AWS re:Invent

DECEMBER 2 – 6, 2024 | LAS VEGAS, NV

CMP207

AWS-accelerated computing enables customer success with generative AI

Dvij Bajpai
(he/him)
Senior Product Manager
AWS

Kirmani Ahmed
(he/him)
AI Engineering Lead
Meta

Samantha Pham
(she/her)
Principal Product Manager
AWS

Agenda

01 Industries, trends, and use-cases

02 Key customer needs

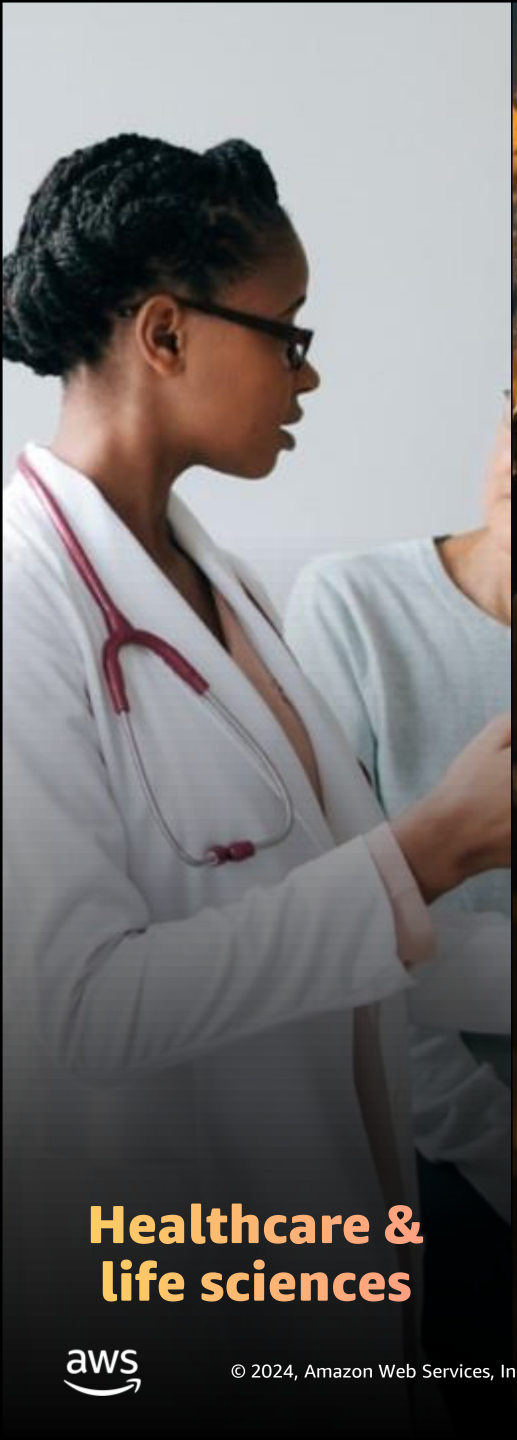
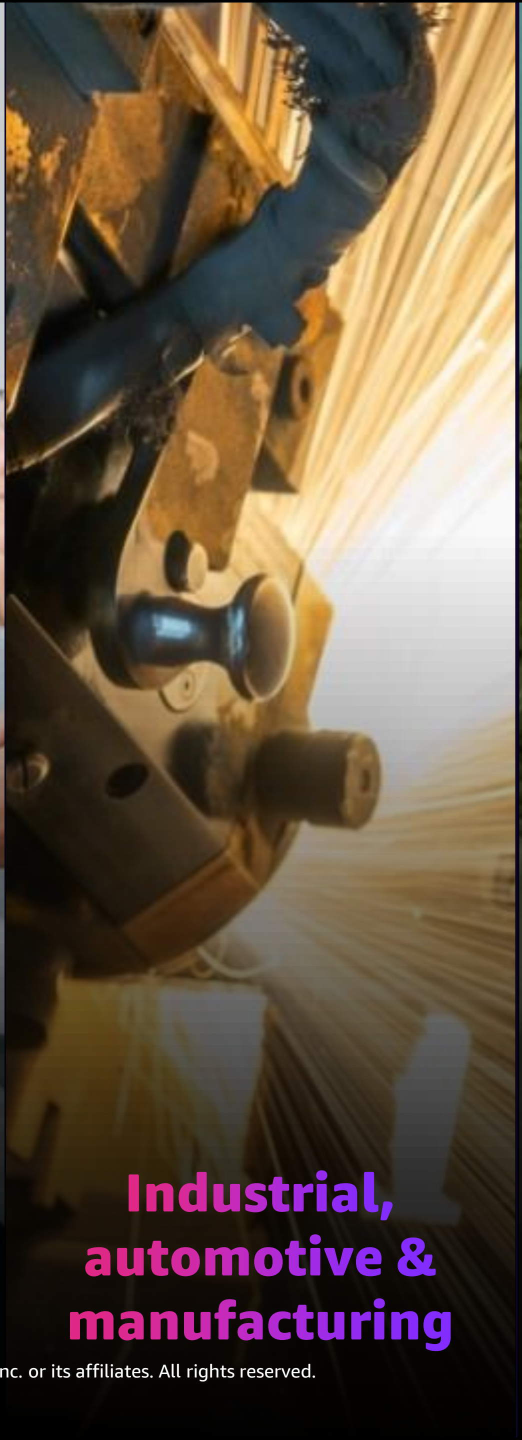
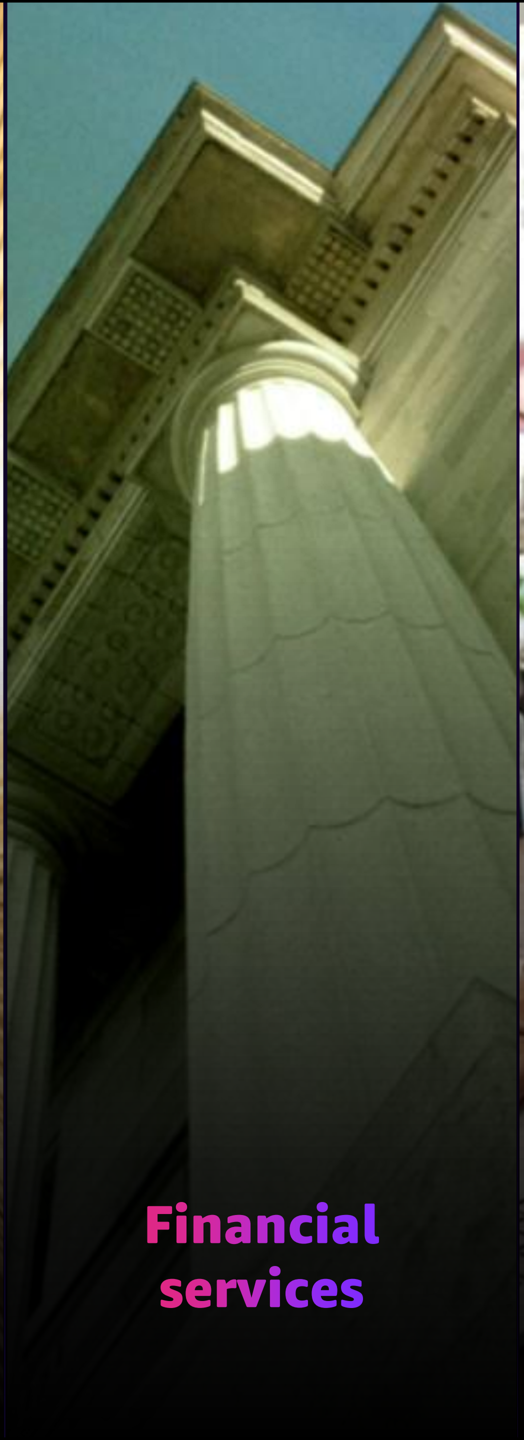
03 Amazon EC2 capabilities

04 AWS generative AI stack

05 Customer references

06 Meta case study

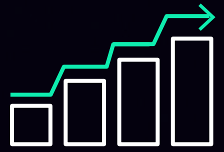
Industries, trends, and use-cases


Healthcare &
life sciences

Industrial,
automotive &
manufacturing

Financial
services

Retail

Media &
entertainment

Trends in AI/ML innovation

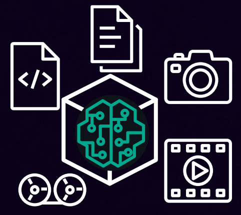


Increased scale in LLM training

Pre-training workloads leverage 10k+ GPUs at scale

Global deployment for LLM Inference

Global access to industry-leading LLMs

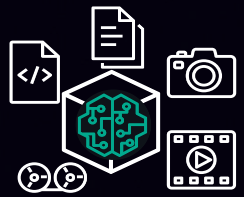


Emergence of multimodal models

Leading LLMs can now interpret and generate images and video

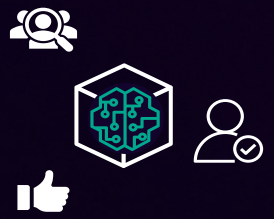
AI/ML model types and architectures

Multimodal LLMs

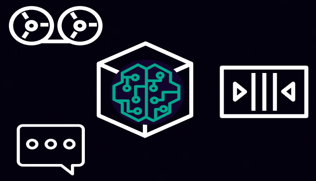


Large language models (LLMs)

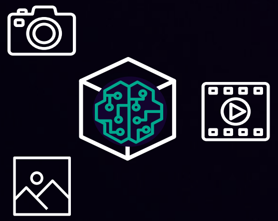
Recommender systems



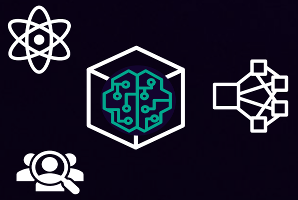
Speech models



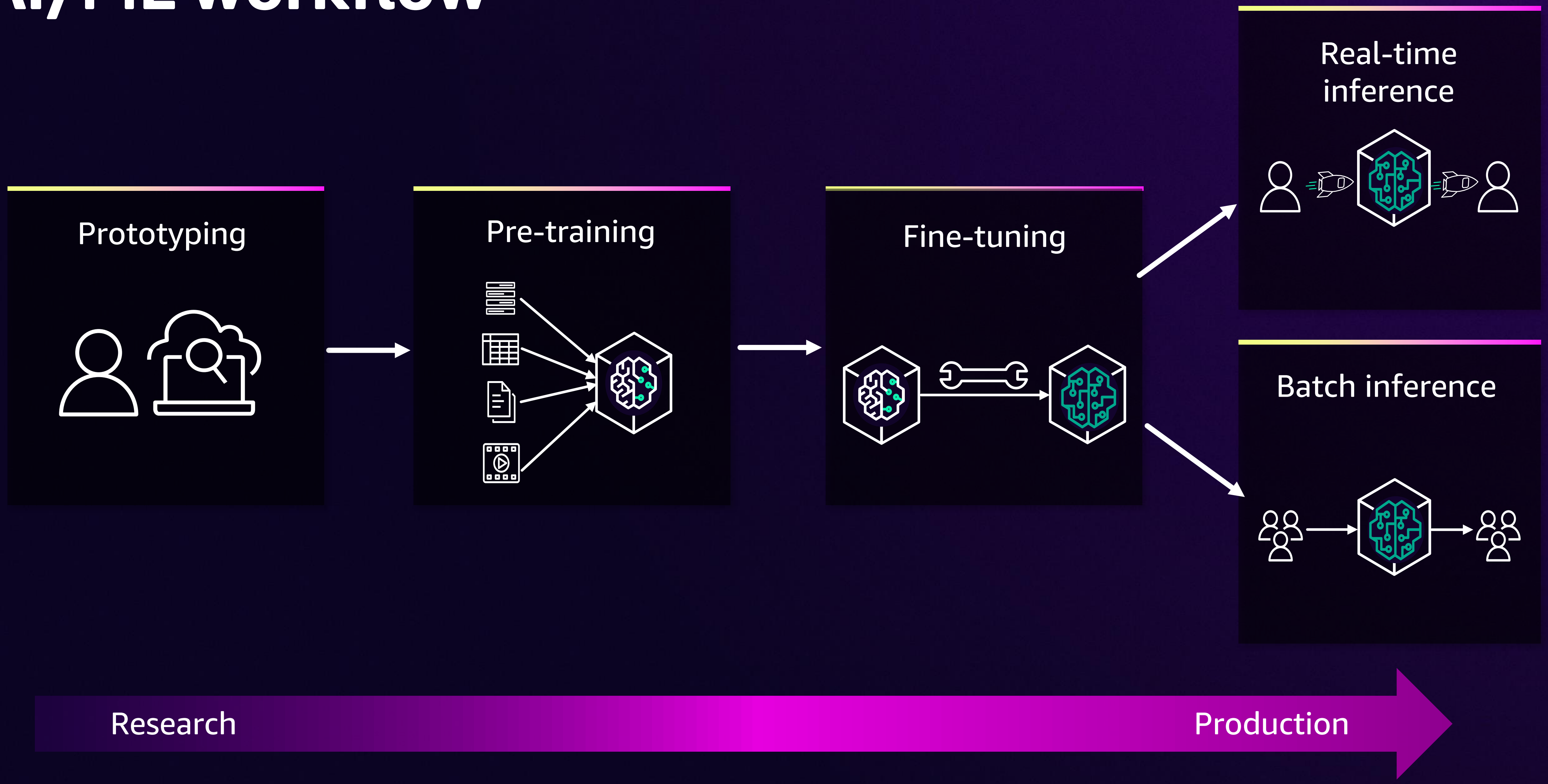
Vision models



Graph neural networks



AI/ML workflow



Key customer needs

Performance

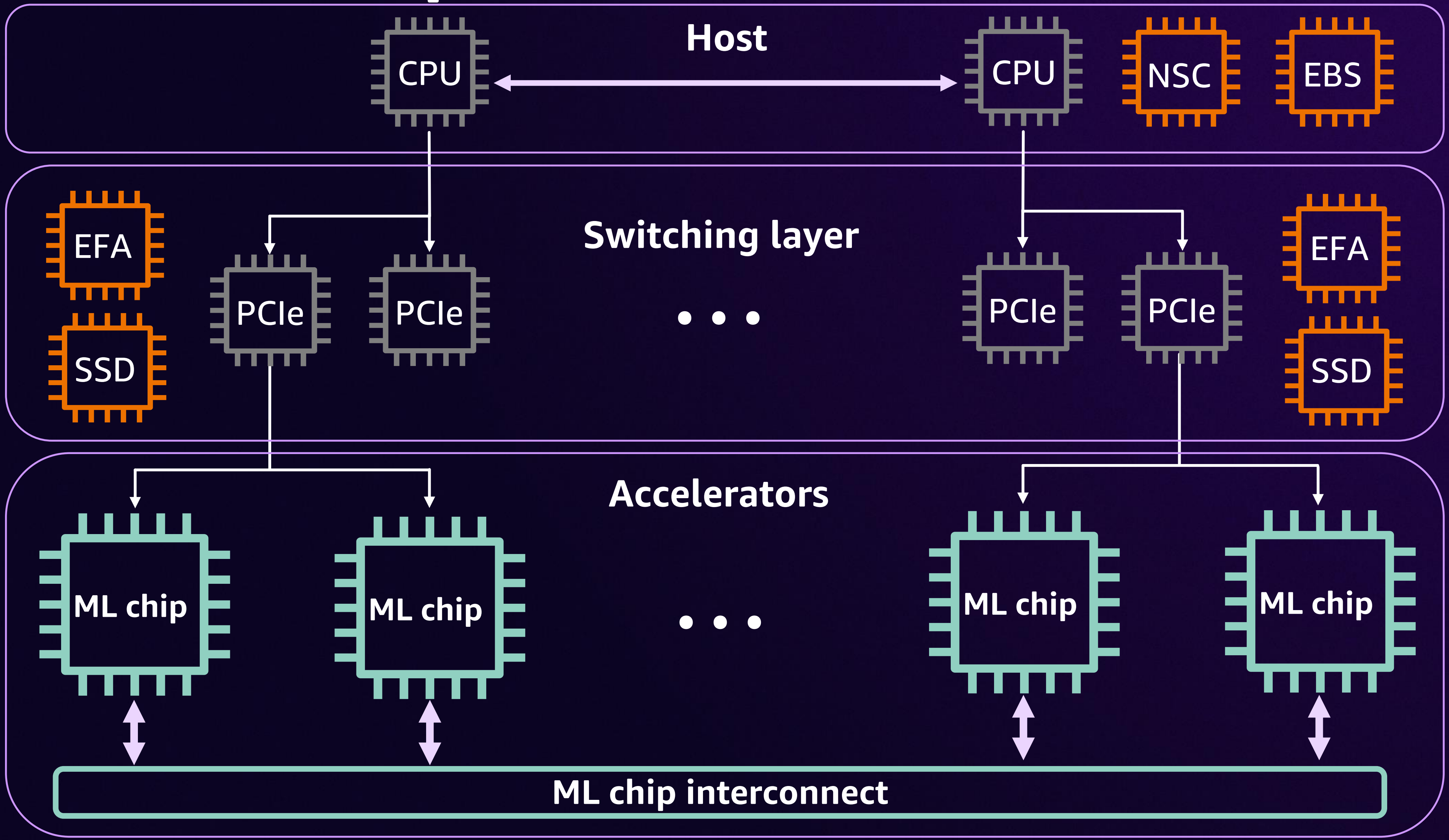
Cost

Security

Ease of Use

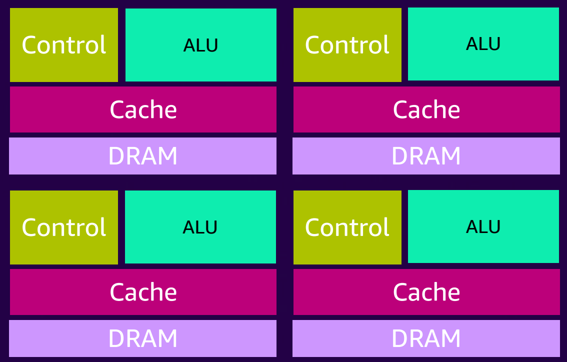
Amazon EC2 capabilities

Accelerated compute architecture



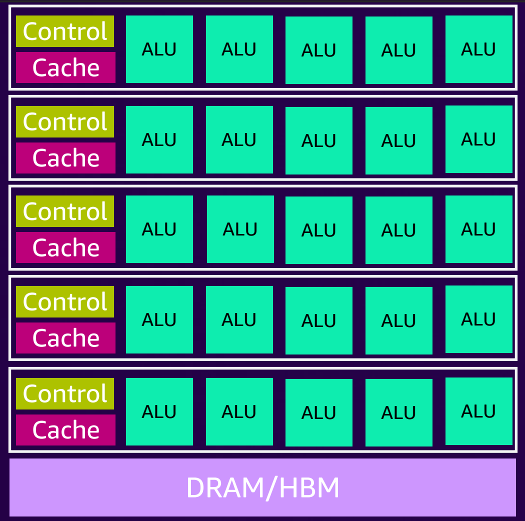
CPUs vs. GPUs vs. ASICs for DL acceleration

CPUs



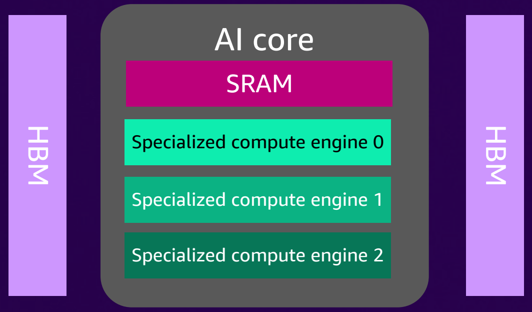
- 10s–100s of processing cores
- Optimized for general-purpose computing
- Time series; linear/logistic regression models

GPUs



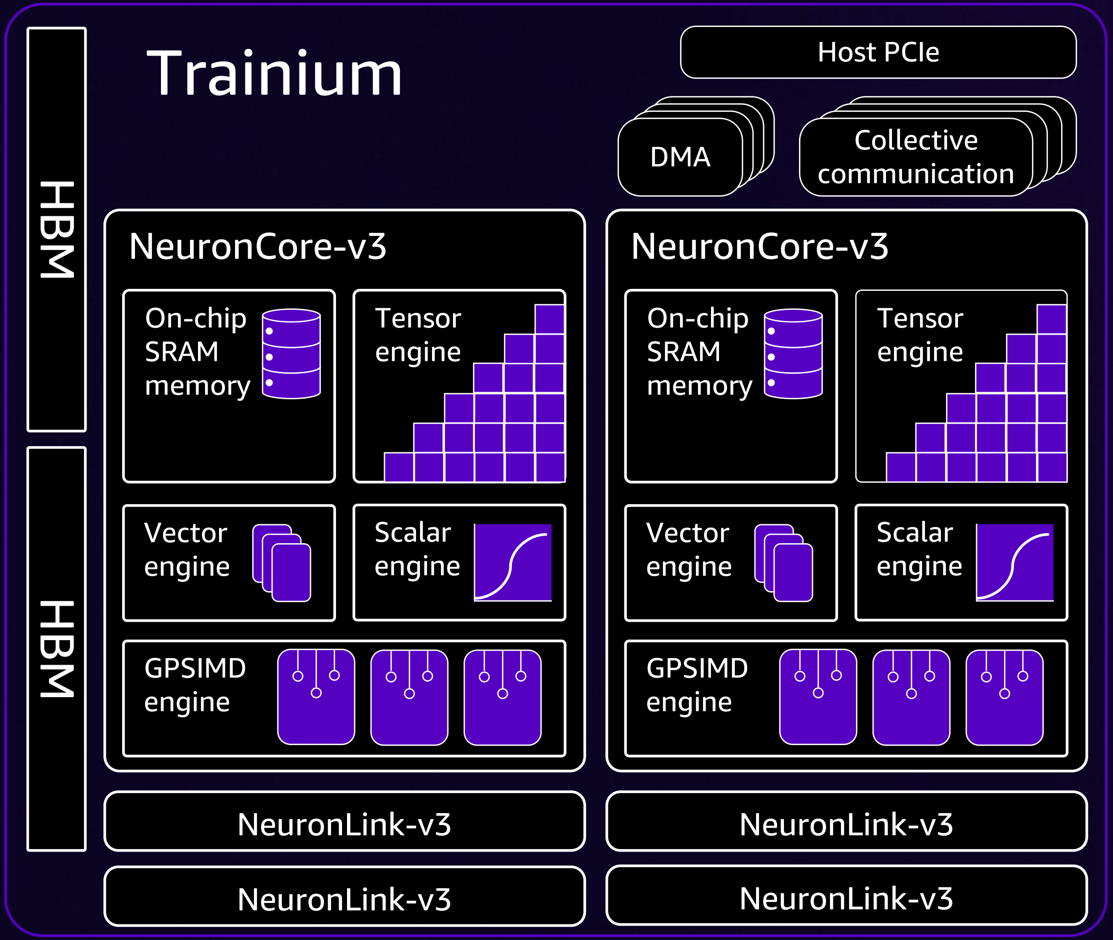
- 1,000s of processing cores
- Highly effective at parallel execution
- LLMs, CV, multimodal models

Custom AI accelerators



- Optimized and custom design for AI acceleration
- Highest price-performance for training and inference

AWS Trainium custom ML Chips



Trn2 powered by Trainium

BF16/FP16
10.4 PFLOPs

TF32
10.4 PFLOPs

FP8
20.8 PFLOPs

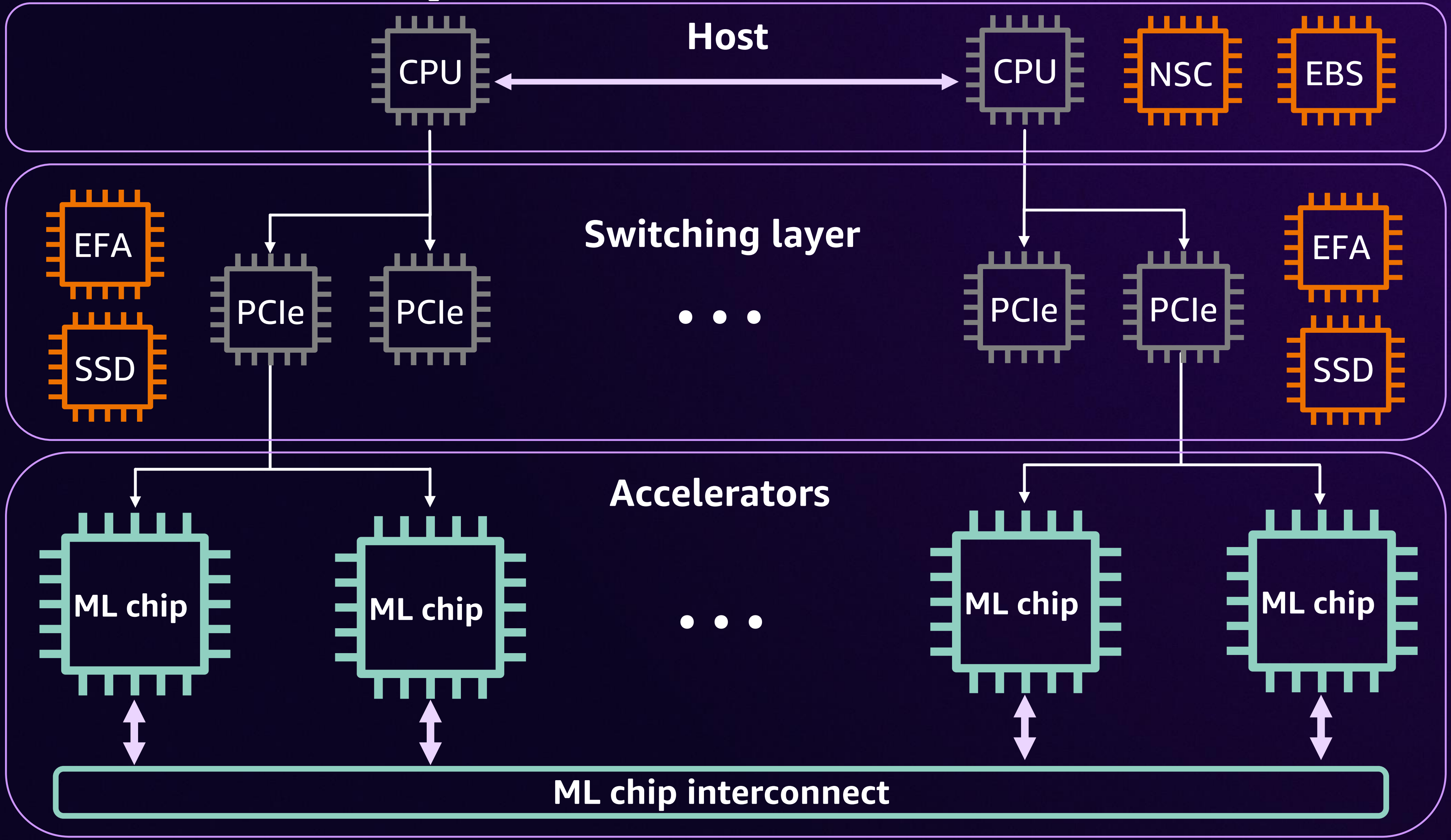
AGGREGATE ACCELERATOR MEMORY
1.5 TB

PEAK MEMORY BANDWIDTH
46 TB/sec

EFA NETWORK CONNECTIVITY
3.2 Tbps

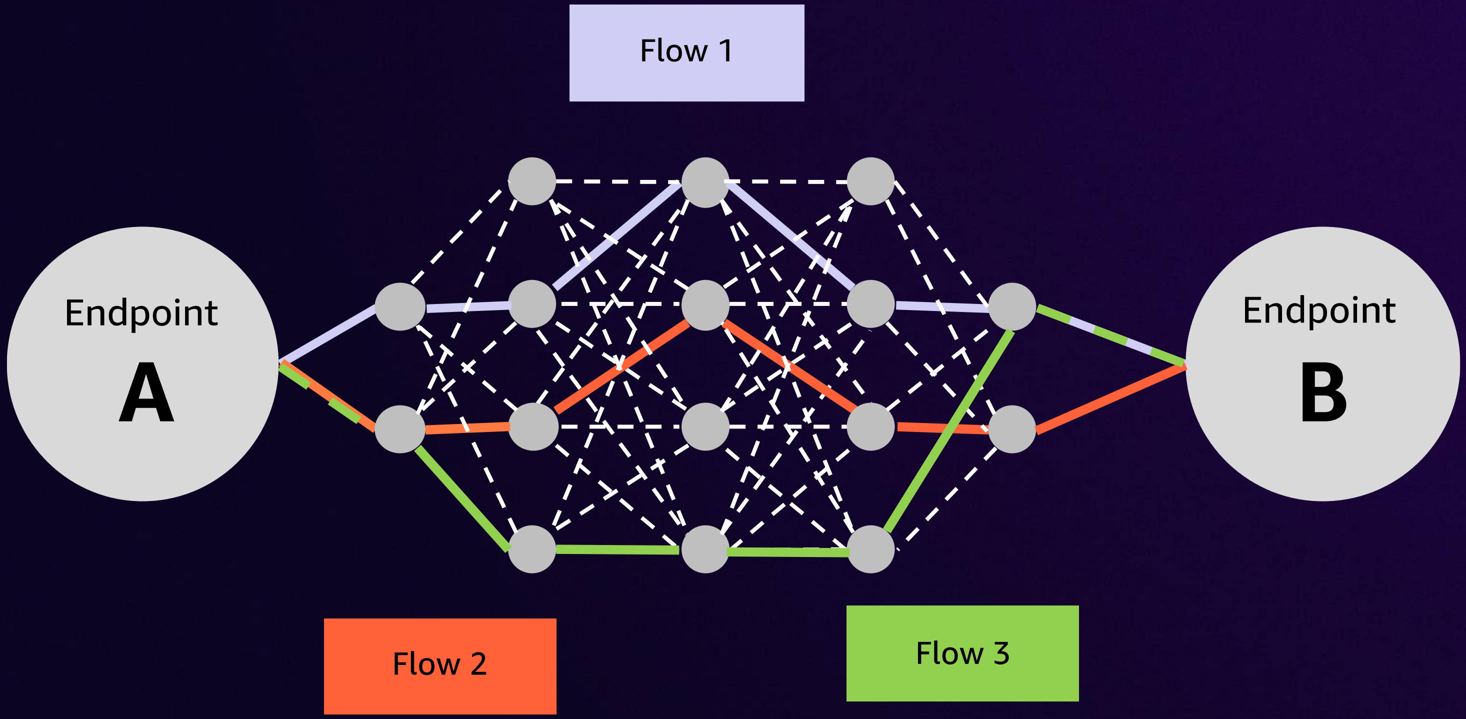
NEURON-CORE V3
NEURON-LINK V3

Accelerated compute architecture



Elastic Fabric Adapter (EFA): How it works

SCALABLE RELIABLE DATAGRAM (SRD)



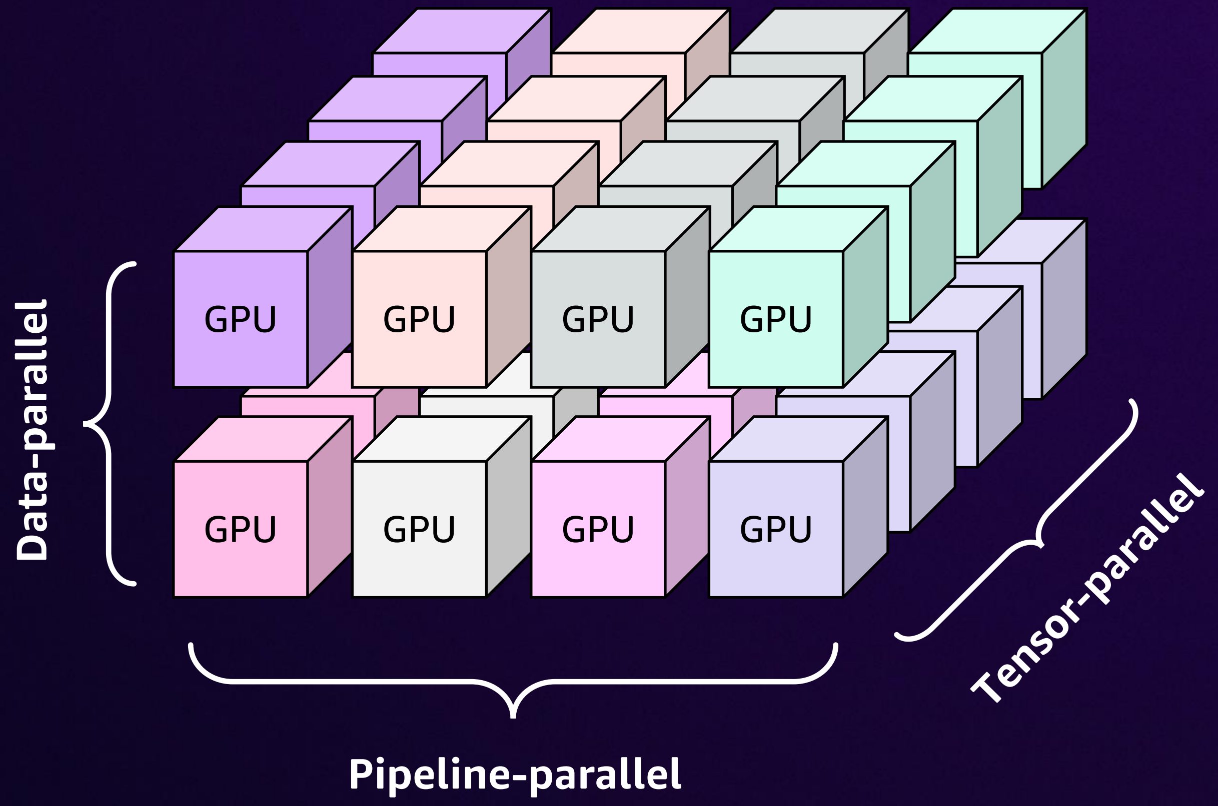
AWS-designed protocol that uses the many paths within the AWS network simultaneously

Designed into AWS Nitro System hardware

AI/ML scaling techniques

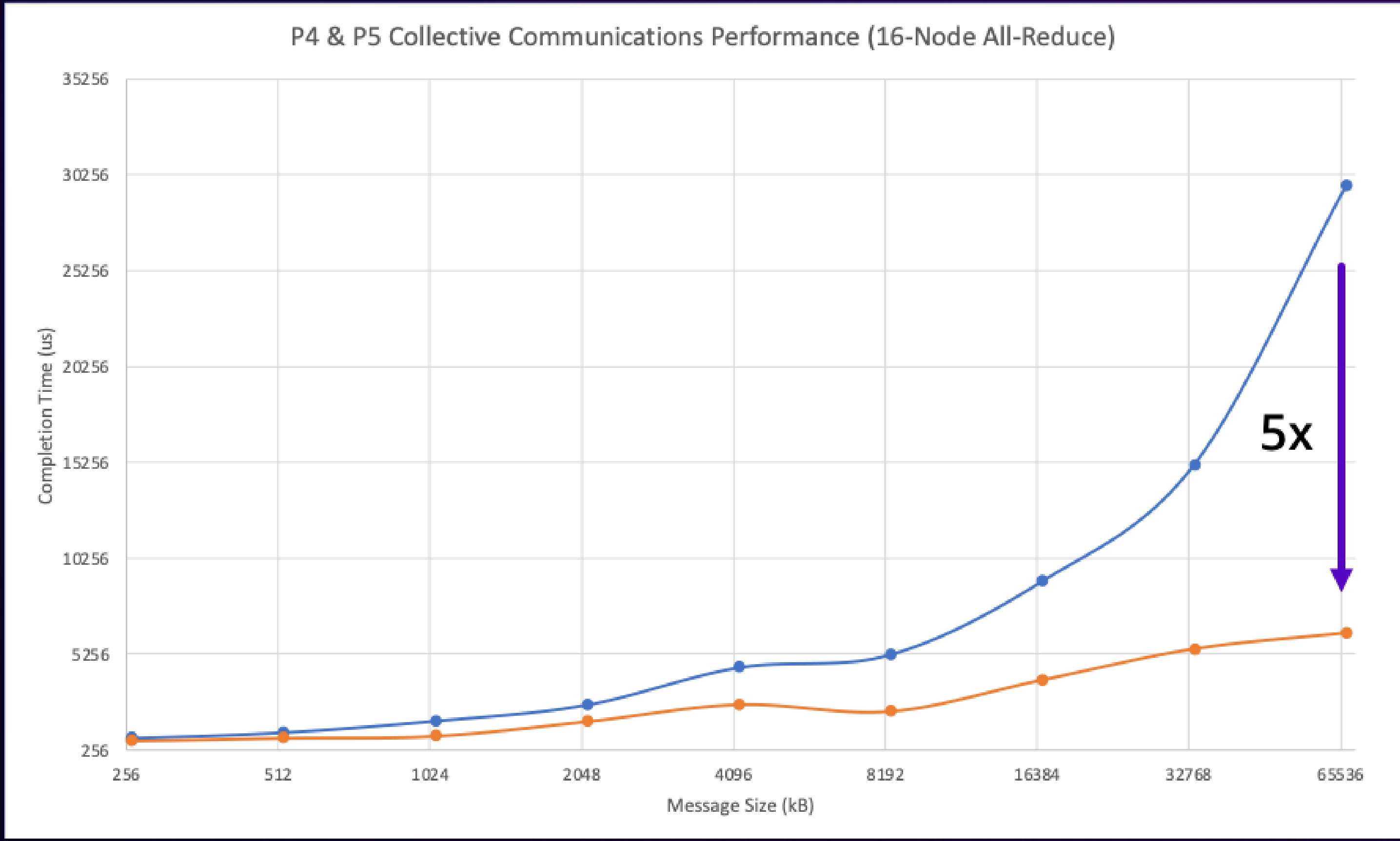
- Accelerate time to train by scaling to 10K+ GPUs
- Increase inference throughput by leveraging more GPUs
- Optimize for ML Chip utilization to maintain throughput/cost

3D parallelism to scale GPU workloads

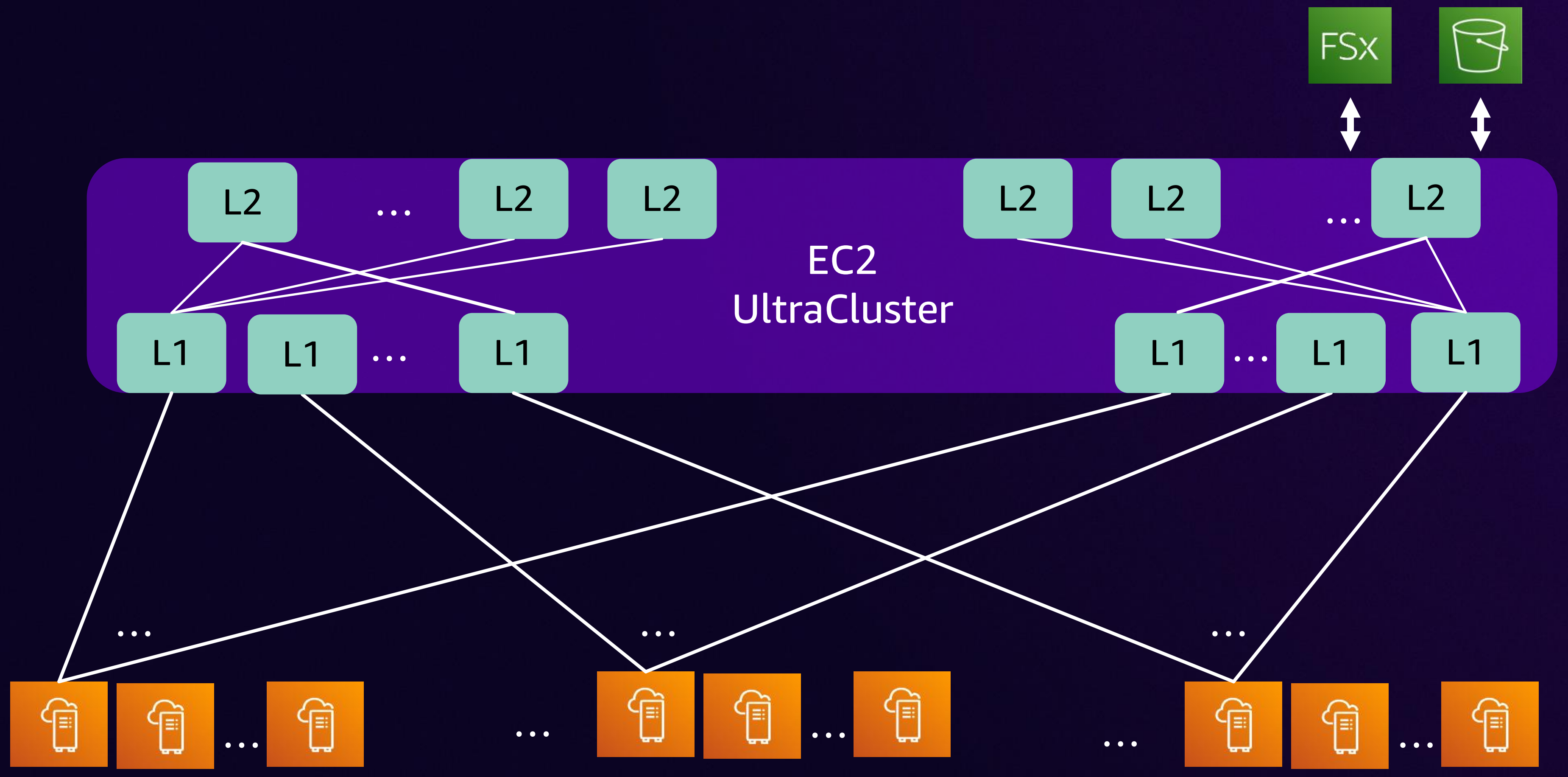


Third-generation Elastic Fabric Adapter (EFAv3)

- SRD protocol purpose-built for scalability in the cloud
- Kernel bypass and GPU-direct RDMA for low-latency; high-throughput communication between GPUs
- Continuing improvements in latency and completion times



EC2 UltraClusters



Nonblocking Pb-scale network infrastructure

Up to 100K ML chips within one EC2 ultracluster

Designed for lower latency with third-generation EFA

High-throughput, low-latency storage through Amazon FSx and Amazon S3

Industry-leading portfolio of storage services

Block

- Local instance storage
- Elastic Block Store

File

- Amazon EFS
- Amazon FSx for NetApp ONTAP
- Amazon FSx for Windows File Server
- Amazon FSx for Lustre
- Amazon FSx for OpenZFS
- Amazon file cache

Object

- Amazon S3
- Amazon S3 Glacier

Data transfer and migration

- AWS Storage Gateway
- AWS DataSync
- AWS Transfer Family
- AWS Snowball
- AWS Snowcone

Backup

- AWS Backup

aws

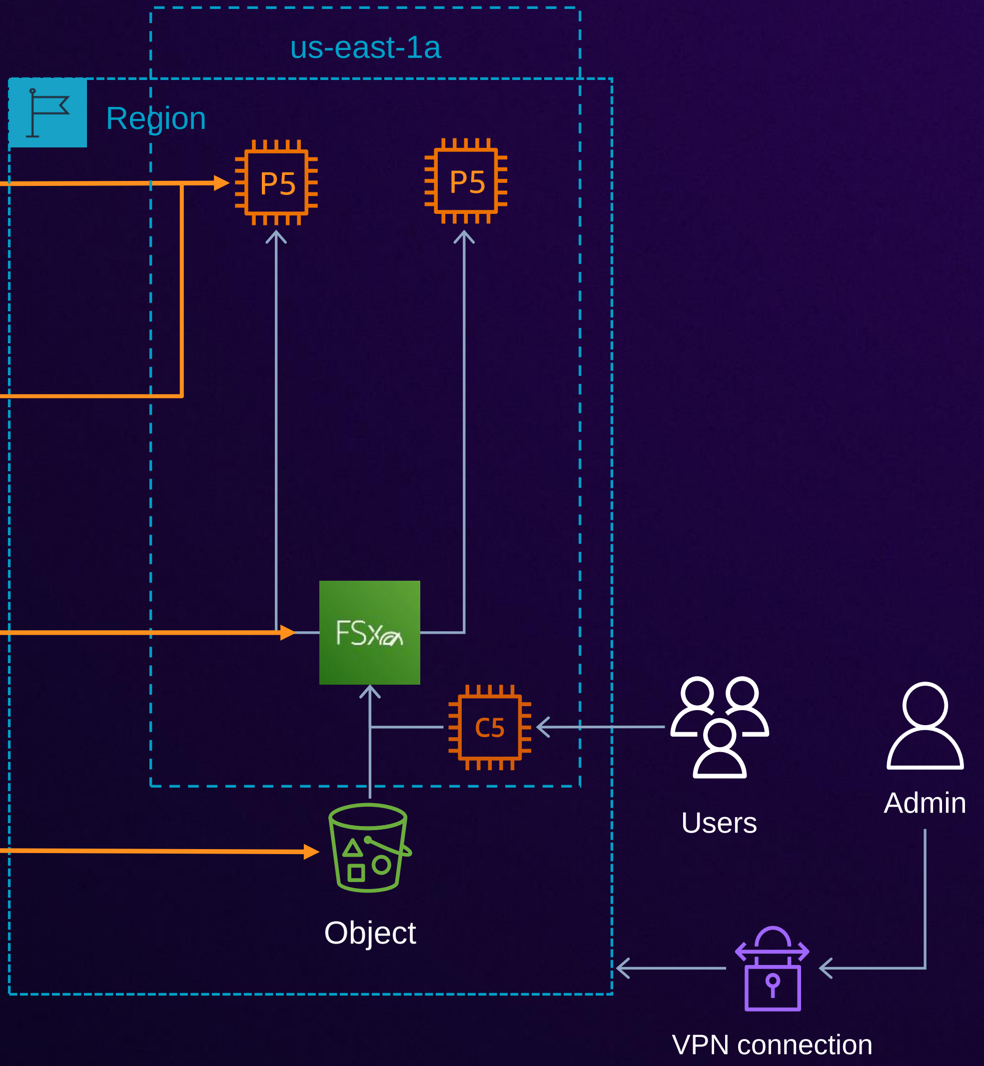
Industry-leading portfolio of storage services

Local instance storage
- Checkpoints, temporary data

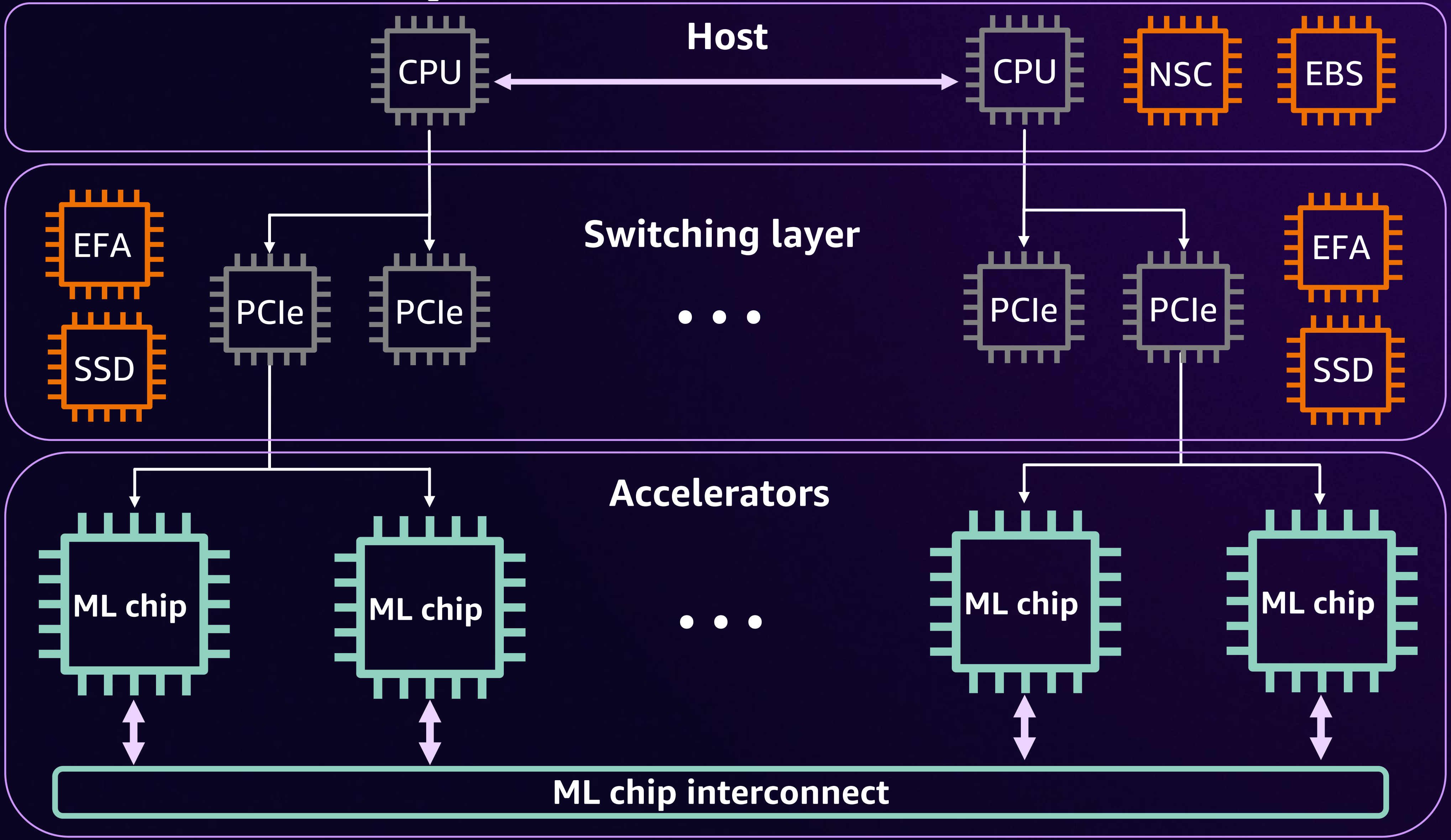
Elastic Blocks Store
- Checkpoints, temporary data

FSx for Lustre
- Shared data sets, checkpoints

Amazon S3
- Training data, checkpoints



Accelerated compute architecture



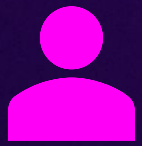
Protecting ML model weights

GOAL: PROTECT AND MAKE IP CONSUMABLE

ML model provider

Encrypted IP

- Model consumers seek to keep prompts and completions confidential
- Model providers seek to protect proprietary model weights
- Data owners seek to protect proprietary data



Data owner and model consumer

Encrypted data

AWS Nitro System

Nitro card

Local NVMe storage
Elastic Block Storage
Networking, monitoring, and security

Nitro security chip

Integrated into motherboard
Protects hardware resources

Nitro hypervisor

Lightweight hypervisor
Memory and CPU allocation
Bare metal-like performance

VPC encryption

VPC-encrypted traffic is anonymized and privacy-preserving

AWS Nitro Enclaves

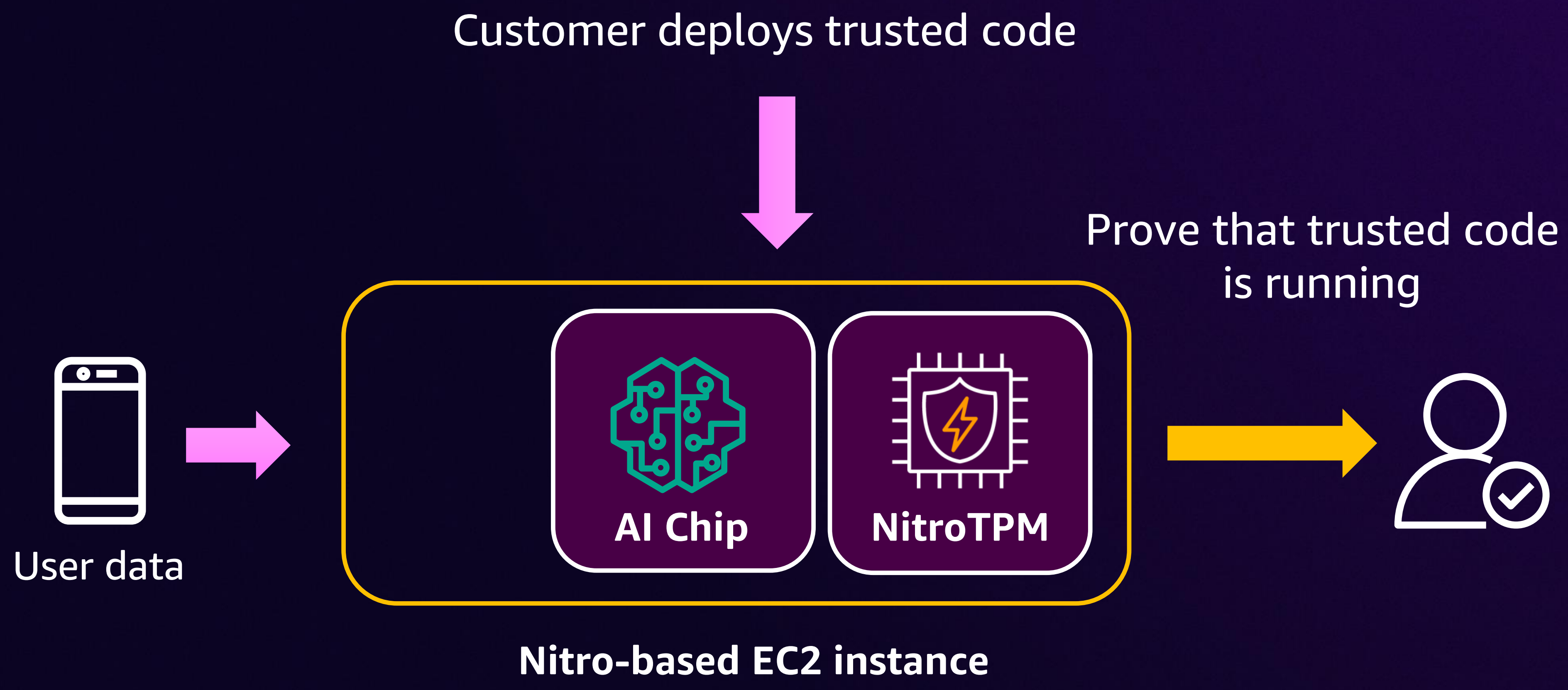
Isolated environments for highly sensitive data processing

NitroTPM

TPM 2.0 specification
Cryptographic attestation of instances integrity

Confidential Inferencing

ASSURANCE THAT USER DATA IS ONLY OPERATED IN A TRUSTED ENVIRONMENT



EC2 accelerated compute instances for AI/ML

AI/ML accelerators and ASICs

Inf1 | Inf2 | Trn1 | Trn2 | DL1 | DL2q

GPUs

G5 (A10G) | G6 (L4) | G6e (L40S) | P4 (A100) | P5 (H100) | P5e (H200) | P5en (H200) | B200 | GB200

Announced

Trainium
Inferentia

H100, H200, B200, GB200, A100, L40S, L4, A10G

Cloud AI100 Standard

Radeon GPU
Xilinx accelerator
Xilinx FPGA

Gaudi accelerator

G-series instances

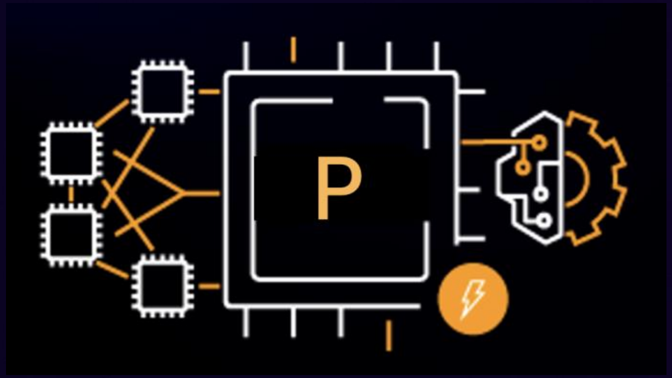
Compute and graphics optimized GPUs

Flexibility with multiple instance sizes

Great for single GPU or single node workloads

Instance	GPU	GPU memory	CPU	vCPU	Instance memory	Networking	Local storage
G6	Up to 8 NVIDIA L4	Up to 192 GB GDDR6	AMD Milan	Up to 192	Up to 768 GB	Up to 100 Gbps	Up to 7.6 TB SSD
G6e	Up to 8 NVIDIA L40S	Up to 384 GB GDDR6	AMD Milan	Up to 192	Up to 1.536 TB	Up to 400 Gbps	Up to 7.6 TB SSD

P-series instances



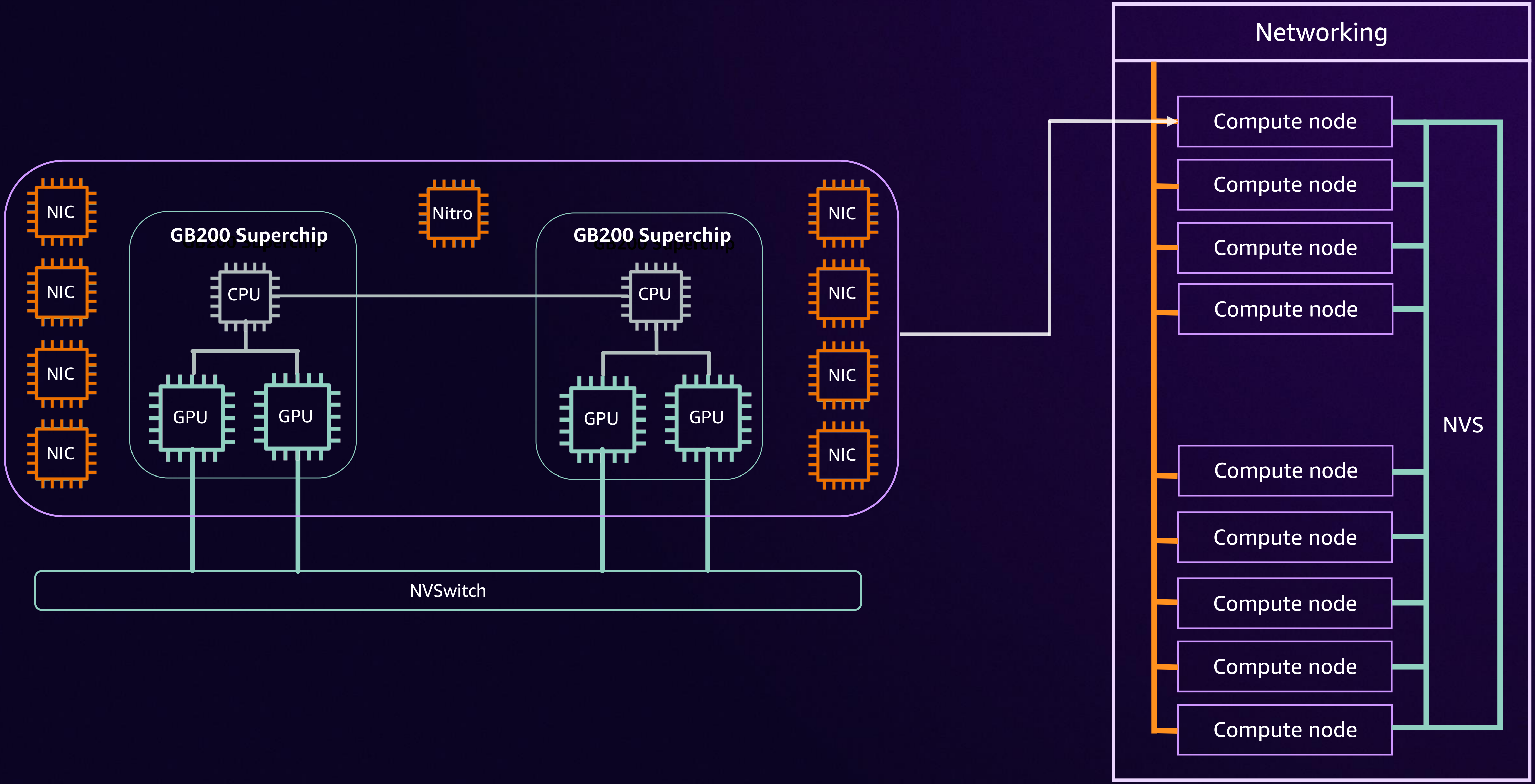
Optimized for AI training and inference

Deployed within EC2 UltraClusters for scale-out

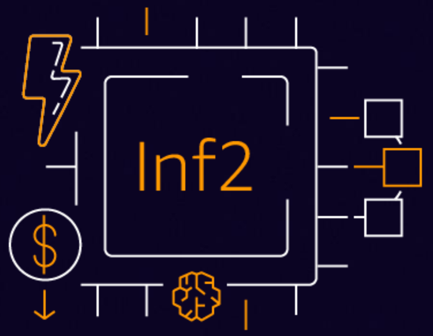
Great for single-node or distributed workloads

Instance	GPU	GPU memory	CPU	vCPU	Instance memory	Networking	Local storage
P5	8 NVIDIA H100	640 GB	AMD Milan	192	2 TB	3200 Gbps EFAv2	30 TB SSD
P5e	8 NVIDIA H200	1128 GB	AMD Milan	192	2 TB	3200 Gbps EFAv2	30 TB SSD
P5en	8 NVIDIA H200	1128 GB	Intel SPR	192	2 TB	3200 Gbps EFAv3	30 TB SSD

GB200 architecture



Inf-series instances



Powered by AWS Inferentia custom ML chips

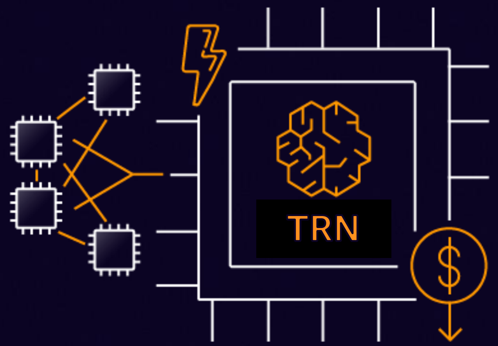
High performance at the lowest cost for generative AI models

Support for ultra-large generative AI models using NeuronLink

9.8 TB/s aggregated accelerator memory bandwidth

Instance	Accelerators	Accelerator memory	NeuronLink	vCPU	Instance memory	Networking
Inf2.xlarge	1	32 GB	N/A	4	16 GB	Up to 15 Gbps
Inf2.8xlarge	1	32 GB	N/A	32	128 GB	Up to 25 Gbps
Inf2.24xlarge	6	192 GB	Yes	96	384 GB	50 Gbps
Inf2.48xlarge	12	384 GB	Yes	192	768 GB	100 Gbps

Trn-series instances



Powered by AWS Trainium custom ML chips

Optimized for large-scale training distributed workloads

TRN2 Ultraservers with extended NeuronLink for trillion-parameter AI

Neuron Kernel Interface (NKI) for custom operators

Instance	Accelerators	Accelerator memory	vCPU	Instance memory	Networking
Trn1.32xlarge	16	512 GB	128	512 GB	800 Gbps EFAv2
Trn1n.32xlarge	16	512 GB	128	512 GB	1600 Gbps EFAv2
Trn2.48xlarge	16	1536 GB	192	2 TB	3200 Gbps EFAv3

AWS generative AI stack

Generative AI stack

APPLICATIONS THAT LEVERAGE LLMs AND FMs

Amazon Q AWS App Studio

TOOLS TO BUILD WITH LLMs AND OTHER FMs

Amazon Bedrock

Guardrails | Agents | Studio | Customization | Custom Model Import | Amazon Models

MANAGED SERVICES FOR FM TRAINING AND INFERENCE

Amazon SageMaker Amazon EKS Amazon ECS AWS Batch

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE

Trainium Inferentia GPUs

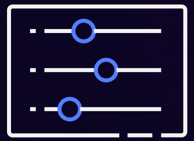
EC2 UltraClusters EFA EC2 Capacity Blocks Nitro Neuron

Amazon SageMaker Studio

FULLY INTEGRATED DEVELOPMENT ENVIRONMENT (IDE) FOR MACHINE LEARNING

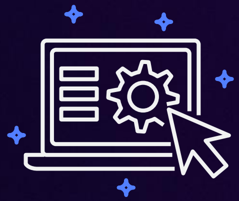
Collaboration at scale

Share notebooks without tracking code dependencies



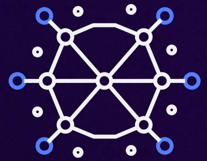
Easy experiment management

Organize, track, and compare thousands of experiments



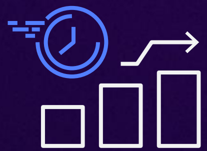
Automatic model generation

Get accurate models with full visibility and control without writing code



Higher-quality ML models

Automatically debug errors, monitor models, and maintain high quality



Increased productivity

Code, build, train, deploy, and monitor in a unified visual interface

Amazon SageMaker for Training

Fast and cost-effective ML model training

Experiment management and model tuning
Save weeks of effort by automatically tracking training runs and tuning hyperparameters

Debug and profile training runs
Use real-time metrics to correct performance problems

Distributed training
Complete distributed training up to 40% faster

Training compiler
Accelerate training times by up to 50% through more efficient use of GPUs

Managed spot training
Reduce the costs of training by up to 90%

Amazon Bedrock

The easiest way to build and scale generative AI applications with powerful tools and foundation models

Choice of leading FMs through a single API

Model customization

Retrieval Augmented Generation (RAG)

Agents that execute multistep tasks

Security, privacy, and data governance

Amazon Bedrock

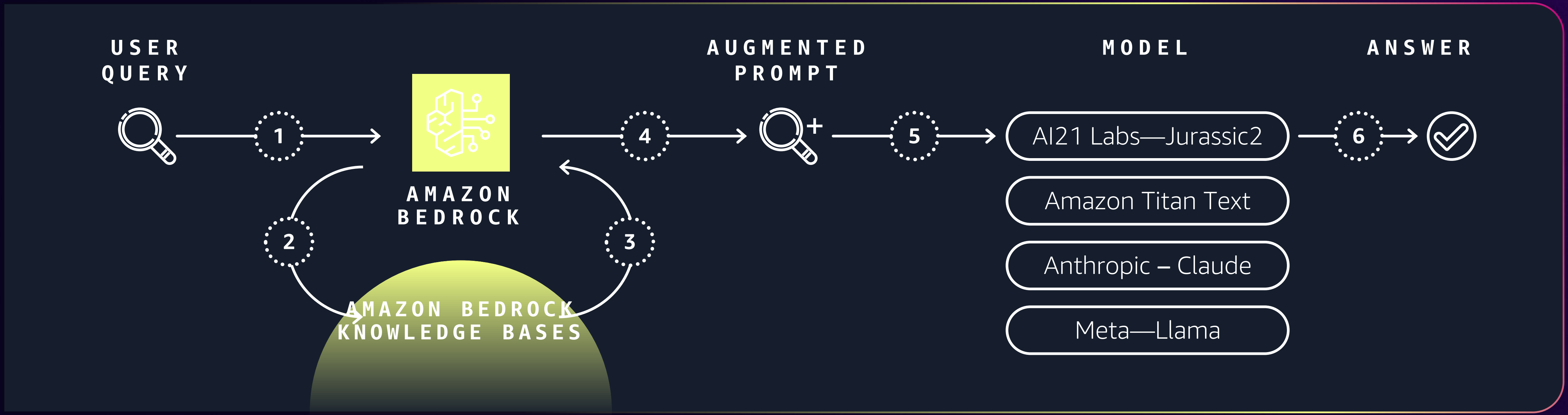
BROAD CHOICE OF MODELS

AI21 labs	amazon	ANTHROP\C	cohere	Meta	MISTRAL AI_	stability.ai
Contextual answers, summarization, paraphrasing	Text summarization, generation, Q&A, search, image generation	Summarization, complex reasoning, writing, coding	Text generation, search, classification	Q&A and reading comprehension	Text summarization, classification, completion, Q&A code generation	High-quality images and art
JURASSIC JAMBA	AMAZON TITAN	CLAUDE	COMMAND + EMBED	LLAMA	MISTRAL MIXTRAL	STABLE DIFFUSION

aws

Amazon Bedrock Knowledge Bases

NATIVE SUPPORT FOR RAG



Securely connect FMs to data sources for RAG to deliver more relevant responses

Fully managed RAG workflow, including ingestion, retrieval, and augmentation

Built-in session context management for multiturn conversations

Automatic citations with retrievals to improve transparency

Meta

CMP 207

Accelerating Wearables Multimodal AI on AWS

Kirmani Ahmed
(He/Him)

AI Engineering Lead
Meta Wearables AI

Ray-Ban Meta


WIP
RAY-BAN META WAYFARER ESSILORLUXOTTICA LTD 0001/7500
TC 00:00:40:14
Kühl & Han®

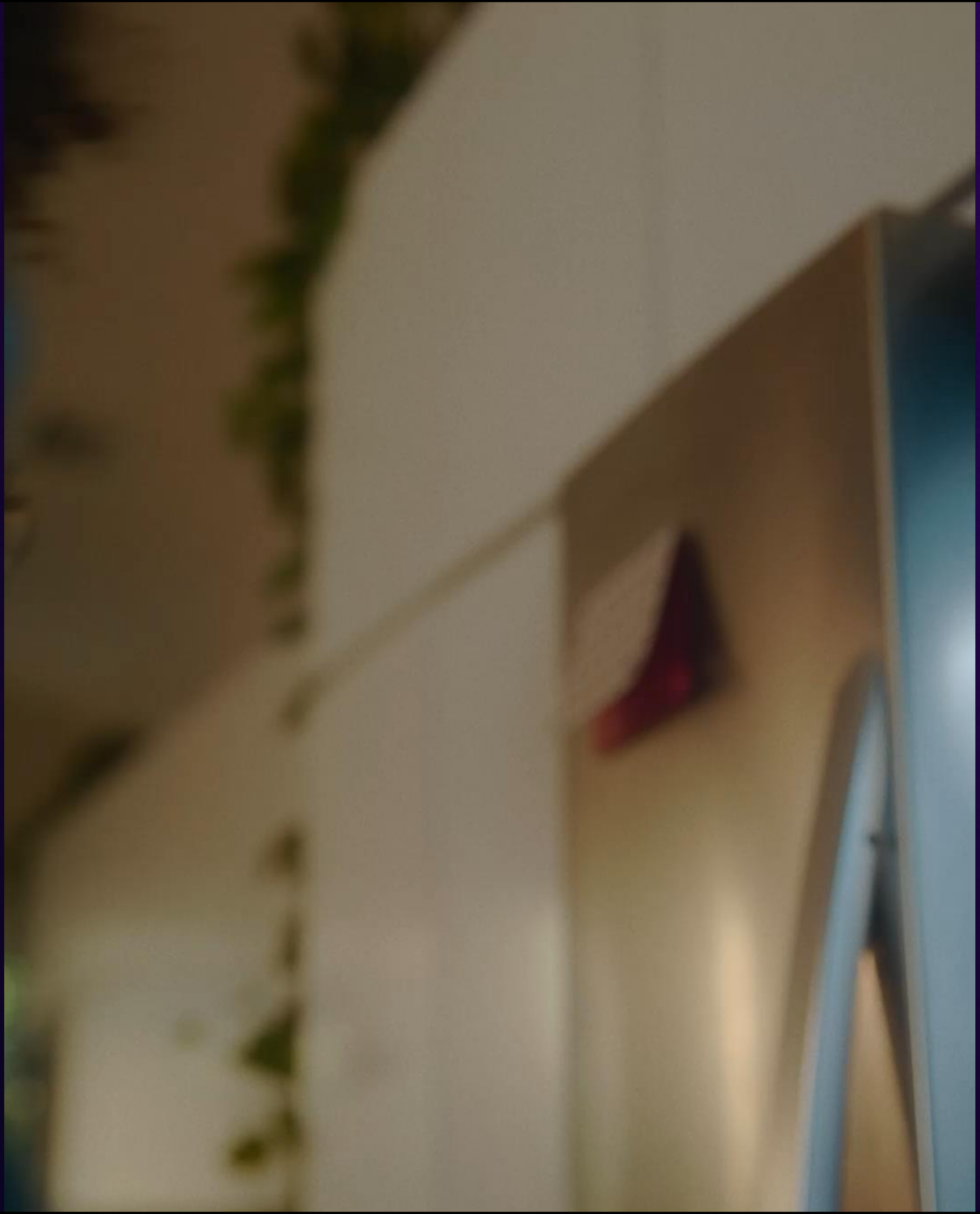
Ray-Ban Meta

Ray-Ban **Meta AI**

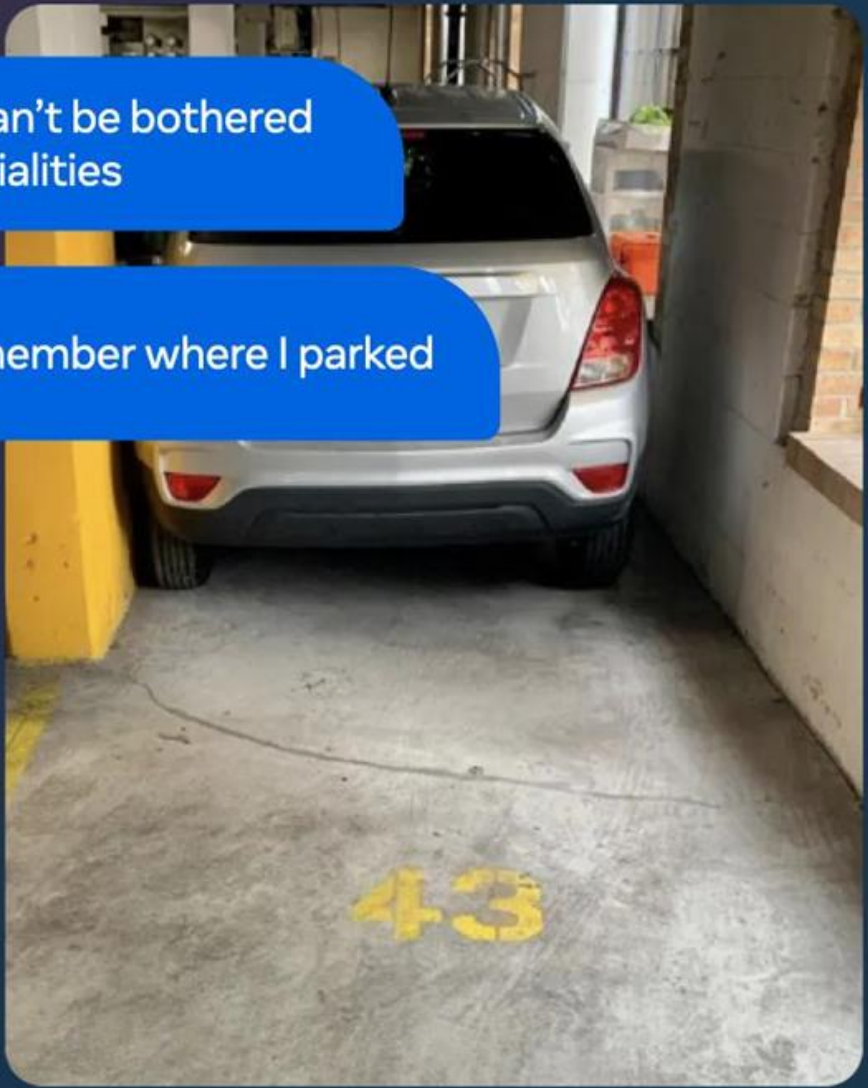
Launched in April ’24

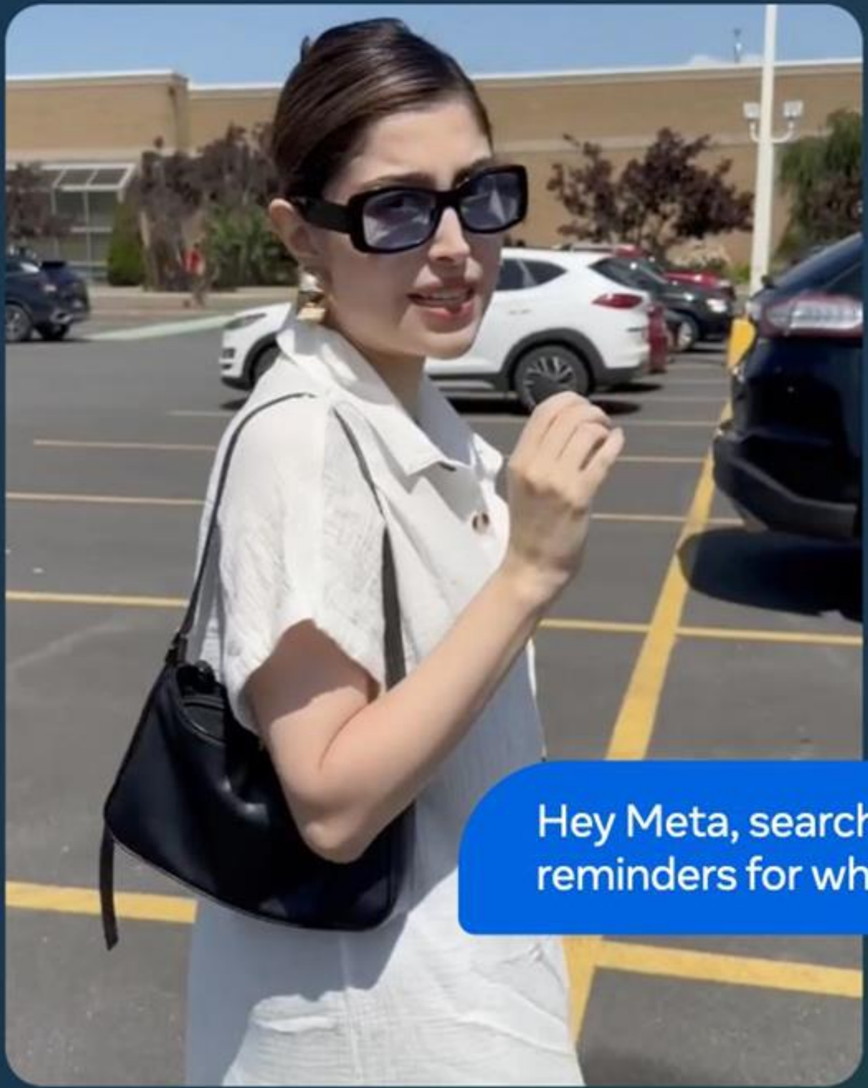
“Meta's Ray-Ban Smart Glasses are Better Than We Thought”

“Meta's Ray-Bans are a turning point”



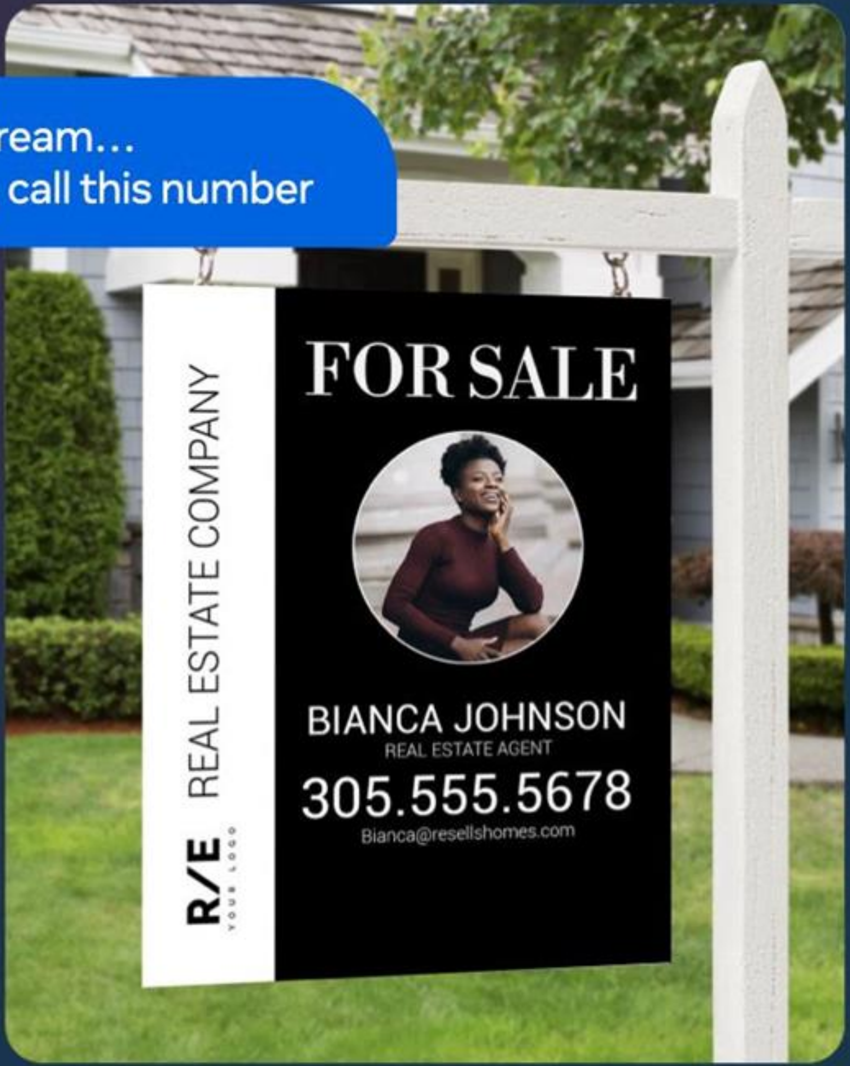
Launched at Connect ’24





Launched at Connect '24

One can dream…
Hey Meta, call this number



Hey Meta, scan this QR code

How did we Move Fast to get here?

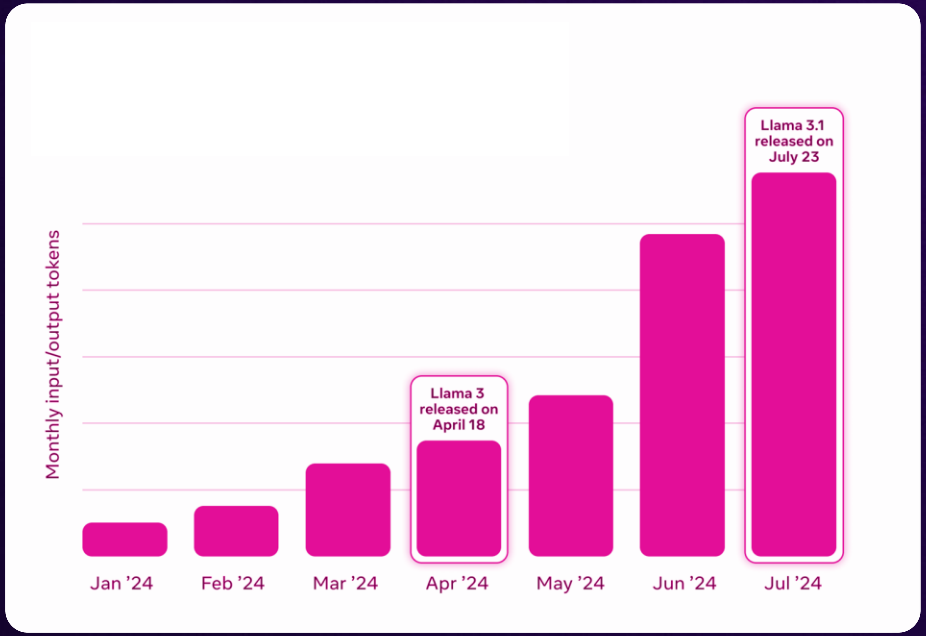
It all starts with **LLama**

"A Giant Leap in Open Source AI"

Feb '23

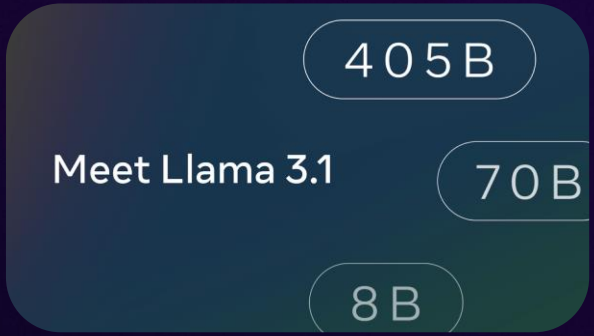
July '23

April '24

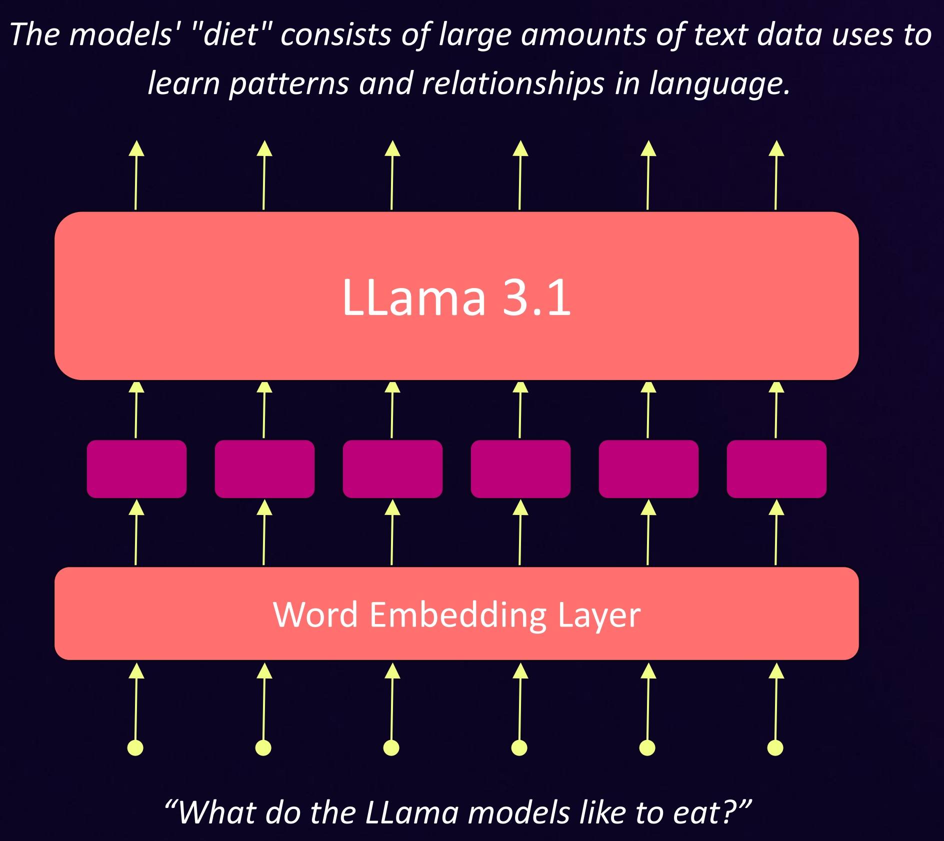


LLama is the leading open source model family

- 350 million downloads to date (10x more YoY)
- 20 million times in the last month alone
- **LLama turbocharged businesses & the Cloud AI ecosystem (10x YoY token volume)**



Quick primer on **LLama**



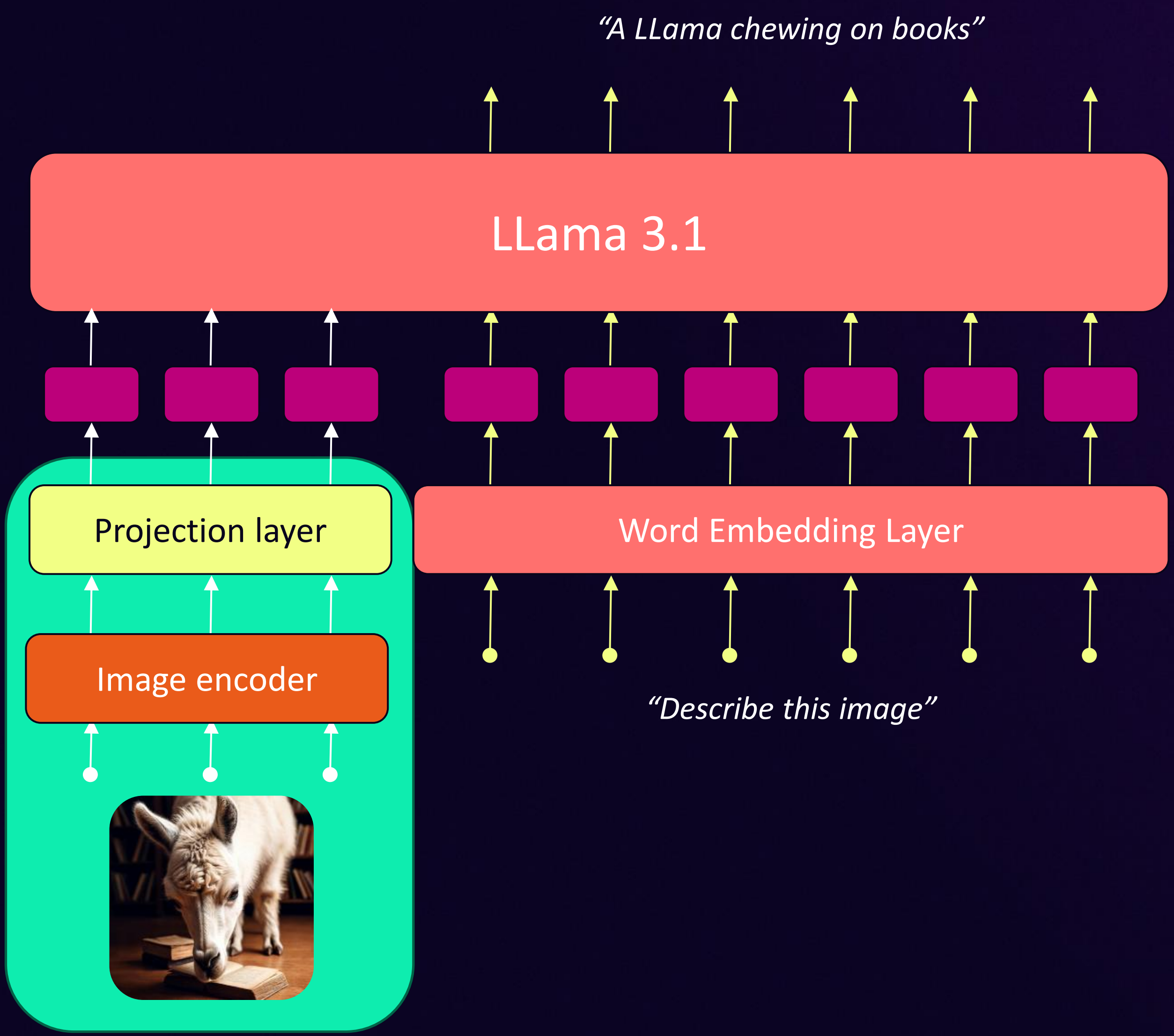
- LLama is a large language model family (8B, 70B, 405B)
- **Text in ⇒ Text out**
- Efficiently trained using 15T tokens, 24K GPUs, 400 TFLOPS/GPU

How can we **make LLMs see (and hear)?**

Chatbot

Multimodal

Key insight: **making LLMs see (and hear)**



Multimodal LLMs

- Text + Image/Audio in ⇒ Text out
- Need a "modality encoder" to convert images into LLM tokens
- Trained on billions of (Image, Text) data pairs

Multimodal LLMs: How to build one

Flamingo: a Visual Language Model for Few-Shot Learning

Jean-Baptiste Alayrac*,‡ Jeff Donahue* Pauline Luc* Antoine Miech*
Iain Barr† Yana Hasson† Karel Lenc† Arthur Mensch† Katie Millican†
Malcolm Reynolds† Roman Ring† Eliza Rutherford† Serkan Cabi Tengda Han
Zhitao Gong Sina Samangooei Marianne Monteiro Jacob Menick
Sebastian Borgeaud Andrew Brock Aida Nematzadeh Sahand Sharifzadeh
Mikolaj Binkowski Ricardo Barreira Oriol Vinyals
Karen Simonyan*,‡

* Equal contributions, ordered alphabetically, † Equal contributions,
‡ Equal senior contributions

DeepMind

ADEPT

Announcements Research

Fuyu-8B: A Multimodal Architecture for AI Agents

ohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell
a, Arushi Somani, Sağnak Taşırlar

cing Fuyu-8B - a small version of the
el that powers our product.

CogVLM: Visual Expert for Pretrained Language Models

Weihan Wang [*1‡] Qingsong Lv [*2] Wenmeng Yu [2] Wenyi Hong [1‡] Ji Qi [1‡] Yan Wang [2] Junhui Ji [2]
Zhuoyi Yang [1‡] Lei Zhao [2] Xixuan Song [1‡] Jiazheng Xu [1‡] Keqin Chen [3‡] Bin Xu [1] Juanzi Li [1] Yuxiao Dong [1]
Ming Ding [2] Jie Tang [1]

Abstract

We introduce CogVLM, a powerful open-source visual language foundation model. Different from the popular *shallow alignment* method which maps image features into the input space of language model, CogVLM bridges the gap between the frozen pretrained language model and image encoder by a trainable visual expert module in the attention and FFN layers. As a result, CogVLM enables a deep fusion of vision language features without sacrificing any performance on NLP tasks. CogVLM-17B achieves state-of-the-art performance on 17 classic cross-modal benchmarks, including 1) image captioning datasets: NoCaps, Flicker30k, 2) VQA datasets: OKVQA, TextVQA, OCRVQA, ScienceQA, 3) LVLM benchmarks: MM-Vet, MMBench, SEED-Bench, LLaVABench, POPE, MMMU, MathVista, 4) visual grounding datasets: RefCOCO, RefCOCO+, RefCOCOg, Visual7W. Codes and checkpoints are available at https://github.com/THUDM/CogVLM.

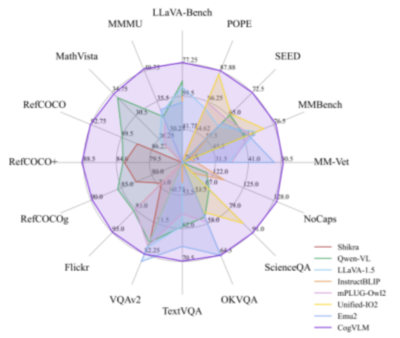


Figure 1. The performance of CogVLM on a broad range of multi-modal tasks in comparison with existing approaches.

Multimodal LLMs: **How to build one *fast***

Accelerating **AI Explorations** into **Insights**

Top cloud compute needs to *Move Fast*

1. “Bare-metal” software platform
2. Highly Reliable & Available
3. High Performance-to-Cost ratio

Building Meta
Wearables Multimodal AI

The Multimodal Recipe

1. Picking the right architecture
2. Training a base model
3. **Scaling the model (the hardest)**

Step 1

Multimodal LLMs: Picking the right architecture

	Cross-attention	Decoder only
Modeling approach	Connect image patches with multi-head attention	Feed image embeddings as input alongside Text
Training complexity	Needs changes to LLM (unfrozen LLMs hard to control)	No changes to LLM (frozen during training)
Computational complexity	Superior computational efficiency for hi-res images	Overloads input context with image tokens
Quality	Higher accuracy for complex reasoning tasks	Achieves higher accuracy in OCR-related tasks

OSS variants

- LLaVa
- Fuyu
- Qwen2-VL
- Pixtral
- Molmo
  MM1.5
- Baichuan
- CogVLM
- …

Key AWS win: The agile SW stack allowed rapid ingestion & evaluation of OSS models and benchmarks

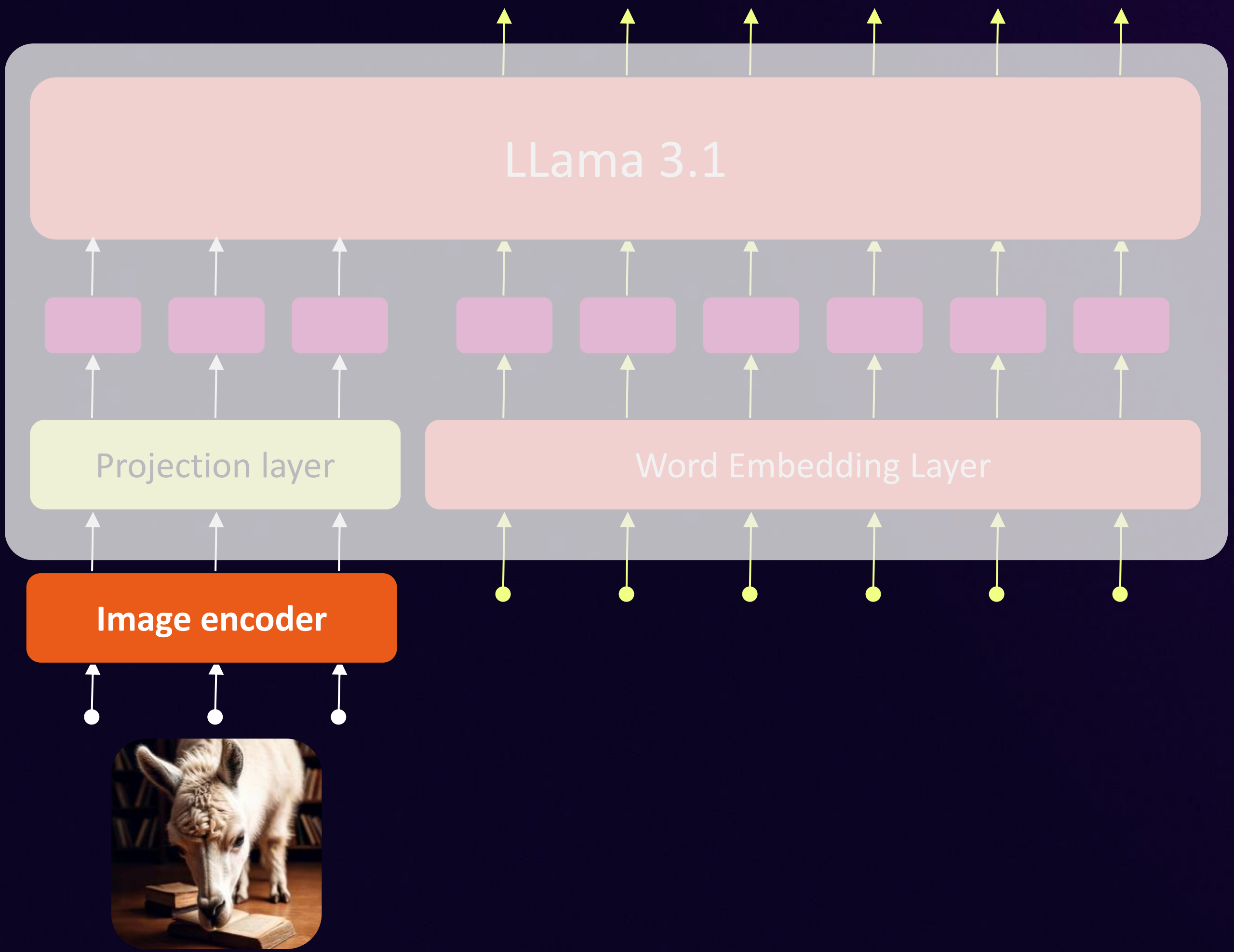
AnyMAL
Any-Modality Augmented Language Model

Shane Moon*, Andrea Madotto*, Zhaojiang Lin*, Tushar Nagarajan*,
Matt Smith, Shashank Jain, Chun-Fu Yeh, Prakash Murugesan,
Peyman Heidari, Yue Liu, Kavya Srinet, Babak Damavandi, Anuj Kumar

(* First Authors)

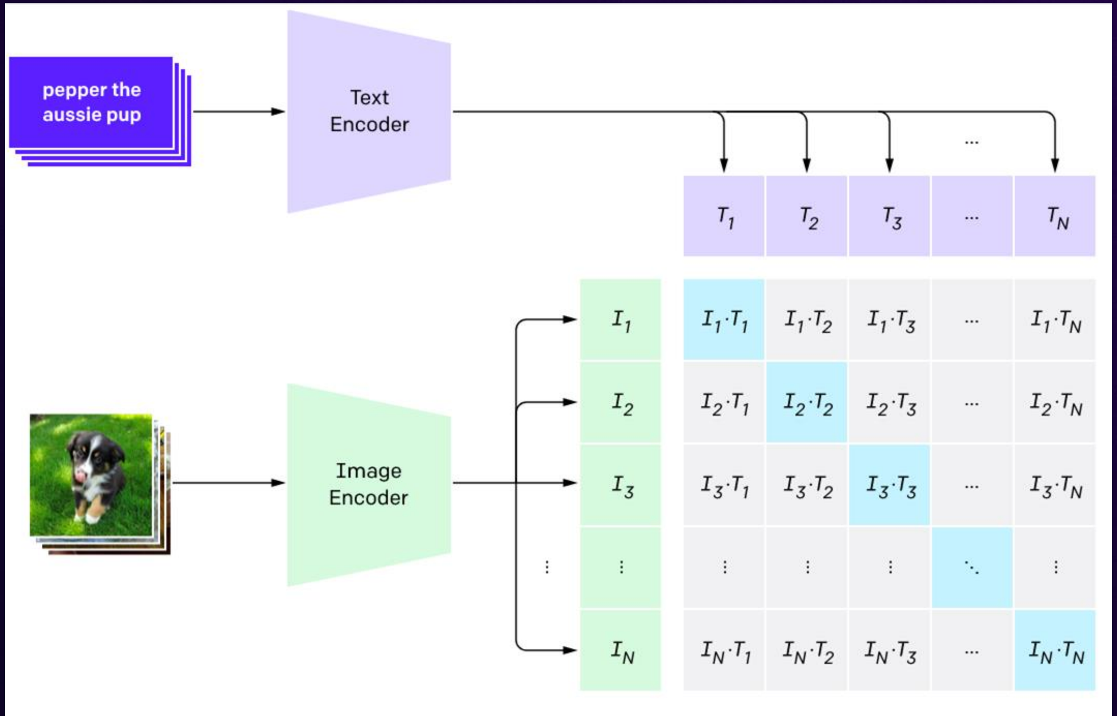
Step 2

Training AnyMAL model: **Vision encoder**



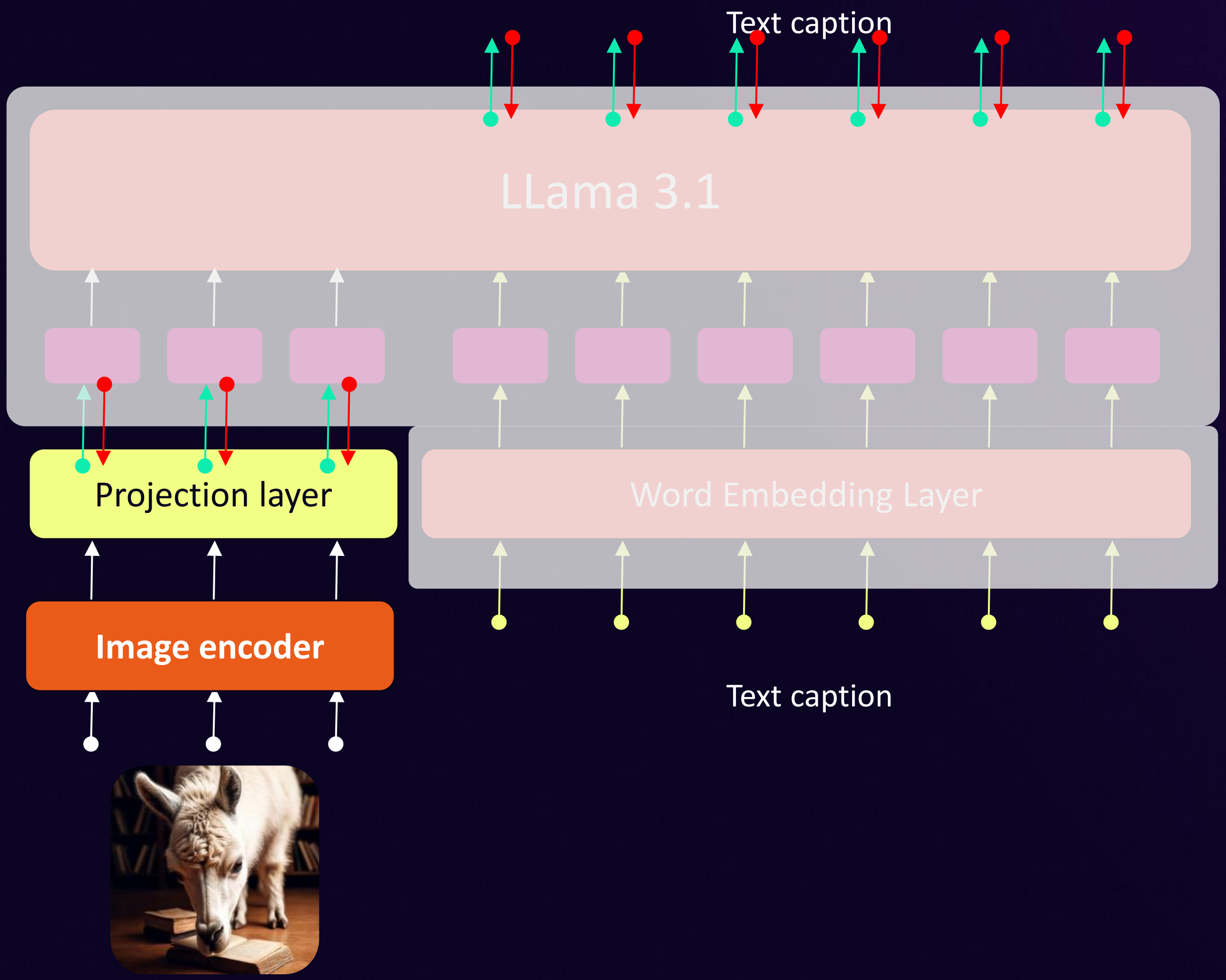
Training Modality Encoder

Trained with MM Contrastive loss (text & other modality) for the best alignment in the text space



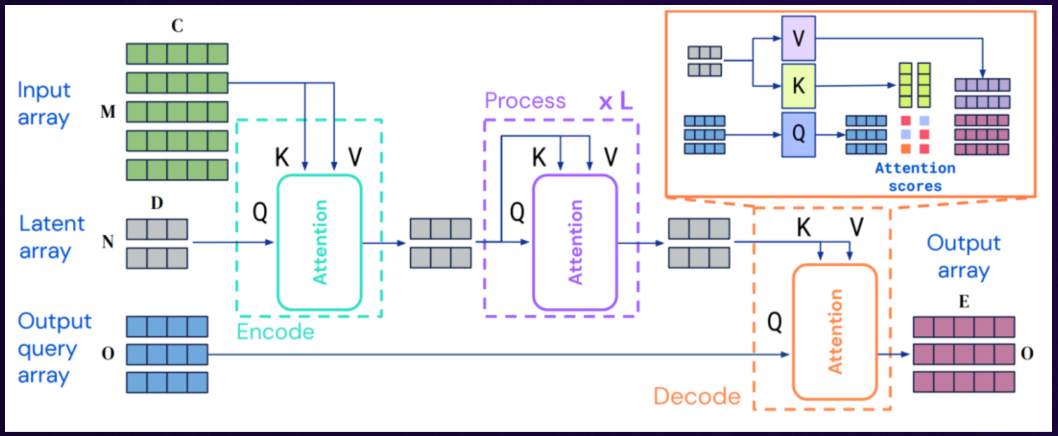
*OpenAI, "**CLIP**: Connecting Text and Images", 2021*
*Moon et al., "**IMU2CLIP**", EMNLP, 2023*

Training AnyMAL model: **Projection layer**

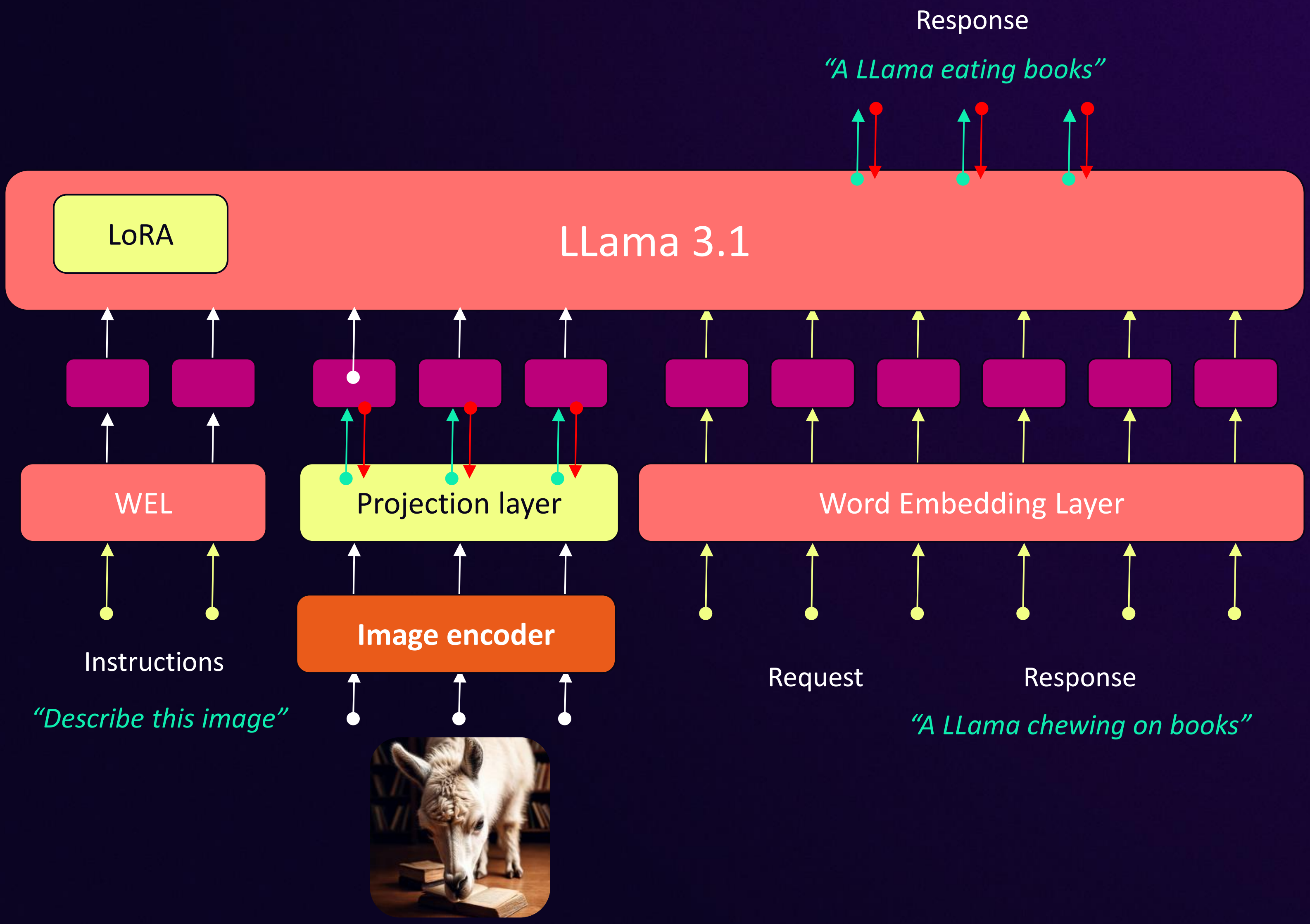


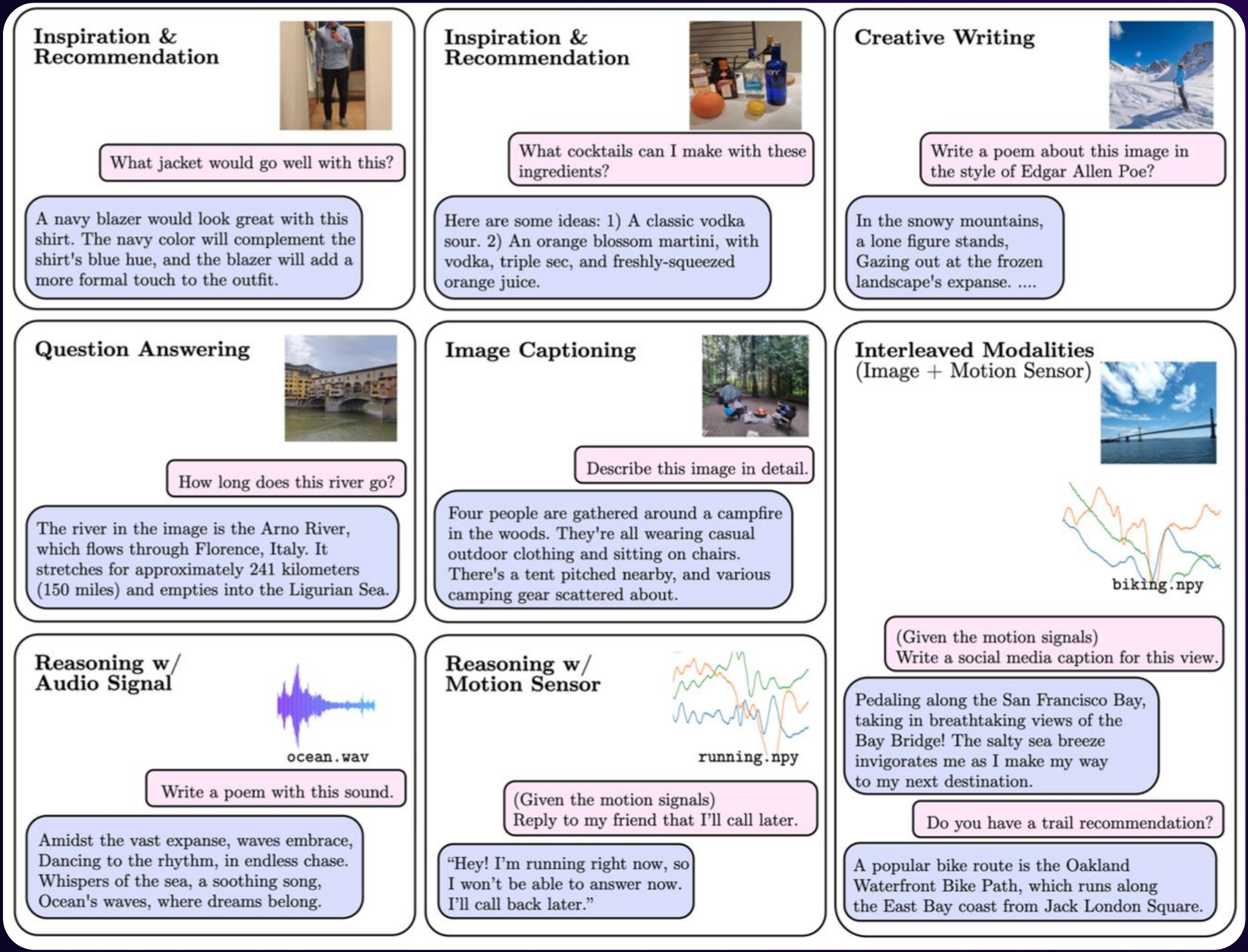
Projection Layers

Perceiver Resampler to resample patch embeddings into a sequence of LLama-compatible tokens

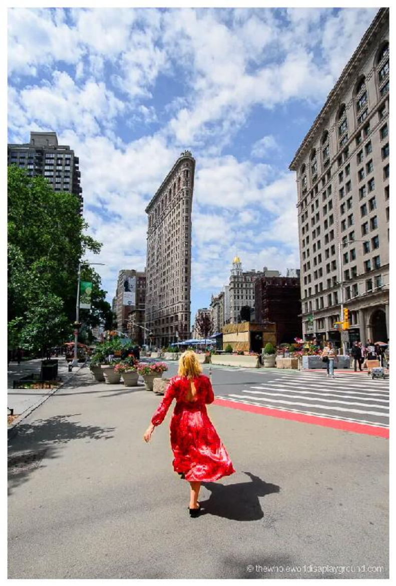


Training AnyMAL model: Instruction following



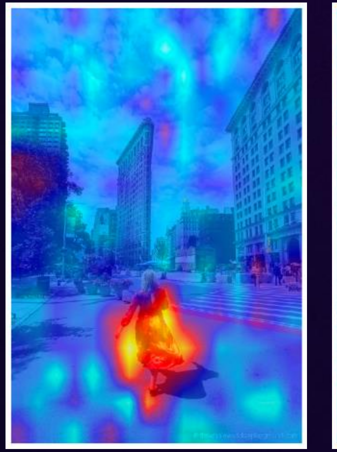

Inspiration & Recommendation
What jacket would go well with this?
A navy blazer would look great with this shirt. The navy color will complement the shirt's blue hue, and the blazer will add a more formal touch to the outfit.
Inspiration & Recommendation
What cocktails can I make with these ingredients?
Here are some ideas: 1) A classic vodka sour. 2) An orange blossom martini, with vodka, triple sec, and freshly-squeezed orange juice.
Creative Writing
Write a poem about this image in the style of Edgar Allen Poe?
In the snowy mountains, a lone figure stands, Gazing out at the frozen landscape's expanse. ....
Question Answering
How long does this river go?
The river in the image is the Arno River, which flows through Florence, Italy. It stretches for approximately 241 kilometers (150 miles) and empties into the Ligurian Sea.
Image Captioning
Describe this image in detail.
Four people are gathered around a campfire in the woods. They're all wearing casual outdoor clothing and sitting on chairs. There's a tent pitched nearby, and various camping gear scattered about.
Interleaved Modalities (Image + Motion Sensor)
biking.npy
(Given the motion signals) Write a social media caption for this view.
Pedaling along the San Francisco Bay, taking in breathtaking views of the Bay Bridge! The salty sea breeze invigorates me as I make my way to my next destination.
Do you have a trail recommendation?
A popular bike route is the Oakland Waterfront Bike Path, which runs along the East Bay coast from Jack London Square.
Reasoning w/ Audio Signal
ocean.wav
Write a poem with this sound.
Amidst the vast expanse, waves embrace, Dancing to the rhythm, in endless chase. Whispers of the sea, a soothing song, Ocean's waves, where dreams belong.
Reasoning w/ Motion Sensor
running.npy
(Given the motion signals) Reply to my friend that I'll call later.
"Hey! I'm running right now, so I won't be able to answer now. I'll call back later."

What does the model see?



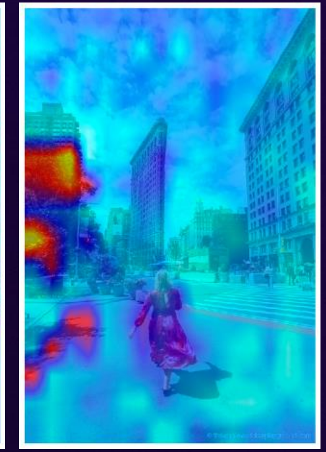
input image

What is the name of that hotel?

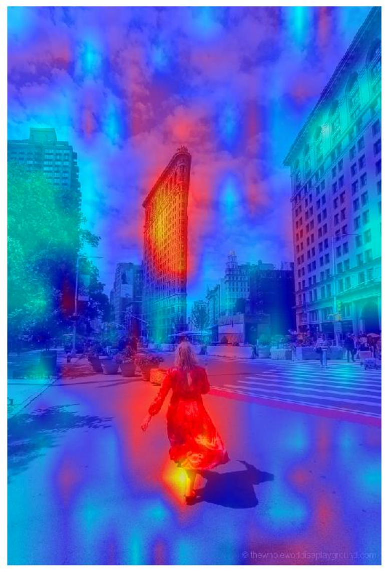


women in red dress

how is the weather?



What is that tree?

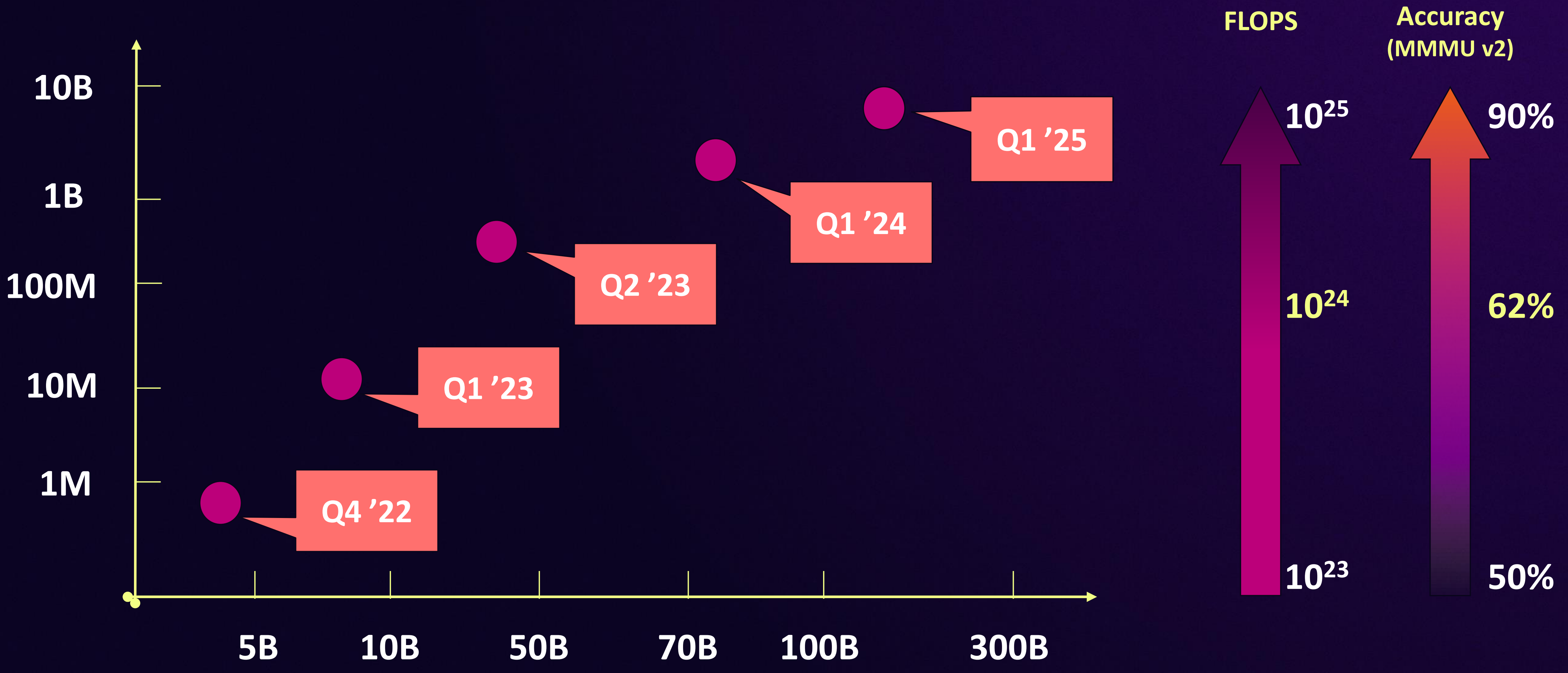


Saliency detection (this|that)

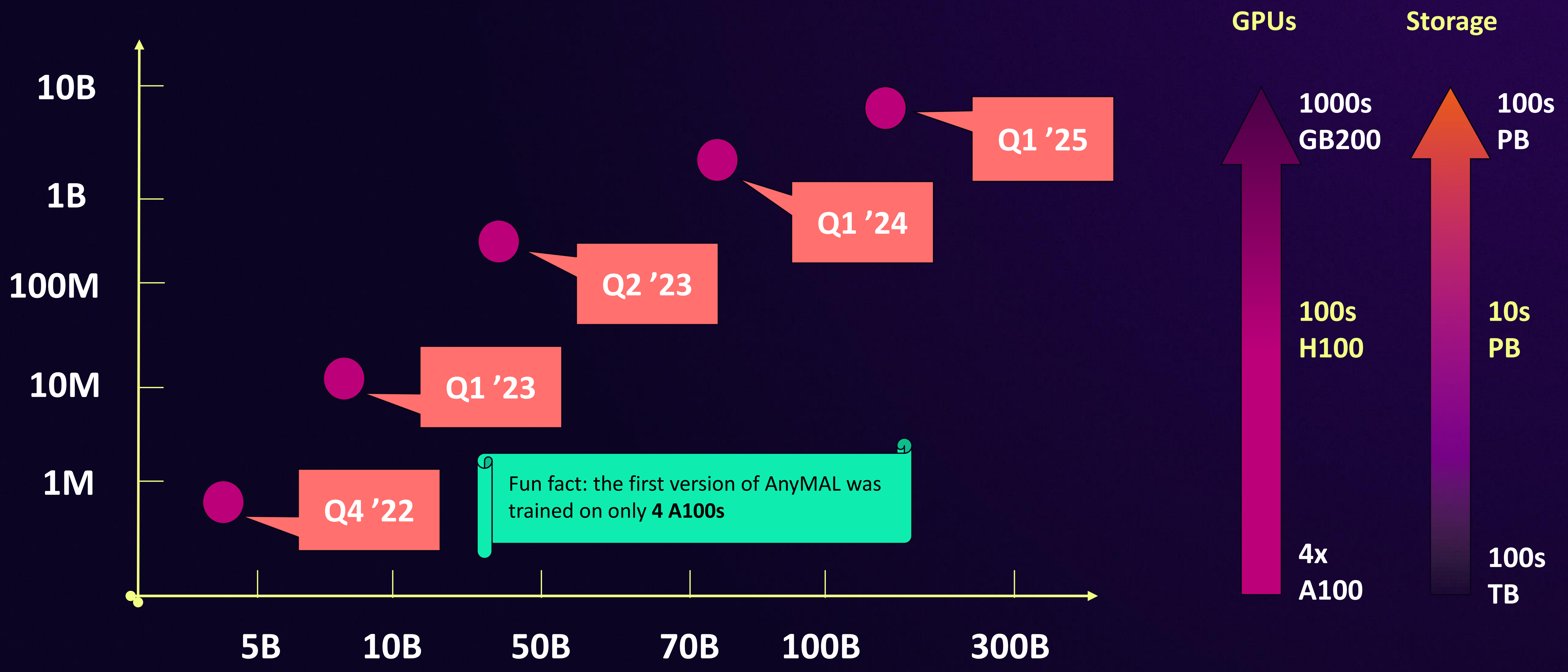
Challenges & Lessons:

Scaling training on AWS Cluster

Model Scaling - Size and Data volume



Model Scaling - Cloud compute needs



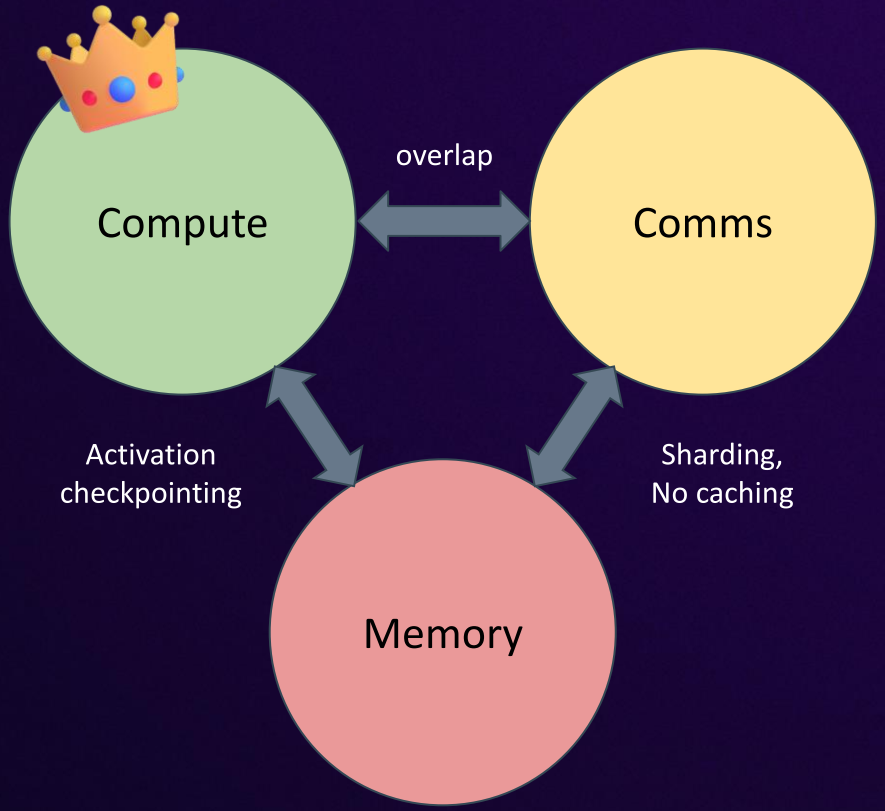
Cloud compute challenges

1. Reliability
2. Scalability
3. Efficiency

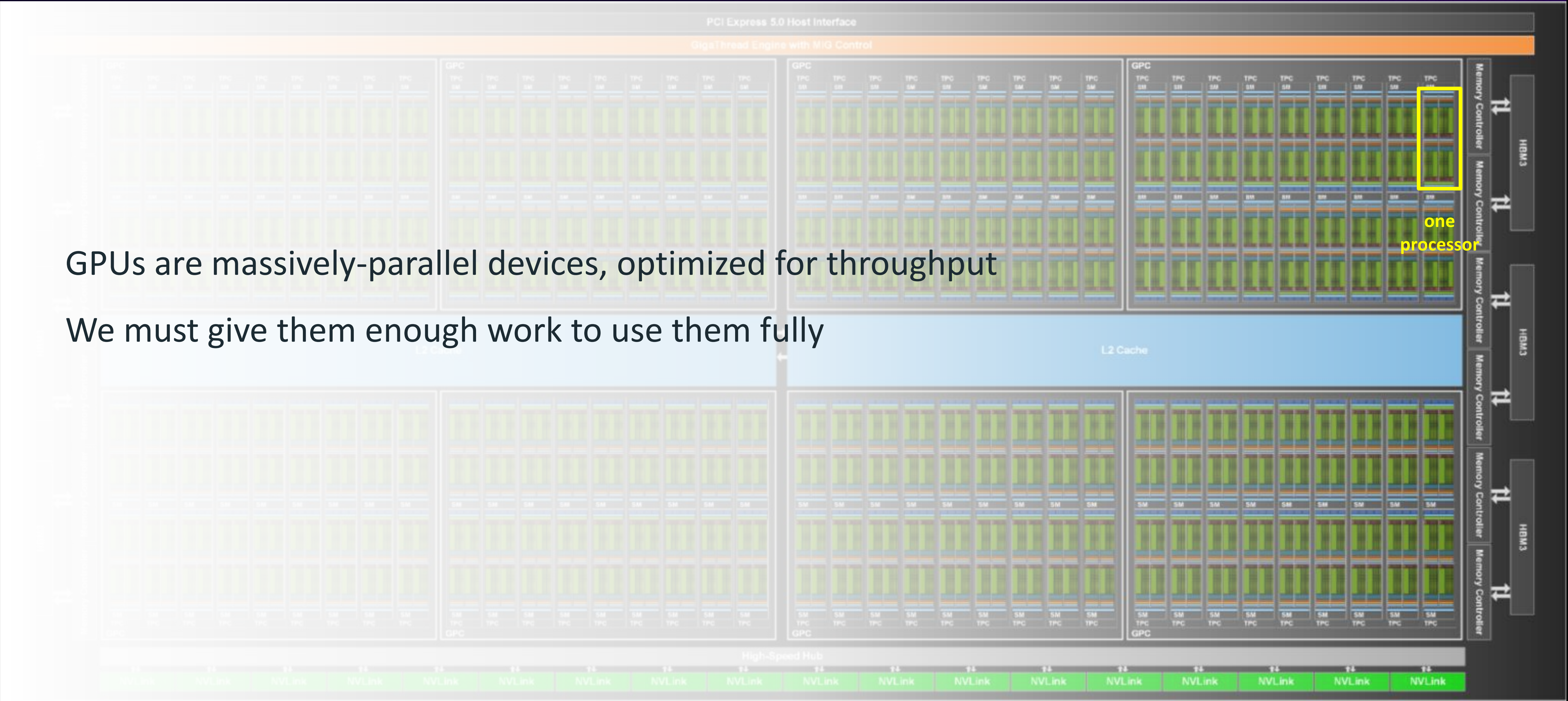
"Always be computing"

- every AI engineer

- Memory is a hard limit
- Comms can be free (after enough effort)
- **Compute is King**



Always be computing



GPUs are massively-parallel devices, optimized for throughput

We must give them enough work to use them fully

Cloud compute challenge #1 - **Reliability**

3 main issues:

1. Flaky large scale file system
2. Faulty nodes
3. Mismatched SW dependencies

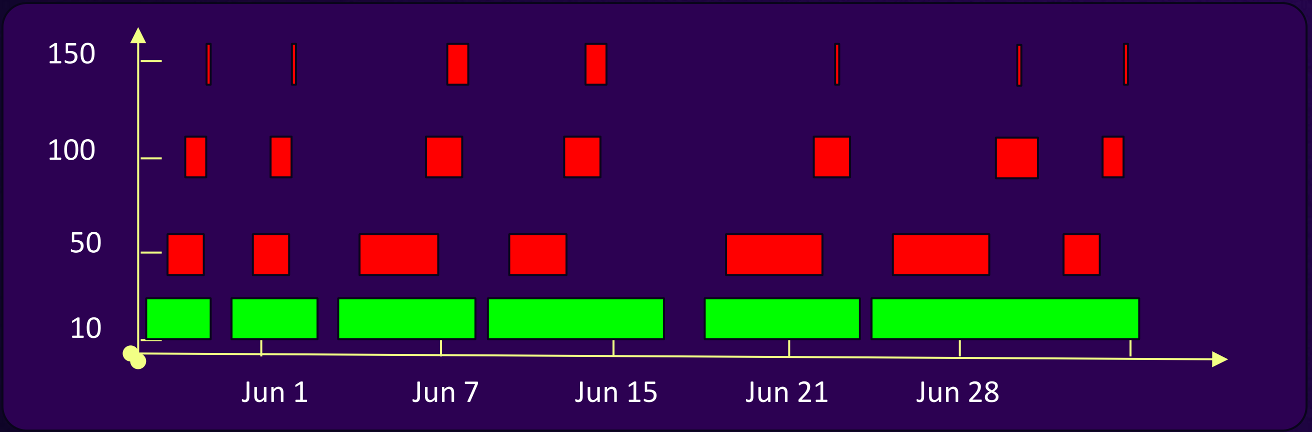
Mitigations:

HW fixes, Tooling update, training workarounds

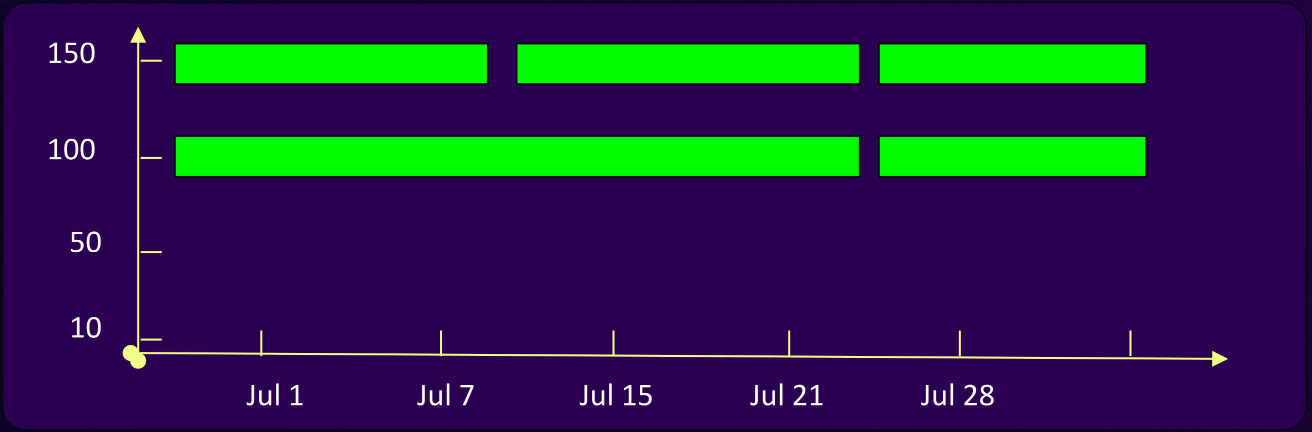
Outcome:

Cluster utilization improved from ~10% → 200+%

Key win: AWS and Meta HPC teams



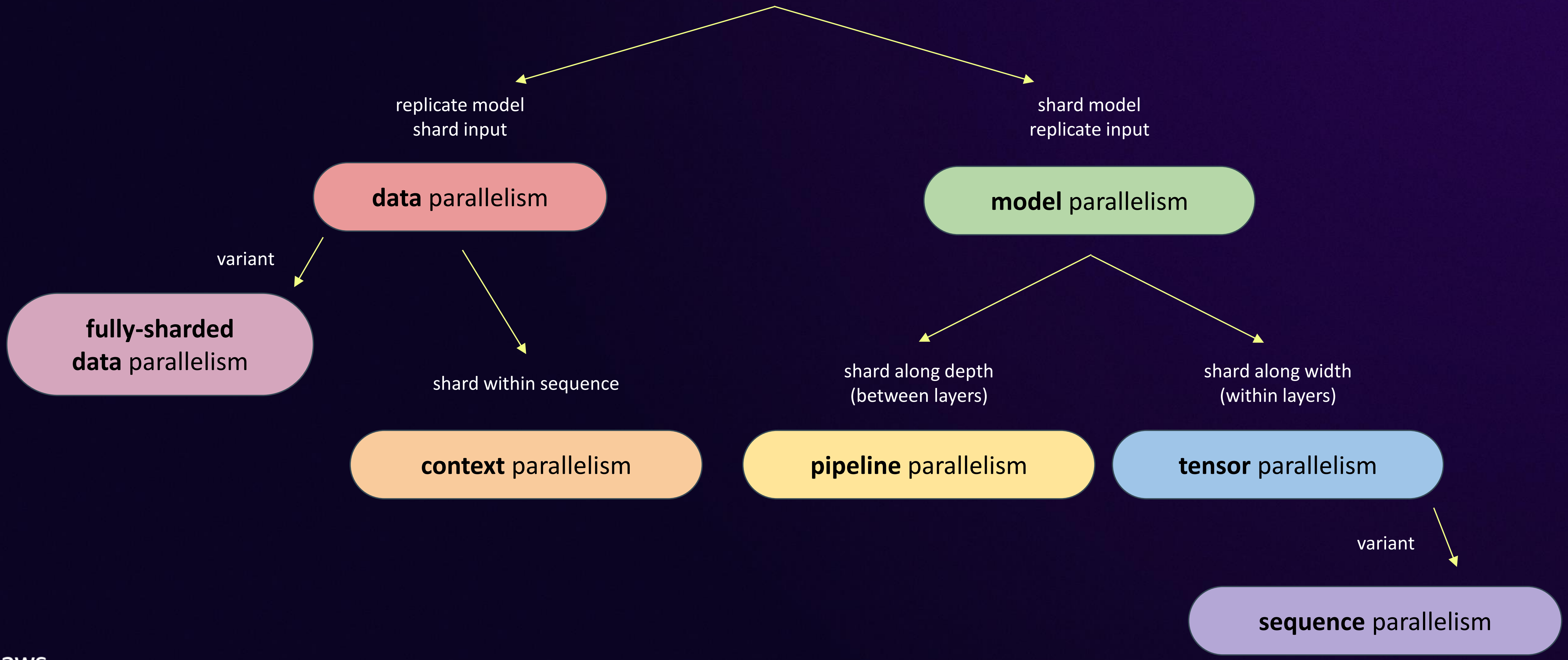
All jobs > 50 nodes failed



Consistently running jobs on 100+ nodes

Cloud compute challenge #2 - Scaling Training

“Parallelize, Parallelize, Parallelize (compute and comms)”



Cloud compute challenge #2 - Scaling Training

- **FSDP (TBD speed up)**. Sharding model over data-parallel workers:
  - LLama layers
  - PerceiverIO layers
  - CLIP layers
  - LoRA weights

- **Tensor Parallelism**
  - 4-way TP reduced training & inference time by 61%

- **Pipeline Parallel, Context Parallel**
  - Increase context length to 128K tokens

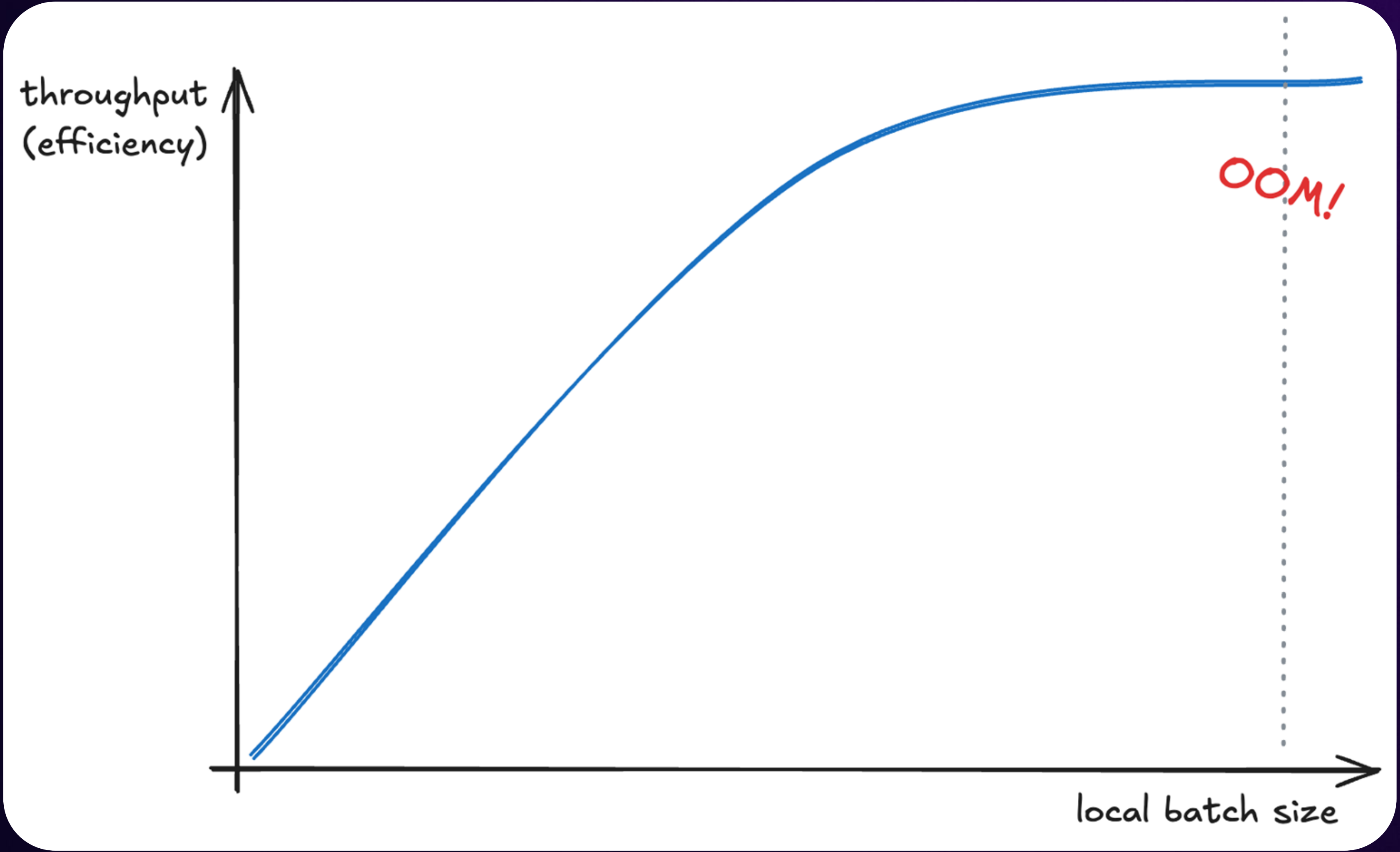
- **Quantization (TBD speed up)**
  - 8 bits & 4 bits (fits 70B model with 8 x 80GB GPUs w/ bsz 4)

Cloud compute challenge #3 - **Efficiency**

Key insight #1:

Efficiency sweet spot

If the global batch size stays fixed, throwing more GPUs at the training will make it less efficient



Cloud compute challenge #3 - **Efficiency**

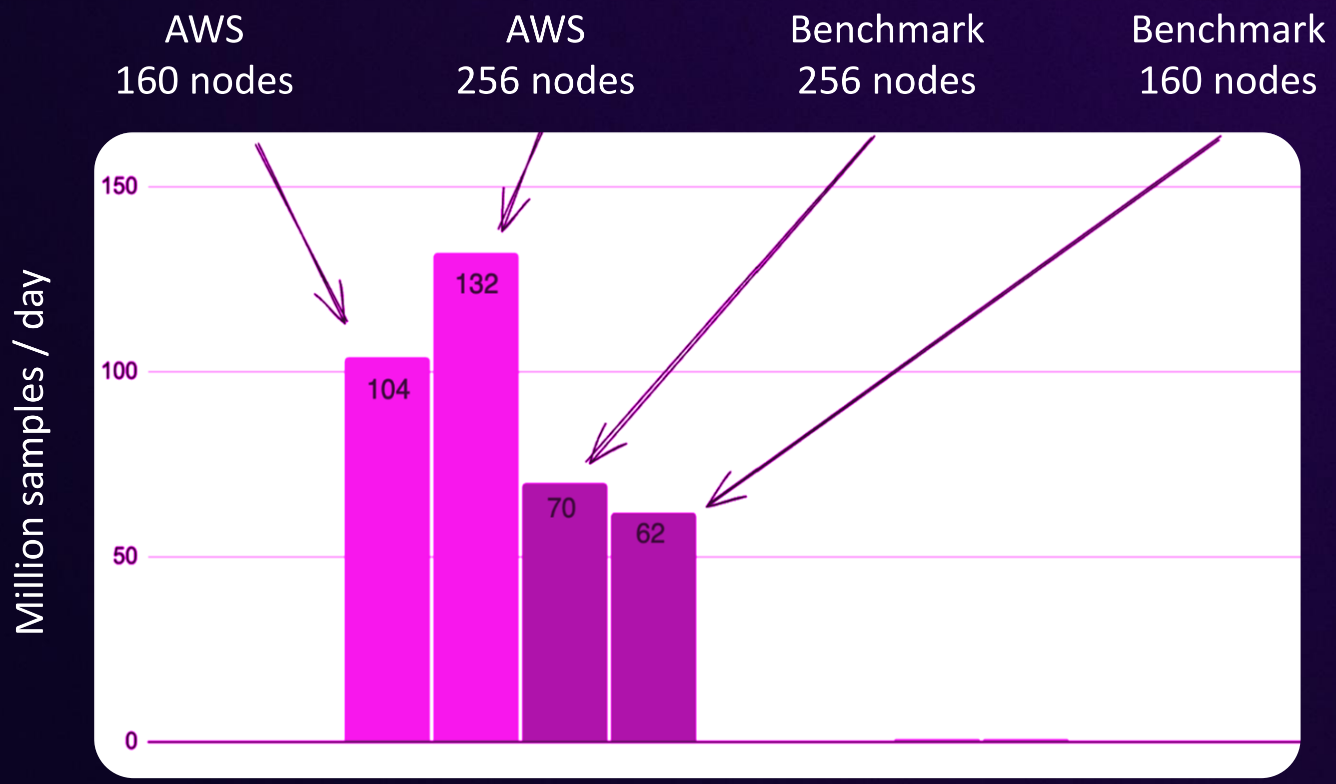
Key insight #2: Inter-GPU Comms sweet spot

The AWS H100 SXM Interconnect b/w is **900 GB/s vs. 600 GB/s** for other Cloud providers.

Outcome:

AWS training speed is ~2x faster than competing cloud benchmarks.

Training throughput - AWS vs. benchmark



Future needs for Multimodal AI

2025+ Look ahead for Multimodal AI

1. Scaling model size (100B → 1T)
2. Support very long context lengths
3. Scaling inference to support user growth

2025+ Look ahead for Multimodal AI

Key levers

1. Fast experimentation (e.g. MoE)
2. Scale compute: H100 → GB200
3. Faster Inference and learning from user feedback

Acknowledgments

AWS team

Meta HPC team

Wearables AI Eng team

Thank you!

Please complete the session survey in the mobile app