OPSW-304-1: AI Open Eco-System

Javier Gonzalez
Organizer - Open-Source Lead
Principal Software Engineer
Samsung Electronics

Prasad Venkatachar
Session Chair
Sr Director – Product & Solutions @ Pliops
IEEE Sr member & BCS Fellow

Panel Introduction

Akshay Subramaniam
Senior AI Developer
Nvidia

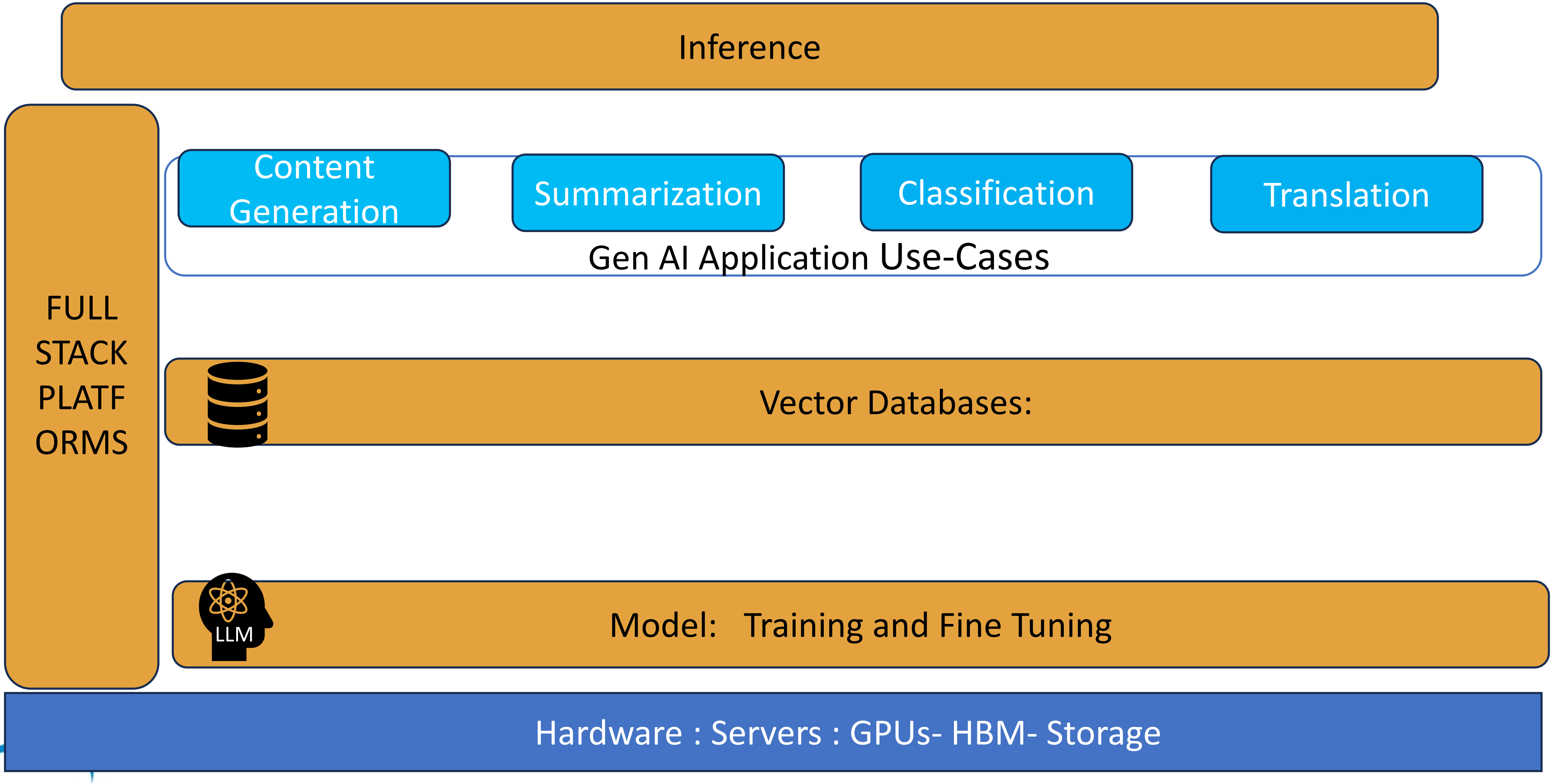
Eric Kern
Distinguished Engineer &
Lenovo AI COE Lead

Sandy Ghai
Group Product Manager
Google Cloud

Yann Collect
Tech Lead
Meta

Nilesh Shah
VP Business Development
ZeroPoint Technologies

Gen AI Stack


Inference
FULL STACK PLATFORMS
Content Generation
Summarization
Classification
Translation
Gen AI Application Use-Cases
Vector Databases:
LLM
Model: Training and Fine Tuning
Hardware : Servers : GPUs- HBM- Storage

Session Discussion Topics - Breakdown

- Session Venue : Ballroom C, Floor 1 Santa Clara Convention Center
- Session Date & Timing: Aug 8th – 1.25 to 2.30 PM PT:
  - Infra Layer - 15 Minutes
  - Model Layer: Training and Fine Tuning – 12 Minutes
  - RAG – 12 Minutes
  - Inference Layer – 12 Minutes
  - Tech Predictions – 5 Minutes
  - Q & A - 15 Minutes

Infrastructure

1) Yann : Back in the day for regular recommendation systems it used to be 8 to maximum 500 GPUs. Now for Gen AI for Llama 2 to 3 training we are looking at 24K(active16K) GPUs for 15 trillion tokens. How does the hyperscaler like Meta manages the systems at Scale and take care of System Resiliency, Power & Cooling.

2) Akshay: let's say Meta builds their Llama4 or What are key innovation in GPUs from Nividia that will help bring down power, energy and power demands, especially Nvidia Blackwell platform promises to 25x less cost and energy consumption than the NVIDIA Hopper architecture. What are major System architecture advantages to accomplish such Goals

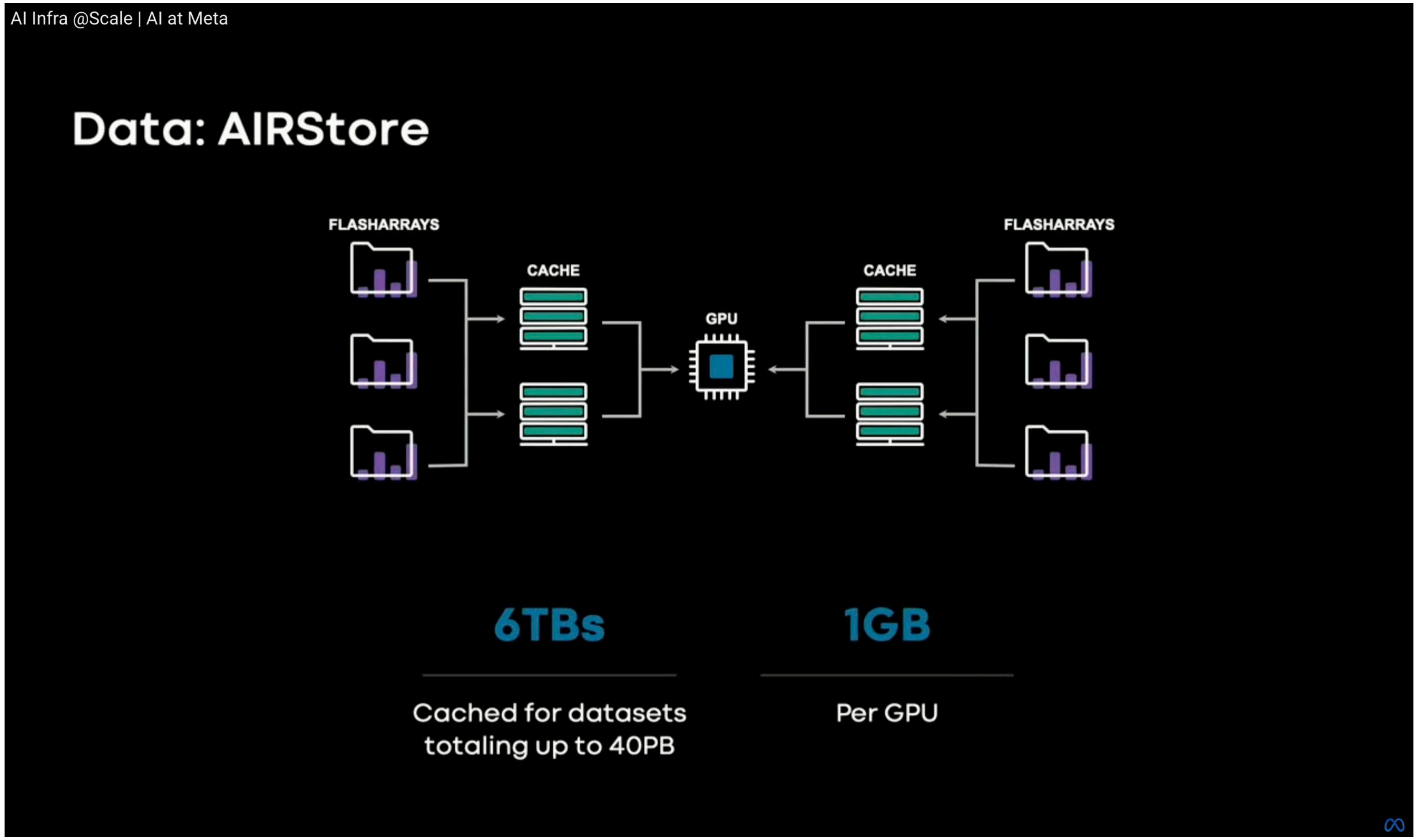
3) Yann & Nilesh: What are the key takeaways from Hyperscalers implementation that are readily applicable for Enterprises.

4) Eric: Lenovo has making great progress from building a GPT in Box to Neptune Water cooling systems. How is the customer experience with liquid cooling systems do they see energy reduction? What are the trade-offs and challenges associated with liquid cooling implementation?

Component	Category	Interruption Count	% of Interruptions
Faulty GPU	GPU	148	30.1%
GPU HBM3 Memory	GPU	72	17.2%
Software Bug	Dependency	54	12.9%
Network Switch/Cable	Network	35	8.4%
Host Maintenance	Unplanned Maintenance	32	7.6%
GPU SRAM Memory	GPU	19	4.5%
GPU System Processor	GPU	17	4.1%
NIC	Host	7	1.7%
NCCL Watchdog Timeouts	Unknown	7	1.7%
Silent Data Corruption	GPU	6	1.4%
GPU Thermal Interface + Sensor	GPU	6	1.4%
SSD	Host	3	0.7%
Power Supply	Host	3	0.7%
Server Chassis	Host	2	0.5%
IO Expansion Board	Host	2	0.5%
Dependency	Dependency	2	0.5%
CPU	Host	2	0.5%
System Memory	Host	2	0.5%

Table 5 Root-cause categorization of unexpected interruptions during a 54-day period of Llama 3 405B pre-training. About 78% of unexpected interruptions were attributed to confirmed or suspected hardware issues.

Yann Schematic Diagram : AIRStore



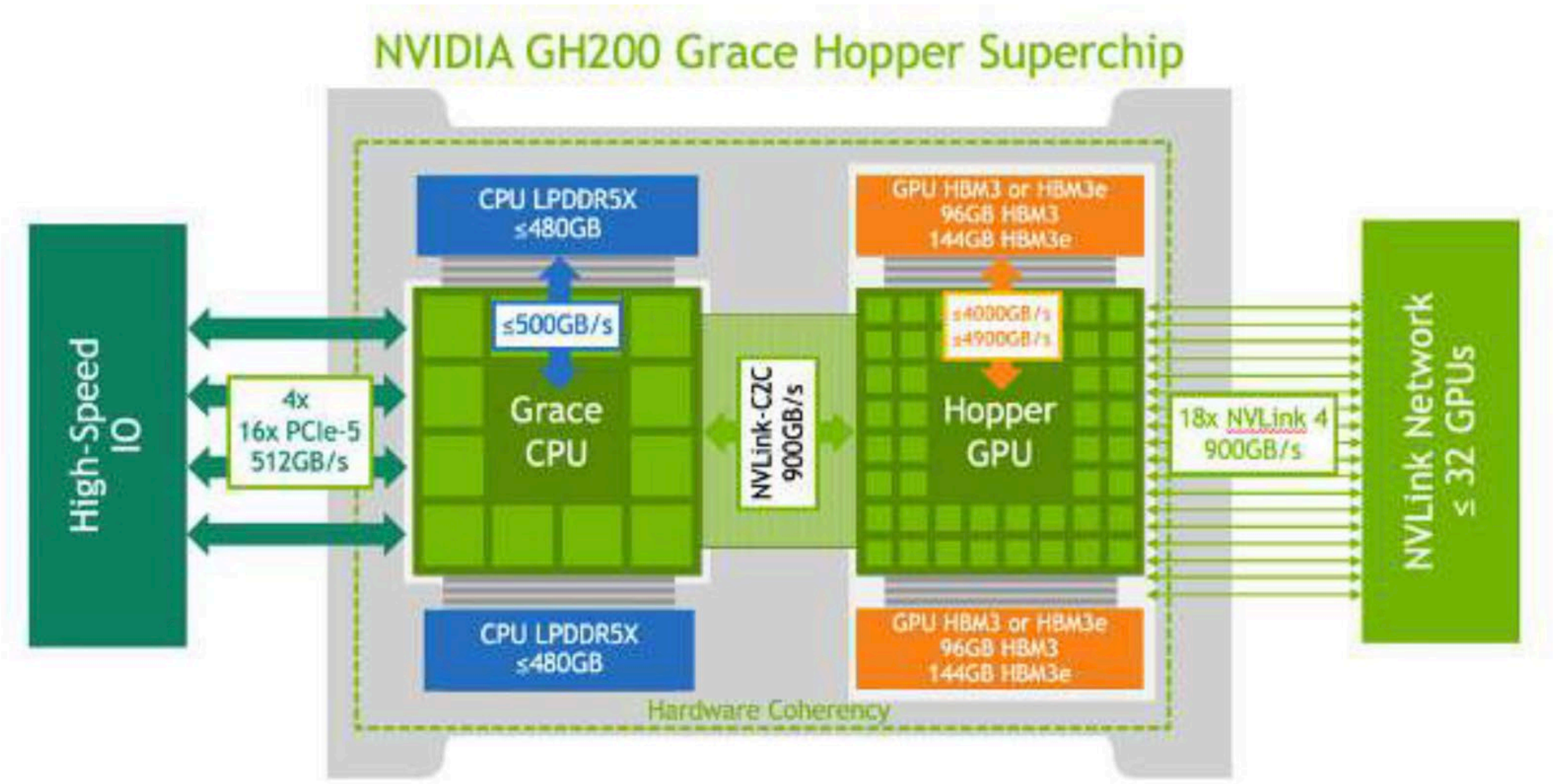


Figure 1. NVIDIA GH200 Grace Hopper Superchip Logical Overview

Compression

- Yann: Model sizes have increased by 250X from initial GPT-2 to latest Llama3 with 400 Billion parameters and Some multimodals like Gemini, GPT-4 potentially might have Trillion parameters. What is the role of Data Compression when model sizes increasing steadily.
- Akshay: How can hardware and software-accelerated data compression techniques improve the efficiency of Gen AI training?
- Nilesh: There is classic mix of opensource and proprietary data/memory compression approaches. How does the industry adopt best of both world approaches

Model Layer : Training and Fine Tuning

- Akshay: Model training is limited to few selective companies. However according to Gartner 2028 more 40 to 50% build their own models due to enterprise needs. How does one arrive at Compute and Memory capacity sizing for a given size of the model for Training.

- Akshay: What are the key hardware and software optimizations one has to factor in Gen AI model training?

- Nilesh: What are the different model compression techniques employed for Fine Tuning

- Akshay: How Small Language Models Excel with Fewer Parameters?

  Llama3(8 Billon), Microsoft Phi-3(3.8, 7 Billion), Google Gemma (2 and 7 Billion), Apple (0.27 Billion)

- Sandy: Adopt Off the shelf models with prompt engineering vs Fine Tuning. How does enterprises conclude Prompt Engineering is sufficient the effort and cost of Fine Tuning is not worth it.

Fine Tuning vs RAG

- Sandy : Fine Tuning vs RAG, Which one organizations choose over other under what scenarios. How does the cost comes into equation.
- Eric: What's Lenovo customer experience in terms of effort and response accuracy Fine Tuning vs RAG.
- Sandy: What are benefits of databases with enhanced vector functionality vs native vector databases
- Eric/Sandy : How do you see customer deploying RAG applications, are they leveraging existing database/schema/tables or building a brand new database.

RAG

- **Sandy:** What are the key techniques to improve the performance of vector search. How is AlloyDB AI innovations helping the customers.
  - SCAN vs HNSW vs IVF
- **Eric:** What are system demands from vector databases compared to traditional databases. How is Lenovo helping to meet the RAG application demands.
- **Sandy:** How does one select effective Chunking Strategy and the role of Embedding model for building RAG systems
- **Sandy:** What are key metrics to evaluate before deploying RAG systems to production.
- **Eric:** What are benefits of Advanced RAG over Naïve RAG and when do we consider implementing it
- **Sandy:** How are Guardrails benefits RAG Applications and what are the tradeoff implementing it
  - Input GuardRails : PII information
  - Output GuardRails : Toxic Responses, Sensitive Information
  - Reliability vs Latency Tradeoff

Inference Layer. : SW & HW Optimization

- **Akshay:** For inference, we need the entire model to fit into memory. we would need to keep over a terabyte of data in memory — this exceeds any GPU in existence today
- **Nilesh:** What are common Inference Optimization Techniques
  - Quantization, KV Caching, Paged Attention
- **Eric:** How is Inference Engines helping customers to boost performance and reduce cost of LLM Deployment.
  - TensorRT, vLLM(NVIDIA AND AMD GPUs)
- **Sandy:** How is Google especially AlloyDB AI helping the customers for Inference optimization

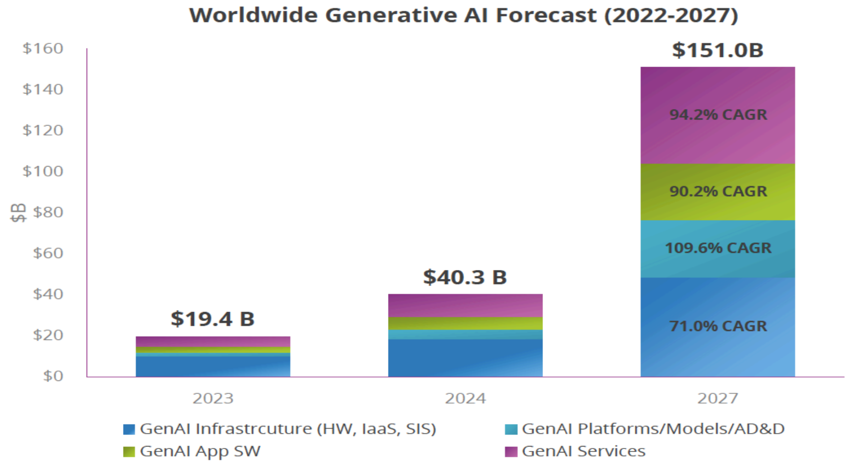
Tech Predictions

Common Question to All:

- What are your Technology predictions for next 3 to 5 Years

- what would like to see changes in Gen AI stack that benefits customers and Industry as a whole for ease of adoption

The Generative AI Market Opportunity



IDC 2023

OPSW-304-1: AI Open Eco-System

Linkedin Post?

Javier Gonzalvez
Organizer: Open-Source Lead
Principal Software Engineer
Samsung Electronics

Prasad Venkatachar
Session Chair & Moderator
Sr Director – Product & Solutions @ Pliops
IEEE Sr member & BCS Fellow

Akshay Subramaniam
Senior AI Developer
Nvidia

Eric Kern
Distinguished Engineer &
Lenovo AI COE Lead

Sandy Ghai
Group Product Manager
Google Cloud

Yann Collect
Tech Lead
Meta

Nilesh Shah
VP Business Development
ZeroPoint Technologies