RBC Capital Markets

Generative AI Update

RBC Imagine™

Thinking further forward

EQUITY RESEARCH | MARCH 19, 2024

For Required Non-U.S. Analyst and Conflicts Disclosures, see page 88.

Disseminated: Mar 19, 2024 19:55EDT; Produced: Mar 19, 2024 19:55EDT

EQUITY I **RESEARCH**

RBC Imagine™: Generative AI Update

March 19, 2024

RBC Capital Markets, LLC

Jonathan Atkin (Analyst) (415) 243-7166 jonathan.atkin@rbccm.com

Rishi Jaluria (Analyst) (415) 633-8798 rishi.jaluria@rbccm.com

Matthew Hedberg (Analyst) (612) 313-1293 matthew.hedberg@rbccm.com

Brad Erickson (Analyst) (503) 830-9488 brad.erickson@rbccm.com

Deane Dray (Analyst) (212) 428-6465 deane.dray@rbccm.com

Bora Lee (Analyst) (212) 618-7823 bora.lee@rbccm.com

Matthew Swanson (Analyst) (612) 313-1237 matthew.swanson@rbccm.com

Shelby Tucker (Analyst) (415) 428-6462 shelby.tucker@rbccm.com

Daniel R. Perlin (Analyst) (410) 625-6130 daniel.perlin@rbccm.com

Ashish Sabadra (Analyst) (415) 633-8659 ashish.sabadra@rbccm.com

Rashim Jain (AVP) (415) 633-8561 rashim.jain@rbccm.com

RBC Dominion Securities Inc.

Paul Treiber (Analyst) (416) 842-7811 paul.treiber@rbccm.com

RBC Europe Limited

Mark Fielding (Analyst) +44 20 7002 2128 mark.fielding@rbccm.com

Royal Bank of Canada, Sydney Branch

Garry Sherriff (Analyst) +61 2 9033-3022 garry.sherriff@rbccm.com

This report is priced as of market close March 15, 2024, ET.
All values in U.S. dollars unless otherwise noted.

For Required Non-U.S. Analyst and Conflicts Disclosures, see page 88.

RBC Imagine™
Thinking further forward

Capital Markets

Table of Contents

We would like to thank Saurabh Ajmera for his contributions to this report.

Section 1

Executive Summary and Key Highlights

RBC Capital Markets

Broad Themes

Generative AI is a type of artificial intelligence technology that can produce various types of content including text, imagery, audio and synthetic data. To train AI models, companies pack thousands of GPUs into datacenters and run them at full capacity for extended periods of time, consuming tremendous amounts of electricity. These developments have multi-faceted implications for datacenters, semiconductor and equipment players, software companies, utilities and energy providers, and regulators and policymakers.

1. **Strong Revenue Trends:** Most AI-leveraged companies have posted strong top-line trends. AI features as a major growth driver for companies such as: AMZN, DLR, GOOG, INTC, META, MSFT, MRVL, NVDA, ORCL, SMCI, DELL and VRT.

2. **AI Investments:** AI is seeing integration into multiple aspects of business operations or product offerings, including cloud platforms & datacenters. At the same time, AI is creating new system design challenges and driving innovation around power consumption and cooling requirements. AWS, Microsoft, Google and Meta are investing heavily in generative AI and Large Language Model (LLM) capabilities, aiming to lead in AI-driven services and infrastructure.

3. **AI Monetization:** Companies are at various stages in determining paths toward monetizing AI, such as offering AI-based services, selling AI-powered products, and leveraging AI for efficiency gains. RBC sees two primary paths toward monetization (direct and indirect). Away from the hyperscalers, enterprises are assessing their AI strategies, and ROIs are largely unknown. RBC's recent work on this topic can be found in *RBC Imagine™: The Software Investor's Handbook to Generative AI*.

4. **GPU Availability**: AI is driving booming demand for advanced chips and accelerating innovation cycles as vendors attempt to meet supply bottlenecks. To provide a sense of scale, AWS, Google and Meta have each referenced >20K GPUs per AI cluster. NVIDIA has achieved notable early progress, particularly with the introduction of the H100 GPU. Elsewhere, AMD introduced the MI300 accelerator family in late 2023, accompanied by aggressive production. While Dell noted that lead times for the H100 have shown improvement, HPE has indicated that the lack of GPUs hindered its revenue growth.

5. **Capex Growth and Datacenter Expansion:** Companies are expanding their datacenter infrastructure to support AI demand while continuing to address their growth needs related to cloud computing, SaaS, e-commerce, social networking, or other services. These include existing hyperscalers, software players, and independent GPU-as-a-service operators. As such, infrastructure-related capex is growing by 20% or more in 2024.

Broad Themes (continued)

6. **AI Topology:** Much of the MW growth in deployments has been driven by requirements related to LLMs. Many LLMs are being deployed in or near existing cloud availability zones by the major CSPs, but at times are placed in more remote areas by CSPs and smaller AI entrants. Topologies around AI inferencing are still evolving, from far edge deployments (even at user devices) to smaller- and medium-sized datacenters closer to the core, depending on the use case. **While model sizes have been expanding exponentially, there is growing interest in smaller model frameworks (e.g., Llama2, phi-2, Orca-2, mistral 7B and various "expert models" or "nimble models") that can deliver a strong user experience with fewer compute resources and energy usage.**

7. **Strategic Partnerships:** Companies formed strategic partnerships to leverage each other's strengths and expertise in AI and related technologies. Examples include Google/Anthropic, Amazon/Anthropic, Google/NVIDIA, Compass/Schneider, Meta/Microsoft, Meta/Dell, Microsoft/Oracle, Microsoft/Mistral, AMD/Mipsology, Oracle/NVIDIA, Cloudflare/NVIDIA, Amazon/Intel, AMD/Microsoft, AMD/Oracle, and AMD/Meta.

8. **Energy Implications:** Energy requirements around AI are placing a greater focus on energy procurement and sustainability as datacenter-based IT increases its percentage usage of the electrical grid from currently ~3% toward mid-single digits over the coming years. Current constraints on the grid and availability of renewable or carbon-free sources are extending the reliance on natural gas. Over time, newer approaches, such as small modular reactors, could see greater focus. For further details, please see the RBC ESG Stratify™ green AI report and RBC Imagine™ utilities enabling AI report.

9. **Datacenter Infrastructure/Industrial Technology Implications:** Within the datacenter, companies are proactively adjusting approaches to accommodate AI workloads that have higher rack densities. This includes refining their use of established containment technologies to employing newer liquid-cooling solutions. We find that many contemporary base designs are compatible with or can be modified to meet AI requirements. That said, direct-to-chip and other liquid cooling approached will increasingly be employed. Further details on liquid cooling, CRAC, PDU, and UPS players can be found in our RBC Imagine™ Multi-Industry report.

10. **Top AI-related Beneficiaries:** Within RBC's current coverage universe, stocks that are substantially affected by AI-related developments include: Accenture (ACN), AES Corp (AES), Amazon (AMZN), Digital Bridge (DBRG), Digital Realty (DLR), Eaton (ETN), Go Daddy (GDDY), Macquarie Technologies (MAQ), Meta Platforms (META), Microsoft (MSFT), Moody's (MCO), NextDC (NXT), nVent (NVT), Schneider Electric (SBGSY), and Verisk (VRSK).

Semiconductor Players – Revenue and Commercial Trends

1. **AMD:** In 4QFY23, AMD's datacenter segment revenue grew 38% year-over-year and 43% sequentially to a record $2.3 billion. Both server CPU and datacenter GPU sales set quarterly and annual revenue records. Despite a soft overall demand environment, server CPU revenue increased year-over-year and sequentially as North American hyperscalers expanded fourth-gen EPYC processor deployments. More than 55 AMD-powered AI, HPC, and general-purpose cloud instances were brought into preview or general availability in the fourth quarter. Enterprise sales accelerated significantly in the quarter as AMD built momentum with Forbes 2000 customers. A growing number of customers are adopting EPYC CPUs for inferencing workloads. The datacenter GPU business exceeded $400 million expectations, driven by a faster ramp for the MI300x. In AI PCs, AMD launched the Ryzen 8000G series processors the first desktop CPUs with an integrated AI engine.

2. **Intel:** In 4QFY23, DCAI revenue was $4 billion, up 4% sequentially. Intel is anticipating CPU compute cores growth to normalize and expects its discrete accelerator portfolio to gain traction, with over $2 billion in the pipeline as 2024 progresses. The company sees AI workload as a driving force behind the $1 trillion semiconductor Total Addressable Market (TAM) by 2030. Intel's Gaudi2 AI accelerators have shown price performance leadership against popular GPUs, with Gaudi3 promising four times the processing power and double the networking bandwidth. Intel will also be producing custom chips for Microsoft, as part of a deal worth more than $15 billion. In 4Q, Intel launched Intel Core Ultra, with dedicated acceleration capabilities across the CPU, GPU and Neural Processing Unit (NPU). OpenVINO adoption grew by 60% sequentially in Q4 and today is a core software layer for AI inferencing on the edge, on the PC and in the datacenter. Intel launched its next-generation 5th Gen Xeon processors. This is the most power-efficient, highest-performant and security-enabled Xeon Intel that the company claims to have delivered, offering a 21% average performance gain over the previous generation.

3. **Marvel Technology:** In 4QFY24, revenue was $1.427 billion, exceeding the midpoint of guidance, growing 1% on a year-over-year and sequential basis. Datacenters were the largest end market, driving 54% of total revenue. The next largest was enterprise networking with 19%, followed by carrier infrastructure at 12%. GAAP gross margin was 46.6%. Non-GAAP gross margin was 63.9%, growing 330 basis points sequentially, driven by a product mix improvement. 800-gig PAM solutions led growth in 4Q. Marvell also benefited from higher sequential demand for storage products as that portion of its datacenter end market continues its recovery. AI was a key driver of datacenter growth in FY24, contributing over 10% of total company revenue vs. 3% in the prior year. Their momentum accelerated throughout the fiscal year with AI revenue well over $200 million in 4QFY24, driven mostly from Optics.

4. **NVIDIA:** Revenue for 4QFY24 was $22.1 billion, up 265% from a year ago and 22% sequentially. Fiscal year revenue was $60.9 billion, up 126% from a year ago. 4Q datacenter segment revenue was a record $18.4 billion, up 27% from the previous quarter and up 409% from a year ago. Full-year revenue rose 217% to a record $47.5 billion. These increases reflect higher shipments of the NVIDIA Hopper GPU computing platform used for the training and inference of LLMs, recommendation engines, and generative AI applications, along with InfiniBand networking solutions. In the fourth quarter, large cloud providers represented more than half of the datacenter revenue, supporting both internal workloads and external customers. The company estimates that 40% of its revenue now comes from inference workloads.

Semiconductor Players – Revenue and Commercial Trends (continued)

NVIDIA:

- At the GTC conference in March 2024, NVIDIA announced its latest product lineup focused on advancing AI, HPC, and cloud services. The new offerings include the NVIDIA Blackwell architecture, featuring the DGX GB200 and DGX B200 servers within the DGX SuperPOD system; the NVIDIA Quantum-X800 InfiniBand and NVIDIA Spectrum-X800 Ethernet switches; and the advanced networking capabilities offered by eight NVIDIA ConnectX-7 NICs and two BlueField-3 DPUs.

- New Quantum-X800 and Spectrum-X800 series switches, both capable of managing extreme performance demands for AI-intensive infrastructure. Each platform offers end-to-end throughput capacities of 800Gb/s, redefining network performance standards in AI, cloud, data processing, and HPC applications. Initial adopters of Quantum InfiniBand and Spectrum-X Ethernet include Microsoft Azure, Oracle Cloud Infrastructure and CoreWeave. A range of infrastructure and system vendors worldwide will make these state-of-the-art switches available, including Aivres, DDN, Dell Technologies, Eviden, Hitachi Vantara, Hewlett Packard Enterprise, Lenovo, Supermicro, and VAST Data.

- Blackwell-based products will be available from partners starting later this year. AWS, Google Cloud, Microsoft Azure and Oracle Cloud Infrastructure will be among the first cloud service providers to offer Blackwell-powered instances, as will NVIDIA Cloud Partner program companies Applied Digital, CoreWeave, Crusoe, IBM Cloud and Lambda. Sovereign AI clouds will also provide Blackwell-based cloud services and infrastructure, including Indosat Ooredoo Hutchinson, Nebius, Nexgen Cloud, Oracle EU Sovereign Cloud, the Oracle US, UK, and Australian government clouds, Scaleway, Singtel, Yotta Data Services' Shakti Cloud and YTL Power International.

- GB200 will also be available on NVIDIA DGX Cloud. AWS, Google Cloud and Oracle Cloud Infrastructure plan to host new NVIDIA Grace Blackwell-based instances later this year.

- Cisco, Dell, Hewlett Packard Enterprise, Lenovo and Supermicro are expected to deliver a wide range of servers based on Blackwell products, as are Aivres, ASRock Rack, ASUS, Eviden, Foxconn, GIGABYTE, Inventec, Pegatron, QCT, Wistron, Wiwynn and ZT Systems.

- Additionally, a growing network of software makers, including Ansys, Cadence and Synopsys—global leaders in engineering simulation— will use Blackwell-based processors to accelerate their software for designing and simulating electrical, mechanical and manufacturing systems and parts.

Semiconductor and GPU Developments at the Major Hyperscalers

1. **Amazon:**
   - Amazon unveiled the latest generation of its chips for model training (Trainium2) and inferencing (Graviton4).
   - Trainium2 is designed to deliver up to 4x better performance and 2x better energy efficiency than the first-generation Trainium. It can scale up to 100K chips in AWS' EC2 UltraCluster product.
   - The Arm-based Graviton4 is intended for inferencing. Amazon claims Graviton4 provides up to 30% better compute performance, 50% more cores and 75% more memory bandwidth than one previous-generation Graviton processor, Graviton3. running on Amazon EC2.
   - Amazon is actively developing Gen AI apps across its businesses, with several already launched and others in development. The P5 instance was recently launched using the Nvidia H100 GPU.
   - AWS will offer NVIDIA Grace Blackwell GPU-based Amazon EC2 instances and NVIDIA DGX Cloud to accelerate performance of building and running inference on multi-trillion parameter LLMs. AWS will offer the NVIDIA Blackwell platform, featuring GB200 NVL72, with 72 Blackwell GPUs and 36 Grace CPUs interconnected by fifth-generation NVIDIA NVLink.

2. **Google:**
   - Google has announced a performance-optimized version of its Tensor Processing Unit (TPU) called the TPU v5p, designed to reduce the time commitment associated with training LLMs. Anthropic will use the latest TPU chip 'at scale' to train its generative artificial intelligence models, including its Claude LLM.

3. **Microsoft:**
   - Microsoft announced the general availability of its H100 virtual machines in 3QCY23.
   - Microsoft has announced two custom-built, in-house AI chips that could be used in datacenters. This includes the Azure Cobalt 100 CPU and the Azure Maia 100 AI Accelerator. The Microsoft Azure Maia AI Accelerator is optimized for AI tasks and generative AI. The chips will start to roll out in 1Q-2024 to Microsoft's datacenters, initially powering Microsoft Copilot and Azure OpenAI.
   - Microsoft has also announced offering AMD's MI300X GPU in preview to its customers.
   - Microsoft will also offer NVIDIA Grace Blackwell GB200 and advanced NVIDIA Quantum-X800 InfiniBand networking for Azure customers. Microsoft also announced the general availability of its Azure NC H100 v5 VM virtual machine (VM) based on the NVIDIA H100 NVL platform and designed for midrange training and inferencing.

4. **Oracle:**
   - Oracle has announced that it will offer the AMD 300 MIX GPU in its AI datacenters.

5. **Meta:**
   - Meta plans to significantly expand its compute infrastructure, targeting 350,000 H100s and a total of 600,000 H100 equivalents including other GPUs. Meta has also provided details of its two new AI datacenter scale clusters, each containing 24,576 Nvidia Tensor Core H100 GPUs, that the company is using to train its Llama 3 large language AI model.
   - Reuters reported that Meta plans to deploy a new version of a custom chip into its datacenters this year, according to an internal company document seen by Reuters.
   - Meta will also be using AMD MI300X in its AI datacenters.

Hyperscalers – Revenue and Commercial Trends

1. **Alibaba:** The revenue mix of Alibaba's Cloud Intelligence Group continues to improve as with the reduced contribution of low-margin, project-based contracts. The group posted growth of public cloud revenue from external customers, with revenue reaching RMB 28.1 billion in 3QFY24, an increase of 3%. The steady improvement in adjusted EBITA, which increased by 86% to RMB 2.4 billion, reflects an improving product mix through a focus on public cloud and operating efficiency.

2. **Amazon:** AWS revenue grew 13% year-over-year in 4QFY23, approaching an annualized revenue run rate of $100 billion. AWS added more than $1.1 billion in incremental quarter-over-quarter revenue. AWS' operating income was $7.2 billion, an increase of $2 billion year-over-year. The operating margin for the quarter was 29.6%, up more than 500 basis points year-over-year.

3. **Google:** Google Cloud revenues were $9.2 billion for the quarter, up 26%. The Cloud team is focused on bringing the benefits of Gemini, Google's AI technology, to enterprises and governments globally.

4. **Microsoft:** In 2QFY24, Microsoft Cloud revenue was $33.7 billion, exceeding expectations and growing 24%. The gross margin percentage was 72%, relatively unchanged year-over-year. Intelligent Cloud segment revenue was $25.9 billion, increasing 20%. Azure and other cloud revenue grew 30%,including growth from AI services. Microsoft has announced a new multiyear partnership with Mistral, the second company to offer a commercial language model on Azure after OpenAI.

5. **Oracle:** 3QFY24 Cloud revenue (SaaS and IaaS, excluding Cerner) reached $4.4 billion, up 26%, with total cloud revenue, including Cerner, at $5.1 billion. This quarter marked the first time total cloud revenue surpassed total license support revenue. Infrastructure revenues stood at $5.4 billion, up 13%, with OCI Gen2 infrastructure cloud services revenue growing 52% to an annualized $6.7 billion. OCI consumption revenue increased by 63%, limited only by ongoing supply constraints.

Hyperscalers – Investment and Capex Trends

1. **Amazon:** In 2023, Amazon's full year capex was $48.4 billion, which was down $10.2 billion year-over-year, primarily driven by lower spend on fulfillment and transportation. However, in 2024, Amazon anticipates capex to increase year-over-year, primarily driven by increased infrastructure capex, supporting growth of their AWS business, including additional investments in generative AI and LLMs. The mix of investment in 2023 was tied to infrastructure, mostly supporting AWS but also supporting its core Amazon businesses, which was about 60% of its spend. Consensus expectations for Amazon's capex for CY-2024/CY-2025 is ~$59 billion/$63 billion.

2. **Google:** Google's reported capex in 4QFY23 was $11 billion, driven overwhelmingly by investment in its technical infrastructure with the largest component for servers followed by datacenters. The step-up in capex in Q4 reflects Google's outlook for the extraordinary applications of AI to deliver for users, advertisers, developers, cloud enterprise customers, and governments globally and the long-term growth opportunities that AI offers. In 2024, Google expects investment in capex will be notably larger than in 2023. Consensus expectations for Google's capex for CY-2024/CY-2025 is ~$41.5 billion/$44 billion.

3. **Meta:** Capital expenditures, including principal payments on finance leases, were $7.9 billion in 4QFY23 driven by investments in servers, datacenters, and network infrastructure. Meta anticipates 2024 capex to be in the range of $30 billion to $37 billion, a $2 billion increase Y/Y at the midpoint, driven by investments in servers, including both AI and non-AI hardware, and datacenters. Consensus expectations for Meta's capex for CY-2024/CY-2025 is ~$35 billion/$37 billion.

4. **Microsoft:** 2QFY24 capital expenditures, including finance leases, were $11.5 billion, lower than expected due to a delay in delivery for a third-party capacity contract from Q2 to Q3. Cash paid for PP&E was $9.7 billion. These datacenter investments support cloud demand, including the need to scale AI infrastructure. For the 3Q outlook, capital expenditures are expected to increase materially on a sequential basis, driven by investments in cloud and AI infrastructure. Consensus expectations for Microsoft's capex for CY-2024/CY-2025 is ~$46 billion/$50 billion.

5. **Oracle:** Oracle spent $2.1 billion on capex in 3QFY24, with a total fiscal year forecast of $7 billion to $7.5 billion and an anticipation of approximately $10 billion in capex for fiscal year '25. Oracle is prioritizing the expansion of cloud capacity to address the backlog and demand. Consensus expectations for Oracle's capex for CY-2024/CY-2025 is ~$10.3 billion/$10.2 billion.

Highlights from the Datacenter Equipment Supply Chain

Recent commentary from Eaton, nVent, Schneider Electric, Supermicro, Wesco, Celestica, Bloom Energy and Vertiv

- **Growth Drivers:**
  - All companies cited growth drivers such as digital transformation, artificial intelligence, and the energy transition. These trends are expected to drive demand for their respective product offerings. Schneider Electric, Supermicro, nVent and Vertiv emphasized the importance of sustainability and the need for environmentally friendly solutions.
  - Datacenter segment demand is strong across the industry, with Schneider Electric, Wesco, Vertiv, Eaton, and Supermicro experiencing growth in this segment due to increased deployments.
  - Eaton, nVent, SuperMicro, Wesco, Schneider Electric, Celestica, and Bloom Energy have all reported double-digit organic growth in their datacenter and hyperscaler businesses.

- **Cooling:**
  - Vertiv, Eaton, nVent, SuperMicro, Wesco, Schneider Electric, Celestica, and Bloom Energy have all reported strong demand for their cooling and power solutions in the AI datacenter segment.
  - Supermicro anticipates that 20 percent or more of worldwide datacenters will need to and will move to liquid cooling in the next several years to efficiently cool datacenters that use the latest AI server technology.
  - nVent has been adding capacity for liquid cooling and it is looking to double its capacity.
  - Immersion cooling, while offering extremely high heat dissipation capacity, has comparatively more limited scalability.
  - Vertiv acquired CoolTera to address the market opportunity in the datacenter sector.

- **Power Density Requirements:**
  - Eaton, nVent and Supermicro highlighted solutions to address high-power density infrastructure requirements, while Vertiv focuses on power management and distribution to support demanding applications.

Energy and Utility Implications of Generative AI

1. RBC’s Utilities team finds that datacenters represent half of the U.S. power demand growth through 2026, increasing forecasts by over 50% in some regions. Furthermore, demand should accelerate beyond 2026 as the grid tries to catch up to a large backlog of projects.

2. The relative locational flexibility of large AI training datacenters opens up opportunities in markets that have spare generation and transmission capacity. However, all else being equal, established hubs retain a scale advantage.

3. Absent meaningful spare capacity, bringing new large-scale datacenters to market takes time. For example, delays in transmission capacity can add three to five years to a new datacenter timeline alone.

4. Datacenters should drive more demand for renewables, but gas still plays a substantial role. While clean energy is a key focus to hyperscalers, many of which have made 100% clean power commitments, reliable baseload gas generation will continue to play a prominent role in the electricity mix.

5. Many utilities plan to add a considerable amount of gas generation, in addition to renewables and storage, to their fleets to meet the growing demand over the coming years.

6. For further details, see
   - RBC ESG Stratify™ green AI report
   - RBC’s RBC Imagine™ utilities enabling AI report.

Information Services: Key GenAI Themes

- **Evolutionary Near-term/Revolutionary Longer-term**. The rollout of GenAI powered products started in late 2023 and has continued into 2024. Based on announcements from our covered companies, we expect GenAI powered products to provide improved content discovery, increased cross-selling/upselling, and market share growth. Although we expect material monetization to begin in 2025. We believe companies with proprietary content (examples: MCO, SPGI, MSCI, EFX, VRSK) will drive monetization and allow for premium pricing. While we expect near-term margin headwinds in FY24 due to elevated investments in GenAI, we expect our companies will be able to partially offset these costs through improved developer productivity and accelerated innovation through GenAI adoption.
- **Co-pilots to GenAI native.** First generation GenAI products have GenAI functionality embedded within existing applications that improve content discovery through enhanced semantic search, improving clients' productivity by summarizing content using LLMs, extracting structured data from unstructured content, and automating the presentation of the information by charting or reports. We expect next-gen products to be GenAI native, similar to Cloud native and mobile-first. This would result in new/enhanced use cases or analyses based on the data, which wouldn't have been possible historically.
- **Executive support and top-down push are key to success**. Many of our covered Information Services companies have received executive level support to develop/sell and incorporate internally GenAI related products. As highlighted by SPGI’s new AI leadership team and their AI accelerator program. In the case of MCO, the Chief Product Officer and Head of Moody's Analytics have been spearheading their GenAI programs.
- **Productivity efficiency to drive innovation, but margin expansion likely elusive**. Most Information Services companies are focused on improving developer efficiency using GenAI, which should enhance new product innovation and increase speed to market. MCO has highlighted in some cases a ~40% reduction in software development time using GenAI. They are also working on lowering operating costs and improving internal efficiencies, e.g., reducing customer support costs. GenAI helped lower FactSet's CallStreet earnings call transcript summaries time by 90%+, and FDS expects to reduce client calls for FactSet codes (~50% of the daily calls) which would drive significant cost savings. ADP expanded its call summarization AI-tool to its service associates and has already started to see productivity gains with shorter handle times and improved customer service levels. We expect these savings will likely be reinvested back in technology. With that said, we remain skeptical of a material margin expansion from GenAI, similar to the elusive Cloud savings.
- **GenAI to improve the user experience**. Enhancing user experience is at the core of companies' GenAI strategies, with ADT planning to integrate GenAI technology more extensively into its ADT+ platform and MSCI developing an "Ask MSCI" search box offering. FDS’s AI Assistant to facilitate quicker decision-making. FICO utilizes AI/ML for fraud detection and is increasingly looking to incorporate explainable AI into its offerings that meet the requirements of regulators. Initial feedback and engagement with DNB's AiBE have been promising, with plans to combine it with Data Bricks and DUNS to create even better capabilities.

For more detailed information please see our note: Information Services GenAI Update - Product Launches, Monetization, Investments, and Cost Savings.

IT Services – AI Perspective

- AI/GenAI offer both opportunities and threats to the IT Services sector, which we break down into two views:
  - 1) “technology prerequisites” are required to fully achieve the benefits of what AI/GenAI offers and should drive incremental & derivative demand and
  - 2) businesses will have to self-cannibalize segments of their operations, specifically in digital CX and content moderation, in order to successfully pivot to a long-term viable growth model

- Prerequisites, in our opinion, include the correct infrastructure on which skilled talent can build full-stack solutions
  - Correct infrastructure is a “modern data foundation as part of the enterprise digital core” (using ACN’s terminology), which includes an **enterprise-wide data platform** and **cloud-based infrastructure**
  - Data must be **domain-specific, curated, cross-functional** and **shared** throughout the organization

- The BPO space historically has had to deal with 3%-8% automation every year, but GenAI has accelerated the need to offset increased automation (ranging as high as 30%+), thus likely resulting in revenue pressures in the near-term, but creating opportunities longer-term

- Content creation online is staggering and with the help of AI/GenAI the amount of content is expected to surge, creating opportunities for content moderation longer-term

Section 2

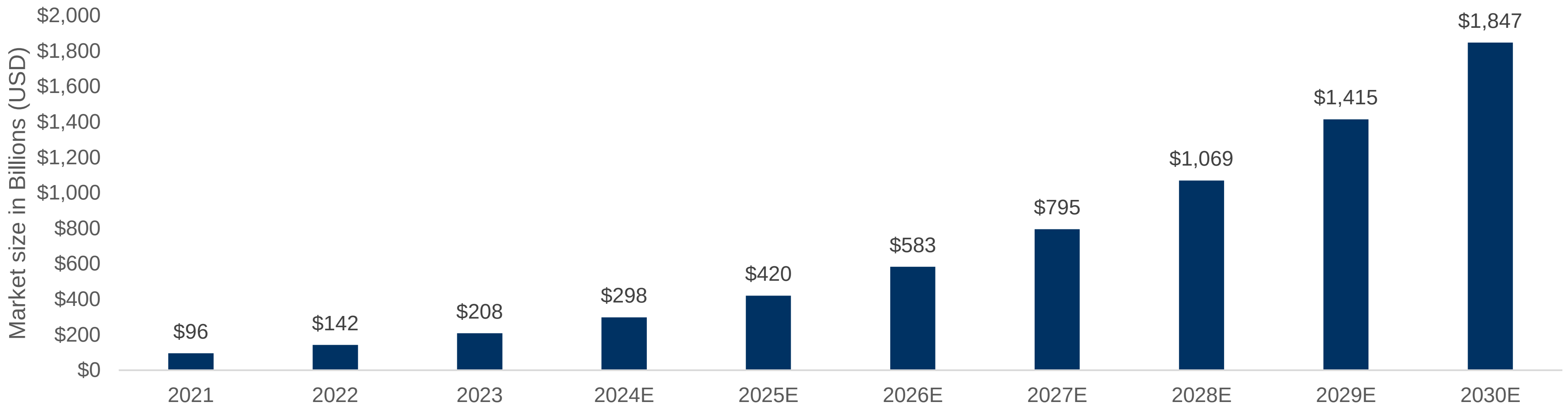
Generative AI Growth and Model Overview

- AI Adoption and Growth in Scale
- Generative AI Scale and Power Intensity
- Large Language Models (LLMs) – Training and Inferencing
- The Rise of Small Language Models (SLMs) and Nimble Models

RBC® Capital Markets

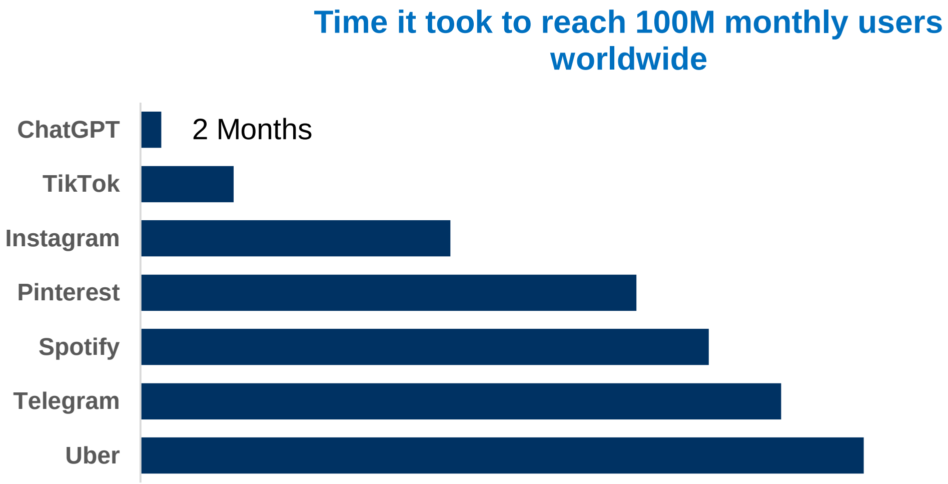
AI – Large-scale and Rapid Fast Adoption

AI GLOBAL MARKET SIZE (2021-2030)



Source: Statista

GENERATIVE AI ADOPTION IS THE FASTEST TECHNOLGY ADOPTION ON RECORD

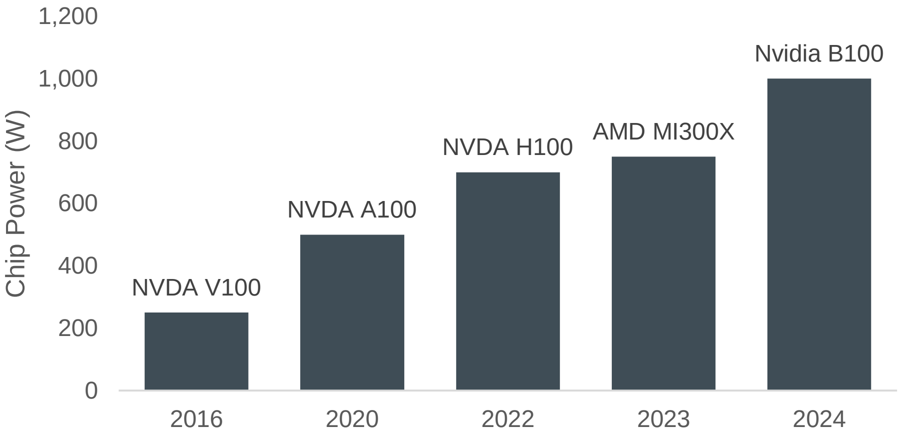


Source: IBM Global AI Adoption Index 2022, IDC Worldwide Artificial Intelligence Spending Guide

Generative AI Scale and Power Intensity

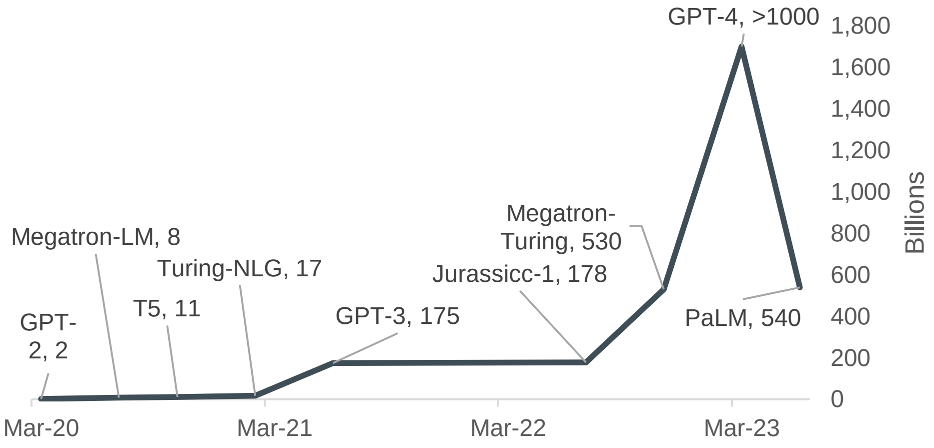
Chip power growth

Power consumption (W) per chip (1)



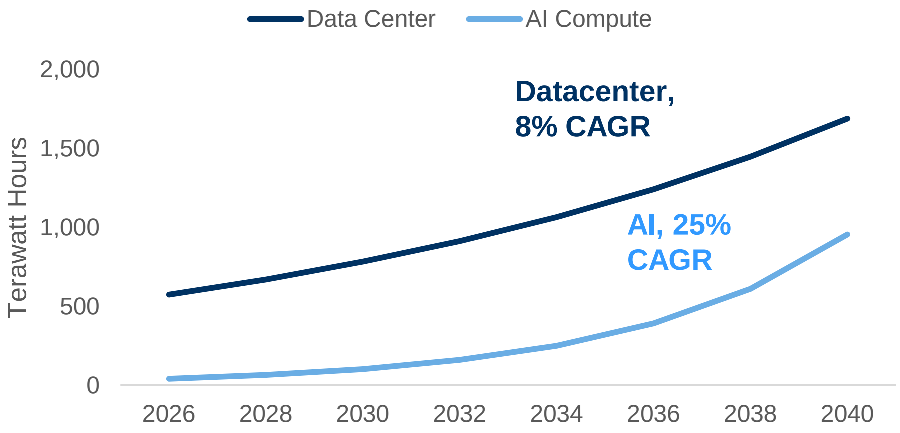
Large models are growing exponentially

Model size (in billions of parameter) (2)



Datacenters will need far more power

Datacenter Power Consumption (TWh) (3)

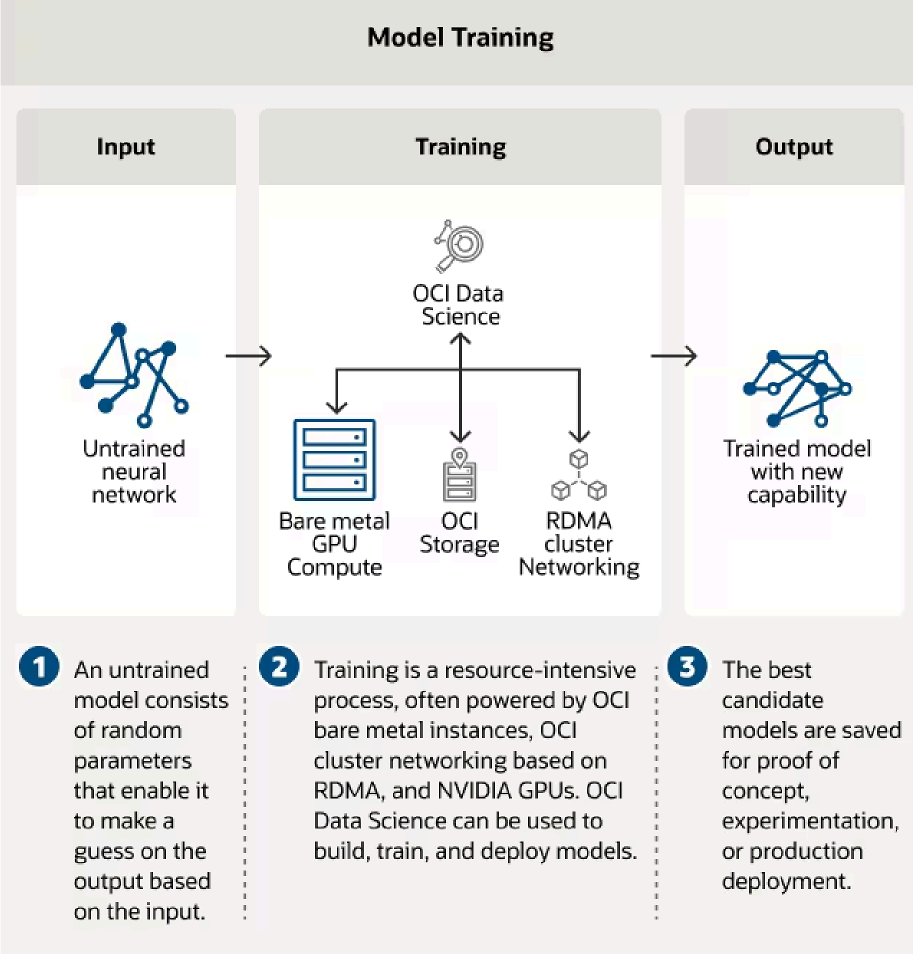


Small scale models

Launch	Small Language models	Model size (in billions of parameter)
Feb-23	Meta's Llama 7B	7
Feb-23	Meta's Llama 13B	13
Oct-23	Mistral 7b	7.3
Nov-23	MSFT Orca-2 7B	7
Nov-23	MSFT Orca-2 13B	13
Dec-23	MSFT Phi -2	2.7
Mar-24	Inflection 2.5	-

Sources: RBC Capital Markets, DigitalBridge, Intel, NVIDA, tom's hardware (2) AvidThink

Large Language Model (LLM) Training

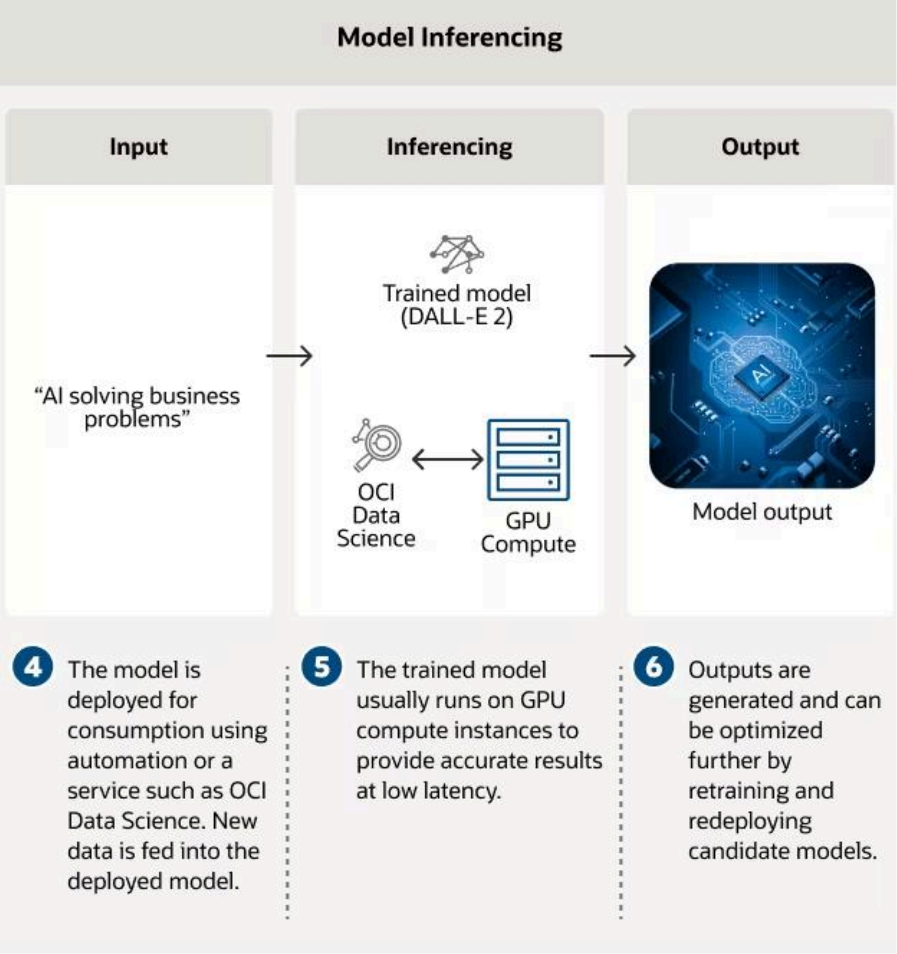


1. An untrained model consists of random parameters that enable it to make a guess on the output based on the input.
2. Training is a resource-intensive process, often powered by OCI bare metal instances, OCI cluster networking based on RDMA, and NVIDIA GPUs. OCI Data Science can be used to build, train, and deploy models.
3. The best candidate models are saved for proof of concept, experimentation, or production deployment.

- LLMs are trained with billions of parameters that require an extensive amount of computational capacity on specialized chips known as GPUs. The process is both compute and bandwidth intensive.
- **Memory-bound inferencing:** LLMs require significant amounts of memory for weights and input data, leading to memory-bound inferencing.
- **Two distinct phases:** LLMs have two distinct phases: the prompt phase and token phase.
- **KV cache:** The KV cache is used during the token phase to replace a quadratic computation with a linear memory lookup, reducing computational intensity and improving performance.
- **Batching:** Batching is necessary to achieve good reuse of weights and reduce memory bandwidth requirements.
- **Model sizes:** While model sizes have been expanding exponentially, there is growing interest in smaller model frameworks (e.g., Llama2 and various “expert models” or “nimble models”) that can deliver an excellent user experience with a vastly smaller model.
- LLMs developed by the hyperscalers are mostly being deployed at or near existing cloud availability zones or campus clusters.
- In the case of GPU-as-a-Service players, LLMs are being deployed across a more varied set of locations, including tier 2 or remote areas.

Source: Oracle and RBC Capital Markets

Large Language Model (LLM) Inferencing


Model Inferencing
Input
Inferencing
Output
"AI solving business problems"
Trained model (DALL-E 2)
OCI Data Science
GPU Compute
AI
Model output
4
The model is deployed for consumption using automation or a service such as OCI Data Science. New data is fed into the deployed model.
5
The trained model usually runs on GPU compute instances to provide accurate results at low latency.
6
Outputs are generated and can be optimized further by retraining and redeploying candidate models.

- **Generative AI – Inference Models**
  - AI Inferencing topologies are still evolving, with locations ranging from far-edge/endpoints to smaller deployments at network POPs or medium-sized datacenters in smaller markets, depending on use case.
  - Inferencing for LLMs is split into two phases: prefill and decode.
  - The prefill phase is highly compute and memory intensive, requiring up to 10 Petaflops just to get to the first token.
  - LLMs Inference Decode: The decode phase is sensitive to latency, including network latency and memory bandwidth. Tokens are processed one at a time, requiring high speed and low latency to ensure good response time.

- **Distributed inferencing:** To handle large models and datasets, distributed inferencing is required, which introduces additional communication overhead and requires careful consideration of interconnect bandwidth and latency.

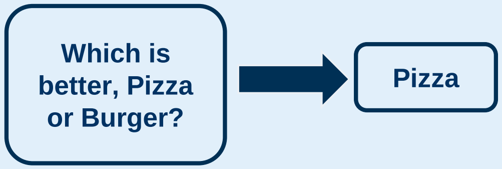
- **Next steps for improved performance:** Drive cost-effective inferencing through advancements in compute, memory, bandwidth, and software efficiency, as well as smaller models for edge accelerators to free up cloud capacity. Data compression and model pruning (to remove unimportant neural network connection) can reduce compute and memory requirements.

Source: Oracle and RBC Capital Markets

AI Inferencing and LLM Applications

Prompt Phase and Token Phase

Prompt Phase



Which is better, Pizza or Burger?

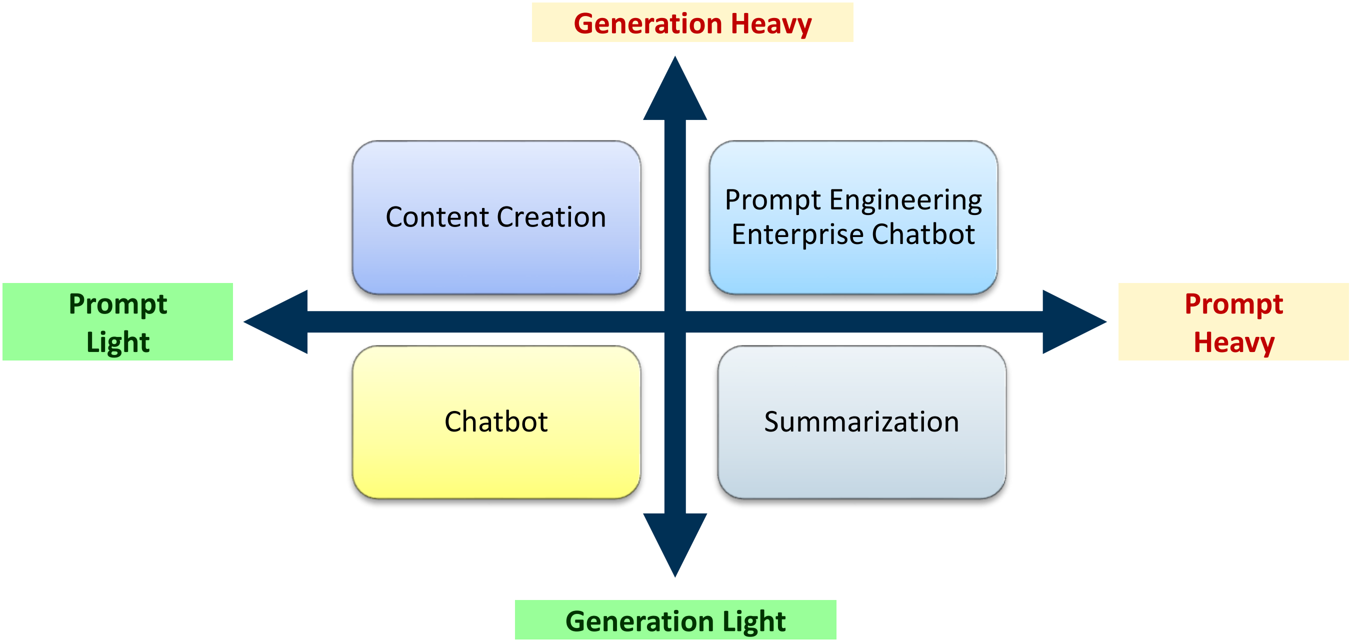
Pizza

- All input tokens are processed in parallel to generate the first output token
- Compute Intensive
- Power Intensive

Token Phase

- Serialized token generation
- Memory Intensive
- Lower power utilization

LLM Application Profiles



Source: Microsoft

Large Language Model (LLM) Training and Inferencing – Infrastructure Requirements

Stages	Description	Compute	Memory Capacity	Memory Bandwidth	Network Latency Sensitivity	Network bandwidth
Large Language Model (LLM) Training	LLMs are trained using a vast amount of data to learn billions of parameters. They require a huge amount of scattering gather transactions across thousands of GPUs to get sufficient flops to process these jobs. This process is super compute-intensive and requires a lot of network bandwidth.	High	Medium	Medium	Medium	High
LLMs Inference Prefill	Once an LLM has been trained, a base exists on which the AI can be used for practical purposes. By querying the LLM with a prompt, the AI model inference can generate a response, which could be an answer to a question, newly generated text, summarized text or a sentiment analysis report. Inference in LLMs involves two stages: prefill and decode. The prefill stage is where the tokens in the input prompt are processed in parallel. The prefill phase is very compute and memory intensive, requiring up to 10 Petaflops just to get to the first token.	High	Medium to High	Low	Low	Low
LLMs Inference Decode	The decode stage is where text is generated one 'token' at a time in an autoregressive manner. The decode phase is very sensitive to latency, including network latency and memory bandwidth as it requires a significant amount of memory to store the intermediate results.	Low	High	High	High	Low
Ranking and Recommendation Training	Ranking and Recommendation Training in LLMs is a process that involves training the model to rank and recommend items or choices based on the input it receives. Ranking and recommendation models, have larger vetting models that are mapped across many machines. This results in a high demand for network bandwidth due to the many collectives that are being instantiated in these models.	Low to Medium	Low to Medium	Medium	Medium	High
Ranking and Recommendation Inference	Ranking and Recommendation Inference in LLMs is the process where the trained model uses what it has learned to make predictions or recommendations. This is a high-demand, high-transactional workload. Although these models require a lot of memory capacity, efforts are made to minimize the amount of compute bandwidth and compute capacity required through tuning and training.	Low to Medium	High	Medium	Low	Low

AI Model Overview

GenAI Models by provider

	Text	Image	Audio or Music	3-D	Video	Protein or DNA sequences
Microsoft			VALL-E	RODIN Diffusion	GODIVA	MoLeR
OpenAI	GPT-4	DALL-E 3	Jukebox	Point-E	Sora	
Meta	LLaMA 2	Make-a-scene	AudioGem	Builder Bot	Make-a-video	ESMFold
Google/DeepMind	Gemini	Imagen	MusicLM	DreamFusion	Imagen Video	AlphaFold2
Stability AI	StableLM	Stable Diffusion 2	Dance Diffusion			LibreFold
Amazon	Lex		Deep Composer			
Apple				GAUDI		
NVIDIA	MT-NLG	Edify		Edify	Edify	MegaMol BART
Cohere	Family of LLMS					
Anthropic	Claude					
AI21	Jurassic-2					

Closed Source	Open Source	Closed Source, available through APIs

Source: McKinsey, Infratil Investors Day, RBC Capital Markets

The Rise of Small Language Models (SLMs) and Nimble Models

Smaller, cost-effective AI models provide an alternative to larger, resource-intensive models. Examples include: Meta’s Llama2 7b, Mistral 7b, and Microsoft’s Orca-2, Inflection 2.5, Microsoft’s Phi-2, Hugging Face’s Zephyr, and Google DistilBERT.

Advantages of SLMs

- Efficiency & cost-effectiveness: Offer significant performance with less computational power.
- Flexibility & customization: Ideal for specific applications with limited datasets.
- Security & privacy: Simplified structure reduces vulnerability to attacks.

Examples of SLM Innovation

- DistilBERT: Lighter and faster, with variants from Mini to MobileBERT.
- Orca 2: Enhanced reasoning abilities using synthetic data.
- Phi-2: Tailored for edge devices and the cloud, excels in text generation, language translation, and informative question-answering.
- GPT-Neo & GPT-J: Open-source alternatives with affordable computing needs.
- Microsoft claims certain SLMs, like Microsoft's Phi-2, demonstrate that it's possible to achieve state-of-the-art performance in areas such as mathematical reasoning and logical reasoning, comparable or even superior to larger models.
- Mistral AI's Mistral 7B and Mixtral 8x7B: Mixtral 8x7B, which is in beta, has nearly 47 billion parameters but processes input and generates output at the speed and cost of a 13-billion-parameter model, according to Mistral.

Challenges & Considerations

- LLMs vs. SLMs: High operational costs and potential bias in LLMs; reduced capabilities and need for fine-tuning in SLMs.
- Potential: Microsoft's SLMs demonstrating state-of-the-art performance in logical reasoning.

Section 3

Software Players and Generative AI Monetization

- Software Company Highlights
- AI Monetization Strategies
- Generative AI Use Cases
- Enterprise Adoption Challenges

RBC Capital Markets

Highlights – AI Deployment and Monetization by Software Companies

1. **Adobe (ADBE)**

   Adobe introduced generative credits as part of its Creative Cloud subscription plans, directly monetizing its AI innovations by integrating them into existing subscription models. The release of new AI models like Firefly and the monetization through generative credits showcase a strategy of enhancing and monetizing existing product offerings with AI capabilities. While credits are available, management's focus is on up-tiering customers deeper into the product stack. Over time we could see more consumption revenue as more compute intensive models such as video, 3D and vector ramp.

2. **IBM**

   IBM introduced WatsonX as an enterprise-ready AI data model platform. Monetization strategies include embedding AI into software solutions, third-party ecosystem partnerships, and leveraging consulting services for adoption. IBM also launched watsonx.gov for governance and compliance, with IBM forming the AI Alliance with Meta for advancing responsible AI. IBM recently introduced its Granite foundation models within Watsonx.ai for developers to build from and offer IP indemnity for its foundation models. IBM's Software and Consulting segments plan to capitalize on AI initiatives and cloud modernization projects.

3. **ServiceNow (NOW)**

   GenAI products have significantly contributed to new annual contract values. The company raised subscription revenue guidance for 2024 due to strong customer adoption of AI products (Pro Plus). ServiceNow charges extra for its Now Assist chatbot, available in the professional-plus pricing plan, highlighting a strategic approach to monetizing AI capabilities.

4. **Salesforce (CRM)**

   Salesforce is transforming customer operations for the AI future with its Einstein 1 Platform. Einstein Copilot provides a customizable AI assistant for CRM, enabling the use of private and trusted data without expensive model training. Management indicates the momentum in the UE+ bundle (Einstein 1) is showing substantial returns from the pricing uplift seen when attracting new customers onto the platform.

5. **Snowflake (SNOW)**

   The company focuses on simplifying AI for customer use, with significant demand for the Document AI and Copilot offerings. There is deep integration of AI into Snowflake, including a partnership with Mistral for hosting models and efforts to ensure accuracy in AI-generated answers. The company's monetization approach is to add value to existing offerings and potentially charge for advanced AI capabilities.

6. **Workday (WDAY)**

   Workday adopts a multipronged approach to AI monetization, including leveraging an innovation index at renewal times, selling AI-driven products like Talent Optimization, and developing an AI marketplace for partner solutions. This approach reflects a combination of adding value to existing subscriptions, selling dedicated AI solutions, and earning from a curated marketplace.

GenAI Monetization Strategies

- There are two primary paths to monetizing GenAI: direct and indirect. Direct monetization involves charging explicitly for GenAI products and features, while indirect monetization occurs when GenAI increases platform usage and boosts overall revenue.

- A hybrid approach that combines both direct and indirect monetization could be a successful way to monetize GenAI. This might involve providing basic GenAI functionality for free but limiting access to advanced features or charging for excessive usage.

- The effectiveness of monetizing GenAI depends on factors such as the specific industry, the type of GenAI being used, and the level of differentiation between the GenAI solution and competitors.

- Some companies may struggle to monetize GenAI due to intense competition from established players like Microsoft. In these cases, alternative pricing models or go-to-market strategies may be necessary.

- Monetizing GenAI requires careful consideration of customer needs, usage patterns, and willingness to pay. Companies must balance the benefits of GenAI against the cost of implementation and maintenance to ensure sustainable profitability.

More details here

GenAI Use Cases

Code Generation, Documentation, and QA

- Generative AI can write, complete, and vet sets of software code. It is handling bug fixes, test generation, and various types of documentation. It is also assisting non-developers by creating code from their natural language scenario-based queries.
- *Example Solutions: Code Snippets AI, ChatGPT, Google Bard, Tabnine*

Product and App Development

- Generative AI is used to code various kinds of apps and write product documentation for these apps. It is also going into projects like semiconductor chip development and design. Generative AI foundation models and APIs are also being used to develop new and fine-tuned generative AI models and products.
- *Example Solutions: MOSTLY AI, Stability AI, AI21 Labs, GPT-4*

Blog and Social Media Content Writing

- LLMs are capable of creating appropriate and creative content for blogs, social media accounts, product pages, and business websites. Many of these models enable users to give instructions on article tone and voice, input past written content from the brand, and add other specifications so new content is written in a way that sounds human and relevant to the brand’s audience.
- *Example Solutions: Jasper, Notion AI, Phrasee, HubSpot Content Assistant*

Inbound and Outbound Marketing Communication Workflows

- Generative AI solutions can create and send the content for these communications. In some cases, they can also automate the process of moving these contacts to the next stage of the customer lifecycle in a CRM platform. These types of assistive generative AI tools are increasingly popping up in both CRM and project management platforms.
- *Example Solutions: Twain, Salesforce Einstein GPT, HubSpot ChatSpot*

Graphic Design and Video Marketing

- Generative AI is capable of generating realistic images, animation, and audio that can be used for graphic design and video marketing projects. Some generative AI vendors also offer voice synthesis and AI avatars so you can create marketing videos without actors, video equipment, or video editing expertise.
- *Example Solutions: Diagram, Synthesia, Lightricks, Rephrase.ai*

Source: eweek.com

GenAI Use Cases

Entertainment Media Generation

- This type of technology is being used to create the graphics for movies and video games, the audio for music and podcast generation, and the characters for virtual storytelling and virtual reality experiences. With many of these tools, an actual human does not need to go on camera, edit footage, or even speak in order to create believable content.
- *Example Solutions: Stability AI's Stable Diffusion, Plask, Charisma, Latitude Voyage*

Performance Management and Coaching

- Generative AI can be used in several business and employee coaching scenarios. As an example, contact center call documentation and summarization, when combined with sentiment analysis, gives managers the information they need to assess current customer service rep performance and coach employees on ways to improve.
- *Example Solutions: Anthropic Claude, Gong, CoachHub AIMY*

Business Performance Reporting and Data Analytics

- Generative AI can work through massive amounts of text and data to quickly summarize the main points, it is becoming an important piece of business intelligence and performance reporting. It's especially useful for unstructured and qualitative data analytics, as these types of data usually require more processing before insights can be drawn.
- *Example Solutions: SparkBeyond Discovery, Dremio, Narrative BI*

Customer Support and Customer Experience

- Generative AI chatbots and virtual assistants can handle customer service questions at all hours of the day. Chatbots have been used for customer service for many years, but generative AI advancements are giving them additional resources to provide comprehensive and more human answers without the help of a human customer support representative.
- *Example Solutions: Gridspace, IBM Watson Assistant, UltimateGPT, Zendesk Advanced AI, Forethought SupportGPT*

Pharmaceutical Drug Discovery and Design

- Generative AI technology is being used to make drug discovery and design processes more efficient for new drugs. With this new development, scientists are beginning to generate novel molecules, more effectively discover disordered proteins, and design and predict clinical trial results.
- *Example Solutions: Insilico Medicine, Entos, Aqemia, New Equilibrium Biosciences*

Source: eweek.com

GenAI Use Cases

Medical Diagnostics and Imaging

- Image generation and editing tools are increasingly being used to optimize and zoom into medical images, allowing medical professionals to get a better and more realistic look at certain areas of the human body. Some tools even perform medical image analysis and basic diagnostics on their own.
- *Example Solutions: Paige.ai, Google Med-PaLM 2, ChatGPT and GPT-4*

Consumer-Friendly Synthetic Data Generation

- Generative AI can be used to create synthetic data copies of actual sensitive data, allowing analysts to analyze and derive insights from the copies without compromising data privacy or compliance. With these accurate data copies, data analysts and other members of an enterprise team can develop AI models and score those models without compromising actual business or consumer data.
- *Example Solutions: Syntho Engine, Synthesis AI, MOSTLY AI, Infinity AI*

Smart Manufacturing and Predictive Maintenance

- Generative AI is quickly becoming a staple in modern manufacturing, helping workers create more innovative designs and meet other production goals. In the realm of predictive maintenance, generative models can generate to-do lists and timelines, make workflow and repair suggestions, and simplify the process of assessing complex data from sensors and other parts of the assembly line.
- *Example Solutions: Biomatter, Clarifai, C3 Generative AI Product Suite*

Fraud Detection and Risk Management

- This type of technology can analyze large amounts of transaction or claims data, quickly summarizing and identifying any patterns or anomalies in that data. With these capabilities, generative AI is a great supporting tool for fraud detection, underwriting, and risk management in finance and insurance scenarios.
- *Example Solutions: Simplifai InsuranceGPT, Docugami, ChatGPT*

Optimized Enterprise Search and Knowledge Base

- Both internal and external search are benefitting from generative AI technology. For employees and other internal users of business tools, generative AI models can be used to scour, identify, and/or summarize enterprise resources when users are searching for certain information about their job or project. These tools are designed to not only search typical sources, like company files, but also company applications, messaging tools, and web properties.
- *Example Solutions: Glean, Coveo Relevance Generative Answering, Elasticsearch Relevance Engine*

Source: eweek.com

GenAI Use Cases

Professional Services

- Accenture is using generative AI to help its clients create smarter business strategies, roadmaps, and operations. Examples include helping a major oil and gas company implement tools from Microsoft Azure and OpenAI, designing an AI-powered search engine for Spain’s Ministry of Justice, and using generative AI to automatically review and triage emails for a multinational bank.

Life Sciences

- PathAI provides AI-driven solutions for precise pathology processes, ranging from lab services to deploying algorithms for clinical trials and diagnoses. Its AI models, trained and validated with over 15 million annotations, boost sample analysis efficiency and accuracy, helping assess therapy efficacy and expedite drug development for complex diseases.

Travel & Hospitality

- Expedia’s beta ChatGPT-powered travel planner lets users ask questions and get recommendations on travel, lodging, and activities. It also saves suggested hotels and venues through an intelligent shopping feature.

E-commerce & Retail

- Shopify now offers Shopify Magic to help retailers generate product descriptions and other product-related content with artificial intelligence.

Source: eweek.com

GenAI Use Cases

Medical Scribe

- By generating patient notes, creating summaries, and extracting key details, AI is helping free up skilled workers’ time to focus on patient care.

Field Engineer Assistant

- GenAI systems that can access large amounts of technical information, manuals, guides, notes and more. Engineers can then describe a problem and GenAI can help with trouble shooting guidance.

Marketing Assistant

- Marketing Assistant GenAI systems that can quickly and automatically generate personalized and contextualized marketing content to target individuals. By using natural language processing and machine learning techniques, it can create content that is more relevant and resonates more with the target users, leading to better conversion rates. Marketing professionals can then focus on strategy and optimize their campaigns.

Fintech & Software Development

- Stripe is using OpenAI’s GPT-4 to power better documentation, summarization, and query management for developers that use Stripe Docs. Stripe is also helping OpenAI and several other generative AI companies better monetize their products with Stripe Billing, Stripe Checkout, Stripe Tax, Revenue Recognition, and Link.

Source: eweek.com

Enterprise Adoption Challenges

High costs associated with GenAI

- Currently, GenAI workloads are expensive due to various factors such as GPU shortages and high levels of capex investments.
- However, as companies become more efficient and build their own hardware, costs will decrease. Additionally, monetization will ramp up, leading to better margins.

"Hallucinations" or incorrect responses from GenAI systems

- To minimize hallucinations, guardrails can be put in place to prevent LLMs from answering questions outside of their trained domain.
- Human intervention can also be used to review common cases of hallucination and retrain the model.

Data privacy and residency concerns

- Companies can use local models or open-source models deployed in a private cloud environment to ensure that their data does not train the central model.
- This can address both data privacy and data residency concerns.

Lack of domain expertise to fully leverage GenAI's capabilities

- Software companies can bring domain expertise to LLMs, which can help customers get 70% of the way to the finished product.
- This creates an opportunity for vertical software vendors, department-specific solutions, and use case-specific applications.

More details here

Section 4

Cloud/Hyperscale Financial Highlights

- Earnings Highlights
- Cloud Revenue Growth Trend at Major Cloud Service Providers
- Capex Trends at Major Hyperscale Service Providers

Amazon Web Services (AWS) – Earnings and Other Highlights

- **AWS Segment Performance:**
  - AWS revenue grew 13% year-over-year in Q4, approaching an annualized revenue run rate of $100 billion. AWS added more than $1.1 billion in incremental quarter-over-quarter revenue. AWS' operating income was $7.2 billion, an increase of $2 billion year-over-year. The operating margin for the quarter was 29.6%, up more than 500 basis points year-over-year.
- **Capex:**
  - Amazon defines its capital investments as a combination of capex plus equipment finance leases. In 2023, full-year capex was $48.4 billion, which was down $10.2 billion year-over-year, primarily driven by lower spend on fulfillment and transportation. In 2024, Amazon anticipates capex to increase year-over-year, primarily driven by increased infrastructure capex, to support growth of its AWS business, including additional investments in generative AI and LLMs.
- **Cloud Regions and Datacenters**
  - New AWS European Sovereign Cloud: Announced to meet EU compliance and offer operational independence for the public sector and regulated industries.
  - New AWS Canada West (Calgary) Region: Marked AWS' expansion as the first major cloud provider in Western Canada.
  - At the end of 2023, AWS had 105 Availability Zones across 33 regions globally, with plans for 12 more AZs and four more regions.
- **Artificial Intelligence:**
  - Amazon built custom AI training (Trainium) and inference (Inferentia) chips. Trainium2 offers four times faster training performance and three times more memory capacity versus the first generation of Trainium.
  - Amazon launched Bedrock, a service for companies to leverage an existing LLM, customize it with its own data, and leverage AWS' security other features all as a managed service. Bedrock has thousands of customers using the service after just a few months.
- **Designing Datacenter Networks for Generative AI:**
  - The advent of generative AI has necessitated the design of datacenter networks that can handle high throughput and low latencies. This has been addressed by launching instances with higher networking throughput.
  - Recently, the P5 instance was launched using the H100 GPU. It provides 3200 gigabits per second on a single EC2 instance.
  - AWS has developed UltraCluster 2.0 to cater to the massive growth in machine learning workloads, offering a flatter and wider network fabric optimized for P5 and future ML accelerators, significantly reducing latency and increasing bandwidth. Customers can scale up to 20,000 H100 GPUs in a single EC2 UltraCluster.

Source: Company reports

Microsoft – Earnings and Other Highlights

- **Cloud Revenue Growth:** Microsoft Cloud revenue was $33.7 billion, growing 24% Y/Y and 22% Y/Y in constant currency. The gross margin percentage was 72%, relatively unchanged year-over-year. Intelligent Cloud segment revenue was $25.9 billion, increasing 20% Y/Y and 19% Y/Y in constant currency, surpassing expectations with better-than-expected results across all businesses. Server products and cloud services revenue grew 22% Y/Y and 20% Y/Y in constant currency. Azure and other cloud services revenue grew 30% Y/Y and 28% Y/Y in constant currency, including 6 points of growth from AI services. For the 3Q outlook, Intelligent Cloud is expected to generate revenue of $26 billion to $26.3 billion, or growth between 18% and 19% Y/Y. Revenue growth will continue to be driven by Azure.

- **Capex:** Capital expenditures, including finance leases, were $11.5 billion, lower than expected due to a delay in delivery for a third-party capacity contract from Q2 to Q3. Cash paid for PP&E was $9.7 billion. These datacenter investments support cloud demand, including the need to scale AI infrastructure. For the 3Q outlook, capital expenditures are expected to increase materially on a sequential basis, driven by investments in cloud and AI infrastructure.

- **Gen AI, LLMs, SLMs:** Azure leverages AMD, NVIDIA, and its own silicon, Azure Maia. Azure AI provides access to a wide selection of foundation and open source models, including both LLMs and SLMs, all integrated with infrastructure, data, and tools on Azure. Azure AI now has 53,000 customers, over a third of whom are new to Azure over the past 12 months. Microsoft has also built SLMs, which offer performance comparable to larger models but are small enough to run on a laptop or mobile device.

- **Impact of Inferencing on Revenues:** Most of what is seen in the AI revenues is inferencing, including via API calls. Small batch training, such as fine-tuning, is a minor part. Growth was also seen in GitHub Copilot, and a growing number of third parties are using it in some smaller ways for training.

- **New AI Chips:** Microsoft has announced two custom-built in-house AI chips that could be used in datacenters. This includes the Azure Cobalt 100 CPU and the Azure Maia 100 AI Accelerator. The Microsoft Azure Maia AI Accelerator is optimized for AI tasks and generative AI. The chips will start to roll out in 1Q-2024 to Microsoft's datacenters, initially powering the company's services such as Microsoft Copilot or Azure OpenAI Service.

Source: Company reports

Google Cloud – Earnings and Other Highlights

- **Google Cloud Segment Performance:**
  - Google Cloud revenues were $9.2 billion for the quarter, up 26% Y/Y.
  - Full-year revenues of $33 billion were up 26% versus the prior year, with strong 4Q momentum.
  - The Cloud team is focused on bringing the benefits of Gemini to enterprises and governments globally.

- **Capital Expenditures:**
  - Google's reported capex in the fourth quarter was $11 billion, driven mostly by investment in technical infrastructure with the largest component for servers followed by datacenters.
  - The step-up in capex in Q4 reflects Google's outlook for AI applications to deliver for users, advertisers, developers, cloud enterprise customers, and governments globally.
  - In 2024, Google expects investment in capex will be notably larger than in 2023.

- **Generative AI:**
  - Management estimates that more than 70% of GenAI unicorns are using Google Cloud.
  - Google closed the year by launching Gemini, a new series of models aimed at fueling the next generation of advances. Gemini is engineered to understand and combine text, images, audio, video, and code in a natively multimodal way, and it can run on everything from mobile devices to datacenters.
  - Google is already experimenting with Gemini in Search, where it's making its Search Generative Experience, or SGE, faster for users. It has seen a 40% reduction in latency in English in the U.S.
  - Google's conversational AI tool Bard is now powered by Gemini Pro and is more capable at tasks such as like understanding, summarizing, reasoning, coding, and planning. It is now in over 40 languages and over 230 countries around the world. Looking ahead, Google will be rolling out an even more advanced version for subscribers powered by Gemini Ultra.

- **New Products, Vertex AI:**
  - Throughout 2023, Google introduced thousands of product advances, including broad GenAI capabilities across its AI infrastructure, its Vertex AI platform, and its new Duet AI agents. In Q4, this helped drive new and expanded relationships with many leading brands.
  - Vertex AI is an enterprise AI platform that helps customers discover, customize, augment, and deploy over 130 GenAI models. Vertex AI has seen API requests increase nearly 6x from H1 to H2 last year.

Source: Company reports

Meta – Earnings and Other Highlights

- **Compute Infrastructure Expansion:** The company anticipates future AI models will require substantially more computing power, continuing a trend of state-of-the-art models needing roughly 10x the compute each year. Investments will focus on advanced clusters, novel datacenter designs, and custom silicon. Meta plans to significantly expand its compute infrastructure, targeting 350,000 H100s and a total of 600,000 H100 equivalents including other GPUs. This expansion is a response to previous underestimations for GPU clusters, particularly noted with the introduction of Reels. Reuters has reported that Meta plans to deploy a new version of a custom chip into its datacenters this year.

- **Long-term Development Focus:** Meta’s strategy has been to build an open-source general infrastructure while keeping its specific product implementations proprietary. This includes its Llama models and industry tools such as PyTorch. Ongoing development efforts include research for future Llama models (5, 6, 7) aimed at achieving full general intelligence.

- **Leveraging Unique Data and Feedback Loops:** Meta benefits from vast amounts of publicly shared content and interaction data across its platforms, enhancing AI system improvements. The significance lies not just in the training corpus but in establishing effective feedback loops with users to refine AI services rapidly.

- **Capex:**
  - Capital Expenditures for 2023: Capex totaled $7.9 billion, driven by investments in servers, datacenters, and network infrastructure, reflecting Meta's commitment to AI and the metaverse as core components of its long-term vision.
  - 2024 Capital Expenditure Guidance: Anticipated to be between $30 billion and $37 billion, marking an increase from previous projections. Meta attributed this adjustment to a better understanding of AI capacity demands for foundational research and product development.

- **Meta GPU clusters:** Meta provided details of its two new AI datacenter scale clusters, each containing 24,576 Nvidia Tensor Core H100 GPUs, that the company is using to train its Llama 3 large language AI model. These clusters, which are an expansion from the previously used 16,000 Nvidia A100 GPUs, mark a significant step forward in supporting larger and more complex AI models. By the end of 2024, Meta aims to expand its infrastructure to include 350,000 Nvidia H100 GPUs, significantly increasing compute power. Meta has developed these clusters using its Open Rack v3 hardware, designed for flexibility and improved power and rack architecture in datacenters. This setup allows for power shelves to be installed anywhere in the rack, facilitating various rack configurations. Meta has also optimized the number of servers per rack within these clusters to achieve a balance of throughput capacity, rack count reduction, and power efficiency.

Source: Company reports

Oracle Cloud – Earnings and Other Highlights

- **Financial Performance:**
  - OCI is now the largest contributor to Oracle's revenue growth.
  - Cloud revenue (SaaS and IaaS, excluding Cerner) reached $4.4 billion, up 26% Y/Y, with total cloud revenue, including Cerner, at $5.1 billion.
  - Infrastructure revenues stood at $5.4 billion, up 13% Y/Y, with OCI Gen2 infrastructure cloud services revenue growing 52% Y/Y to an annualized $6.7 billion.
  - OCI consumption revenue increased by 63% Y/Y, limited only by ongoing supply constraints.

- **Capex:**
  - Oracle spent $2.1 billion on capex in Q3, with a total fiscal year forecast of $7 to $7.5 billion and an anticipation of approximately $10 billion in capex for fiscal year '25.
  - Oracle is prioritizing the expansion of cloud capacity to address its backlog and demand.

- **Cloud and Datacenter Expansion:**
  - Oracle now operates 68 customer-facing cloud regions, including 47 public cloud regions, with more under construction.
  - Oracle is aggressively expanding its datacenter footprint to meet surging demand for cloud services.
  - The Oracle Alloy initiative allows partners to offer customized cloud services using Oracle's infrastructure. This approach is gaining traction, particularly in Japan, and supports Oracle's vision of widespread adoption and customization of its cloud services.

- **AI and GenAI:**
  - Oracle and NVIDIA have expanded their partnership, with Oracle's Gen2 AI infrastructure business booming.
  - Oracle emphasized its strong position in GPU access and capabilities, ensuring they can meet the computational demands of its AI-driven services.
  - Management commented: "*We've got at least 40 new AI bookings that are over $1 billion that haven't come online yet*."

Source: Company reports

Alibaba Cloud – Earnings and Other Highlights

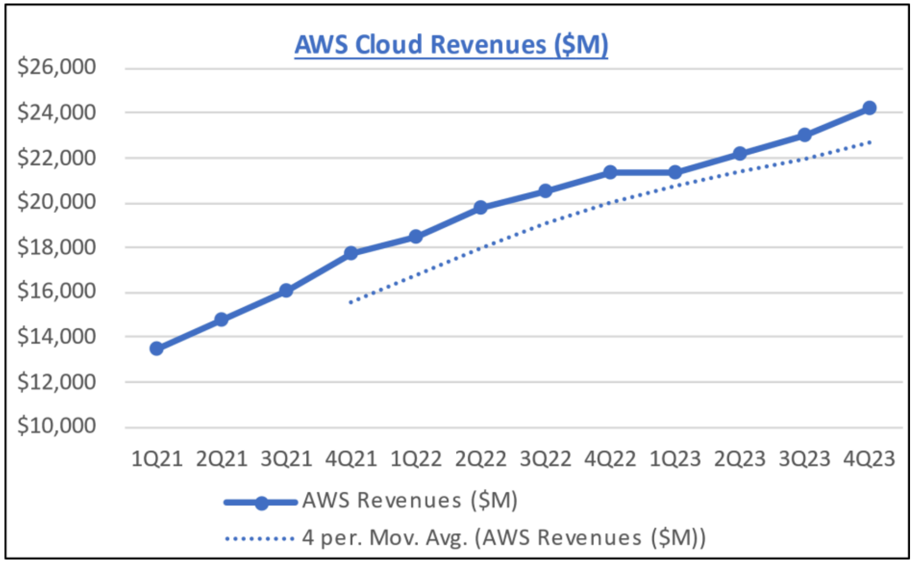
- **Cloud Growth and Demand:**
  - In F1Q-2024 / CY4Q-2023, Alibaba’s Cloud intelligence group revenue increased by 4% Y/Y, driven by expanding businesses and clients across various sectors.
  - Management indicated CDN demand decreased as online activities resumed after the pandemic.
  - The revenue growth was also partly offset by a significant contribution from one major client that led to a reduction in overall revenue.
  - Adjusted EBITDA experienced an impressive increase of 106%, mainly attributed to reduced costs for DingTalk, improvements in product mix, and enhanced operational efficiency (such as better server utilization).
  - There is considerable potential for cloud infrastructure adoption in China, which still lags behind the U.S. in terms of development.

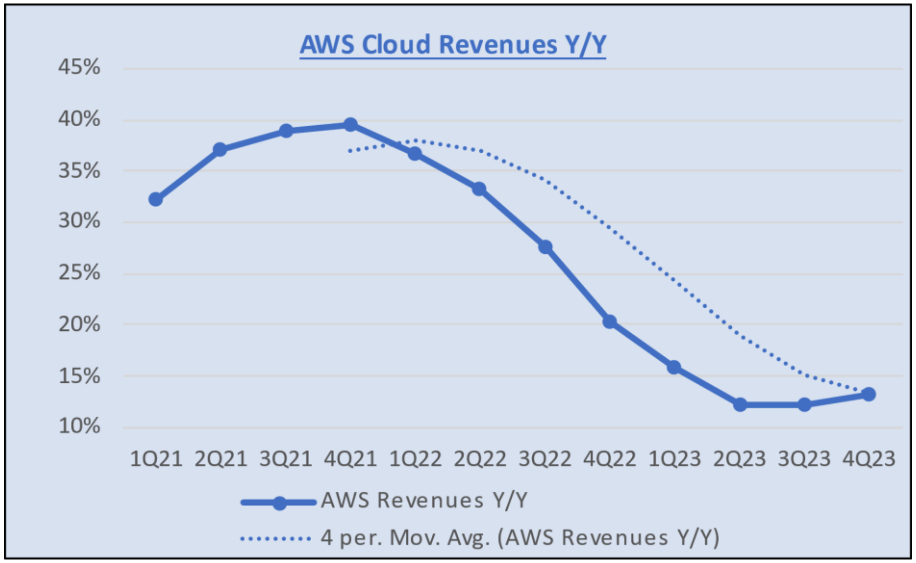
- **AI Opportunity**:
  - In the recent quarter, there was a significant demand for model training and associated AI services using cloud infrastructure. Management indicated that due to global supply chain constraints, this demand could only be partially met.
  - Alibaba believes that the opportunities presented by AI services are still in their infancy.
  - It has been developing its own LLM called Tongyi Qianwen and is testing ways to leverage this model to enhance search and advertising.
  - Alibaba has created a thriving open-source online community in China dedicated to models and related tools and services. This platform is highly popular among developers and currently hosts over 1,000 AI models, such as Meta's newly released Llama 2 and Alibaba's own open-source model.
  - Alibaba Cloud aims to capitalize on AI by offering high-performance, cost-effective infrastructure to other AI companies. The company believes that the AI landscape is evolving rapidly, with various models being developed and refined.
  - Alibaba plans to offer Model as a Service (MaaS) on top of its existing IaaS and PaaS infrastructure.
  - The company sees potential for greater synergy between Alibaba Cloud and the Taobao and Tmall Group, driven by AI.

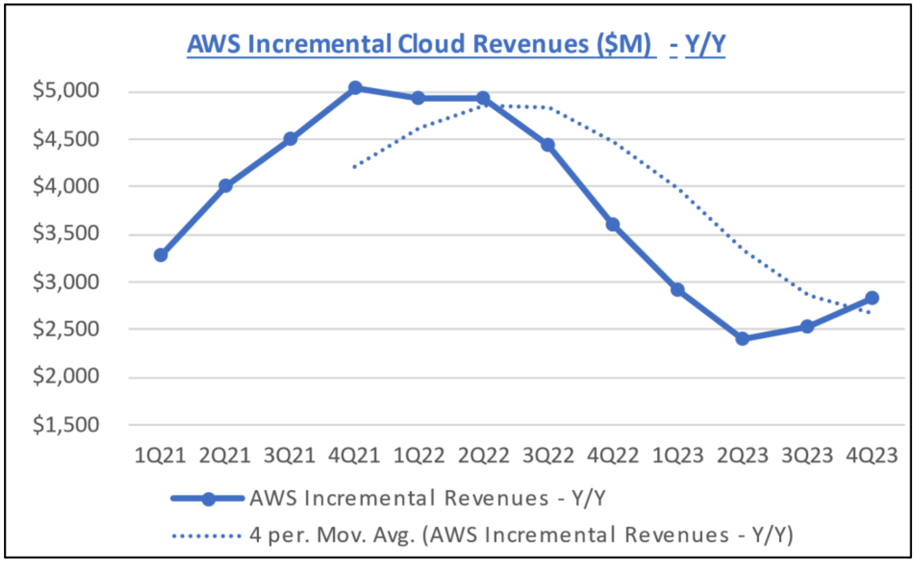
Source: Company reports

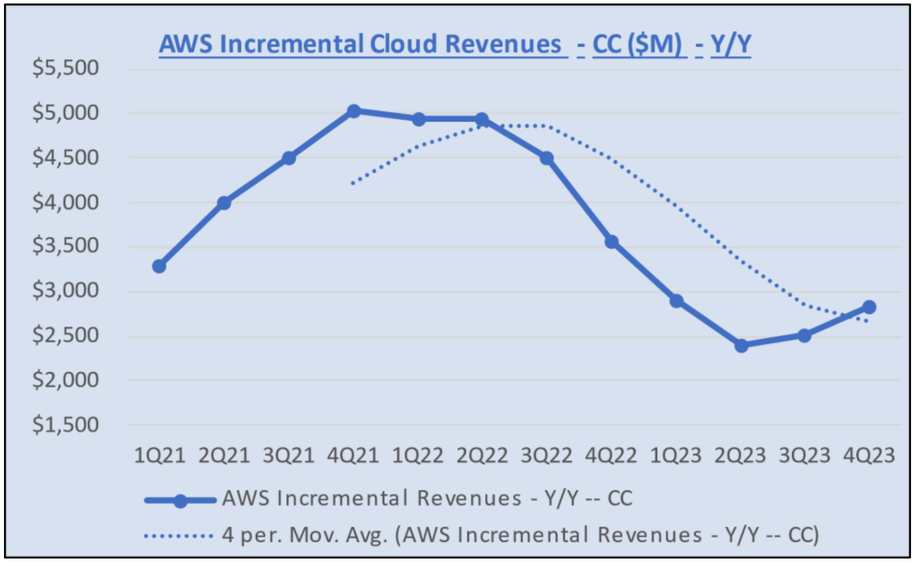
Cloud Revenue Growth – Amazon Web Services (AWS)

AWS revenue grew 13% Y/Y in 4Q23, approaching an annualized revenue run rate of $100 billion. AWS added more than $1.1 billion in incremental Q/Q revenue.


AWS Cloud Revenues ($M)


AWS Cloud Revenues Y/Y


AWS Incremental Cloud Revenues ($M) - Y/Y


AWS Incremental Cloud Revenues - CC ($M) - Y/Y

Source: Company reports, RBC Capital Markets

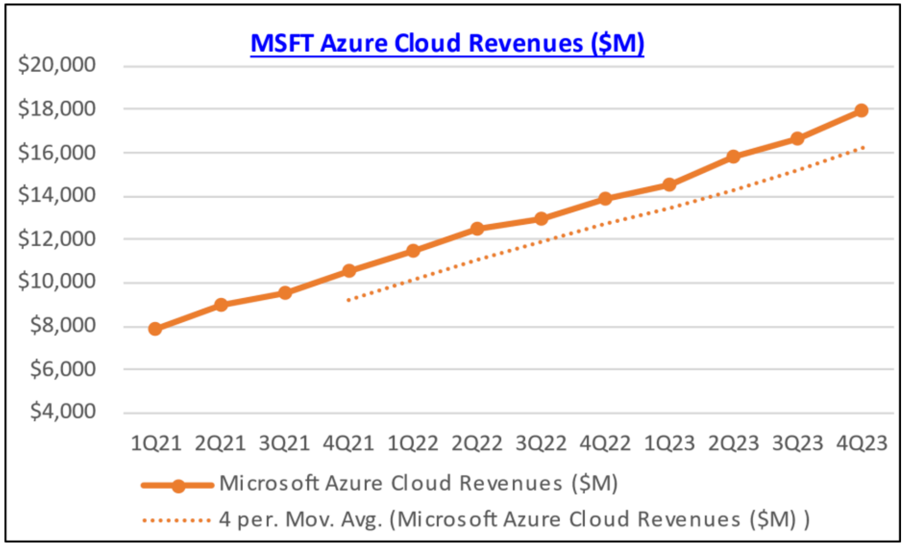
Cloud Revenue – Segment Definitions by Operator

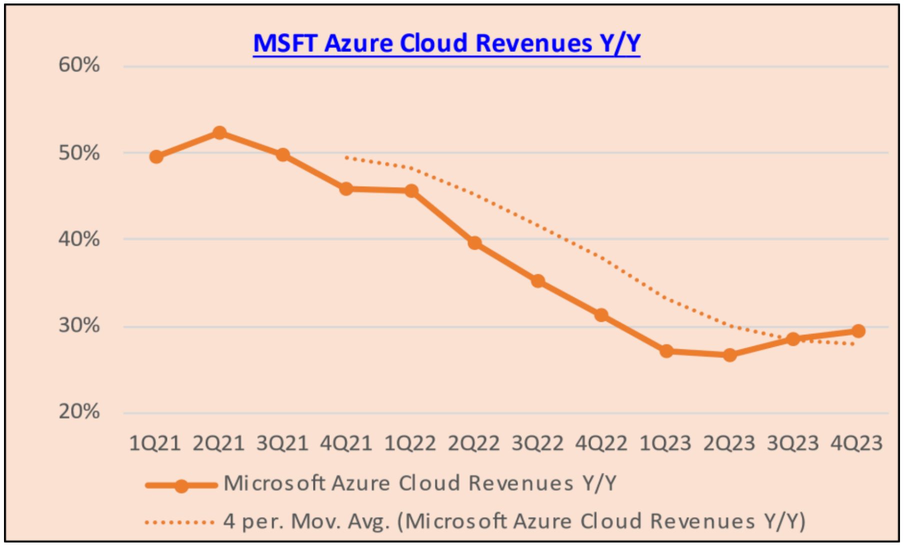
Cloud Provider	Cloud Segment Description
Amazon (AWS)	Amazon reports revenue and capex for the AWS segment. The AWS segment consists of amounts earned from global sales of compute, storage, database, and other services for start-ups, enterprises, government agencies, and academic institutions.
Microsoft (Intelligent Cloud)	Microsoft reports revenue and growth rate for its Intelligent Cloud segment. The Intelligent Cloud segment consists of the company's public, private, and hybrid server products and cloud services. This segment primarily comprises: - Server products and cloud services, including Azure and other cloud services; SQL Server, Windows Server, Visual Studio, System Center, and related Client Access Licenses ("CALs"); and Nuance and GitHub. - Enterprise Services, including Enterprise Support Services, Microsoft Consulting Services, and Nuance professional services.
Microsoft (Azure & other cloud services)	Microsoft separately also reports the revenue growth rate % of "Azure and other cloud services".
Google Cloud	Google's Cloud segment includes Google Cloud offerings, including Google Cloud Platform and Google Workspace. Google Cloud revenues are comprised of the following: - Google Cloud Platform, which includes fees for infrastructure, platform, and other services; - Google Workspace, which includes fees for cloud-based communication and collaboration tools for enterprises, such as Gmail, Docs, Drive, Calendar and Meet; and other enterprise services. Google Cloud is consistently called out as growing faster than Workspace within the Cloud reporting segment.
Oracle Cloud	Oracle Cloud services revenues include revenues earned by providing customers access to Oracle Cloud applications and infrastructure technologies via cloud-based deployment models that Oracle develops, provides unspecified updates and enhancements for, deploys, hosts, manages and supports and that customers access by entering into a subscription agreement with Oracle for a stated period. Oracle Cloud Services arrangements are generally billed in advance of the cloud services being performed; generally they have durations of 1-3 years. Cloud services revenues represented 32%, 25% and 22% of Oracle's total revenues during fiscal 2023, 2022 and 2021.
Alibaba Cloud	Till June 2023: Alibaba's Cloud segment is comprised of Alibaba Cloud and DingTalk. The Cloud businesses primarily generate revenue from the provision of public cloud services and hybrid cloud services to Alibaba's enterprise customers. Post June 2023: Alibaba Cloud is included in the new Cloud Intelligence Group segment that also includes DingTalk and other businesses.

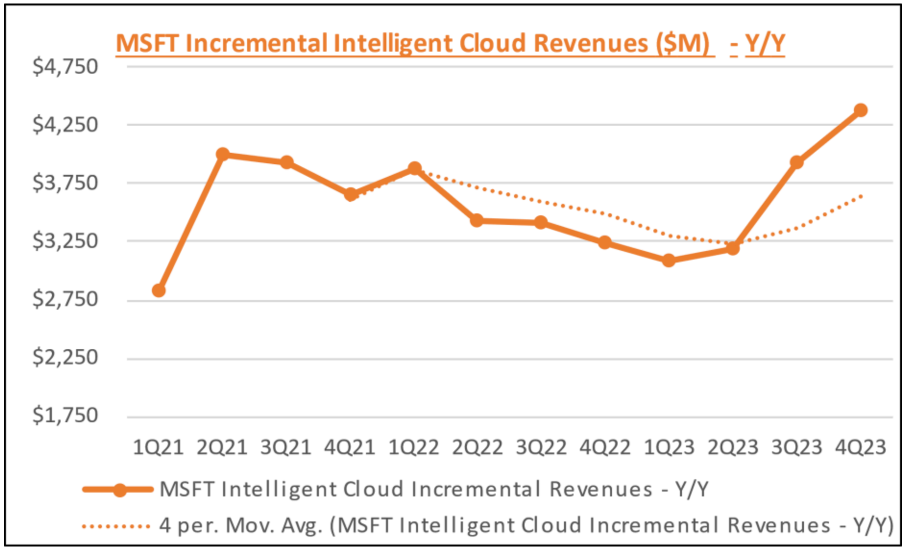
Source: Company reports, RBC Capital Markets

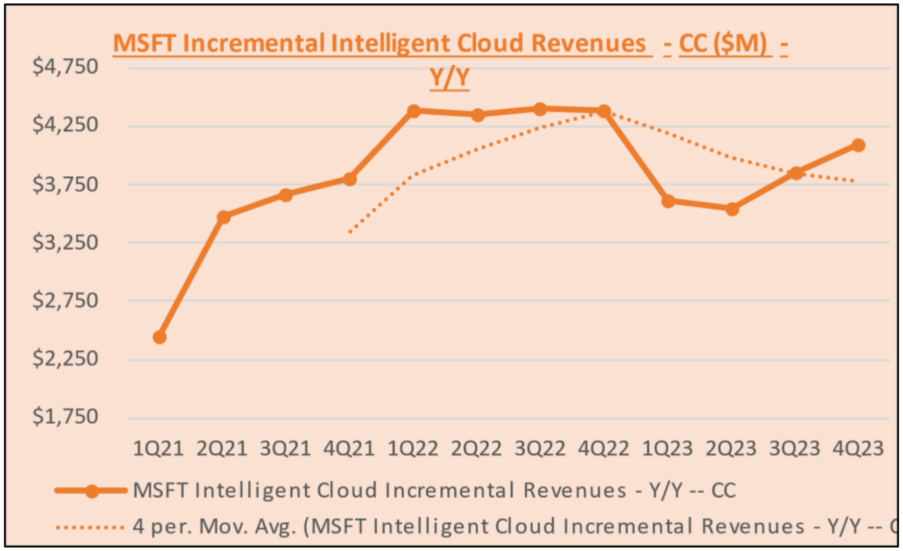
Cloud Revenue Growth – Microsoft Cloud/Azure

Azure & other cloud services revenue grew 30% and 28% Y/Y constant currency in FY2Q24/CY4Q23, with ~6 points increase from AI services. Both AI and non-AI Azure services drove outperformance.





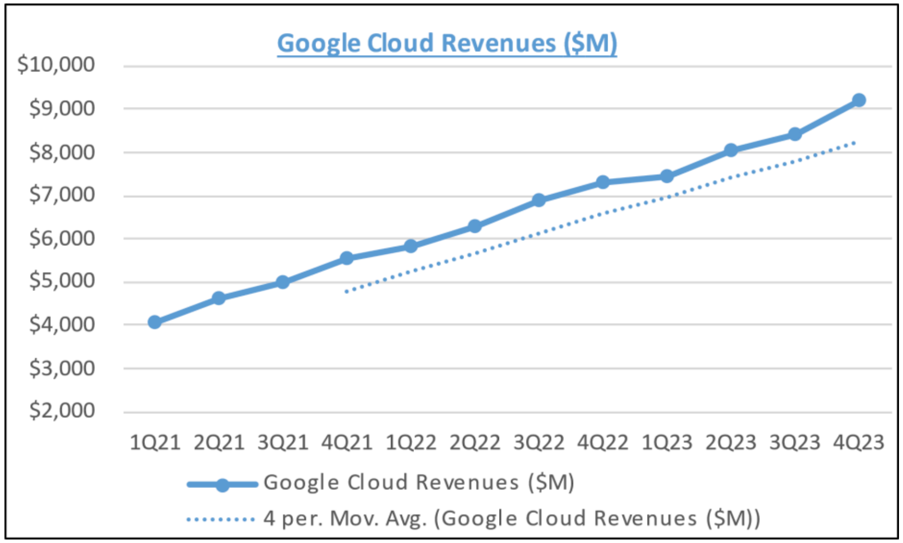


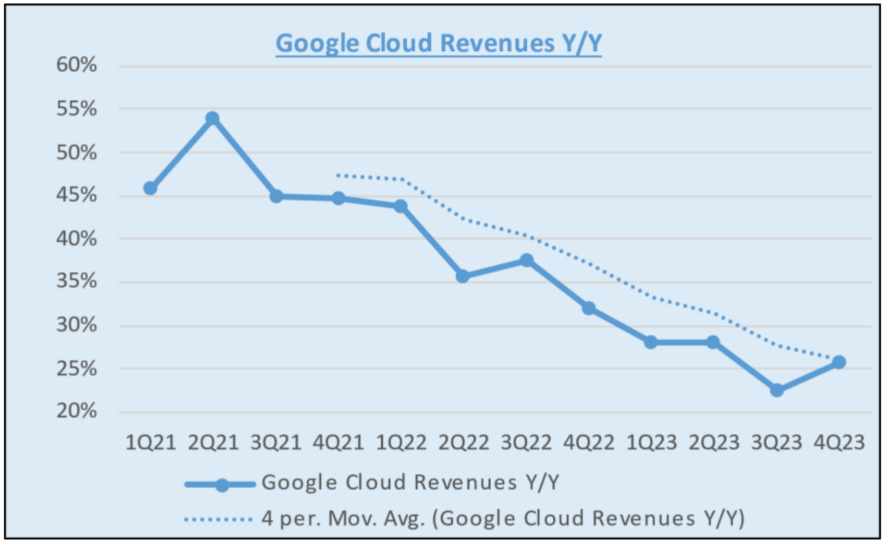


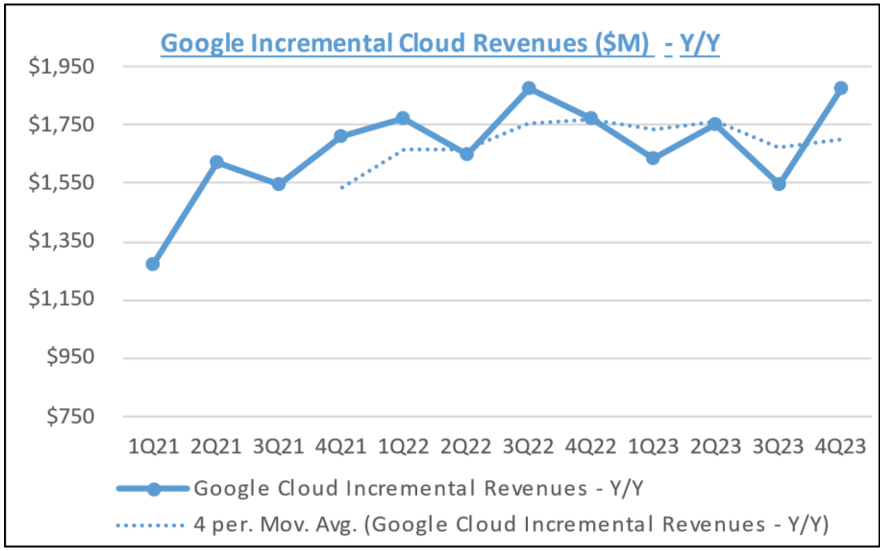
Source: Company reports, RBC Capital Markets

Cloud Revenue Growth – Google Cloud

Google Cloud segment revenues increased 26% Y/Y to $8.2B in 4Q23. Management indicated that GCP revenue growth remained strong across geographies, industries and products. The momentum of GCP is increasing with a significant contribution from AI.



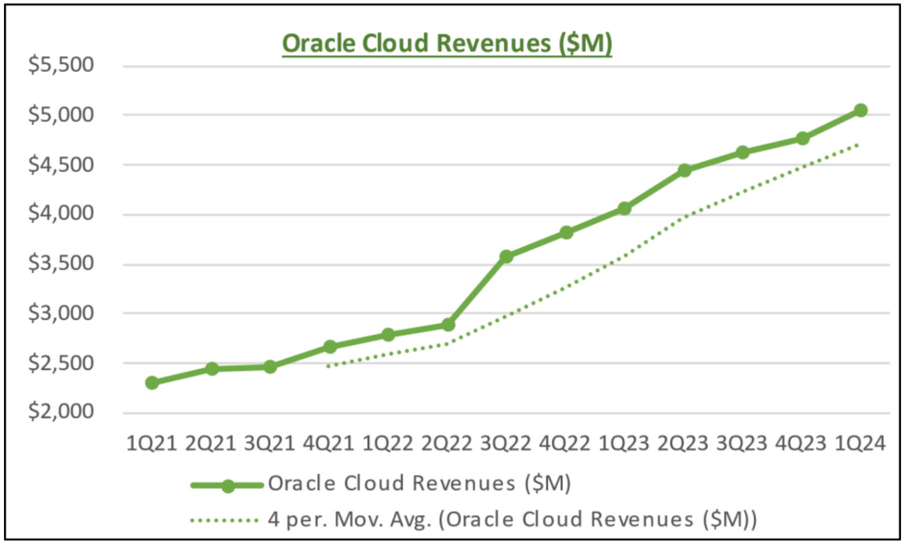


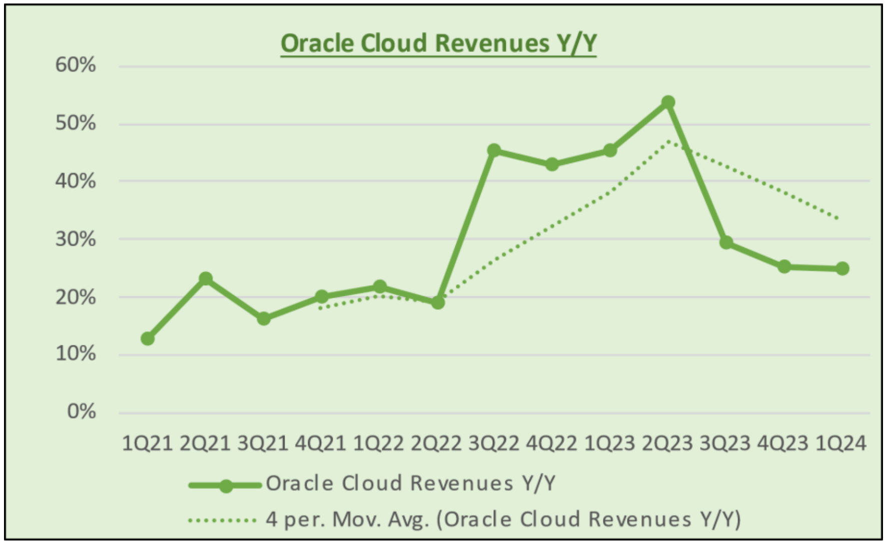


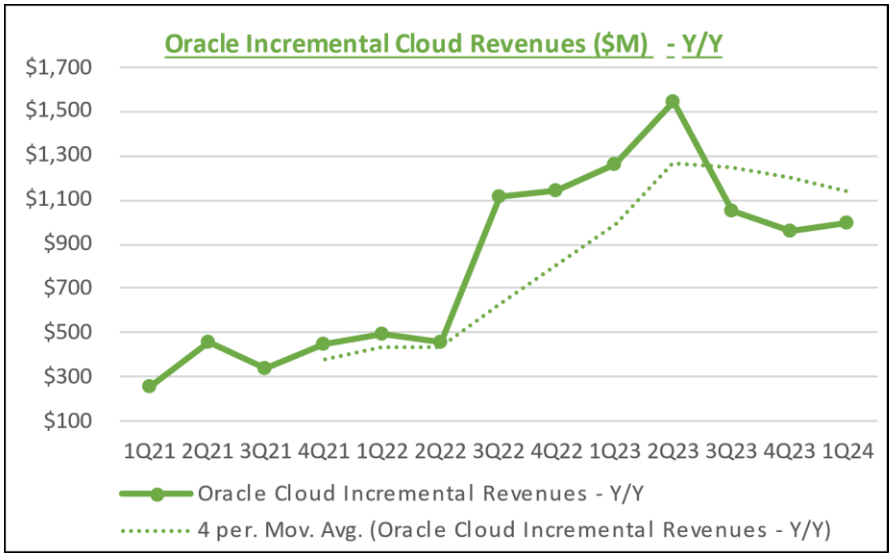
Source: Company reports, RBC Capital Markets

Cloud Revenue Growth – Oracle Cloud

Oracle’s 1Q24 (3QFY24, quarter ending Feb. 29, 2024) Cloud revenue (IaaS plus SaaS) was $5.1 billion, a 25% Y/Y increase in USD and a 24% Y/Y increase in constant currency.



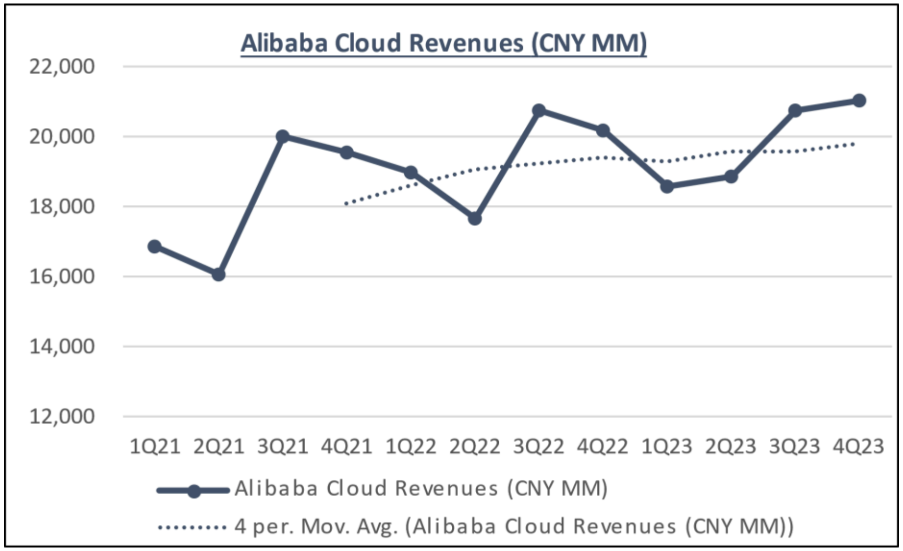


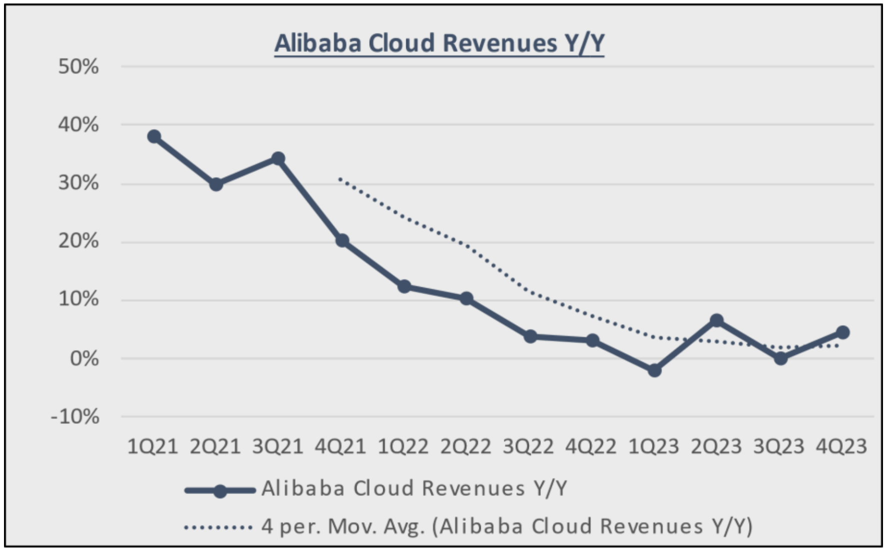


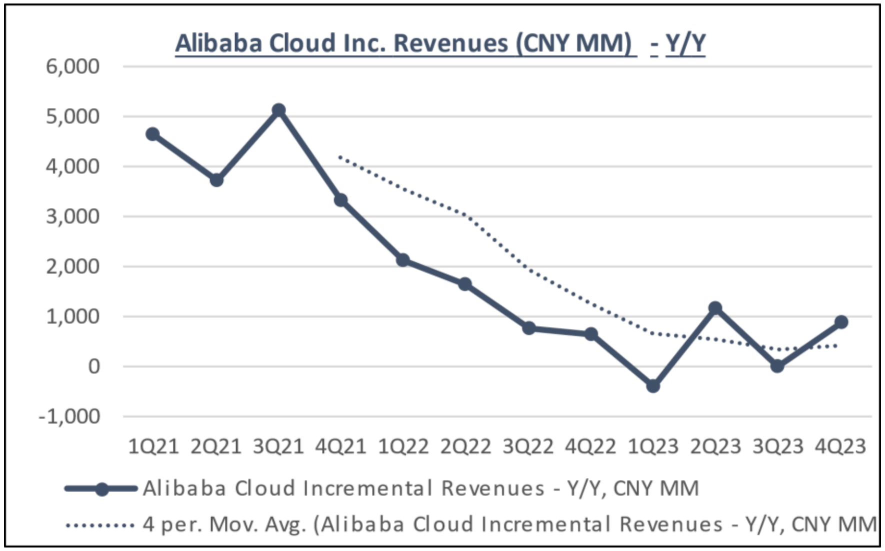
Note: The abrupt increase in 3Q22 was primarily because of the “Cerner” acquisition.
Source: Company reports, RBC Capital Markets

Cloud Revenue Growth – Alibaba Cloud

Alibaba's revenue from Cloud Intelligence Group was RMB28,066 million (US$3,953 million) in 4Q23 (December quarter), growth of 3% Y/Y. Management noted that it continues to improve revenue quality by reducing the revenue from low-margin, project-based contracts.







Source: Company reports, RBC Capital Markets

Capex Trends at Major Companies Driving AI-Related Capex

- Amongst the hyperscalers we track below, total capex is estimated to grow from ~$184 billion in 2023 to $220 billion in 2024 (+20% increase).
- Microsoft, Google, Amazon, Meta and Oracle currently show the largest ramp in 2024E capex.

(in $M)	CY2020	CY2021	CY2022	CY2023A	CY2024E	CY2025E	2024-2023	2024-2023 ($M)
Amazon	$57,976	$72,325	$60,836	$51,579	$58,560	$63,606	14%	$6,982
Microsoft	$21,557	$23,216	$24,768	$35,795	$46,284	$50,349	29%	$10,489
Google	$22,281	$24,640	$31,485	$31,758	$41,505	$44,274	31%	$9,747
Meta/Facebook	$15,163	$18,690	$31,431	$28,251	$34,717	$37,453	23%	$6,466
Apple	$8,702	$10,388	$11,692	$10,531	$9,931	$9,310	-6%	($600)
Alibaba Group	$4,986	$6,525	$7,729	$5,448	$5,874	$6,363	8%	$426
Oracle	$1,833	$3,118	$6,678	$7,963	$10,361	$10,187	30%	$2,398
Tencent	$5,219	$4,613	$3,288	$4,889	$5,408	$5,846	11%	$519
Hewlett Packard Enterprise	$2,328	$2,613	$3,292	$2,841	$2,915	$2,965	3%	$74
IBM	$2,618	$2,062	$1,346	$1,746	$1,807	$2,040	3%	$61
Baidu Inc	$779	$1,715	$1,201	$1,271	$1,290	$1,325	1%	$19
SAP SE	$816	$701	$877	$953	$1,003	$1,068	5%	$50
salesforce.com	$710	$717	$798	$817	$757	$869	-7%	($60)
Total Capex	**$144,967**	**$171,323**	**$185,422**	**$183,842**	**$220,412**	**$235,655**	**20%**	**$36,569**
Y/Y change	27%	18%	8%	-1%	20%	7%		

Source: Company reports, S&P Capital IQ estimates (Total Capex)

Section 5

Datacenter, Infrastructure and Equipment Players

- AI-Related Commentary from Datacenter and other Infrastructure Operators
- AI-Related Commentary from the Equipment and Infrastructure Supply Chain

RBC® Capital Markets

Recent AI-Related Commentary – Datacenters

DigitalBridge

- AI workloads constitute about 40% of the company's datacenter pipeline today, including LLM requirements. Datacenter connectivity demand is already being driven by these AI workloads.
- Management believes AI edge leasing is probably 3 to 4 years out.
- Datacenters dominate capex with $11 billion of the $15 billion budgeted at the portfolio companies in 2024. North America and Europe are the most active markets today, but the rest of the world is accelerating.

Digital Realty Trust

- Digital Realty has seen a growing demand for AI. The company has launched a high-density colocation offering to support private AI deployments.
- The company has been retrofitting existing deployments to fit out for customers needing AI, equipping its capacity blocks that already have a chilled water loop, which is the major most important ingredient for bringing liquid proximate to the chip.
- As AI evolves, the company expects to move from training to inference, and from public- to enterprise-driven consumption.

Equinix

- Equinix has a three-pronged AI strategy: 1) Magnetic AI service provider deployments to support on-ramps, inference nodes, and smaller-scale training needs. 2) Expanding the xScale portfolio, including in North America, to pursue strategic large-scale AI training deployments with hyperscalers and other players. The company has recently leased 90 MW of capacity across six assets in EMEA and APAC. 3) Positioning Platform Equinix as the place where private AI happens, allowing customers to place compute resources in proximity to data and leverage public cloud capabilities while maintaining control of high-value proprietary data.
- Equinix recently announced its expanded partnership for NVIDIA DGX private cloud. This service provides customers a fast and cost-effective way to adopt AI infrastructure. Early wins in this partnership include a Fortune 100 global biopharma company.
- Equinix has the ability in 45 markets or roughly 100 datacenters to support liquid cooling.

NEXTDC

- NEXTDC reported a significant surge in demand for AI-related services, driven by both training and inference requirements. The pipeline has exceeded 1GW, with deal sizes surpassing 100+ MW. Sovereignty concerns are a key driver of localized AI demand. NEXTDC estimates that demand for AI in training and inference is three to five times larger than the base level cloud demand. This means that requirements for AI deployments range from 5 MW to 10 MW up to 100 MW to 300 MW in size.
- NEXTDC announced plans for deploying AI factories in Australia, and followed up with the announcement of S6, designed exclusively for AI factories and sovereign AI, located in Artarmon, Sydney.

Source: Company reports

Recent AI-Related Commentary – Datacenters and Fiber

Other players:

- **CDC Datacenters (owned by Infratil)** commenced construction of additional 117 MW of capacity in response to strong customer demand for generative AI workloads. This was on top of the 265 MW already under construction taking the total under construction to over 380 MW, almost doubling CDC's capacity over a 2-year period. CDC also plans to build over 570 additional MW of capacity by 2029. Earlier in March 2024, CDC confirmed contracting over 200 MW of capacity for the 8 datacenters currently under construction.

- **Fiber:** In the view of a fiber CEO we recently spoke to, AI will be as impactful as cloud was over the last decade. There are initial signs of meaningfully larger requirements for connectivity to large datacenter campuses. The high end of long-haul fiber requirements has increased from 24 fibers to as high as 144 fibers, and within large metro areas the requirements to connect datacenter campuses have increased from 48 fibers to as high as 432 fibers. AI revenue generation appears most promising as an integrated offering into other applications (and a premium charged for the service) vs. standalone AI revenue streams that could be somewhat more complicated or less durable.

Source: Company reports

Recent AI-Related Commentary – Equipment and Infrastructure Supply Chain

Celestica

- **Capital Expenditures and Expansion Plans for 2024:** In 2024, capital expenditures are projected to slightly increase to between 1.75% and 2.25% of revenues, with a higher spend anticipated in the first half of the year. This increase is aimed at expanding capacity at Keysight to meet the surging demand for AI/ML compute and High-Performance Systems (HPS) programs. Specifically, over 100,000 square feet of additional capacity is being developed in Thailand, with phase one expected to be operational in Q1 2024 and phase two by the first half of 2025. This expansion is partially financed by a co-investment with a hyperscaler customer to cater to the demand for specialized datacenter products.
- **Communications and Enterprise (CCS) Segment Demand and Growth:** The demand within the CCS segment remains robust, largely fueled by investments from hyperscalers and the datacenter infrastructure sector. This is driven by the increasing demand for AI and ML applications, with the segment seeing 32% revenue growth in 2023 compared to 2022, amounting to nearly $2.9 billion. This growth accounted for 62% of the total CCS segment revenues, a significant increase from 51% the previous year. The company anticipates this growth trend to continue into 2024, supported by the long-term and structural nature of this investment cycle, potentially offering several years of strong demand for the CCS segment.
- **AI and Datacenter Refresh Cycle:** A shift in capex spending towards AI applications has been observed, with some customers prioritizing datacenter refreshes to accommodate AI applications over traditional cloud applications. As silicon availability improves, there is also a trend towards adopting 400G solutions in anticipation of 800G availability in 2025. The company is still in the early stages of the upgrade cycle for datacenters to support generative AI and LLM processing, indicating a prolonged period of infrastructure upgrades ahead.

Dell

- **AI-Optimized Servers and Growth**: In Q4, Dell reported a nearly 40% sequential increase in AI-optimized server orders, with shipments totaling $800 million and a backlog nearing $2.9 billion. This demand surge was driven by the constrained supply of GPUs and heightened interest in AI GPUs like the H200 and MI300X. Dell projects FY '25 revenues to range between $91 billion and $95 billion, aiming for 5% growth with a midpoint of $93 billion. Dell anticipates mid-teens growth in its Infrastructure Solutions Group (ISG), driven by AI, alongside modest growth in its Client Solutions Group (CSG).
- **Storage Opportunities and AI:** Dell highlighted the integral role of high-performance storage (PowerScale and ObjectScale) in AI model training and inference, emphasizing the need for high bandwidth to prevent GPU idling. This aligns with the trend towards on-prem and edge computing, where AI applications are expected to proliferate due to latency considerations.
- **Comments on NVIDIA B100/B200 GPUs:** Dell management expressed excitement over the B100 and B200 projects. Direct-to-chip cooling should be driven by 1000-watt energy densities per GPU (expected with the B200 set to launch next year).

Source: Company reports

Recent AI-Related Commentary – Equipment and Infrastructure Supply Chain

Eaton

- There is a growing trend towards AI-centric datacenters, which require more powerful and dense electrical infrastructure. This is driving increased demand for Eaton's products and services.
- With the shift towards AI-centric datacenters, the content opportunity for electrical equipment is expected to grow by 5 times compared to conventional datacenters.
- Eaton's Electrical Global segment reported 4% organic growth in Q4 compared to flat growth in Q3, driven by strength in datacenter, industrial, and commercial & institutional (C&I) markets. Operating margin improved by 10 percentage points YoY.
- Order Volume: Orders rose 1% on a rolling 12-month basis, particularly strong in datacenter, IT, utility, MOEM, and industrial markets.
- Datacenter Market Acceleration: Eaton expects the datacenter market to experience robust growth in the coming years, with a projected compound annual growth rate (CAGR) of around 16% over the next five years. This projection is backed by strong order trends, as evidenced by a 30% increase in orders in the past year, with negotiations showing even greater growth.
- Supply Chain Improvements: Management has seen a return to historical supply chain performance levels, overcoming electronic component bottlenecks and addressing intermittent supply issues through internal capacity building and supplier collaborations. While most systemic supply issues are resolved, occasional challenges persist with specific suppliers and components.

nVent

- **Liquid cooling and power distribution units (PDUs)** should drive >40% of the total revenue growth in the Datacenter Connectivity segment.
- **Data Solutions** saw strong double-digit organic growth. This growth is complemented by strategic acquisitions aimed at expanding the product portfolio. A significant portion of the Data Solutions business, exceeding 40%, is concentrated on liquid cooling and PDUs. Additionally, other areas like cable management trays and leak detection also show promising growth.
- **Expanding Liquid Cooling Capacity:** The company announced plans to double its liquid cooling capacity by mid-year, driven largely by hyperscale demand. The company is also working on standardizing these solutions to cater to a broader market, including colocation services and smaller edge computing setups.
- **Supply Chain:** Despite facing productivity challenges against a backdrop of supply chain constraints, the company saw improvements in 2023, driven by material and logistics productivity. However, nVent acknowledges the need for further productivity enhancements to return to pre-2019 levels.
- **New Rear Door Cooling Product:** nVent introduced the RDHX PRO, a rear door cooling (RDC) unit designed to enhance datacenter cooling efficiency. This innovative solution is tailored for high-density racks up to 78kW. RDC solutions are gaining popularity for their ability to retrofit improved cooling capacity on a pay-as-you-go basis.

Source: Company reports

Recent AI-Related Commentary – Equipment and Infrastructure Supply Chain

Schneider Electric

- **Supply Chain Adaptation and Industry Performance:** The company has witnessed a new equilibrium in supply chains, marked by significant reshoring trends influenced by trade dynamics and political factors. This shift has beneficially impacted Schneider Electric, especially in power distribution and digitization, despite a slight weakness in automation. The strategic response to these supply chain challenges underscores the company's resilience and adaptability.
- **Demand Outlook and Trends:** Looking ahead to 2024, Schneider Electric anticipates continued strong market demand driven by structural megatrends. The company expects significant growth contributions from system offers, with a keen focus on datacenters, grid infrastructure, and process industries.
- **Datacenters and Networks:** The company has observed a recovery in the market for datacenters, hyperscalers, and colocation services, buoyed by strong demand and a healthy backlog. Schneider Electric projects a continued growth rate of over 10% in the datacenter and network sectors, emphasizing the enduring importance of these segments to the company's overall growth strategy.

Supermicro

- Growth in FQ2-2024 revenues was attributed to strong demand for AI and rack-scale Total IT solutions, alongside an improving supply chain. Significant design wins, orders, and a backlog from major datacenters and other customers fueled this surge.
- **Datacenter Segment Growth:** The company reported substantial growth in enterprise/channel and OEM appliance/large datacenter revenues, with AI/GPU and rack-scale solutions making up over half of the quarterly revenues. Supermicro is increasing its rack delivery capacity and advanced liquid cooling solutions to meet demand.
- **Accelerated Demand Driven by AI:** Supermicro is experiencing accelerating demand for AI solutions, entering a phase of increased customer wins. The company recently raised more than $2.0 billion to support this growth and is preparing to significantly expand its AI portfolio with upcoming products from NVIDIA, AMD, and Intel.
- The company's position in liquid cooling allows it to provide capacity for future demand, while also supporting air cooling and hybrid cooling options. Its focus on time-to-delivery (TTD) continues to improve, with dedicated capacity for manufacturing 100-120kW racks with liquid cooling expected by the June quarter.

Source: Company reports

Recent AI-Related Commentary – Equipment and Infrastructure Supply Chain

Vertiv

- **Datacenter Segment:** Management highlighted that in 2023 datacenters accounted for roughly 75% of its sales. The datacenter segment was further divided, with half of this 75% stemming from Hyperscale and Colocation (Cloud/Colo) services. Management projects growth rates of 14% and 17%, respectively, in those categories. These areas were the fastest-growing within Vertiv's portfolio, driven by increasing demands for power, thermal management, services, and white space solutions.
- **New Facility Orders and Geographic Expansion:** The majority of Vertiv's new orders in recent quarters were for new installations, with a significant portion of growth occurring in the Cloud/Colo market segment. This trend was observed on a global scale, with the Americas leading in AI acceleration. However, the EMEA region was catching up, albeit with a 6-12 month lag. Vertiv anticipates that North America will continue to lead in AI deployment, with other regions to follow suit.
- **AI-Driven Market Dynamics and Enterprise Demand:** Management anticipates AI demand to expanding into the enterprise market, recognizing this trend as a natural progression of AI's general adoption.
- **Hybrid architecture:** While certain products are directly tied to AI applications—like liquid cooling systems designed for GPUs—the applicability of many other products spans a broader spectrum, including both AI and non-AI uses. The distinction between AI-specific and general applications remains complex, with many datacenters adopting hybrid models that support diverse workloads.
- **Liquid Cooling Technology and Market Opportunities:** Management emphasized the increasing importance and adoption of liquid cooling technology, particularly direct-to-chip cooling, as a response to the next generation of GPUs. This drove Vertiv's strategic acquisition of CoolTera, enhancing its offerings in the direct-to-chip cooling space. The company sees direct-to-chip cooling as the predominant technology for the foreseeable future for applications requiring intense computational power. Vertiv is currently seeing more demand for direct-to-chip cooling than immersion cooling.
- **Supply Chain and Capacity:** Concerning supply chain challenges and the company's ability to meet demand, management indicated that Vertiv’s capacity is designed to have a 25% to 30% leeway to accommodate peaks in demand. They maintain a cautious approach towards capacity expansion, ensuring that they can respond to increased demand without overextending.

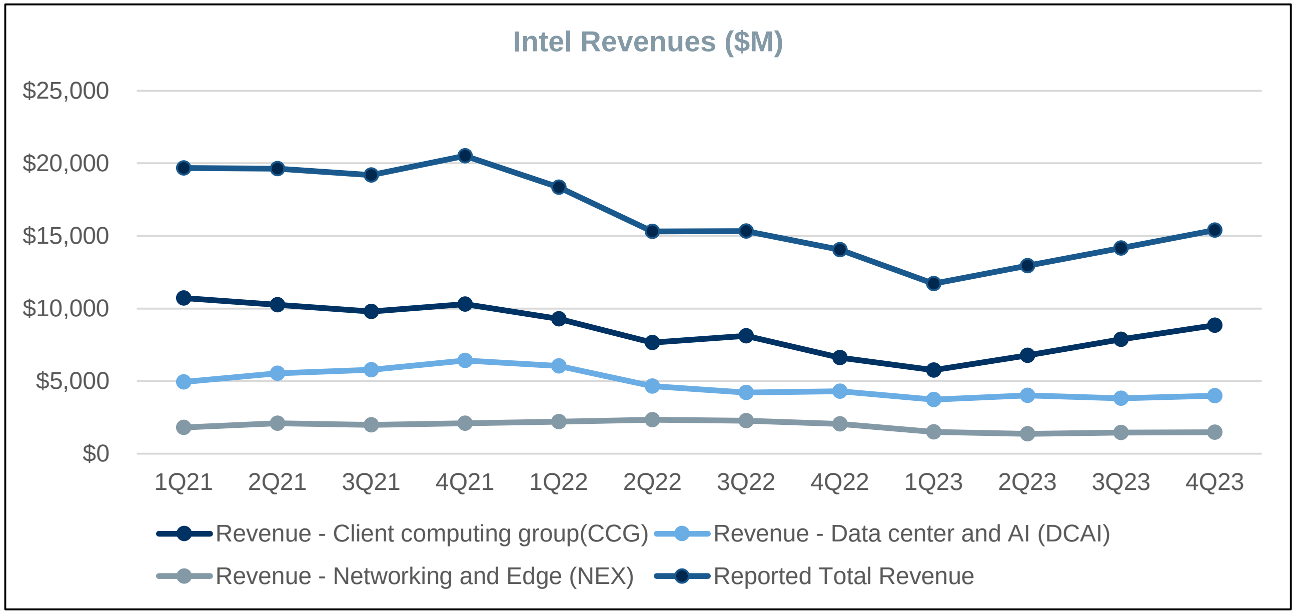
Source: Company reports

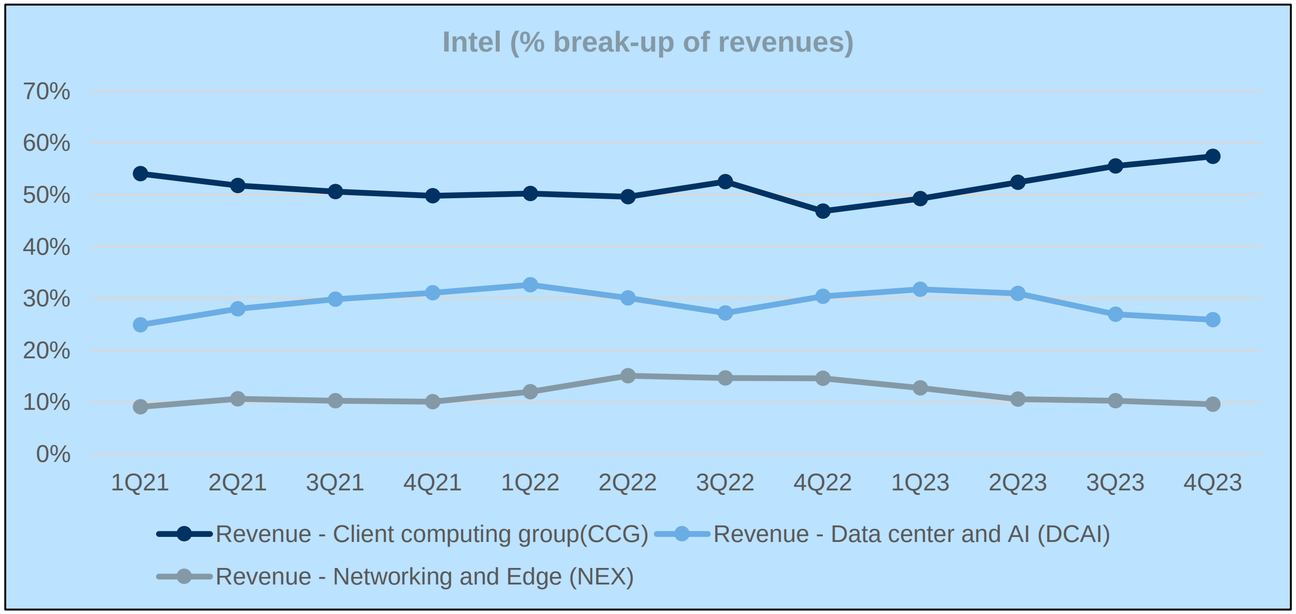
Section 7

Semiconductor Players

- ➢ Financial Highlights
- ➢ Recent AI and Cloud-Relevant Developments

Intel Corporation (INTC) – Revenue

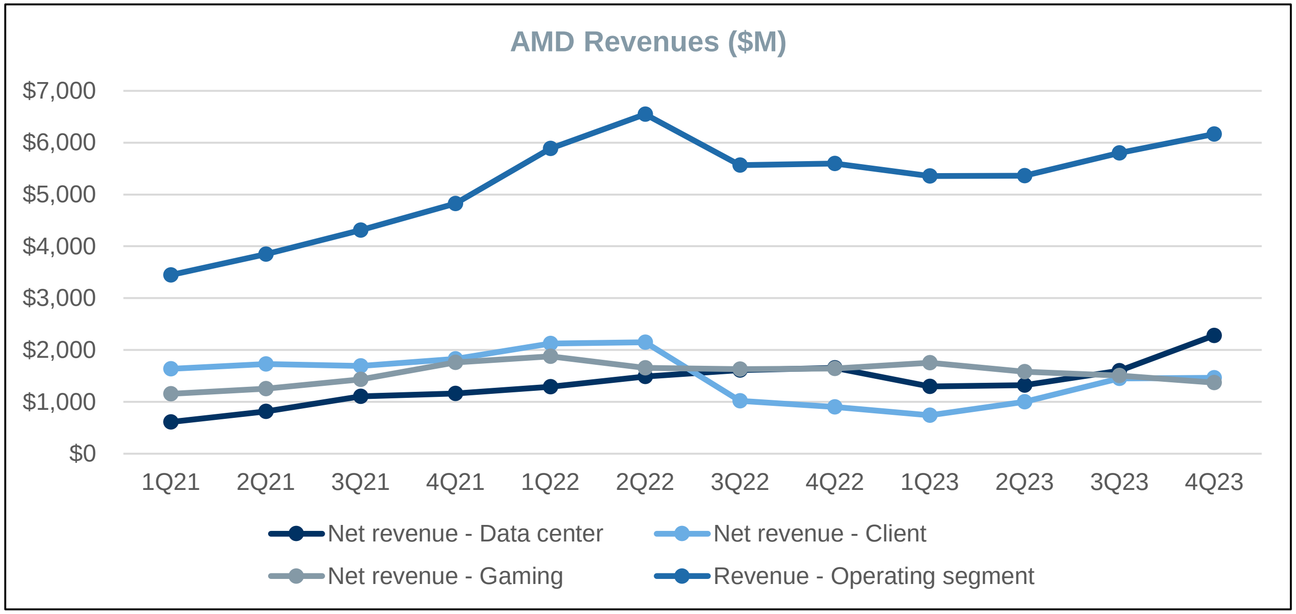


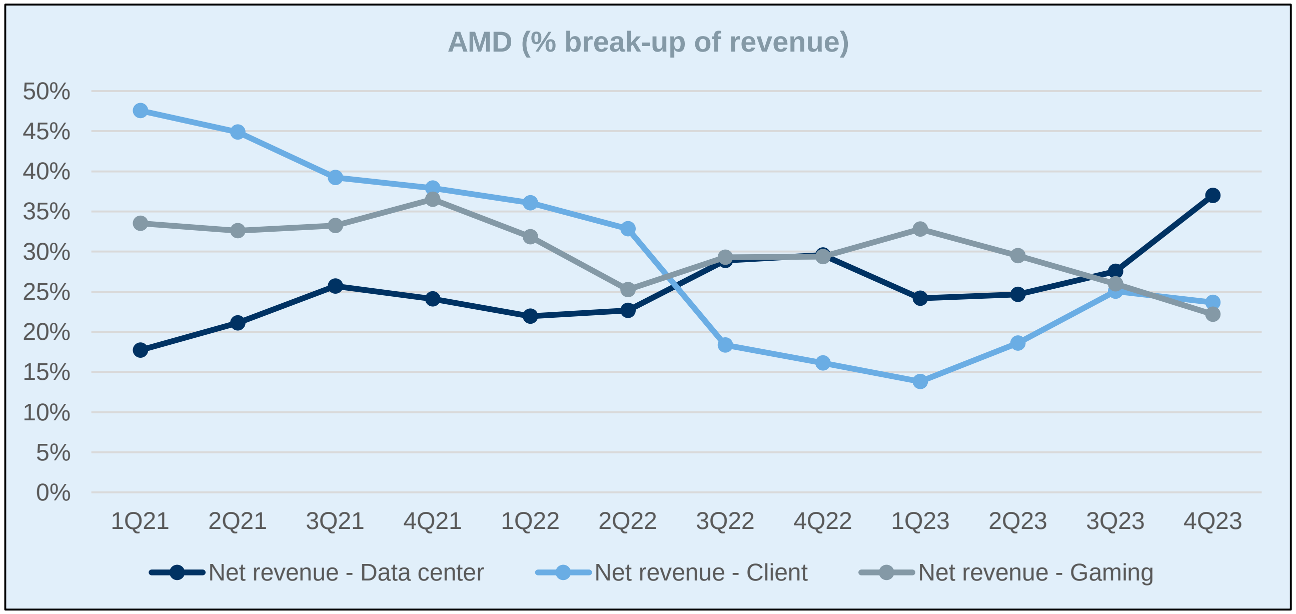


- Total revenue for 4QFY23, increased 10% Y/Y and increased 9% Q/Q.
- CCG revenue increased 12% Q/Q and 33% Y/Y. CCG includes products designed for end-user form factors, focusing on higher growth segments of 2 in 1, thin-and-light, commercial and gaming, and growing other products such as connectivity and graphics..
- DCAI revenue increased 4% Q/Q but decreased 10% Y/Y, driven by a decrease in server revenue.
- NEX revenue increased 1% Q/Q and was down 24% Y/Y, as customers tempered purchases to reduce existing inventories and adjust to a lower demand environment across product lines.
- CCG/DCAI/NEX segments represented ~57%/26%/10% of total revenues for 4QFY23.

Source: Company reports, Visible Alpha

Advanced Micro Devices, Inc. (AMD) – Revenue

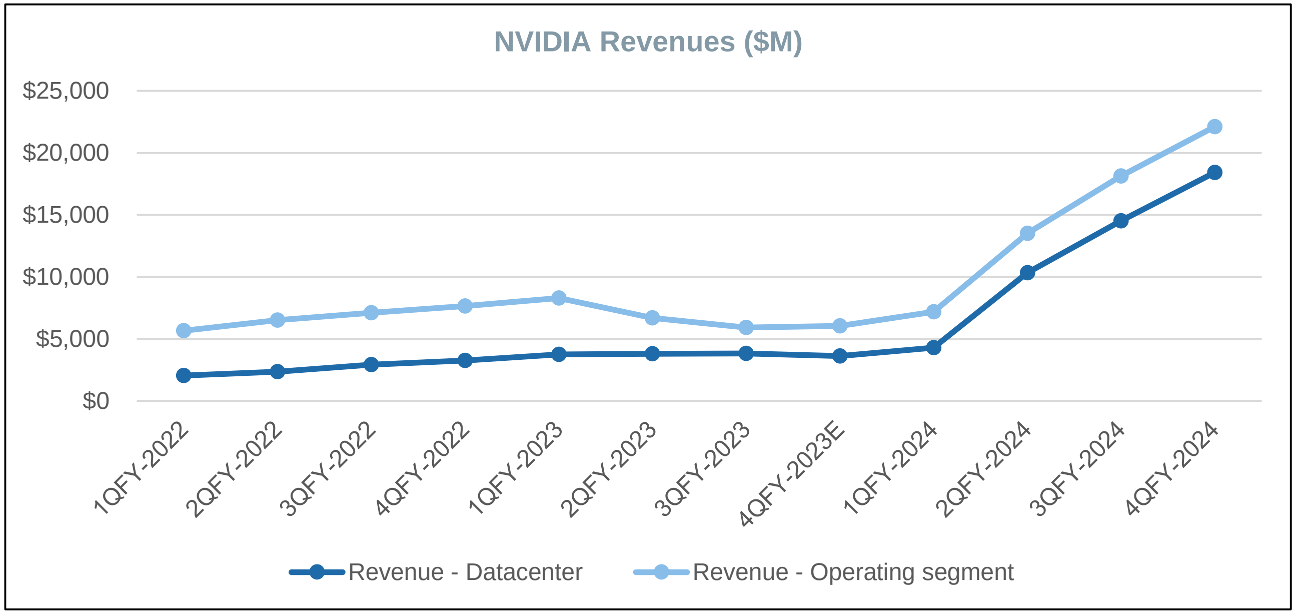


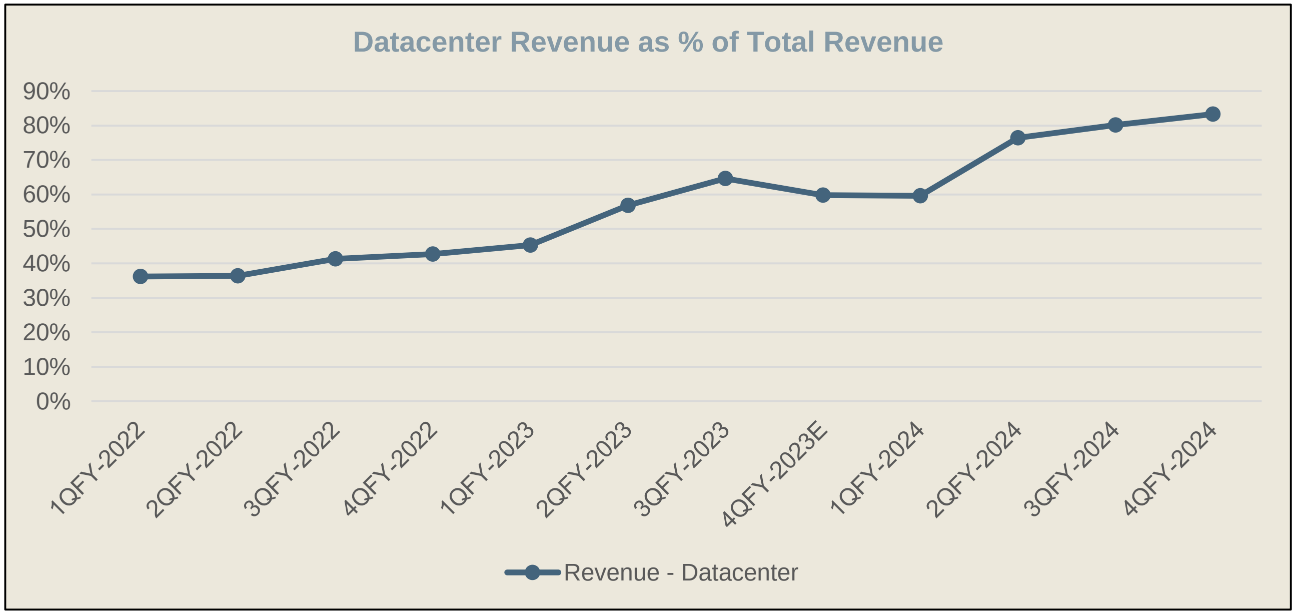


- Total revenue for 4QFY23, increased 10% Y/Y and was up 6% Q/Q.
- Datacenter segment revenue increased 38% Y/Y and 43% Q/Q, driven by strong growth in AMD Instinct™ GPUs and 4th Gen AMD EPYC™ CPUs.
- Client segment revenue increased 62% Y/Y and was up +1% Q/Q driven by an increase in AMD Ryzen™ 7000 Series CPU sales.
- Gaming segment revenue declined 17% Y/Y and 9% Q/Q, due to a decrease in semi-custom revenue, partially offset by an increase in AMD Radeon™ GPU sales.
- Datacenters/Client/Gaming segments represented ~37%/24%/ 22% of total revenues for 4QFY23.

Source: Company reports, Visible Alpha

NVIDIA Corporation (NVDA) – Revenue

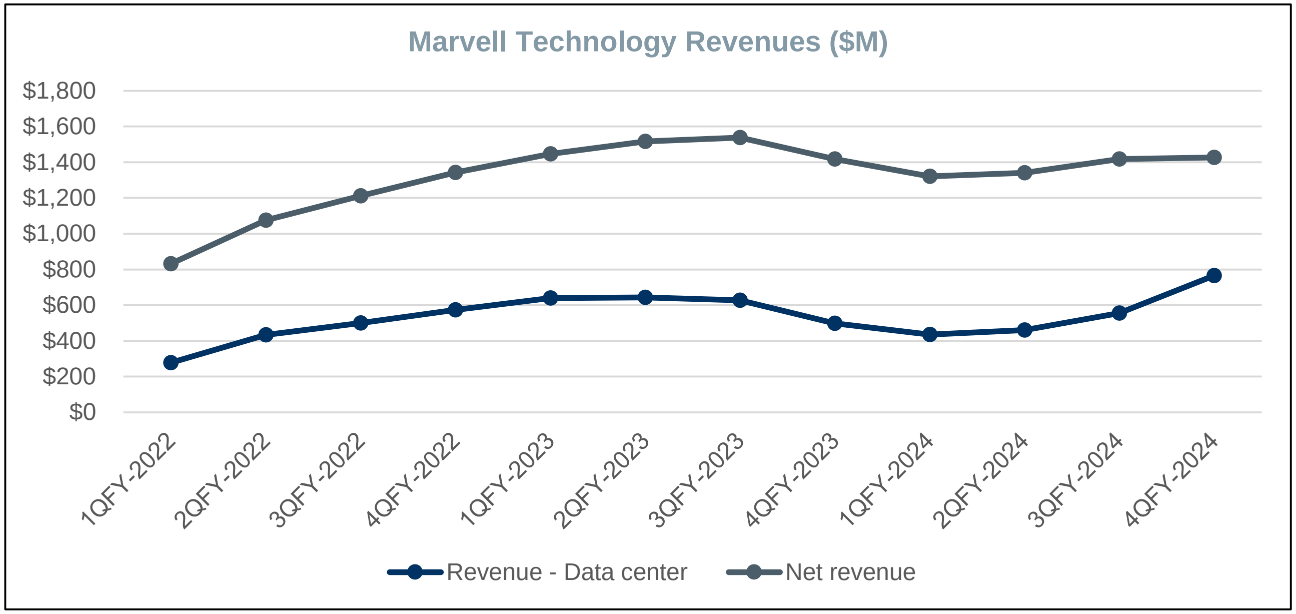


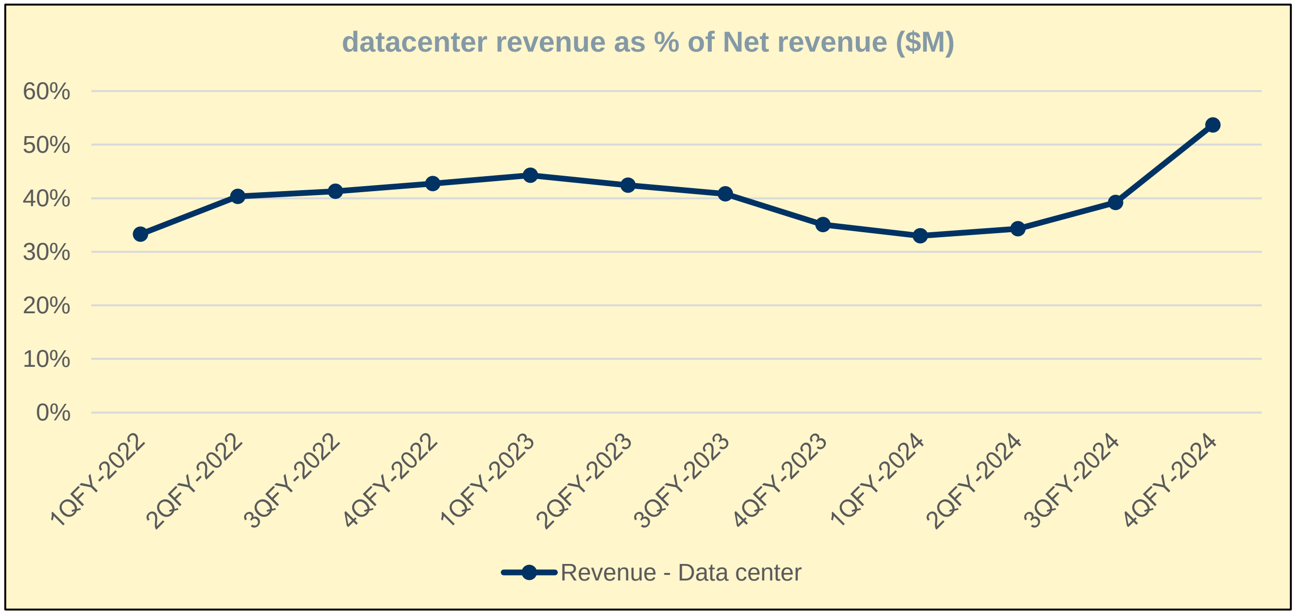


- Revenue for 4QFY24 was $22.1 billion, up 265% Y/Y and up 22% Q/Q.
- Datacenter revenue was up 409% Y/Y and up 27% Q/Q, led by CSPs and large consumer internet companies. These increases reflect higher shipments of the NVIDIA Hopper GPU computing platform used for the training and inference of LLMs, recommendation engines, and generative AI applications, along with InfiniBand end-to-end solutions
- Datacenter segment represented ~83% of the total revenues for 4QFY24 compared to 60% as of $QFY23.
- Networking exceeded a $13 billion annualized revenue run rate. NVIDIA's end-to-end networking solutions define modern AI datacenters.

Source: Company reports, Visible Alpha

Marvell Technologies (MRVL) – Revenue





Source: Company reports, Visible Alpha

- Revenue for 4QFY24 was $1,427 million, up ~1% Y/Y and up ~1% Q/Q.
- Datacenter revenue increased 54% Y/Y and up 38% Q/Q. The strong revenue growth in the quarter was driven by the cloud portion of its datacenter end market.
- Datacenter was its largest end market, driving 54% of total revenue. The next largest was enterprise networking with 19%, followed by carrier infrastructure at 12%, consumer at 10% and auto industrial at 5%.
- Marvell's 800-gig PAM solutions led its growth in the fourth quarter. It also benefited from higher sequential demand for its storage products as that portion of its datacenter end market continues its recovery.
- AI was a key driver of its datacenter growth in fiscal 2024, contributing over 10% of total company revenue, well above its initial forecast. This is a substantial increase from ~3% in the prior year.
- Marvell expects its overall datacenter revenue to grow in the low-single digits sequentially on a percentage basis.

Semiconductors – Recent AI and Cloud-Relevant Announcements

Mar-24	Nvidia's CEO confirms next DGX will be liquid cooled	Nvidia CEO Jensen Huang has confirmed that the next iteration of the company's DGX server family will be liquid cooled.
Mar-24	SK Hynix could invest $1 billion in advanced chip packaging	Chip maker SK Hynix is planning to make further investments in South Korea to expand and improve its advanced chip packaging technology. According to a report in Bloomberg, the investment will be used to further advances such as lowering power consumption and improving performance to meet the growing demands of AI applications.
Mar-24	AMD's China-focused AI chip sale blocked by U.S. Commerce Department	The U.S. Commerce Department has blocked AMD from selling an AI chip specifically tailored to the U.S. market, claiming that it is still too powerful. According to a report by Bloomberg, despite being designed to meet U.S. export restrictions, the department has told AMD it must obtain a license from the Bureau of Industry and Security if it wants to sell the chip in China.
Mar-24	Alibaba cuts cloud services costs by up to 55%	Alibaba Group Holding announced major price cuts across its cloud computing services, as reported by Bloomberg. As of February 29, the Chinese cloud provider is reducing costs for more than 100 products including data storage and elastic computing, with the largest reductions reaching 55%. The average price cut sits at around 20%.
Mar-24	GPU shortages curtailed HPE's first-quarter revenue	During its conference call discussing financial results, Hewlett Packard Enterprise (HPE) CEO Antonio Neri stated that a lack of graphics processing units (GPUs) hindered its revenue growth. The shortage has persisted, causing delays in product deliveries due to increased customer demands for infrastructure setup. However, CFO Marie Myers noted that although supplies are still limited, they are gradually improving as the company implements corrective measures.
Mar-24	Meta reveals details of two new 24k GPU AI clusters	Meta has shared the details of the hardware, network, storage, design, performance, and software that make up its two new 24,000-GPU datacenter scale clusters that the company is using to train its Llama 3 large language AI model. The new training clusters are based on Meta's AI Research SuperCluster (RSC), which was unveiled in 2022. Developed to support AI research and development in areas such as natural language processing, speech recognition, and image generation, the newly announced clusters both contain 24,576 Nvidia Tensor Core H100 GPUs. This is a significant increase over the original clusters, which contained 16,000 Nvidia A100 GPUs. Meta said this increase allows the clusters to support larger and more complex models than the RSC, paving the way for advancements in generative AI product development.

Source: Company reports, RBC Capital Markets

Semiconductors – Recent AI and Cloud-Relevant Announcements

Mar-24	AI chip business Groq acquires Definitive Intelligence to bolster GroqCloud unit	AI chip startup Groq has acquired Definitive Intelligence for an undisclosed sum. The acquisition will support the company's newly formed GroqCloud unit, which will be led by Definitive Intelligence co-founder and CEO Sunny Madra. In a statement, Groq said that Madra and the GroqCloud team will initially focus on expanding capacity, improving efficiency, forming partnerships, and building out the company's developer platform. The company is also formalizing a Groq Systems business unit, which will focus on providing its hardware to public sector organizations.
Feb-24	Kyndryl and Google Cloud team up on AI solutions	Kyndryl and Google Cloud have agreed an extended partnership which will see the digital infrastructure services provider offer its clients access to Google's AI tools, including its Gemini LLMs. Kyndryl also plans to leverage the Google Cloud Cortex Framework to help businesses gain more insights from enterprise resource planning data stored on Google Cloud. It also plans to bolster its internal AI skills base, and will offer staff more access to training and bootcamps through the Google Cloud Cloud Academy for Kyndryl.
Feb-24	Google updates Gemini, launches 1.5 Pro generative AI model	Two months after the initial launch of its Gemini generative AI model, Google has started to roll out an updated version of its multi-modal model for text, image, and audio interactions. Available in three variants, Nano, Pro, and Ultra, Gemini 1.5 Pro is the first model Google has offered up for early testing. Gemini 1.0 Ultra was launched by the company just last week, powering Google's Bard chatbot that will now also be known as Gemini. The version of the chatbot powered by Ultra has been dubbed Gemini Advanced. The 1.5 Pro model has also been built using a Mixture-of-Experts (MoE) architecture, a model architecture that combines multiple parameter subsets, or 'expert models' to generate outputs. Google says this makes the model faster and more efficient to run.
Feb-24	Nvidia CEO Jensen Huang predicts datacenter spend will double to $2 trillion	Nvidia CEO Jensen Huang believes that over the next four to five years, a trillion dollars' worth of datacenter infrastructure and hardware will be built across the world. "There's about a trillion dollars' worth of installed base of datacenters. Over the course of the next four or five years, we'll have $2 trillion worth of datacenters that will be powering software around the world."
Feb-24	Nvidia plans custom chip venture	Nvidia plans to launch a new business unit focused on designing custom chips for clients including the cloud hyperscalers. Reuters reports that the company has already discussed designing bespoke chips for Amazon, Meta, Microsoft, Google, and OpenAI.

Source: Company reports, RBC Capital Markets

Semiconductors – Recent AI and Cloud-Relevant Announcements

Feb-24	Nvidia and Cisco partner on AI solutions for datacenters	Nvidia and Cisco have partnered up to offer organizations cloud-based and on-premises AI infrastructure, networking, and software solutions for the datacenter.
Jan-24	U.S. gov't plans new cloud rules to thwart China AI development	The U.S. government wants to introduce “know your customer” (KYC) style rules for the cloud computing industry which would compel cloud providers to track businesses running workloads in their datacenters. The proposals appear to be the latest salvo in the tech trade war between the U.S. and China.
Jan-24	Google Cloud and Hugging Face partner for open-source AI	Google Cloud is partnering with open-source code and model platform Hugging Face for open-source AI software development using Google's cloud infrastructure. Developers will be able to use Google's technical infrastructure, including its in-house AI chips—or tensor processing units (TPUs)—as well as the technical integration required to use the Nvidia H100 GPUs, which is expected to be completed in weeks, to access Hugging Face's open-source AI software including models and data sets.
Jan-24	Meta to operate "600,000 H100 GPU equivalents of compute" by year-end	Meta expects to field a fleet of 600,000 GPUs by the end of 2024. CEO Mark Zuckerberg told The Verge that the number includes some 340,000 Nvidia H100s, alongside A100s and other AI chips.
Jan-24	Global semiconductor revenue declined 11% in 2023	Global semiconductor revenue saw a year-over-year decline of 11 percent in 2023 to $534 billion, according to a January 16, 2024 press release from Gartner. In total, the combined semiconductor revenue of the top 25 semiconductor vendors declined by 14.1% in 2023, with only nine of those 25 companies posting revenue growth during the same period.
Dec-23	Intel launches new Xeon processors for datacenters	MIntel has launched its fifth-generation Xeon Scalable processors, describing the portfolio of products as “the best CPU for AI” at the company’s AI Everywhere event in New York. According to Intel, Xeon is also the only mainstream datacenter processor with built-in AI acceleration, delivering up to 42% higher inference and fine-tuning on models as large as 20 billion parameters, without the need for additional accelerators. The new chips can contain up to 64 cores and have an L3 cache of 320MB, over double the size of the previous generation’s L3 cache of 112.5MB. By combining this larger cache with a simpler chiplet architecture, Intel says this new generation of Xeon processors provides a 1.21× performance gain over its previous offering.
Dec-23	AMD launches MI300X GPUs and MI300A APUs	AMD has announced the availability of its MI300 accelerators and processors to help power advancements in generative AI. The company says that the AMD Instinct MI300X and the AMD Instinct MI300A have better memory capacity and are more energy efficient than their predecessors. The MI300-series has been designed to train and run LLMs, with AMD claiming they are the highest-performance accelerators in the world for generative AI.

Source: Company reports, RBC Capital Markets

Semiconductors – Recent AI and Cloud-Relevant Announcements

Nov-23	Nvidia plans to release new artificial intelligence chips, namely the HGX H20, L20, and L2 tailored for Chinese market	Nvidia plans to release new artificial intelligence chips, namely the HGX H20, L20, and L2, specifically designed for the Chinese market. This comes less than a month after U.S. officials tightened rules on selling high-end AI chips to China. These chips would include most of Nvidia's newest features for AI work. However, some of their computing power measures have been reduced to comply with new U.S. rules.
Nov-23	Anthropic to use Google chips in expanded partnership	AI startup Anthropic will be one of the first companies to use new chips from Alphabet Inc.'s Google, deepening their partnership after a recent cloud computing agreement. Anthropic will deploy Google's Cloud TPU v5e chips to help power its LLM, named Claude, the companies said on Wednesday. Such software uses a flood of data to train AI interfaces, letting them field questions and generate conversational text.
Nov-23	AWS Announces Amazon EC2 Capacity Blocks for ML Workloads	AWS has introduced EC2 Capacity Blocks for Machine Learning (ML), a new consumption model that provides customers with reserved access to hundreds of NVIDIA GPUs located in Amazon EC2 UltraClusters optimized for high-performance ML workloads. Customers can specify their desired cluster size, start date, and duration when using EC2 Capacity Blocks, ensuring reliable, predictable, and uninterrupted access to GPU compute capacity for their critical ML projects.
Oct-23	US Curbs Nvidia Sales to China	US announced sweeping updates to export curbs designed to block China's access to advanced computer chips, changes that will restrict the sale of semiconductors that Nvidia Corp. designed specifically for the Chinese market. The latest curbs target Nvidia's A800 and H800 chips, a senior US official said. The new rules also require companies to notify the US government before selling chips that fall below the controlled threshold.
Oct-23	Nvidia to make Arm-based PC chips in major new challenge to Intel	Nvidia has quietly begun designing central processing units (CPUs) that would run Microsoft's Windows operating system and use technology from Arm Holdings, Reuters reported citing people familiar with the matter. Nvidia and AMD could sell PC chips as soon as 2025, one of the people familiar with the matter said.
Sep-23	Oracle Cloud Infrastructure (OCI) has made Nvidia H100 Tensor Core GPUs generally available on OCI Compute	The H100 GPUs will be available as OCI Compute bare-metal instances, targeting large-scale AI and high-performance computing applications. The H100s have been found to improve AI inference performance by around 30 times, and AI training by four times compared to that of Nvidia's A100 GPUs, though are notoriously hard to get hold of. The OCI Compute shape includes eight H100 GPUs, each with 80GB of HBM2 GPU memory and 3.2Tbps of bisectional bandwidth. The shape also includes 16 local NVMe drives with 3.84TB each and the 4th Gen Intel Xeon CPU processors with 112 cores and 2TB of system memory.
Aug-23	AmpereOne lands first in Google Cloud C3A instances	AmpereOne has its first cloud instance at Google. The companies announced that the Google C3A instances are being powered by AmpereOne chips a big step for both companies. These are now C3A compute instances. The initial Ampere instances were T2A Tau instances. The Tau line was an entry into Google Cloud. The compute line is Google Cloud elevating the stature of Ampere CPUs.

Source: Company reports, RBC Capital Markets

Semiconductors – Recent AI and Cloud-Relevant Announcements

Aug-23	Tenstorrent, an AI and RISC-V chip company, raised $100 million	Tenstorrent, an AI and RISC-V chip company, has raised $100 million from investors including Hyundai Motor Group, Kia, and a Samsung investment fund. The investment was structured as a debt that will convert to stock at a later date. Tenstorrent builds scalable artificial intelligence accelerators for both the cloud and Edge, hoping to compete with Nvidia's GPUs, and is developing a RISC-V CPU. It also licenses its designs to other companies. Prior to the deal, Tenstorrent had already raised $234.5 million to date with a valuation of $1 billion in its last round.
July-23	AWS announced new EC2 P5 instances based on Nvidia's latest H100 GPUs	Amazon Web Services (AWS) now offers customers access to Nvidia's latest H100 GPUs as Amazon EC2 P5 instances. Nvidia and Amazon claim that the P5 instances are up to six times faster at training large-language models than the A100-based EC2 P4 instances and can cut training costs by 40 percent. Each P5 instance features eight H100 GPUs capable of 16 petaflops of mixed-precision performance, 640 GB of memory, and 3,200 Gbps networking connectivity. Customers will be able to scale their P5 instances to over 20,000 H100 GPUs. Customers already using P5 instances include Cohere, Hugging Face, Pinterest, and Anthropic. Hyperscalers have struggled to procure Nvidia's latest GPU due to supply shortages, but Nvidia has offered priority access to smaller cloud companies that aren't building their own AI chips like CoreWeave and Lambda Labs.
Jun-23	AWS and Oracle announced new instances/VMs based on AMD's 4th-Gen EPYC processors	Amazon Web Services launched EC2 M7a Instances in preview, with general availability expected by Q3. According to AWS, these instances deliver up to 50 percent more performance than M6a instances. These new instances feature AMD's 4th-generation 'Genoa' EPYC processors. The instances support AVX3-512, VNNI, and BFloat16 and feature Double Data Rate 5 (DDR5) memory, which provides 50 percent higher memory bandwidth compared to DDR4 memory to enable high-speed access to data in memory. Oracle also announced Genoa-powered E5 Instances, with general availability starting in July.
Jun-23	AMD announces a new CPU targeted at hyperscale datacenter users	On June 13, 2023, AMD announced hyperscaler-focused Epyc 97X4 processor, featuring 128 Zen 4c cores. The Epyc 97X4 processor line uses the new Zen 4c architecture, a 'cloud-native' version of Zen 4 that optimizes for both power efficiency and performance. Previously codenamed Bergamo, the CPU features up to 128 Zen 4c cores per socket and is fabricated on TSMC's 4nm process node. AMD is now shipping the new CPU to hyperscale customers 'at scale'.
May-23	NVIDIA announces DGX GH200 AI Supercomputer	In May 2023, Nvidia announced a new DGX class, the GH200, for generative AI workloads. The DGX GH200 connects up to 256 Grace Hopper Superchips into a single 144TB GPU system. The superchip is itself a combination of Nvidia's Grace Arm CPU and Hopper GPU, connected by the NVLink C2C chip-to-chip interconnect.
May-23	Ampere launches its custom chips	On May 18, 2023, Ampere Computing announced its AmpereOne chip for cloud providers and enterprises constructing their own private clouds. The chip, featuring 192 cores and a custom ARM-compatible design, is built to balance high-performance with energy efficiency.

Source: Company reports, RBC Capital Markets

Semiconductors – Recent AI and Cloud-Relevant Announcements

May-23	Meta announces AI training and inference chip project	On May 18, 2023, Meta announced plans for its own custom accelerator chip, MTIA, alongside a new “AI-optimized datacenter design” and a “16,000 GPU supercomputer” dedicated to AI research. The Meta Training and Inference Accelerator (MTIA) is an inference accelerator that will enable faster processing of compute-intensive features in the AI services that Meta builds for its users. Meta says that building its own chips will offer granular improvements in performance, power efficiency and cost when they are deployed in 2025. MTIA will be used to support the workloads of internal AI models. The MTIA accelerator is fabricated at TSMC using a 7nm process and runs at 800 MHz, with a thermal design power (TDP) of 25W.
Apr-23	Microsoft is said to be developing an 'Athena' AI chip for large-language models	According to several media sources, Microsoft is developing its own internal artificial intelligence chip, codenamed Athena. The Information has reported that the semiconductor has been in the works since 2019 and is available to a small group of Microsoft and OpenAI employees for testing. The 5nm-process node Athena is reportedly built for training software such as LLMs, which are core to the generative AI surge seen in recent months. But the growth of those models has been held back by GPU shortages at Nvidia.
Apr-23	The EU green lights $47 billion Chips Act	On April 18, 2023, the European Union moved forward with the €43 billion ($47 billion) Chips Act, which hopes to double the EU's global market share in semiconductors from 10% to at least 20% by 2030. The European Council and European Parliament reached a provisional political agreement on the regulation, creating a semiconductor objective within the Digital Europe Program.
Apr-23	Google makes significant progress in chip market	Google has made significant steps in the chip market in the last year. On Feb 23, 2023, the company announced that it was readying two Arm CPUs for its cloud service. In Oct 2022, the company released the E2000 chip in partnership with Intel. It has also developed the Argos video encoding semiconductor for YouTube.
Mar-23	Chinese web giant Baidu backs RISC-V for the datacenter	Chinese RISC-V upstart StarFive has revealed that Chinese web giant Baidu has become an investor, to advance use of the open-source processor design in the datacenter. StarFive said that it would “work with Baidu to promote the implementation of different forms of high-performance RISC-V products in datacenter scenarios”. RISC-V is an open-source architecture, whereas ARM is proprietary. This means that any designer who wants to include an ARM CPU into their design (for example, a SoC) must pay royalties to ARM Holdings. RISC-V, on the other hand, is open-source and does not require any royalties or licensing.
Mar-23	Nvidia launches DGX Cloud to offer GPU Supercomputers-as-a-Service	On March 34, 2023, Nvidia launched DGX Cloud to offer GPU Supercomputers-as-a-Service. Offered through existing cloud providers, the DGX Cloud services provide access to dedicated clusters of Nvidia DGX hardware, which can be rented on a monthly basis. Each instance of DGX Cloud features eight Nvidia H100 or A100 80GB Tensor Core GPUs for a total of 640GB of GPU memory per node. DGX Cloud instances start at $36,999 per instance per month.

Source: Company reports, RBC Capital Markets

Section 8

GPU-Focused Topics

- GPU Descriptions and Specifications
- GPU Pricing
- GPU Availability
- GPUaaS Company Profiles
- Recent Developments

RBC® Capital Markets

GPU Descriptions

Nvidia GH200 “Grace Hopper”

- The NVIDIA GH200 Grace Hopper Superchip, combining a Hopper GPU (H100) and Grace CPU, offers improved memory and bandwidth. It excelled in its first MLPerf industry benchmarks, leading in various fields including computer vision, speech recognition, and generative AI.

NVIDIA H100 GPU

- H100 is a new generation of datacenter GPU that is based on the NVIDIA Hopper architecture. H100 has more CUDA cores, Tensor Cores, and RT Cores than NVIDIA A100, which enables it to handle larger and more complex AI and HPC workloads. H100 supports PCIe Gen5 and NVL PCIe Gen5, which are faster and more efficient interconnect technologies than PCIe Gen4 and SXM4, which are supported by NVIDIA A100. H100 has a larger memory size and uses HBM3 memory type, which is more advanced and has higher bandwidth than HBM2e memory type, which is used by NVIDIA A100. NVIDIA H100 delivers up to nine times faster AI training and 30 times faster inference than NVIDIA A100, depending on the application and the model size.

Nvidia A100 GPU

- Flagship datacenter GPU based on Ampere architecture; designed for AI and HPC workloads; 7nm manufacturing process; 40GB or 80GB memory options; peak performance up to 19.5 TFLOPS FP32; 3rd gen NVLink/NVSwitch interconnects; MIG GPU partitioning and multi-GPU scaling deliver flexibility and scalability.

Nvidia L40 GPU

- Based on Nvidia’s Ada Lovelace architecture, is a newer GPU optimized for AI and graphics performance in datacenters, designed to offer excellent power efficiency for enterprises integrating AI into their operations; delivers 91.6 teraFLOPS of FP32 performance.

AMD GPUs

- AMD introduced the M1300A APU and M1300X GPU for AI and HPC workloads, optimized for LLMs; MI300A features 128GB HBM3 memory and 24 Zen 4 CPU cores; MI300X offers up to 192GB HBM3, 153 billion transistors, and 5.2TB memory bandwidth, claimed to be the fastest GPU for generative AI.

Source: Paul Morrison@LinkedIn

GPU Descriptions

Intel AMX

❿ Built-in matrix multiply accelerators in Intel Xeon Scalable processors designed to improve AI training and inference performance directly on CPUs; shown to enhance AI inference on Alibaba Cloud and throughput for BERT model with Tencent - provides way to accelerate AI workloads natively on Intel CPUs vs. specialized accelerators like Habana’s Gaudi2 and Greco.

Intel FPGA and ASICS

❿ Intel’s Infrastructure Processing Units (IPUs) offload tasks like security and virtualization from CPUs to improve efficiency; 2nd-gen 200G IPUs include FPGA-based Oak Springs Canyon and ASIC Mount Evans co-developed with Google; support common IPDK programming framework; future roadmap includes 400G and 800G IPUs.

Intel Habana Gaudi2

❿ Discrete AI training accelerator from Intel’s Habana Labs, upgraded from Gaudi to 7nm process, 24 tensor cores vs 10, 96GB memory vs 32GB, and 48MB SRAM vs 32MB; shows up to 3.2x performance of Gaudi and 2.8x throughput of Nvidia A100 for AI workloads.

Intel Habana Greco

❿ Discrete AI inference accelerator from Habana Labs, also moved to 7nm process; upgraded to LPDDR5 memory for 5x bandwidth vs Goya and 128MB on-chip memory vs 50MB; lower 75W TDP vs 200W for Goya allows higher density deployments.

Intel Xeon CPUs

❿ 4th gen Xeon Scalable processors designed to unlock new performance levels for breadth of AI workloads; Xeon CPU Max Series with HBM delivers up to 4.8x better AI performance; most built-in accelerators like DL Boost and AVX-512; new Efficient-core architecture optimized for AI efficiency; HBM memory improves performance; designed to deliver improved inference and training performance across wide range of AI applications.

AMD-Xilinx Versal AI Core

❿ Xilinx’s Versal series represents strategic shift from FPGAs to integrated platform chips with programmable logic, AI engines, scalar/adaptable engines, advanced I/O, video decoders, and NoC; provides over 100x compute of current server CPUs for AI Inference and wireless acceleration.

Source: Paul Morrison@LinkedIn

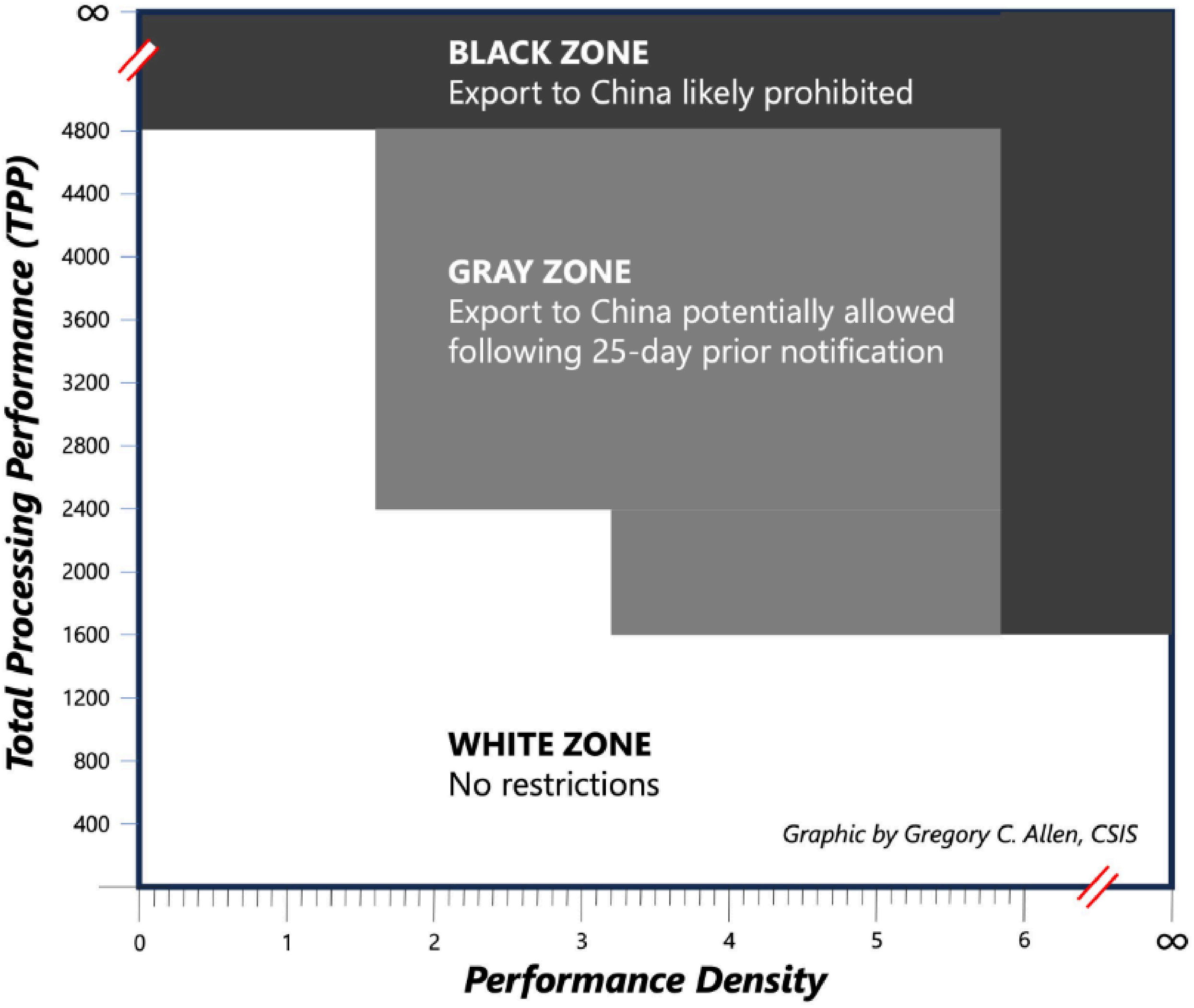
GPU Specifications

	GPU	GPU Arch.	CUDA Cores	Memory	Memory Bandwidth	TFLOPS	Power	Efficiency	Average Pricing
NVIDIA	**H20**	Hopper	NA	96GB	4.0 TB/s	296	400 W	Tailored for Chinese market and to comply with U.S. export requirements, according to ChinaStarMarket.cn.	NA
NVIDIA	**L20**	Ada Lovelace	NA	48GB	864 GB/s	239	275 W	Tailored for Chinese market and to comply with U.S. export requirements, according to ChinaStarMarket.cn. 20% faster than the H100 when it comes to inferencing	NA
NVIDIA	**L2**	Ada Lovelace	NA	24GB	300 GB/s	193	NA	Tailored for Chinese market and to comply with U.S. export requirements, according to ChinaStarMarket.cn. 20% faster than the H100 when it comes to inferencing	NA
NVIDIA	**H100**	**Hopper**	**14,592**	**80GB**	**3.4 TB/s**	**1,979**	**700 W**	**Organizations using NVIDIA H100 GPUs obtain up to a 30x increase in AI inference performance and a 4x boost in AI training compared with tapping the NVIDIA A100 Tensor Core GPU.**	**$40-50K**
NVIDIA	**A100**	**Ampere**	**6,912**	**80GB**	**1.6 TB /s**	**312**	**400 W**	**DRAM utilization efficiency at 95%**	**$10-15K**
NVIDIA	**L40s**	Ada Lovelace	18,176	48GB	864 GB/s	733	300 W	Comparable to H100 but suitable for Inferencing. Suitable for smaller- to medium-sized AI workloads. NVIDIA L40S GPU achieves up to a 20% performance boost for generative AI workloads and as much as a 70% improvement in fine-tuning AI models compared with the NVIDIA A100.	-
NVIDIA	**L40**	Ada Lovelace	18,176	48GB	864 GB/s	362	300 W	Comparable to H100 but suitable for Inferencing. Suitable for smaller- to medium-sized AI workloads. NVIDIA L40S GPU achieves up to a 20% performance boost for generative AI workloads and as much as a 70% improvement in fine-tuning AI models compared with the NVIDIA A100.	$8-10K
NVIDIA	**A40**	Ampere	10,752	48GB	696 GB/s	299	300 W		$5-6K
NVIDIA	**V100**	Volta/Tesla	5,120	16GB	900 GB/s	130	300 W		$1-2K
NVIDIA	**A6000**	Ampere	10,752	48GB	768 GB/s	310	300 W	RTX A6000 is up to 2X more power efficient than Turing GPUs.	~$6-7K
NVIDIA	**A5000**	Ampere	8,192	24GB	768 GB/s	222	230 W	-	~$2K
AMD	**MI 210**	CDNA 2.0	6,656	64GB	1.6 GB/s	181	300 W		
AMD	**MI 250X**	CDNA 2.0	14,080	128GB	3.2 GB/s	383	500 W		
AMD	**MI 300**	CDNA 3.0	14,080	288 GB	9.8 GB/s	2,400	600 W		
Intel	**Gaudi 2**	-	-	96GB	2.5 GB/s	700	650 W		

"CUDA Core" is specific to NVIDIA's GPU architecture. For AMD and Intel GPUs, the equivalent term would be "shading units" or "Tensor Processor Cores (TPCs)" respectively. "TFLOPS" stands for "TeraFLOPS", which is a measure of computing speed and is a unit of measure for the computational power of a GPU. It stands for "trillions of floating-point operations per second".

Source: Company reports

Biden Administration Chip Export Restrictions (Oct. 2023)



BLACK ZONE
Export license required, reviewed with presumption of denial for country group D:5 (includes China); reviewed with presumption of approval for D:1 and D:4 countries

GREY ZONE
Export License officially required, but the Notified Advanced Computing (NAC) license exemption may apply.

If seeking to use NAC for export to Country Group D:5 (includes China), exporter must notify the Dept. of Commerce 25-days prior to concluding written purchase order with buyer. NAC license exemption may or may not be granted at the discretion of the Dept. of Commerce Bureau of Industry and Security

WHITE ZONE
Export license not required unless other regulations (e.g., entity listings) apply

Source: CSIS and Semianalysis.com

Illustrative GPU Pricing

H100

Cloud	GPU Type	GPU Arch	GPUs	GPU RAM	vCPUs	RAM	On-demand	Per-GPU
Lambda	H100 (80 GB)	Hopper	1	80	26	200	$2.49	$2.49
Lambda	H100 (80 GB)	Hopper	8	640	208	1800	$27.92	$3.49
Latitude.sh	H100 (80 GB)	Hopper	4	320	128	768	$11.96	$2.99
Latitude.sh	H100 (80 GB)	Hopper	8	640	128	1536	$22.42	$2.80
Oracle Cloud	H100 (80 GB)	Hopper	8	640	-	-	$80.00	$10.00
CoreWeave	H100 (80 GB)	Hopper	1	80	48	256	$4.25	$4.25
CoreWeave	HGX H100 (80 GB)	Hopper	1	80	48	256	$4.76	$4.76
Vultr	HGX H100 (80 GB)	Hopper	8	640	224	2048	$18.40	$2.30
Vultr	NVIDIA GH200	Hopper	1	96	72	480	$3.99	$3.99

A100

Cloud	GPU Type	GPU Arch	GPUs	GPU RAM	vCPUs	RAM	On-demand	Per-GPU
AWS	A100 (80 GB)	Ampere	8	640	96	1152	$40.97	$5.12
AWS	A100 (40 GB)	Ampere	8	320	96	1152	$32.77	$4.10
Azure	A100 (40 GB)	Ampere	8	160	96	896	$27.20	$3.40
Azure	A100 (80 GB)	Ampere	8	640	96	1900	$37.18	$4.64
Datacrunch	A100 (80 GB)	Ampere	8	640	176	960	$14.80	$1.85
GCP	A100 (40 GB)	Ampere	8	320	96	680	$29.36	$3.67
Jarvislabs	A100 (40 GB)	Ampere	8	320	56	256	$8.80	$1.10
Lambda	A100 (40 GB)	Ampere	8	320	124	1800	$10.32	$1.29
Lambda	A100 (80 GB)	Ampere	8	640	240	1800	$14.32	$1.79
Oblivus Cloud	A100 (80 GB)	Ampere	8	640	32	128	$20.40	$2.55
Oblivus Cloud	A100 (40 GB)	Ampere	8	320	32	128	$19.12	$2.39
Oracle Cloud	A100 (40 GB)	Ampere	8	320	64	2048	$24.40	$3.05
Oracle Cloud	A100 (80 GB)	Ampere	8	640	128	2048	$32.00	$4.00
Paperspace	A100 (40 GB)	Ampere	8	320	96	720	$24.72	$3.09
Paperspace	A100 (80 GB)	Ampere	8	640	96	720	$25.44	$3.18
RunPod	A100 (80 GB)	Ampere	8	640	112	1006	$15.12	$1.89
CoreWeave	A100 40 GB	Ampere	4	160	32	256	$8.24	$2.06
CoreWeave	A100 80 GB	Ampere	4	320	32	256	$8.84	$2.21
Vultr	A100 80 GB PCIe	Ampere	8	640	96	960	$20.83	$2.60
Vultr	A100 80 GB HGX (3 Yr reserved)	Ampere	8	640	112	2048	$11.85	$1.48

Source: Company reports

MLPerf Training Results

MLCommon's MLPerf Training benchmark suite measures how fast systems can train models to a target quality metric. Each benchmark measures the wall clock time required to train a model on the specified dataset to achieve the specified quality target. To account for the substantial variance in ML training times, final results are obtained by measuring the benchmark a specified number of times, discarding the lowest and highest results, and averaging the remaining results.

MLPerf - Version 3.1 - Model GPT3 (LLM)

Availability	Organization	Accelerator	# Accelerators	Processor	# Processors	Result Latency (In minutes)
Available Cloud	NVIDIA	NVIDIA H100-SXM5-80GB	10752	Intel(R) Xeon(R) Platinum 8462Y+	2688	3.92
Available Cloud	Azure+NVIDIA	NVIDIA H100-SXM-80GB	10752	Intel(R) Xeon(R) Platinum 8480C	2688	4.01
Available Cloud	NVIDIA	NVIDIA H100-SXM5-80GB	10240	Intel(R) Xeon(R) Platinum 8462Y+	2560	4.07
Available Cloud	NVIDIA	NVIDIA H100-SXM5-80GB	8192	Intel(R) Xeon(R) Platinum 8462Y+	2048	4.87
Available Cloud	NVIDIA	NVIDIA H100-SXM5-80GB	6144	Intel(R) Xeon(R) Platinum 8462Y+	1536	6.03
Available Cloud	NVIDIA	NVIDIA H100-SXM5-80GB	4096	Intel(R) Xeon(R) Platinum 8462Y+	1024	8.57
Available on-premise	NVIDIA	NVIDIA H100-SXM5-80GB	768	Intel(R) Xeon(R) Platinum 8480C	192	40.63
Available Cloud	Google	TPU-v5e	4096	AMD EPYC 7B13	1024	44.68
Available on-premise	NVIDIA	NVIDIA H100-SXM5-80GB	512	Intel(R) Xeon(R) Platinum 8480C	128	58.30
Available on-premise	Intel-HabanaLabs	Intel Gaudi2	384	Intel(R) Xeon(R) Platinum 8380	96	153.58
Available on-premise	Intel-HabanaLabs	Intel Gaudi2	256	Intel(R) Xeon(R) Platinum 8380	64	223.91

Source: MLCommons

MLPerf Training Results

MLPerf - Version 3.1 - Model Bert-Large (NLP)

Availability	Organization	Accelerator	# Accelerators	Processor	# Processors	Result Latency (In minutes)
Available on-premise	NVIDIA	NVIDIA H100-SXM5-80GB	3472	Intel(R) Xeon(R) Platinum 8480C	868	0.12
Available on-premise	NVIDIA	NVIDIA H100-SXM5-80GB	64	Intel(R) Xeon(R) Platinum 8480C	16	0.90
Available on-premise	Dell	NVIDIA H100-SXM5-80GB	16	Intel(R) Xeon(R) Platinum 8470	4	3.44
Available on-premise	Supermicro	NVIDIA H100-SXM5-80GB	8	Intel(R) Xeon(R) Platinum 8468	2	5.38
Available on-premise	Supermicro	NVIDIA H100-SXM5-80GB	8	AMD EPYC 9654	2	5.40
Available Cloud	Azure	NVIDIA H100-SXM-80GB	8	Intel(R) Xeon(R) Platinum 8480C	2	5.43
Available on-premise	NVIDIA	NVIDIA H100-SXM5-80GB	8	Intel(R) Xeon(R) Platinum 8480C	2	5.47
Available on-premise	Dell	NVIDIA H100-SXM5-80GB	8	Intel(R) Xeon(R) Platinum 8480+	2	5.51
Available on-premise	Dell	NVIDIA H100-SXM5-80GB	8	Intel(R) Xeon(R) Platinum 8480+	4	5.79
Available on-premise	Dell	NVIDIA H100-SXM5-80GB	4	Intel(R) Xeon(R) Platinum 8468	2	11.01
Available on-premise	Dell	NVIDIA H100-SXM5-80GB	4	Intel(R) Xeon(R) Platinum 8480+	2	11.19
Available on-premise	Supermicro	NVIDIA H100-SXM5-80GB	4	Intel(R) Xeon(R) Platinum 8470Q	2	11.49
Available on-premise	Intel-HabanaLabs	Intel Gaudi2	8	Intel(R) Xeon(R) Platinum 8380	2	13.27
Available on-premise	Intel-HabanaLabs	Intel Gaudi2	8	Intel(R) Xeon(R) Platinum 8380	2	14.12

Source: MLCommons

MLPerf Training Results

MLPerf - Version 3.1 - Model ResNet (Image Classification)

Availability	Organization	Accelerator	# Accelerators	Processor	# Processors	Result Latency (In minutes)
Available Cloud	NVIDIA+CoreWeave	NVIDIA H100-SXM5-80GB	3584	Intel(R) Xeon(R) Platinum 8462Y+	896	0.18
Available on-premise	NVIDIA	NVIDIA H100-SXM5-80GB	64	Intel(R) Xeon(R) Platinum 8480C	16	2.50
Available on-premise	Dell	NVIDIA H100-SXM5-80GB	16	Intel(R) Xeon(R) Platinum 8470	4	7.55
Available on-premise	Dell	NVIDIA H100-SXM5-80GB	8	Intel(R) Xeon(R) Platinum 8480+	2	13.41
Available on-premise	Supermicro	NVIDIA H100-SXM5-80GB	8	AMD EPYC 9654	2	13.55
Available on-premise	Supermicro	NVIDIA H100-SXM5-80GB	8	Intel(R) Xeon(R) Platinum 8468	2	13.55
Available on-premise	NVIDIA	NVIDIA H100-SXM5-80GB	8	Intel(R) Xeon(R) Platinum 8480C	2	13.58
Available Cloud	Azure	NVIDIA H100-SXM-80GB	8	Intel(R) Xeon(R) Platinum 8480C	2	13.82
Available on-premise	Dell	NVIDIA H100-SXM5-80GB	8	Intel(R) Xeon(R) Platinum 8480+	4	13.93
Available on-premise	Intel-HabanaLabs	Intel Gaudi2	8	Intel(R) Xeon(R) Platinum 8380	2	15.92
Available on-premise	Intel-HabanaLabs	Intel Gaudi2	8	Intel(R) Xeon(R) Platinum 8380	2	16.42
Available on-premise	Supermicro	NVIDIA H100-PCIe-80GB	8	Intel(R) Xeon(R) Platinum 8490H 60-Core Processor	8	21.87
Available on-premise	Dell	NVIDIA H100-SXM5-80GB	4	Intel(R) Xeon(R) Platinum 8468	2	26.57
Available on-premise	Supermicro	NVIDIA H100-SXM5-80GB	4	Intel(R) Xeon(R) Platinum 8470Q	2	26.69
Available on-premise	Dell	NVIDIA H100-SXM5-80GB	4	Intel(R) Xeon(R) Platinum 8480+	2	26.92

Source: MLCommons

GPU Availability – Microsoft Azure

Instance	GPUs type	Region available - Spot pricing
NC-series	NVIDIA Tesla accelerated platform	East US, East US 2, North Central US, South Central US, West US 2, UK South, North Europe, West Europe, US Gov Arizona, US Gov Virginia, Australia East, Southeast Asia
NCsv2-series	NVIDIA Tesla P100 GPUs	East US, South Central US, West US 2, West Europe, Southeast Asia
NCsv3-series	NVIDIA Tesla V100 GPUs	Central US, East US, East US 2, South Central US, West US, West US 2, West US 3, UK South, Switzerland North, Qatar Central, Korea Central, Japan East, Israel Central, Central India, France Central, North Europe, West Europe, Canada Central, Brazil South, US Gov Arizona, US Gov Virginia, Australia East, East Asia, Southeast Asia
NCas_T4_v3 Series	NVIDIA T4 GPUs	Central US, East US, East US 2, North Central US, South Central US, West US, West US 2, West US 3, UK South, Korea Central, Japan East, Israel Central, Central India, South India, Germany west Central, North Europe, West Europe, Canada Central, Brazil South, US Gov Virginia, Australia Central, Australia Central 2, Australia East, Southeast Asia
NC A100 v4 series	NVIDIA Ampere A100 80GB PCIe GPUs	Central US, East US, East US 2, South Central US, West US, West US 2, West US 3, UK South, Japan East, Italy North, Central India, France Central, North Europe, West Europe, Australia East, Southeast Asia
NCads A10 v4 series	Nvidia A10 GPU	South Central US, West US 3, West Europe
NGads V620 series	AMD RadeonTM PRO V620 GPUs	East US 2, West US 3, Sweden Central, West Europe
NV-series	NVIDIA Tesla accelerated platform	East US, East US 2, North Central US, South Central US, West US 2, UK South, Japan East, Central India, North Europe, West Europe, US Gov Arizona, US Gov Texas, US Gov Virginia, Australia East, Southeast Asia
NVv3-series	NVIDIA Tesla M60 GPUs	East US, East US 2, South Central US, West US, West US 2, UK South, UAE North, Switzerland North, Norway East, Norway west, Japan East, Central India, France Central, North Europe, West Europe, Brazil South, US Gov Arizona, US Gov Virginia, Australia East, Southeast Asia, South Africa North
NVv4-series	AMD Radeon Instinct MI25 GPU	East US, East US 2, North Central US, South Central US, West US 2, West US 3, UK South, Korea Central, Japan East, Italy North, Central India, North Europe, West Europe, Canada Central, US Gov Arizona, US Gov Virginia, Australia East, Southeast Asia
NVads A10 v5 series	Nvidia A10 GPU	Central US, East US, North Central US, South Central US, West US, West US 2, West US 3, UK South, UAE North, Sweden Central, Qatar Central, Korea Central, Korea South, Japan East, Italy North, Israel Central, Central India, Germany West Central, France Central, North Europe, West Europe, Canada Central, Brazil South, US Gov Virginia, Australia East, East Asia, Southeast Asia, South Africa North
NDs-series	NVIDIA Tesla P40 GPUs	East US, South Central US, West US 2, West Europe, Southeast Asia
NDv2 series	NVIDIA V100 Tensor Core GPUs	East US, South Central US, West US 2, Sweden Central, West Europe, US Gov Arizona, US Gov Virginia, Southeast Asia
ND A100 v4 series	NVIDIA Ampere A100 Tensor Core GPUs	East US, East US 2, South Central US, West US 2, West US 3, Italy North, West Europe, US Gov Virginia
NDm A100 v4 series	NVIDIA Ampere A100 GPUs	Central US, East US, North Central US, South Central US, West US, West US 2, West US 3, UK South, UK West, UAE Central, UAE North, Switzerland North, Sweden Central, Poland Central, Norway East, Japan East, Japan West, South India, Germany North, France Central, West Europe, Canada Central, Canada East, Brazil South, Australia East, South Africa North, South Africa West

Source: Company reports, RBC Capital Markets

GPU Availability – AWS

Instance Family	GPUs type	Region available - Spot pricing	Region available - Reserved instance pricing
P5	NVIDIA H100 Tensor Core	US East (Ohio), US East (N. Virginia), US West (Oregon)	Not available
P4d	NVIDIA A100 Tensor Core	US East (Ohio), US East (N. Virginia), US West (Oregon), AWS GovCloud (US-West), Asia Pacific (Seoul), Asia Pacific (Singapore), Asia Pacific (Tokyo), Europe (Frankfurt), Europe (Ireland)	US East (N. Virginia), US East (Ohio), US West (Oregon), AWS GovCloud (US-West), Asia Pacific (Seoul), Asia Pacific (Singapore), Asia Pacific (Tokyo), Europe (Frankfurt), Europe (Ireland)
P4de	NVIDIA A100 Tensor Core	US East (N. Virginia), US West (Oregon), Israel (Tel Aviv)	US East (N. Virginia), US West (Oregon), Israel (Tel Aviv)
P3	NVIDIA Tesla V100	US East (Ohio), US East (N. Virginia), US West (Oregon), Canada (Central), AWS GovCloud (US-West), Asia Pacific (Seoul), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Tokyo), Europe (Frankfurt), Europe (Ireland), Europe (London)	US East (N. Virginia), US East (Ohio), US West (Oregon),Canada (Central), AWS GovCloud (US-West), Asia Pacific (Seoul), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Tokyo), Europe (Frankfurt), Europe (Ireland), Europe (London)
P3dn	NVIDIA Tesla V100	US East (N. Virginia), US West (Oregon), AWS GovCloud (US-East), AWS GovCloud (US-West), Asia Pacific (Tokyo), Europe (Ireland)	US East (N. Virginia), US West (Oregon), AWS GovCloud (US-East), AWS GovCloud (US-West), Asia Pacific (Tokyo), Europe (Ireland)
p2	NVIDIA K80	US East (Ohio), US East (N. Virginia), US West (Oregon), AWS GovCloud (US-West), Asia Pacific (Mumbai), Asia Pacific (Seoul), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Tokyo), Europe (Frankfurt), Europe (Ireland)	US East (N. Virginia), US East (Ohio), US West (Oregon), AWS GovCloud (US-West), Asia Pacific (Mumbai), Asia Pacific (Seoul), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Tokyo), Europe (Frankfurt), Europe (Ireland)
dl1	Gaudi accelerators	US East (N. Virginia), US West (Oregon)	US East (N. Virginia), US West (Oregon)
trn1	AWS Trainium accelerators	US East (Ohio), US East (N. Virginia), US West (Oregon)	US East (Ohio), US East (N. Virginia), US West (Oregon)
inf1	AWS Inferentia accelerators	US East (Ohio), US East (N. Virginia), US West (N. California), US West (Oregon), Canada (Central), AWS GovCloud (US-East), AWS GovCloud (US-West), Asia Pacific (Hong Kong), Asia Pacific (Mumbai), Asia Pacific (Seoul), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Tokyo), Europe (Frankfurt), Europe (Ireland), Europe (London), Europe (Milan)Europe (Paris), Europe (Stockholm), Middle East (Bahrain), South America (Sao Paulo)	US East (N. Virginia), US East (Ohio), US West (N. California), US West (Oregon), Canada (Central), AWS GovCloud (US-East), AWS GovCloud (US-West), Asia Pacific (Hong Kong), Asia Pacific (Mumbai), Asia Pacific (Seoul), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Tokyo), Europe (Frankfurt), Europe (Ireland), Europe (London), Europe (Milan), Europe (Paris), Europe (Stockholm), Middle East (Bahrain), South America (Sao Paulo)
inf2	AWS Inferentia2 accelerators	US East (Ohio), US East (N. Virginia), US West (Oregon)	US East (N. Virginia), US East (Ohio), US West (Oregon)
g5g	NVIDIA T4G Tensor Core	US East (N. Virginia), US West (Oregon), Asia Pacific (Seoul), Asia Pacific (Singapore), Asia Pacific (Tokyo), Europe (Frankfurt), Europe (Spain)	US East (N. Virginia), US West (Oregon), Asia Pacific (Seoul), Asia Pacific (Singapore), Asia Pacific (Tokyo), Europe (Frankfurt), Europe (Spain)
g5	NVIDIA A10G Tensor Core	US East (Ohio), US East (N. Virginia), US West (Oregon), Canada (Central), Asia Pacific (Jakarta), Asia Pacific (Mumbai), Asia Pacific (Seoul), Asia Pacific (Sydney), Asia Pacific (Tokyo), Europe (Frankfurt), Europe (Ireland), Europe (London), Europe (Stockholm), Israel (Tel Aviv), Middle East (UAE), South America (Sao Paulo)	US East (N. Virginia), US East (Ohio), US West (Oregon), Canada (Central), Asia Pacific (Jakarta), Asia Pacific (Mumbai), Asia Pacific (Seoul), Asia Pacific (Sydney), Asia Pacific (Tokyo), Europe (Frankfurt), Europe (Ireland), Europe (London), Europe (Stockholm), Israel (Tel Aviv), Middle East (UAE), South America (Sao Paulo)
g4dn	NVIDIA T4 Tensor Core	US East (Ohio), US East (N. Virginia), US West (N. California), US West (Oregon), Canada (Central), AWS GovCloud (US-East), AWS GovCloud (US-West), Africa (Cape Town), Asia Pacific (Hong Kong), Asia Pacific (Mumbai), Asia Pacific (Osaka), Asia Pacific (Seoul), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Tokyo), Europe (Frankfurt), Europe (Ireland), Europe (London), Europe (Milan), Europe (Paris), Europe (Stockholm), Middle East (Bahrain), South America (Sao Paulo)	US East (N. Virginia), US East (Ohio), US West (N. California), US West (Oregon), Canada (Central), AWS GovCloud (US-East), AWS GovCloud (US-West), Africa (Cape Town), Asia Pacific (Hong Kong), Asia Pacific (Mumbai), Asia Pacific (Osaka), Asia Pacific (Seoul), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Tokyo), Europe (Frankfurt), Europe (Ireland), Europe (London), Europe (Milan), Europe (Paris), Europe (Stockholm), Middle East (Bahrain), South America (Sao Paulo)
g4ad	AMD Radeon Pro V520	US East (Ohio), US East (N. Virginia), US West (Oregon) Canada (Central), Asia Pacific (Tokyo), Europe (Frankfurt), Europe (Ireland), Europe (London)	US East (N. Virginia), US East (Ohio), US West (Oregon), Canada (Central), Asia Pacific (Tokyo), Europe (Frankfurt), Europe (Ireland), Europe (London)
g3s	NVIDIA Tesla M60	US East (Ohio), US East (N. Virginia), US West (Oregon), AWS GovCloud (US-West), Asia Pacific (Seoul), Asia Pacific (Sydney), Asia Pacific (Tokyo), Europe (Frankfurt), Europe (Ireland), Europe (London)	US East (N. Virginia), US East (Ohio), US West (Oregon), Asia Pacific (Seoul), Asia Pacific (Sydney), Asia Pacific (Tokyo), Europe (Frankfurt), Europe (Ireland), Europe (London)

Source: Company reports, RBC Capital Markets

GPU Availability – Google Cloud and Oracle Cloud

Instance	GPUs type	Region available (Google) - Spot pricing
g2-standard-4	NVIDIA L4	asia-east1 (Taiwan), asia-northeast1 (Tokyo), asia-northeast3 (Seoul), asia-south1 (Mumbai), asia-southeast1 (Singapore), europe-west1 (Belgium), europe-west2 (London), europe-west3 (Frankfurt), europe-west4 (Netherlands), us-central1 (Iowa), us-east1 (South Carolina), us-east4 (Virginia), us-west1 (Oregon), us-west4 (Las Vegas)
a2-highgpu-1g	NVIDIA A100 40 GB	asia-northeast1 (Tokyo), asia-northeast3 (Seoul), asia-southeast1 (Singapore), europe-west4 (Netherlands), me-west1 (Israel), us-central1 (Iowa), us-east1 (South Carolina), us-west1 (Oregon), us-west3 (Utah), us-west4 (Las Vegas)
a2-highgpu-1g	NVIDIA A100 80 GB	asia-southeast1 (Singapore), europe-west4 (Netherlands), us-central1 (Iowa), us-east4 (Virginia), us-east5 (Ohio)
N1 machine series	NVIDIA T4	asia-east1 (Taiwan), asia-east2 (Hong Kong), asia-northeast1 (Tokyo), asia-northeast3 (Seoul), asia-south1 (Mumbai), asia-southeast1 (Singapore), asia-southeast2 (Jakarta), australia-southeast1 (Sydney), europe-central2 (Warsaw), europe-west1 (Belgium), europe-west2 (London), europe-west3 (Frankfurt), europe-west4 (Netherlands), me-west1 (Israel), northamerica-northeast1 (Montréal), southamerica-east1 (São Paulo), us-central1 (Iowa), us-east1 (South Carolina), us-east4 (Virginia), us-west1 (Oregon), us-west2 (California), us-west3 (Salt Lake City), us-west4 (Las Vegas)
N1 machine series	NVIDIA V100	asia-east1 (Taiwan), europe-west4 (Netherlands), us-central1 (Iowa), us-east1 (South Carolina), us-west1 (Oregon)
N1 machine series	NVIDIA P100	asia-east1 (Taiwan), australia-southeast1 (Sydney), europe-west1 (Belgium), europe-west4 (Netherlands), us-central1 (Iowa), us-east1 (South Carolina), us-west1 (Oregon)
N1 machine series	NVIDIA P4	asia-southeast1 (Singapore), australia-southeast1 (Sydney), europe-west4 (Netherlands), northamerica-northeast1 (Montréal), us-central1 (Iowa), us-east4 (Ashburn, Virginia), us-west2 (Los Angeles)
N1 machine series	NVIDIA K80	asia-east1 (Taiwan), europe-west1 (Belgium), us-central1 (Iowa), us-east1 (South Carolina), us-west1 (Oregon)

Instance	GPUs type	Availability (Oracle)
Large scale-out AI training, data analytics, and HPC		
BM.GPU.H100	NVIDIA H100	Expected in December 2023 in London and Chicago Regions
BM.GPU.A100	NVIDIA A100	Generally available in many regions
Smaller to medium-sized AI workloads		
BM.GPU.L40S	NVIDIA L40	Expected in 2024
Small AI training, inference, streaming, gaming, and virtual desktop infrastructure		
VM.GPU.A10	NVIDIA A10	Generally available, including US East (Ashburn), US West (Phoenix), US West (San Jose), Canada Southeast (Toronto), UK South (London), Germany Central (Frankfurt), France Central (Paris), Saudi Arabia West (Jeddah), Japan East (Tokyo), Japan Central (Osaka), Singapore (Singapore)
VM.GPU3 and BM.GPU3	NVIDIA V100	Generally available in most regions
VM.GPU2 and BM.GPU2	NVIDIA P100	Generally available in most regions

Source: Company reports, RBC Capital Markets

GPUaaS – Company Profiles

Following is a compilation of several of the GPU as a Service players pursuing AI-driven growth strategies. Financial sponsorship, capitalization, and scale of these companies vary widely. Determinants of success include, in our view: access to GPUs, cost of capital (some companies have cost of debts or all-in costs of capital well into the mid teens or higher), and customer and revenue profile (purely on-demand vs. long-term contractual revenues.

Company	Description
Applied Digital (Nasdaq: APLD)	▪ Applied Digital, headquartered in Dallas, designs, develops, and operates next-generation datacenters in North America, offering digital infrastructure solutions to the high-performance computing (HPC) industry. ▪ The company has facilities in Jamestown and Ellendale, North Dakota, and a third site under construction in Garden City, Texas. ▪ Applied Digital offers AI Cloud services through Sai Computing, initially provided from its 9 MW HPC Jamestown facility. ▪ The company has three business segments: next-generation datacenter colocation services, AI GPU cloud services, and Blockchain datacenters.
Clever Cloud	▪ Clever Cloud is a European PaaS. ▪ Headquartered in Nantes, France, it has Datacenters in Europe and North America. ▪ It has launched GPU-based instances under the Clever Grid brand, designed for machine learning and using Nvidia GeForce GTX 1070 for robust hardware acceleration.
CoreWeave	▪ CoreWeave, founded in 2017, is a specialized cloud provider that offers GPU-accelerated compute resources on demand. CoreWeave operates datacenters in three U.S. regions: US East (Weehawken, NJ), US Central (Chicago, IL), and US West (Las Vegas, NV). It has signed leases totaling over 200 MW of capacity in the U.S. and is expanding to Europe and potentially other regions. The company has previously said that it expects to operate 14 datacenters. ▪ NVIDIA is one of its investors, along with a wide range of other equity sponsors and creditors, including Magnetar Capital, Fidelity Management and Research Co, Investment Management Corporation of Ontario, Jane Street, J.P. Morgan Asset Management, Nat Friedman, Daniel Gross, Goanna Capital, Zoom Ventures, Blackstone, Coatue, DigitalBridge, BlackRock, PIMCO, Carlyle and Great Elm.
Crusoe Energy	▪ Crusoe Energy is a company that aims to align computing with climate change by providing oil and gas companies with a cost-effective solution to eliminate natural gas flaring. The company offers computing services for AI models, mining cryptocurrencies, and other compute-intensive activities. ▪ CrusoeCloud is a cloud computing platform optimized for energy-intensive HPC workloads, offering clean and low-cost GPU cloud computing solutions. ▪ The company uses a distributed modular network across Montana, North Dakota, and Colorado to build a 200 MW cloud network.

Source: Company reports, RBC Capital Markets

GPUaaS – Company Profiles

Company	Description
Denvr Dataworks	▪ Denvr Dataworks, founded in 2020, offers bespoke cloud services for artificial intelligence, machine learning, deep learning, and GPU accelerated data science applications. ▪ It offers up to 4 NVIDIA A100s in a reserved node for up to 14 days for free. Its technology supports hybrid cloud scenarios and integration with DevOps pipelines and APIs in private-cloud, hybrid-cloud, and edge computing settings. ▪ Denvr provides modular datacenters with liquid immersion cooling and integrated waste heat recovery, offering high efficiency design and low costs. ▪ The company operates full racks of HGX nodes in liquid immersion. Based in Calgary, Alberta, Denvr has 38 employees.
DigitalOcean/ Paperspace	▪ Paperspace is a cloud computing platform that focuses on GPU-accelerated virtual machines and machine learning models. Established in 2014, it has served 650,000 users and operates datacenters in New York City, Santa Clara, and Amsterdam. ▪ The company offers high-performance GPU tooling for small and medium-sized businesses to test, build, and scale AI models in the cloud. ▪ Paperspace offers a variety of GPU and CPU types for remote machines and provides services like pre-configured Notebook environments for AI/ML model exploration and fine-tuning. ▪ The company generates revenue through a flexible pricing model that charges customers for the utilization of their instances, depending on the type of GPU instance used. The management team and key personnel are co-founded by Dillon Erb and Daniel Kobran.
Fluidstack.io	▪ FluidStack is a cloud-based technology company that enables internet-connected devices to become cloud servers. Founded in 2017, it is based in London, UK, and has access to over 50,000 GPUs from a global datacenter network. ▪ FluidStack aggregates underutilized GPU capacity from datacenters worldwide, offering cloud computation at a reduced cost. ▪ FluidStack provides cloud-based business intelligence software for enterprise data, offering services to train, render, and scale user bases on the world's most cost-efficient GPU cloud. ▪ Its revenue model is similar to Airbnb, connecting businesses, researchers, and hobbyists to the world's largest network of individual datacenters. FluidStack offers competitive rates for NVIDIA A100 and H100 GPUs, claiming that users can reduce their cloud bill by over 70%.
Gcore Labs	▪ Gcore is a Luxembourg-based public cloud and content delivery network (CDN) company founded in 2014. It operates in 11 regions worldwide and has datacenters in various cities. The company's management team includes Colin Sampson, Vsevolod Vayner, Anatoliy Platonov, Alina Galiautdinova, Ahmed Swelam, and others. ▪ Gcore's GPU capabilities are powered by NVIDIA A100 and H100 GPUs, which are designed to accelerate AI tasks with exceptional GenAI capabilities. Gcore Cloud offers solutions for training models and executing inference on NVIDIA GPUs. ▪ The company's revenue model is based on its AI GPU Cloud Infrastructure services, charging customers for using bare metal servers and virtual machines powered by NVIDIA A100 and H100 GPUs.

Source: Company reports, RBC Capital Markets

GPUaaS – Company Profiles

Company	Description
Jarvislabs.ai	▪ Jarvislabs.ai is a GPU cloud platform designed for AI researchers and practitioners, enabling them to build intelligent models while the platform manages infrastructure. ▪ The platform offers on-demand instances and bare metal servers for 30 or more days, with pricing varying based on GPU type, number of GPUs, and storage. ▪ The company is headquartered in Coimbatore, Tamil Nadu, India and offers various GPU-powered instances for AI model training and deployments.
Lambda Labs	▪ Lambda Labs is an AI infrastructure company based in San Francisco that manufactures and sells hardware for AI, machine learning, and deep learning applications. ▪ The company, which initially offered GPU desktop assembly and server hardware solutions, has since expanded to offer Lambda Cloud as a GPU platform. ▪ The virtual machines are pre-equipped with deep learning frameworks, CUDA drivers, and a dedicated Jupyter notebook. ▪ Lambda Labs claims to be used by 10,000+ research teams and has raised more than $430 million in funding. ▪ The company's investors include Crescent Cove, Mercato Partners, 1517 Fund, Bloomberg Beta, Gradient Ventures, Cloudera cofounder Jeff Hammerbacher, OpenAI cofounder Greg Brockman, Y Combinator president Garry Tan, Thomas Tull's US Innovative Technology fund, B Capital, and SK Telecom.
NextGen Cloud	▪ NextGen Cloud is a European cloud Infrastructure-as-a-Service (IaaS) company that focuses on building high-performance computing and GPU infrastructure using state-of-the-art NVIDIA hardware. ▪ The company offers GPU-as-a-Service (GPUaaS) through the Hyperstack platform. It raised $14 million in funding and partnered with DARMA Capital and Moore and Moore Investments Group to finance its AI Supercloud project. The company has built a GPU fleet in Europe, including NVIDIA H100s, and plans to build an AI Supercloud in Europe with over 20,000 NVIDIA H100 Tensor Core GPUs by June 2024. ▪ It has committed $576 million in hardware orders as part of its $1 billion investment.
Omniva	▪ Omniva plans to build AI-focused GPU-filled cloud datacenters in the Middle East and Europe. ▪ The company is led by former AWS, Google and Meta executives. ▪ The Kuwaiti royal family has provided backing for the venture.
Outscale (Dassault Systèmes)	▪ Outscale, a subsidiary of Dassault Systèmes, offers enterprise-class cloud services, including TINA OS, which automate cloud resources, allowing organizations to easily deploy, manage, and increase their cloud platforms. ▪ In June 2019, Dassault Systèmes acquired a majority stake in OutScale, a leading European provider of cloud-based high-performance computing (HPC) and AI services. ▪ The acquisition aimed to expand Dassault's presence in the cloud computing market and enhance its digital transformation solutions.

Source: Company reports, RBC Capital Markets

GPUaaS – Company Profiles

Company	Description
Runpod.ai	▪ RunPod is a company that provides serverless GPU computing for AI inference and training, offering users the option to pay by the second for their compute usage. ▪ The platform is designed to scale dynamically, meeting the computational needs of AI workloads from smallest to the largest scales. ▪ RunPod offers a range of GPUs, including H100, A100, and L40, for AI inference and training.
Scaleway/Iliad	▪ Scaleway is a European cloud solution provider offering a range of products, including bare metal and serverless, for data processing, artificial intelligence, rendering, and video encoding. ▪ French telecommunications operator Iliad holds a 30% stake in Scaleway. The investment has provided Scaleway with significant financial backing and access to Illiad's network infrastructure, helping it expand its cloud services and customer base. Scaleway is headquartered in Paris and operates datacenters in Paris, Amsterdam, and Warsaw.
Sustainable Metal Cloud (SMC) - Firmus/ STT GDC	▪ Sustainable Metal Cloud (SMC) offers a GPU-centric IaaS for deep learning AI and visual computing workloads. ▪ SMC uses Firmus' proprietary, scaled, immersion-cooled platform, the 'HyperCube', hosted within global STT GDC locations. ▪ The platform can host up to 130kW per 42RU and has been designed with industry players such as NVIDIA. The partnership aims to reduce power usage, CO2 emissions, and petaflops per watt for AI workloads.
Vultr	▪ Vultr is a technology company that provides a cloud platform for developers, offering remote access to data, control panel, APIs, instances deployment, and applications acceleration. Established in 2014, it has 32 datacenter locations across six continents. ▪ The company offers GPU options for HPC, machine learning, and gaming applications, including NVIDIA GPUs. ▪ Vultr's GPU products and services include Talon Cloud GPU, Bare Metal, Kubernetes Engine, and Marketplace. Vultr's GPU revenue model is based on pay-as-you-go pricing with hourly billing, with discounts for reserved instances and long-term contracts.

Source: Company reports, RBC Capital Markets

Recent Developments Related to AI and GPU based Private Companies

Company	Date	Development
CoreWeave	Mar. 2024	CoreWeave has leased 16MW of data center space from Bitcoin mining and digital infrastructure provider Core Scientific.
CoreWeave	Dec. 2023	CoreWeave has purchased thousands of Dell PowerEdge servers to enhance its cloud infrastructure. These servers will feature Nvidia's H100 Tensor Core GPUs to provide AI and generative AI services. The servers will handle various workloads, including machine learning, visual effects rendering, and large-scale simulations.
CoreWeave	Sep. 2023	Per Bloomberg, CoreWeave has sold a minority stake worth ~$7B, with Fidelity Investments purchasing the majority of the $500M in employee-owned shares that were tendered. Additionally, the company is expected to generate ~$1.5B in revenue by 2024.
CoreWeave	Sep. 2023	In 3Q23, Digital Realty signed a 32 MW lease with CoreWeave in Portland, where CoreWeave plans to deploy 32,000 Nvidia H100 GPUs.
CoreWeave	Aug. 2023	CoreWeave co-founder Brannin McBee said that CoreWeave had $30M in revenue 2022 and projected revenues for 2023 are ~$500M.
CoreWeave	Aug. 2023	CoreWeave has secured a $2.3B debt financing facility, led by Magnetar Capital and funds managed by Blackstone Tactical Opportunities. The financing is backed by NVIDIA H100 GPUs, which serve as collateral for the loan.
CoreWeave	Jul. 2023	CoreWeave unveiled the world's fastest AI supercomputer built in partnership with NVIDIA, measured by the industry standard benchmark test called the MLPerf. CoreWeave's publicly available supercomputing infrastructure trained the new MLPerf GPT-3 175B LLM in under 11 minutes, which was more than 29x faster than the next best competitor and 4x larger than the next best competitor.
CoreWeave	Jul. 2023	CoreWeave announced plans to spend $1.6B on a datacenter in Plano, Texas. The company will have to invest at least $800m a year for the next two years to be eligible for a tax rebate passed by the Plano City Council.
CoreWeave	Jun. 2023	CoreWeave has co-developed a commercially available cluster of 3,584 NVIDIA H100 Tensor Core GPUs with startup Inflection AI.
CoreWeave	Jun. 2023	Microsoft signed a multi-year deal with CoreWeave to use its Datacenters for some of its Azure AI workloads. CoreWeave currently offers three Datacenter regions: US East in Weehawken, New Jersey; US West in Las Vegas, Nevada; and US Central in Chicago, Illinois.
CoreWeave	May 2023	Nvidia invested in the company as part of a $221M round and gave CoreWeave priority access to GPUs. In May 2023, CoreWeave raised another $200M.

Source: Company reports, RBC Capital Markets

Recent Developments Related to AI and GPU based Private Companies

Company	Date	Development
Lambda Labs	Feb. 2024	Lambda Labs has raised $320 million in a Series C funding round. The round was led by billionaire Thomas Tull's US Innovative Technology fund, with participation from new investors B Capital, and SK Telecom, and existing investors including Crescent Cove, Mercato Partners, 1517 Fund, Bloomberg Beta, and Gradient Ventures. The funding round values the company at $1.5 billion. Lambda said this new financing will allow the company to further accelerate the growth of its GPU cloud, providing AI engineering teams with access to Nvidia GPUs
Lambda Labs	Oct. 2023	Lambda Labs is nearing a $300 million funding round led by Baire Thomas Tull, founder of Legendary Entertainment. The company forecasts revenues of $250 million for 2023 and nearly $600 million for 2024, amid a surge in demand from the generative AI boom.
Lambda Labs	Mar. 2023	Lambda raised $44 million from several investors, including OpenAI co-founder Greg Brockman.
Lambda Labs	Mar. 2023	Lambda Labs co-founder and CEO Stephen Balaban said that "Lambda is building the best cloud in the world for training AI" and that it has seen extreme growth in its cloud product over the past couple of years.
Crusoe Energy	Dec. 2023	Crusoe Energy has announced a deal to locate GPU processors in atNorth's ICE02 datacenter in Iceland. The GPUs will deliver Crusoe's cloud service and run on renewable power, but the announcement is a departure from Crusoe's normal business model, which is to site its own containerized micro datacenters at sites where there is stranded energy, such as otherwise-wasted natural gas at oil wells.
Crusoe Energy	Oct. 2023	Crusoe announced significant expansion of cloud business with new capacity and $200 million in new financing. The new capacity includes NVIDIA H100 Tensor Core GPUs delivering an order-of-magnitude leap in performance to power AI, as well as additional NVIDIA A100 TensorCore GPUs, connected with high-speed NVIDIA Quantum-2 InfiniBand networking.
Crusoe Energy	Oct. 2023	Crusoe announced a commitment from investment firm Upper90 for asset-backed financing for the purposes of securing additional GPUs. The most recent debt financing from Upper90 will allow Crusoe to continue rapid investment in the infrastructure needed to scale its AI cloud offering. The additional GPU capacity is expected to be available to customers in 1Q24.
Crusoe Energy	Apr. 2022	Crusoe closed a $350 million Series C equity offering and also secured credit facilities expandable up to $155 million with SVB Capital, Sparkfund, and Generate Capital to provide additional debt capital for energy systems related to flare mitigation.

Source: Company reports, RBC Capital Markets

Recent Developments Related to AI and GPU based Private Companies

Company	Date	Development
Gcore	Feb. 2024	Gcore announced the launch of FastEdge, a cutting-edge serverless product revolutionizing application deployment and performance. Designed for cloud-native development, FastEdge is a low-latency, high-performance solution for creating responsive and personalized applications without the complexities of server management.
Gcore	Feb. 2024	Gcore has launched a speech-to-text service that translates English to Luxembourgish. By employing AI, Gcore has successfully developed an advanced machine learning model for speech-to-text translation, powered by its Gcore edge AI technology.
Gcore	Oct. 2023	Gcore announced the launch of its Generative AI Cluster powered by NVIDIA A100 and H100 Tensor Core GPUs.
Gcore	Jul. 2023	Gcore launched an AI Cloud cluster in Newport, Wales, its third such deployment after the Netherlands and Luxembourg.
Gcore	Apr. 2023	Gcore partnered with Nvidia to offer GPU-powered cloud computing services for AI and machine learning applications.
Gcore	Mar. 2023	Gcore raised $50 million in Series B funding to expand its global infrastructure and develop new cloud-based services.
Applied Digital	Feb. 2024	Applied obtained an unsecured loan with a maximum principal amount of $20 million. The company intends to use the loan, which has a two-year term and 12.5% interest rate, to provide additional liquidity to fund the buildout of its HPC datacenters. Additionally, the company has secured $16 million in site-level financing for its Jamestown HPC facility, which is expected to close in the coming weeks.
Applied Digital	Oct. 2023	Applied indicated that it had scaled scale up its commitment with Character.ai all the way to 10K GPUs and is now expanding to 16K GPUs and beyond for 2024.
Applied Digital	Jun. 2023	Applied announced securing its second AI customer with an agreement worth up to $460 million over 36 months.
Applied Digital	May 2023	Applied Digital Corporation secured its first major AI customer, Character.AI, with an agreement worth up to $180 million over a 24-month period. The service, which uses NVIDIA H100 GPUs, went online in June and is expected to be fully operational by the end of the year.
Clever Cloud	Sep. 2023	Clever Cloud has appointed Jean-Baptiste Piacentino as a Cloud Diplomat to promote its technological vision and the technical interests of European cloud players.
Denvr Dataworks	Aug. 2023	Dell Technologies and Denvr Dataworks are partnering to accelerate the adoption of GenAI by combining Dell PowerEdge XE9680 server security with Denvr Dataworks' high-performance cloud computing for AI.

Source: Company reports, RBC Capital Markets

Recent Developments Related to AI and GPU based Private Companies

Company	Date	Development
NextGen Cloud	Feb. 2024	NexGen Cloud and AQ Compute are planning a net zero "AI Supercloud" that will be hosted in the latter's datacenter in Oslo, Norway.
NextGen Cloud	Jan. 2024	WEKA is partnering with NexGen Cloud, to provide the high-performance infrastructure foundation underpinning its forthcoming AI Supercloud, as well as the on-demand services offered by Hyperstack, NexGen Cloud's GPUaaS platform.
NextGen Cloud	Sep. 2023	NexGen Cloud plans to invest $1 billion in Europe's first AI Supercloud deployment, providing a dedicated platform for European technology companies, organizations, and governments to execute sensitive AI applications and research within European jurisdiction and privacy laws.
NextGen Cloud	Aug. 2023	NexGen Cloud has launched Hyperstack, an NVIDIA GPU-accelerated cloud platform for European scale-ups.
Vultr	Mar. 2024	Vultr has expanded its Seattle cloud datacenter region with Nvidia H100 GPU clusters. The GPU clusters are available to customers both on demand and through reserved instance contracts.
Vultr	Feb. 2024	Vultr announced the launch of Vultr CDN. This next-generation content delivery service pushes content closer to the edge without compromising security. Building atop Vultr's global infrastructure spanning six continents, Vultr now enables global content and media caching, empowering Vultr's worldwide community of over 225K developers with turnkey services for scaling their websites and web applications..
Vultr	Sep. 2023	Vultr announced the launch of the Vultr GPU Stack and Container Registry to enable global enterprises and digital startups alike to build, test and operationalize AI models at scale—across any region on the globe.

Source: Company reports, RBC Capital Markets

Recent Developments Related to AI and GPU based Private Companies

Company	Date	Development
Paperspace (acquired by DigitalOcean)	Jan. 2024	Digital Ocean announced virtualized availability of NVIDIA H100 Tensor Core GPUs on its Paperspace platform.
Paperspace (acquired by DigitalOcean)	Nov. 2023	DigitalOcean indicated that Paperspace contributed $3 million to its 3Q23 revenues, and exceeded the $1 million mark in monthly revenue in September 2023. The monthly recuring revenue (MRR) of $2.8 million was reported for the quarter. By the end of FY23, it is expected to be just over $3M in MRR, representing an annual recuring revenue of $12M.
Paperspace (acquired by DigitalOcean)	Jul. 2023	DigitalOcean, a cloud hosting provider, acquired Paperspace for $111 million. Paperspace will remain a standalone business unit within DigitalOcean.
Jarvis Labs	Nov. 2023	JarvisLabs.ai closed an undisclosed funding round led by investment firm Y Combinator. The funds will reportedly be used to further develop the company's AI technology and expand its team.
Firmus	Jun. 2023	ST Telemedia Global Data Centres (STT GDC), a Singapore-based datacenter provider, announced a significant investment into a global venture with Firmus Technologies.

Source: Company reports, RBC Capital Markets

Required Disclosures

Companies mentioned

Accenture Public Limited Company (NYSE: ACN US; $ 374.60; Outperform)

Adobe Inc. (NASDAQ: ADBE US; $492.46; Outperform)

The AES Corporation (NYSE: AES US; $15.01; Outperform)

Alphabet Inc. (NASDAQ: GOOGL US; $141.18; Outperform)

Amazon.com, Inc. (NASDAQ: AMZN US; $174.42; Outperform)

Celestica Inc. (NYSE: CLS US; $43.97; Outperform)

Digital Realty Trust, Inc. (NYSE: DLR US; $140.86; Outperform)

DigitalBridge Group, Inc. (NYSE: DBRG US; $18.48; Outperform)

Eaton Corporation Public Limited Company (NYSE: ETN US; $297.90; Sector Perform)

Equifax Inc. (NYSE: EFX US; $252.94; Sector Perform)

Equinix, Inc. (NASDAQ: EQIX US; $850.39; Outperform)

GoDaddy Inc. (NYSE: GDDY US; $117.19; Outperform)

International Business Machines Corporation (NYSE: IBM US; $191.07; Outperform)

Macquarie Technology Group Limited (ASX: MAQ AU; AUD78.36; Sector Perform)

Meta Platforms, Inc. (NASDAQ: META US; $484.10; Outperform)

Microsoft Corporation (NASDAQ: MSFT US; $416.42; Outperform)

Moody's Corporation (NYSE: MCO US; $384.16; Outperform)

MSCI Inc. (NYSE: MSCI US; $544.74; Outperform)

NEXTDC Limited (ASX: NXT AU; AUD17.51; Outperform)

nVent Electric PLC (NYSE: NVT US; $69.96; Outperform)

S&P Global Inc. (NYSE: SPGI US; $422.81; Outperform)

Salesforce, Inc. (NYSE: CRM US; $294.33; Outperform)

Schneider Electric SE (NXT PA: SU FP; EUR215.00; Underperform)

ServiceNow, Inc. (NYSE: NOW US; $743.91; Outperform)

Snowflake Inc. (NYSE: SNOW US; $156.97; Outperform)

Verisk Analytics, Inc. (NASDAQ: VRSK US; $234.52; Outperform)

WESCO International, Inc. (NYSE: WCC US; $160.13; Sector Perform)

Workday, Inc. (NASDAQ: WDAY US; $268.27; Outperform)

Required Disclosures

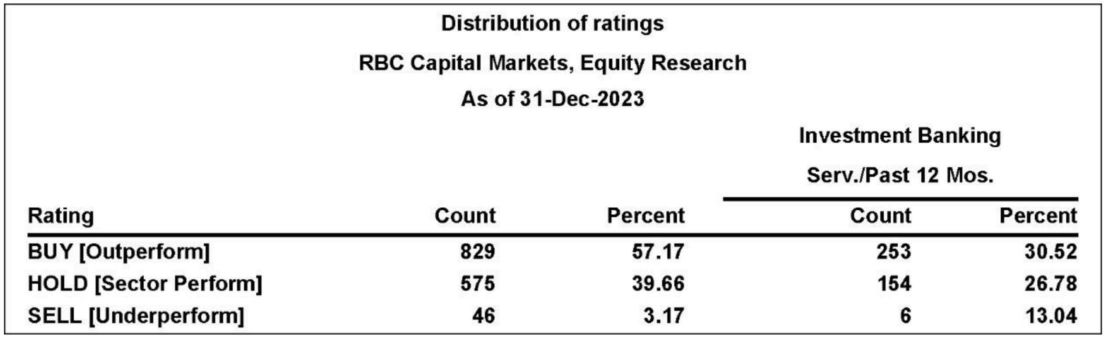
Non-U.S. Analyst Disclosure
One or more research analysts involved in the preparation of this report (i) may not be registered/qualified as research analysts with the NYSE and/or FINRA and (ii) may not be associated persons of the RBC Capital Markets, LLC and therefore may not be subject to FINRA Rule 2241 restrictions on communications with a subject company, public appearances and trading securities held by a research analyst account.

Conflicts Disclosures
This product constitutes a compendium report (covers six or more subject companies). As such, RBC Capital Markets chooses to provide specific disclosures for the subject companies by reference. To access conflict of interest and other disclosures for the subject companies, clients should refer to https://www.rbccm.com/GLDisclosure/PublicWeb/DisclosureLookup.aspx?entityId=1. These disclosures are also available by sending a written request to RBC Capital Markets Research Publishing, P.O. Box 50, 200 Bay Street, Royal Bank Plaza, 29th Floor, South Tower, Toronto, Ontario M5J 2W7 or an email to rbcinsight@rbccm.com.

The analyst(s) responsible for preparing this research report received compensation that is based upon various factors, including total revenues of the member companies of RBC Capital Markets and its affiliates, a portion of which are or have been generated by investment banking activities of the member companies of RBC Capital Markets and its affiliates.

With regard to the MAR investment recommendation requirements in relation to relevant securities, a member company of Royal Bank of Canada, together with its affiliates, may have a net long or short financial interest in excess of 0.5% of the total issued share capital of the entities mentioned in the investment recommendation. Information relating to this is available upon request from your RBC investment advisor or institutional salesperson.

Distribution of Ratings
For the purpose of ratings distributions, regulatory rules require member firms to assign ratings to one of three rating categories - Buy, Hold/Neutral, or Sell - regardless of a firm's own rating categories. Although RBC Capital Markets' ratings of Outperform (O), Sector Perform (SP), and Underperform (U) most closely correspond to Buy, Hold/Neutral and Sell, respectively, the meanings are not the same because our ratings are determined on a relative basis.



Distribution of ratings
RBC Capital Markets, Equity Research
As of 31-Dec-2023

			Investment Banking Serv./Past 12 Mos.	
Rating	Count	Percent	Count	Percent
BUY [Outperform]	829	57.17	253	30.52
HOLD [Sector Perform]	575	39.66	154	26.78
SELL [Underperform]	46	3.17	6	13.04

Required Disclosures

Explanation of RBC Capital Markets Equity Rating System

An analyst's "sector" is the universe of companies for which the analyst provides research coverage. Accordingly, the rating assigned to a particular stock represents solely the analyst's view of how that stock will perform over the next 12 months relative to the analyst's sector average.

Ratings

Outperform (O): Expected to materially outperform sector average over 12 months.

Sector Perform (SP): Returns expected to be in line with sector average over 12 months.

Underperform (U): Returns expected to be materially below sector average over 12 months.

Restricted (R): RBC policy precludes certain types of communications, including an investment recommendation, when RBC is acting as an advisor in certain merger or other strategic transactions and in certain other circumstances.

Not Rated (NR): The rating, price targets and estimates have been removed due to applicable legal, regulatory or policy constraints which may include when RBC Capital Markets is acting in an advisory capacity involving the company.

Risk Rating: The **Speculative** risk rating reflects a security's lower level of financial or operating predictability, illiquid share trading volumes, high balance sheet leverage, or limited operating history that result in a higher expectation of financial and/or stock price volatility.

Conflicts Policy

RBC Capital Markets Policy for Managing Conflicts of Interest in Relation to Investment Research is available from us on request. To access our current policy, clients should refer to https://www.rbccm.com/global/file-414164.pdf or send a request to RBC CM Research Publishing, P.O. Box 50, 200 Bay Street, Royal Bank Plaza, 29th Floor, South Tower, Toronto, Ontario M5J 2W7. We reserve the right to amend or supplement this policy at any time.

Dissemination of research

RBC Capital Markets endeavors to make all reasonable efforts to provide research content simultaneously to all eligible clients, having regard to local time zones in overseas jurisdictions. RBC Capital Markets provides eligible clients with access to Research Reports on the Firm's proprietary INSIGHT website, via email and via third-party vendors. Please contact your investment advisor or institutional salesperson for more information regarding RBC Capital Markets' research.

For a list of all recommendations on the company that were disseminated during the prior 12-month period, please click on the following link: https://rbcnew.bluematrix.com/sellside/MAR.action

The 12 month history of Quick Takes can be viewed at https://www.rbcinsightresearch.com/.

Analyst Certification

All of the views expressed in this report accurately reflect the personal views of the responsible analyst(s) about any and all of the subject securities or issuers. No part of the compensation of the responsible analyst(s) named herein is, or will be, directly or indirectly, related to the specific recommendations or views expressed by the responsible analyst(s) in this report.

Third-party disclaimers

The Global Industry Classification Standard (“GICS”) was developed by and is the exclusive property and a service mark of MSCI Inc. (“MSCI”) and Standard & Poor’s Financial Services LLC (“S&P”) and is licensed for use by RBC. Neither MSCI, S&P, nor any other party involved in making or compiling the GICS or any GICS classifications makes any express or implied warranties or representations with respect to such standard or classification (or the results to be obtained by the use thereof), and all such parties hereby expressly disclaim all warranties of originality, accuracy, completeness, merchantability and fitness for a particular purpose with respect to any of such standard or classification. Without limiting any of the foregoing, in no event shall MSCI, S&P, any of their affiliates or any third party involved in making or compiling the GICS or any GICS classifications have any liability for any direct, indirect, special, punitive, consequential or any other damages (including lost profits) even if notified of the possibility of such damages.

RBC Capital Markets disclaims all warranties of originality, accuracy, completeness, merchantability or fitness for a particular purpose with respect to any statements made to the media or via social media that are in turn quoted in this report, or otherwise reproduced graphically for informational purposes.

Disclaimer

RBC Capital Markets is the business name used by certain branches and subsidiaries of the Royal Bank of Canada, including RBC Dominion Securities Inc., RBC Capital Markets, LLC, RBC Europe Limited, RBC Capital Markets (Europe) GmbH, Royal Bank of Canada, Hong Kong Branch, Royal Bank of Canada, Singapore Branch and Royal Bank of Canada, Sydney Branch. The information contained in this report has been compiled by RBC Capital Markets from sources believed to be reliable, but no representation or warranty, express or implied, is made by Royal Bank of Canada, RBC Capital Markets, its affiliates or any other person as to its accuracy, completeness or correctness. All opinions and estimates contained in this report constitute RBC Capital Markets' judgement as of the date of this report, are subject to change without notice and are provided in good faith but without legal responsibility. Nothing in this report constitutes legal, accounting or tax advice or individually tailored investment advice. This material is prepared for general circulation to clients and has been prepared without regard to the individual financial circumstances and objectives of persons who receive it. The investments or services contained in this report may not be suitable for you and it is recommended that you consult an independent investment advisor if you are in doubt about the suitability of such investments or services. This report is not an offer to sell or a solicitation of an offer to buy any securities. Past performance is not a guide to future performance, future returns are not guaranteed, and a loss of original capital may occur. RBC Capital Markets research analyst compensation is based in part on the overall profitability of RBC Capital Markets, which includes profits attributable to investment banking revenues. Every province in Canada, state in the U.S., and most countries throughout the world have their own laws regulating the types of securities and other investment products which may be offered to their residents, as well as the process for doing so. As a result, the securities discussed in this report may not be eligible for sale in some jurisdictions. RBC Capital Markets may be restricted from publishing research reports, from time to time, due to regulatory restrictions and/ or internal compliance policies. If this is the case, the latest published research reports available to clients may not reflect recent material changes in the applicable industry and/or applicable subject companies. RBC Capital Markets research reports are current only as of the date set forth on the research reports. This report is not, and under no circumstances should be construed as, a solicitation to act as securities broker or dealer in any jurisdiction by any person or company that is not legally permitted to carry on the business of a securities broker or dealer in that jurisdiction. To the full extent permitted by law neither RBC Capital Markets nor any of its affiliates, nor any other person, accepts any liability whatsoever for any direct, indirect or consequential loss arising from, or in connection with, any use of this report or the information contained herein. No matter contained in this document may be reproduced or copied by any means without the prior written consent of RBC Capital Markets in each instance.

Additional information is available on request.

To U.S. Residents: This publication has been approved by RBC Capital Markets, LLC (member FINRA, NYSE, SIPC), which is a U.S. registered broker-dealer and which accepts responsibility for this report and its dissemination in the United States. Any U.S. recipient of this report that is not a registered broker-dealer or a bank acting in a broker or dealer capacity and that wishes further information regarding, or to effect any transaction in, any of the securities discussed in this report, should contact and place orders with RBC Capital Markets, LLC.

To Canadian Residents: This publication has been approved by RBC Dominion Securities Inc. (member CIRO). Any Canadian recipient of this report that is not a Designated Institution in Ontario, an Accredited Investor in British Columbia or Alberta or a Sophisticated Purchaser in Quebec (or similar permitted purchaser in any other province) and that wishes further information regarding, or to effect any transaction in, any of the securities discussed in this report should contact and place orders with RBC Dominion Securities Inc., which, without in any way limiting the foregoing, accepts responsibility for this report and its dissemination in Canada.

To U.K. Residents: This publication has been approved by RBC Europe Limited ('RBCEL') which is authorized by the Prudential Regulation Authority and regulated by the Financial Conduct Authority ('FCA') and the Prudential Regulation Authority, in connection with its distribution in the United Kingdom. This material is not for general distribution in the United Kingdom to retail clients, as defined under the rules of the FCA. RBCEL accepts responsibility for this report and its dissemination in the United Kingdom.

To EEA Residents: This material is distributed in the EU by either RBCEL on an authorised cross-border basis, or by RBC Capital Markets (Europe) GmbH (RBC EG) which is authorised and regulated in Germany by the Bundesanstalt für Finanzdienstleistungsaufsicht (German Federal Financial Supervisory Authority) (BaFin).

To Persons Receiving This Advice in Australia: This material has been distributed in Australia by Royal Bank of Canada, Sydney Branch (ABN 86 076 940 880, AFSL No. 246521). This material has been prepared for general circulation and does not take into account the objectives, financial situation or needs of any recipient. Accordingly, any recipient should, before acting on this material, consider the appropriateness of this material having regard to their objectives, financial situation and needs. If this material relates to the acquisition or possible acquisition of a particular financial product, a recipient in Australia should obtain any relevant disclosure document prepared in respect of that product and consider that document before making any decision about whether to acquire the product. This research report is not for retail investors as defined in section 761G of the Corporations Act.

To persons receiving this from Royal Bank of Canada, Hong Kong Branch: This document is distributed in Hong Kong by Royal Bank of Canada, Hong Kong Branch which is regulated by the Hong Kong Monetary Authority and the Securities and Futures Commission. This document is not for distribution in Hong Kong, to investors who are not “professional investors”, as defined in the Securities and Futures Ordinance (Cap. 571 of Hong Kong) and any rules made under that Ordinance. This document has been prepared for general circulation and does not take into account the objectives, financial situation, or needs of any recipient. Past performance is not indicative of future performance. WARNING: The contents of this document have not been reviewed by any regulatory authority in Hong Kong. Investors are advised to exercise caution in relation to the investment. If you are in doubt about any of the contents of this document, you should obtain independent professional advice.

To persons receiving this from Royal Bank of Canada, Singapore Branch: This publication is distributed in Singapore by the Royal Bank of Canada, Singapore Branch, a registered entity licensed by the Monetary Authority of Singapore. This publication is not for distribution in Singapore, to investors who are not “accredited investors” and “institutional investors”, as defined in the Securities and Futures Act 2001 of Singapore. This publication has been prepared for general circulation and does not take into account the objectives, financial situation, or needs of any recipient. You are advised to seek independent advice from a financial adviser before purchasing any product. If you do not obtain independent advice, you should consider whether the product is suitable for you. Past performance is not indicative of future performance. If you have any questions related to this publication, please contact the Royal Bank of Canada, Singapore Branch.

To Japanese Residents: Unless otherwise exempted by Japanese law, this publication is distributed in Japan by or through RBC Capital Markets (Japan) Ltd. which is a Financial Instruments Firm registered with the Kanto Local Financial Bureau (Registered number 203) and a member of the Japan Securities Dealers Association (“JSDA”) and the Financial Futures Association of Japan (“FFAJ”).