EQUITY I **RESEARCH**

RBC Generative AI Update

13 November 2023

RBC Capital Markets, LLC

Jonathan Atkin (Analyst) (650) 218-8653 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

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

RBC Dominion Securities Inc.

Maxim Matushansky (Analyst) (416) 842-4455 maxim.matushansky@rbccm.com

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

Royal Bank of Canada, Sydney Branch

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

RBC Capital Markets

This report is priced as of market close 10 November 2023.
All values in U.S. dollars unless otherwise noted.

For Required Non-U.S. Analyst and Conflicts Disclosures, please see page 74.

Disseminated: Nov 13, 2023 08:31EST; Produced: Nov 13, 2023 08:31EST

Table of Contents

We would like to highlight the contributions of Saurabh Ajmera to this report.

Section 1

Executive Summary and Key Highlights

RBC Capital Markets

Broad Themes

1. **Strong Financial Performances:** Most companies in the broader sector have posted strong financial performances. AI features as a major growth driver for companies such as: AMZN, DLR, INTC, META, MSFT, MRVL, NVDA, ORCL, and VRTV.

2. **AI Investments:** AI is seeing integration into multiple aspects of business operations, including cloud platforms & datacenters (as outlined below), and even on the opex side (in both sales and marketing, as well as R&D). At the same time, AI is creating new system design challenges around power consumption and cooling requirements.

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 major hyperscalers, there are a plethora of challenges enterprises are grappling with in their adoption of AI.

4. **GPU Shortages**: Continuing challenges in procuring GPUs is pressuring the ability to meet demand for AI. Shortages are being managed through strategic partnerships and investments.

5. **Capex Growth and Datacenter Expansion:** Companies are heavily investing in expanding their datacenter infrastructure to support the growing demand for AI, cloud, and other services (SaaS, e-commerce, social networking). These include existing hyperscalers and independent GPU-as-a-service operators. As such, capex levels are growing, with companies increasing investments in datacenters, AI infrastructure, and cloud computing.

Broad Themes (continued)

6. **AI Topology:** Much of the larger MW growth has been see in requirements related to Large Language Models, which are being established in or near existing cloud availability zones by the major cloud provider, and in more remote areas by other participants. Topologies around AI inferencing will likely be quite varied, from far edge end-point 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 and various "expert models" or "nimble models") that can deliver an excellent user experience with a vastly smaller model.**

7. **Strategic Partnerships:** Companies are forming 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, AMD/Mipsology and Nod.ai.

8. **Industrial/Energy Implications.** Energy requirements around AI are placing a greater focus on energy procurement and sustainability. Energy and other resource constraints will create challenges for all industry participants. Over time, newer approaches, such as small modular reactors, could see greater focus. As it pertains to datacenter design, liquid cooling solutions to enhance rack densities needed for Gen AI are seeing greater focus, though many datacenter operators are not changing their baseline designs, which they find can meet current AI requirements.

9. **Top AI-related Beneficiaries:** Within RBC's coverage universe, stocks that are substantially affected by AI-related developments include: Alphabet (Google), Amazon (AMZN), Digital Bridge (DBRG), Digital Realty (DLR), Meta Platforms (Meta), NextDC (NXT), WIX.com (WIX), Xero (XRO).

Chip Makers - Revenue and Commercial Trends

1. **AMD:** In 3QFY23, total revenue increased by 4% Y/Y and 8% Q/Q. The Datacenter segment, accounting for ~28% of total revenues, declined 1% Y/Y but was up 21% Q/Q, driven by the accelerated adoption of 4th Gen AMD EPYC CPUs during the quarter. The Client segment, representing ~25% of total revenues, saw a significant increase of 42% Y/Y and 46% Q/Q, driven by an increase in AMD Ryzen 7000 Series CPU sales.

2. **Intel:** In 3QFY23, total revenue declined by 8% Y/Y but saw a 9% Q/Q increase. Datacenter and AI (DCAI) revenue, representing 27% of total revenues, decreased by 5% Q/Q and 10% Y/Y due to a decrease in server revenue in a softening CPU Datacenter market. Network and Edge (NEX) revenue, accounting for 10% of total revenues, increased by 6% Q/Q but was down 32% Y/Y as customers reduced purchases to adjust to a lower demand environment across product lines.

3. **Marvel Technology:** In 2QFY24, revenue was $1,341 million, marking a decrease of 12% Y/Y but a slight increase of 1% Q/Q. The Datacenter segment, which represented ~34% of total revenues, saw a decrease of 29% Y/Y but an increase of 6% Q/Q. Despite sequential growth in storage data center revenue from a low base in 1Q and expected modest growth in 3Q, the recovery in datacenter storage has been significantly delayed due to depressed end market demand and high customer inventory. However, accelerated sequential revenue growth is anticipated from overall cloud in 3Q, surpassing last quarter's performance, driven by strong growth from Cloud AI and standard cloud infrastructure. Marvell is enabling AI with a broad range of solutions, including PAM4-based optical DSPs and AECs, DCI products, low-latency, high-capacity Ethernet switches, and custom silicon for compute acceleration.

4. **NVIDIA:** In 2QFY24, revenue was $13.5 billion, marking a significant increase of 101% Y/Y and 88% Q/Q. The Datacenter segment, which represented ~76% of total revenues, saw a 171% Y/Y and 141% Q/Q increase, led by CSPs and large consumer internet companies. This was primarily driven by strong demand for the NVIDIA HGX platform based on Nvidia's Hopper and Ampere GPU architectures, largely reflecting the strong ramp of the company's Hopper-based HGX platform. Data Center Compute grew 195% Y/Y and 157% Q/Q, while Networking was up 94% Y/Y and up 85% Q/Q, primarily on strong growth in InfiniBand infrastructure to support Nvidia's HGX platform.

Hyperscalers - Revenue and Commercial Trends

1. **Alibaba:** Alibaba Cloud revenue grew 4% y/y, despite the challenges of lower demand from video streaming, remote working and learning, and a top customer. There is strong demand for model training and related AI services on Alibaba Cloud, which were only partially fulfilled due to near-term supply chain constraints. Alibaba will continue to upgrade its models and pursue an open-source strategy to drive adoption and usage of its computing power.

2. **Amazon:** AWS reported a 12% y/y revenue growth in 3Q23, reaching $23.1 billion and achieving an annualized revenue run rate of $92 billion. AWS announced several new initiatives and collaborations in generative AI, enhancing its product portfolio and customer satisfaction. AWS and Anthropic, a leading LLM maker, have formed a partnership to use AWS's custom chips, Trainium and Inferentia, for generative AI. The partnership will also involve joint development of Trainium and Inferentia technology, which will enhance their price performance benefits for customers. AWS is seeing significant growth in generative AI, with many companies using it to transform their customer experiences.

3. **Google:** Google Cloud reported a 22% y/y revenue growth in 3Q24, reaching $8.4 billion and serving more than 60% of the world's largest companies. There has been a sevenfold increase in active generative AI projects on Vertex AI, with customers such as Highmark Health using AI to create more personalized member materials. Management mentioned that they are seeing strong demand for GPUs, which are being used for tasks such as machine learning and scientific simulations.

4. **Microsoft:** Microsoft Cloud revenue was $31.8 billion and grew 24% and 23% in constant currency. MSFT announced the general availability of our next-generation H100 virtual machines. More than 18,000 organizations now use Azure OpenAI service, including new to Azure customers. Microsoft's Azure Arc enables customers to run apps across different cloud environments, attracting 21,000 customers, up 140% y/y. Microsoft is the only other cloud provider to offer Oracle's database services, simplifying the migration of on-prem Oracle databases to Azure. MSFT claimed to have its AI services deployed in more regions than any other cloud provider.

5. **Oracle:** Oracle's FQ124 Cloud revenue (IaaS plus SaaS were: $4.6 billion, up 30% in USD, and up 29% in constant currency. Oracle claims that its cloud offers superior speed and cost advantages over other clouds, thanks to its RDMA-interconnected NVIDIA superclusters. Oracle said that its GPU prices are lower than the cost of other clouds, but still profitable for its 100% automated cloud.

Hyperscalers - Capex Trends

1. **Amazon:** Amazon defines its capital investments as a combination of CapEx plus equipment finance leases. These investments were $50 billion for the trailing 12-month period ended September 30, down from $60 billion in the comparable prior year period. For FY23, Amazon expects capital investments to be ~$50 billion compared to $59 billion in FY22. The company expects fulfillment and transportation CapEx to be down y/y, partially offset by increased infrastructure CapEx to support growth of our AWS business, including additional investments related to generative AI and large language model efforts.

2. **Google:** Expects a modest y/y increase in FY23 vs. FY22 spend at ~$32 billion. This includes investments in GPUs, proprietary TPUs, and datacenter capacity. Google reported an $8 billion CapEx in 3Q, mainly for its technical infrastructure and AI compute, but the growth was muted by the timing of supplier payments. Google expects to increase its CapEx in 4Q and FY24, as it sees many opportunities in AI across Alphabet. FY24 aggregate CapEx will be above FY23.

3. **Meta:** Meta's capital expenditures in 3Q23 were $6.8 billion, mainly for servers, Datacenters and network infrastructure. This was lower y/y due to less spending on server and Datacenter construction and payment timing. Meta expects its capital expenditures in 2023 to be between $27 billion and $29 billion, slightly lower than its previous estimate of $27 billion to $30 billion. Meta anticipates its capital expenditures in 2024 to be between $30 billion and $35 billion, with growth driven by investments in servers, both non-AI and AI hardware, and Datacenters with the new architecture it announced last year. Meta's main investment priority in 2024 is AI, both in engineering and compute resources. It will reduce the number of non-AI projects and shift more people to work on AI.

4. **Microsoft:** Microsoft expects capex to increase sequentially each quarter throughout FY-2024, including datacenters, physical infrastructure, servers, CPUs, GPUs, and networking equipment. Consensus expectation for Microsoft's capex for CY-2023/CY-2024 is ~$36/$44 billion. The company remains committed to investing for the cloud and AI opportunity while also maintaining our disciplined focus on operating leverage.

5. **Oracle:** Oracle expects FY24 CapEx to be similar to the previous year ($8.7B in FY23), with investments focused on expanding its cloud infrastructure and developing new AI capabilities. The company remains cautious about pacing its investments appropriately and in line with booking trends.

Hyperscalers - GPU Trends

1. **Amazon:**
   - There is currently a shortage of chips in the industry, making it difficult to obtain enough GPUs. This makes Trainium and Inferentia chips more attractive which offer better price performance compared to other options.
   - AWS 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 6x faster at training large-language models than the A100-based EC2 P4 instances and can cut training costs by 40%.
2. **Google:**
   - A3 AI supercomputers powered by NVIDIA's H100 offered as part of Google Cloud's AI-optimized infrastructure, but GPU shortages affecting the industry.
3. **Microsoft:**
   - Microsoft announced the general availability of our next-generation H100 virtual machines.
   - Growth was ahead of expectations in intelligent cloud segment, primarily driven by increased GPU capacity and better-than-expected GPU utilization of its AI services .
   - Microsoft is reportedly developing its own AI chip ('Athena') that will power the technology behind AI chatbots like ChatGPT.
4. **Oracle:**
   - Oracle's GPU is a low-margin business. In some cases, Oracle's GPU prices are lower than the cost of other cloud providers, and that this is very profitable for Oracle.
5. **Meta:**
   - Meta has 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.
   - Meta's new Datacenter architecture is expected to improve its cost efficiency and capacity planning for CPU and GPU.
6. **Alibaba:**
   - Alibaba Cloud has unveiled its latest chip development platform, Wujian 600. It has been created to help global developers in the design of high-performance Systems-on-Chip (SoCs) for edge-AI computing leveraging the RISC-V instruction-set architecture.

Semiconductor Highlights

1. **AI Chips:**
   - Intel, AMD, Marvell, and NVIDIA are all investing heavily in AI chips and platforms.
   - Intel is developing its Gen AI platform, AMD is working on AI processors, and NVIDIA has been a pioneer in AI-focused GPUs.
   - Marvell's AI chips are being used for datacenter and AI applications, delivering strong revenue growth.
   - All companies are actively targeting the AI market with their product offerings and partnering with other companies to accelerate AI adoption.

2. **GPUs:**
   - Intel, AMD, Marvell, and NVIDIA are all investing in GPUs for various applications, including AI, high-performance computing, and gaming.
   - Intel and NVIDIA are experiencing some impact from GPU shortages but remain confident in their ability to meet demand.
   - AMD and Marvell's stance on GPU shortages is not explicitly mentioned, but the industry-wide impact has affected most companies.

3. **Training and Inferencing:**
   - Training AI models requires massive amounts of computational power and data, and all four companies are addressing this challenge with their respective hardware and software solutions.
   - Nvidia's GPUs and CUDA framework are popular choices for training AI models, while AMD's Radeon Pro and Instinct lines are gaining traction in this area. Intel's Xeon Phi line is optimized for high-performance computing and AI training, and Marvell's ThunderX2 platform includes support for training and inference workloads.
   - All four companies offer inference solutions, either through dedicated hardware or software libraries. Nvidia's TensorRT and CUDA are widely used for inference, while AMD's ROCm and Intel's OpenVINO provide alternative options. Marvell's ThunderX2 platform includes support for inference workloads.
   - NVIDIA’s AI platform set new standards in the latest MLPerf benchmarks. The AI supercomputer, NVIDIA Eos, powered by 10,752 NVIDIA H100 Tensor Core GPUs, completed a GPT-3 model training benchmark in just 3.9 minutes, a nearly 3x improvement from the previous record.

4. **Artificial Intelligence:**
   - Intel, AMD, Marvell, and NVIDIA are actively involved in developing and promoting AI technologies, including deep learning, natural language processing, and computer vision.
   - They are partnering with leading AI research institutions and startups to advance the state of the art in AI.
   - Each company has its own AI software frameworks and libraries, such as Nvidia's CUDA, AMD's ROCm, Intel's OpenVINO, and Marvell's TensorFlow Lite.

Highlights – Common Themes from Parts of the Datacenter Supply Chain

Recent commentary from Eaton, nVent, Schneider Electric, Super Micro, Wesco, and Vertiv

- **Growth Drivers:**
  - All six 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.
- **Sustainability**:
  - Schneider Electric, Super Micro, nVent and Vertiv emphasized the importance of sustainability and the need for environmentally friendly solutions.
- **Datacenters**:
  - Datacenter segment demand is strong across the industry, with Schneider Electric, Wesco, Vertiv, Eaton, and Super Micro experiencing growth in this segment due to increased datacenter deployments.
  - Vertiv says many datacenters have been designed to be multipurpose. So, the majority of its products will support many types of compute, including AI.
- **Artificial Intelligence as a driver**
  - AI revenue percentage continues to grow.
  - Vertiv mentions AI and edge computing as significant growth drivers, while NVent highlights the use of AI and data analytics in its solutions for critical infrastructure.
  - Super Micro emphasizes its position in the AI market by offering high-performance computing solutions for AI applications.
- **Cooling**:
  - Liquid cooling technologies are gaining importance, with Schneider Electric, NVent, and Super Micro highlighting solutions for cooling and efficient heat dissipation in datacenters.
  - Super Micro 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 they are looking to double their capacity.
  - Chip or rack-level liquid cooling advantages include higher heat dissipation capacity and lower energy consumption compared to predominately air-cooled approaches, but requires more requires specialized equipment.
  - Immersion cooling, while offering extremely high heat dissipation capacity, has comparatively more limited scalability.
- **Power Densities/Requirements**:
  - Eaton, nVent and Super Micro highlighted solutions to address high-power density infrastructure requirements, while Vertiv focuses on power management and distribution to support demanding applications.
  - Schneider Electric continues to see strong demand across most of its portfolio, driven by strong secular trends, particularly in its Systems business in segments like datacenter and utilities. The management anticipates strong future growth due to the increasing need for electrical content and a technological shift from hyperscalers to colocations and edge.

Section 2

Generative AI Sector Growth

- Gen AI Basics
- LLM Training and Inferencing
- Infrastructure Requirements
- Gen AI Use Cases
- Monetization Strategies
- Capex Trends

RBC Capital Markets

Generative AI Basics

- Generative AI is a type of artificial intelligence technology that can produce various types of content including text, imagery, audio and synthetic data. Generative AI interfaces include ChatGPT (Microsoft), Bard (Google), Dall-E, and others.
- 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. For instance, to train GPT-4, Microsoft and OpenAI used more power than it would take to supply 10,000 homes for two months.
- AI inferencing requirements will entail a more decentralized architecture, with inference nodes arising in varied points in the network.

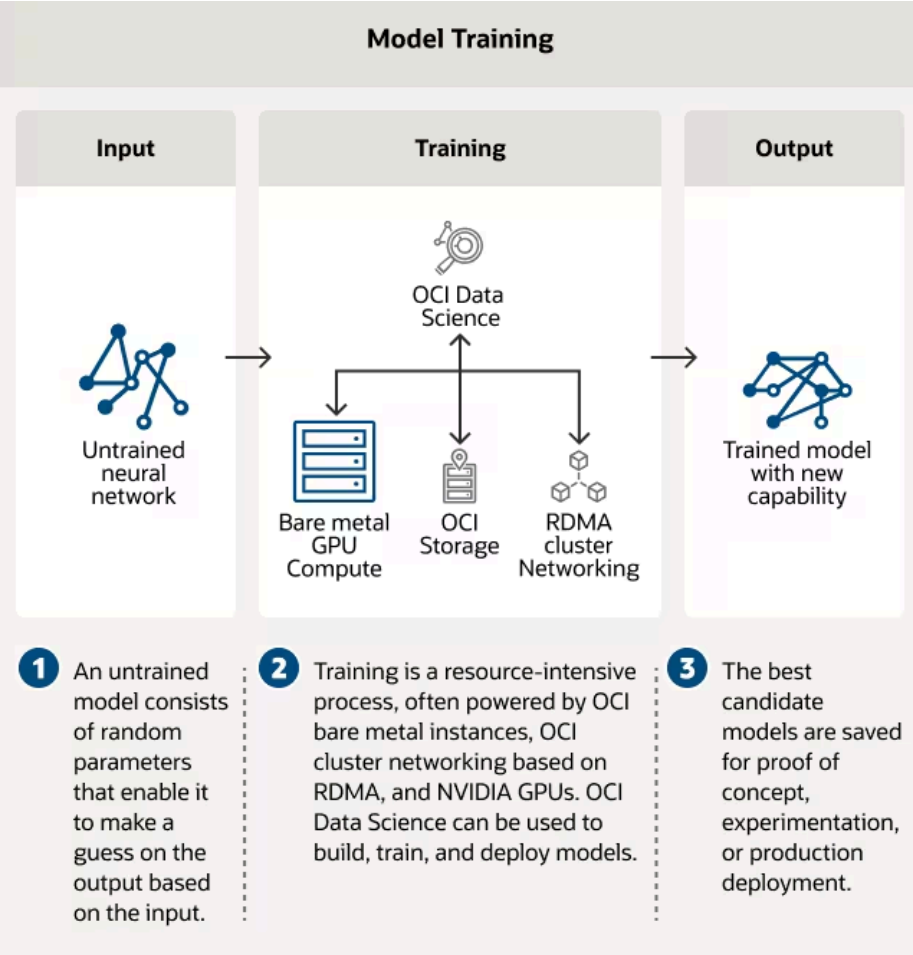
Datacenter Implications

- While a conventional hyperscale datacenter deployment averages 8-10 kW per rack, with some GPU architectures requiring 20 kW/rack or more, a standard NVIDIA system for AI workloads uses 25 kW to 28 kW.
- Thus far in 2023, we estimate it has provided a roughly 50% tailwind to hyperscale demand for datacenter capacity. Energy supply will be a key variable determining the velocity of when this demand can be met.
- Existing datacenter designs have kept up with power density requirements through localized cooling techniques.
- Regulatory implications around generative AI are as yet unclear, with data sovereignty, security, and privacy playing key roles that will influence the sector.

AI Monetization for Software/Internet/Cloud Players

- RBC sees 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.
- Charging for GenAI solutions as a separate SKU (e.g. M365 Copilot and EinsteinGPT). This can incorporate direct pricing such as on a per-seat or consumption basis (e.g. MSFT GitHub Copilot and Azure OpenAI services).
- Adding GenAI capabilities as a premium tier in order to drive greater consumption or upgrades. This can include indirect pricing benefits such as when GenAI drives more usage of a platform, uplifting revenue (e.g. MSFT Azure or MDB Atlas, which can benefit from the velocity of application development).
- Integrating GenAI into the base product and not charging for it, instead using it as a tool to drive competitive differentiation and gross retention. More details here.
- Enterprise adoption challenges include: 1) high costs, 2) "hallucinations" or incorrect responses from GenAI systems (requiring guardrails or human intervention), 3) data privacy and residency concerns, and 4) lack of domain expertise to fully leverage GenAI capabilities.

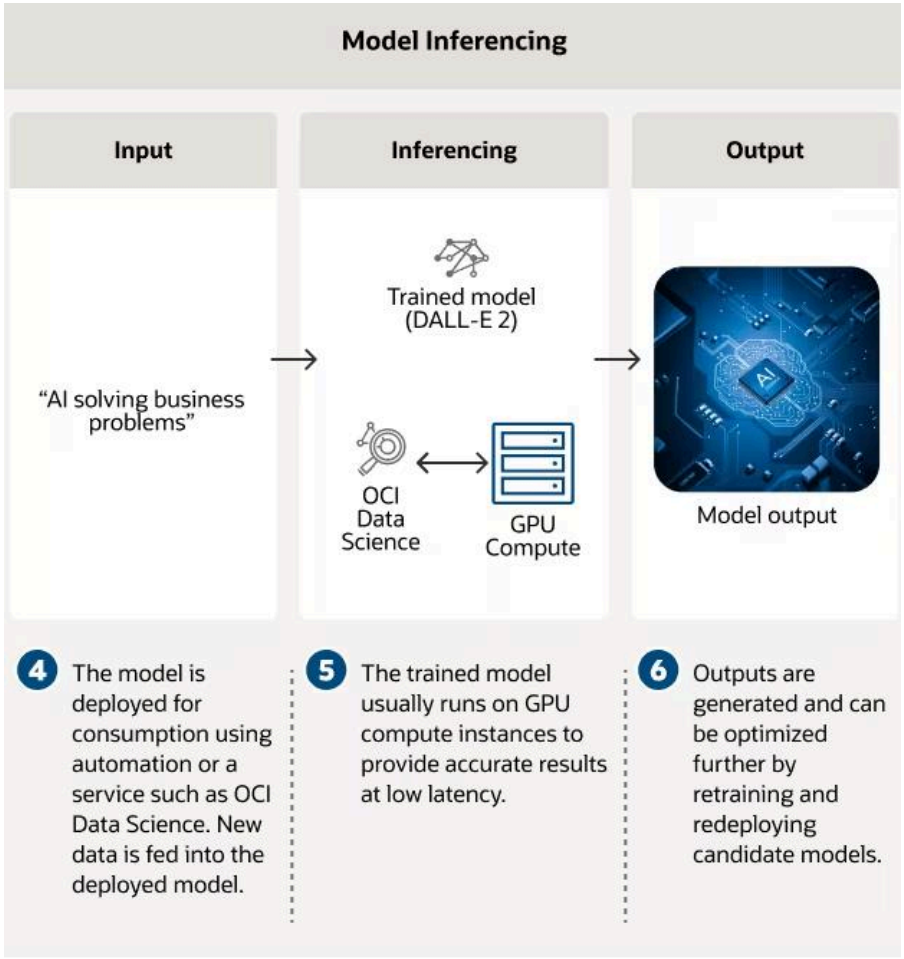
Large Language Model (LLM) Training



- Large language models are trained with billions and trillions 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.
- **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: prompt phase and tprompt phase is compute-bound, while the token phase is memory-bandwidth bound.
- **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.
- Large language models (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 token phase. The providers, 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



- **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

Large Language Model Training and Inferencing – Infrastructure Requirements

Stages	Description	Compute	Memory Capacity	Memory Bandwidth	Network Latency Sensitivity	Network bandwidth
Large Language Model (LLM) Training	Large Language Models (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 Large Language Models (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 Large Language Models (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

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	• Large language models (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

Gen AI – Workplace 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.

Source: Oracle.com

Gen AI Use Cases – Some Examples

Professional Services	• Accenture • 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	• Nvidia • Nvidia has released its BioNeMo Drug Discovery Cloud Service, which uses large language modeling to advance and speed up drug discovery, protein engineering, and research in genomics, chemistry, biology, and molecular dynamics.
Travel & Hospitality	• Expedia • 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 • Shopify now offers Shopify Magic to help retailers generate product descriptions and other product-related content with artificial intelligence.
Fintech & Software Development	• Stripe • 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

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

Overcoming 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 solves 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

AI - Large-scale Opportunity with Very Fast Adoption

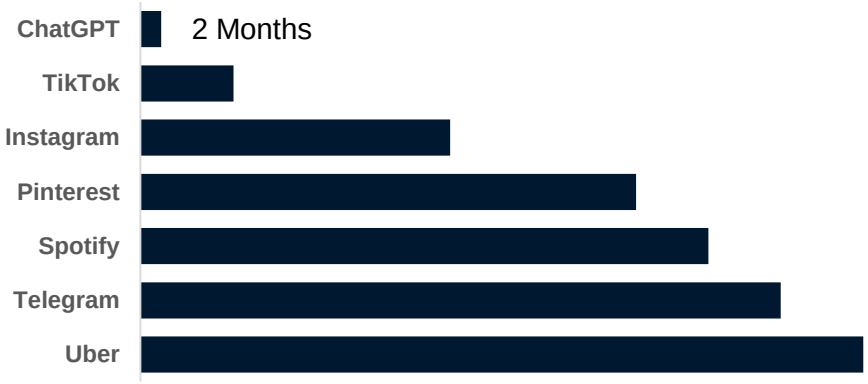
AI IS A CLOUD-SCALE OPPORTUNITY

"This could be big or bigger than the cloud"
DigitalBridge Datacenter portfolio CEO

Sources: DigitalBridge Estimates, Structure Research HSC Market Share Report Q1 2023

GENERATIVE AI ADOPTION IS THE FASTEST ON RECORD

Time it took to reach 100M monthly users worldwide

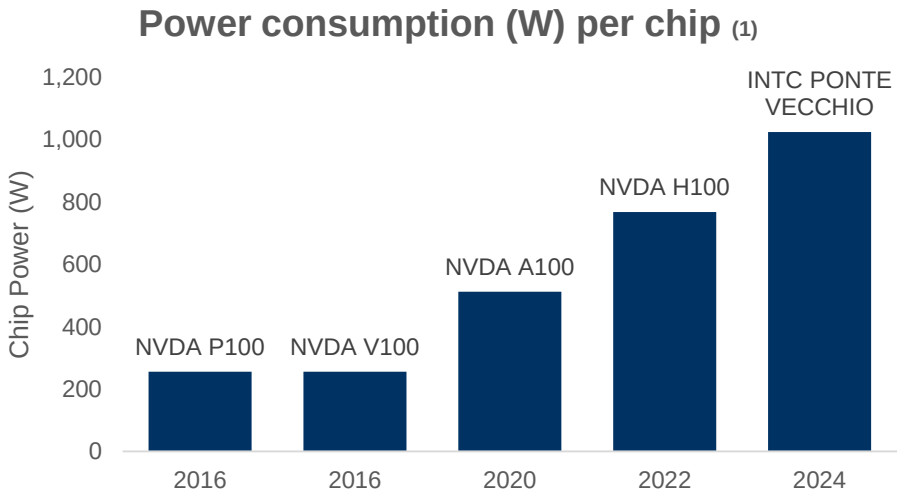


Sources: DBRG August 2023 Presentation, IBM Global AI Adoption Index 2022, IDC Worldwide Artificial Intelligence Spending Guide

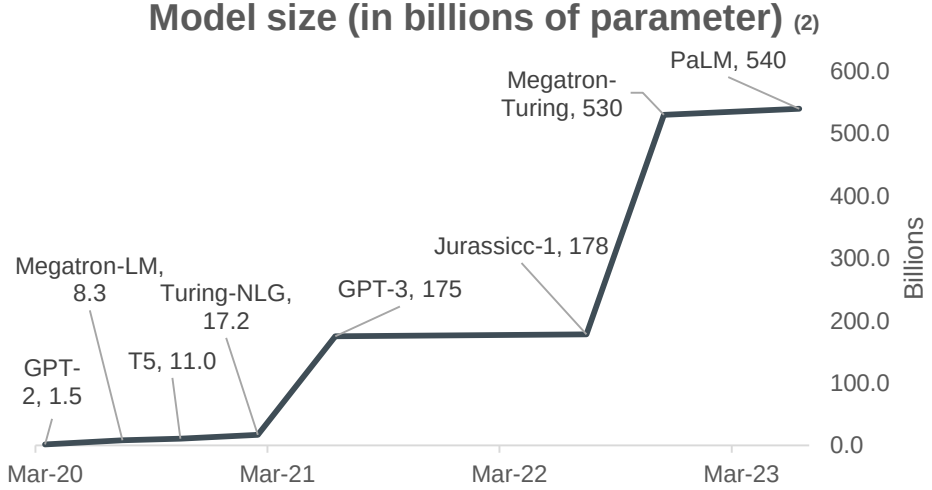
Generative AI Workloads Are Power Intensive

By 2040, ~80% of all datacenter power is expected to be consumed by AI

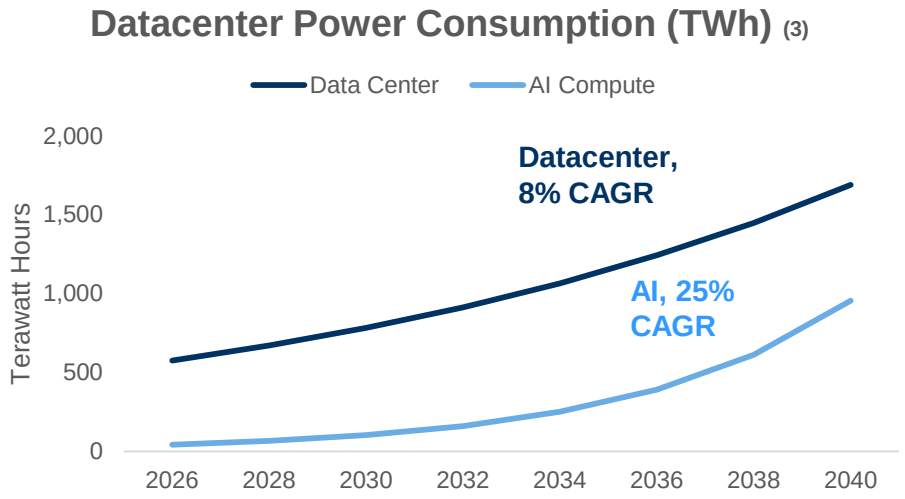
Chip power growth



Large models are growing exponentially



Datacenters will need far more power



Sources: DBRG August 2023 Presentation, (1) INTC, NVIDA, (2) AvidThink

Capex Trends at Major Companies Driving AI-Related Capex

- Alibaba, Microsoft, Meta and Amazon currently should show the largest ramp in 2024E capex
- Amongst the hyperscalers we track below, total capex should grow from $180B in 2023 to $211B in 2024 (17% increase)

(in $M)	CY2020	CY2021	CY2022	CY2023E	CY2024E	2024E-2023E	2024E-2023E ($M)
Amazon	$57,976	$72,325	$60,836	$50,810	$57,782	14%	$6,972
Microsoft	$21,557	$23,216	$24,768	$35,866	$44,479	24%	$8,612
Google	$22,281	$24,640	$31,485	$31,716	$36,338	15%	$4,623
Meta/Facebook	$15,163	$18,690	$31,431	$28,040	$33,159	18%	$5,119
Apple	$8,702	$10,388	$11,692	$10,590	$10,792	2%	$202
Alibaba Group	$4,986	$6,525	$7,729	$3,073	$6,767	120%	$3,694
Oracle	$1,833	$3,118	$6,678	$7,964	$8,750	10%	$785
Tencent	$5,219	$4,613	$3,288	$4,511	$4,936	9%	$425
HPE	$2,328	$2,613	$3,292	$2,730	$2,800	3%	$70
IBM	$2,618	$2,062	$1,346	$1,815	$1,994	10%	$179
Baidu Inc	$779	$1,715	$1,201	$1,169	$1,218	4%	$49
SAP SE	$816	$800	$874	$935	$1,062	14%	$127
salesforce.com	$710	$717	$798	$843	$914	8%	$71
Total Capex	**$144,967**	**$171,422**	**$185,419**	**$180,063**	**$210,992**	**17%**	**$30,928**

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

Section 3

Cloud/Hyperscale Financial Highlights

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

RBC Capital Markets

Earnings and Other Highlights – Amazon Web Services (AWS)

- **AWS segment performance:**
  - AWS reported a 12% y/y revenue growth in 3Q23, reaching $23.1 billion and achieving an annualized revenue run rate of $92 billion.
- **Capex:**
  - Amazon defines its capital investments as a combination of CapEx plus equipment finance leases. These investments were $50 billion for the trailing 12-month period ended September 30, down from $60 billion in the comparable prior year period. For FY23, Amazon expects capital investments to be ~$50 billion compared to $59 billion in FY22.
  - The company expects fulfillment and transportation CapEx to be down y/y, partially offset by increased infrastructure CapEx to support growth of our AWS business, including additional investments related to generative AI and large language model efforts.
- **Artificial Intelligence:**
  - AWS is seeing strong growth in generative AI, with a unique and broad approach that's resonating with customers.
  - The company believes that generative AI will be a tens of billions of dollars opportunity for AWS in the medium to long term.
  - Management noted that generative AI business is growing fast and is already significant, but the market is still in its infancy in regard to monetization.
- **Cost Optimization:**
  - Many customers are taking advantage of enhanced price-performance capabilities in AWS, resulting in significant cost optimization.
  - There's a growing trend of customers migrating from hourly on-demand rates to 1-3 year commitments (savings plans), indicating long-term confidence in AWS.
- **GPUs:**
  - AWS has invested heavily in GPUs and AI, and the company has developed its own custom AI chips for training (Trainium) and inference (Inferentia).
  - There is currently a shortage of chips in the industry, making it difficult to obtain enough GPUs. This makes Trainium and Inferentia chips more attractive which offer better price performance compared to other options.
- **Cloud Regions and Availability Zones:**
  - AWS continued to expand AWS's infrastructure footprint to support customers by launching the AWS Israel (Tel Aviv) Region and a new AWS Local Zone in Phoenix, Arizona.

Source: Company reports

Earnings and Other Highlights – Microsoft

- **Revenue Highlights:** In Q124 Microsoft Cloud revenue was $31.8 billion and grew 24% and 23% in constant currency, ahead of expectations. Microsoft Cloud gross margin percentage increased slightly year-over-year to 73%, 1 point better than expected, primarily driven by improvement in Azure. For the Intelligent Cloud segment -- Revenue was $24.3 billion, increasing 19% and ahead of expectations with better-than-expected results across all businesses. Overall, server products and cloud services revenue grew 21%. Azure and other cloud services revenue grew 29% and 28% in constant currency, including roughly 3 points from AI services.

- **Cloud Workload Optimizations:** Cloud migrations, especially of Oracle databases, which are now available on Azure and attract new customers who want to run their applications on the same cloud platform. AI projects are generating new demand for Azure services, not only for AI meters but also for other cloud meters. More than 18,000 organizations now use Azure OpenAI service, including new to Azure customers. Microsoft continues to see more cloud migrations with Azure Arc. Azure Arc enables customers to run apps across different cloud environments, attracting 21,000 customers, up 140% y/y.

- **Capex:** Microsoft's capital expenditures, including finance leases, were $11.2 billion to support cloud demand, including investments to scale our AI infrastructure. Cash paid for PP&E was $9.9 billion. For FY '24, Microsoft remains committed to investing for the cloud and AI opportunity while also maintaining its disciplined focus on operating leverage.

- **Cloud Regions:** Microsoft has the most comprehensive cloud footprint with more than 60 Datacenter regions worldwide as well as the best AI infrastructure for both training and inference. And company also claims to have AI services deployed in more regions than any other cloud provider.

- **GPUs:** Microsoft announced the general availability of our next-generation H100 virtual machines. Growth was ahead of expectations in intelligent cloud segment, primarily driven by increased GPU capacity and better-than-expected GPU utilization of its AI services.

- **Guidance:** For FY-2H24, assuming the optimization and new workload trends continue and with the growing contribution from AI, Microsoft expects Azure revenue growth in constant currency to remain roughly stable compared to Q2.

Source: Company reports

Earnings and Other Highlights – Google Cloud

- **Google Cloud segment performance:**
  - Revenues were $8.4 billion in 3Q, up 22%. GCP revenue growth remained strong across geographies, industries and products, although the Q3 y/y growth rate reflects the impact of customer optimization efforts.
  - Google Cloud had operating income of $266 million, and the operating margin was 3%.
  - GCP revenue growth in the third quarter was above the growth rate for Cloud overall, with particular strength in infrastructure, data analytics, and security.

- **Capital expenditure:**
  - Google Expects a modest y/y CapEx increase in FY23. This includes investments in GPUs, proprietary TPUs, and datacenter capacity.
  - Google reported an $8 billion CapEx in 3Q, mainly for its technical infrastructure and AI compute, but the growth was muted by the timing of supplier payments.
  - Google expects to increase its CapEx in 4Q and FY24, as it sees many opportunities in AI across Alphabet.
  - FY24 aggregate CapEx will be above FY23.

- **Generative AI:**
  - There has been a sevenfold increase in active generative AI projects on Vertex AI, with customers such as Highmark Health using AI to create more personalized member materials.
  - More than half of all funded generative AI start-ups are Google Cloud customers.
  - Management remains focused on profitable growth and continues to invest aggressively in AI computing and Datacenter capacity.

- **GPU usage:**
  - Management mentioned that they are seeing strong demand for GPUs, which are being used for tasks such as machine learning and scientific simulations.
  - A3 VMs powered by NVIDIA's H100 GPU are generally available, and Google is winning customers with Cloud TPU v5e, its most cost efficient and versatile accelerator to date.

Source: Company reports

Earnings and Other Highlights – Oracle Cloud

- **Financial Performance:**
  - Cloud revenue (IaaS plus SaaS) were $4.6 billion, up 30% in USD, up 29% in constant currency. Cloud Infrastructure (IaaS) revenue were $1.5 billion, up 66% in USD, up 64% in constant currency.
  - The gross margins for Cloud Services and License Support was 78% with IaaS gross margins improving substantially from last year. Oracle's cloud services and license support revenue driven again by its strategic cloud applications, Autonomous Database and its Gen2 OCI. The company's cloud margins improved due to increased scale and efficiency.

- **Capex:**
  - Oracle expects that FY24 CapEx will be similar to this past year's CapEx. The company remains careful to pace its investments appropriately and in line with booking trends, which is why its gross margins are up in its cloud business.
  - Oracle's capex spending is being used to expand its cloud infrastructure and to develop new AI capabilities.

- **Cloud Regions and Availability Zones:**
  - Oracle has 64 cloud regions live with 44 public cloud. regions around the world and another 6 being built. 12 of these public cloud regions interconnect with Microsoft Azure.
  - Oracle has 9 dedicated regions live and 11 more planned, 9 security regions and 12 EU sovereign regions live with increasing demand for more of each.

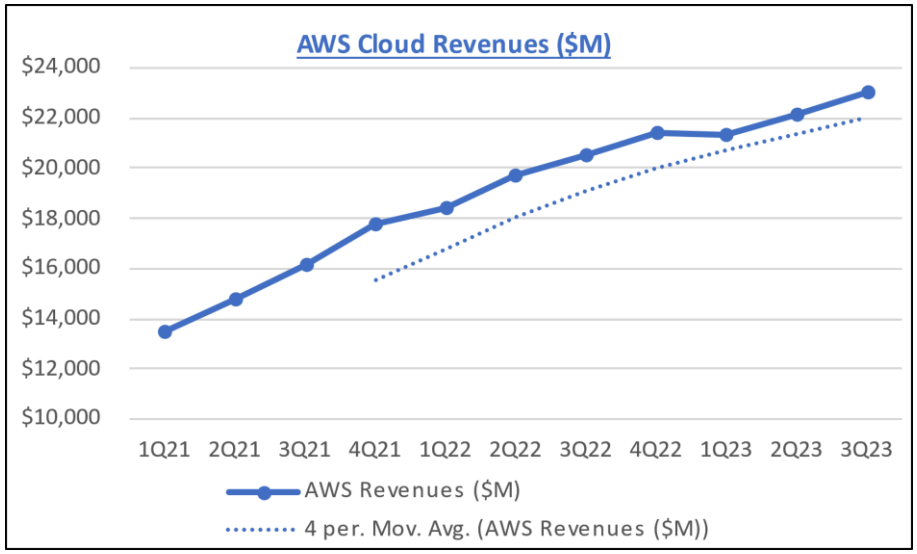
- **GPU:**
  - Oracle’s GPU is a low margin business. In some cases, Oracle’s GPU prices are lower than the cost of other cloud providers, and that this is very profitable for Oracle.

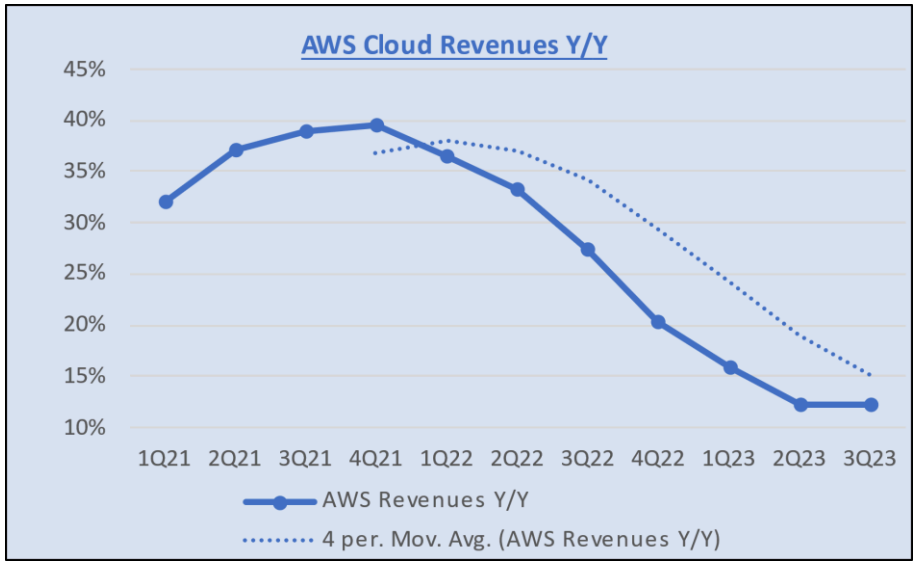
- **Cloud adoption:**
  - Oracle is seeing strong adoption of its cloud services. The company's cloud customer base grew to over 1 million in the fourth quarter of fiscal 2023.
  - The cost advantages, sizing flexibility and deployment optionality of its cloud region continue to make it so compelling in the marketplace to customers.

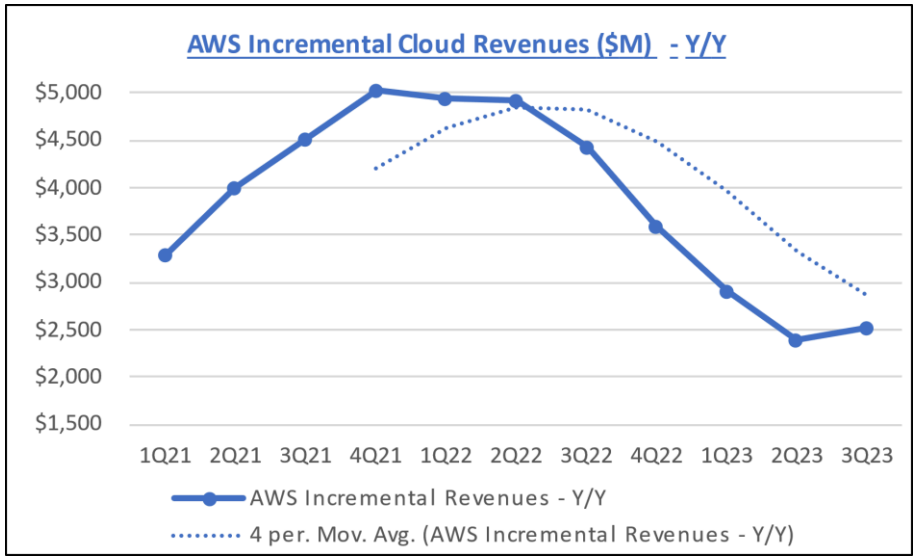
Source: Company reports

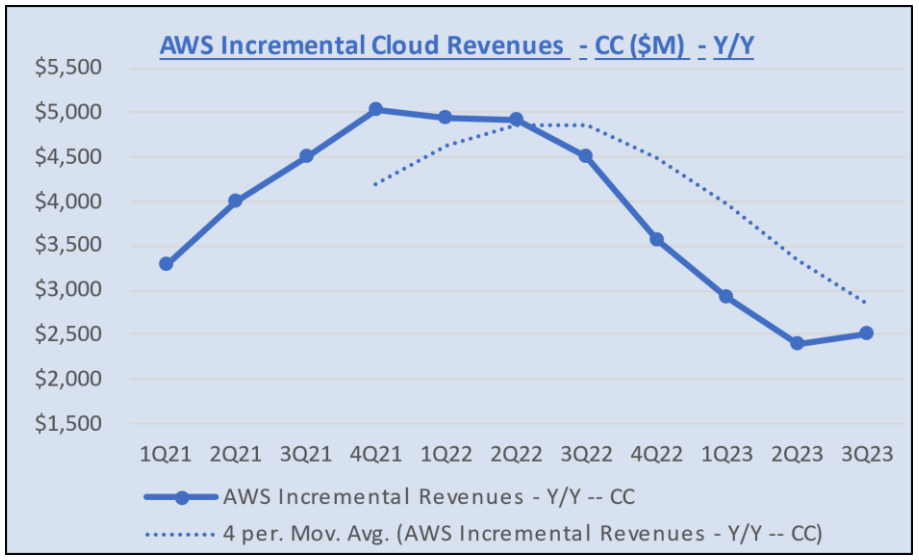
Cloud Revenue Growth – Amazon Web Services (AWS)

AWS segment revenues increased 12.3% y/y to $23.1B in 3Q23. Management indicated that AWS growth is stabilizing as customers started shifting from cost optimization to new workload deployment.





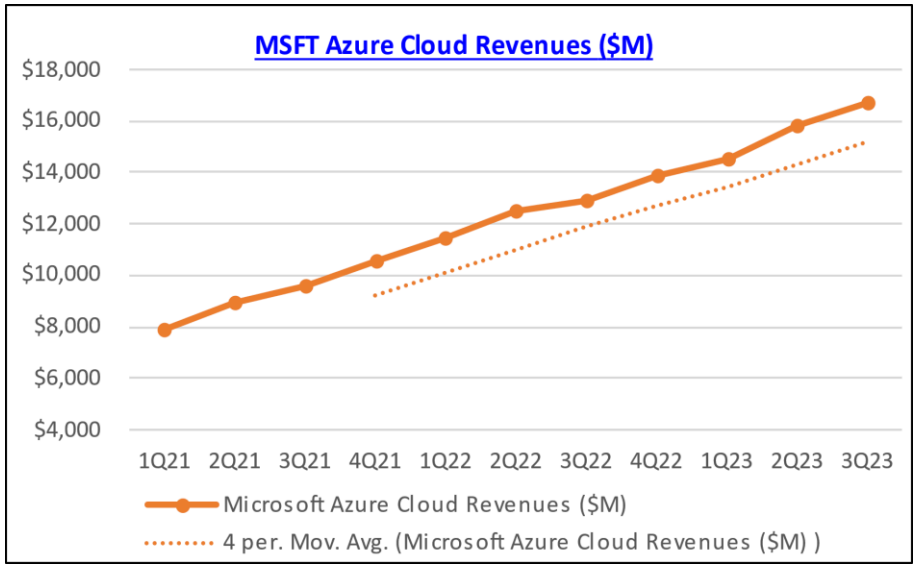


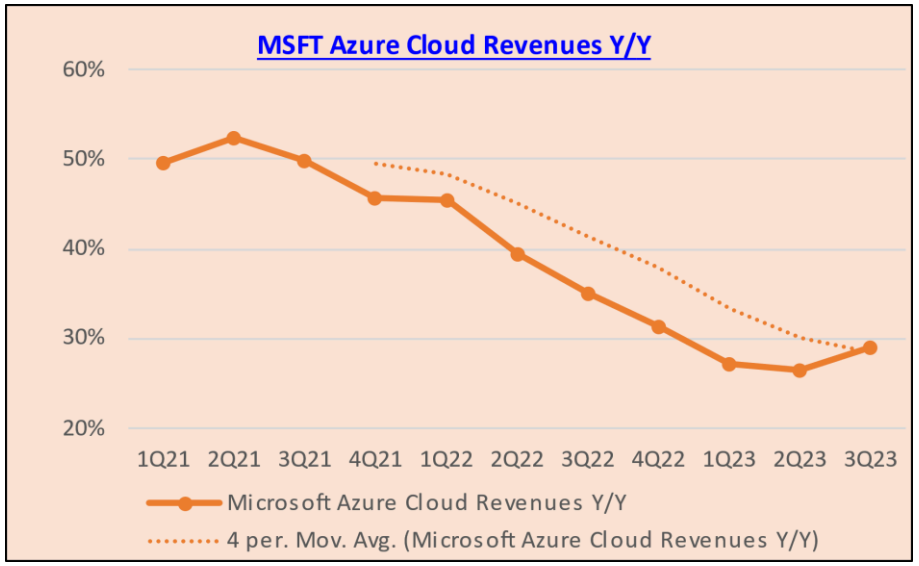


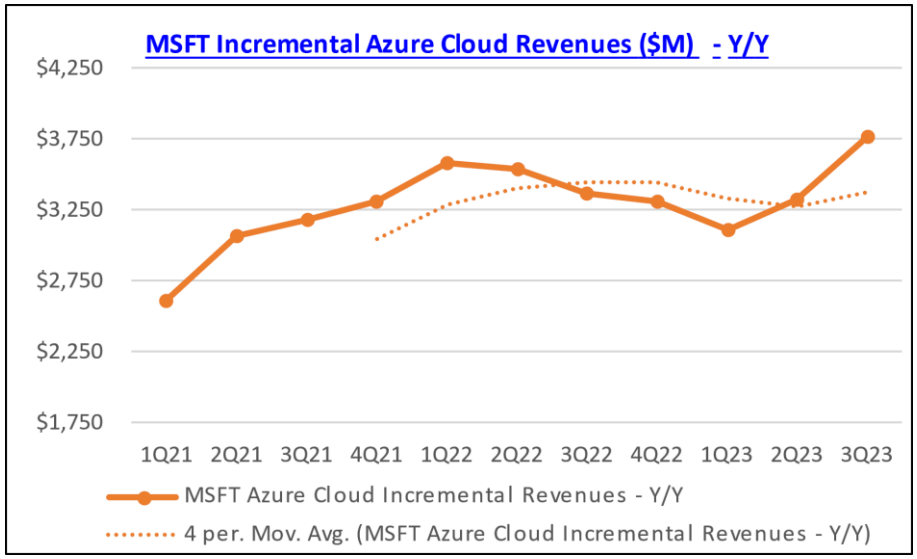
Source: Company reports, RBC Capital Markets

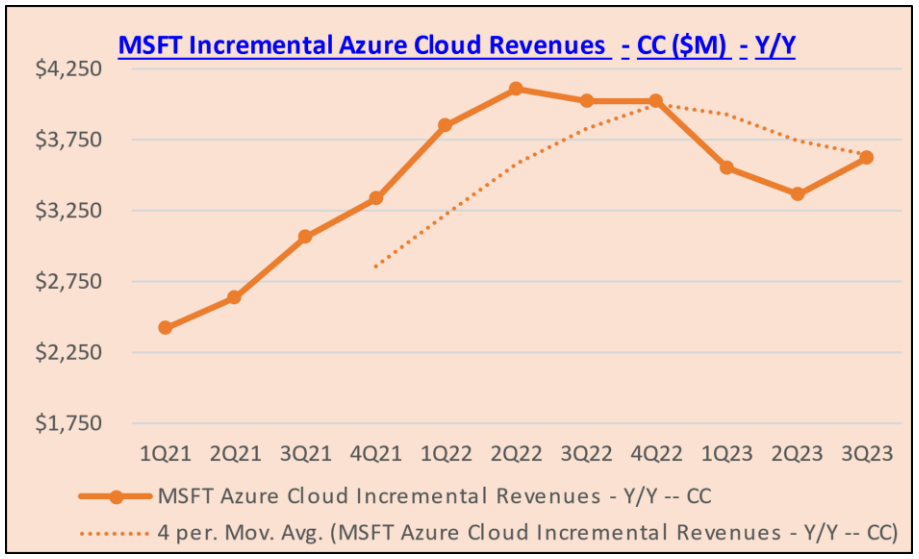
Cloud Revenue Growth – Microsoft Cloud / Azure

Azure & other cloud services revenue grew 29% and 28% Y/Y constant currency in FY1Q24/CY3Q23, with ~3 points increase from AI services. Prior quarters constant currency – F4Q23, 27%; F3Q23, 31%; F2Q23, 38%.





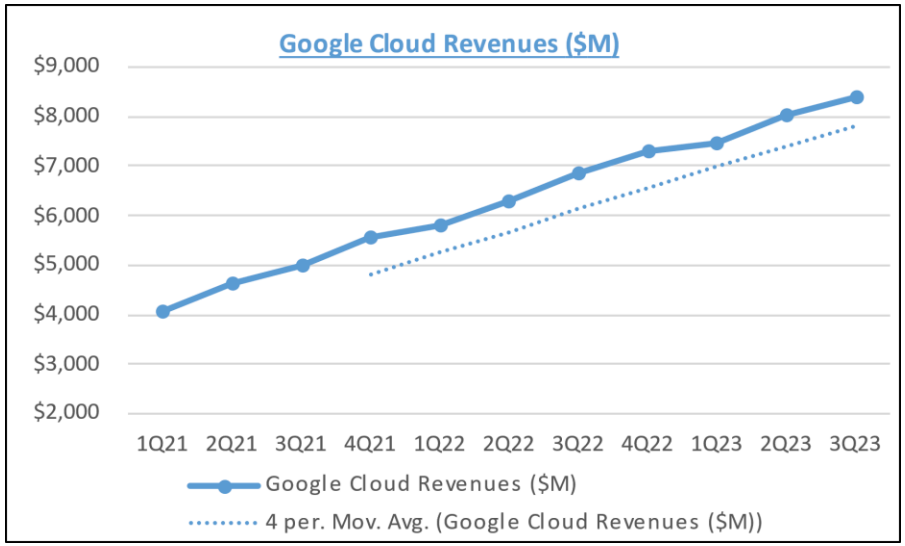


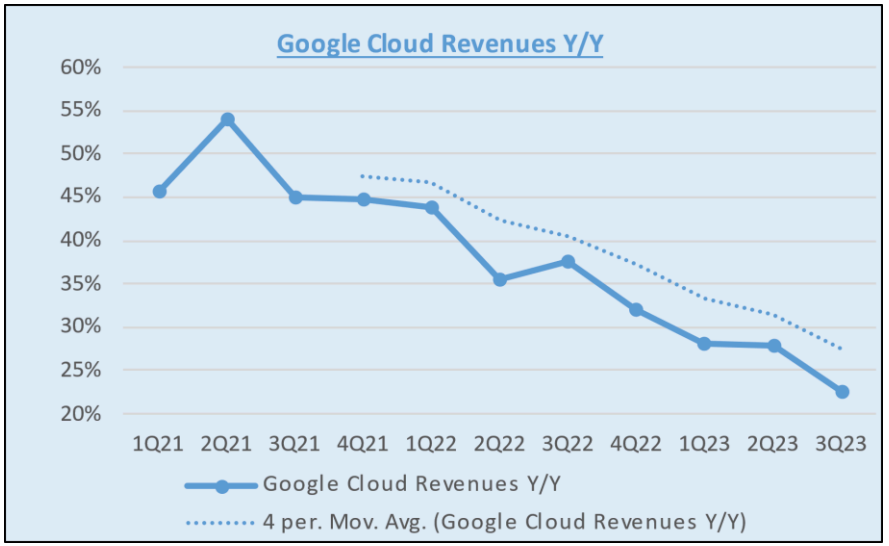


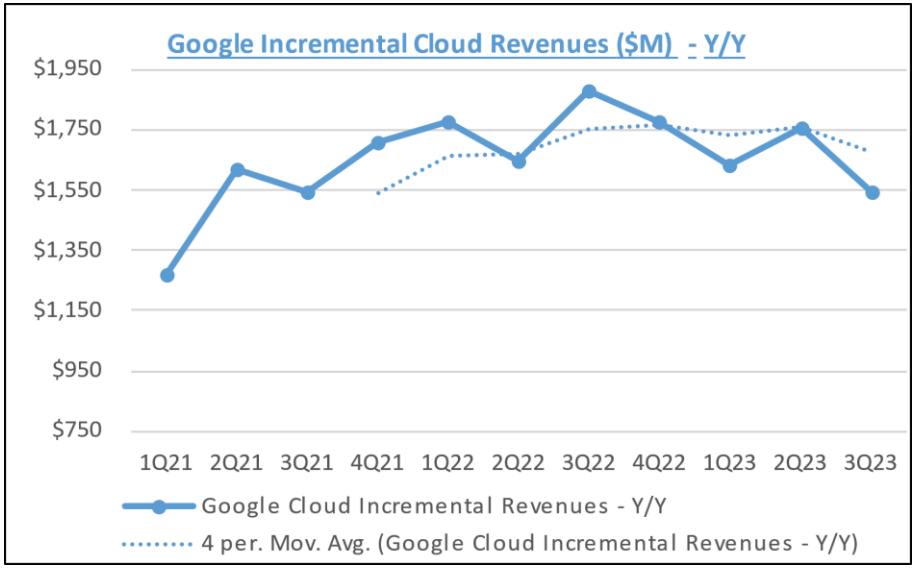
Source: Company reports, RBC Capital Markets

Cloud Revenue Growth – Google Cloud

Google Cloud segment revenues increased 22.5% y/y to $8.4B in 3Q23. Management indicated that GCP revenue growth remained strong across geographies, industries and products, although the Q3 y/y growth rate reflects the impact of customer optimization efforts.



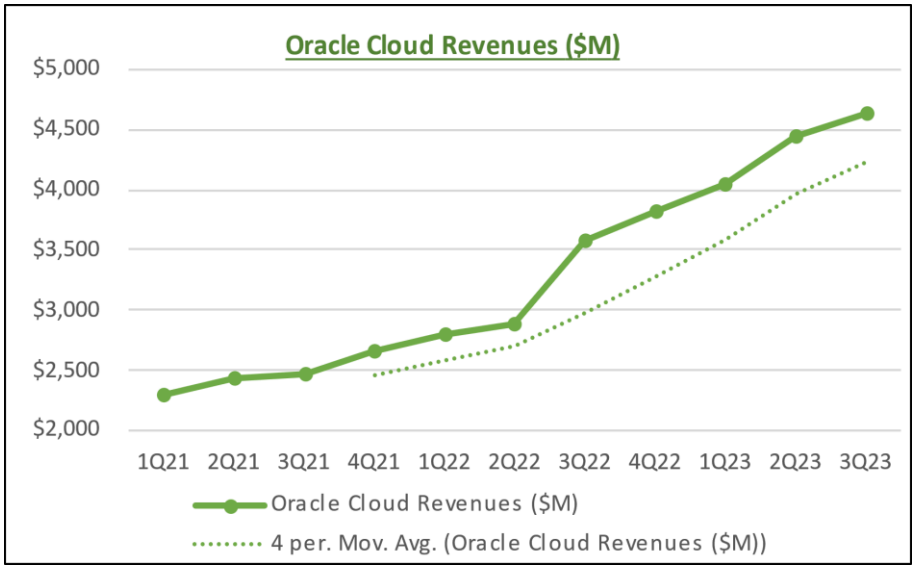


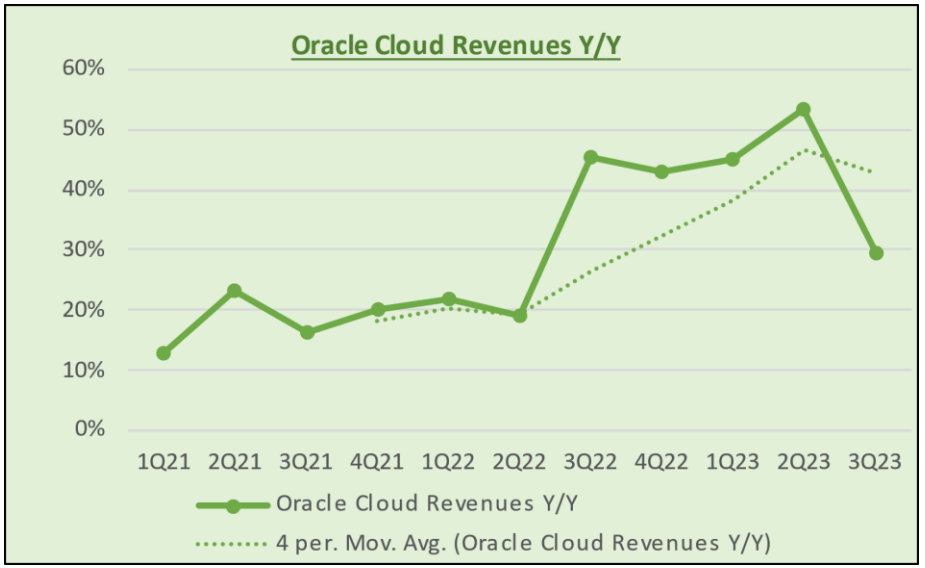


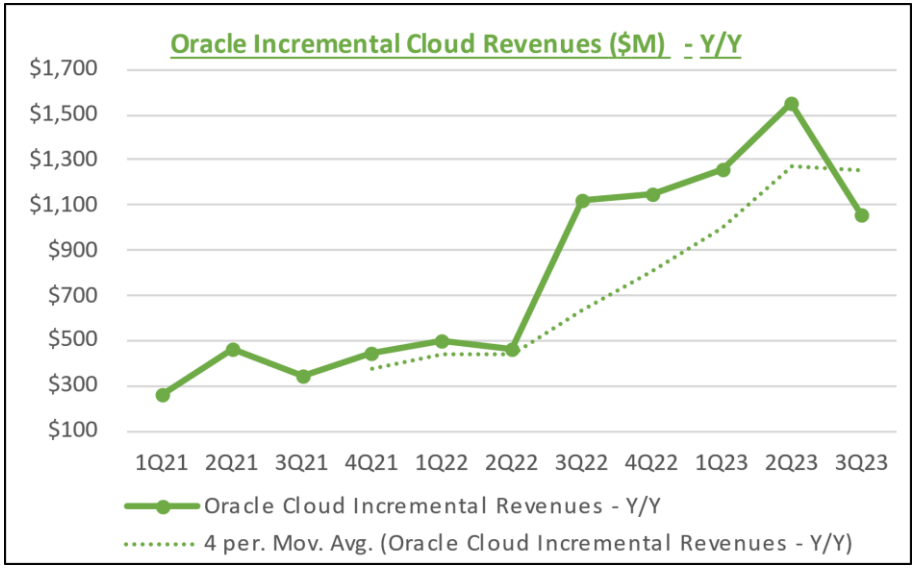
Source: Company reports, RBC Capital Markets

Cloud Revenue Growth – Oracle Cloud

Oracle Cloud revenue (IaaS plus SaaS) increased 30% Y/Y to $4.6 billion for FY1Q24/CY3Q23, ending Aug. 31 2023.





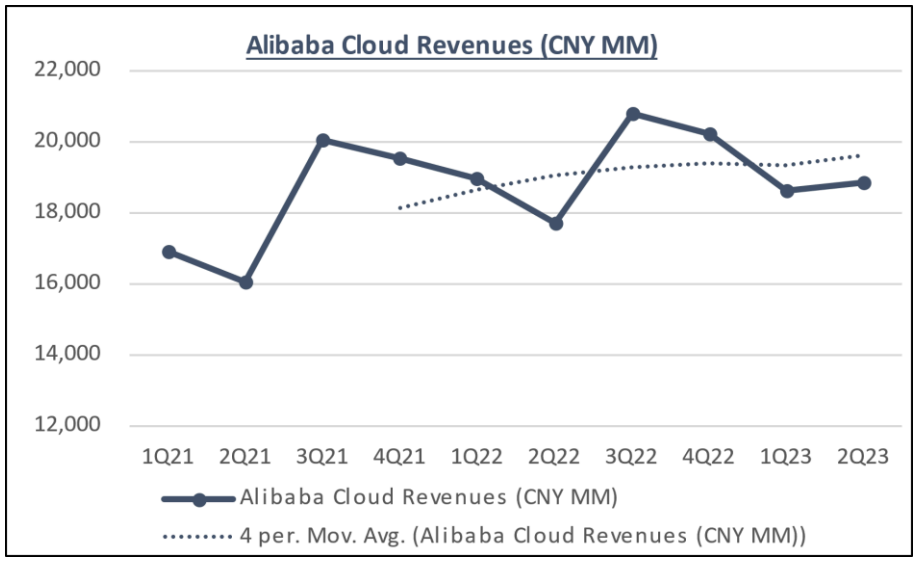


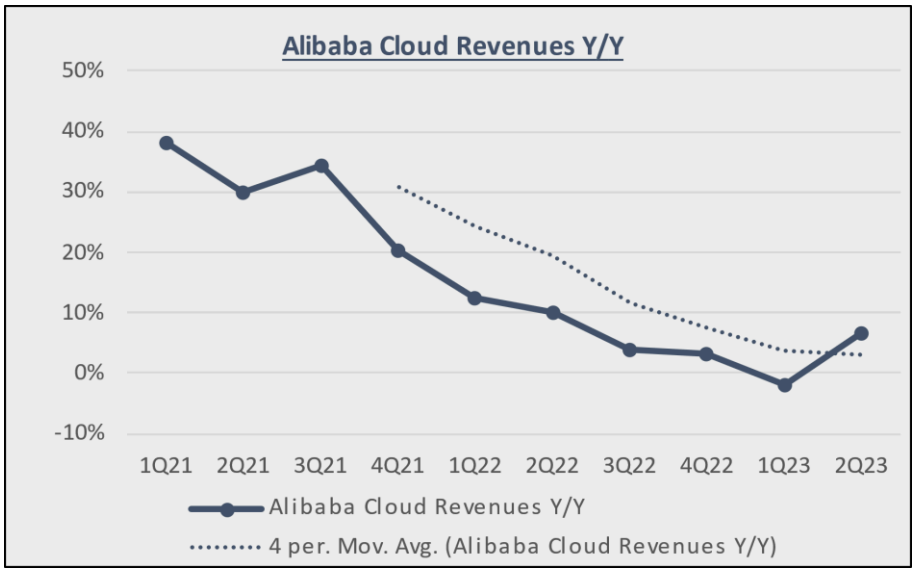
Note: The abrupt increase in 3Q22 was primarily because of the "Cerner" acquisition.

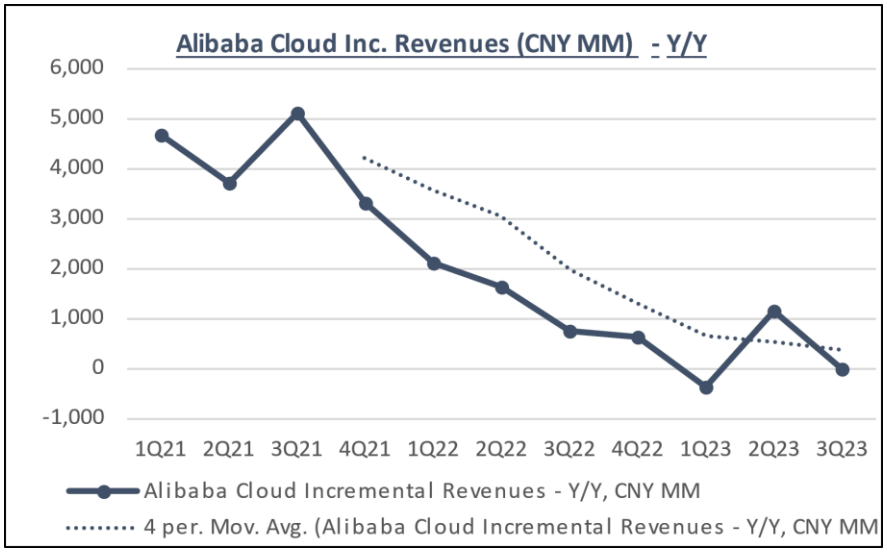
Source: Company reports, RBC Capital Markets

Cloud Revenue Growth – Alibaba Cloud

Alibaba's Cloud revenue grew 4%Y/Y in FY1Q24/CY2Q23, ending June 2023. The growth rate was negatively impacted by the normalization of CDN demand as usage of video streaming, remote working and remote learning came down when offline activities resumed after pandemic measures were lifted. Alibaba will report its FY2Q24/CY3Q23 results on Nov. 16th.







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 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

Section 4

Recent Perspectives on Generative AI From Hyperscalers, Datacenter Operators, Chip Manufacturers, and Other Stakeholders

RBC Capital Markets

Recent AI Related Commentary – Hyperscale & Cloud Companies

Alibaba Cloud

- Alibaba plans to offer Model as a Service (MaaS) on top of its existing IaaS and PaaS infrastructure to provide a stable, secure, high-performance, and cost-efficient computing service for AI model training and related services.
- The company has released a large language pretrained model, Tongyi Qianwen, and plans to launch cloud products and enterprise solutions based on this model.
- Alibaba aims to integrate AI with various businesses within the Alibaba Group, including DingTalk, to offer new AI-based service experiences for users.
- Alibaba plans to continue investing in AI and sees it as a significant opportunity for commercial value.
- The company believes that providing best-in-class services to support new generations of innovation will be key to monetizing this opportunity.
- There is strong demand for model training and related AI services on cloud infrastructure, and the company believes that the growth opportunity driven by AI services has just begun.
- Alibaba has built an open source online community in China for models and related tools and services, which is popular among developers.
- The company has released its own large language model and image model, which have accumulated millions of users.
- Alibaba will continue to upgrade its models and pursue an open source strategy to increase adoption and drive usage of its computing power.

Amazon Web Services (AWS)

- AWS has been a significant player in AI and machine learning, with a history of providing machine learning services, and a large number of customers are utilizing its compute for machine learning training and production.
- The company is at the forefront of generative AI, investing heavily in three key layers: compute for training foundational models and inference, large language models as a service, and actual applications running on top of the models.
- The timing of monetization for generative AI is uncertain, as companies are still in the relatively early stages of exploring and testing the technology.
- AWS is offering services such as Trainium and Inferentia for training and running large-scale generative AI models, and Bedrock for deploying and managing these models in production environments.
- Customers are experimenting with different types of models and sizes to find the right balance of cost and latency for their specific use cases.
- AWS is focusing on democratizing generative AI technology for customers of all sizes and technical abilities.

Recent AI Related Commentary – Hyperscale and Cloud Companies

Google

- There has been a sevenfold increase in active generative AI projects on Vertex AI, with customers such as Highmark Health using AI to create more personalized member materials.
- More than half of all funded generative AI start-ups are Google Cloud customers
- Google is constantly looking for ways to optimize its generative AI models, training costs, and serving costs.
- The company anticipates increased investment in technical infrastructure in the back half of 2023 and continuing growth into 2024, supporting opportunities in AI across Alphabet. This includes investments in GPUs, proprietary TPUs, and datacenter capacity.
- Google is leveraging Generative AI to empower users to boost creativity and productivity. Bard, an experiment in conversational AI, has been rolled out with new features and capabilities available in most of the world and over 40 languages.
- Google Lens capabilities have been added to Bard, enabling users to take images and ask questions, turn them into code, and more. Users have shown great interest in using Bard for coding tasks.
- Google Cloud's investments in AI-optimized infrastructure, large language models, AI platform services, and generative AI offerings are attracting significant customer interest.

Meta

- Meta has invested extensively in AI capacity, enabling it to pursue cutting-edge research and integrate it into its products at an accelerated pace.
- Meta predicts its full-year 2024 capital expenditures between $30 billion and $35 billion, driven mainly by investments in servers, including both non-AI and AI hardware, and Datacenters as it ramps up construction on sites with new architectures.
- Investments in AI, including billions of dollars spent on AI infrastructure, are paying off across ranking and recommendation systems, improving engagement and monetization.
- AI-driven feed recommendations have contributed significantly to increased user engagement (7% increase on Facebook and 6% on Instagram) and advertiser success through Advantage+ shopping campaigns and optimized ad creatives.
- Meta is leveraging AI to move towards using fewer, larger models that enable them to leverage learnings across product surfaces and deploy improvements more quickly, broadly and efficiently.
- Meta has partnered with Microsoft to open source Llama 2, the latest version of their large language model, for both research and commercial use.
- Dell and Meta partner to offer Meta’s Llama 2 models on Dell’s GenAI products and services.
- Meta is focusing heavily on AI development and its potential for future growth. It has achieved significant milestones with its foundation models like Llama 2, which has had over 30 million downloads last month.
- Meta is making progress with sophisticated recommendation AI systems that power various features such as Feeds, Reels, ads, and integrity systems. While generative AI receives more attention, other forms of AI like recommendation AI are equally important and showing rapid improvement.

Recent AI Related Commentary – Hyperscale and Cloud Companies

Microsoft

- Microsoft is helping customers build generative AI applications on top of Azure AI, focusing on maximizing productivity and providing resources for faster adoption of new cloud meters. Generative AI contribution to Azure is measurably improving q/q. Microsoft's capital expenditures, including finance leases, amounted to $11.2 billion, with investments in AI infrastructure included. This represented an acceleration of investment in cloud infrastructure to support the growth of Microsoft Cloud and AI platform demand.. Microsoft has struggled to get enough Nvidia GPUs and has signed a deal with rival CoreWeave to use its facilities to meet demand.
- Microsoft expects AI to contribute significantly to Azure growth, particularly in the areas of machine learning and cognitive services. Microsoft anticipates new workload starts, particularly in the area of AI, to drive future growth.
- Microsoft continues to see more cloud migrations with Azure Arc. The company is meeting customers where they are, helping them run apps across on-prem, edge and multi-cloud environments. Microsoft now has 21,000 Arc customers, up 140% y/y.
- Microsoft announced that it will use Oracle Cloud Infrastructure (OCI) AI infrastructure, along with Microsoft Azure AI infrastructure, for inferencing of AI models to power Microsoft Bing conversational searches. Bing conversational search requires powerful clusters of computing infrastructure that support the evaluation and analysis of search results that are conducted by Bing's inference model.
- Inference models require thousands of compute and storage instances and tens of thousands of GPUs that can operate in parallel as a single supercomputer over a multi-terabit network.
- Microsoft has been leveraging Oracle Interconnect for Microsoft Azure and Azure Kubernetes Service (AKS) to orchestrate OCI Compute.

Oracle

- Oracle's cloud revenue growth was driven by strong demand for its IaaS and SaaS services, including AI-optimized infrastructure and AI platform services.
- Oracle is investing heavily in GPUs and AI, including a partnership with NVIDIA to offer GPU-accelerated AI services on Oracle Cloud Infrastructure (OCI).
- Oracle sees significant potential for generative AI across various industries, including the automotive and pharmaceutical sectors, and notes that its AI development customers have signed contracts worth over $4 billion for AI training capacity in its Generation2 Cloud.
- Oracle's GPU business is low-margin, but it offers highly profitable training services at lower costs than other hyperscalers.
- Oracle expects FY24 CapEx to be similar to the previous year, with investments focused on expanding cloud infrastructure and developing new AI capabilities.
- Oracle reports strong adoption of its cloud services, with over 1 million customers in the fourth quarter of fiscal 2023. Its cloud offerings are attractive due to cost advantages, flexible sizing, and deployment options.
- Oracles' OCI Superclusters can scale up to 4,096 OCI Compute Bare Metal instances with 32,768 A100 GPUs or 16,384 H100 GPUs, and petabytes of high-performance clustered file system storage to efficiently process massively parallel applications.

Recent AI Related Commentary – Datacenters

DigitalBridge:

- Gen AI is driving tremendous demand for datacenters, with pipeline increasing by over 400% to 500%.
- Datacenter workload increases related to AI applications are contributing to the strong performance of portfolio companies (including DataBank, Scala, Switch, and Vantage).
- The company's focus is on supporting AI infrastructure through datacenters and connectivity.
- AI workloads have high power requirements, and the company is strategizing to meet those demands across its datacenter locations.

Digital Realty Trust

- AI has become a significant part of both hyperscale and smaller enterprise customer requirements. The company is witnessing a growing trend of AI-related deployments and increased demand from customers seeking AI capabilities.
- Certain customers are seeking large contiguous capacity for high-density AI and trading models, driving the need for suitable power capacities.
- As AI applications expand, the demand for both training and inferencing services is increasing. Datacenters play a crucial role in supporting the infrastructure requirements for AI training and inferencing workloads.
- DLR announced the launch of its first Nvidia DGX H100-Ready Datacenter in Osaka, Japan.
- DLR rolled out its new high-density colo offering across 28 global metros to support high-performance compute infrastructure, addressing data and AI-related growth challenges.
- In 3Q23, Digital Realty signed a 32 MW lease with CoreWeave in Portland, where CoreWeave plans to deploy 32,000 Nvidia H100 GPUs.

Equinix

- The company has won deals with AI-as-a-service providers, positioning itself to support inference and interconnection to the cloud for AI services.
- Equinix plans to pursue three key vectors for capturing high-value opportunities across the AI value chain: a.) Magnetic AI service provider deployments to support on-ramps, inference nodes, and smaller-scale training needs. B). Expanding the xScale portfolio, including in North America, to pursue strategic large-scale AI training deployments with hyperscalers and other key ecosystem players. c.) Positioning Platform Equinix as the place where private AI happens, allowing customers to place compute resources in proximity to data and seamlessly leverage public cloud capabilities while maintaining control of high-value proprietary data.
- Equinix is already successfully deploying liquid cooling solutions across a range of deployment sizes and densities.

Recent AI Related Commentary – Semiconductors

AMD

- AMD has been investing in its supply chain to meet demand for AI chips and GPUs.
- AMD introduced the MI300 accelerator, for which it anticipates broad customer adoption and a wide range of workloads for the MI300 product, including training and inference workloads. MI300 would be the fastest product to ramp to $1 billion in sales in AMD history.
- The demand for generative AI solutions is high, and the market is expected to grow at a significant rate, potentially reaching a $150 billion market by 2027.
- AMD is executing on a multiyear Ryzen AI road map to deliver leadership compute capabilities built on top of Microsoft's Windows software ecosystem to enable the new generation of AI PCs that will fundamentally redefine the computing experience over the coming years.
- AMD strengthened its AI software capabilities with the strategic acquisitions of Mipsology and Nod.ai. AI software ecosystem expanded, including integration with PyTorch and TensorFlow, and support for Hugging Face models

Intel

- Intel is seeing strong demand across its business segments, particularly in the areas of AI, ML, and cloud computing.
- Intel is investing in GPUs for AI and high-performance computing applications.
- MLPerf benchmark data shows Gaudi2 as a competitive alternative to NVIDIA GPUs.
- Intel is actively engaging with hyperscale cloud providers and next-generation cloud companies with Gaudi offerings.
- Gaudi instances already available on AWS and attracting interest from various AI companies.
- Intel’s 4th Gen Xeon Scalable Processor and Gaudi2 chips are providing strong AI acceleration capabilities for training and inferencing workloads.
- Intel’s AI-enhanced Xeons are primed for model inferencing, enabling seamless infusion of AI into existing workloads. This was visible in the most recent quarter with over 1/3 of 4th Gen Xeon shipments directly related to AI applications. Intel predicts inferencing to be the discussion topic for the industry as we go into '24. That will be done at scale and much of that is going to be done on Xeons.
- As the world moves towards more AI-integrated applications, there's a market shift towards local inferencing. It's a nod to both the necessity of data privacy and an answer to cloud-based inference cost.

Recent AI Related Commentary – Semiconductors

Marvell Technologies

- Marvell is experiencing strong growth in its AI chip business. Marvell's AI initiatives encompass both training and inferencing solutions.
- The company is investing in technologies like CXL to meet the increasing demand for high-performance training and inferencing capabilities.
- Strong traction is observed for AI-specific ASICs (Application-Specific Integrated Circuits) designed for compute offload acceleration.
- Marvell is seeing a rapid shift in its cloud customers' plans as spending on AI infrastructure is becoming a much bigger portion of their CapEx.
- Marvell now expects revenue from AI to exit this year at over a $200 million quarterly revenue run rate or $800 million annualized.
- Increased demand for connectivity between regional Datacenters to support inferencing deployed across multiple locations.

NVIDIA

- Enterprises are racing to deploy generative AI, driving strong consumption of NVIDIA-powered instances in the cloud as well as demand for on-premise infrastructure.
- Cloud service providers are driving strong demand for HGX systems, which represent nearly two decades of full-stack innovation across silicon, systems, interconnects, networking, software, and algorithms.
- Instances powered by NVIDIA H100 Tensor Core GPUs are now generally available at AWS, Microsoft Azure, and several GPU cloud providers, with more on the way soon.
- The outlook for the 3QFY2024 remains strong, with demand visibility extending into next year and continued ramping of supply over the next several quarters.
- There has been no material impact on financial results from recent reports of potential export restrictions on Datacenter GPUs to China, but long-term restrictions could result in a “permanent loss of opportunity” for the US industry.
- Nvidia is prepping three new GPUs for AI and high-performance computing (HPC) applications tailored for Chinese market and to comply with U.S. export requirements, according to media reports. The new units will be based on the Ada Lovelace and Hopper architectures.
- NVIDIA provided that NVIDIA’s AI platform set new standards for AI training and high-performance computing in the latest MLPerf industry benchmarks. A notable achievement is the performance of NVIDIA Eos, an AI supercomputer powered by 10,752 NVIDIA H100 Tensor Core GPUs and NVIDIA Quantum-2 InfiniBand networking. It completed a training benchmark based on a GPT-3 model with 175 billion parameters trained on one billion tokens in just 3.9 minutes. This is a nearly 3x improvement from the previous record of 10.9 minutes set by NVIDIA less than six months ago.

Recent AI Related Commentary – Datacenter 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.
- While AI is not new, its increasing adoption is expected to be an accelerator of growth in the Datacenter market. However, the broader trend of more data and insights requiring more Datacenters is also driving growth.

nVent

- nVent is focusing on developing products in the AI space as part of its Data Solutions business. It invests in R&D to build a more standardized portfolio that allows them to scale AI solutions through distribution channels.
- The company is seeing increased demand for its liquid cooling solutions, particularly in the hyperscale and enterprise markets, and is doubling its capacity to meet this demand. The company believes that liquid cooling is a growing market with a long-term future, driven by the use of more powerful chips and the need for energy efficiency.

Schneider Electric

- The company reported a record third quarter with sales of EUR 8.8 billion, driven by strong trends in electrification and digitization, particularly in Energy Management.
- Energy Management grew by 13% organically, with significant contributions from North America. Demand was strong in Datacenter and infrastructure sectors across all regions.
- The company continues to see strong demand across most of its portfolio, driven by strong secular trends, particularly in its Systems business in segments like Datacenter and utilities.
- Datacenter and networks remained strong, with AI-related orders now being booked and sales and distributed IT returning towards growth after signs of stabilization in the H1. Datacenter and Networks represented ~19% of total group orders based on 2022 data.
- From 2017 to 2022, the company experienced double-digit sales CAGRs in Datacenters, primarily driven by systems software and services, and mid-single digit sales CAGRs in the distributed IT segment, led by products including Transformers, switchgears, UPS, PDUs, Chillers, etc. The management anticipates strong future growth due to the increasing need for electrical content and a technological shift from hyperscalers to colocations and edge.
- The Sustainability business continued to perform strongly with strong demand from customers for decarbonization and energy efficiency.

Recent AI Related Commentary – Datacenter Supply Chain

Super Micro

- Management expects up to 20% of its datacenter deployments to move to liquid cooling. The company's product lineup includes NVIDIA and AMD products, and management expects to meet demand with sufficient supply capacity.
- Lead times for high-end GPUs have improved compared to 90 days ago, but the company is still working to optimize inventory control and product flow-in.
- The company increased its revenue guidance for fiscal year 2024 from $9.5-$10.5 billion to $10-$11 billion.
- The company faced supply challenges for AI GPU and other components, but delivered total solutions and large compute clusters for generative AI workloads, which had high demand and backorders.
- The company also saw growth in AI platforms for rack-scale, liquid-cooling, inferencing, and edge products, especially those based on NVIDIA HGX-H100 and Grace Hopper Superchip.
- The company reported that AI, GPU and rack-scale solutions accounted for more than half of its total revenues this quarter and expected similar performance in the next quarter.

Vertiv

- Macro Environment: The datacenter end market remains strong, with cloud hyperscale and colocation continuing to lead the growth. There is accelerating demand and a strong overall market is expected for the foreseeable future. Different customers are at various points in their demand cycle, but all foresee strong future demand. Enterprise outlook is positive, balanced against some macro concerns. Telecom sector is weak due to delayed investments and a lull in the market after strong 5G investment. The softness in telecom is not new and is expected to continue into 2024.
- Regional Performance: APAC is soft due to a slow Chinese market, offset by encouraging signs in India and rest of Asia. No sharp recovery is expected in China until late 2024. EMEA region was flat in the third quarter, with a significant sequential increase expected in the fourth quarter. Full year organic sales growth is expected to be in the upper single- to low double digits. Americas region continues strong growth with organic net sales up 40%, including 28% from volume and 12% from pricing.
- Artificial Intelligence / Liquid Cooling / GPU: The rise of AI and Gen AI is driving demand for high-density computing and datacenter solutions. AI-related datacenters require specialized cooling solutions, such as liquid cooling and direct-to-chip cooling, to handle the higher power densities. AI activity in the order book is not binary. Different players will have different approaches to AI, and it is expected to show up in a more pronounced manner in 2024. AI is believed to be pervasive and this is just the beginning of a multiyear cycle.
- AI-related pipeline: Retrofit is an opportunity, with conversations about retrofit being encouraging. Participation in both new build and retrofit conversations is active.

Wesco (WCC):

- The company is optimistic about the outlook for Communications & Security Solutions (CSS) business, given the strong fundamentals and the potential acceleration provided by AI and other emerging technologies.
- The acquisition of Rahi Systems resulted in substantial growth in sales to hyperscale datacenter customers.

Section 5

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 data center 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 data centers, 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 large language models; 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	•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	•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	•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	•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	•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	•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

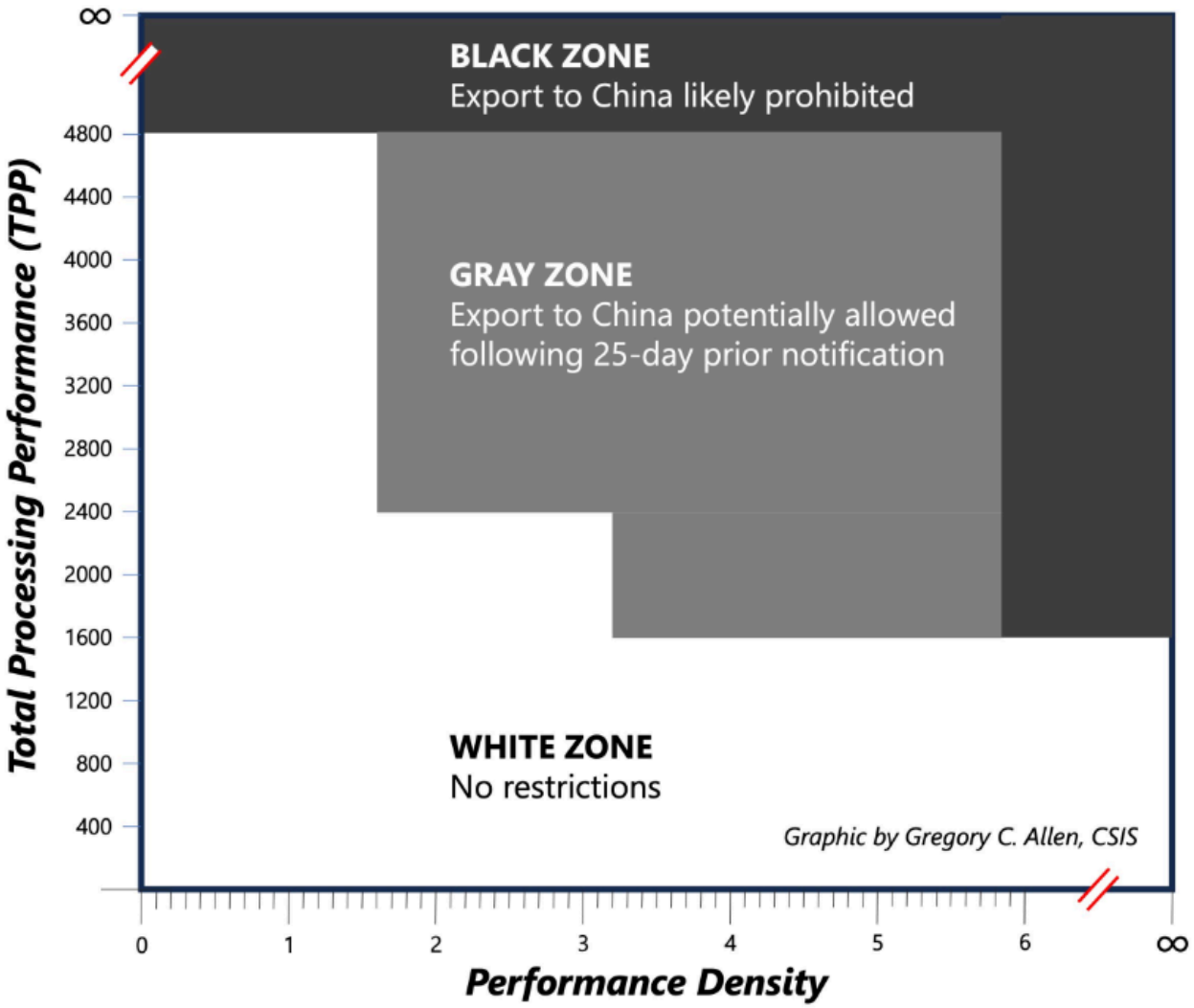
GPU Specifications

	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

CSIS | CENTER FOR STRATEGIC & INTERNATIONAL STUDIES | WADHWANI CENTER FOR AI AND ADVANCED TECHNOLOGIES

Source: CSIS and Semianalysis.com

Illustrative GPU Pricing

H100

Cloud	GPU Type	GPU Arch	GPUs	GPU RAM (GB)	vCPUs	RAM (GB)	On-demand	Per-GPU
Lambda	H100 (80 GB)	Hopper	1	80	26	200	$1.99	$1.99
Lambda	H100 (80 GB)	Hopper	8	640	220	16764	$20.72	$2.59
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

A100

Cloud	GPU Type	GPU Arch	GPUs	GPU RAM (GB)	vCPUs	RAM (GB)	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	$8.80	$1.10
Lambda	A100 (80 GB)	Ampere	8	640	240	1800	$12.00	$1.50
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	$9.84	$2.46
CoreWeave	A100 80 GB	Ampere	4	320	32	256	$10.44	$2.61
Vultr	A100 80 GB	Ampere	8	640	96	960	$20.83	$2.60

Source: Company reports

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 avaialble
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

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. • The executive team includes Wes Cummins as CEO and Chairman, Jason Zhang as co-founder, and David Rench as CFO. Applied Digital offers AI Cloud services through Sai Computing, initially provided from its 9MW HPC Jamestown facility. • The company has three business segments: next-generation datacenter colocation services, AI GPU cloud services, and Blockchain datacenters.
Taiga Cloud (Northern Data Group)	• Taiga Cloud, Europe's largest Generative AI Cloud Service Provider, is part of Northern Data Group, a B2B technology company that provides High Performance Computing (HPC) solutions. • The company has over 19,000 NVIDIA H100, A100, and RTX A6000 GPUs in total, with over 512 GPUs connected into pods of 512. • Taiga Cloud offers flexible, secure access to the latest GPU compute power, providing high-speed, low-latency, clean-energy compute power for organizations to create, train, and deploy new Generative AI solutions.
CoreWeave	• CoreWeave, founded in 2017, is a specialized cloud provider that offers GPU-accelerated compute resources on demand. • The company, which claims to have 45,000 GPUs, is building the next generation public cloud. Its platform accelerates workflows with a cloud-based production pipeline and provides access to compute resources that match the complexity of models. • CoreWeave operates Datacenters in three US regions: US East (Weehawken, NJ), US Central (Chicago, IL), and US West (Las Vegas, NV). NVIDIA is one of its investors, and the company has received significant backing from DigitalBridge Group and Magnetar Capital.
Lambda Labs	• Lambda Labs is an AI infrastructure company based in San Francisco, California, 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 $112.2M in funding over six rounds.

Source: Company reports, RBC Capital Markets

GPUaaS - Company Profiles

Company	Description
Jarvis Labs	• India-based Jarvis Labs, established in 2019, offers quick and easy training of deep learning models on GPU compute instances. • Recognized for its user-friendly setup, it serves over 10,000 AI practitioners and offers quick operations.
Omniva	• A new startup, Omniva, plans to build AI-focused GPU-filled cloud Datacenters in the Middle East and Europe. • The company, still in stealth, is led by Sean Boyle, former AWS CFO until 2020; Kushagra Vaid, former Microsoft VP and distinguished engineer until 2021; and T.S. Khurana, Meta Platforms' vice president of infrastructure until recently. • The Kuwaiti royal family has provided backing for the venture.
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. • The company uses a distributed modular network across Montana, North Dakota, and Colorado to build a 200MW cloud network. • The company's key management personnel include Chase Lochmiller, Cully Cavness, Tara Green, and Timothy Loos. CrusoeCloud is a cloud computing platform optimized for energy-intensive HPC workloads, offering the cleanest and lowest-cost GPU cloud computing solution.
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.
Firmus (Australian immersion cooling specialist) / STT GDC Sustainable Metal Cloud (SMC)	• Firmus, a Singapore-based datacenter provider, has invested in a global venture with firm STT GDC. The venture, Sustainable Metal Cloud (SMC), offers a GPU-centric Infrastructure as a Service (IaaS) for deep learning AI and visual computing workloads. • The SMC uses Firmus' proprietary, scaled, immersion-cooled platform, the 'HyperCube', hosted within global STT GDC locations. This setup delivers sustainable, scalable, high-performance, and cost-effective AI Factories. • The platform can host up to 130kW per 42RU and has been designed with industry leaders like NVIDIA. The partnership aims to reduce power usage, CO2 emissions, and petaflops per watt for AI workloads.

Source: Company reports, RBC Capital Markets

GPUaaS - Company Profiles

Company	Description
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.
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. • Vultr's main competitors include Linode, Kamatera, and Enzu. The company offers GPU options for high-performance computing, 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.
Anthropic	• Anthropic, an AI safety and research company, was founded in 2021 by former OpenAI staffers. • Google invested $300 million in the company in late 2022, gaining a 10% stake and an exclusive cloud contract. • Amazon announced in September 2023 that it would invest up to $4 billion in Anthropic, becoming its primary cloud provider. Anthropic will use AWS as the primary cloud provider for mission-critical workloads, including safety research and future foundation model development. • The company will use Amazon's custom Trainium and Inferentia chips to train and deploy its foundation models. • The company's GPU capabilities are powered by Google Cloud's TPU and GPU clusters and AWS's Trainium and Inferentia chips. Anthropic's GPU products and services include Claude and Claude Instant, foundation models for conversational and text processing tasks.
Vast.ai	• Vast.ai is a cloud computing service based in Los Angeles, aiming to lower the cost of compute-intensive workloads. • The company, led by CEO Renen Hallak, competes with services like DedicatedCore, Amazon Web Services, Elasticsearch, DigitalOcean, Website Free Host, Golem, and Cloudalize. • Vast.ai offers GPU rentals at competitive prices, allowing users to become cloud compute providers and hardware owners to maximize returns on their investments.

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 & Training.
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%.
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. • They offer GPU-as-a-Service (GPUaaS) through the Hyperstack platform and aim to become the world's top supplier of GPUaaS solutions. The company's headquarters are in Salisbury House, London Wall, but sources suggest they are based in Marbella, Spain. • They raised $14 million in funding and partnered with DARMA Capital and Moore and Moore Investments Group to finance their AI Supercloud project. The company has built one of the largest GPU fleets 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. • They have committed $576 million in hardware orders as part of their $1 billion investment.
Outscale backed by Dassault Systems	• Outscale, a subsidiary of Dassault Systèmes, offers enterprise-class cloud services, including TINA OS, which automates cloud resources, allowing organizations to easily deploy, manage, and increase their cloud platform. • In June 2019, Dassault Systèmes acquired a majority stake in OutScale, a leading European provider of cloud-based high-performance computing (HPC) and artificial intelligence (AI) services. • The acquisition aimed to expand Dassault's presence in the cloud computing market and enhance its digital transformation solutions. • With a stake of approximately 80%, OutScale provides computing resources to high-standard customers worldwide and in the best Datacenters in Europe, North America, and Asia.

Source: Company reports, RBC Capital Markets

GPUaaS - Company Profiles

Company	Description
Scaleway backed by 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 minority stake in Scaleway, which has increased its stake to around 30%. • 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.
Clever Cloud	• Clever Cloud is a European PaaS provider offering tailored products and managed services for hosting, deploying, and maintaining applications at a controlled cost. • 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.
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 is led by Vishnu Subramanian, the Founder & CEO. The platform offers various GPU-powered instances for AI model training and deployments.
Denvr Dataworks	• Denvr Dataworks, a private company founded in 2020, offers bespoke cloud services for artificial intelligence, machine learning, deep learning, and GPU accelerated data science applications. • They offer up to 4 NVIDIA A100s in a reserved node for up to 14 days for free. Their technology supports hybrid cloud scenarios and integration with DevOps pipelines and APIs in private-cloud, hybrid-cloud, and edge computing settings. • Denvr provides first-of-its-kind 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.

Source: Company reports, RBC Capital Markets

Recent Developments Related to AI and GPU based Private Companies

Company	Date	Development
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 large language model (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	Oct. 2023	Lambda Labs is nearing a $300M funding round led byBaire Thomas Tull, founder of Legendary Entertainment. SoftBank is also in discussions to invest, but Nvidia is no longer expected to participate. The funding would value the company at $1.5B. The company forecasts revenues of $250m for 2023 and nearly $600m for 2024, amid a surge in demand from the generative AI boom. Along with being cost-competitive to hyperscalers, Lambda was also given preferential access to Nvidia's much sought-after high-end GPUs.
Lambda Labs	Jul. 2023	Nvidia is in talks with Lambda Labs to take an equity stake, which would be part of a $300M raise that could value the company at more than $1B.
Lambda Labs	Mar. 2023	Lambda raised $44M 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 they have seen extreme growth in their cloud product over the past couple of years.
Crusoe Energy	Oct. 2023	Crusoe announced significant expansion of cloud business with new capacity and $200M 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 $350M Series C equity offering and also secured credit facilities expandable up to $155M with SVB Capital, Sparkfund, and Generate Capital to provide additional debt capital for energy systems related to flare mitigation.
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 artificial intelligence (AI) models at scale — across any region on the globe. The GPU Stack supports instant provisioning of the full array of NVIDIA GPUs, while the new Vultr Container Registry makes AI pre-trained NVIDIA NGC models globally available for on-demand provisioning, development, training, tuning and inference.
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

Recent Developments Related to AI and GPU based Private Companies

Company	Date	Development
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 $50M in Series B funding to expand its global infrastructure and develop new cloud-based services.
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 $460M over 36 months.
Applied Digital	May. 2023	Applied Digital Corporation secured its first major AI customer, Character.AI., with an agreement worth up to $180M 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.
Taiga Cloud	Nov. 2023	Northern Data, entered into a loan agreement with a company of the Tether Group, under which it secured a EUR 575M debt financing facility. The facility is unsecured, at standard market conditions and has a term until 1 January 2030. It will enable Northern Data Group to make further investments across its three business lines Taiga Cloud, Ardent Datacenters and Peak Mining. The focus of these investments will be on the acquisition of additional sophisticated hardware allowing Northern Data Group's Taiga Cloud business to further expand its offering as a Generative Artificial Intelligence Cloud Service Provider in Europe.
Taiga Cloud	Oct. 2023	Taiga Cloud, Europe's first and largest Generative AI Cloud Service Provider and part of Northern Data Group, has entered into a strategic European partnership with GIGABYTE, an industry innovator and leader in the enterprise computing market. The partnership positions Taiga Cloud to meet the surging demand for compute power for Generative AI applications. GIGABYTE will supply Taiga Cloud with 20 NVIDIA H100 GPU pods, worth €400M, with customer access beginning late Q4 2023
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
Anthropic	Oct. 2023	Google agreed to invest up to $2B in the artificial intelligence company Anthropic. Google is already an investor in Anthropic.
Anthropic	Oct. 2023	Media reports indicated that Anthropic's Claude 2 LLM was available to use for free in 95 countries. One of the good things about Claude 2 when compared to ChatGPT is the ability to upload files. Anthropic had launched Claude 2 back in July to users in the US and the UK.
Anthropic	Sep. 2023	Amazon has announced a $4B investment in Anthropic, with a $1.25B note that can convert to equity and a $2.75B second note expiring in 1Q24. Anthropic will use AWS as the primary cloud provider for mission-critical workloads, safety research, and future foundation model development.
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. This cost-efficient and scalable platform offers direct-to-compute GPU cloud access, allowing businesses to embrace AI, high-performance computing, and graphics without regulatory risks. Hyperstack participates in the NVIDIA Inception program.
Paperspace (acquired by DigitalOcean)	Nov. 2023	Paperspace is said to be growing in the triple-digit range and brings 36 employees to DigitalOcean. Notably, it has 'elite' status with NVIDIA, which may help it access inventory.
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	Feb. 2023	There were rumors circulating online that Google was acquiring JarvisLabs.ai. However, neither party has confirmed the acquisition, and it remains unverified.
Jarvis Labs	Jan. 2023	The company announced integration with popular project management tool Asana, allowing teams to create and assign tasks within the chatbot itself.
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.

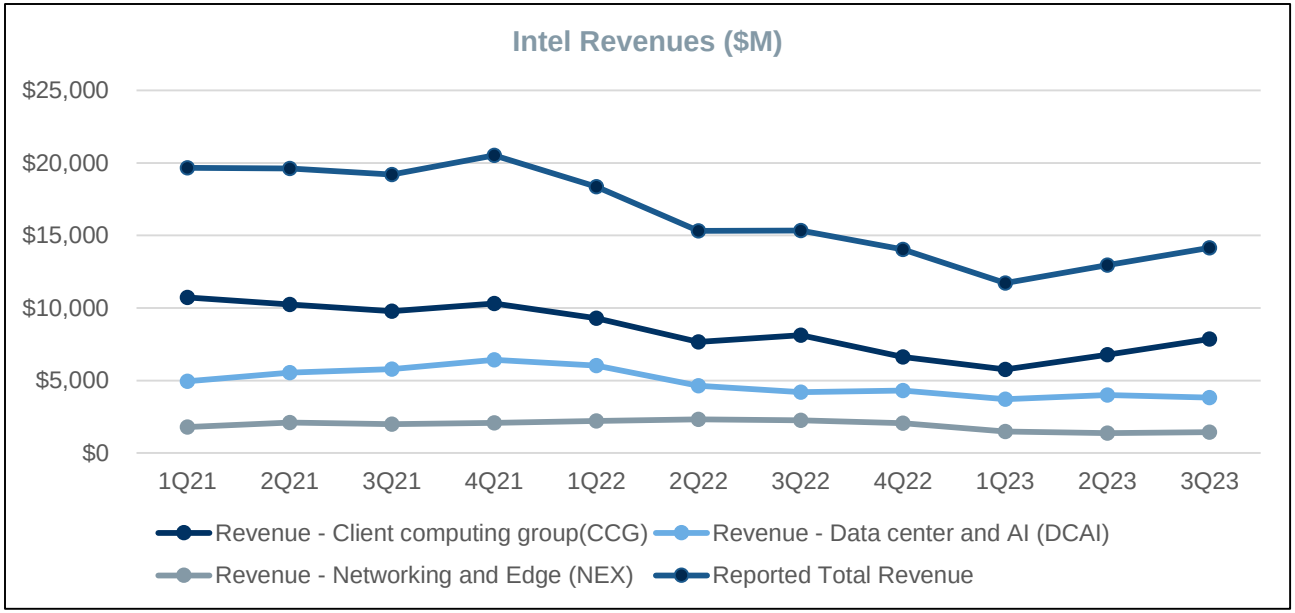
Source: Company reports, RBC Capital Markets

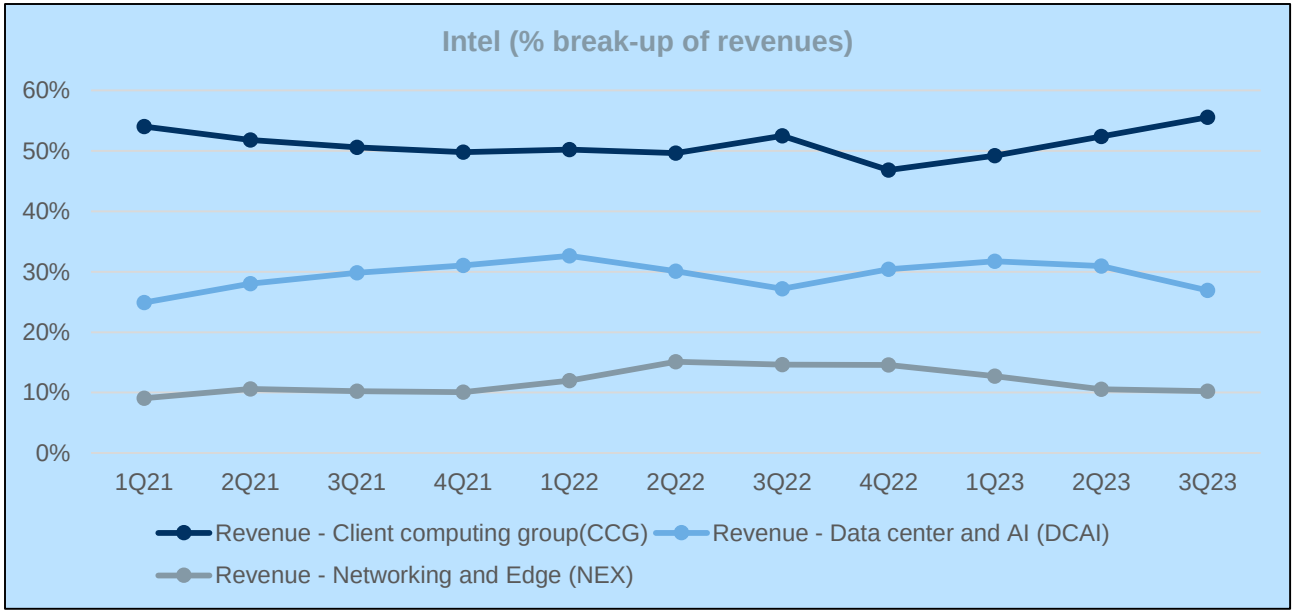
Section 6

Semiconductor-Related Trends

RBC Capital Markets

Intel Corporation (INTC) – Revenue

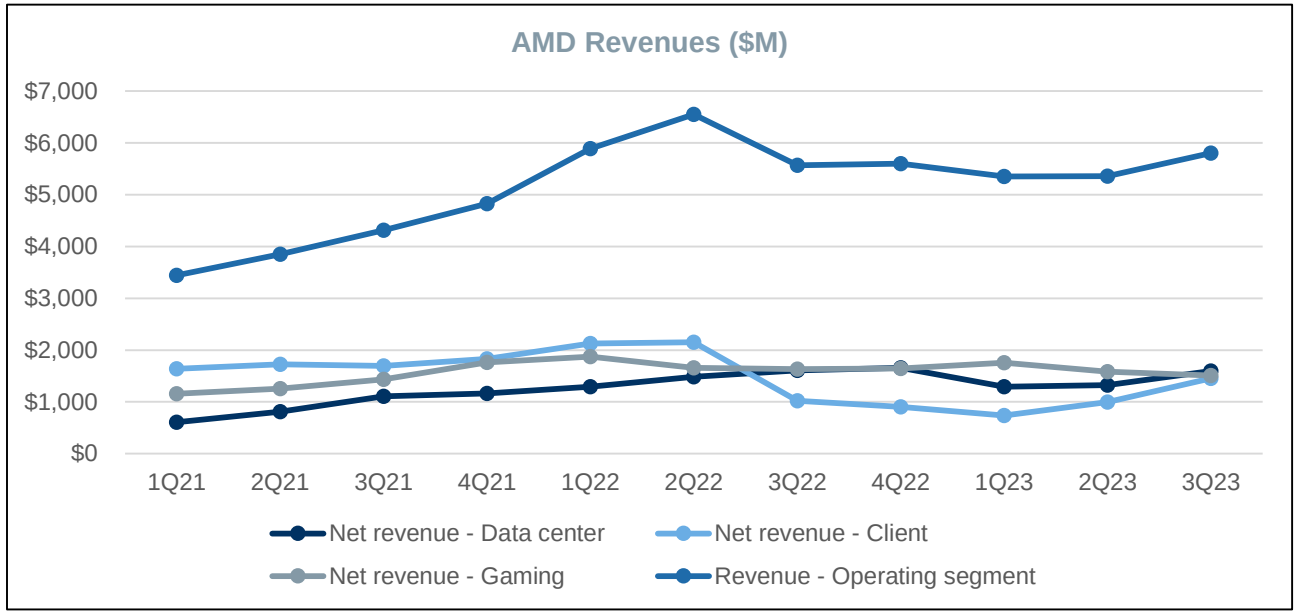


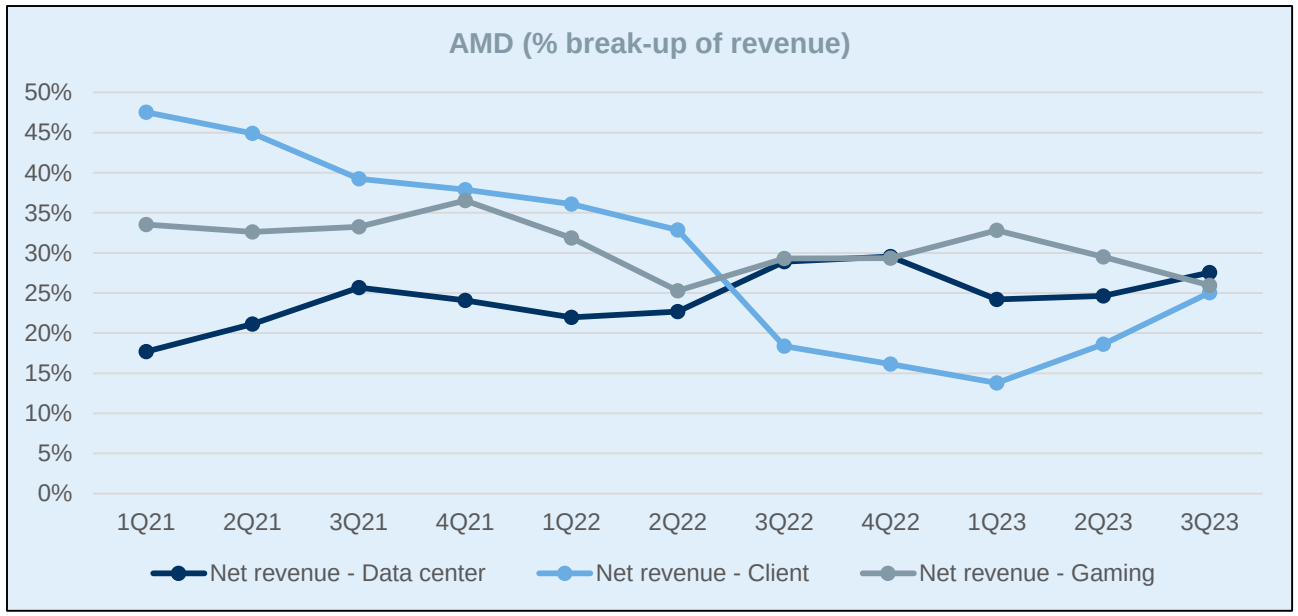


- Total Revenue for 3QFY23, declined 8% Y/Y and increased 9% Q/Q.
- CCG Revenue increased 16% Q/Q but was down 3% Y/Y, primarily on lower desktop volumes.
- DCAI Revenue decreased 5% Q/Q and 10% Y/Y, driven by a decrease in server revenue. Server volume decreased due to lower demand in a softening CPU Datacenter market.
- NEX Revenue increased 6% Q/Q and was down 32% 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 ~56% / 27% / 10% of the total revenues for 3QFY23.

Source: Company reports, Visible Alpha

Advanced Micro Devices, Inc. (AMD) – Revenue

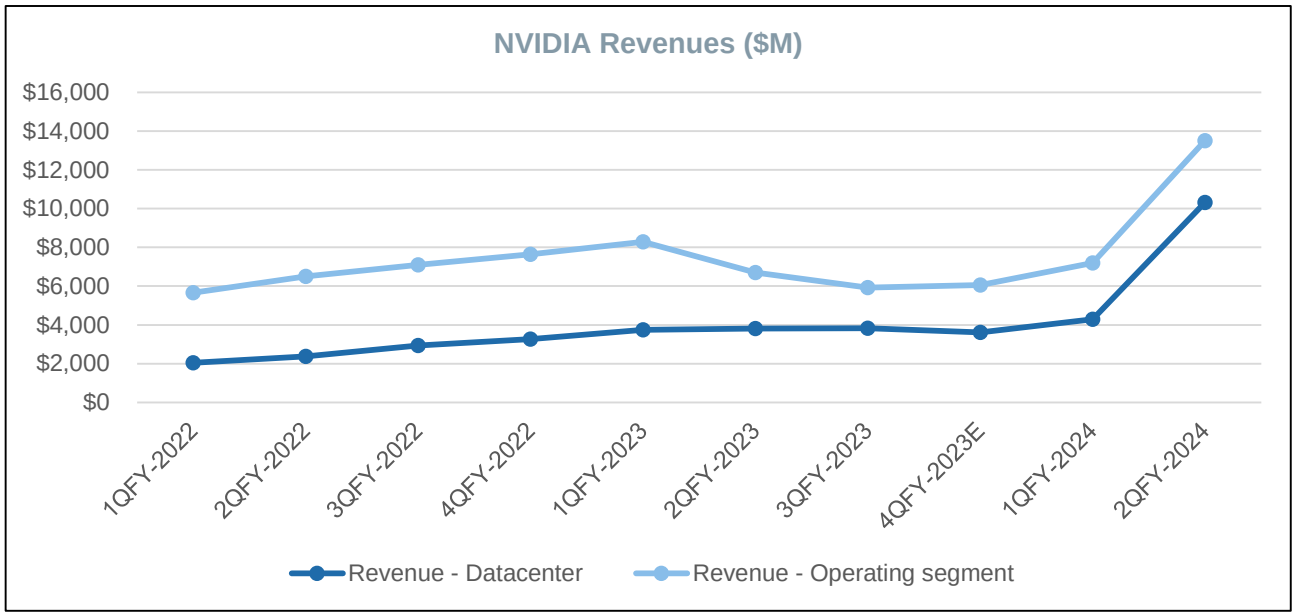


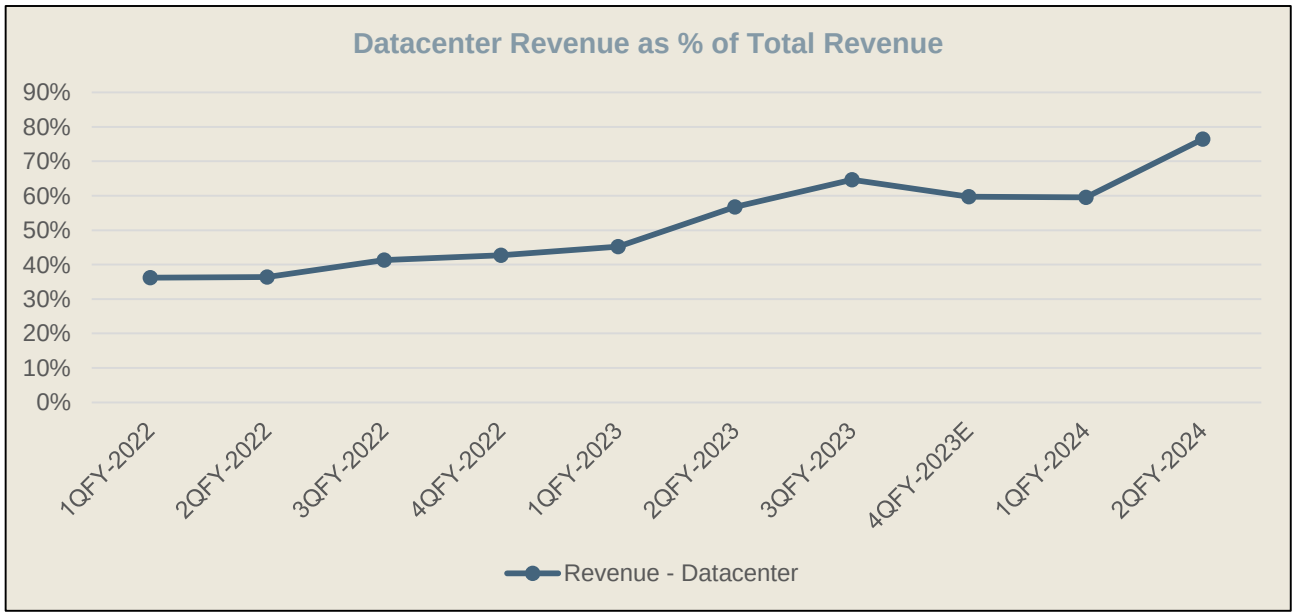


- Total Revenue for 3QFY23, increased 4% Y/Y and was up 8% Q/Q.
- Datacenter Segment Revenue declined 1% Y/Y but was up +21% Q/Q, as 4th Gen AMD EPYC™ CPU as customer adoption of 4th Gen AMD EPYC CPUs accelerated during the quarter.
- Client segment revenue increased 42% Y/Y and was up +46% Q/Q driven by an increase in AMD Ryzen™ 7000 Series CPU sales.
- Gaming segment revenue declined 8% Y/Y and declined 5% Q/Q, primarily due to lower semi-custom sales.
- Datacenters / Client / Gaming segments represented ~28% / 25% / 26% of the total revenues for 3QFY23.

Source: Company reports, Visible Alpha

NVIDIA Corporation (NVDA) – Revenue

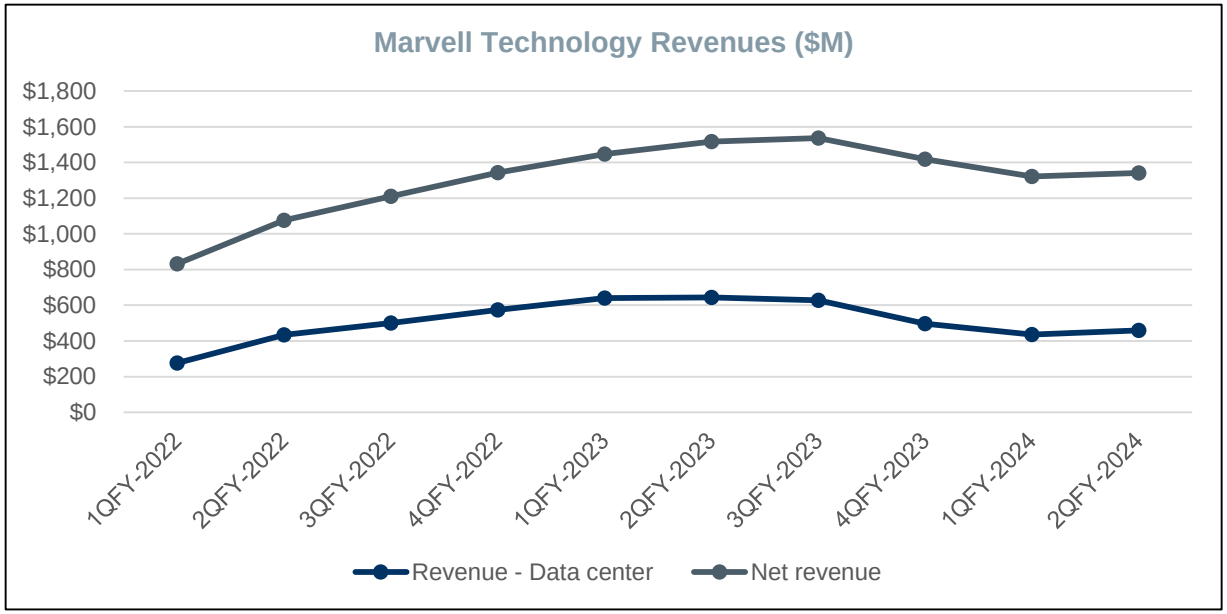


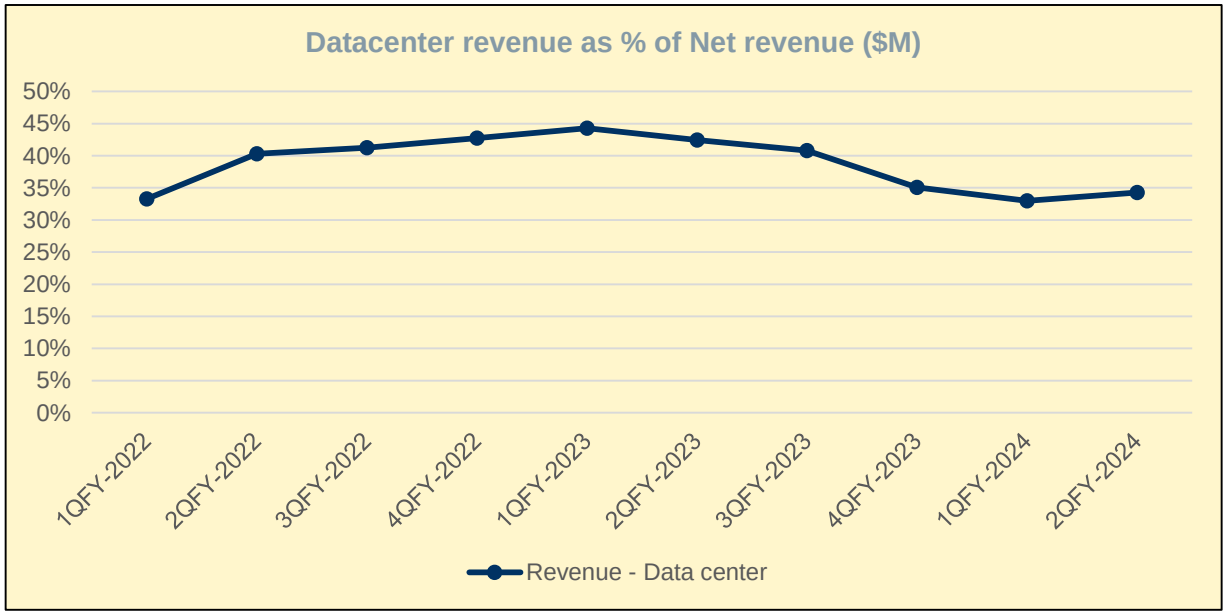


- Revenue for 2QFY24 was $13.5 billion, up 101% Y/Y and up 88% Q/Q.
- Datacenter revenue was up 171% Y/Y and up 141% Q/Q, led by CSPs and large consumer internet companies. Strong demand for the NVIDIA HGX platform based on Nvidia's Hopper and Ampere GPU architectures was primarily driven by the development of large language models and generative AI. Datacenter Compute grew 195% Y/Y and 157% Q/Q, largely reflecting the strong ramp of Company's Hopper-based HGX platform. Networking was up 94% Y/Y and up 85% Q/Q, primarily on strong growth in InfiniBand infrastructure to support Nvidia's HGX platform.
- Datacenter segment represented ~76% of the total revenues for 2QFY24.

Source: Company reports, Visible Alpha

Marvell Technologies (MRVL) – Revenue





- Revenue for 2QFY24 was $1,341 million, down 12% Y/Y and up 1% Q/Q.
- Datacenter revenue was down 29% Y/Y and up 6% Q/Q.\ Datacenter segment represented ~34% of the total revenues for 2QFY24.
- In 2Q, there was sequential growth in storage data center revenue from a low base in 1Q, and modest sequential growth is expected in 3Q. However, due to significantly depressed storage end market demand and high customer inventory, industry expectations for a datacenter storage recovery have been meaningfully delayed.
- For the third quarter, Management expects accelerated sequential revenue growth from overall cloud, surpassing last quarter's performance, driven by continued strong growth from Cloud AI and standard cloud infrastructure.
- Marvell is enabling AI with a broad range of solutions, which include PAM4-based optical DSPs and AECs for connecting accelerator clusters inside AI data centers. DCI products for connectivity between regional data centers, low-latency, high-capacity Ethernet switches for fabric connectivity inside data centers and custom silicon for compute acceleration.

Source: Company reports, Visible Alpha

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 large language model, 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 $234.5m 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 large-language models (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 $47B Chips Act	On April 18, 2023, the European Union moved forward with the €43B ($47B) 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

Required Disclosures

Companies mentioned

Adobe Inc. (NASDAQ: ADBE US; $597.22; Outperform)

Alphabet Inc. (NASDAQ: GOOGL US; $132.59; Outperform)

Amazon.com, Inc. (NASDAQ: AMZN US; $143.56; Outperform)

Celestica Inc. (NYSE: CLS US; $25.48; Outperform)

Digital Realty Trust, Inc. (NYSE: DLR US; $128.75; Outperform)

DigitalBridge Group, Inc. (NYSE: DBRG US; $15.58; Outperform)

Gitlab Inc (NASDAQ: GTLB US; $44.61; Outperform)

HubSpot, Inc. (NYSE: HUBS US; $428.92; Outperform)

Meta Platforms, Inc. (NASDAQ: META US; $328.77; Outperform)

Microsoft Corporation (NASDAQ: MSFT US; $369.67; Outperform)

NEXTDC Limited (ASX: NXT AU; AUD12.42; Outperform)

nVent Electric PLC (NYSE: NVT US; $51.28; Outperform)

Pro Medicus Limited (ASX: PME AU; AUD85.23; Sector Perform)

ServiceNow, Inc. (NYSE: NOW US; $634.76; Outperform)

Wix.com Ltd. (NASDAQ: WIX US; $89.29; Outperform)

Xero Limited (ASX: XRO AU; AUD98.49; Sector Perform)

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 30-Sep-2023

			Investment Banking Serv./Past 12 Mos.	
Rating	Count	Percent	Count	Percent
BUY [Outperform]	820	55.97	250	30.49
HOLD [Sector Perform]	590	40.27	148	25.08
SELL [Underperform]	55	3.75	5	9.09

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

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”).