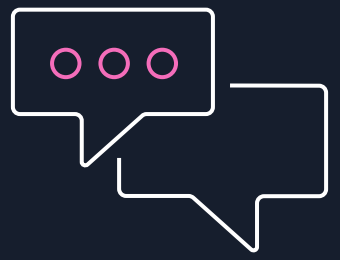
Challenges of training and deploying foundation models at scale

Maxime Hugues Ph.D.
Principal Applied Scientist GenAI

2025-06-16

Generative AI has potential business value



NEW EXPERIENCES

Create new innovative and engaging ways of interacting with your customers and employees

PRODUCTIVITY

Radically improve productivity across all lines of business

INSIGHTS

Extract insights and clear answers from all your corporate information, enabling faster and better decisions

CREATIVITY

Create new content and ideas, including conversations, stories, images, videos, and music

Generative AI Application Examples

Travel & Hospitality
Chatbot / assistants / conversational AI, question answering, search

Healthcare
Protein folding, drug development, personalized medicine, improved medical imaging

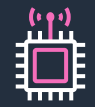
Media and entertainment
Video game generation, upscaling content, face synthesis, film preservation and coloring

Automotive
Autonomous vehicles, design parts for fuel efficiency

Financial services
Risk management, fraud detection, summarize annual reports, review portfolio / risks

Consumer goods
Optimize pricing and inventory, correctly flag product brand and category

Energy and utilities
Design renewable energy sources optimized for geo, predictive maintenance



Technology hardware
Chip design, robotics

Different GenAI player

Model Builder

Startup, Research Institute, Enterprise

Model Tuner

Integrated Software Vendors, Startups, GSI

Model Consumer

Everyone

Most have inexistent or limited access to accelerators

Persona training, deploying and using model

Machine Learning Engineer

Engineer

Scientist

Data scientist

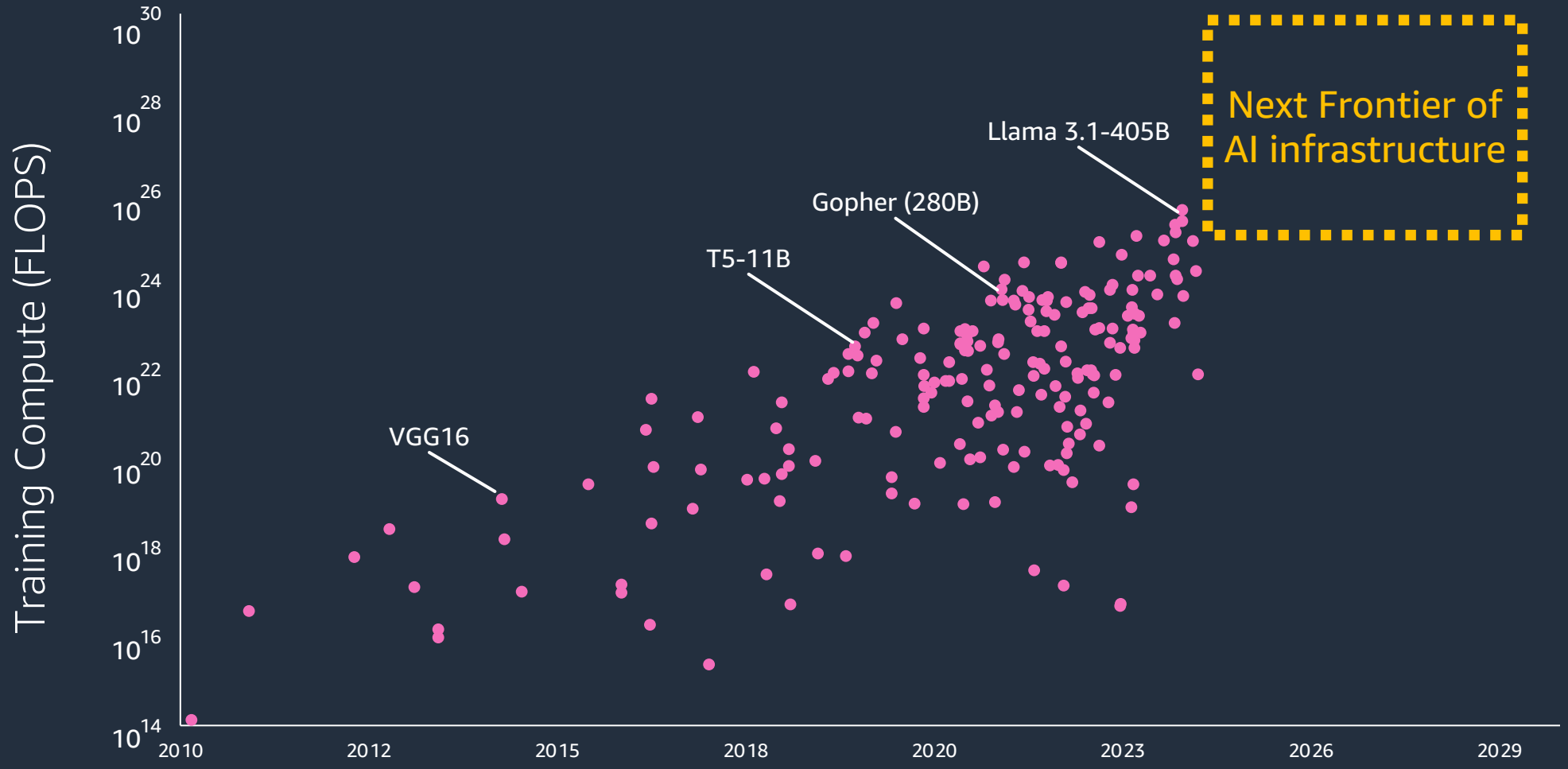
Analyst

Accelerated computing for ML

LARGEST SELECTION OF ML INSTANCES IN THE CLOUD

Model Training Challenges

Scaling compute requires next-generation AI infrastructure



Source: Epoch AI, 'Data on Notable AI Models'. 26 Nov 2024

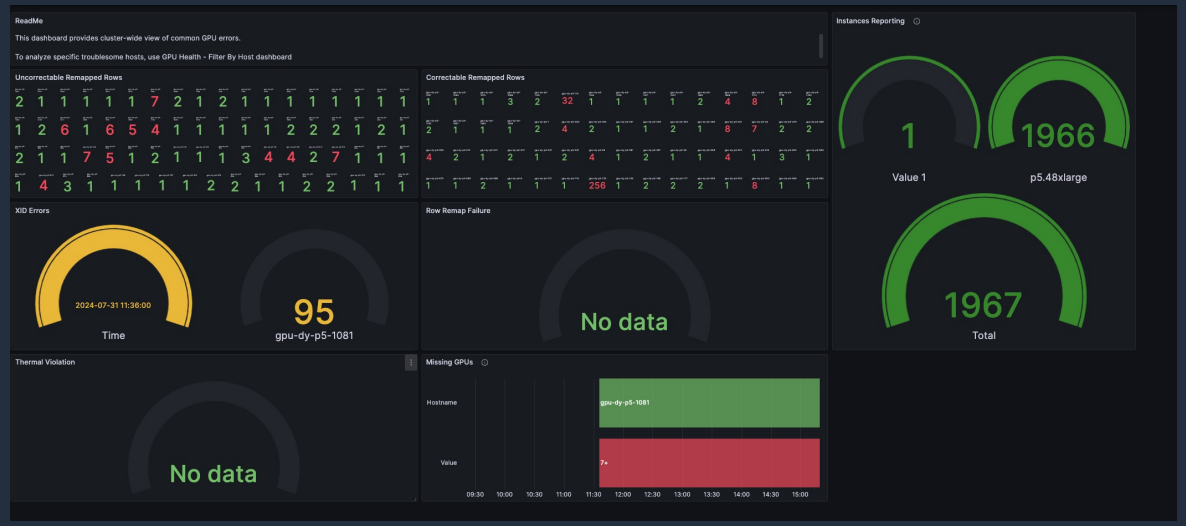
Training at scale

Llama 3 405B was trained during 54 days on **16,000 H100 GPUs** on a multilingual corpus of 15T tokens corpus

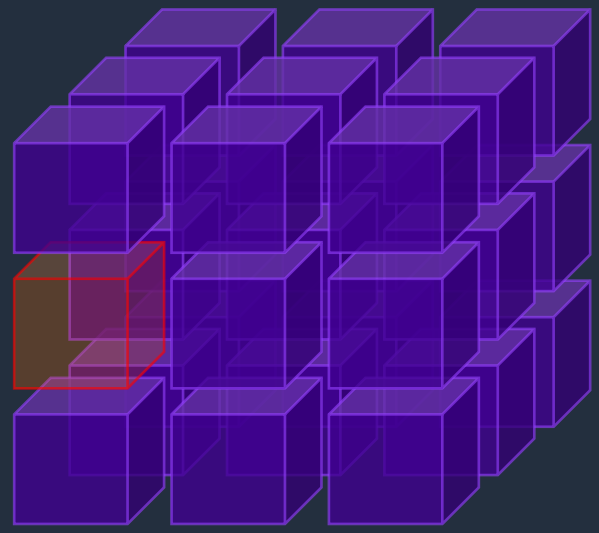
Interrupted every ~2.8h
GPUs accounted for 58.7% of failure

Infrastructure resilience

Observability



Fault detection, hardware replacement



SageMaker Hyperpod
EKS NodeRepair

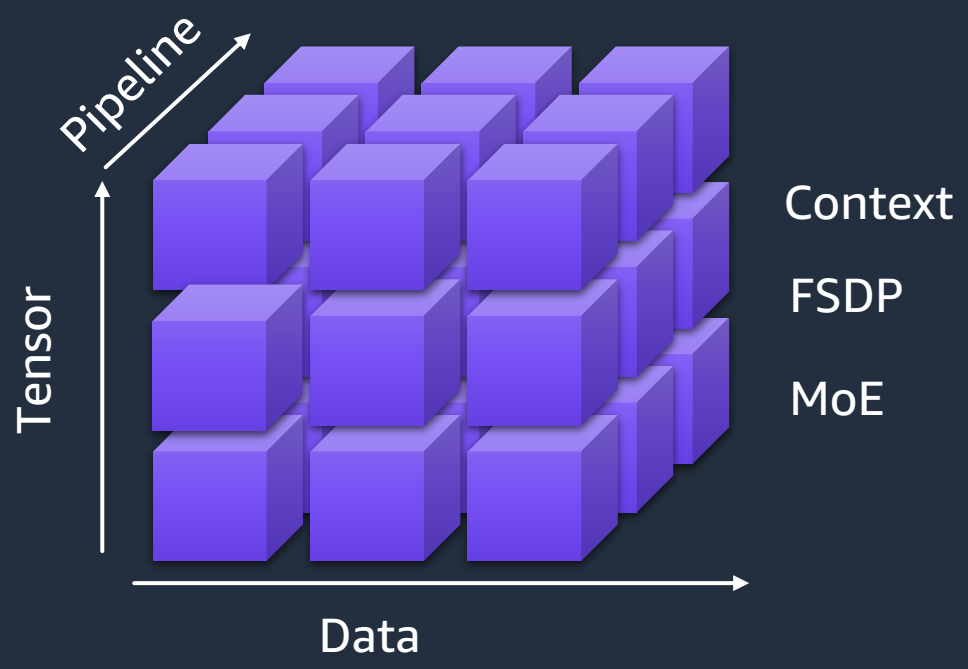
Customer focus on training and business value

Multi-knobs performance

NCCL Parameters

NCCL_COLLNET
NCCL_BUFFSIZE
NCCL_P2P
MCCL_MNVLS

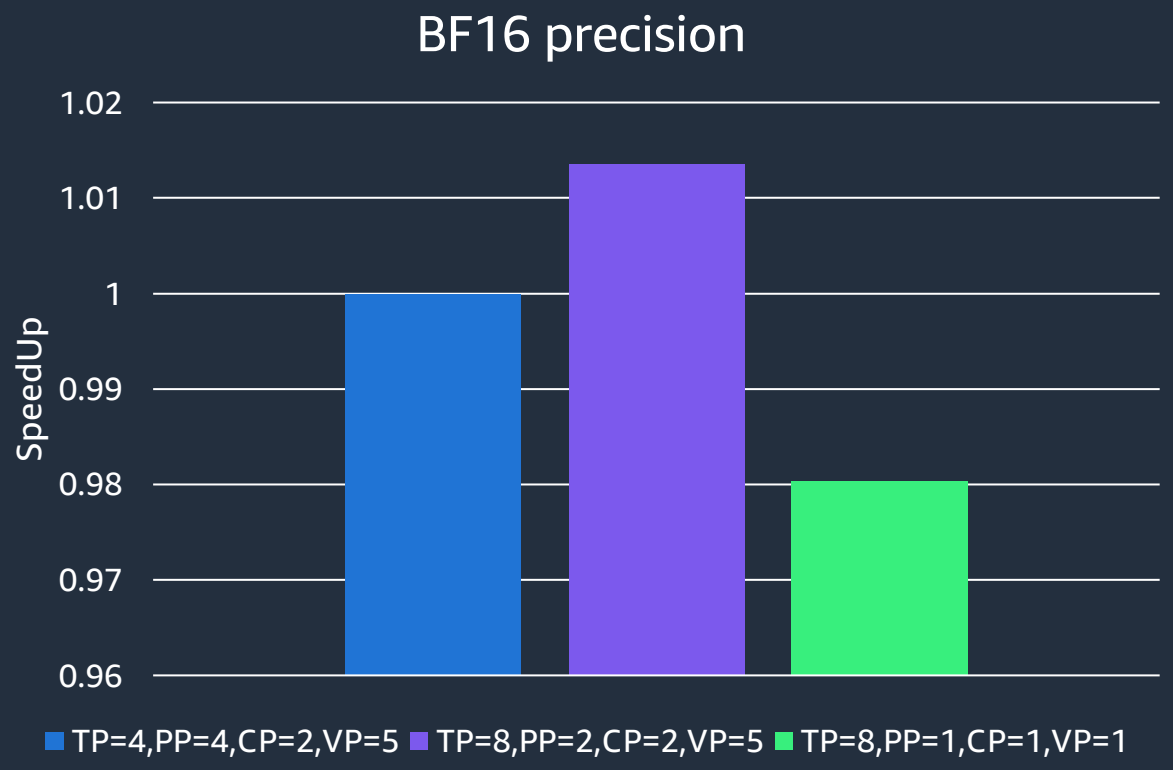
Parallelism

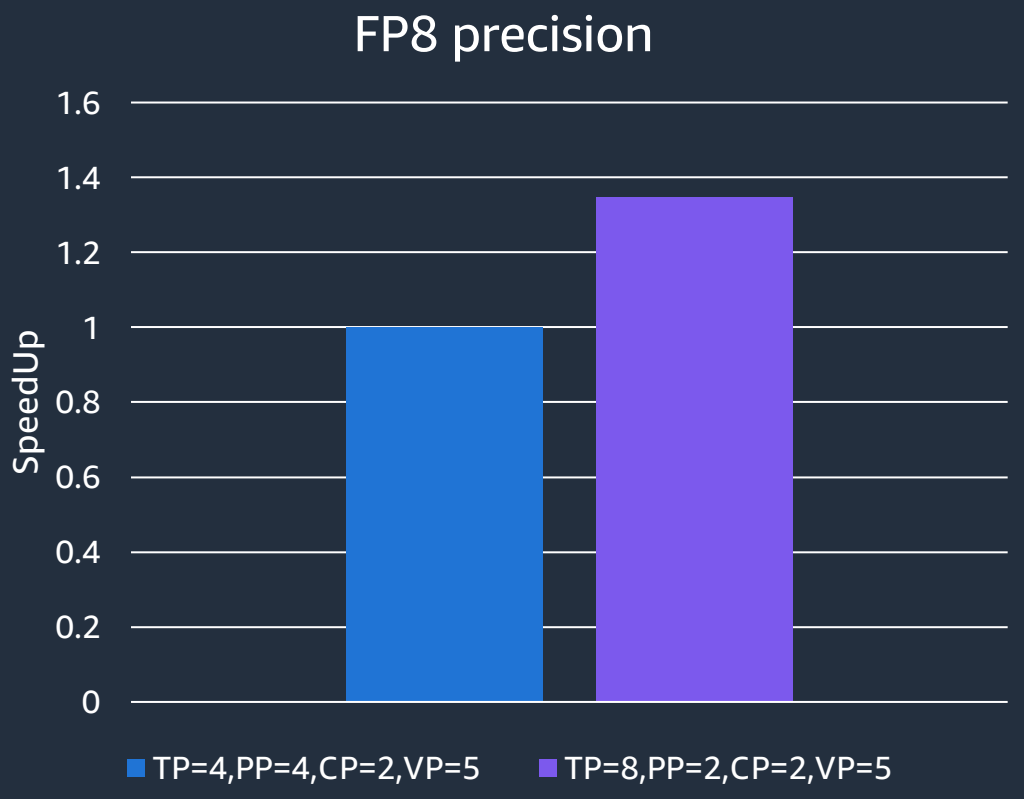


Precision

BF16
FP8+BF16
MoE FP4 + BF16

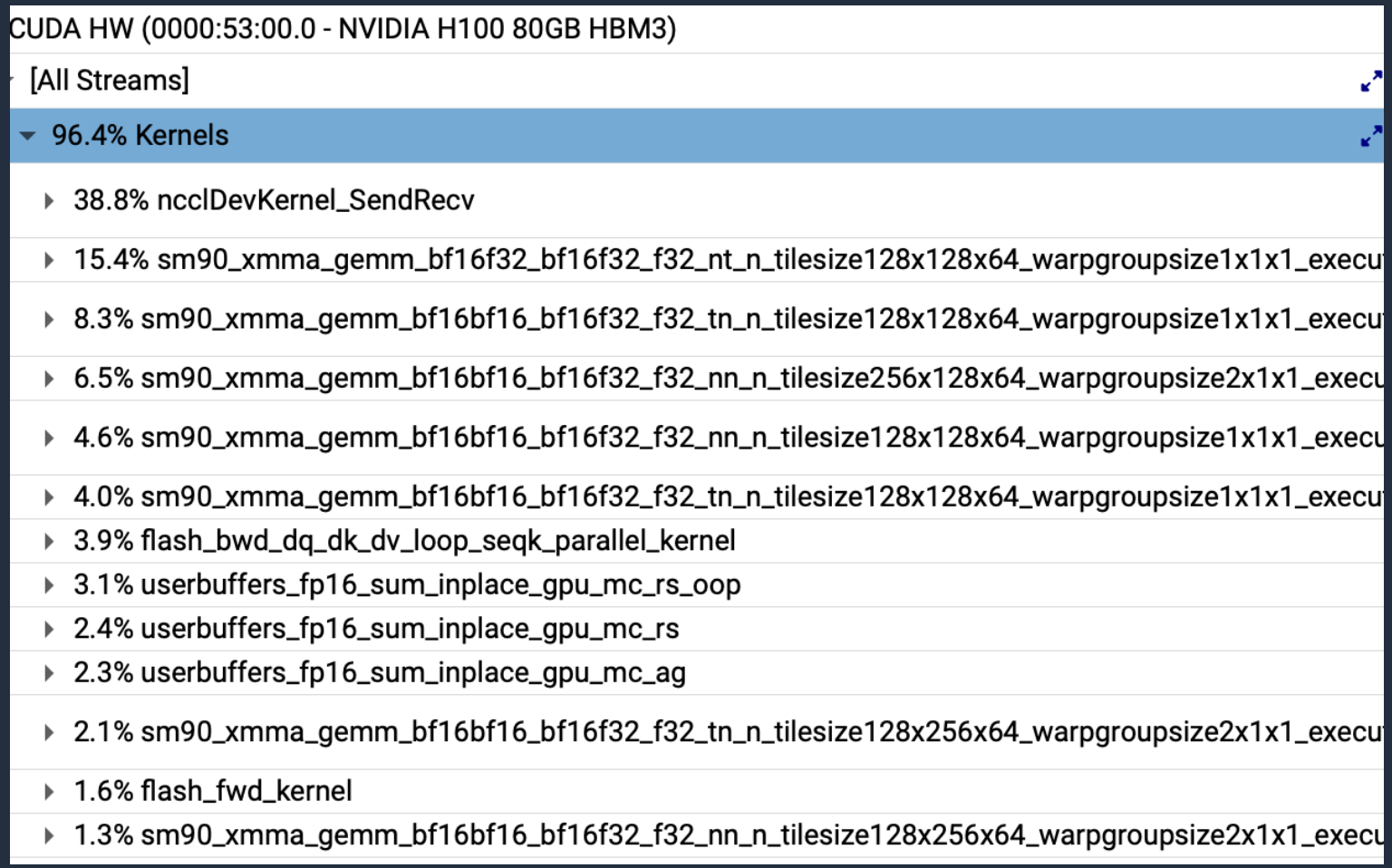
Parallelism impact on training performance





Llama 3.1 70B on 128x H100

LLama3.1 70B Profile



Send/Recv dominant

Model Serving

Model serving type

Offline

Batch processing
Image
Text
Database

Online

ChatBot
Assistant
Medical Imaging
Image creation

Serving model closer to end-users

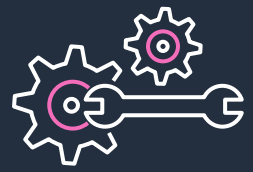
600+

Amazon CloudFront Points of Presence

Build and Scale AI with Amazon Bedrock

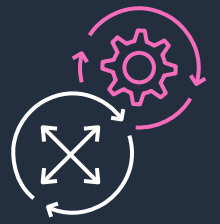
Accelerate development of generative AI applications using FMs through an API, without managing infrastructure

Choose FMs from AI21 Labs, Anthropic, Stability AI, and Amazon to find the right FM for your use case



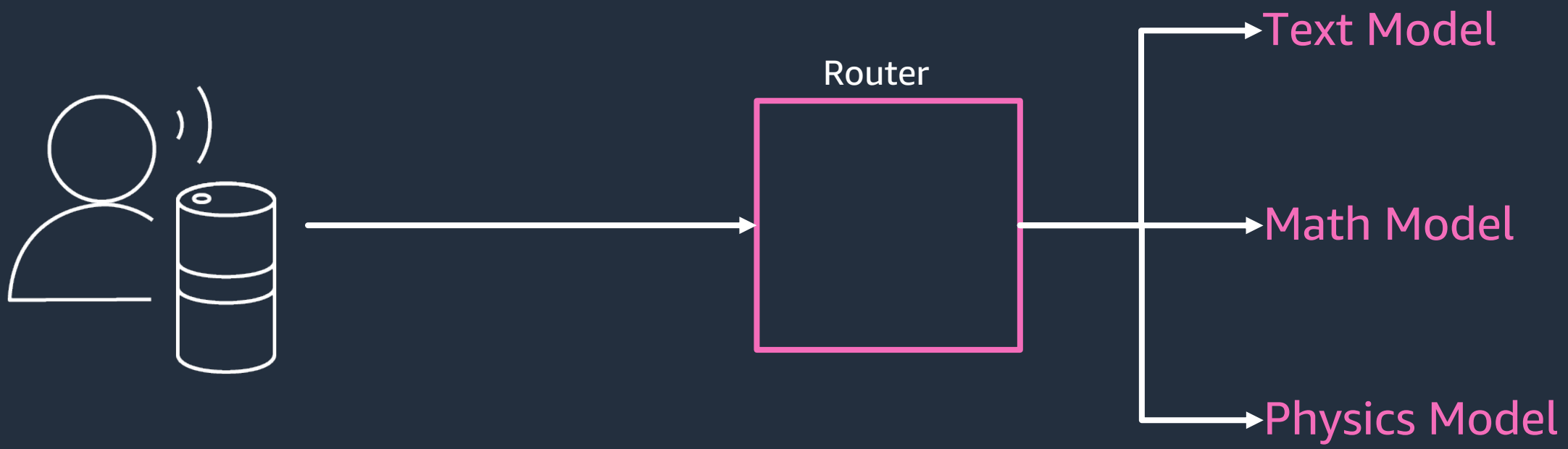
Privately customize FMs using your organization's data

Enhance your data protection using comprehensive AWS security capabilities



Use AWS tools and capabilities that you are familiar with to deploy scalable, reliable, and secure generative AI applications

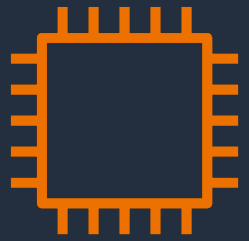
In the wait of AGI



What’s next

Ease of use for performance and results

Abstraction from infrastructure is reality



AI optimizing AI ?

Tuning Infrastructure

Tuning parallelism

HPC Agents

Thank you!