etc.*	*Production will gradually expand for the next-generation **CoWoS-L and HBM3e.*** *It is expected that Blackwell (including GB200, B100, B200, etc.) will drive CoWoS and HBM shipments, leading to **over high double-digit growth.*** *It is expected that AI server unit PSU spec will increase **from 3.3kW to over 5.5kW**, and **liquid cooling** solutions will expand.* *It is expected that **HGX AI servers** will remain the mainstream configuration, with share of **around 50-60%** in 2025.* *In 2025, the GB200 will be initially supplied **to hyper CSPs, followed by NCP and other brand customers** .*
Midstream Manufacturing	- ***ODMs:*** *FII, Inventec, Quanta, Wistron, Wiwynn, Supermicro, etc.*	
Downstream End Customers	- ***Hyper CSPs:*** *Microsoft, AWS, Google, Meta, Oracle, BBAT.* - ***Brands:*** *Dell, HPE, Lenovo, Gigabyte etc.* - ***Others*** ：*CoreWeave, Lambda, Yotta, IBM, NCP related.*	

Source: TrendForce, Aug., 2024

Conclusion

1 *The projected shipments of NVIDIA's high-end GPUs for 2024 total about **3.5 million units**, marking a YoY growth rate of over **120%**. After 2H24, **B-series** will enter the phase of early mass production.*

2 ***ASICs** that CSPs have been developing in-house will also play a key role. In 2024, **AWS** will commit a significant amount of resources into its next-generation ASICs, and the same goes for **Meta and Chinese CSPs**.*

3 *It is expected that in 2025, considering the lower cost and the specific AI application needs, CSPs will still keep **20~30%** of ASIC market share and will not fully adopt GB200, especially **Google, AWS, and Chinese CSPs**.*

4 *It's projected that the release of the NVIDIA's **Blackwell** platform will drive CoWoS and HBM shipments to achieve **over high double-digit growth** in 2025.*

Source: TrendForce, Aug., 2024

TrendForce & TRI Research Areas

- IC
- Memory
- Consumer Electronics
- Green Energy
- Display Panel
- Smartphones & Apps
- LED
- Innovative Technology
- Intelligent Living
- Industry 4.0
- Automotive Technology
- AI
- Communication, Broadband & Networking
- Gov. & Ent. Consulting Service
- Cloud Computing & IoT

Sales & Services

Semiconductor Research

DRAM, NAND Flash, Foundry

SR_MI

SR_MI@TrendForce.com

Green Energy Research

Solar PV

GER_MI

GER_MI@TrendForce.com

Optoelectronics Research

Micro LED, Mini LED, VCSEL, UV, Video Wall, Lighting

OR_MI

OR_MI@TrendForce.com

Display Research

TFT-LCD , OLED , Smartphone , Tablet , NB , Monitor/AIO , TV

DR_MI

DR_MI@TrendForce.com

ICT Application Research

Communication & Broadband, Consumer Electronics, Innovative Technological Applications, Automotive, Industry 4.0, Gov. & Ent.

TRI_MI

TRI_MI@TrendForce.com