CABOT PARTNERS
OPTIMIZING BUSINESS VALUE

Insights from Efficient Generative AI and AI Hardware & Edge AI 2023 Conferences

Jean S. Bozman, President,
Cloud Architects Advisors LLC
M. R. Pamidi, Ph.D., Principal Analyst

September 2023

Cabot Partners Group, Inc.
100 Woodcrest Lane, Danbury, CT 06810.
www.cabotpartners.com
info@cabotpartners.com

Insights from Efficient Generative AI and AI Hardware & Edge AI Conferences

Introduction

The advent of Generative Artificial Intelligence (GenAI) is a catalyst for personal computing and, now, for enterprise computing. This September, twin conferences – both focused on AI – drove home that customers will adopt and use AI/Machine Language (ML), large language models (LLMs), foundation models, customized hardware, and cloud services as they integrate AI into everyday IT processes.

Customers and vendors described a range of GenAI usage models applied for personal productivity (GenAI Analytics) and enterprise use cases that leverage AI/ML analytics and data management. The specific models and toolsets will vary by the customer's use cases and their choice of AI/ML models. How this AI-enabling process plays out will vary, depending on the industries involved and the business use cases customers envision for their AI workloads.

Personal use cases for GenAI are generally intuitive, quickly generating combinations of text, images, audio, and video, especially in GenAI's role as a co-pilot or assistant to a business user or developer. These GenAI uses –proofs of concept and early use cases, are speeding searches of vast quantities of data and data analytics processes. However, it is crucial to monitor the accuracy of the results carefully. Otherwise, the "hallucinations" caused by data errors and bias can skew the results.

By contrast, Enterprise uses of GenAI will require deeper IT skills that select and apply AI to programming (e.g., PERL, Python, R), DevOps, deployments, and maintenance of AI-enabled applications designed for multi-cloud environments. Companies must discover dependencies on their older or aging code and address them for a company's enterprise AI adoption to fit within established data integrity, security, and governance/compliance guidelines.

Use Cases for GenAI

The world changed – and technology shifted – when ChatGPT, a form of GenAI, was made available in Fall 2022. Suddenly, new AI-based decision-making and application-building scenarios became possible using GenAI models and toolsets. GenAI will transform software and use cases across all major industry segments. However, adoption rates will vary greatly, depending on the degree of security, data privacy, reliability, and availability required in a specific sector.

"This is going to touch every industry," said Bratin Saha, Vice President And General Manager of ML and AI Services at Amazon Web Services (AWS). "We've reached an inflection point for AI … Every domain is going to get reinvented with generative AI," he said, adding: "You need a very performant infrastructure." Often, that means tapping specific processor types to match compute resources with specific workloads (CPUs, DPUs, GPUs, IPUs, NPUs, or TPUs).
Each industry will accomplish that transformation in its own way, Saha noted, as customers will continue to support security, data privacy, and governance/compliance requirements, which will vary by sector.

For customers, GenAI will "transform how we interact with technology," said Amin Vahdat, Vice President and General Manager of ML, Systems, and Cloud AI at Google. "Continued advances will bring similar leaps with multi-modal, real-time generation of text, images, audio, and video." Cloud service providers like Google Cloud and its competitors will work to amplify compute efficiency by providing specialized infrastructure that leverages distributed computing and memory. These infrastructure gains will compensate for slower-than-expected gains for DRAM and memory in recent years, Vahdat said. Cloud providers like Google must support customers' demands to use extremely LLMs for AI capacity, fueling visualization/imaging, natural language engines, inferencing, and recommendations.

Key Concerns about Applying GenAI in the Enterprise

Key concerns about integrating GenAI into enterprise workloads include data privacy, security, sustainability, and data sovereignty. These attributes were delivered differently by transactional systems, which have long been the workhorses of enterprise data centers.

Regarding sustainability, cloud providers are highlighting environmental gains and carbon-reduction pledges:

- Google Cloud has publicly committed to operating its cloud services using 24/7 carbon-free energy by 2030.
- Microsoft (Azure) claims that:
  - Its cloud infrastructure is up to 93% more energy-efficient and 98% more carbon-efficient than on-premises solutions.
  - By 2030, its operations will be carbon-negative, water-positive, and zero waste, with the benefit of protecting more land than is used to host the cloud-services infrastructure.
  - By 2050, it will remove all historical carbon generated since Microsoft's founding in 1975.[1]

These claims, if fulfilled, look very impressive. However, they represent the large cloud providers' viewpoints, so the amount of efficiency and sustainability savings will vary by cloud services provider and data centers that adopt AI-based operations to run their company's databases and applications.

New Use Models Combining Cloud Migration + AI/ML

With multi-cloud options for application deployment, cloud-based computing will gain many more migrating workloads. For cloud services, AI-enabled software will aid data management – directing AI queries to specific compute engines. It will help customers 'scale out' queries to distributed resources and improve performance. Essential to this scale-out infrastructure will be directing specific tasks to the 'right' processing resources: CPUs, GPUs, DPUs, NPUs, and TPUs, relying on AI software to make the most efficient choices for mapping incoming requests to computing resources.

This multi-cloud world already exists – and many large customers are already leveraging two or three public cloud service providers (CSPs). IDC data shows that cloud migration has produced a mix of enterprise applications – some delivered via the cloud (public cloud or private cloud) and some delivered on-premises in the data center. After the pandemic began in 2020, cloud migration has accelerated – showing that the customers' choices for deployment are real, supporting cloud deployments when a company's data-protection policies allow it. Today, customers use a mix of compute and storage resources, with specific tasks directed to specific compute resources in the cloud and the data centers they operate.

Vinesh Sukumar, Senior Director and head of AI/ML Product Management at Qualcomm told the conference he believes that cloud economics makes it very challenging for GenAI applications to scale to broader groups of consumers. That's why, he said, the future of AI is hybrid computing -- with AI processing distributed between the cloud and the device, but the challenge of doing so depends on how the orchestration is defined and enabled.[2] He added that initial planning and careful oversight of where workloads are running is vital to gaining timely and accurate results.

[1] "Hardware for the Age of Copilots," Marc Tremblay, Microsoft, *AI Hardware & AI 2023*.
[2] "The Future of AI is 'On-Device,'" Vinesh Sukumar, Qualcomm, *AI Hardware & AI 2023.*

Edge ML: Analysis of Market Readiness

Becky Soltanian, VP of R&D at Sanborn in Mountain View, CA, delved into Edge ML, a branch that runs ML algorithms and models directly on edge devices. It includes devices such as smartphones, IoT things, embedded systems, robots, or other localized computing platforms, rather than relying on remote cloud servers for processing, which would produce delayed results due to distance.

Edge ML has applications in autonomous drones, self-driving cars, industrial robots, and agricultural robots, with currently shipping products that include Spot by Boston Dynamics, DJI drones, Tesla Autopilot, and many robots in manufacturing.

We believe that what is emerging is a hybrid ML model: with Edge ML performing localized processing on edge devices minimizing latency, reducing data transfer, and ensuring privacy -- and Cloud-based ML with centralized processing on remote servers, offering scalability but probably introducing network delays and data security concerns.

Ethics in AI

There have been a lot of controversies in the last few years concerning social media's role in society. The question many in the social blogosphere are asking is this: Do they have to honor freedom of speech? Can they be used maliciously to spread false news, provide "alternative facts" from a parallel universe, etc.?

Amidst these controversies, it was delightful to hear Andrew Ng, Founder and CEO of Landing AI and founder of DeepLearning.AI, tell the AI Hardware Summit audience that ethics are vital to AI's success: "My teams work only on projects that move humanity forward. For example, we kill otherwise financially sound projects on ethical grounds."[3] Similarly, Amazon Web Services calls for *Responsible AI,* stating, "Our commitment to develop AI and ML responsibly is integral to our approach."[4]

What Customers are Saying?

In the gaming industry, accelerated development is very often the goal of companies in that sector. "They can iterate faster and put the code out there by using AI – and who doesn't want to be 30 times more productive?" said Luc Barthelet, CTO of Unity, which sells software that helps its customers generate electronic and video gaming systems. "Completely new genres [of games] are going to appear," using real-time processing of 3D images, language translation, and sound, Barthelet said—all of them supported by AI software. He said an estimated 1.5 million developers are using Unity's AI-based software tools, and "at each step of the way, we are accelerating AI" as a building block of software development.

Praveen Kolli, Staff ML Engineer at the DoorDash food-delivery service, said the new AI technologies must be carefully deployed and tested to ensure the accuracy of predictive results. "You must look forward as well as backward," he told the AI Hardware Summit conference. That is part of the learning curve when using GenAI and Enterprise AI, along with regression testing and comparing results based on AI/ML analytics. So, improving the IT skillsets for on-staff developers – and emphasizing the importance of adding to their AI/ML skills – is becoming increasingly important to the accuracy of results. "Fine-tuning [the AI models] requires a different kind of toolset," Kolli said.

3 "Opportunities in AI," Andrew Ng, Landing AI, AI Hardware & AI 2023.
4 "Enterprise Scale Generative AI," Dr. Bratin Saha, AWS, AI Hardware &Edge AI 2023.

Alternative Solutions to GPUs

Many customers we spoke with expressed dissatisfaction about waiting six months or longer to get their hands on the highest-performing NVIDIA hardware because of the massive demand for the company's hardware and its efforts to catch up on the pent-up demand caused by the COVID-19 pandemic. So, are there any other options for GPUs?

We believe there are other options, even though NVIDIA remains a top choice for GPUs running AI workloads. For instance, not every AI-enabled application will require the fastest GPU. Depending on an application's needs, an ASIC, FPGA, IPU, NPU, TPU, a lower-performing NVIDIA GPU, or a combination of those technologies might be a workable alternative platform for AI/ML (Figure 1).[5]

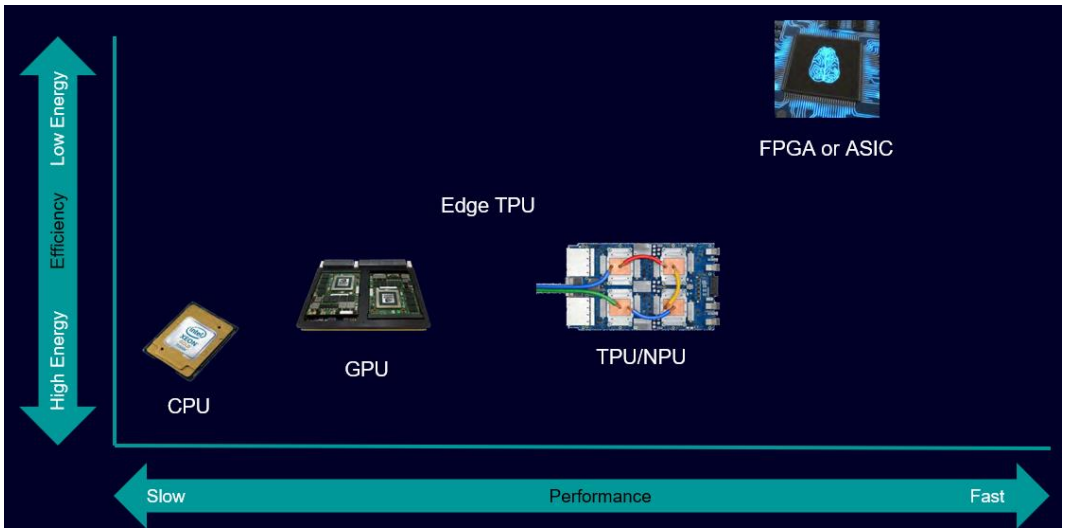

Figure 1: One size does not fit all

We were also impressed with Redwood City, CA-based Numenta, a company founded by Jeff Hawkins and Donna Dubinsky, co-founders of Palm Computing. Numenta CEO Subutai Ahmad told the AI Hardware conference that, based on the firm's extensive neuroscience research, the company has developed the Numenta Platform for Intelligent Computing (NuPIC) for unparalleled scaling of LLMs and CPUs (Figure 2).[6] It discovered that CPUs could support AI/ML processing as an alternative to the industry's GPUs.

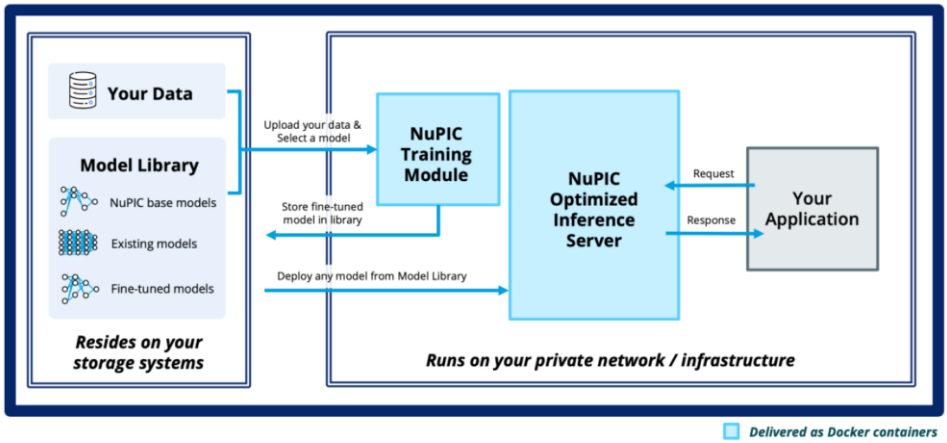

Figure 2: Numenta Architecture

5 "AI in Everything, Everywhere, At the Edge'" Russell Klein and Ankur Gupta, AI Hardware &Edge AI 2023.
6 Numenta Datasheet.

Deployed as a Docker container, NuPIC operates entirely within a customer's infrastructure, ensuring that models and data remain private and under customer control. It runs on various infrastructure options, such as cloud providers or on-premises data centers – and integrates into existing MLOps tools as a lightweight containerized solution. The company claims that Numenta delivers a 64X improvement in GPT throughput compared to current CPUs.

Customers looking for GPU alternatives may evaluate Numenta instead of waiting months for an NVIDIA solution. Depending on their specific application or workload, others will look to companies that provide hybrid solutions, leveraging GPUs, GPU clusters, and specialized silicon.

Future Trends

1. Semiconductor companies are designing *domain-specific artificial intelligence* chips (Figure 3), offering customers, on one hand, a wide choice of products to choose from. On the other hand, interoperability and integration issues with existing infrastructure and the long-term viability of these vendors may pose challenges for end-to-end AI workloads.

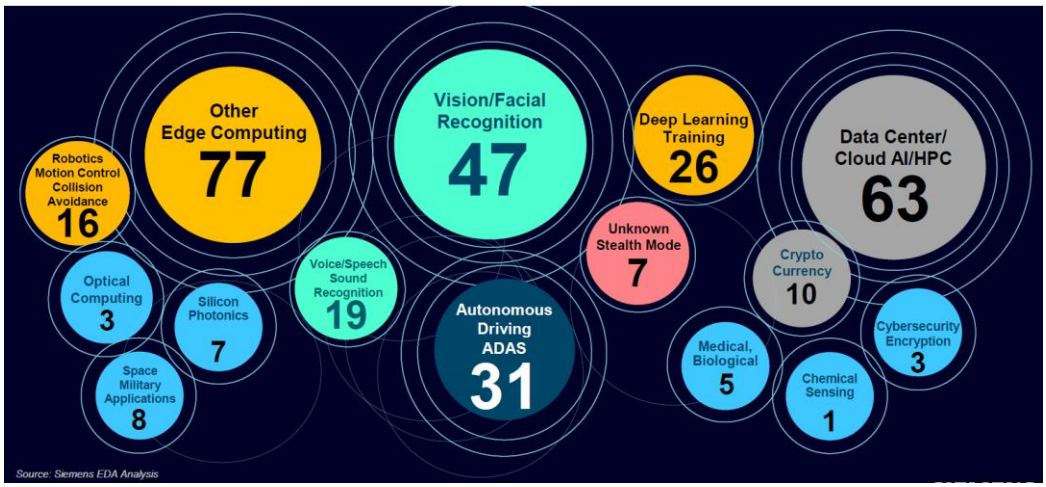

Figure 3: Application-Specific Chip Vendor Landscape

2. Continued pressures on Moore's Law (Figures 4 and 5) and the ever-increasing chip density (Figure 6) will present significant challenges to software developers; it could be that AI – coupled with new types of highly distributed hardware deployments – will come to their rescue.

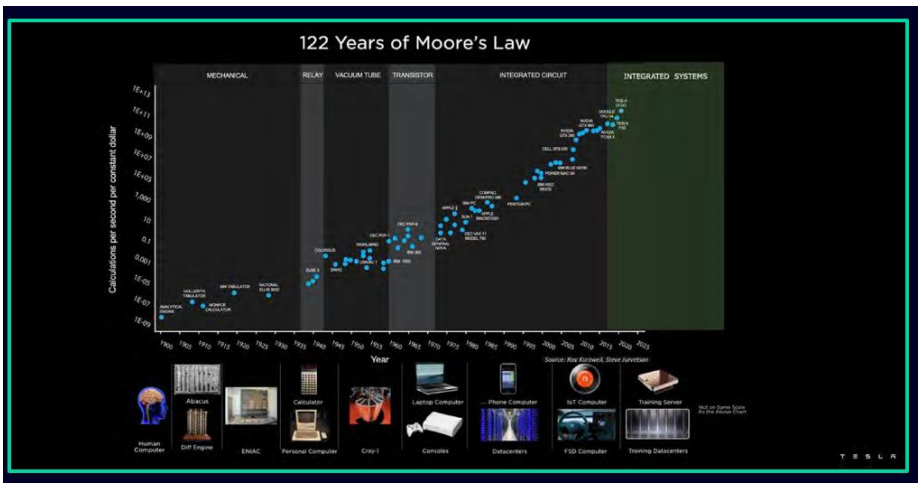

Figure 4: Moore's Law to Date and Its Future[7]

7 "How is AI Revolutionizing Chip Innovation?" Thomas Andersen, Synopsis, AI Hardware &Edge AI 2023.

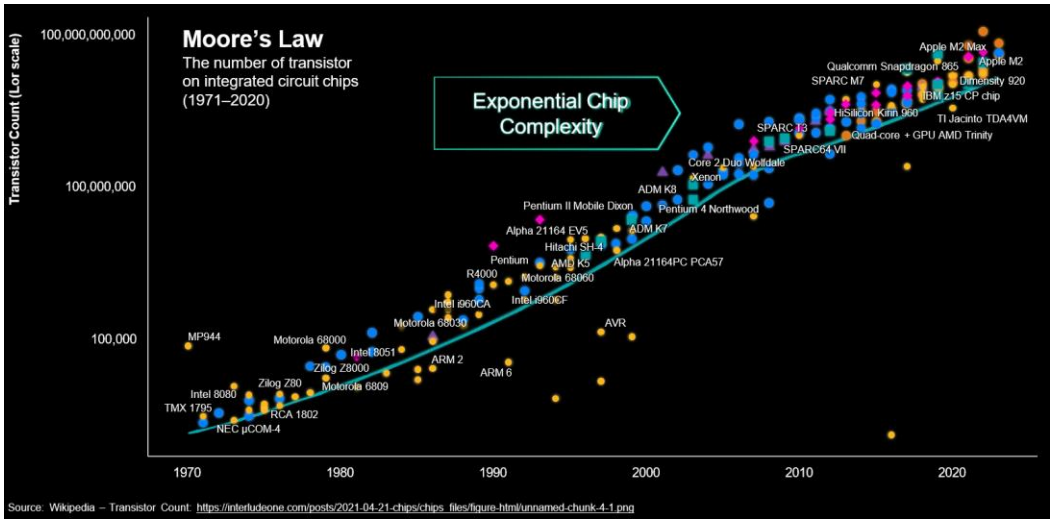

Figure 5: Moore's Law Pushes the Limits of Chip Design[7]

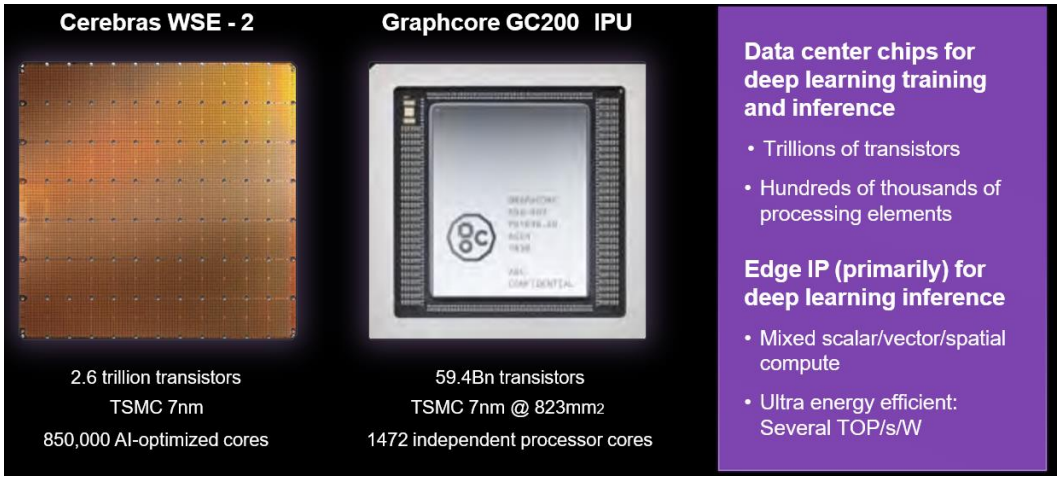

Figure 6: Example of Super Chips[7]

Summary

The case for AI in the enterprise is being constructed using LLMs, AI development toolkits (including those from AWS, Google, and Microsoft), and various compute resources developed for specific processing styles. In the coming years, it will take a while for companies to figure out which approaches to AI work best – and that will be based on actual use cases, not on proofs of concept (PoCs) or beta-level products beginning their march into the marketplace.

The momentum is building for AI as businesses seek better, more efficient outcomes from their IT and cloud investments. For each sector, AI technology is rapidly reinventing IT and changing the industry. "AI is transforming software," said Marshall Choy, Senior VP of Product Management at SambaNova Systems in Palo Alto, California. "Models are the new code."

CABOT PARTNERS

OPTIMIZING BUSINESS VALUE

Cabot Partners Group, Inc.

100 Woodcrest Lane, Danbury, CT 06810

www.cabotpartners.com| info@cabotpartners.com