



# The Influence of AI on Server Market Dynamics: Projections for 2025

**Presenter: Mark Liu/Research Manager**



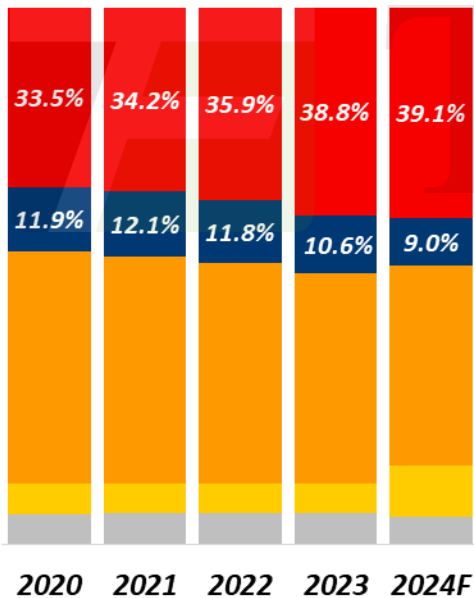
# The Changing Mix of Supply

- ❑ Server DRAM proportion among the three suppliers is expected to increase.
- ❑ Due to falling PC unit demand, PC DRAM has a smaller share of the total DRAM output.
- ❑ Mobile DRAM is still the biggest segmentation in each companies' product mix.

## 2020-2024F Production Mix by Output Chip

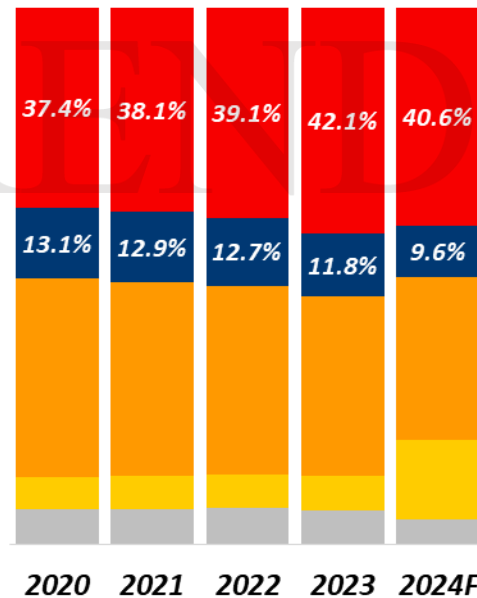
### SAMSUNG

2023 TTL Bit Growth: **-7.9%**  
 2024 TTL Bit Growth: **9.9%**



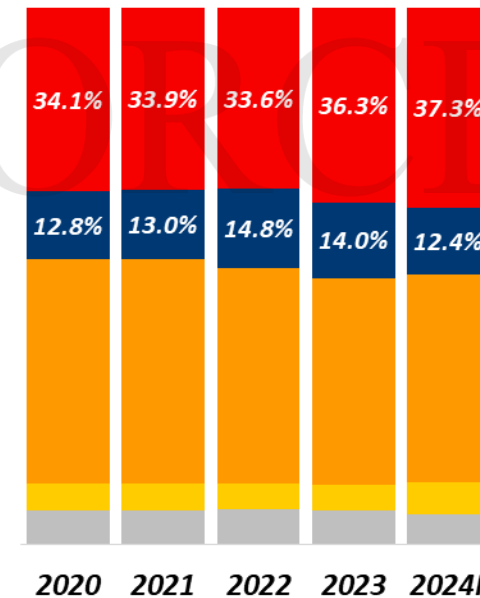
### SK hynix

2023 TTL Bit Growth: **-2.0%**  
 2024 TTL Bit Growth: **12.3%**



### Micron

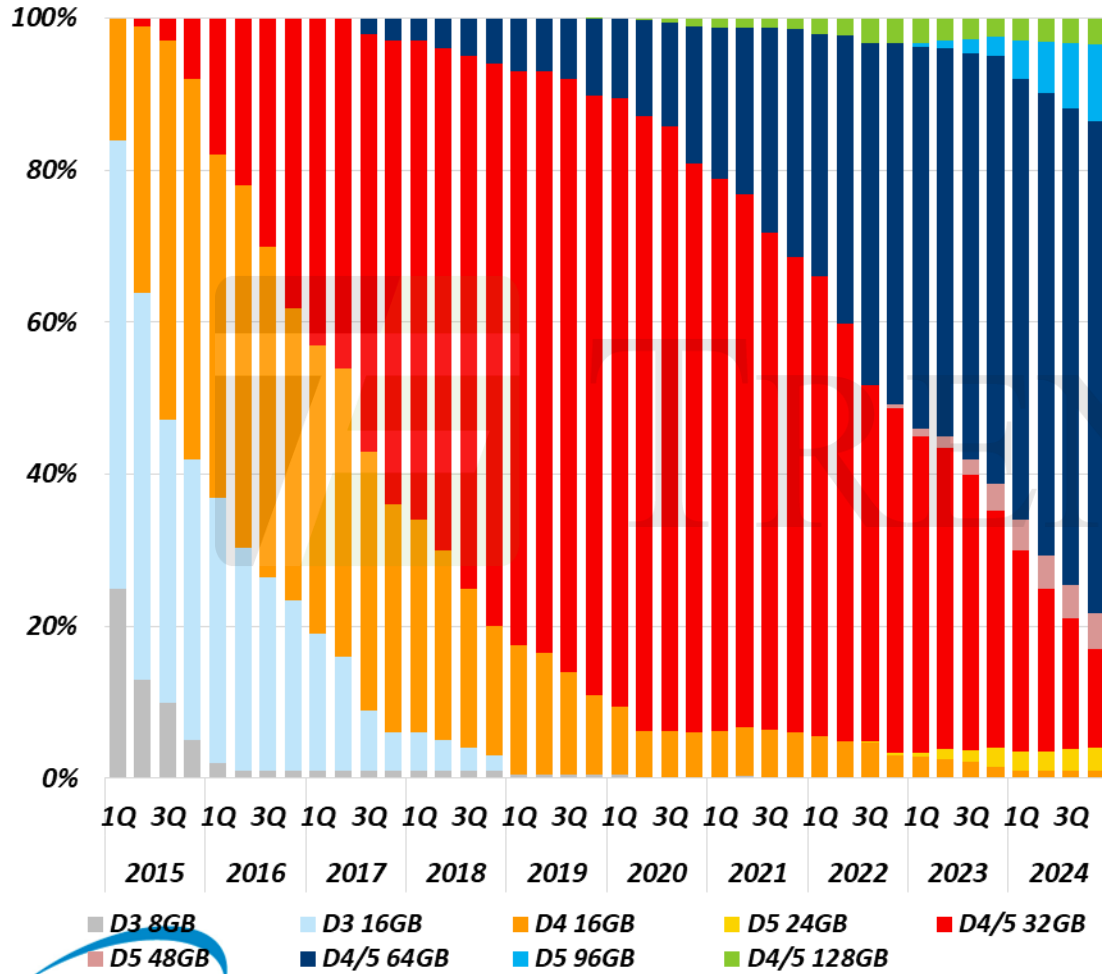
2023 TTL Bit Growth: **-7.0%**  
 2024 TTL Bit Growth: **24.1%**



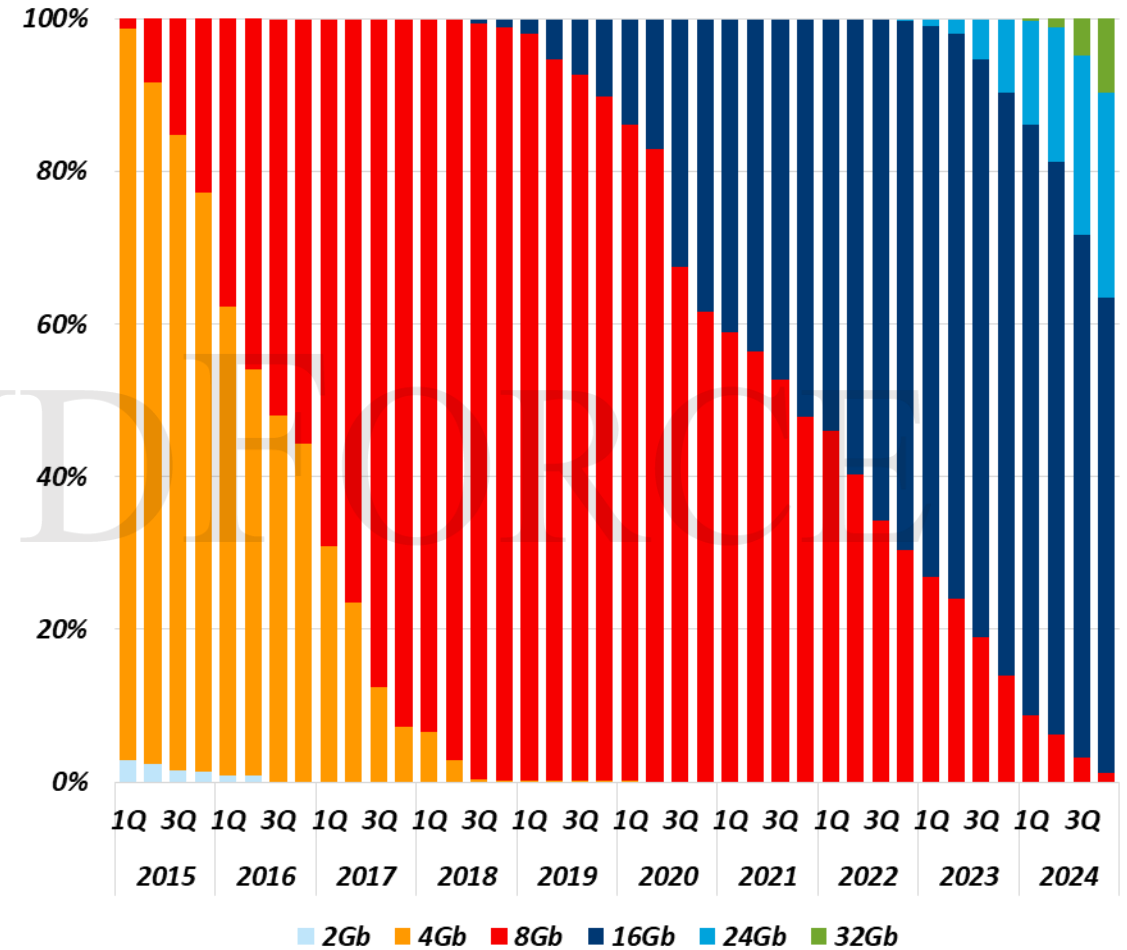
Source: TrendForce, Aug., 2024



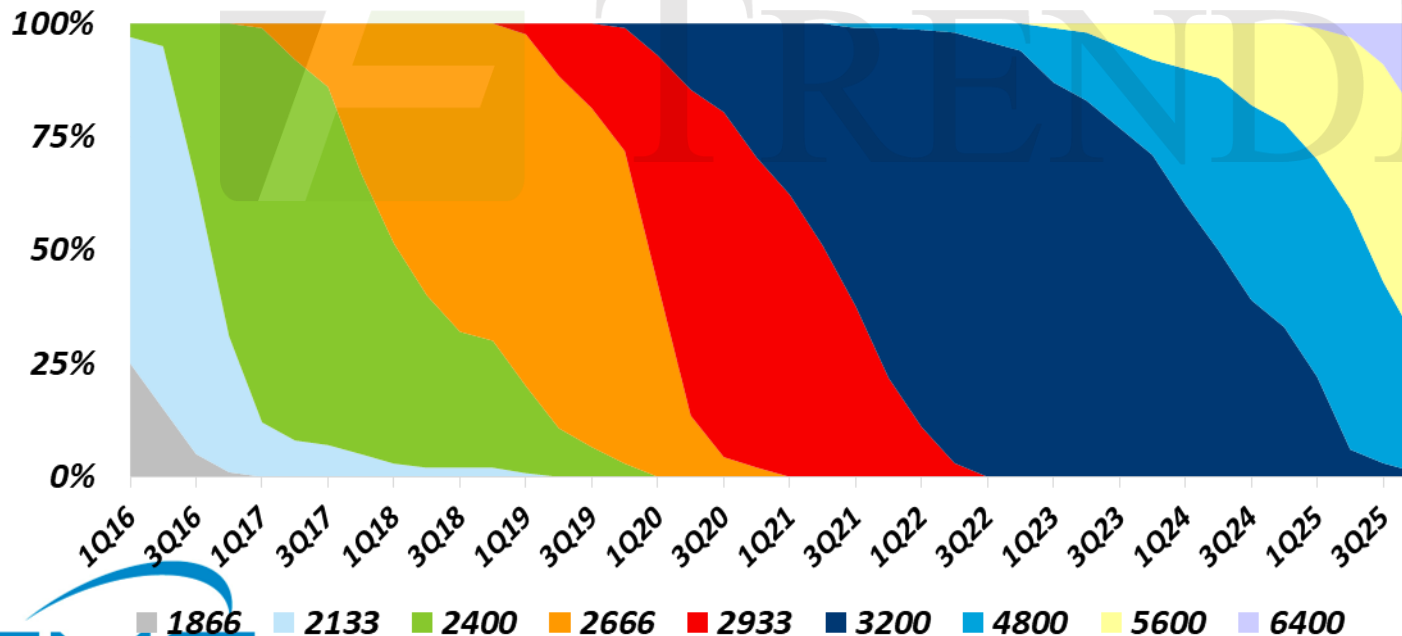
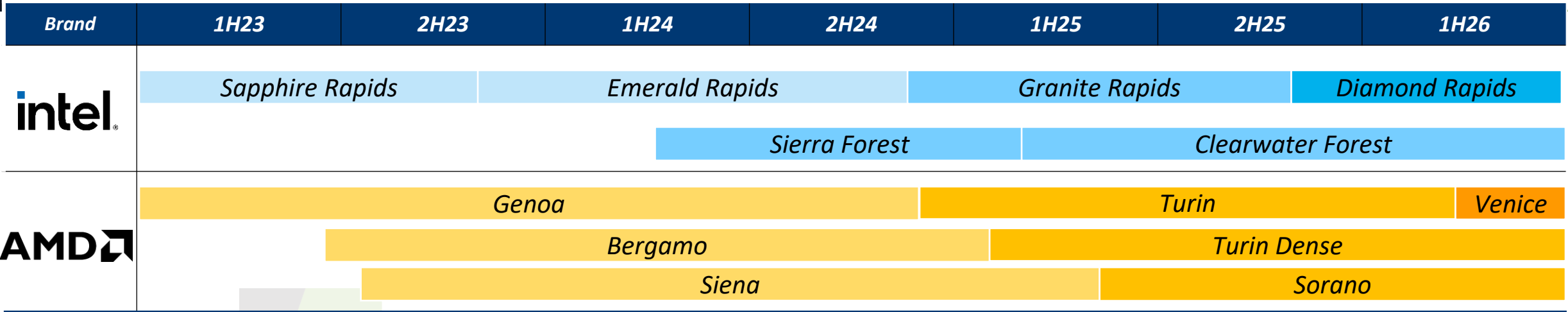
### Server DIMM Shipment Allocation (Type by Density)



### Server DIMM Shipment Allocation (by Density)



# 1Q16-4Q24 Server DIMM Speed Trend



## Key Customer DDR5 Adoption Status:

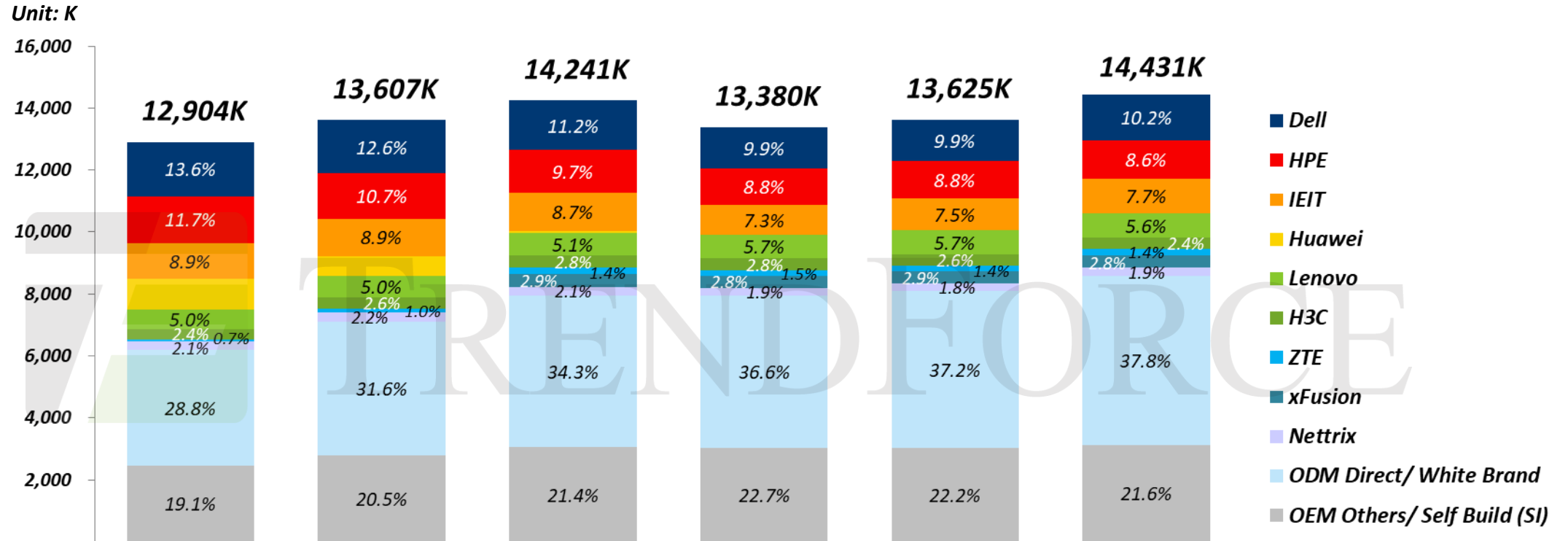
- AWS: at 38% by 2Q24 (Graviton most)
- Meta: at 33% by 2Q24 (AI most)
- Azure: at 17% by 2Q24 (AI most)
- Google: at ~45% by 2Q24 (SPR most)
- HPE: at 18% by 2Q24
- Dell: at ~26% by 2Q24



# 2020-2024F Server Shipment by Brand

Server Shipment continues to grow (without Workstation)

□ ODM Direct outgrows the market as demand of datacenters and cloud computing is strong.

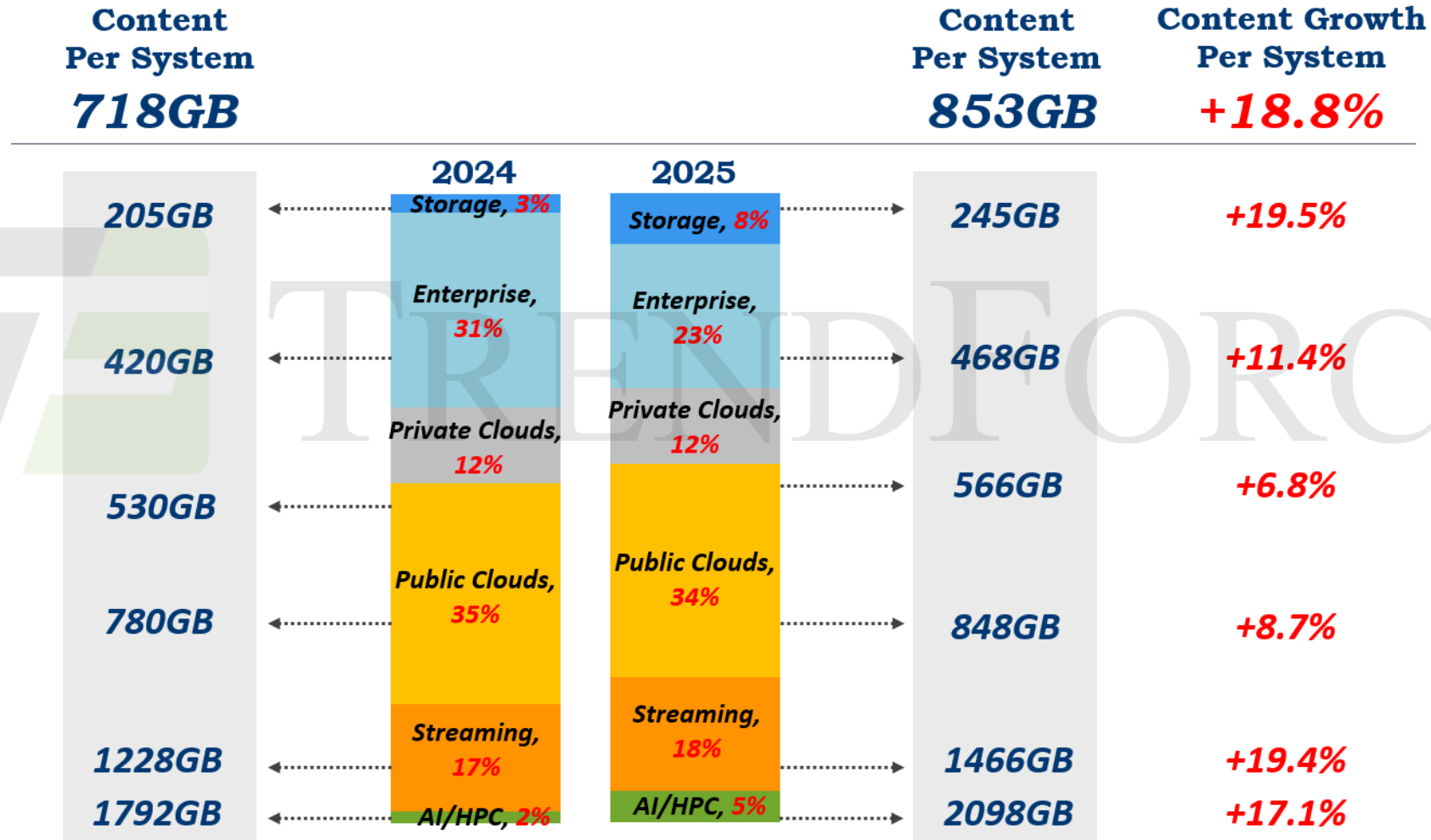


	2020	2021	2022	2023	2024F	2025F
<b>ODM Direct YoY</b>	13.6%	15.6%	13.7%	0.1%	3.5%	7.6%
<b>Total Server YoY</b>	3.5%	5.4%	4.7%	-6.0%	1.8%	5.9%
<b>China Server Domestic Shipment %</b>	27.7%	26.8%	26.7%	24.5%	23.8%	23.1%



# The AI/ HPC Application Drives Server Content Growth by 17.1% in 2025

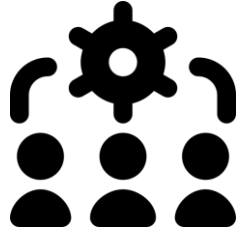
## Server Types and Respective DRAM Content Growth



Source: TrendForce, Aug., 2024

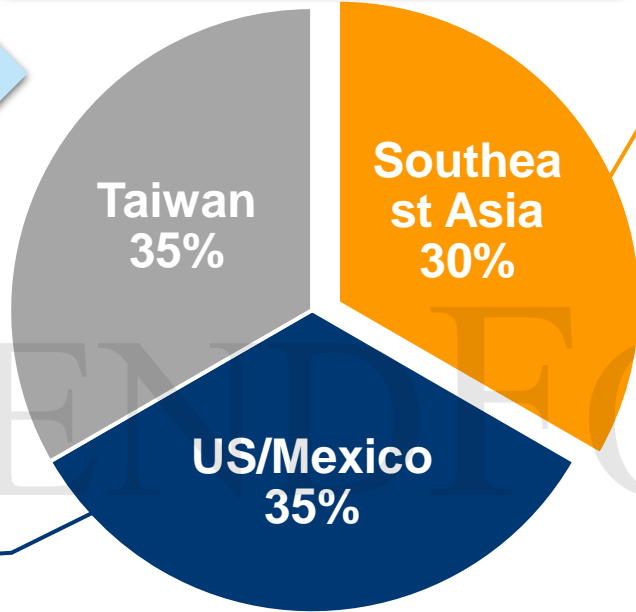
# Emergence of Southeast Asian Market: Transition of Supply Chain under China+1

## 2 Shift of Supply Chain



Projected Ratio of Server Supply Chain by 2027

- US: Foxconn underwent another expansion for its fabs in Wisconsin and Texas in 1H24, while Wistron has initiated pilot runs in California, and Quanta could complete its expansions in California and Tennessee in 2H24.
- Mexico: Foxconn is expected to expand its capacity in Jalisco, and Pegatron is scheduled to begin production for its Juarez fab starting from 2Q24, while Inventec and Wistron are each planning for expansions throughout 2024.

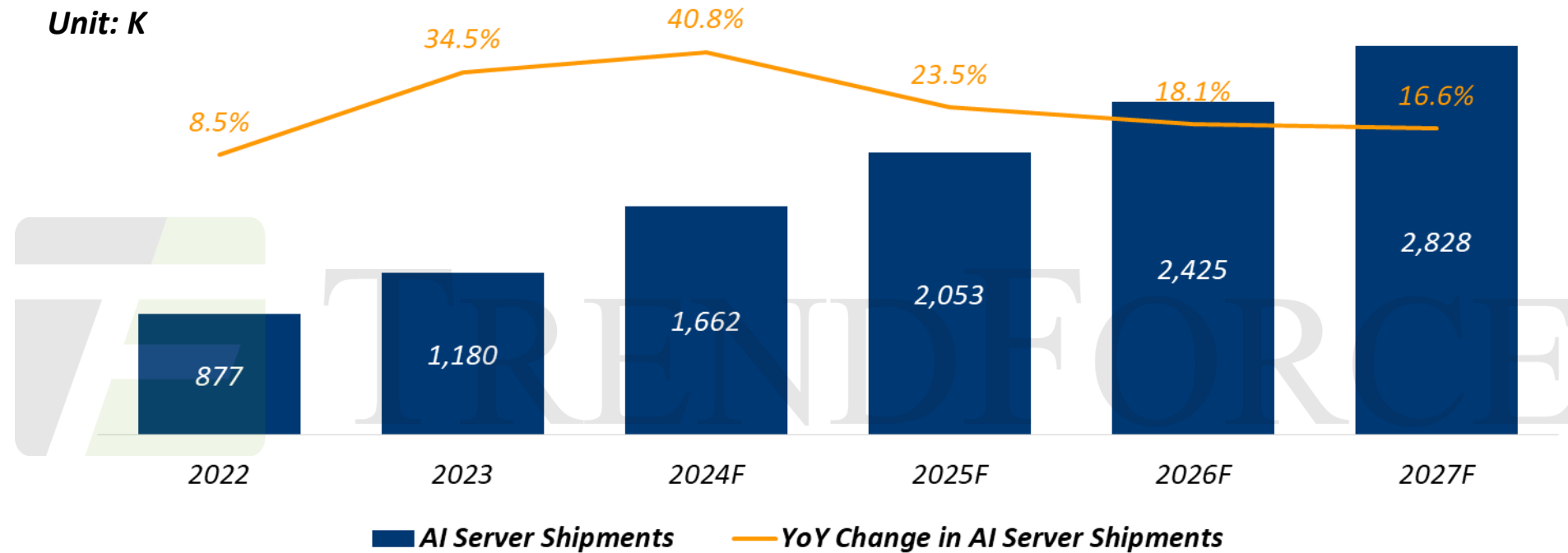


- Malaysia: Wiyynn has activated its server assembly plant in Johor since October 2023.
- Thailand: Quanta has set up shop at Chonburi, which is likely to become a key server production region second to Taiwan. Inventec is expected to complete the construction for its fab in Samut Prakan by the end of 2024, before commencing production since 1Q25.
- Vietnam: MiTAC's new fab in Hanoi is scheduled for completion and initiation of production in 2H24 at the earliest.

- ❑ Taiwan and US/Mexico are projected to each account for 35% of the server supply chain in 2027, while Southeast Asia would ascend to 30% then.
- ❑ ODMs are turning to the US, Mexico, and Southeast Asia under the tendency of "China+1", with concentration on Thailand, Malaysia, and Vietnam for Southeast Asia. Focus on capacity would be placed on SMT and L6 for the preliminary phase, followed by expansions in accordance with the demand for AI servers. CCL supplier Elite Material is expected to complete construction for its fab in Malaysia during 2025, while ITEQ could potentially begin pilot runs for its fab in Thailand since 2H24, and chassis supplier Chenbro is considering to establish corresponding fabs in Mexico, before gradually forming a more integrated supply chain cluster.



# 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.



# Continued Buzz Around AI Servers and the Unveiling of the Blackwell Structure

- GPU shortage eases compared to 2023, but CoWoS capacity continues to chase AI demand. TSMC's total CoWoS capacity is expected to increase by over 150% in 2024.
- NVIDIA remains the leading supplier of mainstream AI server chips, with shipments in 2024 expected to focus on the H series, while the B series is expected to be introduced in 2025.
- In 2025, NVIDIA is project to deliver the GB200 in a rack format to meet the needs of large data centers. The integration of communication and cooling systems will test full system assembly capabilities.
- Server cooling will become critical as data center power usage effectiveness standards grow increasingly stringent.

## Launch of Blackwell Architecture GPU Products

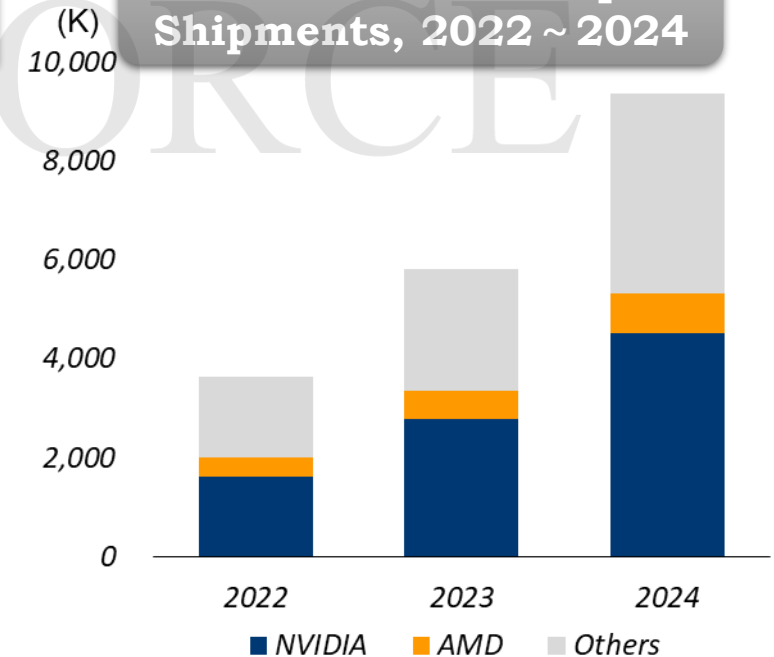


- NVIDIA announced the launch of its latest BLACK architecture GPU products—B100, B200, and GB200—at NVIDIA GTC 2024. Companies such as AWS, Google, Meta, Microsoft, OpenAI, Oracle, and Tesla immediately announced plans to adopt these products.
- Alongside the announcement of the B series, Taiwanese companies including Gigabyte, Inventec, ASUS, Wiyynn, and Pegatron simultaneously introduced servers and solutions that are expected to be implemented in data centers by 2025.

## Rack-Level Output Model

- NVIDIA also introduced the GB200 NVL72 server product for large-scale users, equipped with 36 CPUs and 72 Blackwell GPUs, along with an integrated liquid cooling solution.
- Additionally, NVIDIA unveiled its AI supercomputer DGX SuperPOD—centered around the DGX GB200 system. Each system is equipped with 36 GB200 chips, aimed at enabling AI computing at an ultra-large-scale.

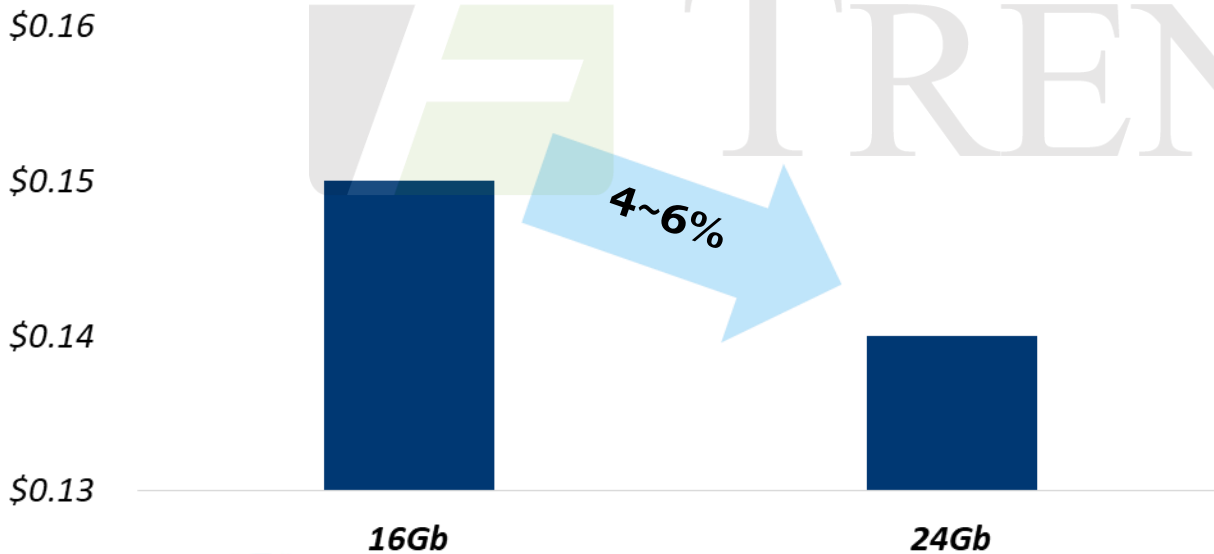
## Estimated AI Chip Shipments, 2022 ~ 2024



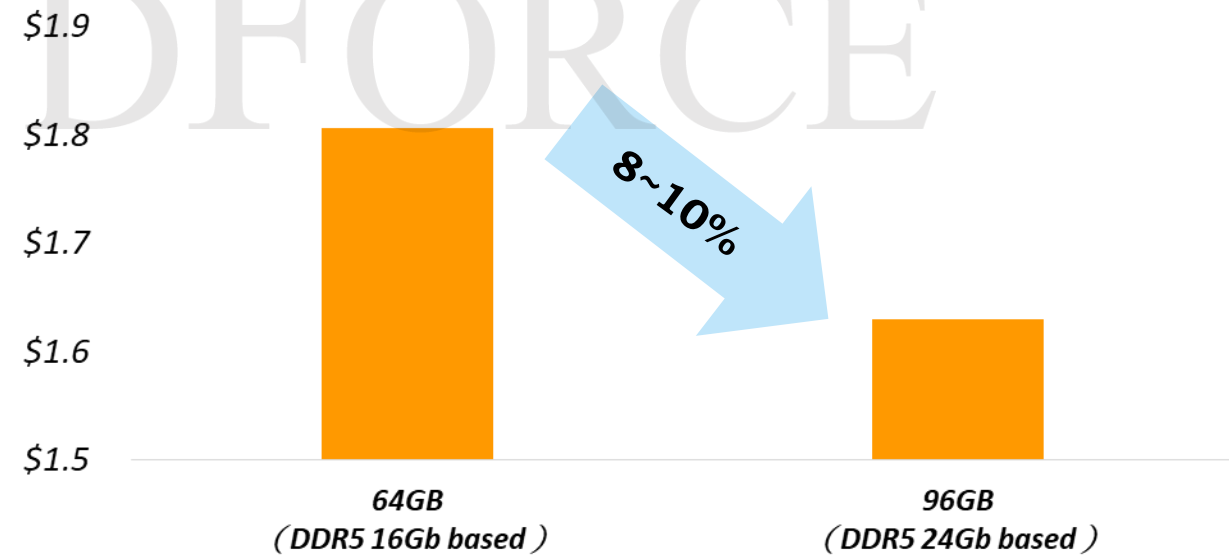
# 24Gb Mono Die Enjoys Cost Benefits over 16Gb on DDR5 Modules

- Despite 24Gb has lower yield compared to 16Gb for the time being, the cost improvement still occurs in both of chip and module.
- 96GB module is the most ideal solution for single socket CPU with high content DRAM requirement. TrendForce expects that suppliers will provide price incentive to justify CSPs' adoption.
- DDR5 24Gb products are exclusive for CSP. TrendForce forecasts that the adoption will be around 20% post mid-2024.

### Fully-loaded Manufacturing Cost Improvement per Gb (Chip)




### Fully-loaded Manufacturing Cost Improvement per GB (Module)



\*Content Production Yield Rate Assumption Based on 1alpha nm





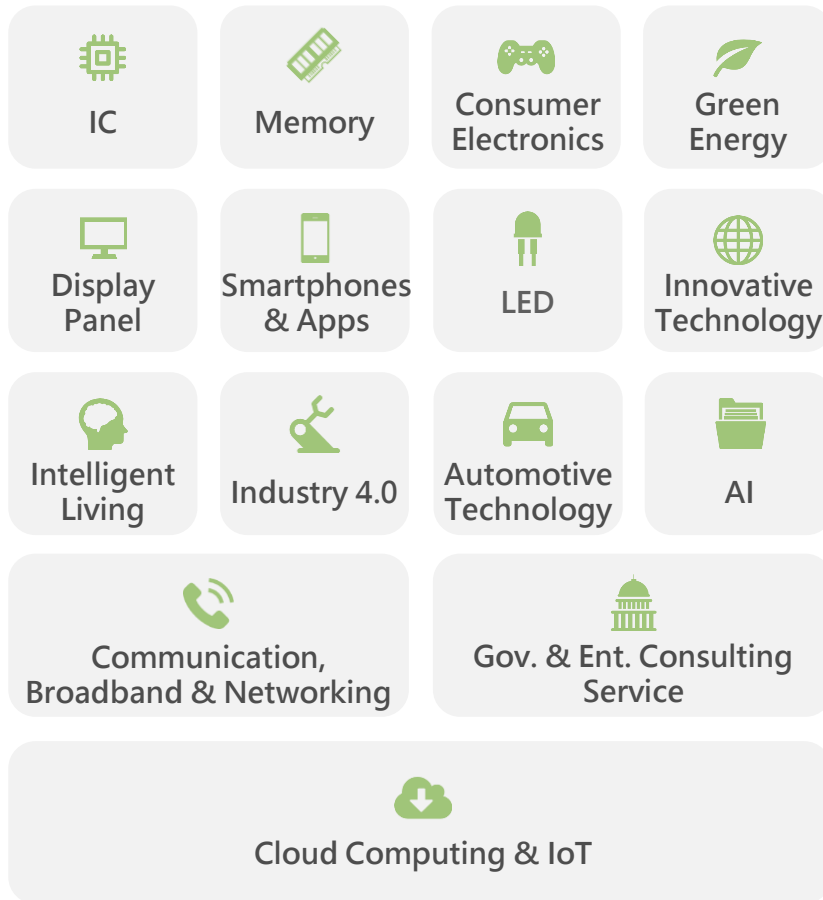
■ **The Ongoing Transformation of the Server Supply Chain** - In addition to Taiwan, the strategic deployment in Southeast Asia and other third-party locations has become a cornerstone of the new partner ecosystem.

■ **CSPs and Corporates are more Focused on Investment in Specific Fields** - In addition to recent investments in AIGC and other related areas, there will be a greater emphasis on themes such as ESG, which will lead to subsequent changes in server infrastructure.

**Server DRAM is shifting to high-capacity chips to meet AI needs**, improving cost structure and profitability. Due to process shrinking difficulties, manufacturers are focusing on high-capacity solutions.

**Next year, AI-related memory will dominate the market, with HBM production increasing to 10%.** Despite lower yield rates and larger chip sizes, manufacturers are expanding HBM capacity. TrendForce predicts DRAM price increases will continue into 2025.

## TrendForce & TRI Research Areas



## 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