Forecast Analysis: AI-Optimized Servers, Worldwide

Published 30 November 2023 - ID G00803017 - 19 min read

By Analyst(s): Adrian O'Connell

Initiatives: Technology Market Essentials

This forecast of AI servers up to 2027 covers a view of customer segments, AI type and training, or inference stages. AI servers are expected to account for 2.7 million units by 2027 and $81 billion of end-user spending, representing a spending compound annual growth rate of 30.1% from 2022 to 2027.

Overview

Forecast Assumptions

- Hyperscalers will prioritize spending on large AI systems over general-purpose systems in 2024.
- Enterprise spending on AI servers will be prioritized at the expense of other server purchases through 2025.
- The average selling price (ASP) for AI servers will grow at a compound annual growth rate (CAGR) of 4.5% from 2022 through 2027, peaking at $32,300 in 2024.

Market Impacts

- Hyperscalers typically account for around 40% of overall server market volumes. Hyperscalers and other service providers will account for 67% of AI server units in 2024. Absolute units will increase, but this proportion will fall to 55% by 2027.
- Enterprise shipments of AI servers will reach almost 460,000 units in 2024, rising to 740,000 by 2027. This represents 24% of AI server shipments in 2024 and 27% in 2027, plus 12% of total enterprise server shipments by 2027.
- The competitive shift on the supply side for accelerators, as well as the increasing proportion of inference-based systems, will result in an erosion of AI server ASPs. The ASP for an AI-optimized server will be $32,300 in 2024 falling to $29,500 by 2027.

Notable Changes

This Forecast Analysis provides a first forecast of the shipment and spending opportunity created by AI servers through to 2027.

AI servers are defined as servers that use specialist workload accelerators to support AI workloads, making them an emerging subset of the total servers that utilize workload accelerators. Accelerators used in high-performance computing (HPC) environments in a non-AI context are excluded. Similarly, the related category of function accelerator cards is also excluded (specifically, smart network interface cards [SmartNICs]).

This forecast only considers servers that will utilize graphics processing units (GPUs), application-specific integrated circuits (ASICs) and field-programmable gate arrays (FPGAs) for AI purposes. For further details on the market for AI-enabled CPUs, please refer to the Forecast: AI Semiconductors, Worldwide, 2021-2027, 3Q23 Update.

The Forecast: AI Semiconductors, Worldwide, 2021-2027, 3Q23 Update and this research shares a consistent view of the AI-optimized hardware landscape, offering specific perspectives on the trends within the markets for components and systems. The Forecast: Servers, All Countries, 2021-2027, 3Q23 Update makes use of the total GPUs, FPGAs and ASICs contained in the semiconductors research. Consistent with that forecast, this research defines an AI server as being one with workload accelerators, designed specifically to execute the algorithms and programmatic models associated with AI-based applications.

The superset of AI workloads across the entire market will go beyond this, with many AI workloads being able to be run on CPUs, or accelerated CPUs without the addition of a dedicated workload accelerator chip. This research excludes systems running these AI workloads.

Accelerators may be preinstalled into servers by server manufacturers, or installed by end users at the time of — or after — system deployment. This research only includes the servers with accelerators installed by the manufacturer at the time of the sale — any postsales upgrades are excluded.

For more details about the forecast methodology used to create this AI server forecast, see Market Definitions and Methodology: Servers.

Forecast Data Summary

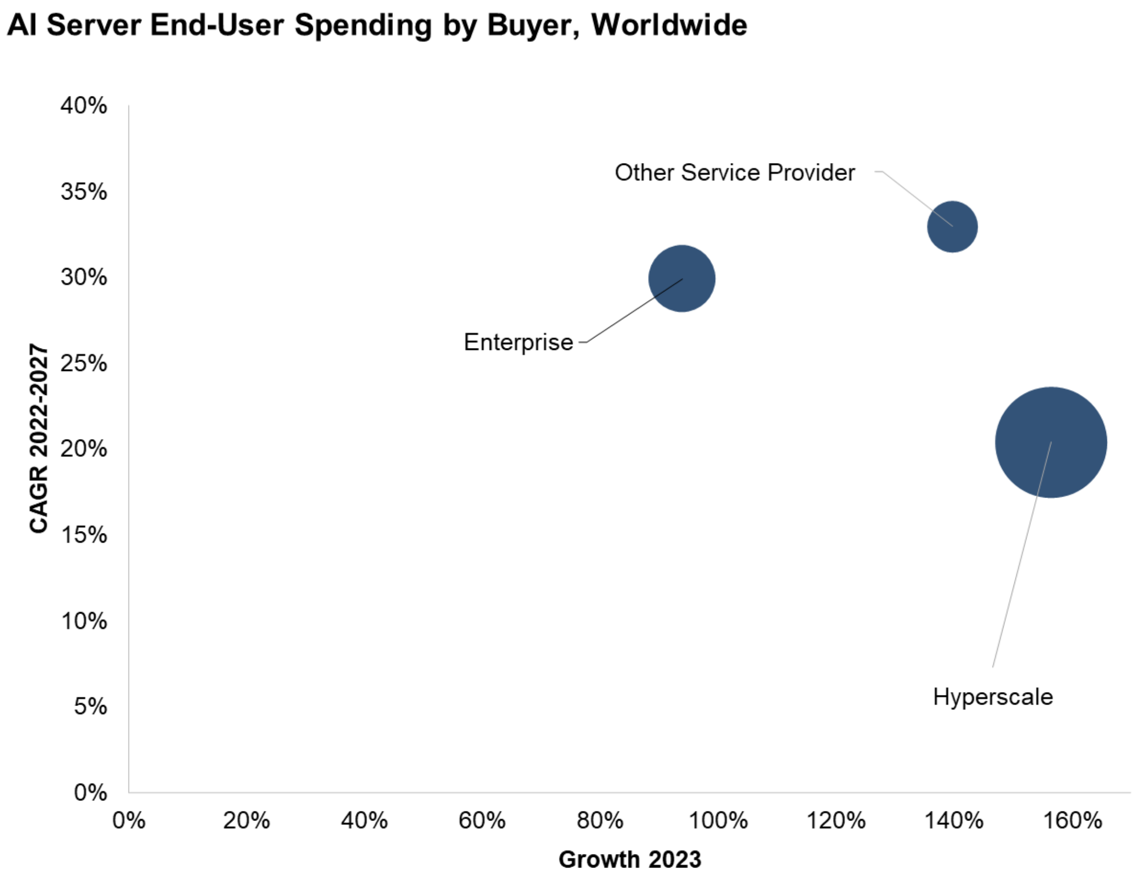
AI servers have been a factor in the market for some time but since the start of 2023, AI servers have come into focus both in terms of volume as well as attention — given all the hype around generative AI (GenAI). AI servers have been gaining traction in a number of areas from computer vision to natural language processing. In addition, with the advent of GenAI broadly and the headlines driven by ChatGPT in particular, it helped push the topic to the forefront for many enterprises.

Although the AI server forecast has strong spending growth, with a CAGR of 30.1% to 2027, this is not a market that is certain to follow a smooth, predetermined adoption curve. There is a metaphorical space race or gold rush going on in this area. Particularly in the consumer hyperscale segment, there is a significant possibility that current investment levels will not be sustained and could even retrench. In terms of broader enterprise adoption, we are assuming after an initial evaluation period, there will be increasing adoption. However, significant adoption within the forecast period is subject to enterprises finding a clear ROI for AI investments and, increasingly, net new use cases, which will be additive to current spending levels.

That said, this forecast represents a relatively conservative view of the opportunity and, although we are conscious of potential downside scenarios, there is also the possibility of upside scenarios too — through broader or faster, enterprise adoption.

Figure 1 shows the forecast by buyer type, from 2022 through 2027.

Figure 1. Forecast AI-Optimized Server End-User Spending by Buyer Type, Worldwide (2022-2027)

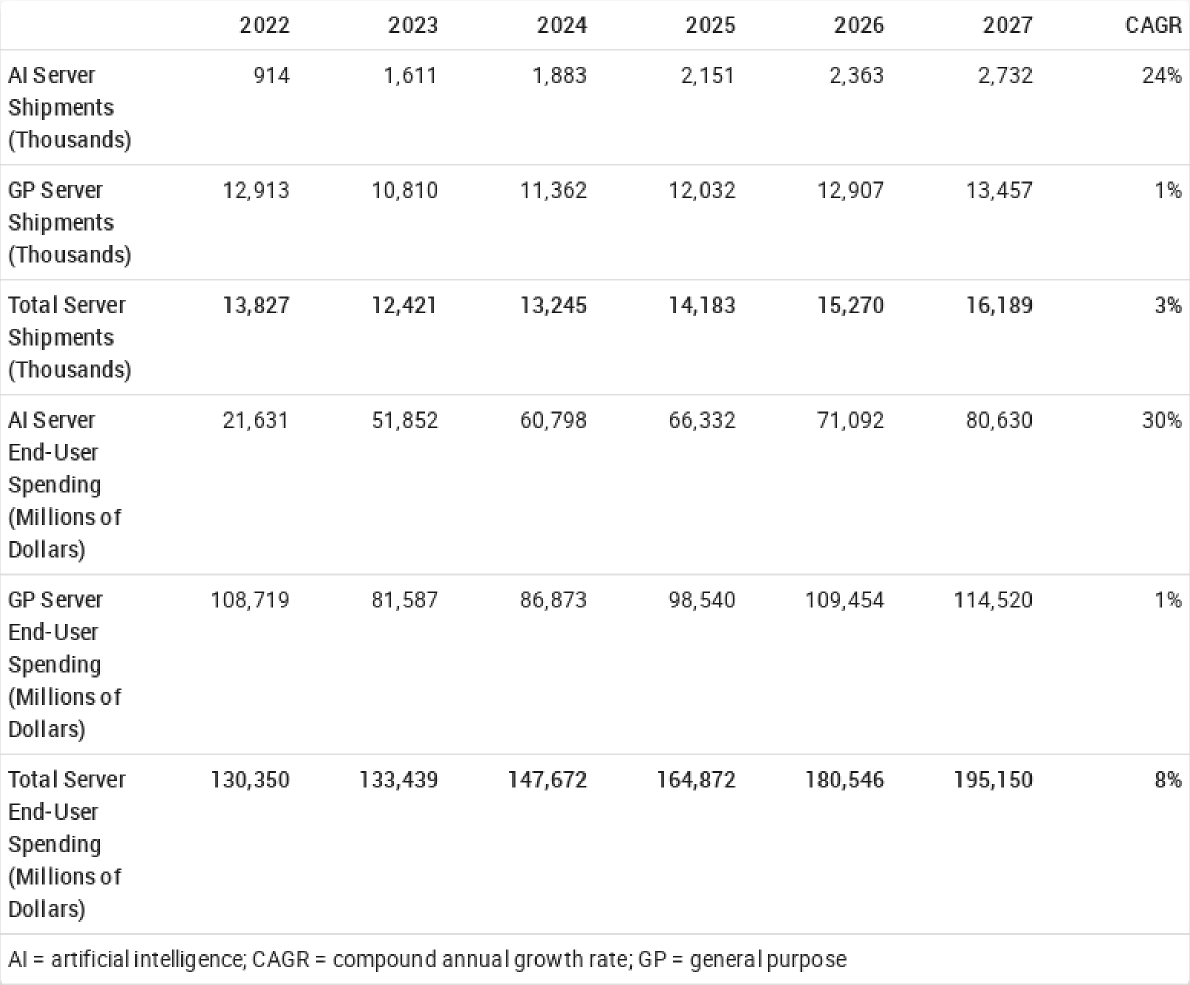


Source: Gartner (November 2023)
AI artificial intelligence
Note: The size of each bubble represents 2021 end-user spending by segment in current U.S. dollars.
ID: 803017

Table 1 shows shipments and end-user spending for AI servers and total servers.

Table 1: Shipments and End-User Spending for AI Servers

(Enlarged table in Appendix)



	2022	2023	2024	2025	2026	2027	CAGR
AI Server Shipments (Thousands)	914	1,611	1,883	2,151	2,363	2,732	24%
GP Server Shipments (Thousands)	12,913	10,810	11,362	12,032	12,907	13,457	1%
Total Server Shipments (Thousands)	13,827	12,421	13,245	14,183	15,270	16,189	3%
AI Server End-User Spending (Millions of Dollars)	21,631	51,852	60,798	66,332	71,092	80,630	30%
GP Server End-User Spending (Millions of Dollars)	108,719	81,587	86,873	98,540	109,454	114,520	1%
Total Server End-User Spending (Millions of Dollars)	130,350	133,439	147,672	164,872	180,546	195,150	8%

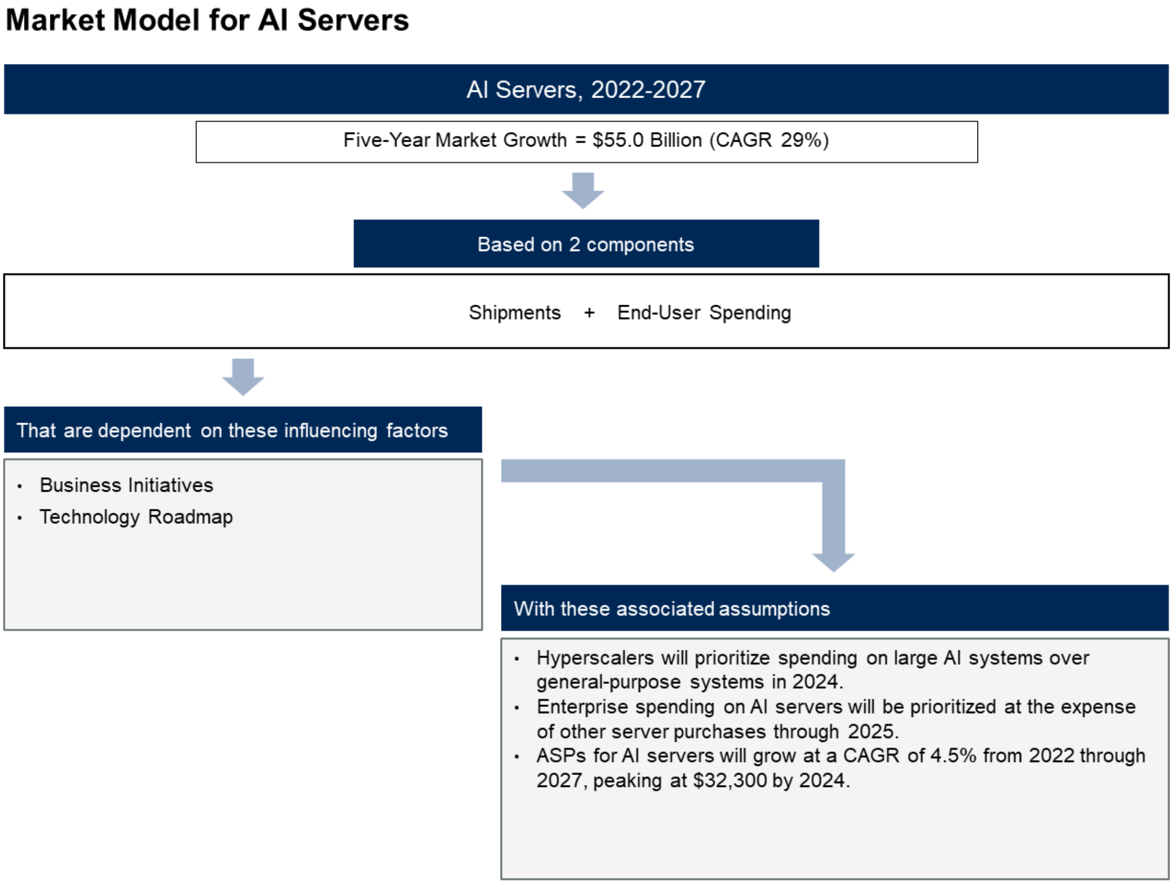
AI = artificial intelligence; CAGR = compound annual growth rate; GP = general purpose

Source: Gartner (November 2023)

Forecast Model Summary

Figure 2 summarizes the key components, influencing factors and forecast assumptions that drive the AI-optimized servers forecast. (See Figure 2 for the market model.)

Figure 2. Market Model for AI Servers Forecast



Market Model for AI Servers

AI Servers, 2022-2027

Five-Year Market Growth = $55.0 Billion (CAGR 29%)

Based on 2 components

Shipments + End-User Spending

That are dependent on these influencing factors

- Business Initiatives
- Technology Roadmap

With these associated assumptions

- Hyperscalers will prioritize spending on large AI systems over general-purpose systems in 2024.
- Enterprise spending on AI servers will be prioritized at the expense of other server purchases through 2025.
- ASPs for AI servers will grow at a CAGR of 4.5% from 2022 through 2027, peaking at $32,300 by 2024.

AI = artificial intelligence; ASP = average selling price; CAGR = compound annual growth rate

Source: Gartner (November 2023)

803017

Influencing Factors and Assumptions

> ### Influencing Factor: Business Initiatives
>
> The customer segments across the server market are all subject to different business conditions and drivers. As such, how they choose to prioritize their investments in order to meet the requirements of those various drivers will influence the overall market direction.

Forecast Assumption: Hyperscalers will continue to prioritize spending on large AI systems over general-purpose systems in 2024.

New — In going through a period of extending the useful life of their various installed hardware assets, the hyperscalers slowed their new hardware spending since the most recent peak in the first half of 2022. This resulted in shipment declines but strong spending growth in 2023 to date (up to 2Q23 market share). Although their general-purpose systems purchases are likely to start growing again in 2024, helped in part by the weak year-over-year comparisons, entering 2024 we still expect hyperscalers to prioritize spending on large AI systems.

Forecast Assumption: Enterprises will primarily be in exploratory mode with a view to AI-optimized servers through to the end of 2024.

New — We expect enterprises to spend 2024 and the first half of 2025 in a largely exploratory and evaluation mode of adoption for AI-optimized servers for GenAI use cases. There are some leading edge enterprises that are already deploying these sorts of systems in production environments, but we expect the mainstream to be more cautious. It's important to recognize the business environment as we approach the end of 2023, which is one of continued high inflation and ongoing business uncertainty. That environmental backdrop is likely to compound an already conservative approach from mainstream enterprises toward significant spending on a new — albeit promising — technology.

Forecast Assumption: Enterprise spending on AI servers will be prioritized at the expense of other server purchases through 2025.

New — In the earlier years of the forecast period, up to and into 2025, we expect enterprise purchases of AI-optimized servers to largely come at the expense of other server spending. For mainstream enterprises we do not expect total server budgets to increase due to AI procurements over this initial time frame.

AI server funding is most likely to come through savings elsewhere (whether projects are reprioritized or simply further extending the system life cycles).

Forecast Assumption: Hyperscale investment levels will continue to grow throughout the forecast period, particularly in the U.S. and China.

New — The prioritization of high-value AI servers, particularly in hyperscale environments, is leading to increased ASPs, which were already inflated by higher component costs. Hyperscale spending is driven both by their own internal requirements for first-party workloads, as well as increasing capacity to meet expected external customer "as a service" demand.

As such, one of the key variables here is how much the hyperscalers want and choose to invest in these capabilities ahead of proven demand from their customers. Hyperscalers are among the largest and most profitable tech companies in the world. Although not immune to the overall economic constraints, like higher interest rates that make the cost of borrowed money more expensive, they do not face the same pragmatic constraints on their investment levels as that of many of their enterprise customers.

A portion of these hyperscale investments is coming from the need to showcase relevant capabilities which may lead to further end-user demand. Future investment that may come from external demand, say fine-tuning some of the large language models (LLMs) is unclear. However, willingness to be able to meet future demand is one of the forecast variables, which has clear potential for both upsides and downsides.

The differing assumptions for enterprise and hyperscale spending inform our expectations in terms of relative buyer proportions over the forecast period.

Table 2 shows shipments for AI servers by buyer type.

Table 2: Shipments for AI Servers by Buyer Type

Units (Thousands)	2022	2023	2024	2025	2026	2027	CAGR
Enterprise	201	296	457	540	615	746	30%
Other Service Provider	113	234	334	372	411	471	33%
Hyperscale	600	1,081	1,092	1,239	1,337	1,516	20%
AI = artificial intelligence; CAGR = compound annual growth rate							

Source: Gartner (November 2023)

Table 3 shows end-user spending for AI servers by buyer type.

Table 3: End-User Spending for AI Servers by Buyer Type (Millions of U.S. Dollars)

End-User Spending	2022	2023	2024	2025	2026	2027	CAGR
Enterprise	4,975	9,645	10,336	11,608	12,797	14,836	24%
Other Service Provider	3,028	7,259	8,208	8,623	9,242	10,079	27%
Hyperscale	13,628	34,949	42,255	46,101	49,053	55,715	33%

AI = artificial intelligence; CAGR = compound annual growth rate

Source: Gartner (November 2023)

> **Influencing Factor: Sourcing Preferences**
>
> There will be several ways that end-users can source their capacity for AI servers and several factors that influence that decision. The choice of how to source that capacity has a bearing on the overall size of the market and who benefits from increasing AI hardware sales — enterprise or hyperscale system providers.

Forecast Assumption: Public cloud will be the primary preference of delivery model for enterprises while they are in the exploratory phase of GenAI.

New — As previously asserted, enterprises will spend the next 12 to 18 months evaluating where AI broadly, and GenAI specifically, may make sense for their own business needs. As they go through the process of gaining a better understanding of the potential benefits, challenges and possible return on investment (ROI), it will make sense to do much of this initial exploratory work in public cloud environments.

Where there is an unclear long-term business case with an understood ROI, using an operating expenditure (opex) public cloud model is likely to have fewer risks, while also being quicker than investing in an on-premises deployment.

Forecast Assumption: In the 2025 through 2027 time frame, enterprise deployments of AI-optimized servers will increasingly be deployed in on-premises environments.

New — Although hyperscaler provisioning for public cloud will account for the majority of near-term spending, there will be an increasing proportion of shipments that will come from enterprise customers. Entering the second half of the forecast period, there will be more demand from mainstream enterprises for AI servers.

Multiple factors around compliance and governance, data sovereignty, security concerns, distributed use cases or simply preference, will result in a growing proportion of on-premises deployments. Implicit within this is that special-purpose, or organization-specific model sizes will be smaller than the most well-known LLMs that have driven 2023. On-premises deployments will be skewed to inference models.

> **Influencing Factor: Technology Preferences**
>
> The technology preferences influencing factor covers the variety of technology factors that determine what types of technologies users will choose to deploy. Often there is more than one option from which the user will have to choose.

Forecast Assumption: Generative AI (GenAI) will account for >70% of hyperscale investments in AI-optimized hardware throughout the forecast period.

New — LLMs like Bard, ChatGPT and M6 have driven much of the AI-optimized server hardware demand to date. We expect those and similar models to account for much of the hyperscale investment over the forecast period. Rather than a once-and-done “build and deploy” approach, we expect the hyperscalers to continue to build on their existing capabilities as each tries to assert leadership (however you choose to define it) in this area. This will be a global issue, with China providers as keen to demonstrate leadership as those based in the U.S.

Forecast Assumption: Enterprise adoption of generative AI will primarily be focused on operational benefits in the first half of the forecast period, up to 2025.

New — As well as being largely exploratory in their approach to GenAI, we expect enterprise attention to focus on the potential cost-saving and operational benefits of GenAI in the short to medium term (now to 2025), as shown in Emerging Tech: Top Use Cases for Generative AI. Operational efficiency was the main business value associated with GenAI by providers.

Forecast Assumption: Generative AI will provide an uplift to traditional AI usage in enterprise environments in 2025 and beyond, due to increased awareness of the potential benefits.

New — As we move into the longer-term time frame of the forecast period (2026 through 2027) we expect the initial focus on operational efficiencies to broaden out into a wider set of use cases, including net new, or additive ones. This will impact overall market spending levels.

Within enterprise budgets, the IT Key Metrics Data 2023: Infrastructure Measures — Executive Summary finds data center infrastructure to account for 14% of IT spending. Due to the increasing importance, and higher cost of AI-optimized systems, we expect enterprises to increase their overall proportion of data center infrastructure spending from 14% to 15% of IT spending by 2027. This assumes a continued trend of cloud migration, which the market has seen for some time.

This also assumes that the increased server budgets will come from savings in software or services budgets. To illustrate the point, if we divide typical IT budgets into three pools of hardware, software and services (specifically, people — both external and internal), then you would typically expect ratios in the range of 20/40/40. (These aren't exact proportions but just illustrative). So, if a modest increase in the hardware budget results in net savings to the overall budget, then that can be a worthwhile investment.

Table 4 captures the thinking in the above assumptions and demonstrates our predictions for GenAI versus traditional AI over the forecast period.

Table 4: Shipments for AI Servers by AI Type

Units (Thousands)	2022	2023	2024	2025	2026	2027	CAGR
Generative AI	195	924	1,092	1,258	1,394	1,626	53%
Conventional AI	719	688	791	893	969	1,106	9%

AI = artificial intelligence; CAGR = compound annual growth rate

Source: Gartner (November 2023)

While Table 5 shows end-user spending for AI servers by AI type.

Table 5: End-User Spending for AI Servers by AI Type (Millions of U.S. Dollars)

End-User Spending	2022	2023	2024	2025	2026	2027	CAGR
Generative AI	8,050	35,453	41,400	45,425	47,120	51,531	45%
Conventional AI	13,582	16,400	19,398	20,907	23,971	29,099	16%
AI = artificial intelligence; CAGR = compound annual growth rate							

Source: Gartner (November 2023)

> **Influencing Factor: Service Delivery Models**
>
> Service delivery models describe how capacity is made available for delivery in the market. As such, the availability of various delivery methods will influence the options available for users to fulfill their capacity requirements.

Forecast Assumption: Over 90% of the AI servers in on-premises environments will be to support inferencing models throughout the forecast period.

New — Although public cloud is expected to be the primary delivery model for AI capacity in the early years of the forecast, the later years are expected to see a growing proportion of deployments in on-premises (including colocation) environments. The vast majority of the deployments are expected to be for inference models.

Forecast Assumption: More than 80% of the servers to support training models will be based in the public cloud throughout the forecast period.

New — The corollary to the previous assumption, is that we expect more than 80% of training to be done in public cloud environments throughout the forecast period. There will be increasing inference undertaken in public cloud environments, but there will not be such a clear advantage for public cloud models over alternatives.

The costs of both hardware and energy, as well as skills challenges, are likely to make on-premises deployments for training suboptimal for the average enterprise.

In overall terms, as stated in Forecast Analysis: AI Semiconductors, Worldwide, more than 60% of accelerators will be used to support inference models.

Table 6 shows shipments of AI servers by AI training/inference.

Table 6: Shipments of AI Servers by AI Training/Inference

Units (Thousands)	2022	2023	2024	2025	2026	2027	CAGR
Training	412	650	672	668	568	507	4%
Inference	503	961	1,211	1,483	1,795	2,225	35%
AI = artificial intelligence; CAGR = compound annual growth rate							

Source: Gartner (November 2023)

Meanwhile Table 7 shows end-user spending for AI servers by training/inference.

Table 7: End-User Spending for AI Servers by Training/Inference (Millions of U.S. Dollars)

End-User Spending	2022	2023	2024	2025	2026	2027	CAGR
Training	13,736	34,373	33,646	31,974	29,700	29,272	16%
Inference	7,896	17,479	27,152	34,357	41,392	51,358	45%
AI = artificial intelligence; CAGR = compound annual growth rate							

Source: Gartner (November 2023)

Influencing Factor: Technology Roadmap

> The technology roadmap influencing factor is a supply-side issue that outlines how the evolution of technologies may have a bearing on the market. For AI servers, the stage of the model deployment, the competitive dynamics in the market and the size of the models used will all have direct bearings on the overall size and direction taken by the market.

Forecast Assumption: Integrated server designs will account for >70% of training servers by 2027.

New — Integrated designs, where the accelerator is integrated onto the motherboard in contrast to a Peripheral Component Interconnect Express (PCIe)-based add-in card, allow the highest performance due to the additional bandwidth this design allows. As such, we assume that systems with this design will be the preference for training systems throughout the forecast period. Inference systems, where performance requirements are comparatively lower, will sometimes utilize integrated designs, but more commonly make use of PCIe options. Price points for PCIe-based systems should also be comparatively lower.

Forecast Assumption: ASPs for AI servers will grow at a CAGR for 4.5% from 2022 through 2027, peaking at $32,300 by 2024.

New — Throughout 2023, the GPU market was supply-constrained. This is expected to ease during 2024 as increasing capacity comes online for NVIDIA. In addition to that supply/demand balance, we expect the competitive landscape for accelerators to change over the forecast period.

NVIDIA will face increasing competition from both their traditional competitors (such as AMD and Intel), as well as emerging competitors, as new products become available throughout the forecast period. Additionally, the hyperscalers will increasingly use their own ASIC designs.

This changing dynamic in terms of accelerator supply is likely to see the ASPs for accelerators peaking in 2023. ASPs for systems will take longer to reduce, as it will take some time for this supply imbalance on the component side to work through to systems pricing. ASPs for server systems are forecast to reach their peak during 2024.

Forecast Assumption: The average model size used by enterprises will be >15 billion in size by 2027.

New — A central assumption underpinning the adoption of generative AI-optimized servers among enterprises is that model sizes will be significantly smaller than the LLMs which have dominated the landscape during 2023. While models like Pathways Language Model (PaLM), Generative Pretrained Transformer 3 (GPT-3) and Large Language Model Meta AI (LLaMA) can be 540 Bytes (B), 175B and 65B (respectively), we're assuming that mainstream enterprises will choose to deploy smaller, domain-specific models that are much smaller than these examples.

It is also a key factor in the size of the system types, and shapes the mix of the AI server sizes (as seen in Table 10).

Table 8 shows shipments of AI servers by accelerator type.

Table 8: Shipments of AI Servers by Accelerator Type

Units (Thousands)	2022	2023	2024	2025	2026	2027	CAGR
GPU	609	1,168	1,322	1,434	1,543	1,752	24%
ASIC/FPGA	305	443	561	717	820	980	26%

AI = artificial intelligence; ASIC = application-specific integrated circuit; CAGR = compound annual growth rate; FPGA = field-programmable gate array; GPU = graphics processing unit

Source: Gartner (November 2023)

Meanwhile Table 9 shows end-user spending for AI servers by accelerator type.

Table 9: End-User Spending for AI Servers by Accelerator Type (Millions of U.S. Dollars)

End-User Spending	2022	2023	2024	2025	2026	2027	CAGR
GPU	18,321	46,661	53,517	56,664	59,404	65,674	29%
ASIC/FPGA	3,310	5,191	7,282	9,667	11,688	14,956	35%

AI = artificial intelligence; ASIC = application-specific integrated circuit; CAGR = compound annual growth rate; FPGA = field-programmable gate array; GPU = graphics processing unit

Source: Gartner (November 2023)

Table 10 shows shipments of GPU AI servers by size.

Table 10: Shipments of GPU AI Servers by Size

Units (Thousands)	2022	2023	2024	2025	2026	2027	CAGR
1 to 3 GPUs	530	958	1,084	1,180	1,272	1,449	22%
4 to 7 GPUs	49	82	106	118	131	152	26%
8 GPUs and Above	30	129	132	136	140	151	38%

AI = artificial intelligence; CAGR = compound annual growth rate; GPU = graphics processing unit

Source: Gartner (November 2023)

An important point to note is the scale of the systems deployed to support AI workloads, with options running from one to two GPUs installed in a PCIe slot, to eight and above GPU designs integrated into the server via the SXM socket.

Tables 10 and 11 give a perspective on how the size of the systems will break down over the forecast period. While Table 11 shows end-user spending for GPU AI servers by size.

Table 11: End-User Spending for GPU AI Servers by Size (Millions of U.S. Dollars)

End-User Spending	2022	2023	2024	2025	2026	2027	CAGR
1 to 3 GPUs	10,167	19,004	23,566	25,678	27,735	31,620	25%
4 to 7 GPUs	2,670	4,497	5,844	6,512	7,199	8,197	25%
8 GPUs and Above	5,484	23,160	24,107	24,474	24,470	25,858	36%
AI = artificial intelligence; CAGR = compound annual growth rate; GPU = graphics processing unit							

Source: Gartner (November 2023)

Acronym Key and Glossary Terms

AI	artificial intelligence
ASIC	application-specific integrated circuit
ASP	average selling price
B	Bytes
CAGR	compound annual growth rate
ChatGPT	Chat Generative Pretrained Transformer
CPU	central processing unit
FPGA	field-programmable gate array
GenAI	generative artificial intelligence
GP	general purpose
GPT-3	Generative Pretrained Transformer 3
GPU	graphics processing unit
HPC	high-performance computing
LLaMA	Large Language Model Meta AI
LLM	large language model
opex	operating expenditure
PaLM	Pathways Language Model
PCIe	Peripheral Component Interconnect Express
ROI	return on investment
SmartNIC	smart network interface card

Recommended by the Author

Some documents may not be available as part of your current Gartner subscription.

Forecast Analysis: Servers, Worldwide

Forecast Analysis: AI Semiconductors, Worldwide

Emerging Tech: Top Use Cases for Generative AI

Emerging Tech: Primary Impact of Generative AI on Business Use Cases

Applying AI — A Framework for the Enterprise

Table 1: Shipments and End-User Spending for AI Servers

	2022	2023	2024	2025	2026	2027	CAGR
AI Server Shipments (Thousands)	914	1,611	1,883	2,151	2,363	2,732	24%
GP Server Shipments (Thousands)	12,913	10,810	11,362	12,032	12,907	13,457	1%
Total Server Shipments (Thousands)	**13,827**	**12,421**	**13,245**	**14,183**	**15,270**	**16,189**	**3%**
AI Server End-User Spending (Millions of Dollars)	21,631	51,852	60,798	66,332	71,092	80,630	30%
GP Server End-User Spending (Millions of Dollars)	108,719	81,587	86,873	98,540	109,454	114,520	1%
Total Server End-User Spending (Millions of Dollars)	**130,350**	**133,439**	**147,672**	**164,872**	**180,546**	**195,150**	**8%**

AI = artificial intelligence; CAGR = compound annual growth rate; GP = general purpose

Source: Gartner (November 2023)

Table 2: Shipments for AI Servers by Buyer Type

Units (Thousands)	2022	2023	2024	2025	2026	2027	CAGR
Enterprise	201	296	457	540	615	746	30%
Other Service Provider	113	234	334	372	411	471	33%
Hyperscale	600	1,081	1,092	1,239	1,337	1,516	20%

AI = artificial intelligence; CAGR = compound annual growth rate

Source: Gartner (November 2023)

Table 3: End-User Spending for AI Servers by Buyer Type (Millions of U.S. Dollars)

End-User Spending	2022	2023	2024	2025	2026	2027	CAGR
Enterprise	4,975	9,645	10,336	11,608	12,797	14,836	24%
Other Service Provider	3,028	7,259	8,208	8,623	9,242	10,079	27%
Hyperscale	13,628	34,949	42,255	46,101	49,053	55,715	33%
AI = artificial intelligence; CAGR = compound annual growth rate							

Source: Gartner (November 2023)

Table 4: Shipments for AI Servers by AI Type

Units (Thousands)	2022	2023	2024	2025	2026	2027	CAGR
Generative AI	195	924	1,092	1,258	1,394	1,626	53%
Conventional AI	719	688	791	893	969	1,106	9%
AI = artificial intelligence; CAGR = compound annual growth rate							

Source: Gartner (November 2023)

Table 5: End-User Spending for AI Servers by AI Type (Millions of U.S. Dollars)

End-User Spending	2022	2023	2024	2025	2026	2027	CAGR
Generative AI	8,050	35,453	41,400	45,425	47,120	51,531	45%
Conventional AI	13,582	16,400	19,398	20,907	23,971	29,099	16%
AI = artificial intelligence; CAGR = compound annual growth rate							

Source: Gartner (November 2023)

Table 6: Shipments of AI Servers by AI Training/Inference

Units (Thousands)	2022	2023	2024	2025	2026	2027	CAGR
Training	412	650	672	668	568	507	4%
Inference	503	961	1,211	1,483	1,795	2,225	35%
AI = artificial intelligence; CAGR = compound annual growth rate							

Source: Gartner (November 2023)

Table 7: End-User Spending for AI Servers by Training/Inference (Millions of U.S. Dollars)

End-User Spending	2022	2023	2024	2025	2026	2027	CAGR
Training	13,736	34,373	33,646	31,974	29,700	29,272	16%
Inference	7,896	17,479	27,152	34,357	41,392	51,358	45%
AI = artificial intelligence; CAGR = compound annual growth rate							

Source: Gartner (November 2023)

Table 8: Shipments of AI Servers by Accelerator Type

Units (Thousands)	2022	2023	2024	2025	2026	2027	CAGR
GPU	609	1,168	1,322	1,434	1,543	1,752	24%
ASIC/FPGA	305	443	561	717	820	980	26%
AI = artificial intelligence; ASIC = application-specific integrated circuit; CAGR = compound annual growth rate; FPGA = field-programmable gate array; GPU = graphics processing unit							

Source: Gartner (November 2023)

Table 9: End-User Spending for AI Servers by Accelerator Type (Millions of U.S. Dollars)

End-User Spending	2022	2023	2024	2025	2026	2027	CAGR
GPU	18,321	46,661	53,517	56,664	59,404	65,674	29%
ASIC/FPGA	3,310	5,191	7,282	9,667	11,688	14,956	35%

AI = artificial intelligence; ASIC = application-specific integrated circuit; CAGR = compound annual growth rate; FPGA = field-programmable gate array; GPU = graphics processing unit

Source: Gartner (November 2023)

Table 10: Shipments of GPU AI Servers by Size

Units (Thousands)	2022	2023	2024	2025	2026	2027	CAGR
1 to 3 GPUs	530	958	1,084	1,180	1,272	1,449	22%
4 to 7 GPUs	49	82	106	118	131	152	26%
8 GPUs and Above	30	129	132	136	140	151	38%

AI = artificial intelligence; CAGR = compound annual growth rate; GPU = graphics processing unit

Source: Gartner (November 2023)

Table 11: End-User Spending for GPU AI Servers by Size (Millions of U.S. Dollars)

End-User Spending	2022	2023	2024	2025	2026	2027	CAGR
1 to 3 GPUs	10,167	19,004	23,566	25,678	27,735	31,620	25%
4 to 7 GPUs	2,670	4,497	5,844	6,512	7,199	8,197	25%
8 GPUs and Above	5,484	23,160	24,107	24,474	24,470	25,858	36%
AI = artificial intelligence; CAGR = compound annual growth rate; GPU = graphics processing unit							

Source: Gartner (November 2023)