AMERICAS TECHNOLOGY: HARDWARE

Traditional & accelerated servers: An industry primer

Katherine Murphy
+1(212)902 1151
katherine.a.murphy@gs.com
Goldman Sachs & Co. LLC

Michael Ng, CFA
+1(212)902 8618
michael.ng@gs.com
Goldman Sachs & Co. LLC

Zorayda Montemayor
+1(212)357 6403
zorayda.montemayor@gs.com
Goldman Sachs & Co. LLC

This report serves as a comprehensive primer on the server industry, designed for both seasoned technology investors and those new to the hardware sector. It provides a foundational analysis of the market's technological evolution over the last 70 years, current competitive dynamics, and the financial frameworks governing the industry. By grounding our analysis in historical trends and current unit economics, this report aims to equip investors with the necessary tools to navigate the near-term volatility and long-term shifts within the global server ecosystem.

Key areas of focus in this primer include:

- 2026-30 server market forecast: 650 Group forecasts the global server market to grow at a +38% 5-yr CAGR to $1.4 tr by 2030, driven by +13% growth in traditional servers ($164 bn in 2030) and +45% growth in accelerated servers ($1.2 tr in 2030). Both the traditional server and accelerated server markets should see double digit growth across all customer verticals, with strength in hyperscaler & neocloud driven by continuous AI infrastructure build outs, while enterprise demand should benefit from data center modernization efforts in support of agentic AI adoption & broader technology refresh. Also includes a summary of the role of a CPU and GPU within AI infrastructure deployments.
- Market structure and vendor dynamics: An examination of the server vendor landscape, where market share varies meaningfully based on the customer vertical. Enterprise market is highly concentrated and dominated by branded OEMs, including Dell, HPE, Lenovo, and IBM, while the hyperscale opportunity is primarily addressed by white box vendors; neoclouds currently use a mix of both OEMs and ODMs.
- Unit economics and impacts of changing commodity costs: A detailed breakdown of traditional and AI server bill of materials (BOM). We estimate that server OEMs typical earn 15% gross margins and 5-7% operating margins on a CPU server sale. DRAM & NAND historically represented ~30% of the bill of materials for a traditional server, but under current industry expectations price increases in 2026/27, memory could represent >60% of the bill of materials for a traditional server by the end of calendar 2026, assuming no offsetting actions.

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.

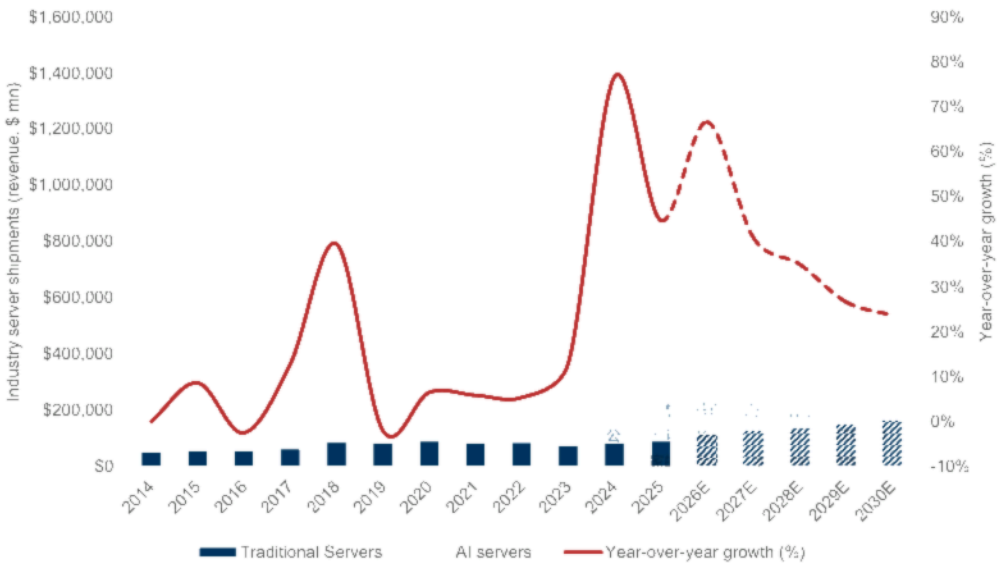
Table of Contents

Defining the server market

Servers are specialized computers that run application-specific services for clients. While PCs & servers share similar architecture, servers are typically designed with more CPU/GPU sockets, memory, and redundancies around power supplies/network interfaces to ensure availability for mission-critical workloads. In the enterprise, servers are used to address the edge or front end of the network for applications such as collaborative business software, web serving, file serving, and email. Cloud providers deploy servers in data centers in massive quantities (on the magnitude of hundreds of thousands) to run applications for cloud computing, search, video delivery, social networking, and more recently - generative AI workloads.

The total server market was ~$284 bn in annual revenue in 2025, up +45% yoy driven by robust growth in AI server demand & higher ASPs. All server market estimates (both market share & forecasts) in this primer are based on 650 Group data unless otherwise noted. 650 Group expects the overall server market to grow at a 38% CAGR over the next five years to ~***$1.4 tr 2030,*** driven by growth driver by AI server demands. While the server market has historically been driven by new technology adoption & infrastructure refresh cycles, the robust AI driven growth realized in 2023 and 2024 marks a significant inflection from historical growth trends. As such, for the purposes of this report, **we examine the traditional server (i.e., general purpose, non-accelerated servers) markets and accelerated server (i.e., high performance server with added GPUs or other custom accelerators) separately.**

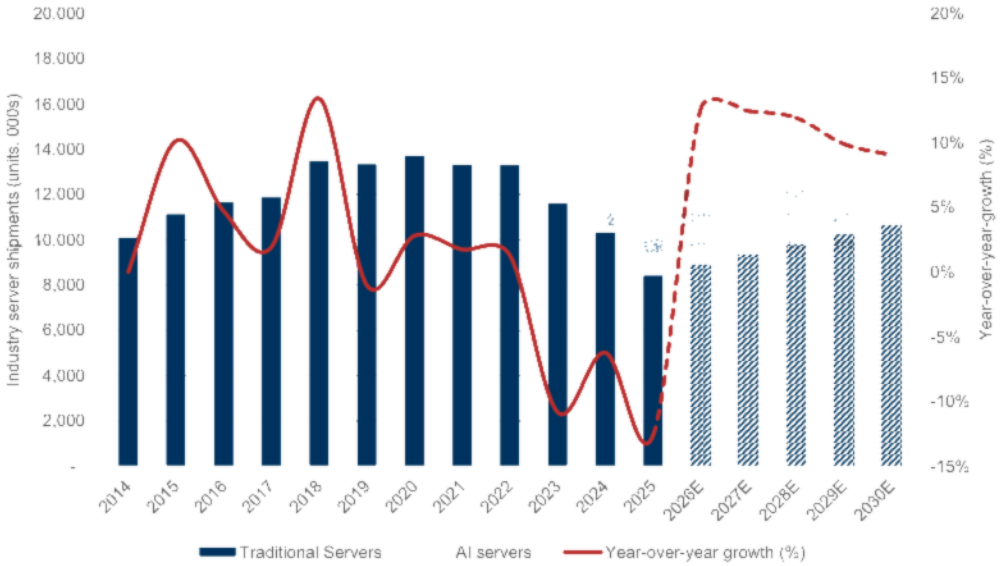
Exhibit 1: The total server market was ~$284 bn in annual revenue in 2025, up +45% yoy driven by robust growth in AI server demand & higher ASPs
Industry server revenue ($, mn) and growth (%)



Source: 650 Group

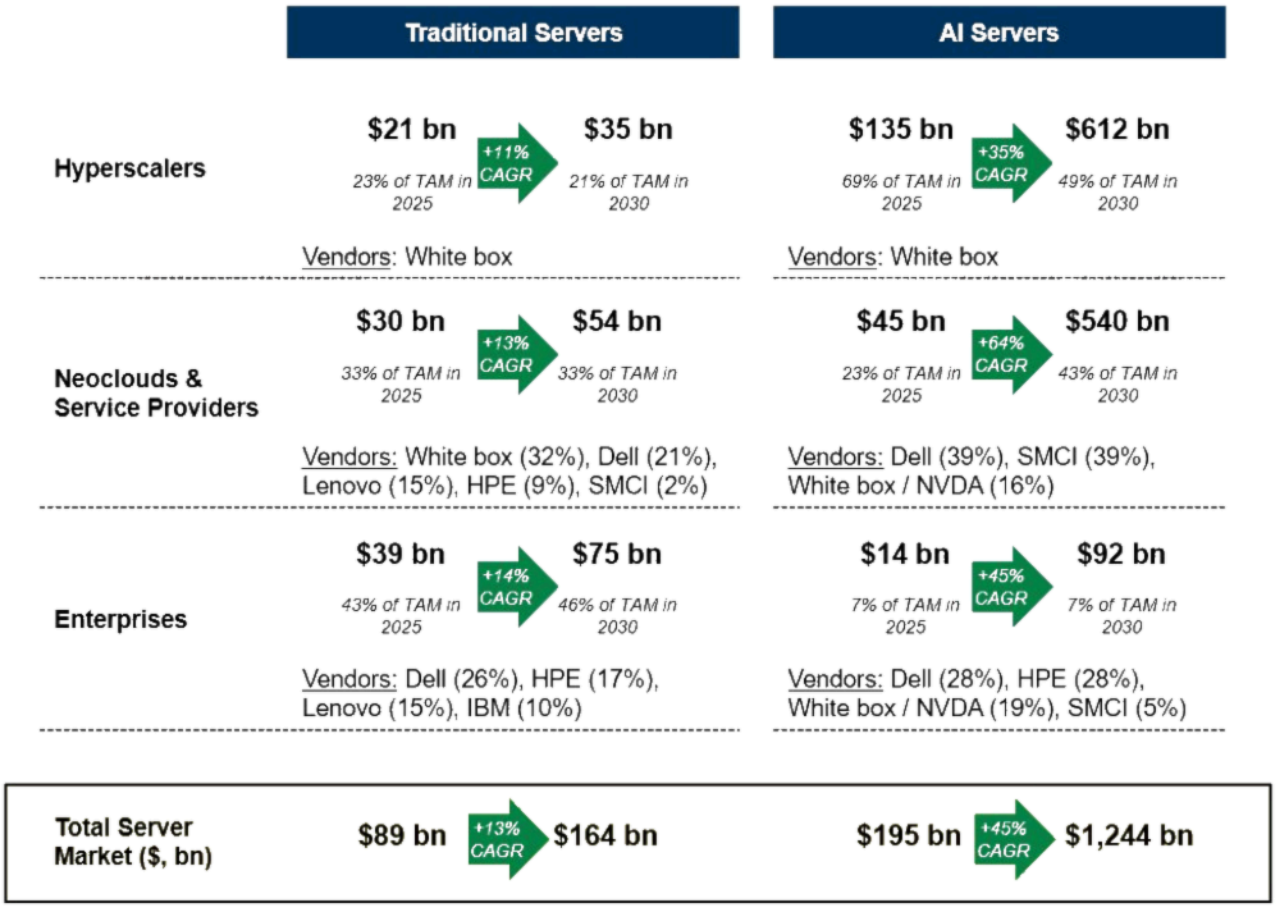
Exhibit 2: Industry server units were down 13% yoy in 2025 driven by AI server demand that was more than offset by a decline in traditional servers; both traditional and AI server units are expected to grow in 2026

Industry server shipments (units, 000s) and growth (%)



Source: 650 Group

Exhibit 3: Vendor market share differs significantly by vertical & product (AI v. traditional)
Server segment summary



Source: 650 Group

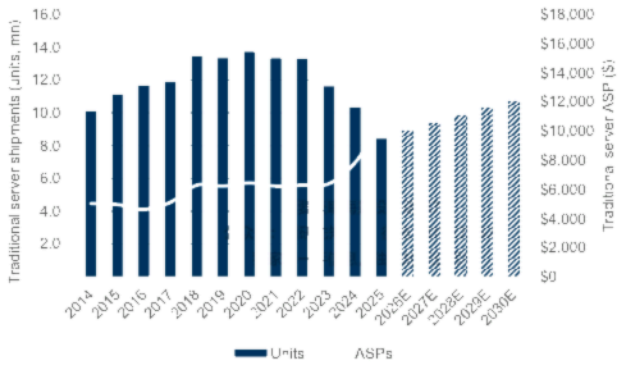
Traditional Servers

The traditional server market was ~$89 bn in annual revenue in 2025 (+10% yoy) with approximately 8.4 mn units sold (-18%) at an average selling price of ~$10,500 (+35% yoy), based on 650 Group. Traditional servers, also referred to as general purpose servers, are deployed across various environments ranging from on-premise server rooms to data centers, often used for more general purpose workloads like web hosting, databases management, file sharing, or application servers (v. specialized workloads like AI/ML). For AI inference and training, non-accelerated servers are used for orchestration and control plane layers. A traditional server includes a CPU, memory, storage, etc. Unlike AI servers, traditional servers do not have an embedded accelerator card (i.e., GPU).

Exhibit 4: The traditional server market was ~$89 bn in annual revenue in 2025 (+10% yoy)...
Industry traditional server revenue ($, mn) and annual growth rate (%)

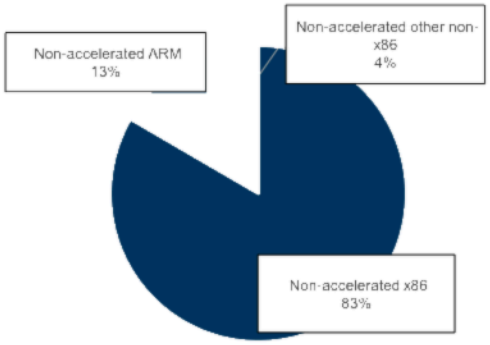
Source: 650 Group

Exhibit 5: ... with approximately 8.4 mn units sold at an average selling price of ~$10,500
Traditional server shipments (units, mn) and ASP ($)



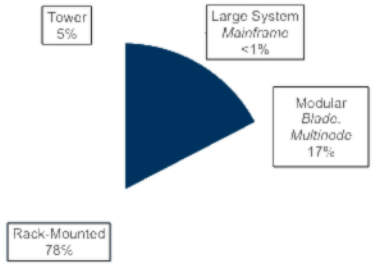
Source: 650 Group

Exhibit 6: Industry standard servers are based on the x86 CPU architecture
2024 traditional (non-accelerated) server shipments by CPU architecture



Source: IDC, Data compiled by Goldman Sachs Global Investment Research

Exhibit 7: The majority of traditional servers are rack-mounted or modular
2024 traditional (non-accelerated) server shipments by form factor

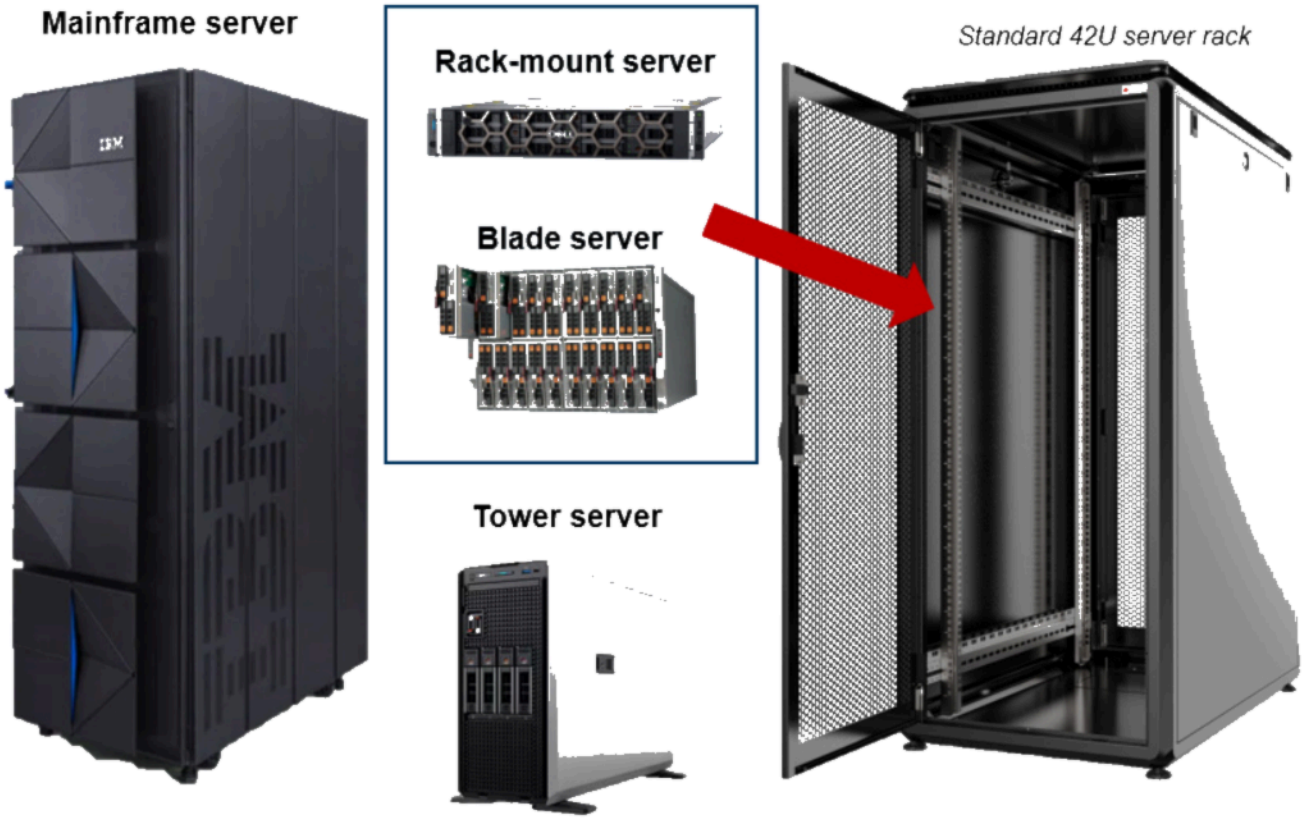


Source: IDC

The earliest servers introduced in the 1950s were mainframe computers, which are high-performance, highly reliable computing systems. Mainframes employ a centralized architecture with large processors and memory capacity, and most importantly, with multiple layers of redundancy which helps guarantee high availability. Although mainframe only represents a minor share of server volumes by unit shipments, mainframes represents a significant share of mission critical workloads for industries that require continuous operations like financial institutions, healthcare companies, and government agencies (IBM reports that 90% of the top 50 banks run on IBM Z, its Mainframe solution). Rack-mounted servers were introduced in the 1990s to address the growing demands of internet infrastructure. Rack-mounted servers are built in standardized units (called rack units ("U"), each 1.75" tall and 19" tall) and installed in a chassis or rack (standard server racks are 42U tall). This standardization allows for increased efficiency, scalability, and easier management; rack-mount servers are typically the more cost-effective compute solutions. Servers come in a variety of form factors but standard CPU servers are typically 1U or 2U, which means up to ~20-40 servers fit on each rack when leaving space for other appliances like power, cooling, networking, and storage. Each rack-mount servers typically contains all of the

components needed to operate as a standalone system (e.g., compute, memory, power, storage, etc), as compared to blade servers that are designed in a more modular style to increase density & therefore processing power. Each blade includes critical components for a server like a CPU and memory & is installed within a specialized chassis which provides shared infrastructure like networking, power, and cooling. Blade server chassis are designed to fit within industry standard server racks. Similarly, multi-node servers are similar to blade servers in their modular design & share resources, except they do not share network fabrics. Tower servers are standalone upright server cabinets - similar in form factor to a traditional desktop computer- but with upgraded components (e.g., compute, memory), typically used by SMBs.

Exhibit 8: Servers are available in several form factors including mainframe, rack-mounted, blade, and tower



Source: Company data, Goldman Sachs Global Investment Research

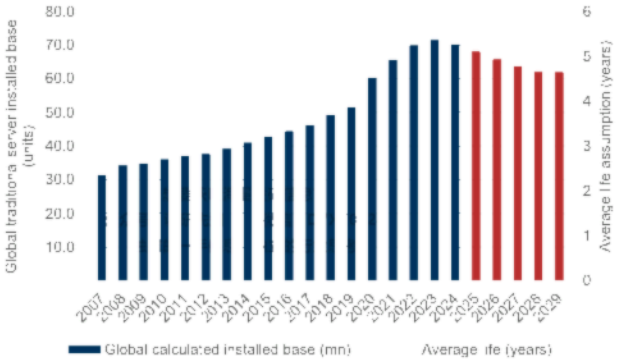
Industry standard servers are based on the x86 CPU architecture, which is the most commonly deployed CPU architecture across network infrastructure, data centers, and cloud providers, representing ~85% of units deployed in 2025, per Gartner. x86 architectures (e.g., Intel Xeon, AMD Epyc) support a wide range of instructions & is compatible with a wide range of software and operating systems. ARM (15% and growing) architectures are designed to compute simpler instructions, which results in a more power efficient solutions that are more cost-effective. Custom silicon designed by hyperscalers (e.g., AWS Graviton, Google Axion) & Nvidia's Grace CPU use ARM-based architectures. Because different CPU architectures have unique instruction sets, software complied for one architecture cannot typically run on another architecture without recompiling, which can be a significant lift for enterprise organizations. Other

CPU architectures include RISC, CISC, and EPIC. GS Semiconductor Analyst James Schneider expects that the x86 instruction set will maintain a majority share of the server CPU market, particularly in enterprise, with some share gains for ARM server CPUs, especially at hypercscalers who are building custom hardware focused on lowering power consumption.

We estimate that there are ~70 mn traditional servers in the global installed base, with an average life of 5.7 years. We estimate the server installed base by assuming an average life for each server in each year (based on Lawrence Berkeley National Lab estimates) and then summing the server shipments for the corresponding number of years prior. Our server shipments estimates are based on a blended average of IDC, 650 Group, and Gartner traditional server forecast (i.e. non accelerated).

- We estimate that average life of servers has been extended from 4.4 years (2003-2019) to 5.7 years today, driven by (1) hyperscalers extending the useful life/depreciation for their servers from 3 years to 5-6 years over the last few years, citing efficiency gains from improved processes and maintenance; and (2) enterprises running their equipment "hot" (i.e., at higher than desirable utilization rates) in a time of increased IT budget pressure from an uncertain macro environment & competing demands (i.e., AI). As a result, we estimate that 32% of the global server installed base in 5+ years old as of 2025, up meaningfully from 8% a decade ago. Note that this portion of the installed base (2019-2020) was put into data centers prior to many of the requirements for AI-infrastructure, which requires significantly more power that traditional servers. DELL reported on their F1Q27 (3 months ending April 30, 2026) earnings that the majority of their installed base of traditional servers were 14th generation, which was first introduced in 2017, or older.

Exhibit 9: We estimate that there are ~70 mn traditional servers in the global installed base, with an average life of 5.7 years.
Global traditional server installed base (units, mn) and average usable life assumption (years)



Source: Goldman Sachs Global Investment Research, Lawrence Berkeley National Lab

Exhibit 10: We estimate that 32% of the global server installed base is 5+ years old as of 2025, up meaningfully from 8% a decade ago...
Global server installed by age (units, mn) and % of servers aged 5 years and older

Source: Goldman Sachs Global Investment Research, Lawrence Berkeley National Lab

Exhibit 11: ... driven in large part by extended useful lives of servers at major hyperscalers
Useful life of servers (for accounting/depreciation purposes) at major US hyperscalers

Source: Company data, Goldman Sachs Global Investment Research

We expect traditional server shipments should normalize to ~10 million units annually (v. 13 mn average in 2018-22) as (1) refresh to newer systems increase efficiency (i.e., densification) and (2) agentic AI drives demand drives expanded server estates.

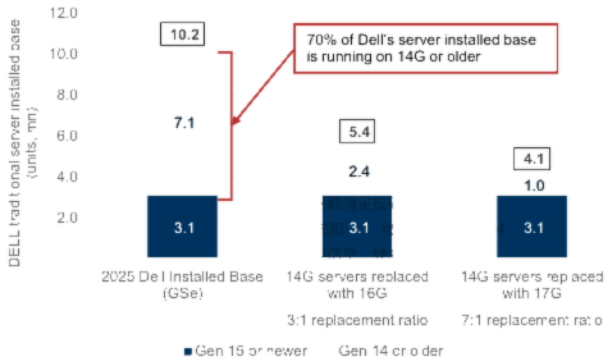
- ***How will server consolidation impact the installed base over the long term?*** HPE has reported that a single Gen12 server can replace 7 Gen11 servers or 14 Gen10 servers while reducing power by 65%. DELL has reported that its 17G consolidates old servers at a 3:1 ratio (15G) or 7:1 ratio (14G), with over 70% of its installed base is still running on 14G servers or older (as of November 2025). In a simplified scenario analysis where all legacy servers are replaced with next-gen, assuming a 3:1 or 6:1 replacement rate on 14G older servers, Dell could shrink its server installed base by up to 60%. We note that this does not account for any growth in workloads, which would necessitate incremental compute capacity being brought online (a partial offset to this consolidation headwind). That said, the consolidation of traditional servers today may lead to a structurally smaller installed base of servers in the medium term, which could limit potential upside in future refresh cycles. If significantly fewer server units ship each year, ASPs & profit margins need to be up significantly to maintain a stable profit pool.

Exhibit 12: DELL has reported that its 17G converts old 14G servers at a 6:1 ratio
Illustrative comparison of 14G and 17G Dell servers specs

	PowerEdge R540	PowerEdge R470	17G v. 14G
Generation	14th Generation	17th Generation	
Year Released	2017	2024	
Form Factor	1U	1U	
Processor	Intel Xeon Silver	Intel Xeon 6 E	
# of Cores	8	64	8X
Memory	16 GB DDR4	32 GB DDR5	2X
Storage	2 TB SATA HDD	960 GB SSD	
Input Power	143 watts	316 watts	2X
Starting Price	$4,677	$10,000	2.1X

Source: Company data, Goldman Sachs Global Investment Research

Exhibit 13: Assuming a 3:1 or 7:1 replacement rate on 14G older servers, Dell could shrink its server installed base by up to 60% in our illustrative scenario analysis
Dell traditional server installed base scenario analysis



Source: Company data, Goldman Sachs Global Investment Research

Exhibit 14: If significantly fewer server units ship each year, ASPs & profit margins need to be up significantly to maintain a stable profit pool
Industry traditional server installed based scenario analysis

	Current industry (2020-25 avg)	3:1 replacement ratio	7:1 replacement ratio
Annual unit shipments (mn)	11.9	4.0	1.7
		Assume 3:1	*Assume 7:1*
ASPs ($)	$8,036	$14,063	$20,090
		Assume +75%	*Assume 150%*
Total Industry Value ($, bn)	**$96**	**$56**	**$34**
EBIT margins (GSe)	7%	12%	20%
	GS estimate	*Implied if profit pool unchanged*	
Industry profit pool	**$6.7**	**$6.7**	**$6.7**

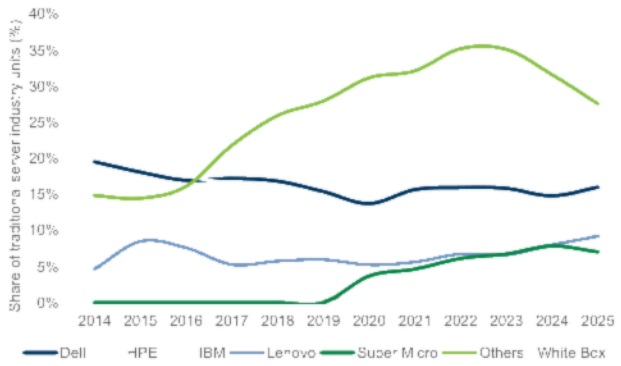
Source: Goldman Sachs Global Investment Research, IDC, Company data

- ***What is the role of traditional servers in AI deployments?*** While GPU servers represent the majority of net-new AI infrastructure spend to date, traditional CPU also play a critical role in AI deployments. First, in massive GPU clusters designed for AI training, CPU servers serve as the head nodes/orchestration layers. While GPUs handle the heavy lifting of parallel processing required for model training (i.e., worker nodes), CPU servers manage the cluster, schedule jobs, route data, and monitor resource allocation. Without traditional servers orchestrating the workflow, the high-performance GPUs would be unable to function efficiently, making CPU infrastructure a key component of any training cluster. Second, agentic AI requires a hybrid approach that blends the probabilistic reasoning of GPUs (statistical pattern recognition and generating tokens) with the deterministic execution of CPUs (rule-based logic, API tool calls, and strict workflow orchestration). Because agentic workflows are heavily tool-dominated and require constant, predictable orchestration, CPUs are better suited to handle these deterministic tasks to prevent latency bottlenecks. This architectural shift will have an impact on hardware configurations; while traditional AI training typically sees a CPU-to-GPU attach rate of 1:4 to 1:8, Intel notes that the demands of agentic inference are driving this attach rate closer to 1:2 or even 1:1. Agentic AI adoption, particularly at the enterprise, is still in the early days and expectations for the timing/magnitude of required

investments in CPU infrastructure remains fluid.

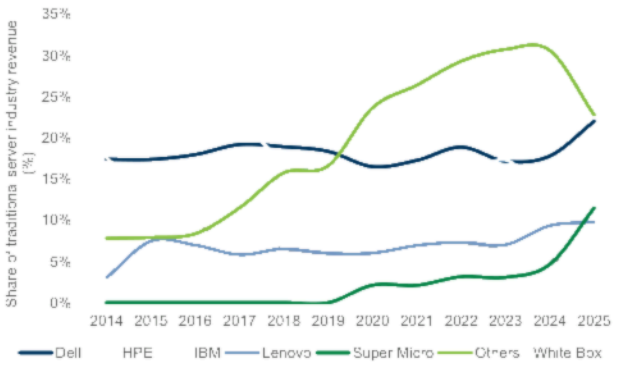
Traditional server vendor market share has shifted from branded OEMs towards white-box as hyperscaler demand grows. Over the past decade, market share as measured by units for the top 5 traditional server vendors (Dell, HPE, IBM, SMCI and Lenovo) had declined from a cumulative 53% in 2014 to 42% in 2025, per 650 Group. We also highlight similar trends in revenue share declining from 63% in 2014 to 57% in 2025. This trend has been driven by shift away from traditional OEMs towards non-branded "white box" units, which accounted for ~23% of traditional server units and ~28% of traditional server revenue in 2025, up from 15% and 8% in 2014, respectively.

Exhibit 15: Over the past decade, market share as measured by units for the top 5 traditional server vendors (Dell, HPE, IBM, SMCI and Lenovo) had declined from a cumulative 53% in 2014 to 42% in 2025
Share of traditional server industry unit shipments by vendor (%)



Source: 650 Group

Exhibit 16: Similarly, revenue share from the top 5 vendors has declined from 63% in 2014 to 57% in 2025
Share of traditional server industry revenue by vendor (%)



Source: 650 Group

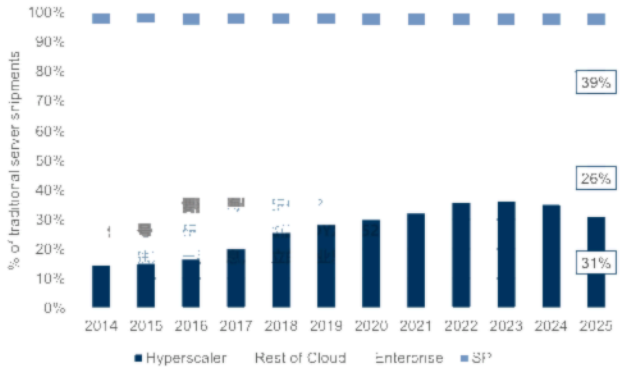
The major driver of this shift is a change in underlying demand across customer verticals, with hyperscale customers (highly indexed to white box) growing traditional server revenue at an 18% 10-year CAGR (2014-24) v. enterprise traditional server revenue down on a -2% 10-year CAGR, primarily driven by the shift of workloads to the public cloud. As of 2025, 650 Group estimates that 72% of the hyperscaler traditional server market was addressed by ODMs/white box providers. Comparatively, 98% of the enterprise server market was addressed by branded OEMs, with the top 5 vendors representing 71% of total revenue.

White-box servers are unbranded solutions, typically assembled with off-the-shelf, commodity components by ODMs. They are typically sold at lower prices with more customization opportunities for customers (customers can specify hardware needs (i.e., memory, CPU, etc) and operating systems). Hyperscalers and other cloud providers operate at a larger scale than traditional enterprises, with server architectures growing into the millions. As compared to traditional enterprise, hyperscalers are focused on maximizing (1) value, with little desire to pay for the extra features on servers beyond what is essential; and minimizing (2) cost, since they pay very little for maintenance and support and as any servers that fail are discarded while workloads get re-allocated to other servers in their scale out data center. ***See 'IT hardware manufacturing industry primer: OEMs, ODMs, CMs' sections in our IT hardware distribution & manufacturing industry primer for more on server manufacturing.***

Traditional server OEMs compete on the basis of pricing, brand recognition, distribution models, and value-added services; the standalone products are often not technically differentiated enough to result in market share gains. For traditional enterprise deployments, reliability is paramount & systems are built to avoid hardware failures at all costs. As a result, enterprises are more willing to pay for built-in redundancies and expensive maintenance contracts:

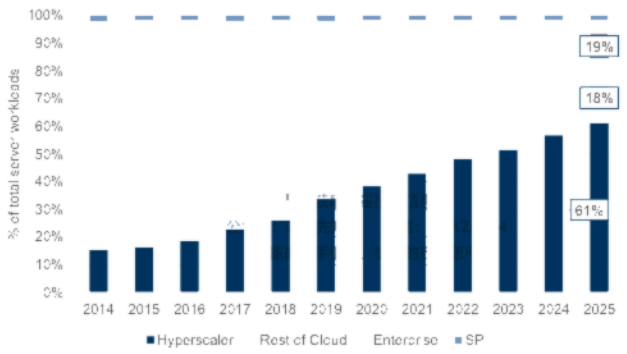
- First, enterprise hardware is designed to be easily serviceable with sensors embedded into appliances to collect critical metrics like temperature, drive performance, and power usage in order to provide visibility into server health & potential failures. Server OEMs like DELL, HPE, SMCI also offer "out-of-band" management technologies within their server appliances that allow administrators to remotely monitor and control servers independently of the CPU. While SMCI's Intelligent Management platform is based on Intel's open-source IPMI (Intelligent Platform Management Interface), both DELL (integrated Dell remote access controller, or iDRAC) and HPE (Integrated lights-out, iLO) have invested in building proprietary embedded server management tools with more advanced features and capabilities. A Principled Technologies study in April 2024 found that DELL's sever management tools has significantly more management & security features than SMCI's equivalent platforms.
- Second, OEMs - especially DELL and HPE - offer extended warranties and enhanced support and maintenance services on their enterprise hardware. For example, Dell's ProSupport and HPE's Tech Care programs extend upon standard hardware warranties to include software configuration, proactive monitoring & issue detection (enabled by hardware enhancements and management tools like iDRAC and iLO noted above), and up to 24/7 technical support online & via phone, with many plans including on-site support as well. To drive attach of maintenance services to hardware sales, Dell's sales force typically start quotes to enterprise customers with a 3-year mission critical ProSupport contract included. In a simple example, attaching a $1,300 ProSupport mission critical contract to a $7,000 Dell PowerEdge R470 server sale can add ~4 percentage points to the gross margins of a total sever sale (assuming hardware is ~15% gross margins & services gross margins are ~40%, in line with DELL's reported F2025 disclosures).
- Third, DELL & HPE also offer professional services, managed services, education, and financial services to enterprise customers as well. For example, for an additional consulting-type fee, DELL offers solution-level design & implementation services (a type of professional service) for the installation of hardware across the entirety of a data center project, including other OEMs hardware. This is beyond the scope of DELL's typical ProDeploy contracts, which generally just cover installation for the particular hardware appliance it is attached to without considerations for the best-practices or sequencing of a data center installation (e.g., servers first, networking second, etc). Our channel checks show that Dell's program level services, a legacy of the support model at EMC, are regarded as best-in-class & a point of differentiation compared to other hardware OEMs. Similarly, HPE offers a comprehensive portfolio of advisory and professional services across its range of compute, storage, and networking appliances.

Exhibit 17: Hyperscalers account for 31% of traditional server shipments as of 2025...
Share of industry traditional server shipments by customer vertical



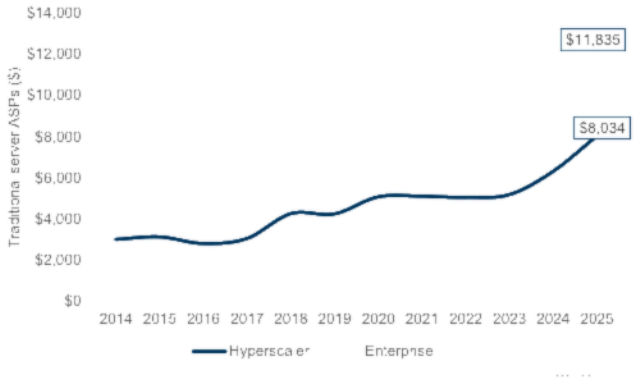
Source: 650 Group

Exhibit 18: ... and 61% of server workloads
Share of industry server workloads by customer vertical



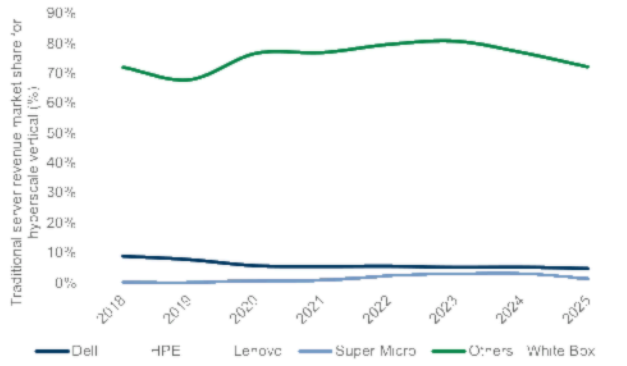
Source: 650 Group

Exhibit 19: Hyperscaler traditional server ASPs are ~30% cheaper than enterprise ASPs in 2025
Traditional server ASP by customer vertical ($)



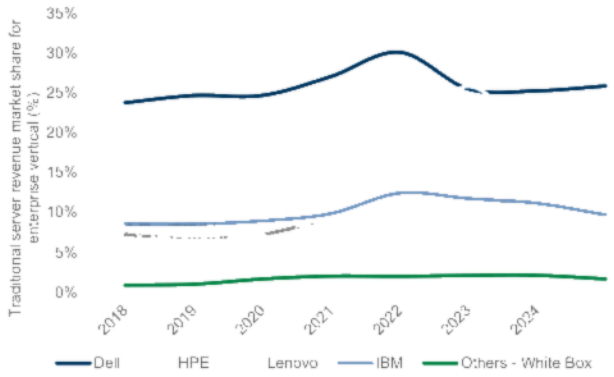
Source: 650 Group

Exhibit 20: 650 Group estimates that 72% of the hyperscaler traditional server market was addressed by ODMs/white box providers
Traditional server revenue market share - hyperscaler customer vertical (%)



Source: 650 Grpup

Exhibit 21: As compared 98% of the enterprise server market addressed by branded OEMs
Traditional server revenue market share - enterprise customer vertical (%)



Source: 650 Group

Traditional server unit economics

Branded vendors typically earn ~15-20% gross margins and ~6-8% operating margins on server hardware sales. Below we present an illustrative bill of materials for an x86 server, with component pricing as of July 2025 to reflect a bill of materials that is unimpacted by the spike in DRAM/NAND prices that started in fall 2025. In this illustrative example, components represent ~90% of the total cost of product, while server brand markups & ODM manufacturing fees represent ~10%. Under current pricing expectations for 2026/27 (see 'How do higher commodity costs impact the server market?' below for more details), DRAM/NAND content in some traditional server configurations can represent >60% of a server BOM. As noted above, we estimate that attaching support services and maintenance contracts to server sales can add 4+ percentage points of gross margin for branded vendors.

Exhibit 22: Traditional server bill of materials
Illustrative bill of materials for a CPU general server

Dual CPU general server

Component	Details	Estimated cost	Share of total BOM
CPU + Graphics	2x Intel Xeon server	$1,850	28%
RAM	1TB	$267	4%
Storage	12 x 512 GB SSD	$1,800	27%
Networking	SmartNIC	$650	10%
Cooling	8x fans, heat sinks	$140	2%
Power	2x 1kw per server	$334	5%
Other	Mainboard, cables, chassis, rail kit	$787	12%
Component costs		**$5,828**	**87%**
ODM Manufacturing/testing fees		$100	1%
Server brand mark up		$750	11%
Labor, freight, warranty,OS		**$850**	**13%**
Total cost of product:		**$6,678**	**100%**

Average selling price:	*OEM gross margins:*
$8,000	*17%*

Component pricing as of July 2025 to reflect a bill of materials that is unimpacted by the spike in DRAM/NAND prices

Source: Goldman Sachs Global Investment Research

Accelerated Servers

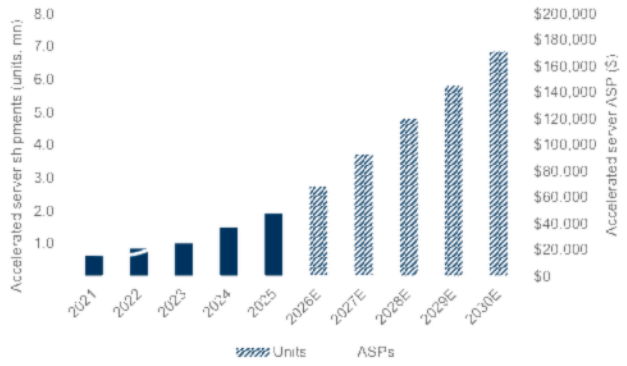
GPUs are well suited for 3D graphics rendering because they can simultaneously process repetitive calculations necessary for complex visuals, like transforming shapes, applying lighting, coloring pixels.

The accelerated server market was ~$195 bn in annual revenue in 2025 (+69% yoy) with approximately 1.9 mn units sold (+27%) at an average selling price of ~$100,000 (+33% yoy), based on 650 Group estimates. Accelerated servers, including those used for AI workloads (AI servers), are equipped with specialized hardware accelerators such as a GPU or custom ASIC[1] that work in tandem with a CPU to speed up data-intensive, complex workloads (collectively referred to as "XPUs"). While CPUs are architected for sequential, serial instruction processing across its cores (Intel's Granite Rapids server CPU has 128 cores), GPUs are designed to process multiple computations simultaneously (i.e., in parallel) across its thousands of cores (Nvidia's B200 GPU has 20,480 CUDA cores and 640tensor cores). As a result, parallel processing is more efficient at handling tasks that can be broken down into smaller, independent parts like analytics on large data sets, scientific simulation, or graphics rendering. Accelerators are not new to server architectures; Nvidia debuted its GeForce 256 GPU for rendering 3D video game graphics on PCs over 25 years ago. Over the last two decades, GPUs and other custom ASICs have been increasingly deployed for a widening range of workloads that also benefit from parallel processing, referred to as GPGPU or general purpose GPU. For the purposes of this report, we define the accelerated server market to include all server units that have an accelerator card. This includes servers used for AI, as well as other machine learning and high performance compute (HPC) use cases.

Exhibit 23: The accelerated server market was ~$195 bn in annual revenue in 2025 (+69% yoy)...
Industry traditional server revenue ($, mn) and annual growth rate (%)

Source: 650 Group

Exhibit 24: ... with approximately 1.9 mn units sold (+27%) at an average selling price of ~$100,000 (+33% yoy)
Traditional server shipments (units, mn) and ASP ($)



Source: 650 Group

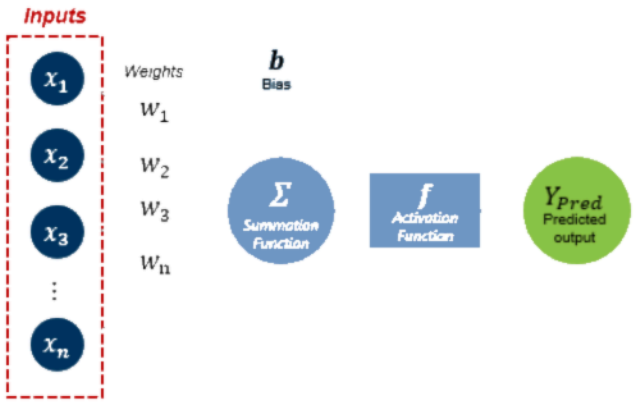
Transformer models, including LLMs like Chat-GPT, are a type of neural network.

AI training requires a sizable investment in accelerated compute. Training AI models involves processing massive amounts of data & conducting billions of mathematical operations to learn patterns and make predictions. In a simple example, a neural network is a basic computational model that is designed to recognize patterns and learn from data by finding connections between different interconnected nodes, or neurons. Within a neural network, the input data (such as an image, text, or numerical figure represented by a matrix) is multiplied against a matrix of weights. These weights determine the specific patterns or features the neural network is looking to detect. The

[1] Application Specific Ingregated Circuit; examples includes the TPU (Tensor Processing Unit) and DPU (Data Processing Unit)

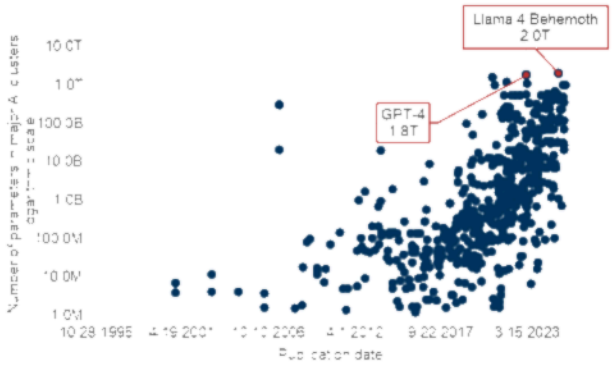
result of this matrix multiplication, often followed by an activation function, determines the activation of neurons in the next layer. As data passes through layers, the network can extract increasingly abstract and complex features. AI models "learn" by updating the network's weights to minimize the difference between predictions and outcomes. Critical for this report, **matrix multiplication is a highly parallelizable operation (many smaller, independent calculations happening simultaneously), which is very well suited for GPUs.** Researchers found that matrix multiplication is used in 90%+ of compute tasks in an LLM. As a result, accelerated server demand has inflected meaningfully in 2023 following the public launch of ChatGPT in November 2022 and continues to grow alongside increased investment by model builders like OpenAI (ChatGPT), Google (Gemini), Anthropic (Claude), and Meta (Llama). The massive size of AI models (Llama 4 Behemoth is reported to have 2.0 trillion parameters) makes accelerated servers not just an advantage but a necessity to train models in a feasible amount of time. Models that would take years to train with CPU compute can be trained in weeks to months with GPUs.

Exhibit 25: A neural network is a basic computational model that is designed to recognize patterns and learn from data by finding connections between different interconnected nodes, or neurons
Illustrative neural network architecture



Source: Goldman Sachs Global Investment Research

Exhibit 26: The massive size of AI models (Llama 4 is reported to have anywhere between 2.0 trillion parameters) makes accelerated servers not just an advantage but a necessity to train models in a feasible amount of time
Parameters in notable artificial intelligence systems



Parameters are variables in an AI system whose values are adjusted during training to establish how input data gets transformed into the desired output; for example, the connection weights in an artificial neural network

Source: Epoch, Goldman Sachs Global Investment Research

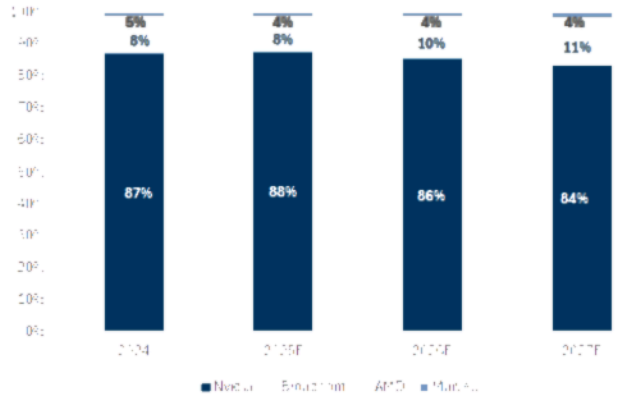
Training AI models involves iterative adjustemnts of model parameters (e.g., back propagation) while AI inference primarily involves a "forward pass" through the neural network

Beyond model training, accelerated compute also plays a role in inference and RAG. First, AI inference is the process of using a trained AI model to make predictions on new data. Although requiring less raw computational power than training an AI model, AI inference still benefits from the parallel architectures of an accelerated server to perform matrix multiplication. During inference, the pre-trained model is fed new input data, which then passes through each layer of its neural network. At each layer, the pre-determined weights of the model are multiplied against the new data (i.e., matrix multiplication), moving through each layer of the network ("forward pass"), which combined with bias addition and activation functions, leads to the output or prediction. Second, RAG or Retrieval Augmented Generation, is the technique of updating or improving on a static AI model with external knowledge basis like a database or web source. RAG techniques improve the accuracy of an LLM by adding a retrieval step (converting a query into a numerical representation or " embedding", then searching for

similar embeddings in a vector database) before the generation step. Generating these vector embeddings, indexing and searching vector data bases, and re-ranking retrieved results can be performed more efficiently with parallel GPU compute than with a CPU. Similar to AI training, AI inference and RAG perform better on accelerated compute hardware, but an XPU is not explicitly required. Nscale found that the Mistral 7B Instruct model can generate 63 tokens/second on a GPU but only ~18 tokens/second when run on CPU.

Most XPUs are sold by Nvidia today, with more custom accelerator adoption expected over time. GS Semiconductor analyst Jim Schneider expects that merchant AI accelerators (e.g., Nvidia GPUs) will retain the majority share of the AI infrastructure market for the foreseeable future because: (1) AI model development prioritizes flexibility and optimization of model training; (2) AI developers are focused on high-performance solutions; and (3) a more mature developer/software ecosystem for merchant silicon. That said, Jim sees growing adoption of custom ASICs to address internal workloads (e.g., Google for Search, Amazon for e-commerce) in the near term and more external workloads, especially inference, in the long run.

Exhibit 27: GS Semiconductor analyst Jim Schneider expects NVDA to maintain the majority revenue share of the AI accelerator market
AI chip designers accelerated compute revenue (% of total)



Source: Company data, Goldman Sachs Global Investment Research

CPUs, memory, and NICs are also critical in accelerated servers. While CPUs are not responsible with the massive parallel processing (handled by the GPU), CPUs are still used in AI tasks such as managing data processing, initial model set up, and orchestrating GPU libraries. Memory (both RAM and storage) is also critical in processing vast amounts of data. In order to avoid bottlenecks in data progressing, accelerated servers must be equipped with ample memory bandwidth, density, and efficiency. Network interface cards, or NICs, are dedicated add-in cards that are programmed to run network software processes and free up the server processing for its primary application tasks. While GPUs represent the majority of an accelerated server's BOM, the costs of other components are still higher on a dollars basis relative to traditional servers, which supports the higher ASP.

Accelerated server unit economics

Branded vendors (e.g., DELL, HPE) typically earn ~8-10% gross margins and ~MSD% operating margins on AI server hardware sales. Below we present an illustrative bill of

materials for an AI server (H100 DGX Server), with component pricing as of July 2025 to reflect a bill of materials that is unimpacted by the spike in DRAM/NAND prices that started in fall 2025. In this illustrative example, components represent ~95% of the total cost of product, while server brand markups & ODM manufacturing fees represent ~5%.

Exhibit 28: AI server bill of materials
Illustrative bill of materials for a CPU general server

H100 DGX Server

Component	Details	Estimated cost	Share of total BOM
CPU	2x Intel Xeon CPU	$4,000	2%
GPU	GPU Baseboard (8x100)	$206,780	86%
RAM	2 TB	$2,000	0.8%
Storage-1	8 x 3.84TB NVMe (Data Cache) + 2x 1.92TB NVMe (OS)	$3,000	1.2%
NIC	ConnectX-7 NIC	$10,400	4%
Motherboard	PCB	$1,504	1%
Cooling-1	Heat sinks for CPUs & GPUs, fans	$980	0%
Power	6 x 3.3 kw per server	$1,800	1%
Chassis	Cables, rails, chassis	$650	0%
Component costs		**$231,114**	**96%**
ODM Manufacturing/testing fees		$3,750	2%
Labor, freight, warranty		$6,250	3%
Non-component costs		**$10,000**	**4%**
Total cost of product:		**$241,114**	**100%**

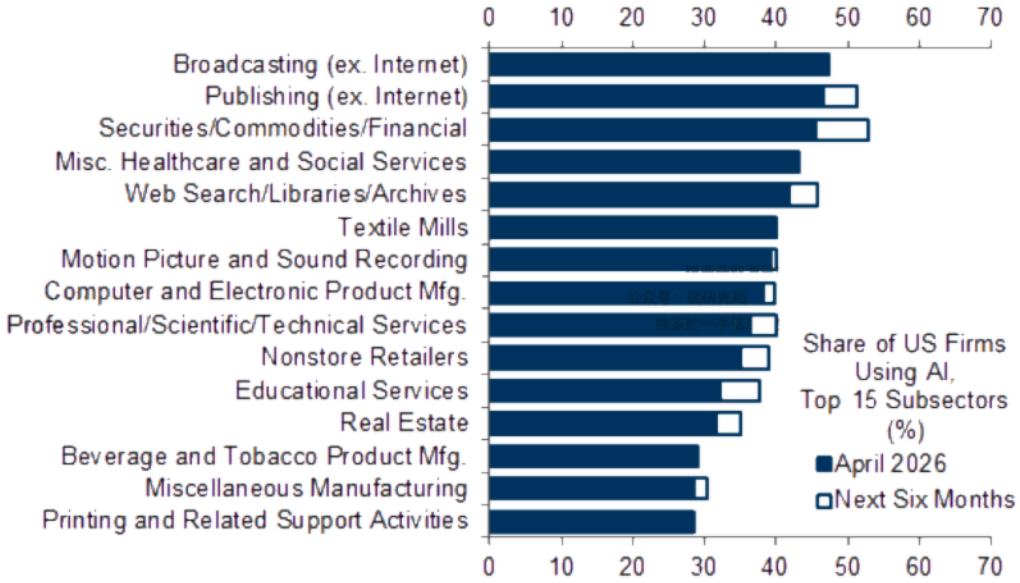
Average selling price:	*OEM gross margins:*
$265,000	*9%*

Component pricing as of July 2025 to reflect a bill of materials that is unimpacted by the spike in DRAM/NAND prices; GPU cost includes HBM pricing.

Source: Goldman Sachs Global Investment Research

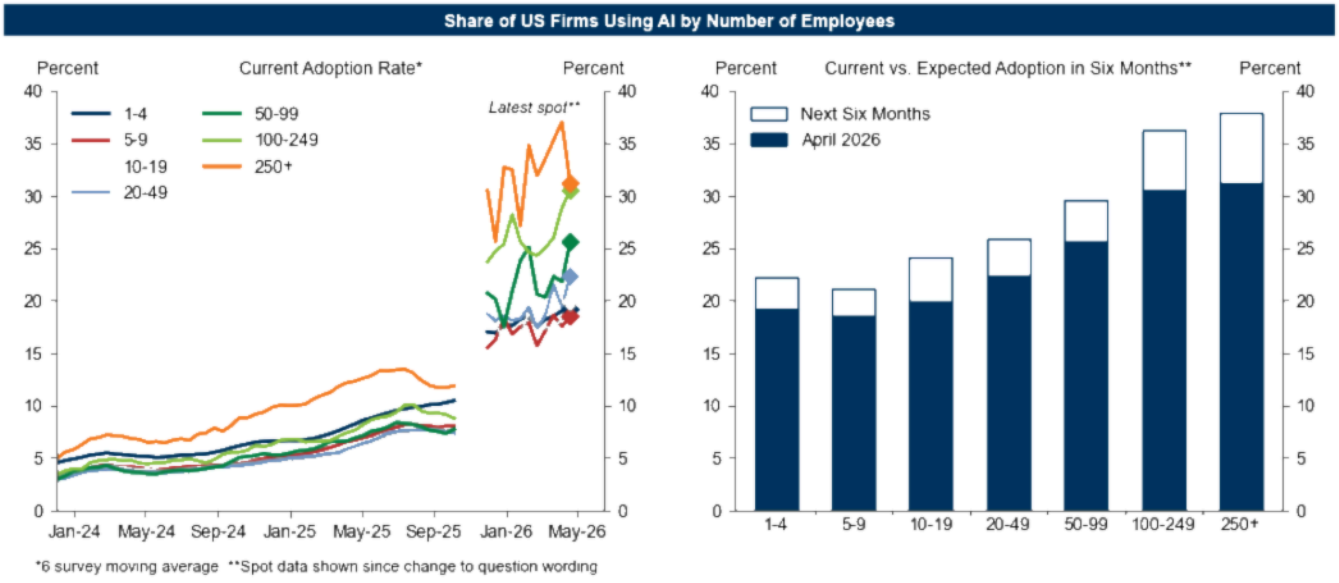
The major buyers of AI server infrastructure today are (1) hyperscalers, who are investing in AI for their own internal workloads (e.g., Search, e-commerce) and to build out capacity for their public cloud operations; **(2) tier 2 neoclouds** who are building of AI compute capacity for GPU rentals (e.g., Coreweave) or for sovereign AI projects; and to a lesser extent **(3) enterprises**, who are experimenting with AI in their business model to drive cost efficiencies and grow revenue opportunity. AI adoption is up across all industries, with GS Research finding that 19.8% of firms have adopted AI as of April 2026, with particular strength in Information, Professional, Education, and Finance sector firms. Adoption rates are highest among medium-sized firms (100-249 employees) and large enterprise (250+ employees) as of April 2026.

Exhibit 29: 19.8% of Firms Across All Industries Have Adopted AI as of April 2026
Share of US firms using AI by sector (%)



Source: Census Bureau, Goldman Sachs Global Investment Research

Exhibit 30: Current Adoption is highest among medium-sized and large enterprises
Share of US firms using AI by number of employees



Source: Census Bureau, Goldman Sachs Global Investment Research

Buying patterns (and therefore vendor market share) differ by customer type. Similar to traditional servers, hyperscalers generally buy AI servers from ODMs, while tier 2 cloud and enterprise customers buy from branded OEMs:

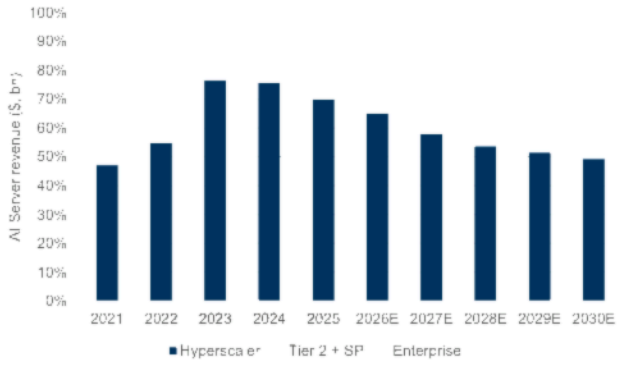
- **Hyperscalers** buy accelerated servers in massive quantities, representing 82% of total AI server shipments in 2025 across 5 companies (Amazon, Alphabet, Meta, Microsoft, Apple). Hyperscalers typically procure custom AI hardware, buying chips directly from merchant silicon providers & then working with an ODM to design custom boxes optimized for their specific needs, like density, cooling, or power. As

such, white box vendors represent more than ~92% of all hyperscaler AI server revenue, followed by Nvidia (~3%) and Supermicro (~3%).

- **Tier 2 cloud and neoclouds** primarily procure AI servers from best-in-breed server OEMs, lead by Supermicro (~39% share as of 2025) andDell (~38%), followed by white box (~10%) and Nvidia (~6%). Server OEMs like DELL & SMCI deliver AI servers that are based on NVDA's reference designs (referred to as an "HGX" server), as well as robust suite of consulting & implementation services. A HGX server is built in accordance with Nvidia's reference architecture around a foundational HGX baseboard (GPUs plus some networking), though OEMs can create distinct products by customizing things like cooling (air v. DLC), CPU & RAM density & configuration, and networking. We expect that the tier 2 cloud customers will continue to rely on the engineering expertise of OEMs, but by 2027, that some of the more sophisticated neoclouds will start to use ODMs for custom or semi custom server designs. 650 Group expects Tier 2 cloud to grow from ~13% of total AI shipments in 2025 to 36% by the end of the decade.

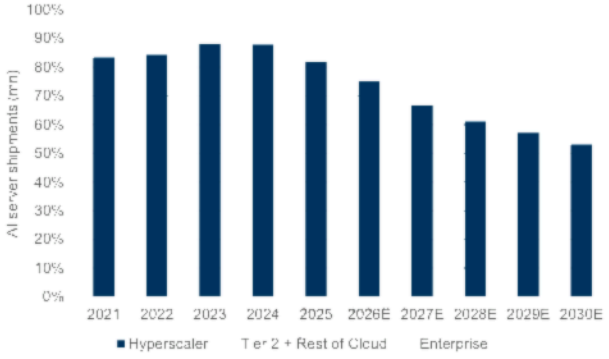
- **Enterprise** AI server demand is still nascent, with enterprise only representing ~5% of AI server shipments in 2025. Demand is addressed by traditional IT hardware providers Dell & HPE (each ~28% share in 2025), followed by Nvidia (~15%), who goes to market in the enterprise vertical with its DGX server platform & its complete portfolio of networking and software. OEMs also add value to their enterprise customers through their comprehensive services portfolio, including advisory, implementation, management, and financing. As such, we generally think of enterprise customers as the most margin accretive to OEMs.

Exhibit 31: Hyperscalers account for the majority of industry AI server revenue as of 2025...
AI server revenue ($, bn)



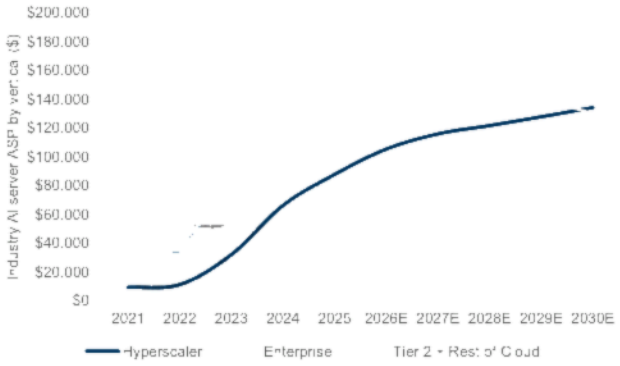
Source: 650 Group

Exhibit 32: .. with the highest share of AI server unit shipments...
AI server shipments (units, mn)



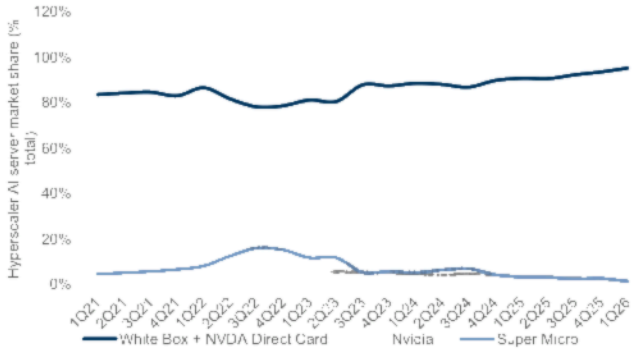
Source: 650 Group

Exhibit 33: ... but the lowest AI server ASP
AI server ASP ($)



Source: 650 Group

Exhibit 34: The hyperscaler AI server market is dominated by white box...
Hyperscaler AI server market share (% of total vertical revenue)

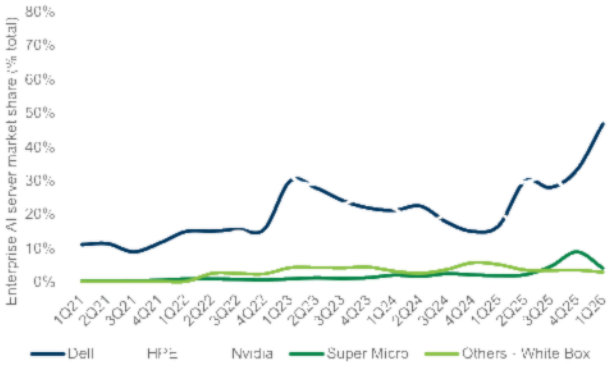


Source: 650 Group

Exhibit 35: ... while the tier 2 market is lead by best-in-breed server vendors DELL & SMCI
Tier 2 Cloud & Service Provider AI server market share (% of total vertical revenue)

Source: 650 Group

Exhibit 36: The enterprise AI server market is more balanced across HPE, DELL, and Nvidia
Enterprise AI server market share (% of total)



Source: 650 Group

Exhibit 37: Summary of key attributes by server manufacturer
Comparison of key attributes across DELL, HPE, SMCI, and whitebox

	Dell	Hewlett Packard	Supermicro	White Box
Hardware design & customization	**High** Large number of SKUs, foucs on reliability & performance	**High** Large number of SKUs, foucs on reliability & performance	**High** Highly customizable, modular open architecture	**High** Low-cost builds with OTS componets enables high degree of customization
Software & management tools	**High** Comprehensive portfolio of software & services	**High** Comprehensive portfolio of software & services	**Medium** Focus on hardware optimization > software & support	**Low** Typically sold as bare metal hardware, users integreat own OS
Ecosystem & partnerships	**High** Deep integrations & partnership across storage, software, and security	**High** Deep integrations & partnership across storage, software, and security	**High** Deep integrations & partnership across storage, software, and security	**Low** Rely on commodity hardware, merchant silicion, open-source software
Customer breadth & bundling	**High** SMBs to large enterprise, Tier 2 cloud	**High** SMBs to large enterprise	**Medium** Tier 2 Cloud & sophisticated enterprise	**Low** Hyperscalers & CSPs
Brand reputation & customer loyalty	**High** High reliability and performance across diverse IT environments	**High** High reliability and performance across diverse IT environments	**Medium** Speed to market advantage & US based engineers, but history of financial reporting & spyware issues	**Low** ODMs are unbranded
Price	**High**	**High**	**Medium**	**Low**

Source: Goldman Sachs Global Investment Research

Attaching services & software support can improve the profitability of AI server sales. In order to maximize the performance & functionality of IT hardware, IT hardware vendors often provide additional services to customers, such as professional services (installation, configuration, migration), lifecycle services (managed services, ongoing support, break/fix), as well as some cloud or as-a-service offerings (DELL APEX, HPE Greenlake, NetApp Keystone). Enterprise customers generally need more support to implement IT projects than tier 2 cloud or hyperscaler customers. These customers typically utilize a combination of OEM-offered services for point solutions and specialized IT systems integrators that work across vendors (e.g., Accenture, IBM, Tata). More sophisticated customers that are building bespoke AI clusters may also engage companies like Penguin (PENG), who has 30+ years of expertise in data center systems integration. This expertise makes PENG a partner of choice for major AI projects like the Meta Research Cluster that are focused on maximizing both time-to-first-token and performance of the underlying GPUs. Penguin's solutions are faster to deploy in part because the company pre-builds, pre-wires, pre-configures, and tests clusters in their factory before shipping to customers (v. drop shipping equipment and assembling on site at peers).

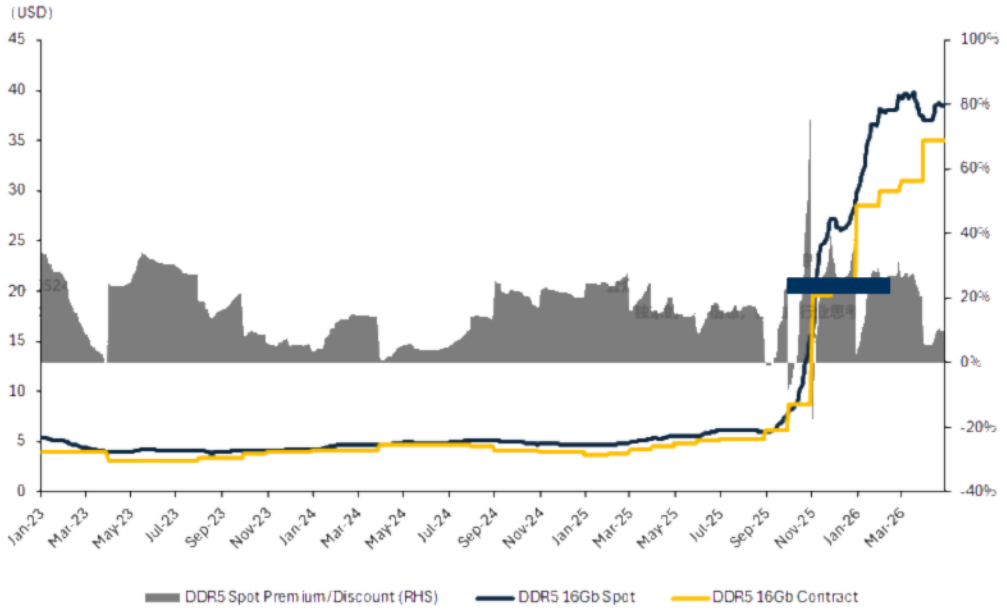
How do higher commodity costs impact the server market?

What's happening in the memory market? *(As of June 2026)* Prices for DRAM (a type of semiconductor memory used in hardware appliances for volatile storage, or "memory") and NAND (non-volatile, persistent semiconductor "storage") have increased significantly (beyond typical memory cycle peak/trough) starting in fall 2025, driven in large part by demand for AI infrastructure. In order to produce ample supply of the more profitable High Bandwidth Memory (HBM) that is in-demand for AI infrastructure buildouts, chipmakers (e.g., Samsung, Micron, SK Hynix have shifted capacity away from standard DRAM & NAND production, leading the supply shortages and sustained price hikes. A HBM wafer uses ~3x the silicon area per gigabyte as compared to conventional DRAM, meaning that the prioritization of HBM has an outsized effect on conventional DRAM output. GS Semiconductor analysts forecast ~316% growth in conventional DRAM pricing in 2026E (as of May 31, 2026). We note that many of the same factors that are driving higher DRAM costs are also impacting NAND flash pricing, with GS Semiconductor analysts estimating a 256% increase in NAND pricing in 2026. As of May 2026, GS semiconductor analysts expect the supply/demand mismatch to persist though through 2028.

Why this matters? DRAM & NAND typically represents ~30% of the bill of materials for a traditional server (Exhibit 40). Under current industry expectations for DRAM pricing, memory could represent >60% of the bill of materials for a traditional server by the end of calendar 2026, assuming no offsetting actions. In an illustrative example, if a traditional server vendor wanted to maintain its gross margin rate of ~15% while absorbing a ~315% DRAM price increase, the vendor would have to raise its prices by over 90%. *This assumes no pricing pressure from other components of the bill of material, like CPUs, which have also been observing price increases from Intel and AMD.*

For AI servers, memory represents a smaller share of the bill of materials because of significant GPU prices (DRAM & NAND ~2%, HBM about 6% of the BoM). As such, relative price increases to offset the rising input costs will be less meaningful.

Exhibit 38: DDR5 16Gb spot is trading at 10% premium vs. latest contract price
DDR5 16Gb spot pricing premium/discount vs. contract



Source: Trendforce

Hardware OEMs have a well-established playbook for managing typical commodity cost volatility, including through de-spec'ing, qualifying new suppliers, building strategic inventory, and passing through higher prices. That said, this playbook has been built around more "typical" memory cycles (i.e., ~3 year period of underbuilding/supply shortages leading to over building/over supply) and therefore ***may not be sufficient to address the current structural memory headwind.***

- **Signing long-term agreements (LTAs).** Server vendors like DELL & HPE are actively asking for LTAs (also called as New Business Model or Strategic Customer Agreements) to secure mid-to-long term volumes. Many of these OEMs have finalized their LTAs, which unlike the past, include higher levels of binding commitments. Based on checks done by our global Semi's team, the LTAs being discussed include sizable prepayments and meaningful penalty clauses for non-compliance which could guarantee stronger binding power. Pricing is being discussed within a pre-determined range with a floor price that guarantees a high margin level on a sustainable basis.
- **Qualifying new suppliers:** In order to qualify a new supplier, vendors conduct a multi-stage validation process to formally certify that a specific supplier's memory modules are reliable, compatible, and performant enough to be used across their range of computer models. This is typically a deep, cross-functional effort across engineering, quality assurance, and supply chain teams that can take 3-6 months.
  - It is important to note that in the case of memory, the conver
    market is dominated by three major players (Samsung, Micro
    who together represent over 90% of the market. Because ent
    OEM's all work with major DRAM suppliers already, that mean
    potential new suppliers are smaller players like Nanya and Wi

人在草木间

or CXMT in Mainland China, which in totality represent <10% of the total market (and therefore, qualifying them as suppliers likely will not fully address supply shortages). Similar to the DRAM market, the NAND market is also highly concentrated across Samsung, SK Hynix, Kioxia, Western Digital, and Micron (>90%). Smaller players like China's YMTC only produce a small amount of global supply.

- **Building strategic inventory:** Server vendors can leverage their balance sheets to build strategic reserves of key components (in some instances, up to 6 months of supply), which can help to offset near-term impacts of fluctuating commodity costs. Building strategic supply of inventory (v. "just in time" manufacturing model where components arrive just in time for production to minimize holding costs) can help vendors hedge prices and assure supply. That said, tying up the balance sheet with inventory limits a vendor's ability to generate interest income or invest elsewhere & carries the risk of inventory obsolescence. HPE and Lenovo have a historically leveraged their balances to strategically build inventory, while DELL operated on a more just-in-time manufacturing model. Server vendors typically sign long-term arrangements with component suppliers with volume purchase arrangements, though pricing is not always guaranteed.
- **Raising prices:** Vendors can pass through higher costs by raising prices, either directly (increasing list price), indirectly (reducing discounting), or through product mix (demand shaping). First, increasing direct list prices are more common in commercial and enterprise markets where customers are less sensitive to small prices increases (v. consumer buyers). Second, vendors can adjust their discounting strategy to increase their effective price by modifying sales, promotions, and rebates. Third, OEMs with more direct sales forces can shape demand by proactively steering customers to products that are more profitable, readily available platforms.

How does the current memory supply shortage impact the server market?
We examine two ways in which the ongoing DRAM memory supply shortage could impact supply, demand, and profitability in the server market.

1. **Supplier shortage:** Demand for conventional memory is significantly ahead of supply, which could pose a headwind to shipments. As of May 2026, GS Semiconductor analysts expect 2026/2027/2028 DRAM supply/demand will be in large undersupply of 5.0%/5.9%/3.9%, making it the most severe year of undersupply in the last 15+ years. As a result, we do not expect server vendors will fully be able to ship to their demand/backlog in the near to medium term. Further, GS semiconductor analysts expect the expanding adoption of LTAs will enhance demand visibility for memory suppliers, leading them to make more efficient capex plans and lower the oversupply risk seen in prior cycles.
2. **Demand elasticity:** In order to protect already-thin profit margins, vendors are likely to pass through higher costs to customers, which may also impact demand. For example, there are reports of server vendors like Dell raising prices by more than 50% year to date in 2026 (*as of May)*. While it is still too early to tell what the demand impact will ultimately look like (as of May 2026), we expect that there will be a limit to how much of a price increase enterprise server customers in particular can absorb before

demand starts to deteriorate given the generally fixed nature of IT budgets. Conversely, we have seen hyperscalers & AI neoclouds consistently raise their capex budget throughout 2026 in order to address inflationary commodity costs in AI infrastructure. Said differently, we do not expect to see increasing server prices impact the elasticity of demand for servers from this customer cohort.