Generative AI

2024

Assessing Opportunities and Disruptions in an Evolving Trillion-Dollar Market

The adoption of generative artificial intelligence (AI) and large language models (LLM) is already rippling through every segment of the technology sector, as incumbents and new entrants reimagine existing end markets to potentially generate $1.3 trillion in revenue by 2032.

- **Compute for Training and Major LLMs:** Nvidia can maintain dominance in data-center compute, driven by its performance lead and strong demand for AI workloads tied to training LLMs. Among the foundational model companies, competition is likely to be intense, with OpenAI-Microsoft, Google, Meta and Anthropic releasing new LLMs.
- **On-Device AI for Inference:** On-device AI could see rapid adoption of features like summarizing text, translating language, generating images and videos with text-based prompts, and voice assistants. Makers of edge devices, such as PCs from Dell, smartphones from Apple and connected cars from Tesla, could be among the beneficiaries of the demand for on-device AI.
- **Incremental Cloud Sales Boost for Hyperscalers:** Hyperscale cloud companies like Microsoft, Amazon.com and Google should get a boost in recurring revenue. Cloud vendors' high capex and availability of GPUs for training and inferencing aids monetization as gen AI is deployed across enterprises.

Featured in This Report: Beginning in Section 3 and throughout the report, Bloomberg Intelligence's interactive market-sizing models are used to forecast growth potential for total generative-AI spending and segments including, hardware, software, training, inference, advertising and services. These calculators are available on the Terminal.

March 28, 2024

Bloomberg Intelligence

BI

Contents

BI

To Contact the Analyst:

Mandeep Singh
msingh15@bloomberg.net
+1-212-617-9560

Section 1. Executive Summary

$1.3 Trillion
AI-driven sales by 2032 from about $64 billion in 2023

43%
Projected CAGR through 2032

$471 Billion
AI Training Market

Racking Up 10-12% of Technology Spending

Generative artificial intelligence is poised to produce $1.3 trillion in revenue – 10-12% of all technology spending – over the next eight years in hardware, software and services and more as businesses supercharge their products. The interfaces and tools leveraging gen AI are early, but we believe some common themes include generating summaries, personalized recommendations, image and video content using conversational user interfaces, and built-in language translation. Training AI through machine learning and neural network algorithms using massive datasets (the large language model, or LLM) will be a huge market, reaching $471 billion in sales by 2032 and boosting demand for accelerators on servers and storage units at data centers. Companies will use the public cloud to deploy generative AI, benefitting hyperscalers like Meta, Microsoft, Amazon and Alphabet, with a projected CAGR of 54% to $309 billion.

Key Research Topics

- **Training Opportunity:** Within hardware, infrastructure spending ($471 billion) could triple that of the device market by 2032, as companies invest in and consume compute and storage services on hyperscalers' clouds to manage intensive workloads used to "train" AI.
- **Importance of Inference:** Smartphone makers such as Apple and auto OEMs like Tesla could reap benefits from the demand for inference-based conversational AI products and vision AI offerings tied to generative AI. Once a machine is trained, inference is used to derive new knowledge or conclusions from existing data.
- **Foundry Positioning for Gen AI:** TSMC's dominance in leading-edge node semiconductor manufacturing means it should keep the lion's share of AI chip production orders from key players such as Nvidia and AMD. That advantage can extend thanks to the company's strong production yield. Also, many AI chip designers prefer TSMC's CoWoS packaging for its superior interconnection density, larger package sizes and cost effectiveness.
- **High-bandwidth memory (HBM) chips:** As AI models become more complex and training more demanding, HBM chips like DDR5 should be more widely adopted. The speed of chip performance improvement required for AI is faster than the evolution of miniaturization and advanced packaging, aiding the role of chip testers.
- **Regulatory Scrutiny for Gen AI Players:** The EU has moved forward with comprehensive regulations on AI, with formal approval of the AI Act in March paving way for implementation in the coming years. The transparency obligations will only apply to general-purpose AI systems that are deemed to pose a "systemic risk." OpenAI's GPT-4 and Google's Gemini are likely to be the first to fall into that category.

Performance and Valuation

BI's AI theme basket encompasses companies across the tech spectrum and is the standout performer over the past six months. Within AI, it's more than just Nvidia, with stock returns for the Magnificent 7 and a number of other players, including Super Micro, AMD and Broadcom, also aided by steady valuation multiple expansion.

Section 2. Catalysts to Watch

Attach Rates, Regulations to Pave Growth Path

Spending on generative AI has quickly become non-discretionary for enterprises, and we expect sharp growth to be driven by steady hardware investments, uptake of chatbots and attached subscriptions for copilot-type offerings. Retrieval-augmented generation (RAG) allows enterprises to leverage proprietary data lakes to enhance accuracy and reduce hallucinations within LLM responses. RAG will likely become a driving factor for enterprise AI adoption, and we may see many hyperscalers intertwine the technology within their own LLM offerings. Already, companies such as Nvidia have seen huge moves in growth forecasts as a result of the push into AI, while others, like Microsoft (Azure consumption and copilots) anticipate robust gains.

2.1 Copilots, Content Generation and Targeted Ads Lead

- **2024:** GPU, accelerator-chip availability improves for training LLMs
- **2024**: New versions of foundational LLMs show greater accuracy
- **2024**: Strong attach rates for copilots launched by software companies
- **2024:** Chatbots disrupt customer service and help save operations costs
- **2024:** RAG makes enterprises comfortable with deploying foundational LLMs
- **2024:** New content-generating tools, ad-targeting improvement from large internet companies
- **2025:** On-device AI gains steam with new features for mobile, PCs, AR/VR
- **2025:** Multimodal LLMs become a feature across internet apps, enterprise software
- **2025:** EU on course to adopt first comprehensive regulations through the AI Act.
- **2023-27:** TSMC's generative AI segment reaches compound annual growth of 50%
- **2027:** AI networking could expand by 5x, driven by specific accelerator requirements
- **2030:** Software spending on generative AI hits $204 billion (10% of the total) from $1 billion in 2022

Section 3. AI Overview

Addressable Markets Appear Ready to Expand

Training foundational LLMs is still the biggest driver of incremental revenue for generative AI, yet traction with GitHub Copilot and growing interest in new apps like Perplexity for consumer search and Sora for prompt-based video generation may continue to expand the addressable market. Generative AI is poised to be a $1.3 trillion market by 2032 as it boosts sales for tech's hardware, software, services, ads and gaming segments, growing at a compound annual rate of about 41%, based on BI's interactive market-sizing model. As the revolutionary technology changes how businesses operate and enhance their products and services, generative AI could expand to 10-12% of total information-technology spending in such segments from less than 1% today.

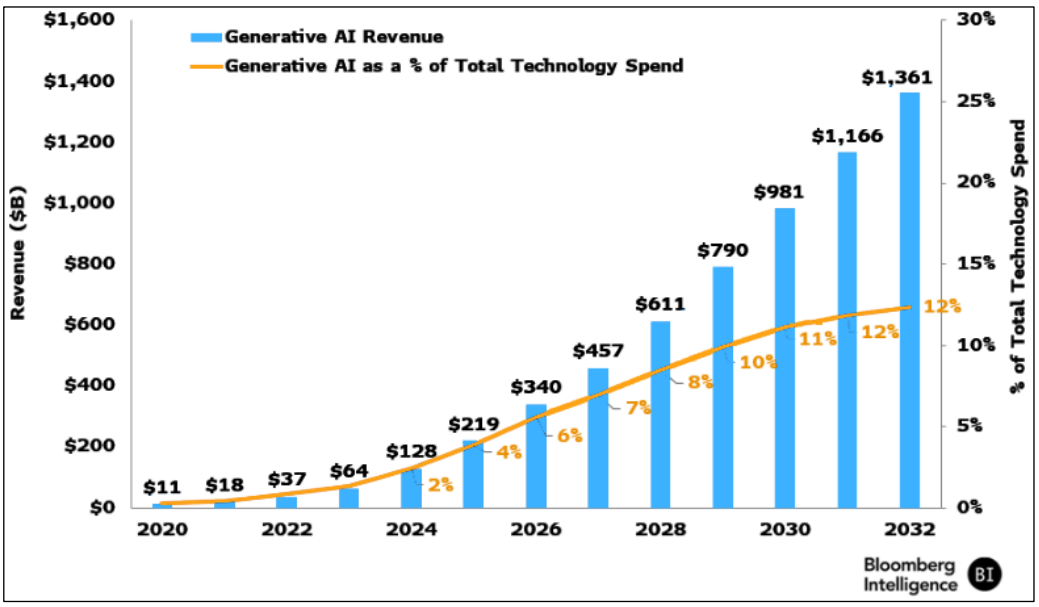
Figure 1: Generative AI Revenue Potential

(in milllions of $)	2023	2027E	2032E	Implied 9 yr. CAGR (%)
Hardware	**$53,105**	**$286,903**	**$639,399**	**32%**
Devices (Inference)	$6,415	$72,703	$168,641	44%
Computer Vision AI Products	$2,749	$19,387	$58,376	40%
Conversational AI Products	$3,666	$53,315	$110,265	46%
Infrastructure (Training)	$46,690	$214,200	$470,758	29%
AI Server	$26,060	$73,984	$105,197	17%
AI Storage	$10,858	$31,707	$56,982	20%
Generative AI Infrastructure as a Service	$9,772	$108,509	$308,579	47%
Compute	$4,343	$69,756	$173,575	51%
Internal Consumption	$1,303	$20,434	$33,312	43%
Hyperscale Consumption	$3,040	$49,322	$140,263	53%
Networking	$3,257	$16,911	$43,832	33%
Inference/Fine-Tuning Cloud	$2,172	$21,843	$91,171	51%
Software	**$5,028**	**$61,680**	**$317,961**	**59%**
Specialized Generative AI Assistants	$2,489	$22,029	$95,259	50%
Enterprise Applications	$1,493	$13,217	$50,011	48%
Consumer/E-Commerce Applications	$995	$8,812	$45,248	53%
Coding, DevOps and Generative AI Workflows	$473	$13,436	$68,763	74%
Generative AI Workload Infrastructure Software	$1,195	$13,885	$80,788	60%
Generative AI Drug Discovery Software	$32	$4,561	$35,091	117%
Generative AI Based Cybersecurity Spending	$11	$3,419	$15,063	124%
Generative AI Education Spending	$829	$4,349	$22,996	45%
Generative AI Based Gaming Spending	**$533**	**$24,890**	**$83,591**	**75%**
Virtual Goods	$133	$8,889	$31,347	83%
Game Design Software	$399	$16,000	$52,244	72%
Generative AI Driven Ad Spending	**$4,624**	**$53,154**	**$206,693**	**53%**
Search	$2,458	$21,006	$67,661	45%
Videos	$1,666	$24,729	$100,941	58%
Messaging	$500	$7,419	$38,091	62%
Generative AI Focused IT Services	**$165**	**$20,451**	**$80,904**	**99%**
Generative AI Based Business Services	**$78**	**$9,705**	**$32,443**	**95%**
Total	**$63,533**	**$456,782**	**$1,360,990**	**41%**

Bloomberg Intelligence BI

Source: BI's forecasts based on data from IDC, eMarketer, Statista

Figure 2: Generative AI Spending



Source: BI's forecasts based on data from IDC, eMarketer, Statista

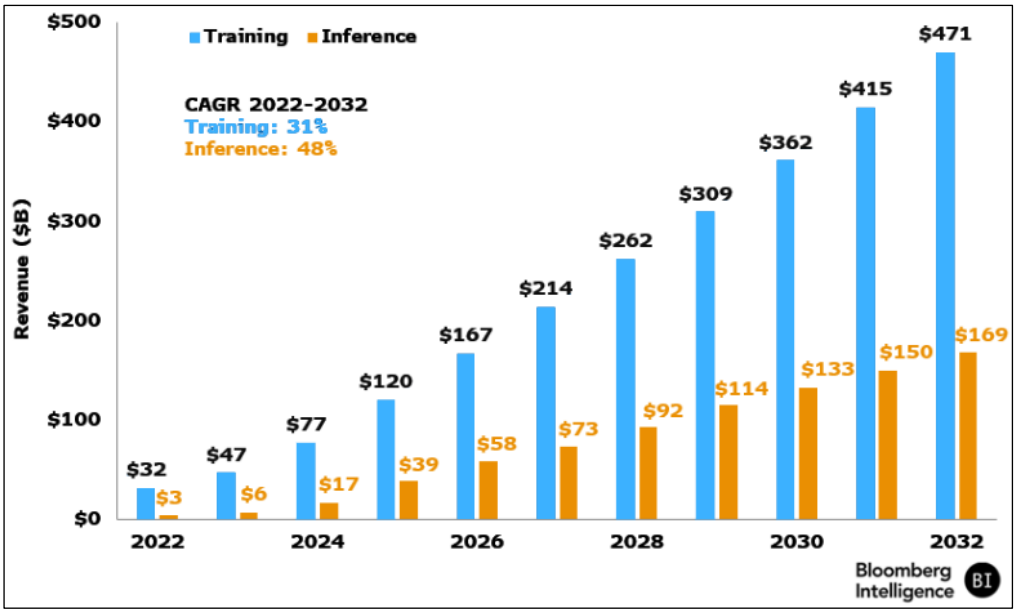
Semiconductors, hardware, cloud software, IT services and advertising companies may be the vanguard of these changes. Yet we also could see new products and services, a displacement of incumbents and the emergence of new categories.

3.1 Training, Then Inference Offer Market Opportunity

Training AI platforms through LLMs based on neural networks with billions of parameters will likely be a bigger part of the market than inference (using previously built models to make predictions or decisions) in the near-term, driving demand for accelerators for servers and storage units at data centers. Training could be the field's largest source of added revenue and nearly a $500 billion market by 2032, encompassing servers, storage and service offerings on the cloud. About 20% of the training segment revenue will be related to fine-tuning foundational LLMs and generative AI inferencing available through existing SaaS applications. The high growth expectations have so far been supported by the rapid change in data-center compute and storage to handle AI workloads across both consumer and enterprise apps. Nvidia has been the most significant beneficiary of the trend, while hyperscalers have all boosted capex to ensure GPU availability for both internal consumption and their cloud businesses.

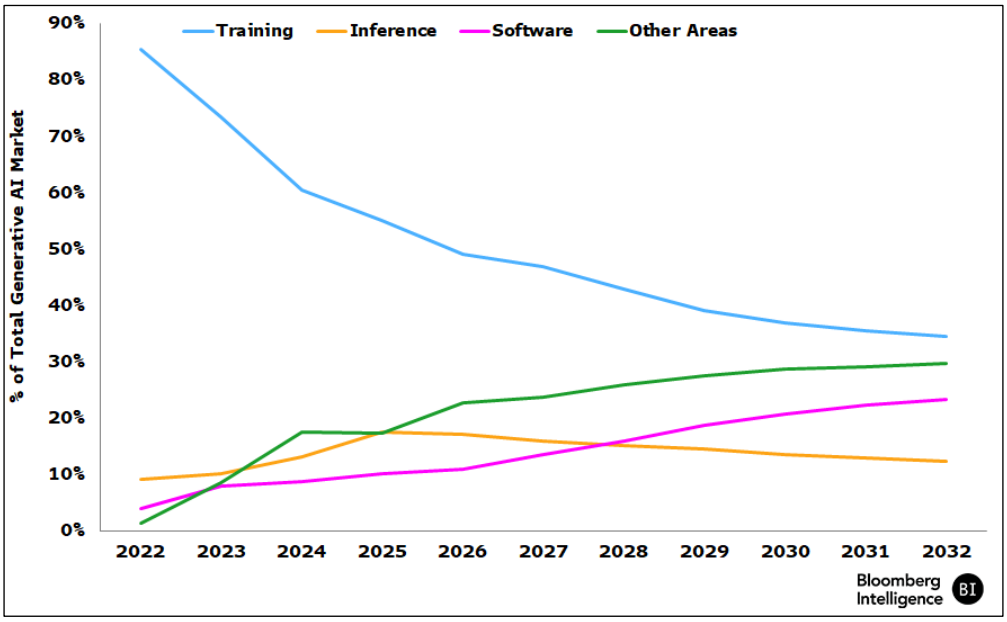
As for inference, computer vision and conversational AI products may emerge as new categories with the availability of LLMs for domain-specific predictions and applications. Such new items can help accelerate growth in the $1 trillion devices market, which has blossomed with smart speakers and wearables.

Figure 3: Training vs. Inference Forecasts



Source: BI's forecasts based on hardware and software data from IDC

Figure 4: Generative AI Market Share



Source: BI's forecasts based on data from IDC, eMarketer, Statista

Within hardware, infrastructure expenditures (for training) will likely be about triple the size of those for devices (inference) as companies spend on servers and storage to manage the intensive workloads required. Nearly 75% of chief information officers in the US surveyed by Bloomberg Intelligence indicated plans to increase their IT-infrastructure budgets in 2024, with about 38% expecting to boost spending by over 11%. Nvidia and Dell may remain among the top server

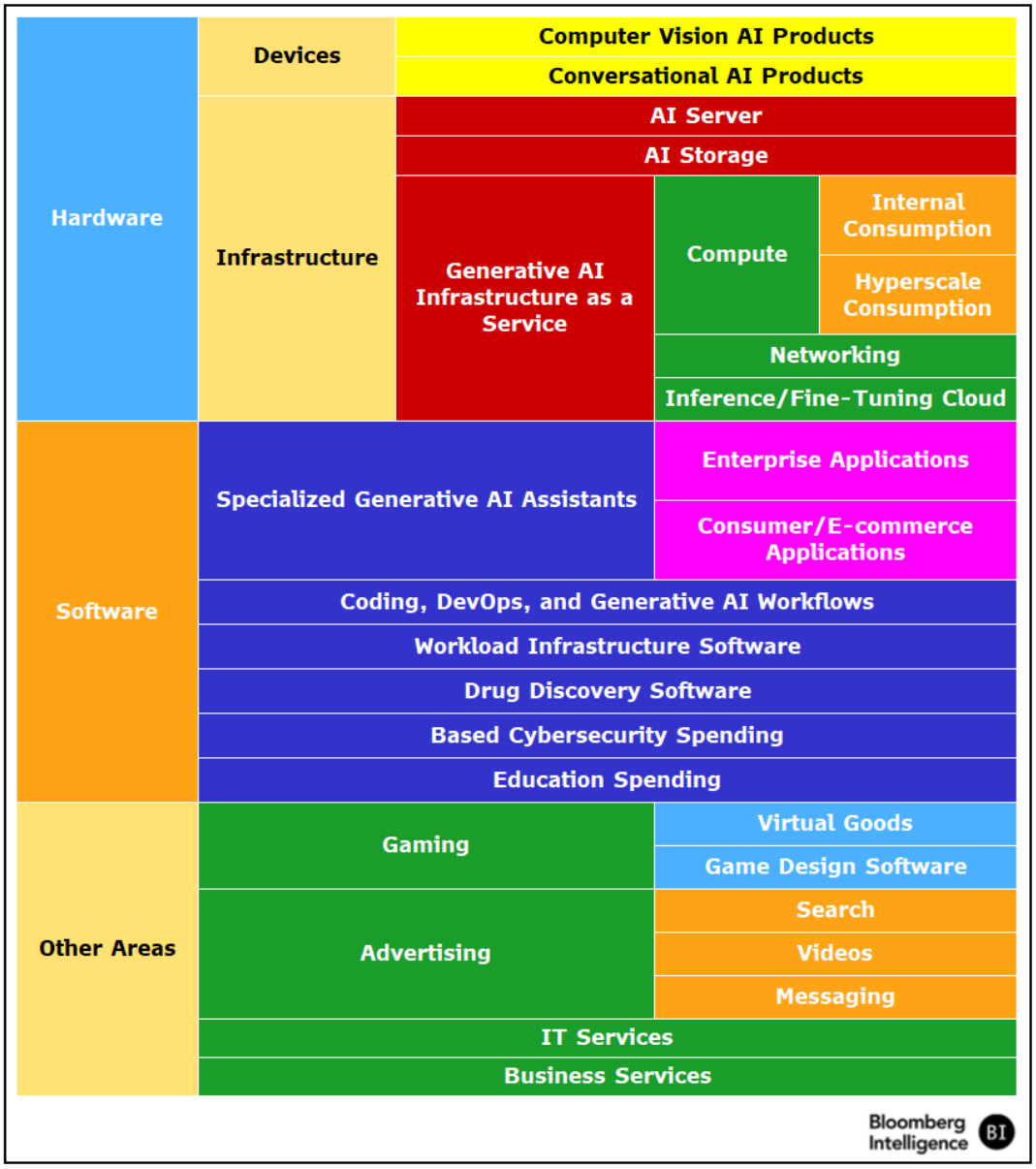
providers for generative AI workloads, with about 51% of CIOs selecting them. Generative AI infrastructure as a service (IaaS) will be key to training LLMs and could add $309 billion in sales over a decade. Within generative AI IaaS, the market for compute resources will likely be the largest at around $174 billion, driven by hyperscale consumption of $140 billion. Networking can add about $44 billion and fine-tuning clouds for specialized use may grow to a $91 billion opportunity. The market for computer-vision AI products is set to grow to $58 billion, while sales of conversational AI products could hit $110 billion. We expect that AI may add $639 billion to the total hardware market by 2032 from roughly $53 billion last year.

As for software, generative AI products may add about $318 billion in spending by 2032, growing at a compound annual rate of 71%. Cybersecurity, drug discovery, AI assistants and coding workflow should be top beneficiaries for generative AI expenditures. Many software peers will likely introduce their own AI copilots to enhance the user experience, with specialized assistant software poised to log $95 billion in sales by the end of the decade. Spending on educational software could be strong in an effort to improve existing learning tools and build new ones. We also expect generative AI to expedite development of gaming and creative software, reducing barriers to entry and creating opportunities for disruption.

On the internet side, generative AI can improve ad targeting and spur the creation of new formats to drive user engagement and increase conversion of ad views to sales. Large companies like Meta and Alphabet rely less on open-internet corpuses than other companies developing foundational LLMs, given their abundance of first-party data to deploy, along with robust capacity to spend that can help train models to improve ad targeting and efficiency. Such enhancements may drive an additional $207 billion for the digital-ad sector by 2032.

In IT and business services, we estimate that generative AI products and tools can add about $113 billion in sales as companies look for new products to drive top-line growth and trim unnecessary costs.

Figure 5: Generative AI Market Overview



Source: Bloomberg Intelligence

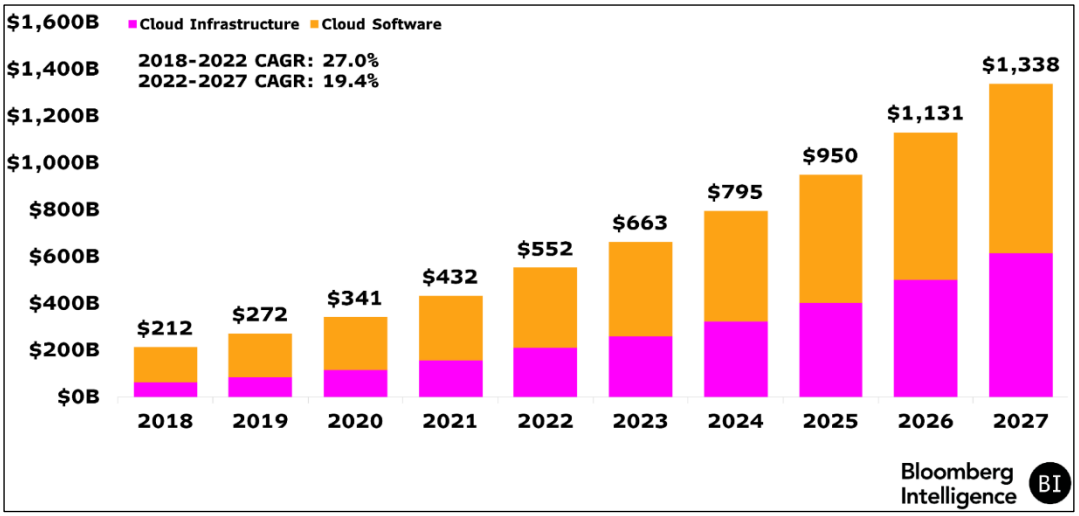
3.2 Cloud to Overtake Server Development

Though servers and storage could be the most prominent segments for generative AI services in the near term, many enterprises doubtlessly will leverage public cloud deployment eventually. We believe hyperscalers will develop in-house foundational LLMs, which will work best on their own cloud infrastructures. Meta, Microsoft, Alphabet, Nvidia, Amazon and other such suppliers may be among the main facilitators for training LLMs. These companies have access to the capital needed to set up the training infrastructure while keeping usage high for their servers to sustain healthy profit margins.

In time, generative AI as a service should be a much bigger market than servers and storage, logging 54% growth compounded annually through 2032 as gains for stand-alone servers and storage taper off. The trend favors expansion for hyperscale cloud suppliers over smaller infrastructure-software peers, mirroring the evolution of the software-, platform- and infrastructure-as-a-service portions of the roughly $663 billion public cloud market.

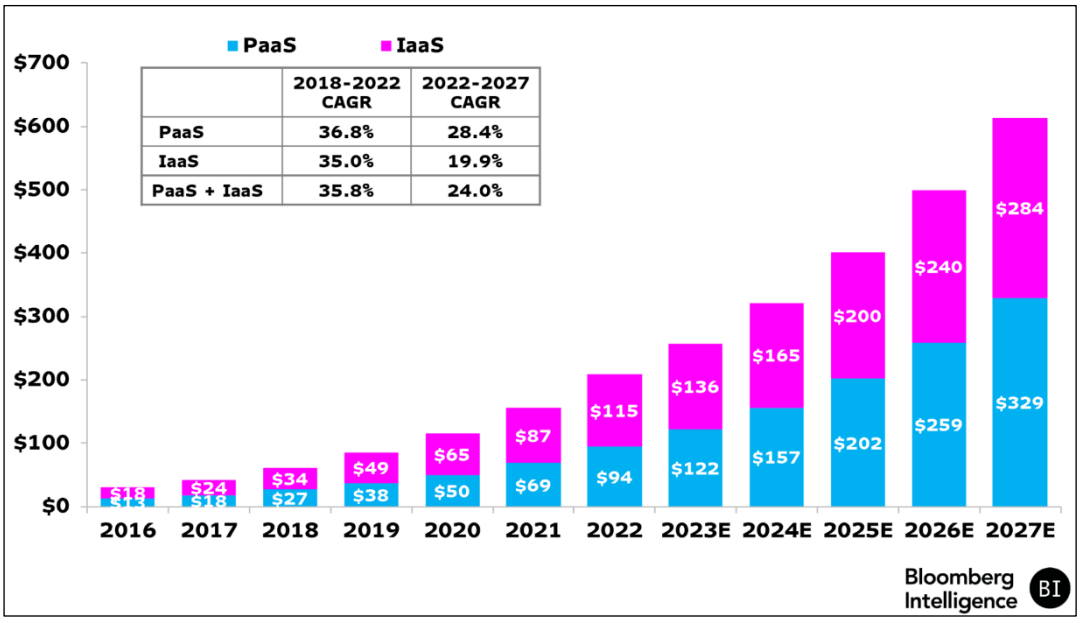
The forecasts in the accompanying graphic are conservative. Though it's highly likely that the enterprise demand shift toward the cloud will gather speed in the coming years, that isn't included in our assumptions.

Figure 6: Total Public Cloud Spending Forecast ($ Billion)



Source: BI's forecasts based on hardware and software data from IDC

Figure 7: IaaS, PaaS Revenue Forecast ($ Billion)

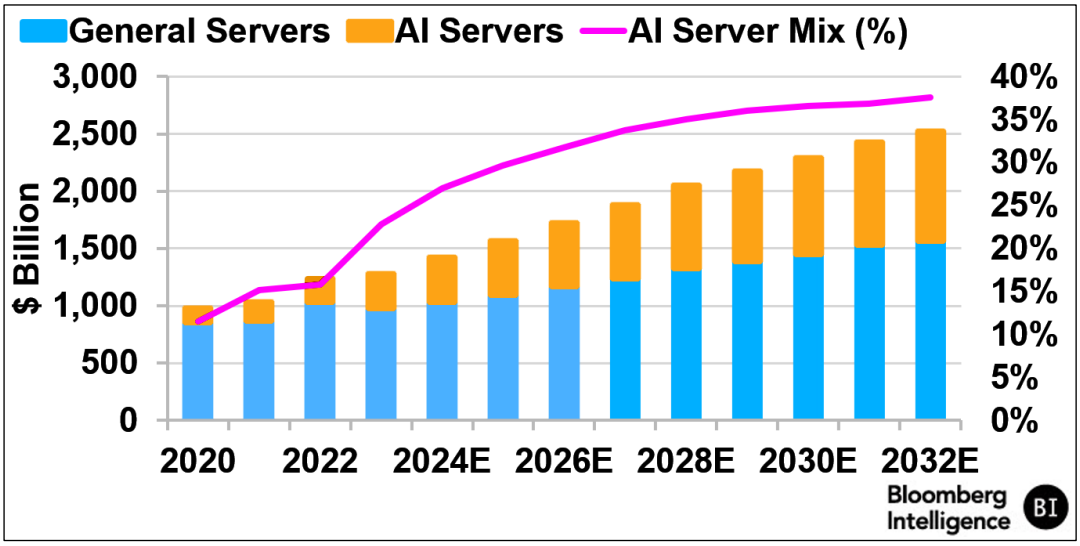


Source: BI's forecasts based on hardware and software data from IDC

3.3 Near-Term Server Demand Should Stay Healthy

The eventual shift to cloud deployment notwithstanding, the explosive demand for generative AI – as evidenced by the insatiable demand for Nvidia GPUs – should fuel significant growth in the infrastructure hardware market, particularly servers, that provides the necessary computing power. The global market for AI servers is set to roughly double to $39.4 billion in 2024 from 2022, according to our estimates, notching 41% average annual growth. AI is poised to contribute more than 20% of global server revenue starting this year, from 15% in 2021. Despite economic headwinds in 2023, spending on AI servers could remain robust, thanks to the ChatGPT-fueled arms race in generative AI.

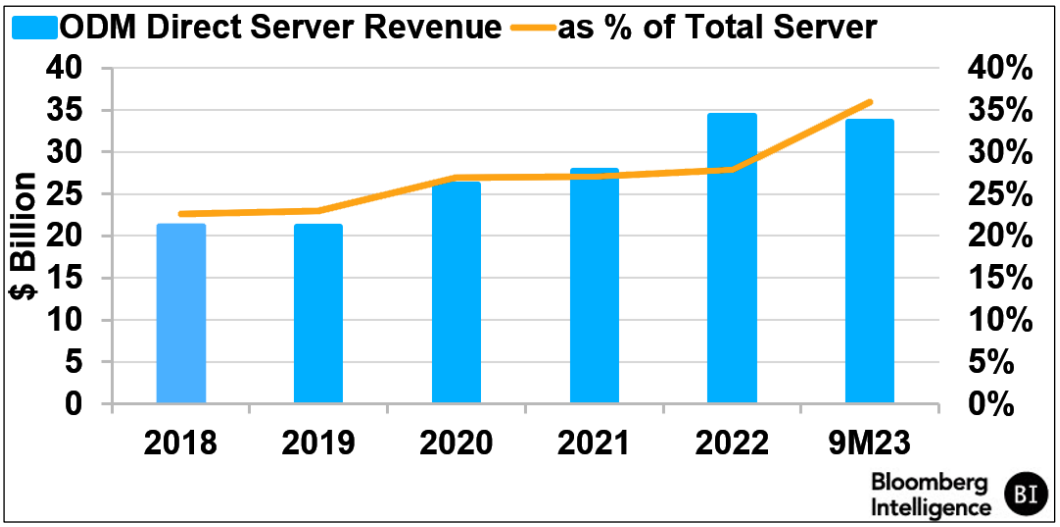
Figure 8: Worldwide AI Server Market Forecast



Source: BI's forecasts based on hardware and software data from IDC

The lion's share of server demand might go to original design manufacturers (ODM) building customized models for major cloud service providers like Microsoft and Google that are providing significant backing and development for AI applications. Their public cloud infrastructures also offer the necessary scalability for AI development in terms of computing and storage capability. Microsoft is a key investor in OpenAI, the owner of ChatGPT, and Microsoft Azure is the exclusive cloud platform for ChatGPT. Wiwynn, a major ODM server maker based in Taiwan, indicated that AI servers accounted for 20% of its revenue in 4Q23. It expects that revenue contribution to rise further in the coming quarters, given the steady launches and ramp-up in shipments of new AI server projects.

Figure 9: ODM-Direct Server Market



Source: IDC

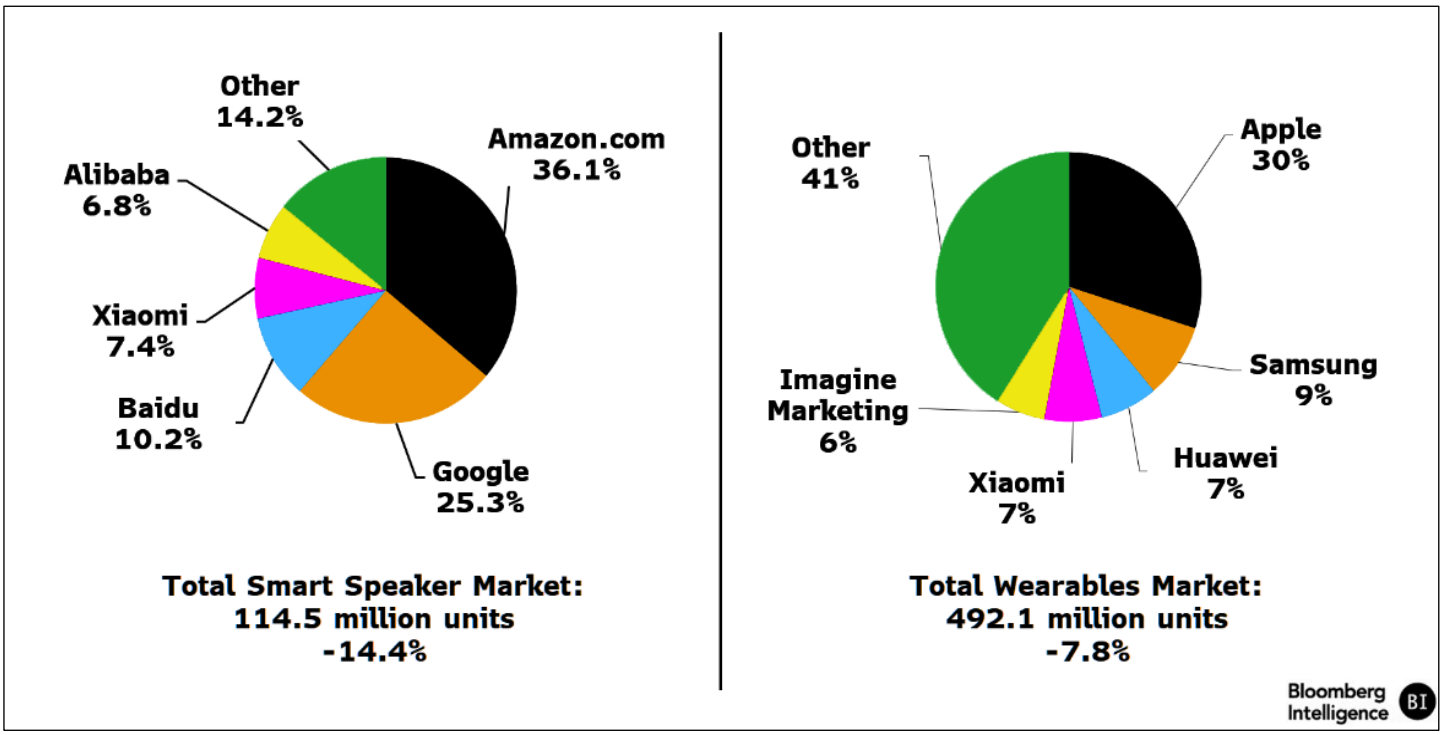
The complex designs of AI servers may help boost profit for related suppliers. While typical servers mainly use Intel and Advanced Micro Devices' x86 central processing units, AI servers use so-called heterogeneous computing architecture, which requires the combination of different processors – such as CPUs, graphic processing units and advanced RISC machine-based (ARM-based) chips – or proprietary application-specific integrated circuits. That mix-and-match approach can optimize system performance and power efficiency but poses a challenge to server designs as each processor has different instruction sets and data transmission cycles. As a result, ODMs with advanced design expertise would have a competitive edge over rivals that don't and might be able to charge more, expanding profitability.

The shift from using general-purpose CPUs to custom accelerators for large dataset workloads is key to why training is poised to be a bigger market (accounting for 35% in 2032) than devices (12%) in generative AI. The use of semiconductor accelerators will likely increase as more companies train their own LLMs, similar to OpenAI's GPT, Meta's Llama and Alphabet's Gemini.

3.4 Faster Hardware Refreshes; Networking a Key

The need for inferencing on so-called edge devices (hardware that controls data flow across the boundary between two networks) could speed refreshes of personal computers and smartphones – which currently aren't well-suited for the heavy processing, memory and storage requirements for AI's LLMs – while spawning new categories beyond wearables and smart speakers. Demand for inference is expected to ramp up as more applications are developed on top of foundational models such as OpenAI's GPT, Google's Gemini and Meta's Llama. We expect more compact versions of these models, such as the Gemini Nano, to be released in the next 1-2 years, allowing edge devices to perform AI workloads natively. These models will likely require less computational power and be less expensive to train.

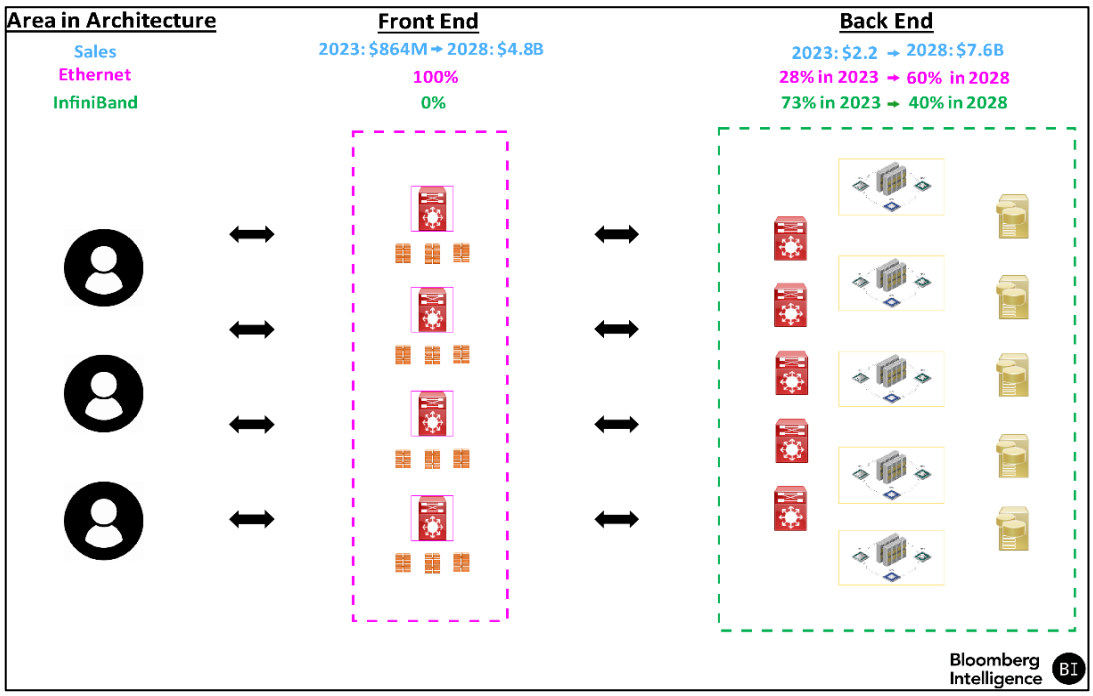
Figure 10: Smart Speakers, Wearables Market 2022



Source: IDC

Networking has emerged as a strategic area of AI infrastructure, along with servers and graphics processing units. It's typically been a bottleneck for hyperscale cloud infrastructure, which companies have aimed to resolve with higher capacity gear. Public cloud generative AI workloads are expected to grow faster than general cloud due to the computing intensity needed to ingest a burgeoning amount of structured and unstructured data for LLMs, supported by clusters of GPU-powered servers that can number in the tens of thousands. This complexity and density of AI architecture requires a separate "back end" AI network that supports high capacities – 800 gigabits or greater – while delivering low-latency, error-free data traffic separate from the general purpose, user-facing "front-end" networks. Given the rapid rise of AI architectures, roughly 18-20% of total Ethernet data center ports are projected to support AI traffic by 2028. AI networking will likely be a $12 billion market within the next 5 years, from $3 billion in 2023, according to our market forecast.

Figure 11: AI Network Architecture Overview

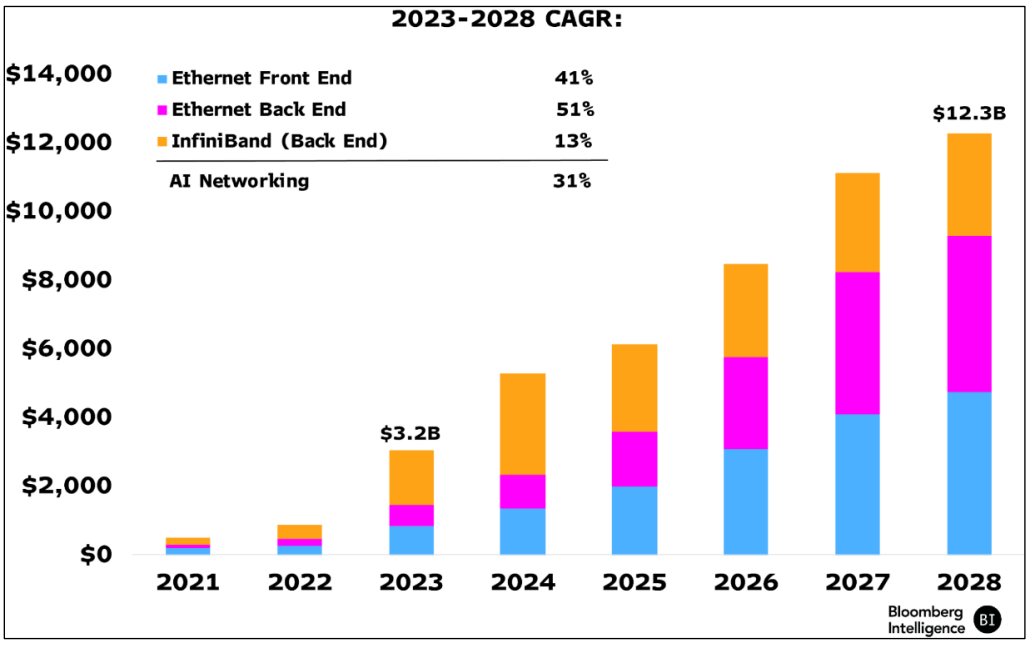


Source: Bloomberg Intelligence

Today, InfiniBand has emerged as the preferred connectivity technology for backend networks, contrasting the ubiquity of the Ethernet protocol powering most of a cloud and corporate data center network. InfiniBand has lineage to high-performance computing and supercomputing environments and the ability for the technology to reliably transport data at high speeds with little data loss, allowing the proprietary technology to account for 73% of AI back-end networking sales in 2023.

Nvidia accounts for nearly all the InfiniBand market through its 2020 acquisition of Mellanox. It leveraged leadership in AI infrastructure – software, GPUs, data processing units (DPUs) and interconnects – to bundle InfiniBand as an integrated product. While Nvidia's bundled approach sharply contrasts with hyperscale clouds' disaggregated approach, a tightly integrated, turnkey AI system could fit well with enterprise infrastructures over the long term.

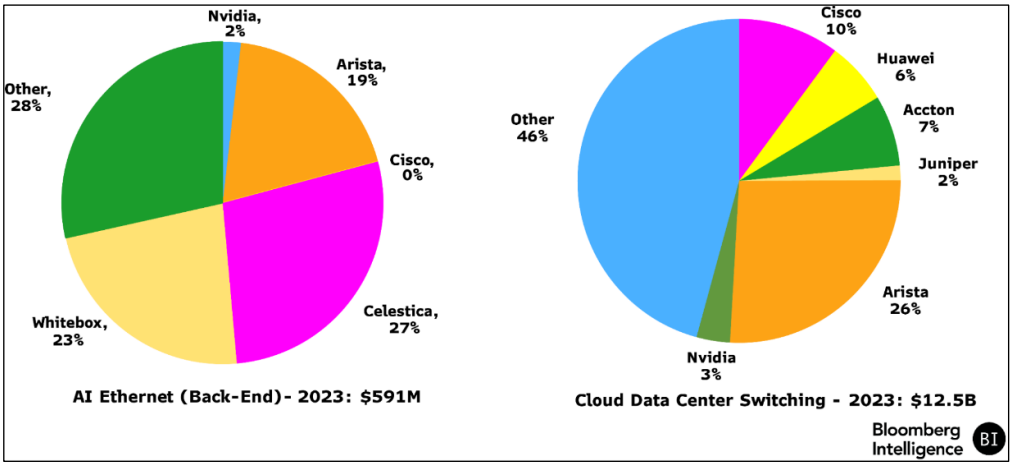
Figure 12: Cloud AI Networking Sales Forecasts



Source: 650 Group

Yet interest in Ethernet technologies is high, especially among cloud customers, and is enhanced by a growing ecosystem of products led by the Ultra Ethernet Consortium. As a result, back-end AI-related Ethernet sales are expected to grow at a 50% compound annual rate for 2023-28 to $4.6 billion. The latest chip and hardware innovations solve the "bursty" and "lossy" nature of Ethernet data traffic, which could make it more attractive than InfiniBand. Hyperscale cloud interest in adopting Ethernet may be high in part because of familiarity with the technology but also because it allows each cloud to create differentiated AI architectures and helps avoid lock-in with the Nvidia ecosystem. Arista's strength in high-speed networking gear positions it well to be a leading beneficiary of the shift toward Ethernet by cloud providers. Nvidia meanwhile is well situated to shift customers to Ethernet for AI networks, thanks to the Spectrum switching gear it gained through its Mellanox acquisition.

Figure 13: AI Ethernet, Total Cloud Switching Market Share



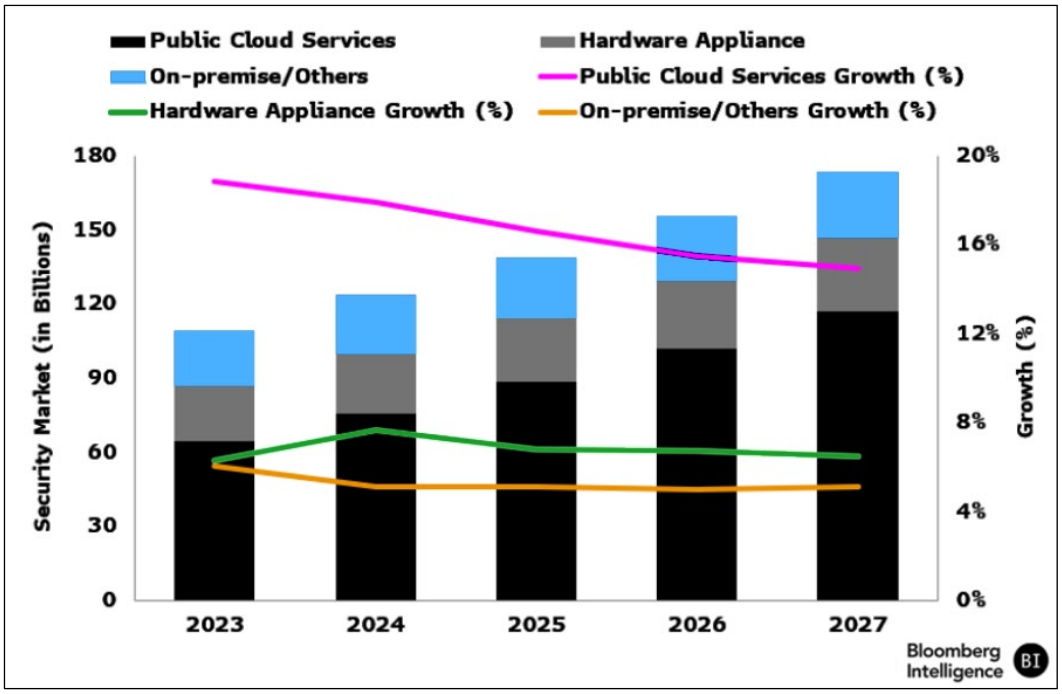
Source: 650 Group, Dell 'Oro

3.5 Digital Transformation Initiatives Spill Over

With the rapid development of cloud-based AI technologies like ChatGPT, the significance of edge AI development is growing and represents a significant step in advancing the AI ecosystem. Edge AI is particularly valuable for real-time decision making and cost savings, which are crucial in areas such as health care, manufacturing and transportation and could lead to a larger user base than cloud-based AI. Our scenario analysis finds that the edge AI semiconductor market could be as much as 3.37x the size of the cloud-based AI market by the end of 2032. Adoption of edge AI can drive significant growth in uptake by the consumer (projected to lead other segments with a 10-year CAGR of 39%), industrial and automation sectors over the next 10 years.

Beyond generative AI, advancements in machine learning and other aspects of artificial intelligence appear likely as well. Oracle has been pushing its autonomous databases for the last few years, which could gain from increased budget allocations to AI. We expect increased availability of such features from other software providers, too, where product patching, security updates and tasks usually assigned to a database administrator are automated using machine learning. It could play a much bigger role in cybersecurity in the coming years, as well, especially in event management, in analyzing irregular patterns inside organizations and in the form of copilots to augment security professionals and boost automation.

Figure 14: Market Size by Deployment Type



Source: IDC

Section 4. Market Disruption

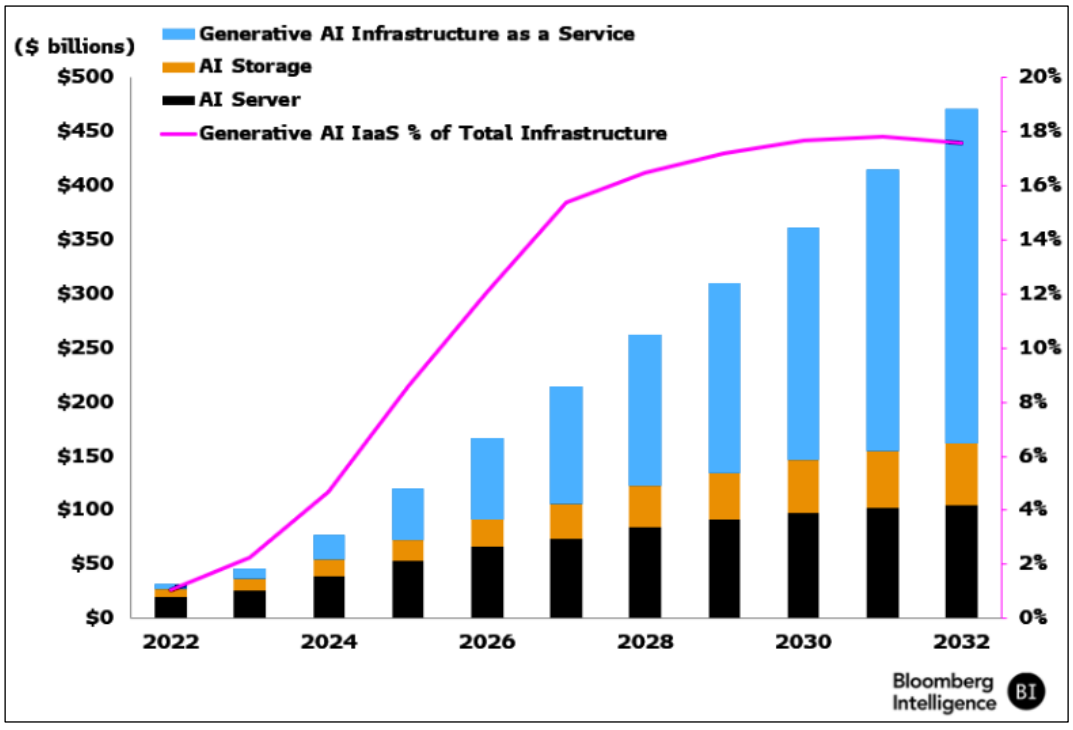
A Shift Coming in Hardware, Ads, Gaming

Generative AI offers opportunities for disruption, especially within hardware, digital advertising and gaming. The computational intensity of training large language models may spark a market-share shift toward advanced RISC machines, potentially making this category the fastest growing within hardware. Alphabet, Meta and other digital-ad giants can improve targeting and brand conversions by implementing machine-learning models based on their vast library of first-party data. Sony, Google, Unity and others in the gaming segment may leverage AI to facilitate development and make the user experience more engaging.

4.1 LLM Training Favors Move to Accelerators

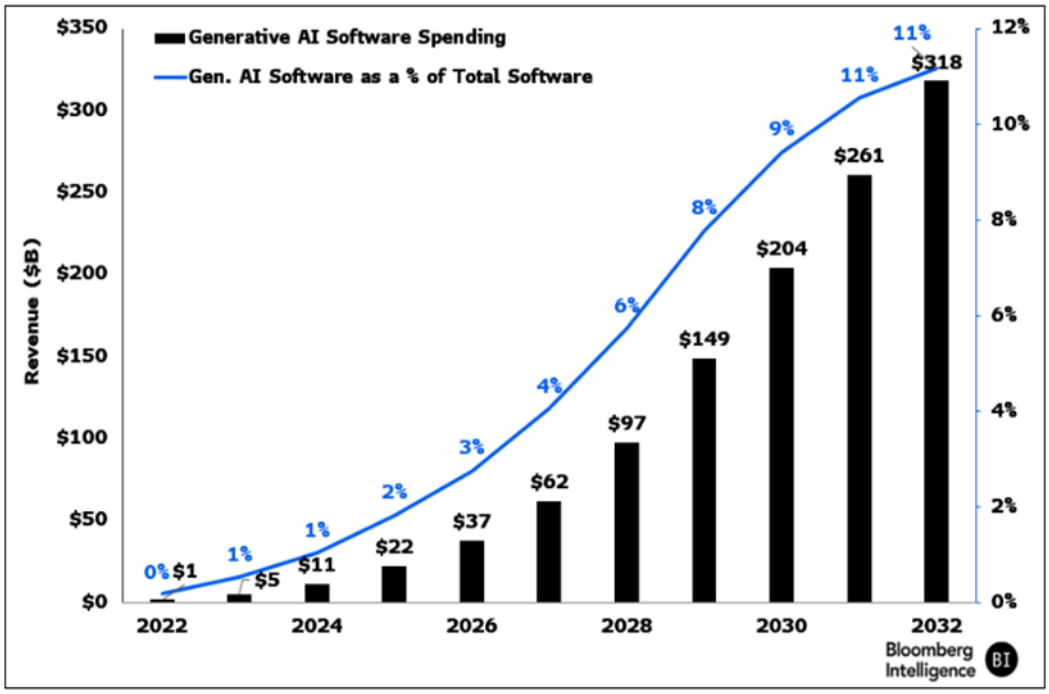
Training large language models could result in big market-share shifts to accelerators based on ARM and RISC instruction sets at the expense of servers based on traditional x86 CPU architecture. A surge in demand for AI servers has made ARMs the fastest growing architecture in AI accelerators, and we believe generative AI as a service will likely be much bigger over time as enterprises leverage the public cloud for deploying LLMs and other forms of advanced AI.

Figure 15: Generative AI as a Service



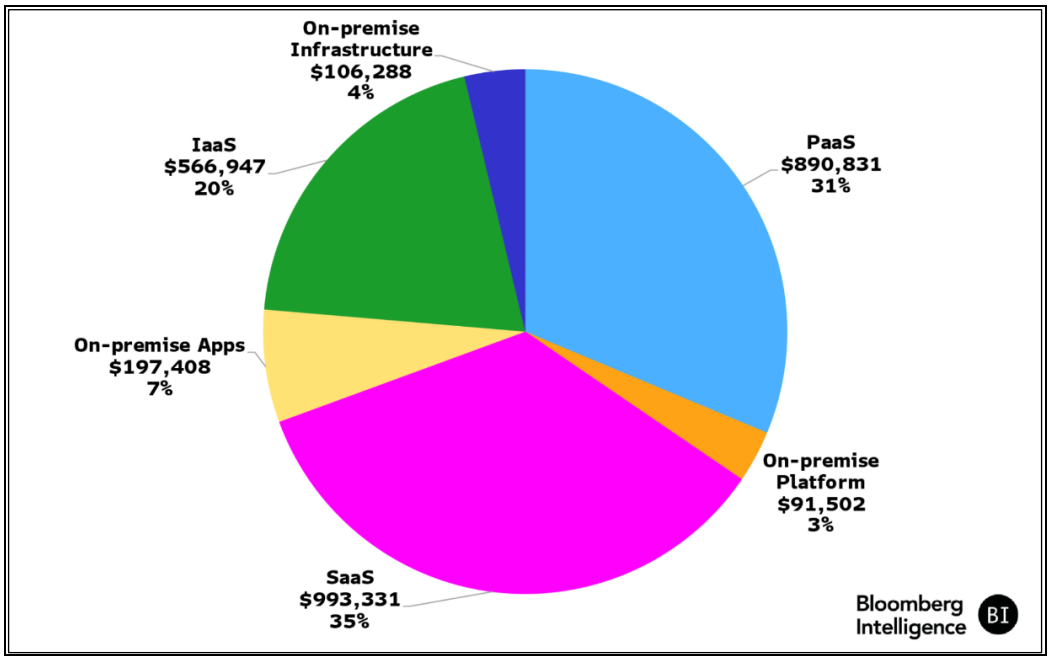
Source: BI's forecasts based on hardware and software data from IDC

Figure 16: Generative AI Software Spending Forecast



Source: BI's forecasts based on hardware and software data from IDC

Figure 17: Software Spending Forecast Breakdown, 2032



Source: BI's forecasts based on hardware and software data from IDC

4.2 Ad Leaders Adobe, Salesforce Aided by First-Party Data

LLMs have extensive computing and storage needs, central reasons that we expect the first phase of experimentation to be performed with hyperscale cloud providers like Google, Microsoft and

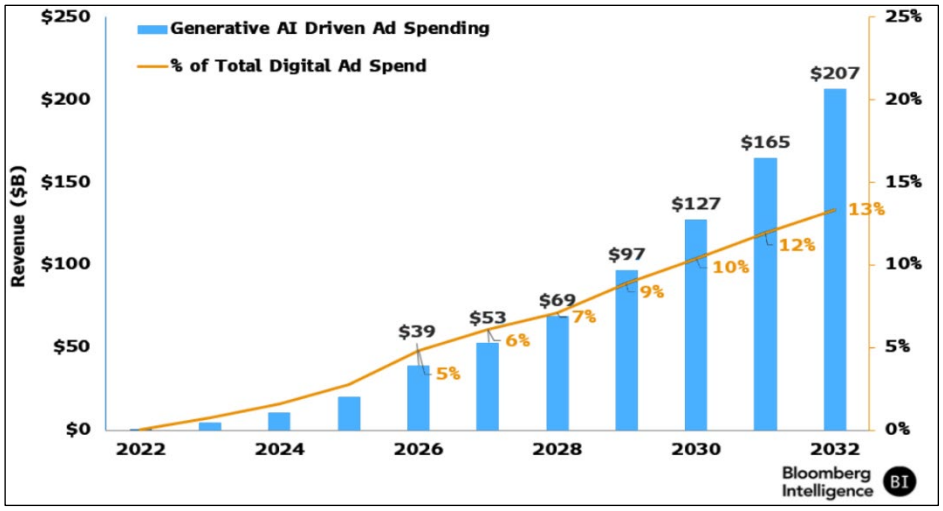
Amazon Web Services. Even at maturity, such companies are likely to gain the most market share, given the scale and cost needed to develop an infrastructure in-house.

OpenAI's lead in training transformer models and its early partnership with Microsoft has aided ChatGPT's adoption over text-based LLMs from Meta, Amazon, Alphabet and Anthropic. With most hyperscalers investing in developing their own foundational LLMs, we believe OpenAI will need to maintain its lead on the algorithm side while ensuring access to training data from the open-internet corpuses from companies such as Wikipedia, Reddit and Stack Overflow. Alphabet's merger of its DeepMind and Google Brain AI units, can spur faster changes that leverage LLMs to maintain user engagement across revenue generators, like its Search, Chrome and Maps applications. Amid the rapid shift to generative AI, large players like Meta, Adobe, Microsoft, Alphabet and Salesforce are better positioned than smaller rivals for two reasons: they have scads of first-party data in hand and ample ability to deploy capital. Each has a leading market share in its category, offering access to large amounts of information to train AI models, driving more accurate and efficient results.

Social-media platforms like Meta should see a lift as AI-generated content rises rapidly, helping to boost engagement and monetization. LLMs and generative AI can accelerate the shift to digital ads from linear TV. We calculate that more time spent online, coupled with ad targeting and personalization could add over $200 billion to the market through 2032. Conversion rates for ads on these platforms might be aided by the growing capabilities of LLMs, which should favor companies with strong presences in cloud infrastructure and that have the most first-party data.

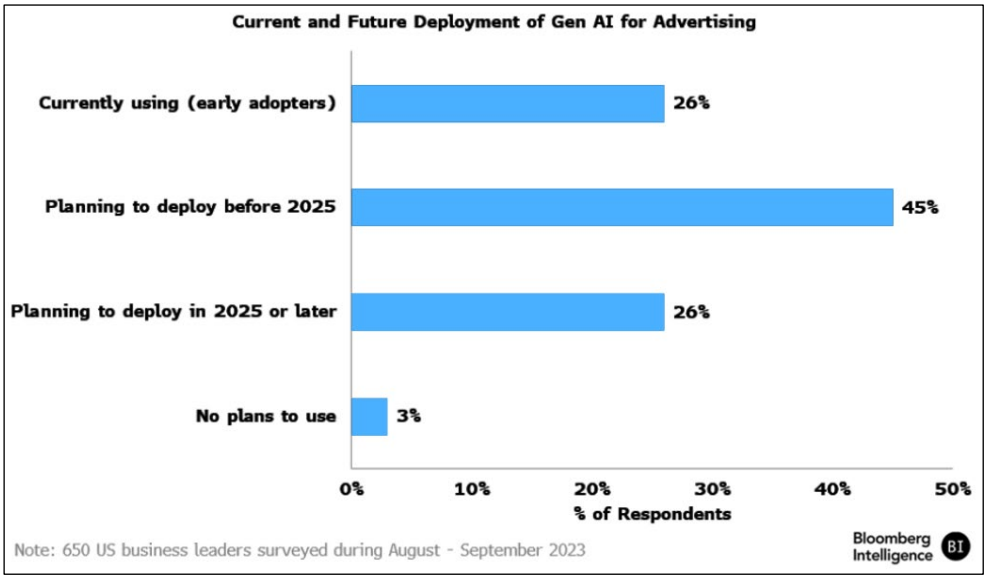
Given the high costs of generative AI infrastructure, OpenAI, Anthropic and Google currently have paid subscriptions for their LLM offerings for heavy users. An ad-supported model is unlikely to be profitable for online search and new tools that leverage deep learning and generative AI due to the higher costs of LLM-based queries. A recent Bloomberg Intelligence survey found that only 13% of respondents were willing to pay for a subscription to use a generative AI tool like ChatGPT. Of those, just 1% said they would spend $20 a month for ChatGPT, with the rest amenable to $6-$10. Among all participants, 93% indicated that they wouldn't lay out more than $10 monthly. The results suggest that lower prices could help boost adoption by 10x for a generative AI subscription. As an illustration: though a freemium version of ChatGPT helped propel it to 100 million monthly active users faster than any consumer app at the time, its conversion to paid subscribers remained in the low single digits.

Figure 18: Generative AI Digital Advertising



Source: BI's forecasts based on digital advertising data from eMarketer

Figure 19: AI Current and Future Advertising Usage



Source: eMarketer

4.3 Sony, Google Parlay New Interfaces for Gaming Design

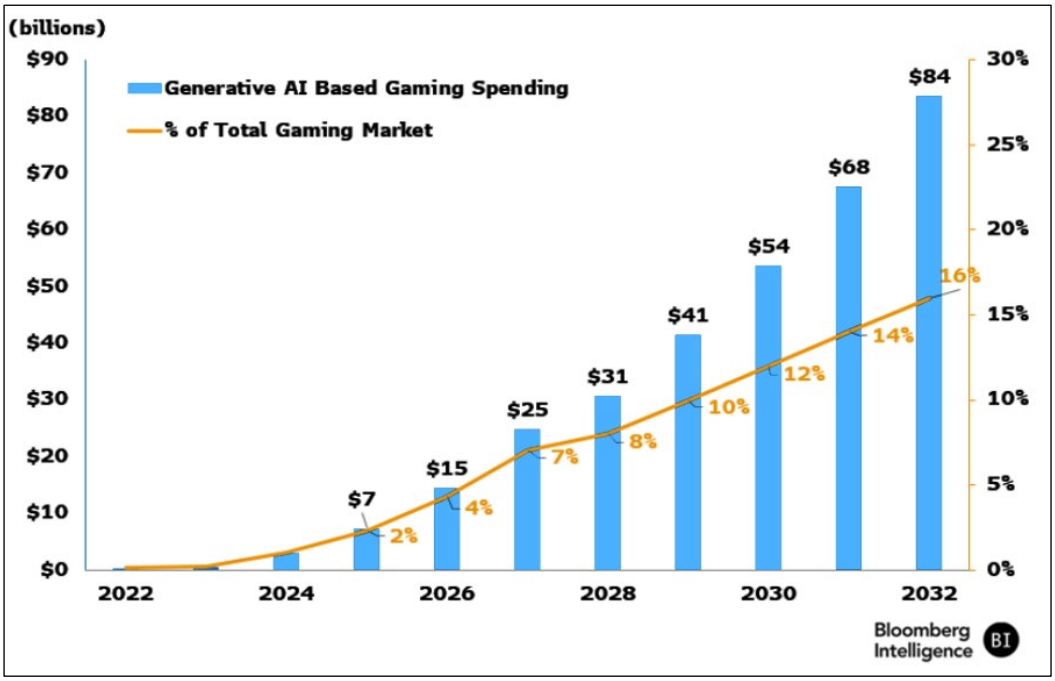
Some startups have already showcased interesting techniques to produce synthetic content — computer-made data that mimics real-world information — based on descriptions and the large amounts of training data available on the open internet. Generative AI can speed that creative process for mobile games, social media and virtual- and augmented-reality applications.

AI tools may rapidly increase gaming data available beyond the high-budget, high-profile major publishers, including those made by users. Developers remain key to gaming and the metaverse beyond the foundational models provided by tools such as OpenAI's GPT, Google's Gemini, Meta's Llama and Anthropic's Claude. Apple and Google's Android, as well as gaming ecosystems like Sony's PlayStation could offer software development kits to leverage LLMs to

ease creation of new content on their platforms. Generative AI might help creative software tools shift to description- and voice-based content from point-and-click user interfaces.

Though Google and Meta have developed LLMs for image generation, they have trailed Stability AI, Midjourney and OpenAI's Dall-E in terms of adoption. OpenAI just released its video-based LLM, Sora, as foundational models become multimodal. Most image-based generative models rely on a diffusion technique and the quality of images rendered depends on training data and the weights assigned to the parameters used. While Adobe has been investing in developing its own generative AI capabilities with the introduction of its Firefly offering, we expect other design- and gaming-software companies to invest in their own generative AI models to leverage proprietary data and distribution.

Figure 20: Generative AI Gaming



Source: BI's forecasts based on hardware and software data from IDC

Section 5. Segment Analysis

Popularity Hinges on Adoption in Existing End Markets

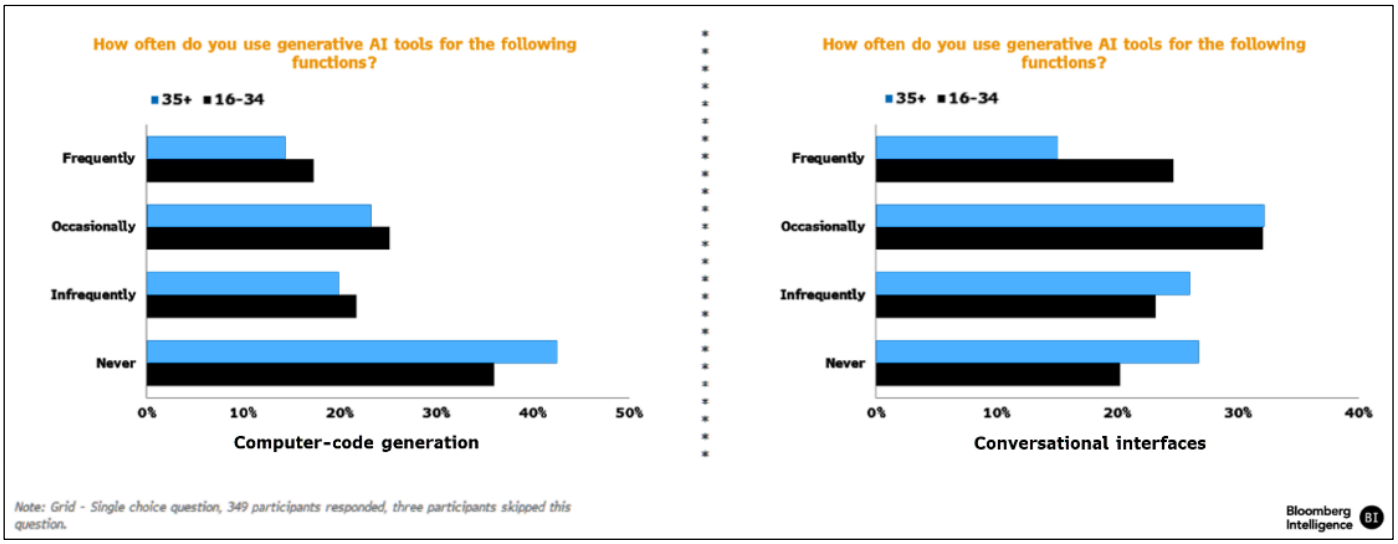
Generative AI is likely to have a far less pronounced effect on the application software industry than infrastructure software in terms of new revenue generated. Within application software, however, we're already starting to see the rise of AI copilots, with companies like Microsoft, Adobe, Snap and others introducing their own versions of this technology in recent months. Bloomberg Intelligence market opportunity analysis shows that a majority of the $318 billion of new software sales tied to generative AI is likely to fall in the infrastructure bucket.

5.1 Copilots Lead the Way Into New Endeavors

Education, drug discovery and specialized AI assistants could be the bigger contributors to new revenue streams within application software. Gaming, IT and business services may be smaller categories. The customer-service and business-process outsourcing subsegment of business services might be heavily affected by AI tools and sales could shrink.

Microsoft 365 Copilot and Adobe's Firefly are examples of two generative AI assistants in application software. Though we don't see them fueling rapid growth in new users, given the high penetration of the applications, it's likely that average revenue per user will increase as the products become stickier to use.

Figure 21: Coding vs. Conversational Interfaces



Source: Bloomberg Intelligence

Over the past few months, we've observed AI assistants, copilots and chatbots deployed into three key areas: software development, core software products and customer service. We view copilots deployed to software development to have the greatest near-term impact to boost coding productivity, helping corporations that are grappling with a shortage of developers. Microsoft's GitHub Copilot has emerged as the leader in this area, with business and enterprise

SKUs priced at $19 and $39 a user per month, respectively. These tools have potential to substantially reduce coding time by recommending lines of code and spotting errors.

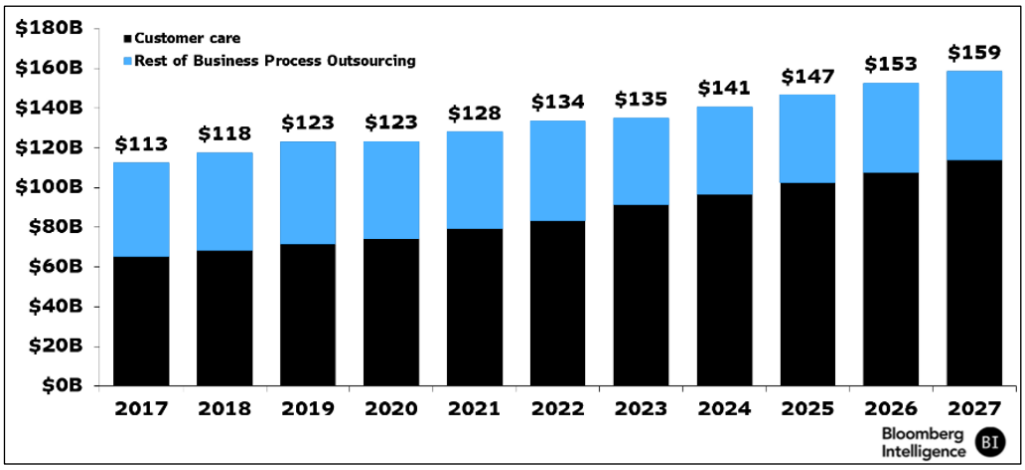
Other software companies like Adobe and Salesforce have embedded generative AI within their core software to make it usable to their end customers. Such copilots can recommend actions based on available data, answer questions and generate output in the software. Examples include expanding an image through text or creating a PowerPoint presentation from a Word file.

Many software companies are also experimenting with different ways to monetize generative AI, especially as the cost to run these workloads is higher than traditional AI. The most common pricing model we've observed is generative AI products are priced out separately. Examples include GitHub Copilot noted above in addition to Microsoft 365 Copilot ($30 a user per month) and Salesforce Sales GPT and Service GPT ($50 a user per month).

Generative AI has also led to broad-based subscription pricing increases in cases like Adobe who lifted pricing for all Creative Cloud plans by roughly 9-10%. In other cases, such as ServiceNow, pricing for its generative AI Pro-Plus SKU was based on customer savings due to AI, in which ServiceNow only reclaimed 10% of the perceived customer savings, equating to roughly a 60% price uplift over their Pro SKU.

Business-process outsourcing services may be more heavily disrupted than IT services, with jobs in customer service and back-office areas displaced by AI assistants. That could lead to near-term pricing pressures, particularly in customer care, which sits at the bottom of the BPO value chain yet is still the largest and fastest growing of its segments. Customer care is forecast to expand 6.5% annually through 2027, compared with 3.5% for all other BPO services, according to IDC.

Figure 22: Business Process Outsourcing Forecasts



Source: BI's forecasts based on hardware and software data from IDC

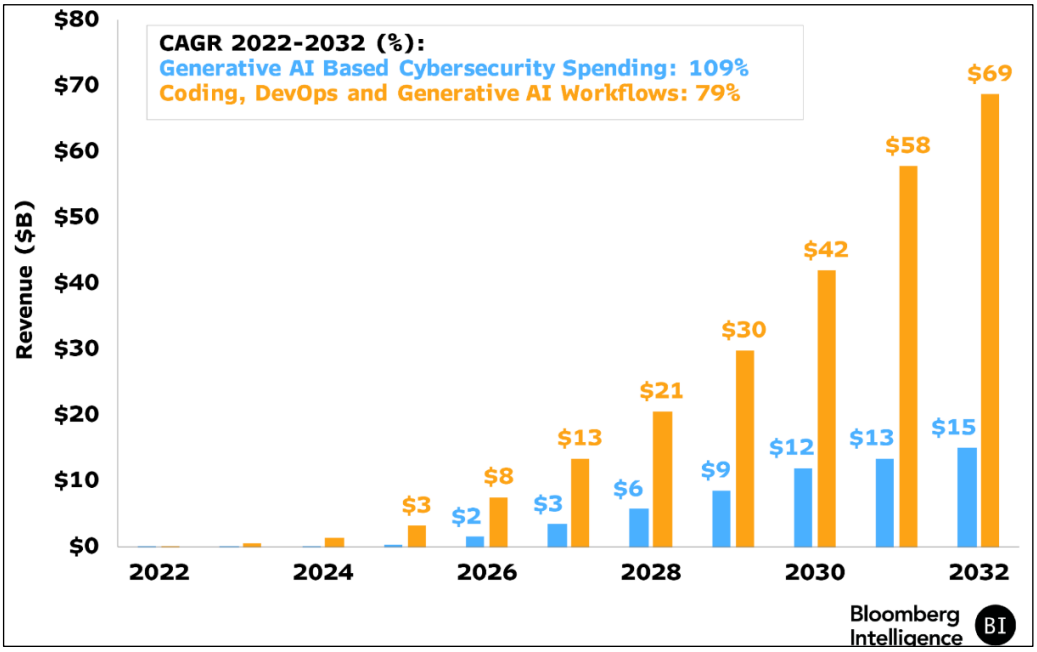
Generative AI will be more of a revenue tailwind for BPO companies that have minimal exposure to customer-care services, like Genpact and EXL Service. Companies that specialize in greater value-added services might leverage efficiency gains from AI to expand their total addressable markets, particularly in areas like data analytics.

5.2 Bolstering Cybersecurity While Improving Retention

The advent of generative AI led to an increase in the number and sophistication of cyberattacks, calling for more automation and efficiency on defense in an industry already struggling from alert fatigue and a lack of skilled talent. While generative AI or large language models are unlikely to drive stand-alone revenue opportunities in cybersecurity and DevOps, the use of copilots coupled with the growing efficacy of pure-play cloud suppliers' offerings could aid gross retention and upselling, which typically is lower in cybersecurity than in other software segments.

The availability of large language models that can ingest extensive amounts of telemetry and threat data from structured and unstructured sources may increase the effectiveness and boost the positions of cloud providers that have proprietary data and were already toward the top of the segment, such as CrowdStrike, SentinelOne and Zscaler. CrowdStrike's launch of a security copilot and its expanded alliance with Amazon Web Services are aimed at using generative AI to improve product effectiveness and gain an early lead in applying the technology to cybersecurity. Sentinel One, Microsoft, Palo Alto Networks and Check Point also launched or are in the process of launching gen-AI based copilots.

Figure 23: Generative AI Cybersecurity, DevOps



Source: BI's forecasts based on hardware and software data from IDC

Though Microsoft is a formidable competitor, the rapid growth of generative AI may increase demand for pure-play cloud security suppliers like CrowdStrike, SentinelOne and Okta as they sit on large amounts of proprietary data and can offer service across multiple clouds, which helps address evolving threats and the potential use of generative AI by bad actors. Google can continue to use its purchase of Mandiant for boosting the security of its cloud offerings compared with hyperscale rivals.

5.3 Data Comes at a Premium

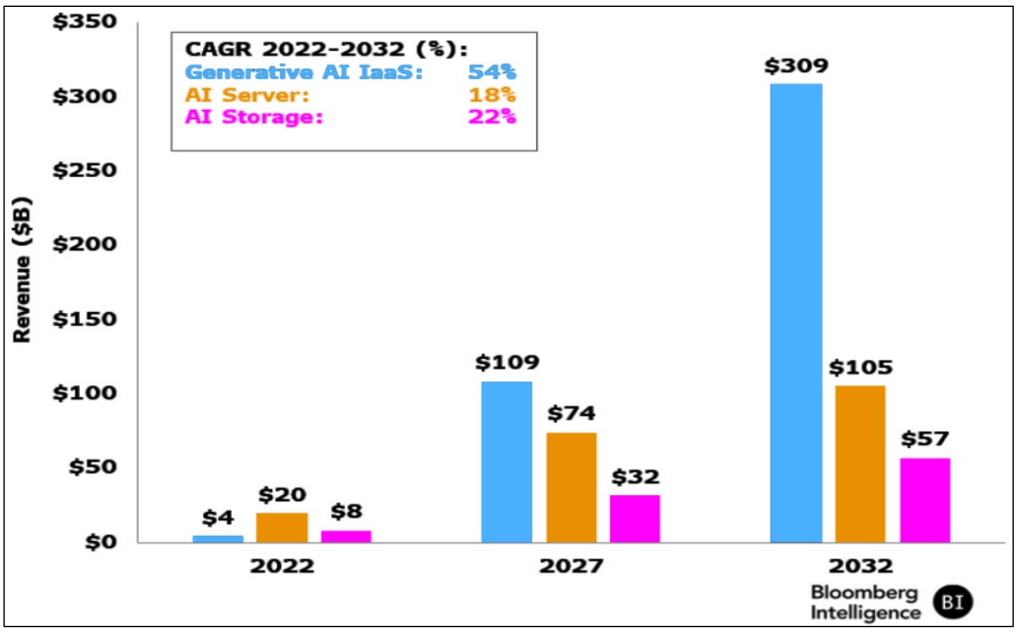
The size and complexity of large language models makes the training process extraordinarily data intensive. Though OpenAI's ChatGPT reached a partnership with Microsoft, it still could be disadvantaged compared with internet enterprises in the volume of training data available.

ChatGPT's initial application was mainly focused on changing the nature of search, which has been dominated by Alphabet's Google. ChatGPT's primary use was to analyze, generate and edit text based on user input. Yet within just a few months, OpenAI realized how powerful generative AI can be, and the platform's scope quickly expanded beyond traditional search. The latest version of ChatGPT can handle data including images, audio and video. Such inputs require vastly more computing resources than text-based LLMs.

The size and complexity of LLMs based on transformers architecture are likely to grow as a result of multimodal input, which can help hyperscale companies including Microsoft-OpenAI, Meta, Google and Amazon to maintain their lead over other foundational LLMs from peers. Given training of LLMs is a recurring process, we expect foundational LLM companies will consolidate to the ones that have a hyperscale infrastructure and large amounts of first-party data to improve the accuracy of the models.

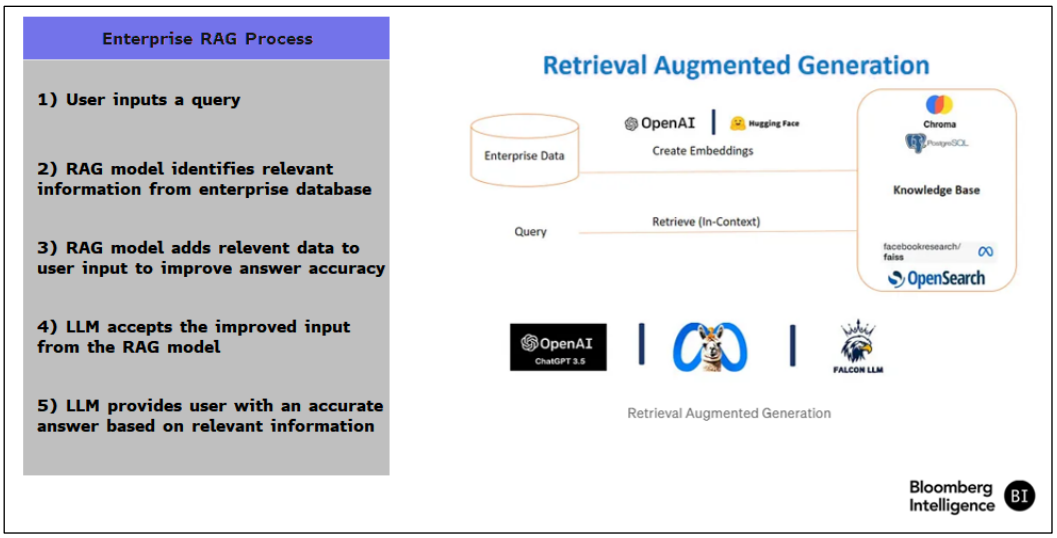
Retrieval-augmented generation (RAG) techniques will likely become key for developing enterprise chatbots. This approach features an additional model (RAG model) that is trained on proprietary enterprise data, adding an element of customization. When a user inputs a query, the RAG model can identify relevant information from the enterprise database and append it to the user query, which then gets passed on to the LLM. The additional context provided by the RAG model helps reduce irrelevant answers and hallucinations from the LLM. Given the recent scrutiny around Alphabet Gemini's image generator, we anticipate hyperscalers will have a heightened focus on reducing historical inaccuracies and questionable responses. As a result, these companies could be first in line to develop a RAG counterpart to their existing models.

Figure 24: Training Forecasts, 2022-32



Source: BI's forecasts based on hardware and software data from IDC

Figure 25: Retrieval-Augmented Generation (RAG) for Enterprises



Source: Medium, Amazon Web Services (AWS)

5.4 AI Fuels a Fifth of Global Server Revenue

Robust gains in the number of ChatGPT active users indicate that generative AI could be among the most important catalysts to growth for the server supply chain in coming years, driving over 20% of global server revenue by 2024 from 15% in 2021, by our calculations.

After its November launch, OpenAI's ChatGPT amassed a base of 1 million users in a week and exceeded 100 million in just two months. OpenAI introduced a subscription service at $20 a month and offered businesses paid access to ChatGPT to expand commercial applications. Companies including Snap, Shopify and Instacart already have integrated ChatGPT into their products.

The server supply chain's original design manufacturers could reap the most demand, since cloud service providers are integrally involved in AI development. AI servers could also drive robust sales for other suppliers with design expertise.

Figure 26: Asia's Major Server Manufacturers

Company	Ticker	Sales YoY			Net Income YoY			Remarks
		2022A	2023E	2024E	2022A	2023E	2024E	
Quanta	2382 TT	13.4%	-14.2%	38.9%	-14.0%	35.8%	29.3%	Management expects double-digit server revenue growth in 2023
Wiwynn	6669 TT	52.0%	-17.4%	43.7%	63.9%	-15.0%	53.3%	AI server revenue contribution could rise in 2024, from 20% in 4Q23
Hon Hai	2317 TT	10.6%	-7.4%	4.6%	1.6%	-5.5%	13.6%	23% of 9M23 sales from cloud/networking business
Inventec	2356 TT	4.2%	-5.4%	12.0%	-6.3%	-7.1%	35.3%	Management expects AI server shipment to double in 2024
Wistron	3231 TT	14.2%	-11.9%	12.9%	6.6%	2.8%	64.8%	Key supplier to Nvidia's GPU baseboards for AI servers
Lenovo	992 HK	-13.5%	-9.3%	8.2%	-20.8%	-41.7%	41.9%	15% of calendar 2023 sales from datacenter related hardware
IEIT	000977 CH	3.7%	1.2%	10.3%	3.9%	-21.0%	35.2%	Largest server vendor in mainland China

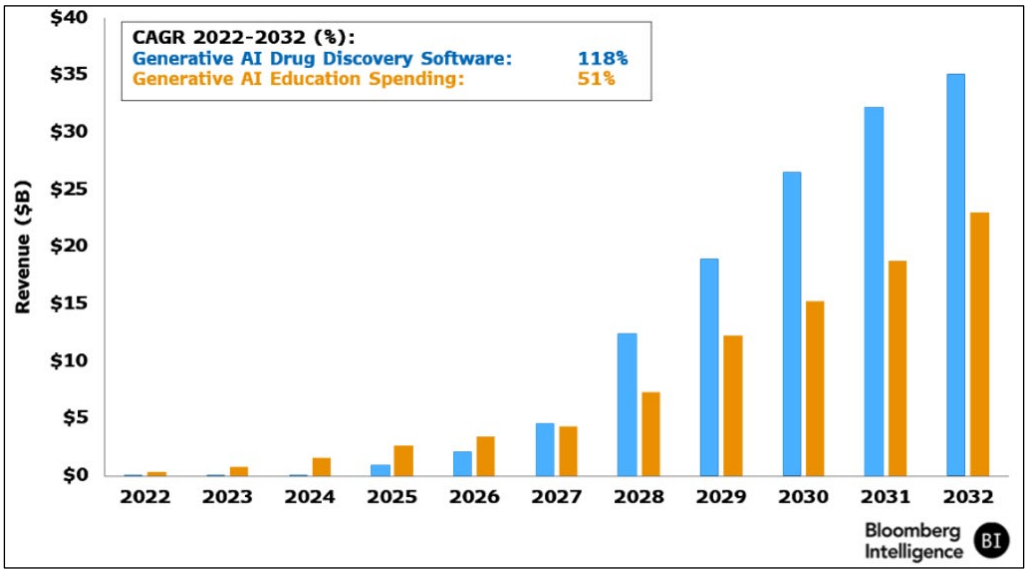
Source: Bloomberg Intelligence

Section 6. Expanding Uses

Spending on Services to Double Annually in Next Decade

Early uses of ChatGPT suggest that generative AI could significantly expand the vertical software market, which currently represents a small fraction of segments such as databases, enterprise resource planning and customer relationship management. Life sciences and education companies might notch rapid growth and emerge as pure-play vendors that benefit from the use of large language models in software, driving productivity. That's in addition to specialized AI-based software assistants that may transform search and other needs to summarize information.

Figure 27: Generative AI Life Sciences, Education Spending



Source: Bloomberg Intelligence

Figure 28: Companies Using Generative AI

	Examples on how companies are leveraging AI
Kroger	Deliver the appropriate promotional offers and discounts to customers at the right time
Target	Power product detail pages and provide more friendly and relevant explanations
T-Mobile	Spends around $2.5 billion on advertising annually, using AI to place ads and optimize media spend
eBay	Enhance sellers' product images to make them more compelling to customers
State Street Corp	Digitize and automate approximately 85% of bank loans settlements
Edison International	Improve inspections, customer experience, and grid planning; Also using AI for research, workflow automations, and code development
Hershey	Optimize the talent profile for certain key jobs
McKesson Corp	Improve patient intake and workflow, system productivity, and several supply chain use cases, including disruption predictions, forecast accuracy algorithms and fraud detection

Bloomberg Intelligence BI

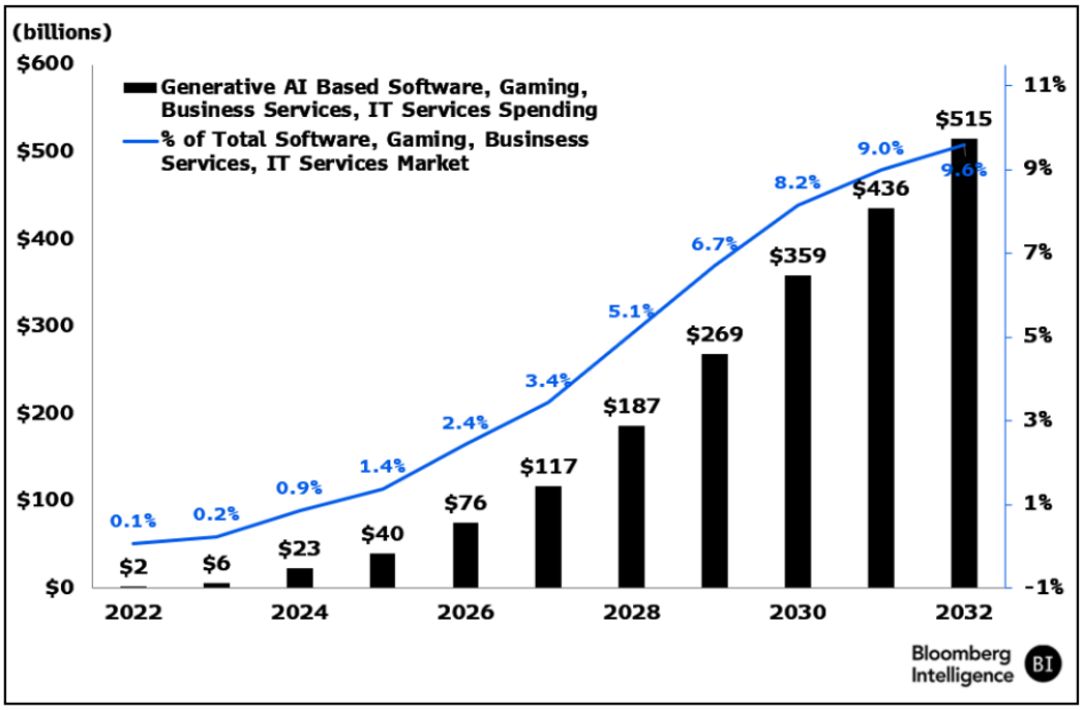
Source: Bloomberg Intelligence

6.1 Infrastructure Outruns Software, Gaming, IT Services

The new AI technology framework could add about $318 billion in software spending by 2032, climbing 71% a year from 2023, our market-opportunity analysis shows. AI assistants, cybersecurity, drug discovery and coding workflows are some of the key categories that may drive the additional outlay.

Revenue opportunities for infrastructure software appear likely to outpace those for application software, gaming and IT services. We expect software retention rates to improve, average revenue per user to rise and development costs to drop, with the software margin likely gaining 200-400 bps over the next 3-5 years as R&D expenses decline. Expansion in the IT services margin could be less pronounced as costs are tied more to human labor. However, as revenue accelerates, we don't expect the pace of hiring to increase at the same rate.

Figure 29: Generative AI Software Spending Forecast

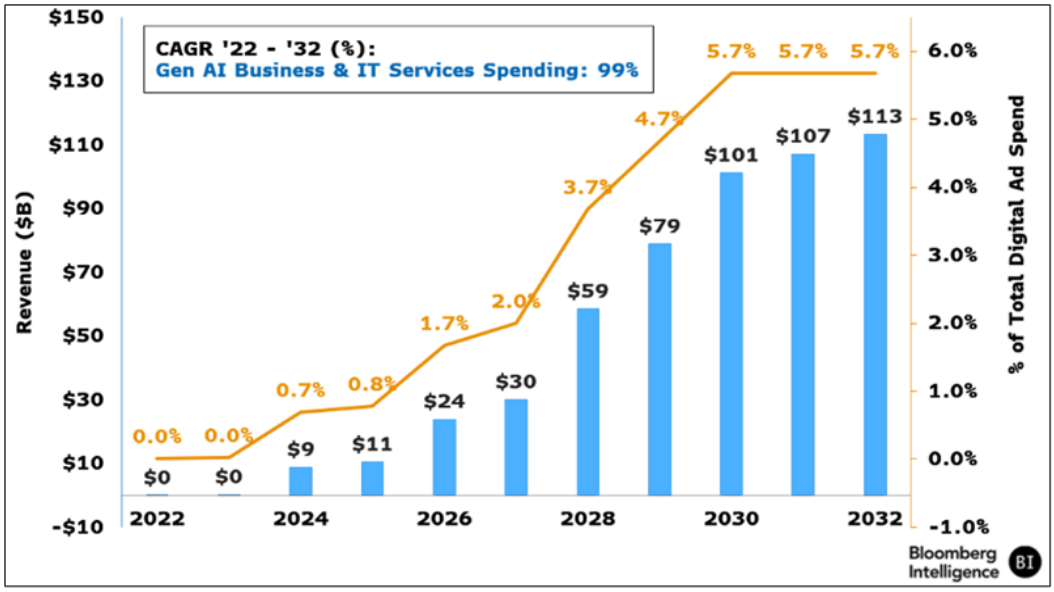


Source: BI's forecasts based on hardware and software data from IDC

Services spending could increase about $113 billion over the decade, doubling annually from 2023. Consulting, data-related services, custom application development and creation of new chatbots would drive the additional expenditures. The current total market for IT and business services combined is roughly $1.2 trillion, which may reach $2.1 trillion in 10 years, assuming 6% annual growth.

Among IT services peers, Accenture stands out given its propensity to invest ahead of the curve in emerging technologies. In 2023, the company announced a three-year $3 billion investment in generative AI that would include training 80,000 workers on AI as well as offering greater solutions and models. We calculate generative AI will contribute roughly $1.5-$2 billion to Accenture's revenue by 2025, after booking $450 million alone in 1Q. Outside of Accenture, practices such as those from IBM, Infosys and Tata Consultancy Services will also likely benefit.

Figure 30: Generative AI for IT & Business Services

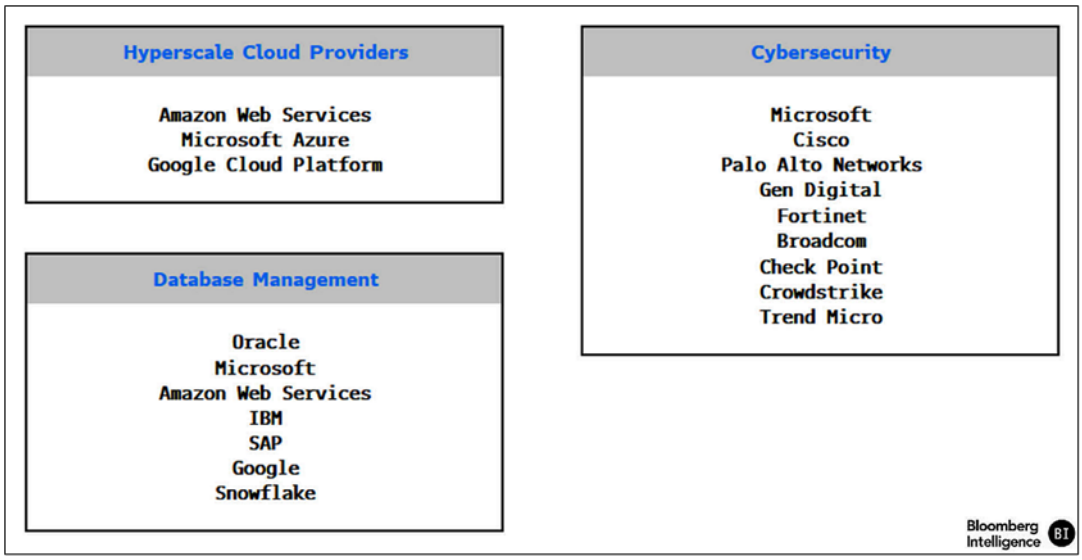


Source: BI's forecasts based on hardware and software data from IDC

Hyperscale cloud providers like Amazon Web Services, Microsoft, Google and Oracle could win increased orders for cloud services, in addition to hybrid cloud providers such as IBM. For IBM, the main source of traction would be in its Red Hat, security services and Watson-related products. Oracle can parlay its leading market share in database-management products. Cisco, Databricks, Snowflake, VMware and ServiceNow also stand out. The leading cybersecurity players including CrowdStrike and Microsoft have launched security copilots.

Similarly, design and gaming companies including Adobe, Unity, Roblox and others are integrating AI into their software to fend off competition from startups that use LLMs.

Figure 31: Major Cloud, Database and Cybersecurity Providers



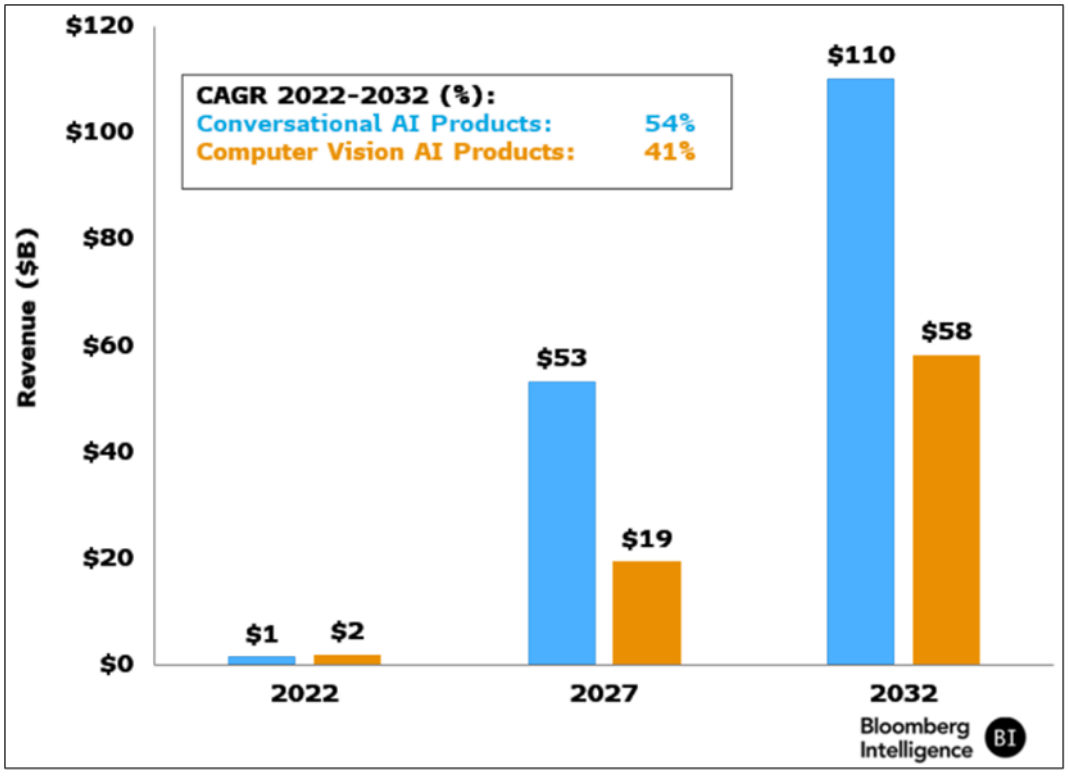
Source: Bloomberg Intelligence

Section 7. Personalizing Technology

Apple, Samsung in the Mix as AI-Enabled Tools Expand

Voice assistants based on conversational AI and computer-vision products may emerge as new categories for inference, given the availability of large language models (LLMs) for domain-specific predictions. Apple, Samsung, Amazon and others may look to conversational AI given how well their existing product offerings mesh with the category. The launch of more compact versions of existing LLMs bodes well for consumer devices given the reduced computational intensity required to run AI workloads. Auto manufacturers like Tesla and GM could invest in computer vision research to drive the next generation of AI in vehicles. Advancements in generative AI and more accurate responses for recently trained LLMs have set the stage for such categories to accelerate the overall $1 trillion devices market, where smart speakers and wearables already are large categories. Mainstream adoption of generative AI could drive a faster refresh cycle for personal computers and smartphones as new versions of these edge devices may be optimized to run generative AI apps natively, likely leveraging a smaller LLM given the processing, memory, and storage requirements. Adoption in Asia, albeit gradual, is being led by technology giants there. In one instance, Alibaba's open-sourced SeaLLM, despite having fewer parameters than OpenAI's ChatGPT-3.5, can process text in Southeast Asia faster and with more accuracy than the latter, thanks to a custom-built training dataset tailored for the region's diverse language profiles and cultural norms.

Figure 32: Inference Forecasts 2022-32



Source: BI's forecasts based on hardware and software data from IDC

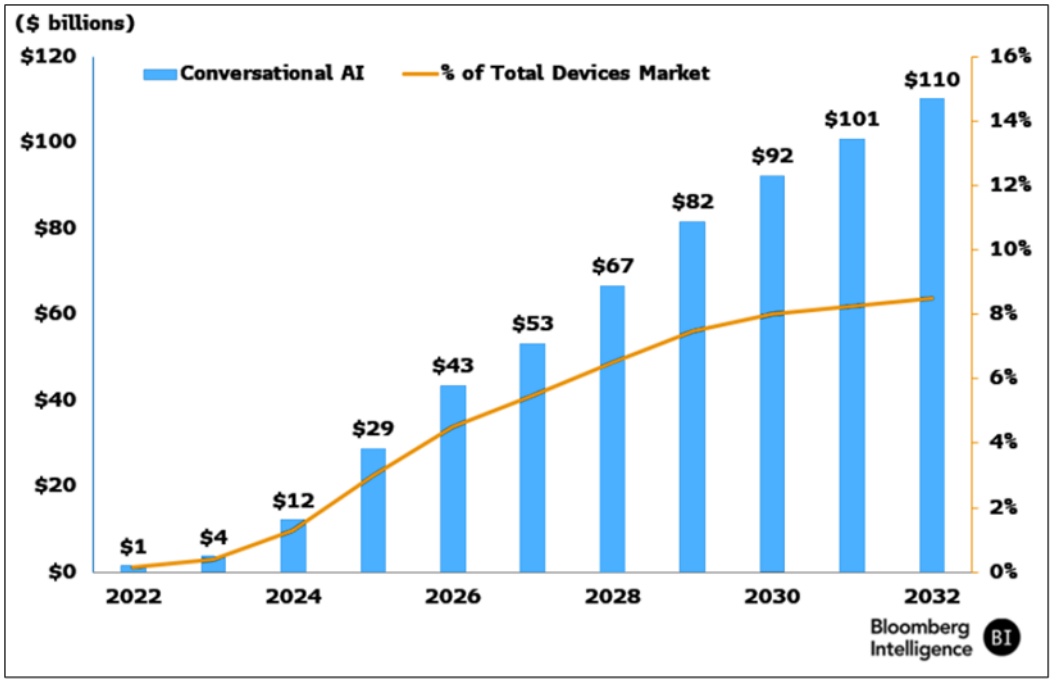
7.1 Getting Personal to Drive Sales

Conversational AI products from hardware makers like Apple and Samsung will likely be tethered to PCs and smartphones, helping to boost upgrades of existing installed bases while driving growth for services. Suppliers like Apple (HomePod), Google (Home) and Amazon (Echo) may enhance their device speakers with assistants, while carmakers including Tesla, BMW, Ford Motor and Volkswagen could incorporate them to boost driver engagement. Conversational AI is much more popular with consumers than generative AI for copilots, according to a recent Bloomberg Intelligence survey, with over 40% of respondents citing frequent use of AI tools for conversational interfaces. We expect these products to grow at roughly a 54% compound annual rate through 2032, in line with the overall generative AI market. Most of the gains will probably come in the latter half of the period as the product category becomes more established.

Computer vision also could become a significant application of generative AI tools. Building LLMs will require large amounts of training data and then need generative AI for deployment in automobiles to run inference functions. We expect most incremental revenue from computer vision to come from hardware, with the category expanding to about $60 billion by 2032, thanks to its application in advanced driver-assistance systems. There may be an even larger impact on related service sales in the medium-to-long term.

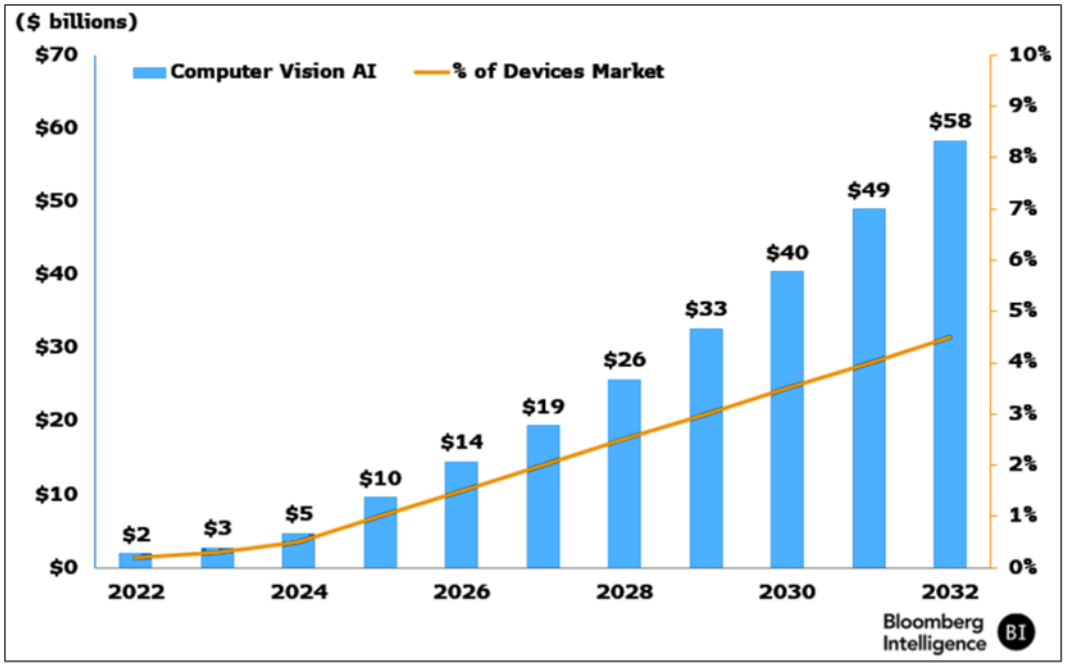
An AI training infrastructure will be essential to run these heavy workloads, sparking demand for high-capacity servers and storage. Most training-related workloads will be new since enterprises currently use general-purpose CPUs for analytics and transactions.

Figure 33: Conversational AI



Source: BI's forecasts based on hardware and software data from IDC

Figure 34: Computer Vision AI



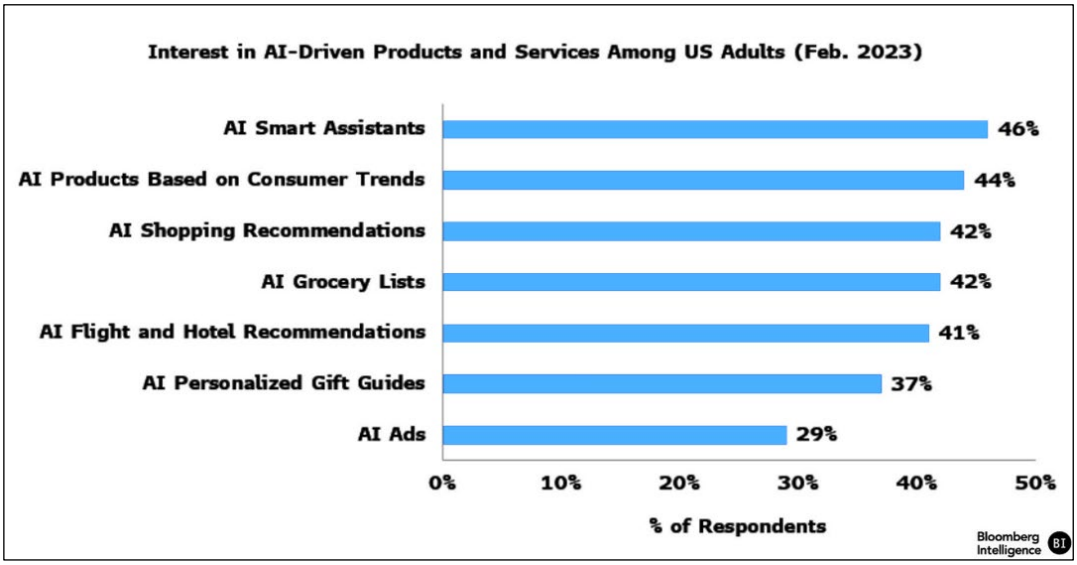
Source: BI's forecasts based on hardware and software data from IDC

7.2 Chatbots Morph Into Personalized Shopping Assistants

Generative AI is paving the way for chatbots to become personalized shopping assistants that take consumer requests and display appropriate brands and products. Snap, Meta, Pinterest and other companies that already have shopping on their platforms are investing in AI chatbots and could implement personalized shopping assistants to boost user adoption of social commerce, fueling monetization opportunities. Enhanced capabilities to handle longer and more flexible search prompts may also benefit retailers. Walmart demonstrated a search-enhancement feature at the Consumer Electronics Show (CES) in January 2024, which accepts less-specific user input and fetches relevant products. For example, the prompt "help me plan a football watch party" resulted in products like chips, salsa, and other game-day essentials being shown. We expect more retailers to add longer search prompt capabilities to their apps or websites, which may significantly expand the number of search queries on all kinds of digital platforms, especially in e-commerce.

Multimodal search could enhance user experience beyond text-based functions, which currently dominate the market. We believe the conversational nature of ChatGPT might reduce ad loads in the near term as summarized responses reduce the need to click on links to find information.

Figure 35: Service Interest



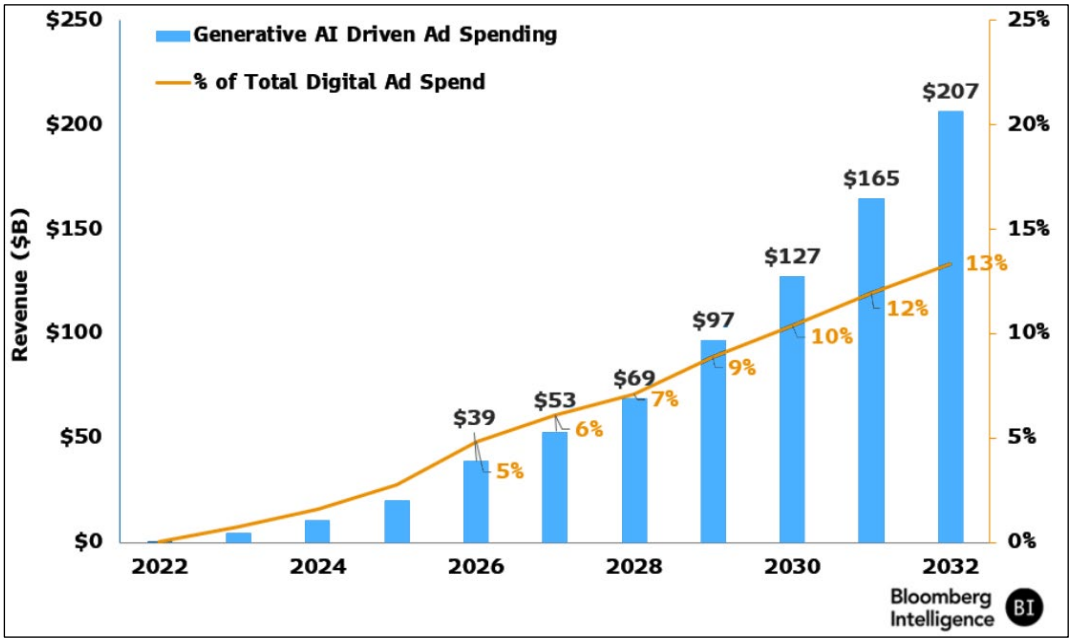
Source: eMarketer

As generative AI and machine-learning algorithms develop and adjust from user input, they cater to the person's tastes, interests and lifestyle, providing a more customized, unique experience and curating new content for social media and search platforms. That can expand availability and engagement, similar to how TikTok uses AI algorithms to recommend content to users.

LLMs may improve ad targeting for large companies that are rich with first-party data. Meta has already pivoted to AI-based recommendations with its Reels offering, helping to offset some of the headwinds from Apple's changes to its identifier for advertisers (IDFA) policy. Meta can continue to develop its Llama LLM and enhance the quality of its advertising campaigns.

Generative AI could also accelerate the shift to digital ads from linear TV, especially since offering personalized versions of advertising can increase efficiency and sales conversion. LLMs should also provide added benefit for existing large media companies as more premium content shifts to streaming from linear TV. Our analysis suggests that the generative AI market may add around $207 billion through 2032, through time spent on platforms, ad targeting and personalization.

Figure 36: Digital Ad Spending



Source: BI's forecasts based on digital advertising data from eMarketer

7.3 Recommendations Are Better Aligned

Integrating recommendations and similar-item features can also boost conversion as it helps shoppers more easily find what they're looking for. Amazon's new AI personal shopping assistant, Rufus, can answer shopper's questions, compare products, help consumers find things for a specific occasion and discover information about a specific product without the buyer needing to read all the details. This, coupled with its customer review summaries feature, can help people find items faster.

Wayfair's Decorify recommends items for a specific space, while Warby Parker uses AI to suggest specific frames based on a shopper's preferences. Similar items and personalized recommendation features for Revolve and Rent the Runway, among others, are aimed at driving brand loyalty.

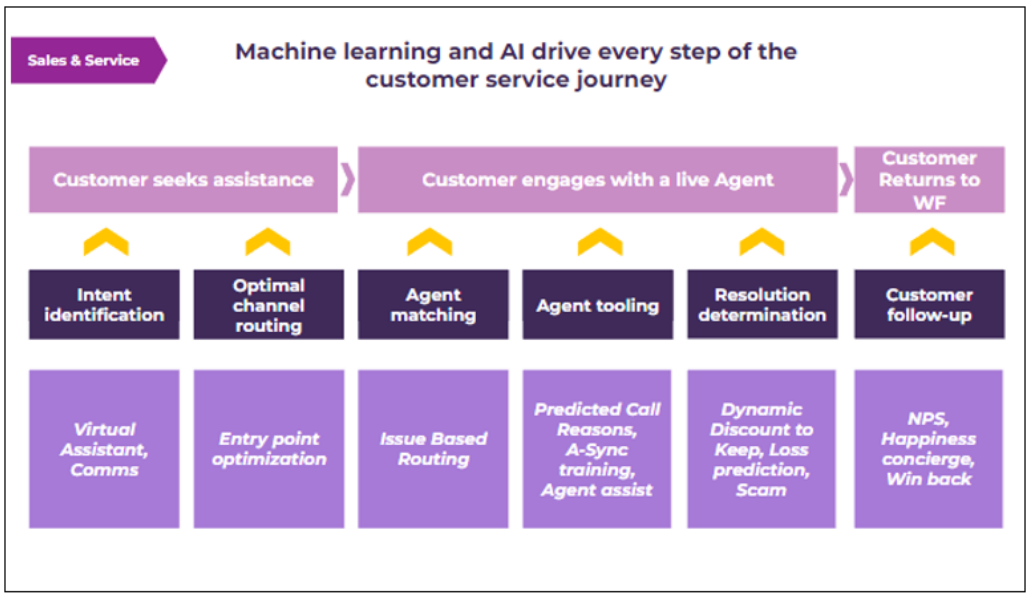
7.4 Improving Customer Service and Experience

Many companies are incorporating generative AI into customer service features, especially chatbots, which can cut costs and drive margin expansion. According to IBM, using chatbots can help reduce costs by 30% and increase customer satisfaction by 25%. In addition to Amazon's AI shopping assistant, many companies are using chatbots to assist shoppers rather than talking to a representative.

Etsy is using chatbot technology to streamline processes for its engineers to drive efficiencies. Wayfair uses AI to match customers with the best agent, predict what kind of service it should provide a customer with a complaint, what kind of discount could persuade someone to keep the item and provide all this information in a timely manner.

Warby Parker similarly uses generative AI to quickly transcribe eyewear prescriptions and help conduct virtual vision tests.

Figure 37: Wayfair's Customer Service Focus with AI



Source: W US Company Filings, Bloomberg Intelligence

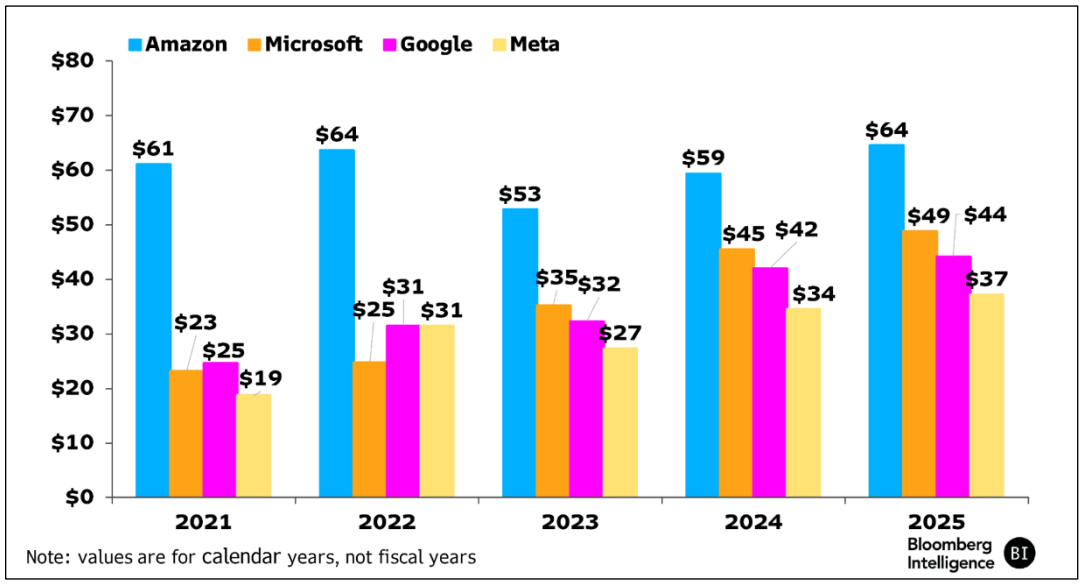
Section 8. Capital Spending Outlook

Appetite for AI to Feed $2.2 Trillion in Capex

Generative AI workloads are intense, which should spur near-term corporate investment in servers and storage. Growth in global software spending has averaged around 10-12% annually for the past few years and, despite a recent slowdown, prospects for the category are much stronger as companies invest in AI. Software makers, in particular, can burnish their product offerings by adding generative AI. As a result, capital expenditures to deploy these technologies could increase, expanding software spending 13% annually in 2022-32 and reaching $2.2 trillion by decade's end.

In the near term, data centers and cloud operators most likely will tolerate increased costs to ensure high-quality performance for AI workloads since malfunctions and system failures could lead to lawsuits, canceled contracts and financial damages. We believe that eventually, most hyperscalers – like Alphabet, Meta and Amazon – will target capex to develop proprietary, foundational large language models that will work best on their own cloud infrastructures. Microsoft's substantial investment in OpenAI's ChatGPT suggests that the software giant is unlikely to develop its own LLM in the near term.

Figure 38: Hyperscaler Capital Spending



Source: Bloomberg Intelligence

Section 9. Processing, Memory Chip Demand

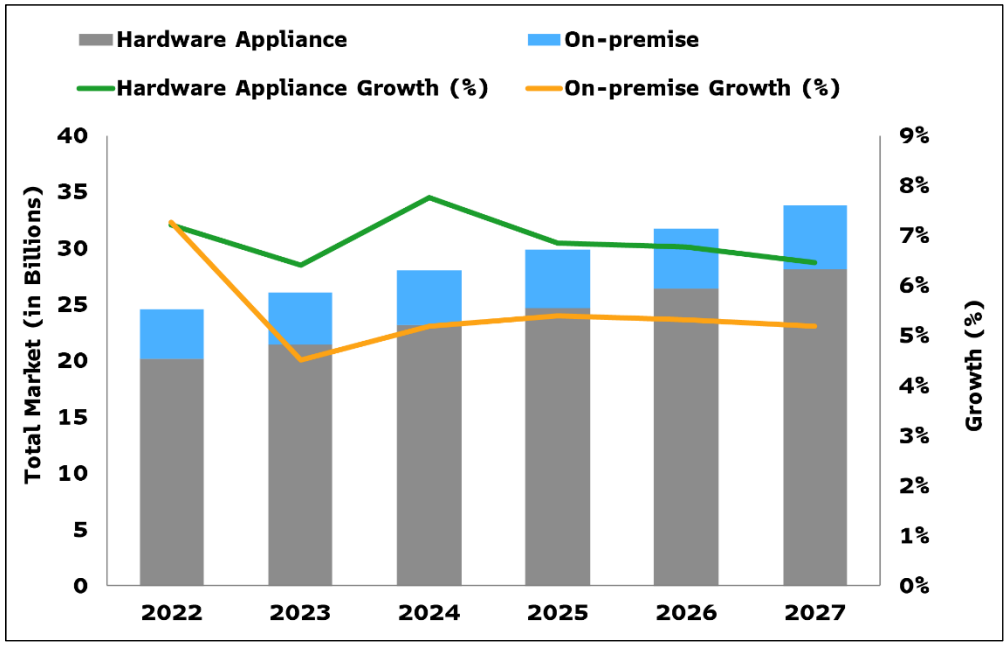
Growth Opportunities Spread Around the Globe

The spread of artificial intelligence could drive demand for graphics processing units (GPUs) and dynamic random-access memory (DRAM), lifting sales at SK Hynix, Samsung Electronics and Micron. We believe memory semiconductors will play a key role in the expansion of the data-center chip market, along with AI accelerators, with each expected to grow more than 15% annually over the next three to five years.

9.1 TSMC Prowess Sets It Apart From Pack of Rivals

Fortinet and Palo Alto Networks may leverage custom semiconductors for generative AI, supporting steady refreshes of their hardware and software firewalls. Fortinet added software-defined wide area network (SD-WAN) functionality to its ASIC chips, helping it to take share from traditional firewall vendors such as Cisco and Check Point. Palo Alto successfully bundled its Prisma, Cortex and virtual firewalls to help enterprise customers secure their on-premise and public-cloud workloads.

Figure 39: Network Security Market



Source: IDC

AMD's energy-efficient AI accelerators position stand to gain amid rising scrutiny over power consumption by data center AI chips.

TSMC is poised for a strong rebound in 2024-25, fueled by robust demand for AI accelerators and significant order gains from leading AI chip designers like Nvidia and AMD. Despite a 9% revenue dip in 2023, we forecast a 22% surge in 2024, buoyed by significant demand for its 3- and 5-nm processing node technologies and 2.5D packaging. AI infrastructure investment and

chip demand growth will be a long-term trend, which will help TSMC overcome headwinds spawned by sluggish growth for PCs and smartphones chips.

The burgeoning AI chip market, catalyzed by rapid advancements in generational AI technologies, will notably benefit TSMC, with the server GPU and AI accelerator market expected to quintuple to $51.6 billion from $10.5 billion in 2022, according to IDC.

Figure 40: BI Scenario Analysis

	2020	2021	2022	2023	BI Scenario 2024E	BI Scenario 2025E	2026E
(Revenue By Applications, $Mn)							
Smartphone	21,917	24,887	29,835	26,146	31,013	35,849	39,434
Sales Mix %	*48%*	*44%*	*39%*	*38%*	*37%*	*35%*	*34%*
HPC	14,938	21,045	31,295	29,992	39,460	49,416	59,480
Sales Mix %	*33%*	*37%*	*41%*	*43%*	*47%*	*49%*	*52%*
Others	8,632	10,904	14,856	13,220	13,824	16,101	16,202
TOTAL Sales	$ 45,487	$ 56,835	$ 75,986	$ 69,358	**$84,298**	**$101,367**	**$115,117**
Growth %		*25%*	*34%*	*-9%*	*22%*	*20%*	*14%*
Sales Consenus Estimation (as of 31/12/23)					**84,085**	**98,783**	**110,357**
					21%	*17%*	*12%*

Key Assumptions:

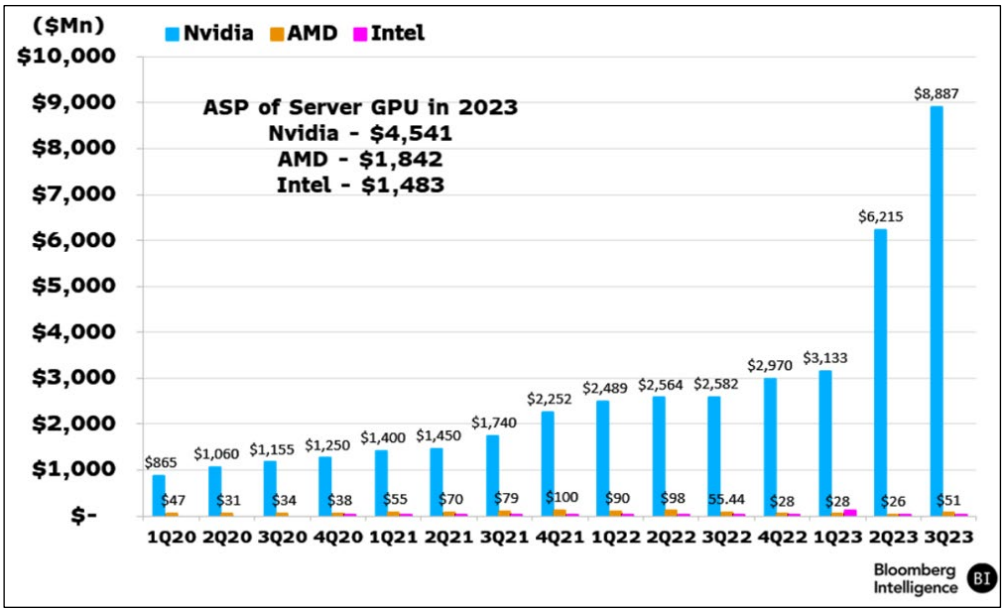
1. Smartphone related revenue to grow at an 15% compound average rate from 2024 to 2026
2. High Performance Computing (HPC), which include AI accelerator, server processors, PC will grow at a 26% compound average rate from 2024 to 2026
3. Total sales contribution from smartphone and HPC chip businesses will be 84% in 2023, 2024 and 87% in 2026
4. TSMC will still secure over 90% of global AI accelerator production orders

Bloomberg Intelligence

Source: Bloomberg Intelligence

TSMC's dominance in leading-edge node semiconductor manufacturing processes positions it to maintain its hold on the lion's share of AI chip production orders from key players such as Nvidia and AMD. That advantage is expected to continue, thanks to the company's strong production yield. Also, many AI chip designers prefer TSMC's CoWoS packaging for its superior interconnection density, larger package sizes and cost effectiveness.

Figure 41: Server GPU Sales Growth 2020-23

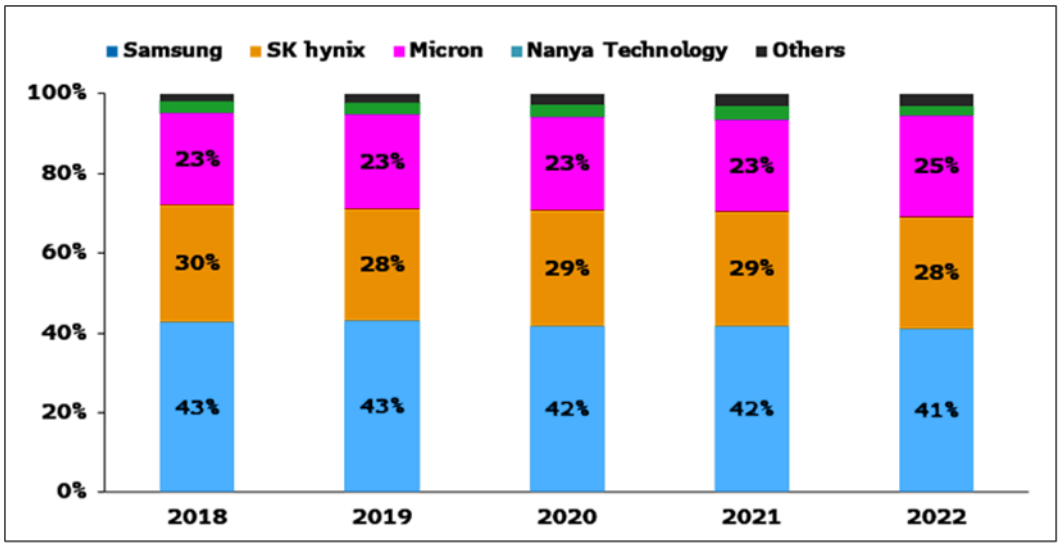


Source: IDC, Bloomberg Intelligence

9.2 Need for Speed Fuels Rapid Performance Gains

High-bandwidth memory (HBM) chips are due for a significant role since rapid performance improvements of GPUs can only be fully realized if memory can supply it with large volumes of data at high speed. As AI models become more complex and training becomes more demanding, HBM chips, such as DDR5, are expected to be more widely adopted. Since SK Hynix announced it will ship the sector's first HBM to Nvidia, it can benefit from increasing demand for

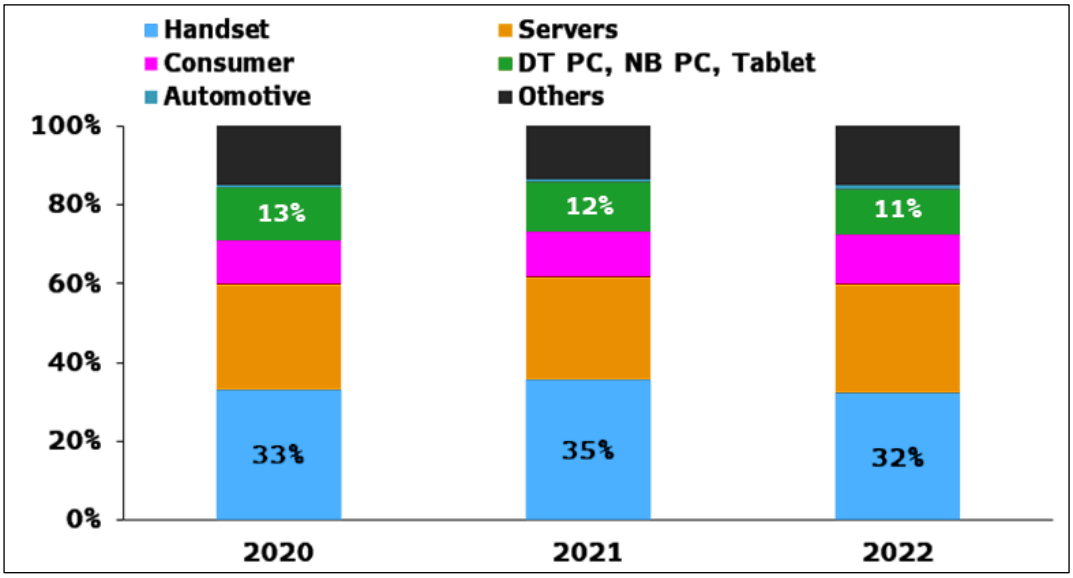
Figure 42: DRAM Bit Demand by Application



Source: Gartner, Bloomberg Intelligence

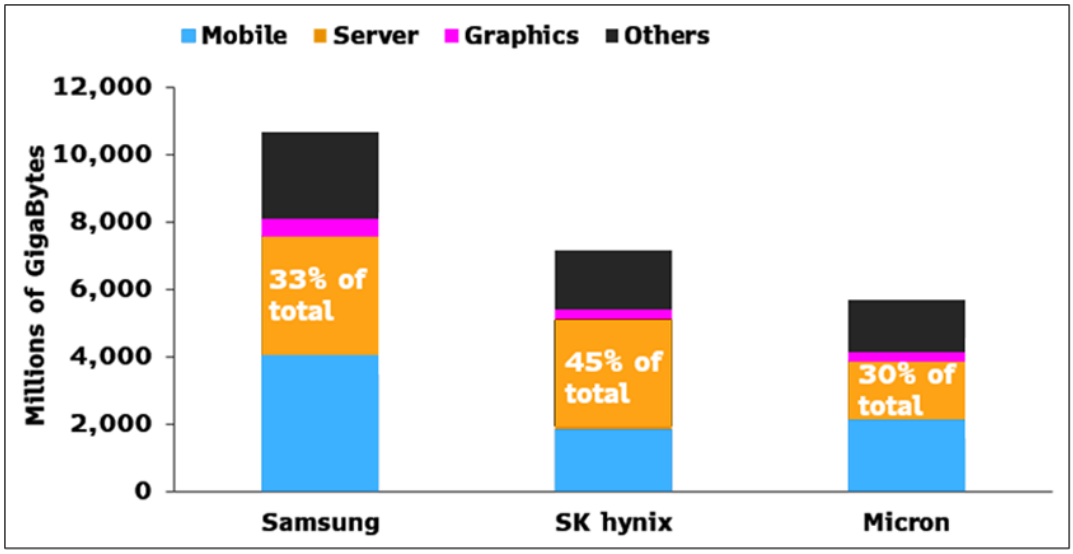
Nvidia's GPUs. Micron could also grow revenue, as it started volume production of HBM chips for Nvidia in early 2024. Results at Samsung, the world's largest DRAM maker, also may rise amid increasing use of GPUs and HBMs.

Figure 43: DRAM Bit Demand by Application



Source: Gartner, Bloomberg Intelligence

Figure 44: Leading DRAM Makers Gigabytes Shipped by Use



Source: IDC, Bloomberg Intelligence

Escalating use of AI for inference applications might heighten the value of DRAM formats like graphic double-data rate (GDDR), which is used in retail PC graphic boards where cost is critical, and low-power-consumption double-data rate (LPDDR), which is mainly deployed in smartphones.

Power consumption and data telecommunications would become excessive if all processing were conducted on servers. That will make it necessary to perform AI tasks on edge devices, which could fuel a surge in DRAM orders for products including PCs, autos, robots, smartphones and security cameras, buoying sales for Samsung, SK Hynix and Micron.

The volume of DRAM used on servers to perform AI's large-scale calculations is smaller than smartphones and PCs. Smartphones and PCs account for about 40% of global DRAM bit demand. It is true that servers account for about 30% of overall DRAM demand and artificial intelligence is just a small portion of that. However, demand for specific tools to make AI processors, HBM chips and AI chip packages could grow robustly and start to contribute to sales and profit growth, as chip customers are expected to rapidly expand production capacity.

Figure 45: Comparison of Different Types of DRAM

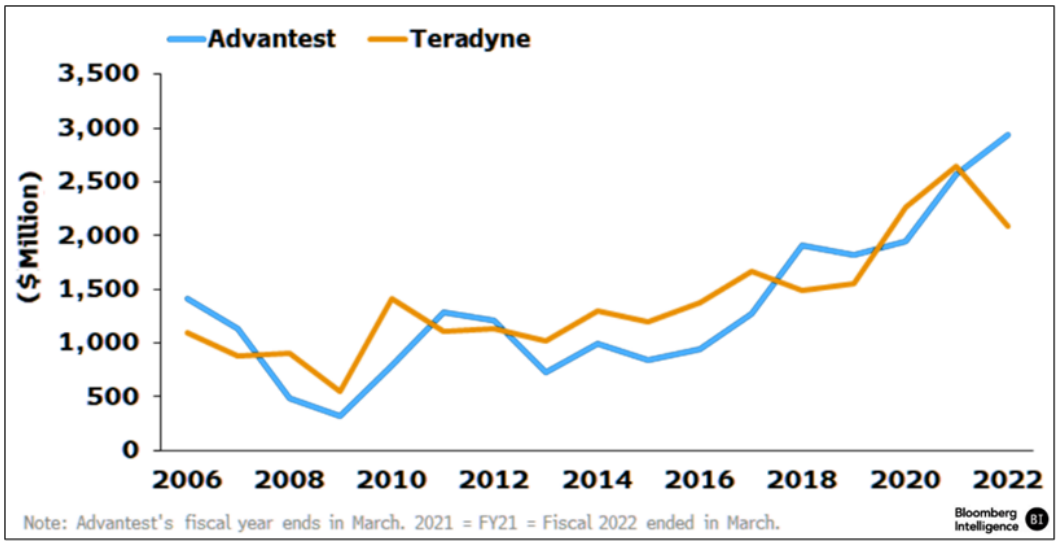
	DDR	LPDDR	GDDR	HBM
Processor	CPU	CPU	GPU	GPU/CPU
End-Device	PC	Mobile	Gaming, PC	Server
Interconnect	DIMM	Package on Package (PoP)	PCB	Interposer
Bandwidth	20-60 GB/s	100-400 GB/s	500-1,000 GB/s	1-3 TB/s

Bloomberg Intelligence

Source: Bloomberg Intelligence

The speed of chip performance improvement required for AI is faster than the evolution of miniaturization and advanced packaging, meaning quality isn't certain. As a result, the role of chip testers that can accurately assess performance and quality might rise sharply. Teradyne has a strong competitive edge in the field. And customers have given high marks to Advantest's T5000 series of memory chip testers and V93000 series of system-on-a-chip testers.

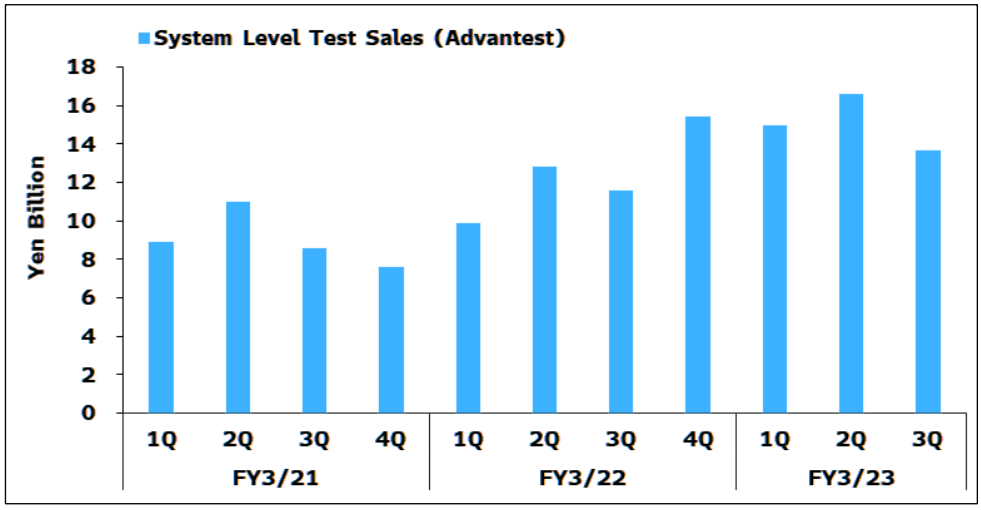
Figure 46: Chip Tester Sales for Advantest, Teradyne



Source: Company data, Bloomberg Intelligence

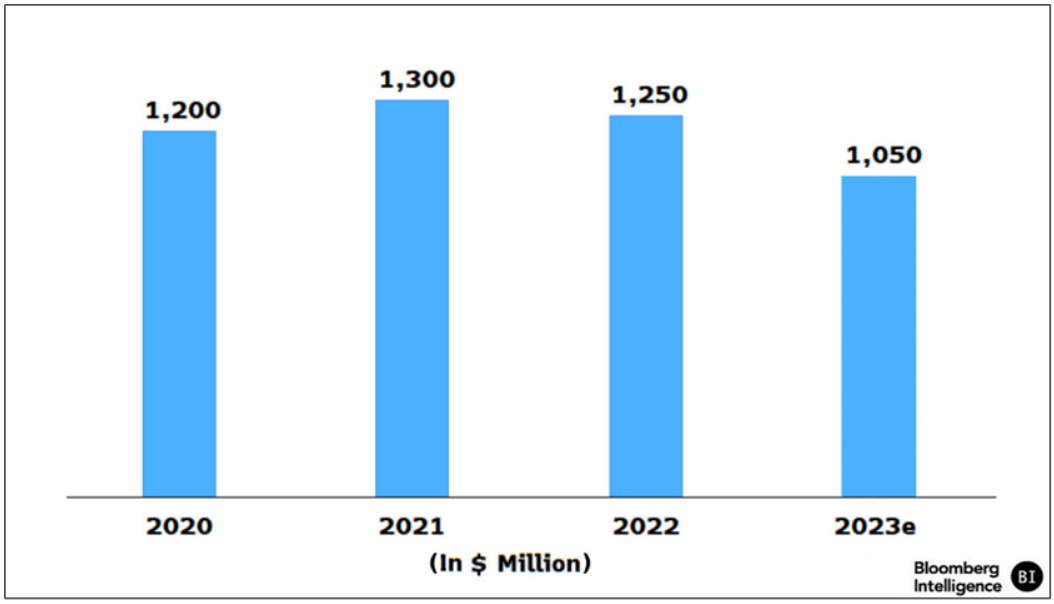
Quality assurance will be necessary not only for single chips, but also for chip packages, and entire system-level tests (SLTs) will become more important. Three-nanometer technology is taking off and chiplet packages, which place several chips or dies on a substrate vertically, could be adopted in a few years. Beyond identifying defective products, tester makers like Teradyne need to reduce the risk of misclassifying good products as deficient, requiring substantial technological capability. Advantest strengthened its SLT business by acquiring Astronics in 2019 and Essai in 2020.

Figure 47: Advantest SLT Sales



Source: Advantest, Bloomberg Intelligence

Figure 48: Memory Tester Market Size



Source: Advantest, Bloomberg Intelligence

Section 10. Regulatory Landscape

EU Farther Ahead Than US; Big Tech Exposed

Trust and content safety for generative AI needs to be strengthened, and large participants including Snap, Meta, TikTok and Alphabet are well positioned to detect and prevent deep fakes created by its misuse. Doing so should improve brand safety for advertisers and increase conversion rates for ad spending on these platforms. Any potential increase in AI-related regulation also could contribute to greater outlays on securing and encrypting data.

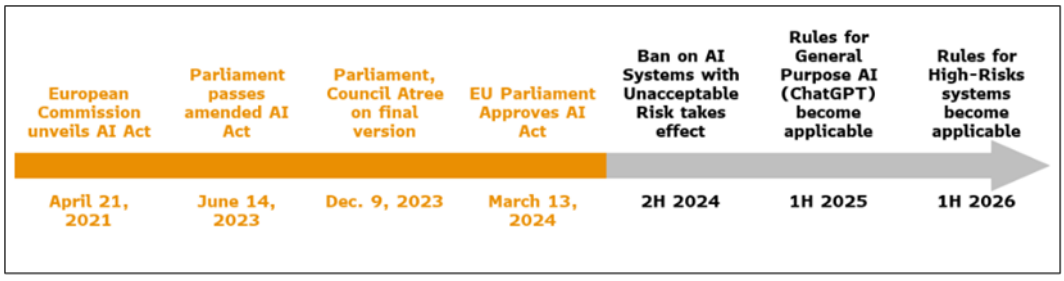
Europe is farther ahead than the US on regulating AI. Given the tool's rapid development, we believe that creating a dedicated agency is one of the few workable regulatory models.

10.1 EU Regulation Takes Broad Approach

The EU has moved forward with comprehensive regulations on AI, with formal approval of the AI Act in March paving way for implementation in the coming years. The final package was far less restrictive than earlier proposals, particularly for developers of generative AI systems, although from mid-2025 those developers will be saddled with potential onerous new transparency obligations. Still, European regulators opted against requiring those developers to get pre-approval before placing a general-AI system on the market. Had those rules been retained in the final rules, it could have frustrated the rollout of AI in the region.

The transparency obligations will only apply to general-purpose AI systems that are deemed to pose a "systemic risk." The EU will use training computing power as a basis for bucketing general-purpose AI models into the systemic-risk category, with an initial cutoff of 10 to the 25th power of floating-point operations per second (FLOPs). OpenAI's GPT-4 and Google DeepMind's Gemini are likely to be the first to fall into that category, according to the European Commission. The FLOPs threshold could be adjusted upward or downward over time.

Figure 49: Timeline of EU Act Legislative Process



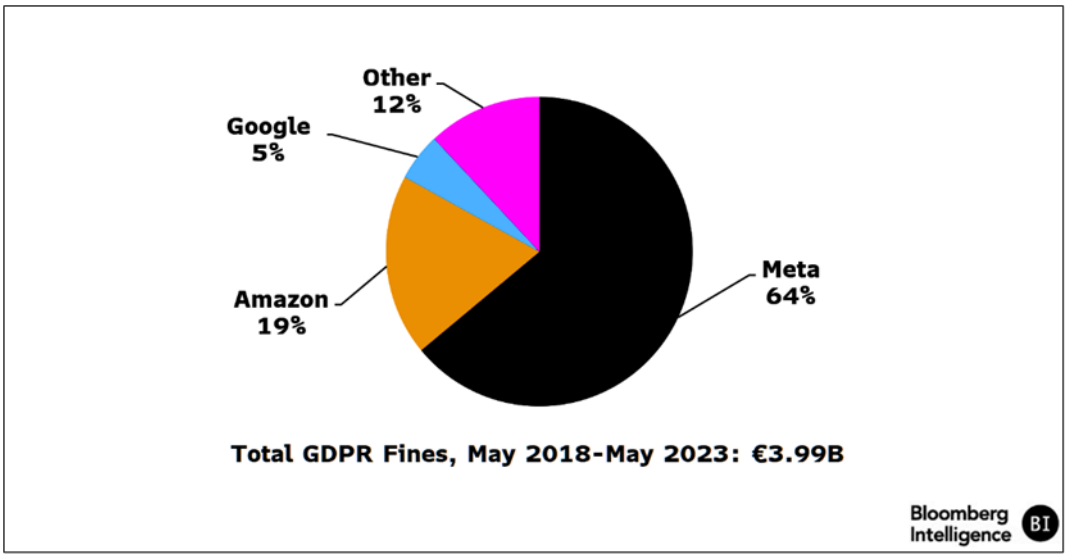
Source: Bloomberg Intelligence

The EU's AI Act buckets systems into categories based on the type of risks their applications pose. The rules will outlaw all "unacceptable risk" applications by the end of the year. That includes use of AI for things such as behavioral manipulation and biometric identification. For "high-risk" applications, which comprise large platforms' recommendation systems, a multistep approach to get approval will take effect in 2026. The European Parliament sought to include

general AI systems as high risk, but, in a boon to the industry, that was rejected. Those systems will instead be subject to transparency requirements from next year. Limited-risk applications (such as chatbots) would simply require disclosure, while minimal-risk applications (like the use of a spam filter) wouldn't have any restrictions.

The AI Act could impose fines of up to 7% of annual turnover, exceeding the 4% maximum under the current General Data Protection Regulation. In its first five years of enforcement, the GDPR resulted in cumulative fines of almost €4 billion, with penalties on Meta making up 64% of the total. Amazon accounted for 19% and Google, 5%. We don't believe the threat of penalties would deter near-term investment in generative AI, given the potential to shape the market.

Figure 50: Meta Bears Brunt of GDPR Enforcement



Source: GDPR Enforcement Tracker, Bloomberg Intelligence

The AI Act passed a plenary vote by the European Parliament on June 14, paving the way for negotiations with the European Council and European Commission. There's no timeline for talks, though there will be considerable pressure to finish by year-end, ahead of May parliamentary elections. After enactment, there would be a two- to three-year transition period for enforcement. As a result, compliance might not be expected before late 2025 and more likely not until 2026. EU policymakers are pushing for a voluntary code of conduct as a stopgap before the AI Act takes effect. If that happens, we expect industry leaders to embrace principles that could help shape the scope of regulation.

The AI Act would add to the EU's increasingly complex regulatory framework on tech. In recent years, the bloc has sought to rein in the wanton collection and use of personal data (the GDPR) and to impose obligations on platforms relating to content moderation (the Digital Services Act) and abuse of market power (the Digital Markets Act). The rules can impose substantial financial penalties and operational remedies that can materially alter a business. Enforcement of the GDPR has been fractured, with Ireland taking the mantle for regulating most platforms. The European Commission monitors for DSA and DMA compliance. Individual national regulatory bodies likely would lead AI Act enforcement.

Figure 51: EU's Recent Tech-Focused Regulations

Regulation	Compliance Date	Purpose	Mostly Likely Targets	Max Fine
General Data Protection Regulation (GDPR)	May 25, 2018	Provide strict rules on the collection and use of personal data	Meta, Amazon and Google have been the most targeted by GDPR regulators	4% of annual revenue
Digital Services Act (DSA)	August 25, 2023	Impose new obligations on platforms to enhance moderation of illegal content	EC has identified 17 "Very Large Online Platforms" and two "Very Large Online Search Engines" that were required to comply with the DSA from August 25, 2023	6% of annual revenue
Digital Markets Act (DMA)	May 2, 2023; March 6, 2024	Provide new competition rules on large, "gatekeeper" platforms	Gatekeepers include large tech platforms, which will then have to comply by March 2024	10% of annual revenue
AI Act	2026 to 2027	Provide broad safeguards on the development and deployment of AI systems	Potentially any developer or implementer of an AI system, but large tech firms once again likely to be focus given early investments into generative AI models	7% of annual revenue

Bloomberg Intelligence BI

Source: Bloomberg Intelligence

10.2 Aggressive Regulatory Tack Unlikely in US

If the US adopted an aggressive regulatory approach – which we doubt will happen – it could dent growth of AI products from a range of companies: chipmakers like Micron and Nvidia; cloud infrastructure providers like Amazon and Oracle; software- and development-tool makers like Adobe, IBM and Microsoft; and platforms that use AI for data, search and ad capabilities, like Alphabet and Meta.

The first federal bipartisan bill – the No Section 230 Immunity for AI Act, which would do little – shows how much work remains before serious AI limits land in the US. The measure would clarify that a federal liability shield, Section 230 of the Communications Decency Act, wouldn't apply to AI, but we think courts would only rarely apply the provision to the technology anyway. More notably, the legislation wouldn't create a federal right to sue over AI harms. Nor would it attempt to say what AI is, defining "general artificial intelligence" so broadly that the law could remove Section 230's shield from many existing online platforms. The provision most likely won't become law without being narrowed significantly.

A second bipartisan bill, the National AI Commission Act, strikes us as a sensible first step for Congress, with a solid chance of becoming law. It would create an independent, bipartisan commission with 20 members to study AI's risks and suggest guardrails. The panel would release three reports over two years, shaping regulation. The commission would include industry representatives, which should ease opposition.

A Senate committee hearing in May generated surprising support for what we see as the most logical – though probably the most disruptive – approach: creating a dedicated federal agency.

Congress is incapable of keeping up with AI's rapid development and though a new agency would struggle as well, it at least would have a shot at monitoring the technology and fashioning legal limits. Such a body could also be more focused than an entity like the Federal Trade Commission, which has oversight of all industries. A new agency would upset the apple cart, however, so it likely would face vigorous industry opposition, and companies that might support it probably would push to narrow the body's powers.

Congress also might consider a licensing model, in which AI applications with a broad reach or the potential to cause severe harm can't be put into operation until they've obtained permission from a new regulator. The approach would face fierce industry pushback, including claims that the move would stifle AI innovation and drive it abroad.

A less intrusive form of regulation would focus on transparency, requiring that AI products be disclosed and labeled, and might require that regulators or researchers be able to monitor data that's collected or used.

Since Congress is unlikely to reach consensus on a disruptive approach – like creating a new agency or developing a licensing model – we expect AI to be governed by existing laws, even if they're not always a good fit. The FTC could monitor for unfair and deceptive practices, for example, and existing sector-specific laws may limit AI application by industry. AI should thrive in the US under such a light touch. Yet unlike social-media companies, AI users probably won't receive broad protection from lawsuits from Section 230, which raises the specter of legal liability.

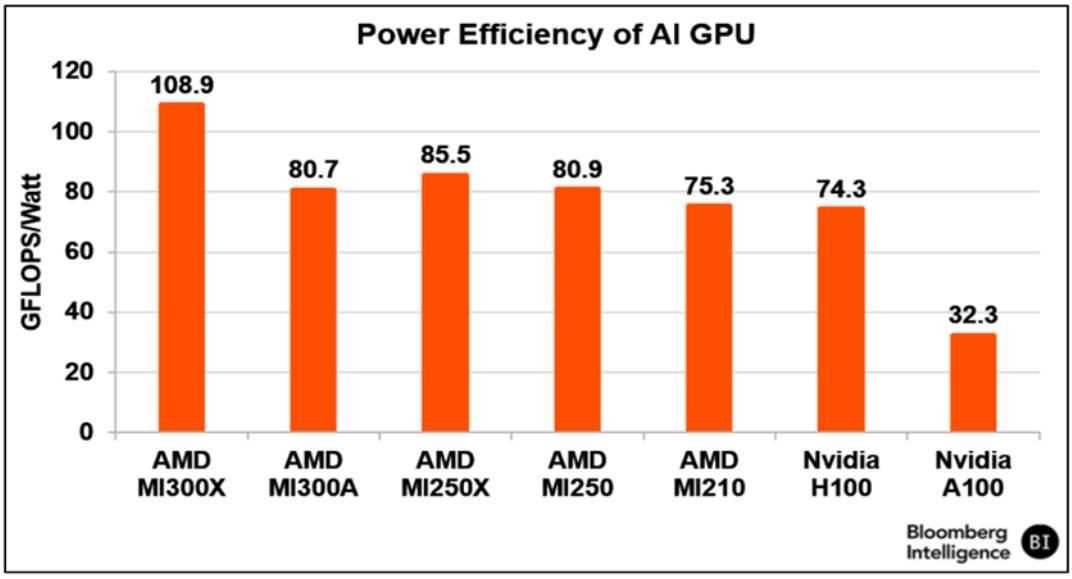
Section 11. ESG Outlook

Reducing Power Consumption; Protecting IP and Privacy

Increased use of graphics processing units for AI inference will significantly boost energy consumption in data-center servers, putting a priority on energy savings to maximize operating efficiency while minimizing power and cooling costs. That could favor AMD over Nvidia as the former's latest MI250X accelerator outperforms Nvidia's H100 by performing a higher peak number of floating-point operations a second per watt.

We believe ARMs can continue to gain market share from x86 processors in data centers. Existing internet workloads largely run on x86 architecture, but most generative AI applications will run on chips like GPUs that can conduct massive parallel processing with low power consumption.

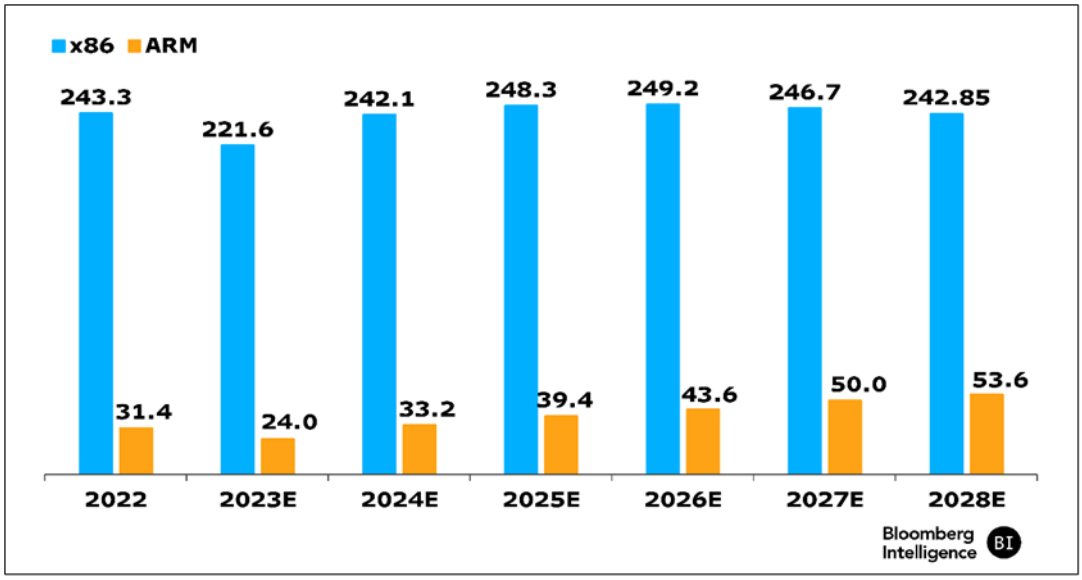
Figure 52: Energy Efficiency of Advanced AI Chips



Source: Company filings, Bloomberg Intelligence

Nvidia's advantage with its CUDA interface is likely to remain for ARM-based processors and may help the company gain market share over x86 processors across data centers, networking and edge devices thanks to the ARM design's power efficiency.

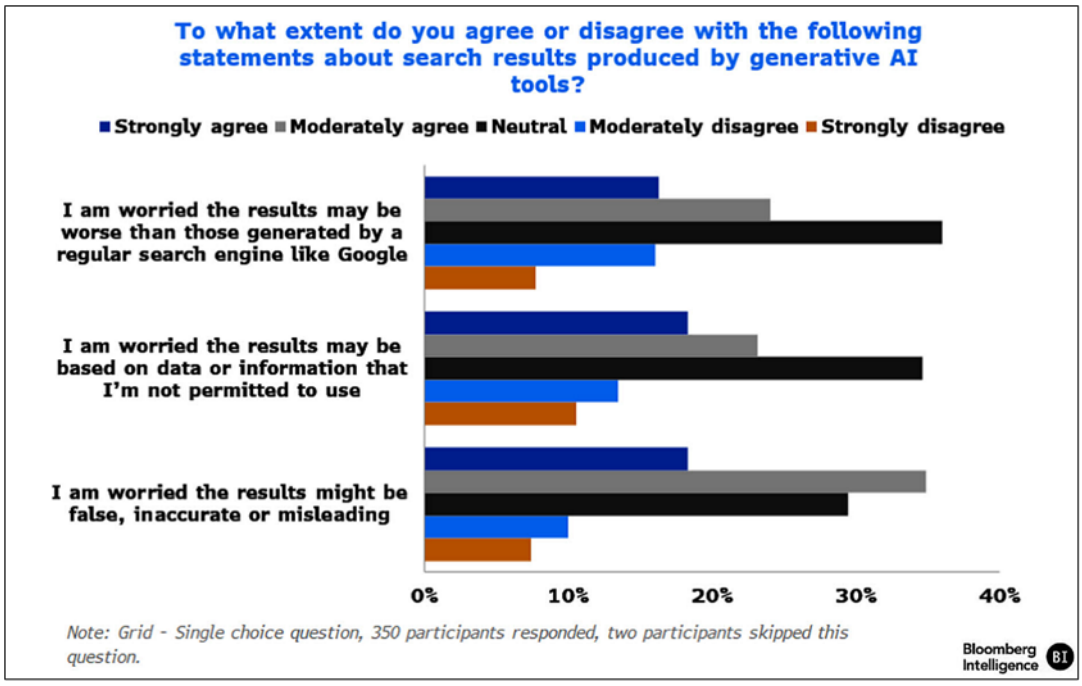
Figure 53: x86 vs. ARM Shipment Forecast



Source: IDC

About 40% of respondents in a BI survey expressed concern over how generative AI's use of information could breach intellectual-property rights. Still, a majority said they would use such tools as long as they provided better results than traditional search functions from Google and other websites. IP concerns will likely decrease over time as companies share more information about how their algorithms are trained using proprietary rather than open web data.

Figure 54: Generative AI Search Result Concerns



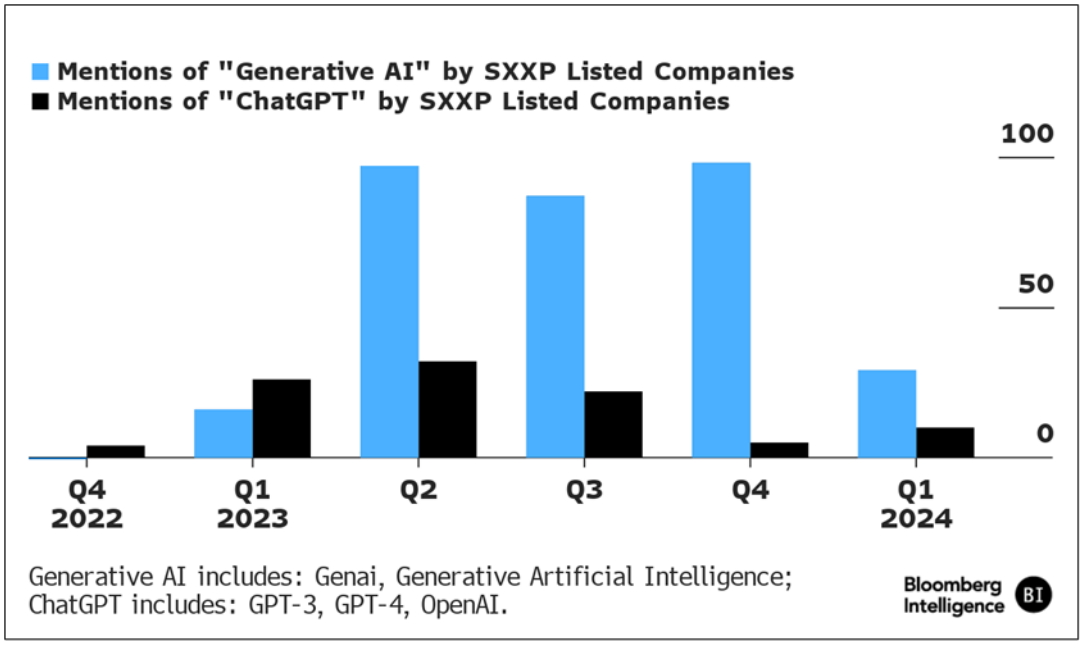
Source: Bloomberg Intelligence

Developers of high-level generative AI systems could retreat from the EU, absent further revisions

to the bloc's proposed regulations. Mentions of generative AI and ChatGPT by European companies exploded this year, suggesting a strong desire to employ the new technology. Yet the European Parliament on June 14 adopted rules that would subject developers of generative AI models to additional constraints, such as transparency rules related to the data sets used for training.

Since all AI essentially is driven by data – collecting lots and using machine learning to produce outputs based on it – US regulators could address related harms by limiting what can be collected and used. That would be similar to data-privacy regulations that have been proposed for social-media companies. Given the parallels, expect internet platforms to vigorously lobby against such limits.

Figure 55: European Firms Eye Potential



Source: DSCO <GO>

Section 12. Performance and Valuation

AI Establishes Itself as Dominant Accelerating Tech Theme

As tech themes pick up speed in every segment, artificial intelligence is living up to the hype as the standout performer, fueling advances in both share prices and valuation multiples across the technology spectrum, from software to hardware, networking, services and more.

12.1 Performance: Knock-On Effect of AI on Themes

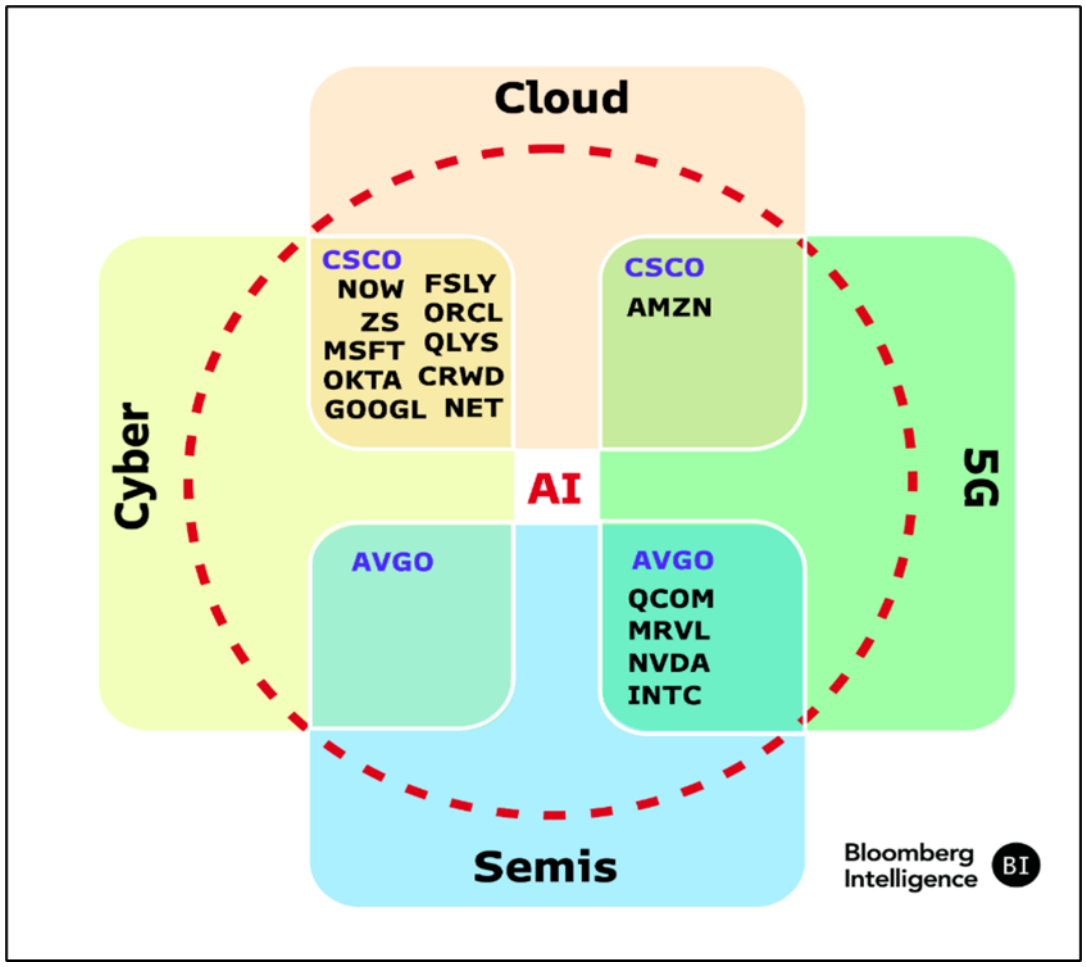
AI is the top performer across BI's accelerating tech theme category – up 65% vs. 26% for the broader group over the past year. One of AI's most powerful implications to the theme landscape is that its proliferation stands to extend its impact to adjacent themes such as 5G, cloud, cybersecurity and semiconductors. AI and AI-adjacent themes, thus, rank at the top of BI's near-term theme radar.

The interconnection between AI, cloud, cyber, semis and 5G is significant, giving rise to a curated group of names we've identified as having strong thematic breadth, related to these much-watched themes. Broadcom and Cisco's appearance in four themes signals strong significance for the growth of AI. Broadcom is seen in AI, 5G, cybersecurity and semiconductors; Cisco in AI, 5G, cybersecurity and cloud. All of our identified AI and AI-adjacent thematic-breadth names are seen in AI, proving how that theme serves as a powerful epicenter.

As a group, these high thematic-breadth names outperformed the S&P 500, Nasdaq 100 and the broader combination of their full BI theme universes, based on year-to-date, 1-year, and 3-year-annualized returns analysis. 41% of the high-breadth names are ultra-large caps, manifesting gains as highly visible hyperscalers. Nvidia, CrowdStrike and Broadcom are the top-performing high-breadth names over the past year, up 276%, 166% and 115%.

Generative AI remains the dominant catalyst of positive estimate revisions and multiple expansion. So far, semiconductor and hardware companies exposed to training foundational models, most notably Nvidia and Microsoft, have been the primary beneficiaries of the trend. Internet companies like Alphabet, Meta and Roblox are investing in developing their own LLMs built atop proprietary data and open-internet data to challenge the adoption of OpenAI's GPT.

Figure 56: AI and Adjacent Themes (5G, Cloud, Cybersecurity, Semiconductors) Breadth Names (3+ Themes Overlap)



Source: Bloomberg Intelligence

12.2 Valuation: Estimates Climb as Product Horizon Broadens

There's been significant multiple expansion for specific semiconductor companies, with Nvidia again leading the way, with the most visible impact on sales growth expectations from AI.

Momentum in top-line estimate revisions could hinge on the pace of product releases such as Microsoft's release of its GitHub and Office copilots. Alphabet has recently released its Gemini LLM and Duet AI copilot while Meta has open-sourced its Llama foundational models to spur adoption. Database and infrastructure software companies such as Oracle, Snowflake, MongoDB and Databricks have continued to ramp up their offerings with vector search capabilities that could stand to benefit from large amounts of data used for training LLMs, which could help drive positive revisions to consensus. However, the benefits that accrue may not be balanced across the players.

Section 13. Company Impacts

With a projected $1.3 trillion in spending by 2032, generative AI's effects will ripple through every industry in the technology sector. Here's a look at how some companies are poised to gain over the coming decade.

Microsoft Among Best Positioned Software Makers

Market Cap (USD Bln)		3,177.25
Share Price (USD)		429.37
Earnings Per Share* (USD)		P/E
2023A	9.68	44.36
2024E	11.66	36.83
2025E	13.38	32.10

* Bloomberg Consensus

$6-9 Billion

Gen AI sales impact in 2024

75%

GitHub Copilot adoption

Company Outlook: Microsoft's vast array of software applications makes it a key beneficiary of the growing digital transition as companies upgrade legacy IT systems. A foothold in cloud infrastructure, coupled with its close relationship with OpenAI, puts Microsoft in a strong position to capitalize on rising demand for generative AI.

AI Impact: Microsoft is better positioned than most software companies to capitalize on the increased adoption of generative AI, given a first-mover advantage from its tie-up with OpenAI, and we calculate this could result in $6-$9 billion in generative AI revenue in calendar year 2024. It's the first large company to launch AI copilots across its product portfolio, from Office to GitHub. Azure, a cloud-infrastructure product, most likely will be Microsoft's main beneficiary of increased AI demand in the long term. Not only does ChatGPT run on Azure, but Microsoft is also making OpenAI's LLMs available on the platform. Search is another area where we believe Microsoft could gain market share steadily over time.

Amazon Can Gain From Training, Inference, Creative

Market Cap (USD Bln)		1,854.93
Share Price (USD)		178.15
Earnings Per Share* (USD)		P/E
2023A	2.83	63.05
2024E	5.07	35.17
2025E	6.23	28.60

* Bloomberg Consensus

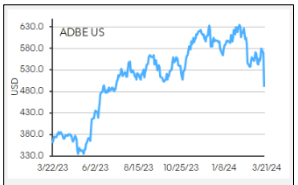
100-200 Bps

AWS growth from AI

Company Outlook: Amazon.com's push for speed, convenience and value, coupled with building momentum across retail, cloud and ads, position it well to deliver on its 1Q plan. Cloud-services growth could accelerate in 2024 with margins also expanding. Improving IT budgets and companies' greater willingness to shift infrastructure to the public cloud remain catalysts for AWS in the longer run. Operating margin may continue to expand on cost cuts, efficiencies and rising contribution from cloud and ads. Amazon's push for pharmacy and grocery are large undertakings that we'll monitor closely.

AI Impact: AWS should see its fair share of new generative AI workloads from training and inference to creating new applications through its Bedrock offering. Like Microsoft's GitHub, AWS also is offering a generative AI-embedded software development product called CodeWhisperer that significantly shortens the time it takes a developer to code. AWS has the largest market share in cloud infrastructure as a service and more than 100,000 customers using its other AI and machine learning services. Though AWS hasn't combined with OpenAI LLMs yet, it does provide its own foundational models in addition to working closely with other providers, such as Anthropic, Stability AI and AI21.

Adobe to Parlay Base of 70 Million Creative Professionals



Market Cap (USD Bln)		226.33
Share Price (USD)		511.25
Earnings Per Share* (USD)		P/E
2023A	16.07	31.81
2024E	18.02	28.37
2025E	20.32	25.16

* Bloomberg Consensus

$0.9-$1.4 Billion

Gen AI sales gain over 2-3 years

25%-Plus

DF copilot adoption Year 1

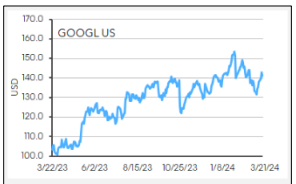
100 Bps

Operating margin expansion in 2024

Company Outlook: Adobe's solid portfolio of digital products could lead to organic sales growth of 12-15% in constant currency over the next three years, along with non-GAAP operating margin of about 45%. We see the company as among the key beneficiaries of increased spending on digital transformation, given its focus on data insights, digital commerce, marketing and content creation. Generative AI product Firefly could help drive up average revenue per user, as Adobe differentiates itself with built-in copyright safeguards for immediate commercial use, amid increasing competition from providers like OpenAI.

AI Impact: Adobe's installed base of around 70 million creative professionals, the highest market share in this category, positions it well to reap benefits from generative AI. The recently launched Firefly creative copilot can substantially reduce the time it takes to create images via text and has already helped create more than a billion visuals through generative fill. The troves of data housed in Adobe's creative cloud suite, which includes Photoshop and Illustrator, make it better positioned than rivals to train its LLMs, and rights to the underlying training content in Adobe Stock can also provide creators with legal peace of mind. Gen AI advancements in digital documents as well as its enterprise front-office software can also aid productivity and could lead to pricing improvements.

Alphabet Leveraging AI Across Product Suite



Market Cap (USD Bln)		1,879.48
Share Price (USD)		147.60
Earnings Per Share* (USD)		P/E
2023A	5.80	25.45
2024E	7.05	20.93
2025E	8.09	18.23

* Bloomberg Consensus

$3-$4 Billion
Boost to Google Cloud

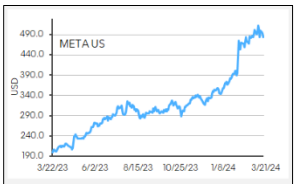
10-15%
Search queries to leverage AI

5-10%
YouTube ad engagement

Company Outlook: Alphabet's improved top-line growth for core search and YouTube segments appear sustainable for the rest of 2024, while demand for generative AI computing bodes well for its Cloud segment profitability. Though Bing-ChatGPT is a risk to the ad business, we believe the company's coming launch of a multimodal large language model and integration of generative AI features into its core search and YouTube products have eased near-term competitive pressure. Network ad sales face pressure amid the removal of cookies, though YouTube ads and subscriptions could see double-digit growth in 2H.

AI Impact: Alphabet is exposed to most segments of the generative AI market, including training, inference and digital ads. Gemini is already being used across the company's offerings, including ad targeting algorithms, Google Cloud Vertex AI, and the Google Pixel 8, the first smartphone with a native generative AI assistant. Cloud AI tools may offer monetization opportunities, while improvements to targeting algorithms can provide a lift to ad pricing. A potential partnership with Apple will further aid Gemini's positioning in the nascent gen AI inferencing market. Though many retailers are rolling out their own enhanced search offerings with longer prompts, the probability of Alphabet losing consumer traffic is relatively low given users' preference for Google's search

Meta Using Llama for Recommendations, May Appeal to Enterprises



Market Cap (USD Bln)		1,295.74
Share Price (USD)		507.76
Earnings Per Share* (USD)		P/E
2023A	14.87	34.15
2024E	20.38	24.91
2025E	23.56	21.55

* Bloomberg Consensus

5-10%
Effect on engagement, impressions growth

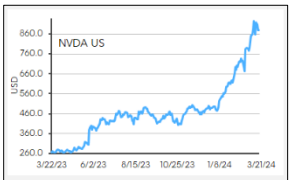
$1-$2 Billion
LLM licensing sales bump

$10 Billion
Click-to-message ad run rate

Company Outlook: Meta may continue to see benefits to user engagement from pivoting to AI-based recommendations, driving stronger impressions growth across its family of apps. The company has leveraged its Llama model to improve ad targeting algorithms while also launching new products like AI Studio and Meta Assistant which can aid monetization efforts. Reels, a high-single-digit contributor to sales, and click-to-messaging ads are both above $10 billion in revenue run rate. User growth will be driven mostly by Instagram and WhatsApp, while Reels could emerge as a contributor to ad pricing this year. Free cash flow may remain squeezed by a slightly higher capex guide and potentially a $20 billion annual loss from Reality Labs in 2024

AI Impact: Meta's scale in running its own data-center infrastructure, coupled with large amounts of training data from its family of apps, has allowed it to build its own foundational LLM, Llama, to compete with offerings from OpenAI, Alphabet, and others. Llama's open-source nature may be more attractive for enterprises looking to build their own functionality on top of the model. The pace at which new content is created for social media and virtual- and augmented-reality applications for the metaverse could speed up with generative AI. Meta may also implement personalized shopping assistants to boost user adoption of social commerce, increasing monetization opportunities. Our analysis suggests that the generative AI market may add almost $207 billion in ad spending through 2032 with time spent on platforms, plus ad targeting and personalization.

Nvidia's GPU Dominance to Stay Intact in Data-Center



Market Cap (USD Bln)		2,360.59
Share Price (USD)		914.35
Earnings Per Share* (USD)		P/E
2024A	12.96	70.55
2025E	24.67	37.06
2026E	29.78	30.70

* Bloomberg Consensus

$116 Billion
Data-center sales by 2025

127 Bps
Gross margin growth

Over 40%
Hyperscaler customer concentration

Company Outlook: Nvidia's data-center business has grown as it dominates the server GPU market, especially for large-language model training workloads. This dominance and associated revenue growth is likely to continue for the next few years building on its new Blackwell product portfolio in both training and inference. But competitive pressures continue to rise from chip rivals like AMD and Intel as well as from in-sourced solutions by its large cloud customers. Nvidia aims to expand its success beyond large cloud customers to large and medium-sized enterprises and develop its full-stack AI software solutions.

AI Impact: Given the high growth expectations in generative AI's training and inference market, Nvidia's data-center segment is poised to continue accelerated gains. As more workloads need to be accelerated amid increasing AI penetration everywhere, GPUs and associated AI networking fabric are likely to become the core computing engine in data centers, both areas where Nvidia is a market share leader. Nvidia faces risk from hyperscale public cloud providers Microsoft, Alphabet and Amazon – which are among the biggest spenders on foundational models – using their own chips for training LLMs to boost margins.

Section 14. Glossary of Terms

These can help decipher highly technical elements:

Advanced RISC Machines (ARM): A processor architecture based on 32-bit reduced instruction set computer.

AI Assistants: A software agent that can perform tasks for a user based on input such as commands or questions. Think Siri or Cortana.

AI Server: Computers used for AI inferencing and training.

AI Storage: Often a software-as-a-service application that performs analysis in a public cloud.

ChatGPT: A free chatbot that can answer just about any question.

Conversational User Interface: Allows people to interact with software, apps and bots as they would another human being. Amazon's Alexa is an example.

Computer Vision: Enables computers and systems to derive meaningful information from digital information, videos and other visual inputs, then act or make recommendations based on that information.

Corpus: Literally "body," this is the collection of billions of data points used to train a large language model.

CPU: Central Processing Unit. Basically, the semiconductor chip that is the essential logic circuitry in a hardware system.

Edge: Deployment of computing and storage resources closer to where data is produced.

Ethernet: Technology to connect devices in a local area network (LAN) or wide area network (WAN). Slower than InfiniBand.

Generative AI: Uses algorithms, such as ChatGPT, to create new content including audio, code, images, text and videos.

GPU: Graphics Processing Unit. A specialized circuit for image and video display.

Hallucination: A response/output from a large language model that is irrelevant or incorrect

Inference: The process of reasoning and making decisions based on available information or data. It follows training to derive new knowledge or conclusions from existing data.

InfiniBand Network: A high-performance, low-latency way to facilitate high-speed communications. Faster than Ethernet

IaaS: Infrastructure as a Service, a business model that offers computing, storage and networking resources on demand.

Large Language Models: Deep learning algorithms that can recognize, summarize, translate, predict, and generate content using massive datasets.

Machine Learning: The use and development of computer systems that learn and adapt without following explicit instructions, by using algorithms and statistical models to analyze and draw inferences from patterns in data.

Neural Network: A method that teaches computers to process data the way the brain does. It's a type of machine-learning that uses interconnected nodes in a layered structure.

ODM: Original Design Manufacturer.

PaaS: Platform as a Service. A type of cloud computing service model that offers a flexible, scalable platform to develop, deploy, run and manage apps.

Personalization: Uses AI and machine learning to analyze a user's data, helping to understand their needs and tailor their experience accordingly.

Retrieval-Augmented Generation (RAG): A method by which a company can supply its proprietary data to a model that can append relevant information to a user query prior to feeding it into a LLM

Training: The process of teaching an AI system to perceive, interpret and learn from data. That way, the AI will later be capable of inferencing—making decisions based on information it's provided.

Section 15. Methodology

Bloomberg Intelligence's interactive market-sizing model helps provide growth forecasts for generative AI. This model, which will evolve on a regular basis, is still in its early stages, and we have provided Terminal clients an interactive calculator to test scenarios (available at BI INET <GO>).

The methodology is based on a bottom-up approach to forecast revenue for areas within hardware, software, digital ads, gaming, IT and business markets. Our forecasts for new segments are anchored to established end-markets that generative AI is likely to disrupt and create new revenue opportunities. The approximate calculations are driven by our assumptions around how generative AI could disrupt these existing end-markets to varying degrees. Figure 57 illustrates the existing end markets with growth assumptions for 2022-27 and 2027-32. Figure 58 shows BI's base-case penetration rates of new AI segments across these end-markets. Figure 59 depicts BI's base-case generative AI revenue forecasts, driven by AI penetration rates found in Figure 58.

Figure 57: Revenue and Growth Forecasts for Existing Technology End-Markets

Worldwide Revenue by Technology Group

(in billions of $)	2023	'23-'27 CAGR(%)	2027E	'27-'32 CAGR(%)	2032E
Hardware	$1,351	6%	$1,674	13%	$3,051
Devices	$916	1%	$969	6%	$1,297
Infrastructure	$434	13%	$705	20%	$1,753
Software	$973	12%	$1,529	13%	$2,846
Application Development & Deployment	$236	17%	$448	17%	$982
PaaS	$122	28%	$330	22%	$891
On-premise	$115	1%	$118	-5%	$92
Applications	$498	10%	$734	10%	$1,191
SaaS	$294	15%	$516	14%	$993
On-premise	$204	2%	$218	-2%	$197
System Infrastructure Software	$239	10%	$347	14%	$673
IaaS	$112	17%	$210	22%	$567
On-premise	$127	2%	$137	-5%	$106
IT Services	$827	5%	$1,023	6%	$1,348
Business Services	$389	6%	$485	6%	$649
Digital Ad Spending	$602	10%	$871	12%	$1,548
Search	$246	9%	$350	10%	$564
Display	$333	10%	$495	14%	$952
Classified and Other	$23	4%	$26	4%	$32
Cybersecurity Spending	$105	13%	$171	12%	$301
Gaming Spending	$266	7%	$356	8%	$522
Life Sciences Spending	$162	9%	$228	9%	$351
Education Technology Spending	$138	12%	$217	12%	$383
Total	$4,813	8%	$6,554	11%	$11,000

Bloomberg Intelligence BI

Source: Bloomberg Intelligence's forecasts based on data from IDC, eMarketer, and Statista

The hardware market is currently valued at $1.35 trillion, split into devices ($916 billion) and data-center infrastructure ($434 billion), based on IDC data. Bloomberg Intelligence expects new

segments for generative AI in this category are AI servers, AI storage and Generative AI infrastructure-as-a-service on the data-center side, and conversational AI and computer vision products on the devices side. The assumptions around the shift in spending to AI servers and storage from traditional servers and storage can be changed in the BI interactive calculator, available on the Terminal. Generative AI infrastructure-as-a-service is how we expect the training compute and storage capacity to be consumed on the cloud. Similarly, for inferencing, conversational AI products and computer vision should emerge as new categories in devices that could be used at home and in cars. For the $973 billion software market, we anticipate new categories to emerge, such as coding copilots, specialized virtual assistants, chatbots and drug discovery software.

Figure 58: Generative AI Penetration Rate (BI Base-Case Assumptions)

AI Penetration (%) Base-Case Assumptions	2023	2027E	2032E
Hardware			
Devices (Inference)			
Computer Vision AI Products	0.3%	2.0%	4.5%
Conversational AI Products	0.4%	5.5%	8.5%
Infrastructure (Training)			
AI Server	6.0%	10.5%	6.0%
AI Storage	2.5%	4.5%	3.3%
Generative AI Infrastructure as a Service			
Compute			
Internal Consumption	0.3%	2.9%	1.9%
Hyperscale Consumption	0.7%	7.0%	8.0%
Networking	0.8%	2.4%	2.5%
Inference/Fine-Tuning Cloud	0.5%	3.1%	5.2%
Software			
Coding, DevOps and Generative AI Workflows Software	0.2%	3.0%	7.0%
Specialized Generative AI Assistants Software			
Enterprise Applications	0.3%	1.8%	4.2%
Consumer/E-Commerce Applications	0.2%	1.2%	3.8%
Generative AI Workload Infrastructure Software	0.5%	4.0%	12.0%
Drug Discovery Software	0.0%	2.0%	10.0%
Cybersecurity Spending	0.0%	2.0%	5.0%
Education Spending	0.6%	2.0%	6.0%
Gaming Spending			
Virtual Goods	0.1%	2.5%	6.0%
Game Design Software	0.2%	4.5%	10.0%
IT Services	0.0%	2.0%	6.0%
Business Services	0.0%	2.0%	5.0%
Digital Ad Spending			
Search	1.0%	6.0%	12.0%
Videos	0.5%	5.0%	10.6%
Messaging	0.2%	1.5%	4.0%

Bloomberg Intelligence BI

Source: Bloomberg Intelligence

15.1 BI's Market-Sizing Conclusions

The 2022-32 forecast scenario for the generative AI market in Figure 59 is based on the size of the end-markets in Figure 57 and penetration rates highlighted in Figure 58. The CAGR assumptions for both-end markets and generative AI impact can be modified to come up with your own scenarios. For example, we assume in our base case that the data-center market ("Infrastructure" in Figure 57) is likely to expand at a 13% CAGR from 2022-27 and 20% from 2027-32. The training segment, which is tied to the data-center market, could hit a CAGR of 46% for 2027 and 29% for 2032, in our scenario. The training segment is further comprised of AI Server, AI Storage and Generative-AI-as-a-service segments. Generative-AI-as-a-service is broken down further into compute, networking, and inference/fine-tuning cloud. The respective market penetration assumptions can be found in Figure 58 above.

Figure 59: Generative AI Revenue Base-Case Forecast by Technology Segment

(in milllions of $)	2023	2027E	20232E	Implied 9 yr. CAGR (%)
Hardware	**$53,105**	**$286,903**	**$639,399**	**32%**
Devices (Inference)	$6,415	$72,703	$168,641	44%
Computer Vision AI Products	$2,749	$19,387	$58,376	40%
Conversational AI Products	$3,666	$53,315	$110,265	46%
Infrastructure (Training)	$46,690	$214,200	$470,758	29%
AI Server	$26,060	$73,984	$105,197	17%
AI Storage	$10,858	$31,707	$56,982	20%
Generative AI Infrastructure as a Service	$9,772	$108,509	$308,579	47%
Compute	$4,343	$69,756	$173,575	51%
Internal Consumption	$1,303	$20,434	$33,312	43%
Hyperscale Consumption	$3,040	$49,322	$140,263	53%
Networking	$3,257	$16,911	$43,832	33%
Inference/Fine-Tuning Cloud	$2,172	$21,843	$91,171	51%
Software	**$5,028**	**$61,680**	**$317,961**	**59%**
Specialized Generative AI Assistants	$2,489	$22,029	$95,259	50%
Enterprise Applications	$1,493	$13,217	$50,011	48%
Consumer/E-Commerce Applications	$995	$8,812	$45,248	53%
Coding, DevOps and Generative AI Workflows	$473	$13,436	$68,763	74%
Generative AI Workload Infrastructure Software	$1,195	$13,885	$80,788	60%
Generative AI Drug Discovery Software	$32	$4,561	$35,091	117%
Generative AI Based Cybersecurity Spending	$11	$3,419	$15,063	124%
Generative AI Education Spending	$829	$4,349	$22,996	45%
Generative AI Based Gaming Spending	**$533**	**$24,890**	**$83,591**	**75%**
Virtual Goods	$133	$8,889	$31,347	83%
Game Design Software	$399	$16,000	$52,244	72%
Generative AI Driven Ad Spending	**$4,624**	**$53,154**	**$206,693**	**53%**
Search	$2,458	$21,006	$67,661	45%
Videos	$1,666	$24,729	$100,941	58%
Messaging	$500	$7,419	$38,091	62%
Generative AI Focused IT Services	**$165**	**$20,451**	**$80,904**	**99%**
Generative AI Based Business Services	**$78**	**$9,705**	**$32,443**	**95%**
Total	$63,533	$456,782	$1,360,990	41%

Bloomberg Intelligence BI

Source: Bloomberg Intelligence's forecasts based on data from IDC, eMarketer and Statista

Research Coverage Team

Lead Analyst:

Mandeep Singh	**Global Technology**	**msingh15@bloomberg.net**

Contributing Analysts

Anurag Rana	**Software, Americas**	**arana4@bloomberg.net**
Nishant Chintala	**Technology, Americas**	**nchintala@bloomberg.net**
Breanne Dougherty	**Strategy**	**bdougherty25@bloomberg.net**
Masahiro Wakasugi	**Hardware, APAC**	**mwakasugi4@bloomberg.net**
Woo Jin Ho	**Hardware, Americas**	**who88@bloomberg.net**
Charles Shum	**Hardware, APAC**	**cshum2@bloomberg.net**
Steven Tseng	**Hardware, APAC**	**htseng18@bloomberg.net**
Tamlin Bason	**Software, EMEA**	**tbason3@bloomberg.net**
Sunil Rajgopal	**Software, Americas**	**srajgopal4@bloomberg.net**
Niraj Patel	**Software, Americas**	**npatel646@bloomberg.net**
Matthew Schettenhelm	**Litigation/Policy, Americas**	**mschettenhel@bloomberg.net**
Kunjan Sobhani	**Hardware, Americas**	**ksobhani@bloomberg.net**
Kevin Tsao	**Software, Americas**	**ktsao14@bloomberg.net**

Copyright & Disclaimer

Copyright

Disclaimer

The data included in these materials are for illustrative purposes only. The BLOOMBERG TERMINAL service and Bloomberg data products (the "Services") are owned and distributed by Bloomberg Finance L.P. ("BFLP") except (i) in Argentina, Australia and certain jurisdictions in the Pacific Islands, Bermuda, China, India, Japan, Korea and New Zealand, where Bloomberg L.P. and its subsidiaries ("BLP") distribute these products, and (ii) in Singapore and the jurisdictions serviced by Bloomberg's Singapore office, where a subsidiary of BFLP distributes these products. BLP provides BFLP and its subsidiaries with global marketing and operational support and service. Certain features, functions, products and services are available only to sophisticated investors and only where permitted. BFLP, BLP and their affiliates do not guarantee the accuracy of prices or other information in the Services. Nothing in the Services shall constitute or be construed as an offering of financial instruments by BFLP, BLP or their affiliates, or as investment advice or recommendations by BFLP, BLP or their affiliates of an investment strategy or whether or not to "buy", "sell" or "hold" an investment. Information available via the Services should not be considered as information sufficient upon which to base an investment decision. The following are trademarks and service marks of BFLP, a Delaware limited partnership, or its subsidiaries: BLOOMBERG, BLOOMBERG ANYWHERE, BLOOMBERG MARKETS, BLOOMBERG NEWS, BLOOMBERG PROFESSIONAL, BLOOMBERG TERMINAL and BLOOMBERG.COM. Absence of any trademark or service mark from this list does not waive Bloomberg's intellectual property rights in that name, mark or logo. All rights reserved. © 2024 Bloomberg.

Bloomberg Intelligence is a service provided by Bloomberg Finance L.P. and its affiliates. Bloomberg Intelligence likewise shall not constitute, nor be construed as, investment advice or investment recommendations, or as information sufficient upon which to base an investment decision. The Bloomberg Intelligence function, and the information provided by Bloomberg Intelligence, is impersonal and is not based on the consideration of any customer's individual circumstances. You should determine on your own whether you agree with Bloomberg Intelligence. Bloomberg Intelligence Credit and Company research is offered only in certain jurisdictions. Bloomberg Intelligence should not be construed as tax or accounting advice or as a service designed to facilitate any Bloomberg Intelligence subscriber's compliance with its tax, accounting, or other legal obligations. Employees involved in Bloomberg Intelligence may hold positions in the securities analyzed or discussed on Bloomberg Intelligence.

About Bloomberg Intelligence

Your go-to resource for making better investment decisions, faster.

Bloomberg Intelligence (BI) research delivers an independent perspective providing interactive data and research across industries and global markets, plus insights into company fundamentals. The BI, team of 475 research professionals is here to help clients make more informed decisions in the rapidly moving investment landscape.

BI's coverage spans all major global markets, more than 135 industries and 2,000 companies, while considering multiple strategic, equity and credit perspectives. In addition, BI has dedicated teams focused on analyzing the impact of government policy, litigation and ESG.

BI is also a leading Terminal resource for interactive data. Aggregated, from proprietary Bloomberg sources and 500 independent data contributors, the unique combination of data and research is organized to allow clients to more quickly understand trends impacting the markets and the underlying securities.

Bloomberg Intelligence is available exclusively for Bloomberg Terminal® subscribers, available on the Terminal and the Bloomberg Professional App.

Take the next step.

For additional information,

press the <HELP> key twice on the Bloomberg Terminal®.

Beijing
+86 10 6649 7500

Hong Kong
+852 2977 6000

New York
+1 212 318 2000

Singapore
+65 6212 1000

Dubai
+971 4 3641000

London
+44 20 7330 7500

San Francisco
+1 415 912 2960

Sydney
+61 2 9777 86 00

Frankfurt
+49 69 92041210

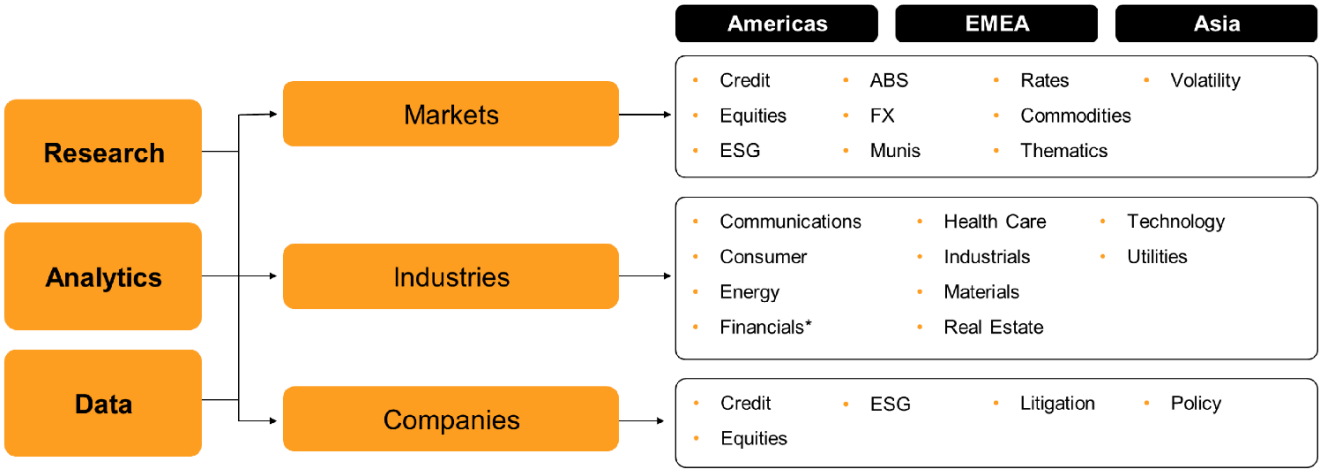
Mumbai
+91 22 6120 3600

Sao Paulo
+55 11 2395 9000

Tokyo
+81 3 4565 8900

Bloomberg Intelligence

Research, analytics and data tools to help you make informed investment decisions



Bloomberg Intelligence by the Numbers.

475
research professionals

135+
industries

500+
data contributors

2,000+
companies

21
markets covered