AMAZON AT MLSYS

Tarun Gopinath

Sr. Principal Eng @Amazon

Amazon & AI

Applications (Alexa; Stores; Prime Video; etc)

SaaS (AWS Quick; Connect; etc)

PaaS (AWS Agent Core; Sage Maker; etc)

IaaS (AWS Compute; Storage, Security; etc)

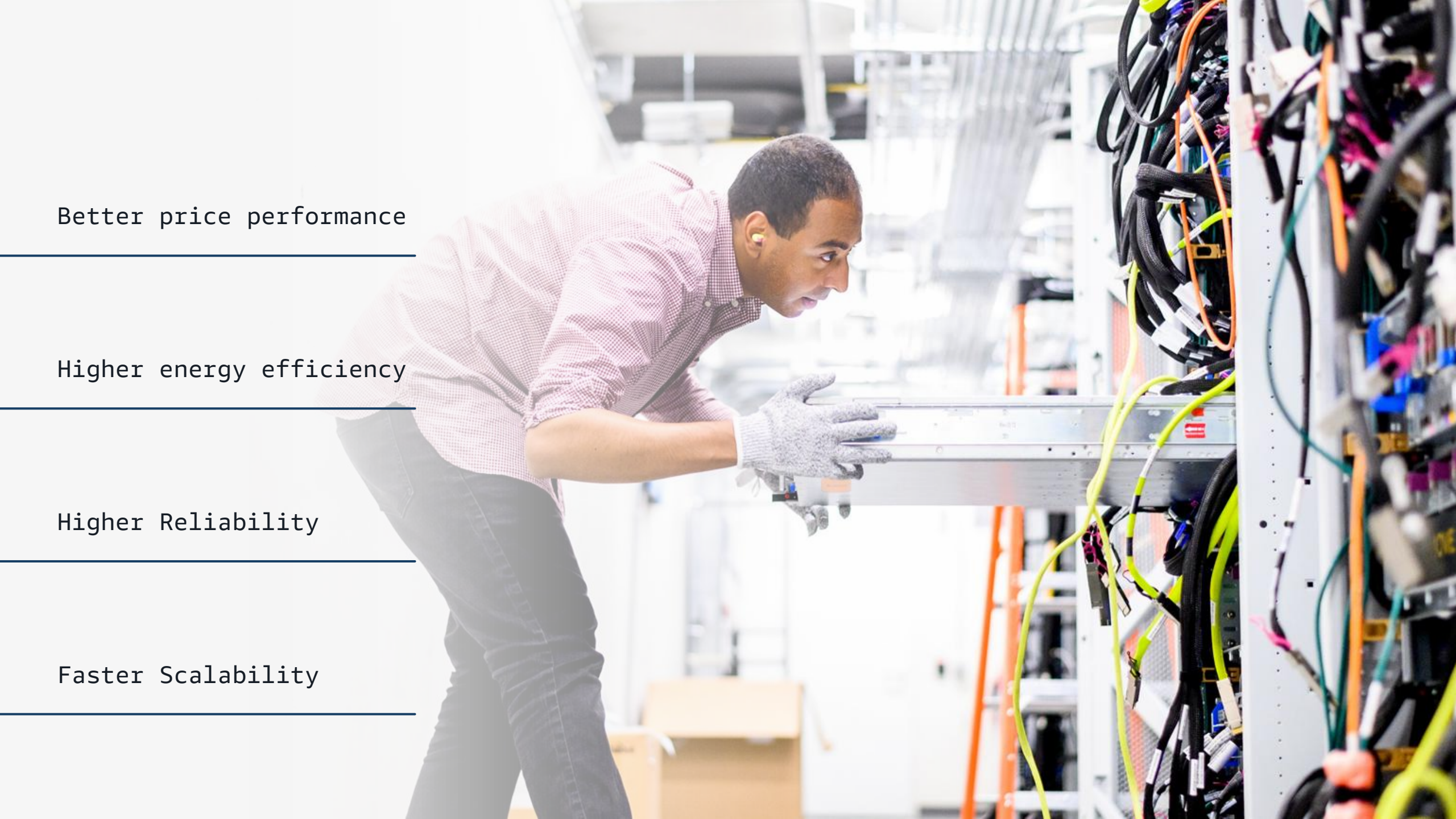
Annapurna Labs

Better price performance

Higher energy efficiency

Higher Reliability

Faster Scalability



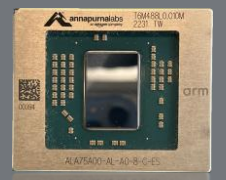
2011 2013 2015 2017 2018 2019 2020 2021 2022 2023 2024 2025

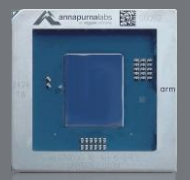

NITRO V1


NITRO V2


NITRO V3


NITRO V4


NITRO V5

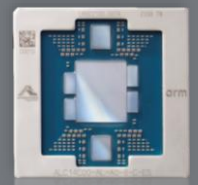

NITRO V6

Silicon that drives the innovation…


GRAVITON


GRAVITON2


GRAVITON3/3E


GRAVITON4


INFERENTIA


TRAINIUM


TRAINIUM2

AWS
Trainium3

4th generation AI chip purpose built deliver the best token economics for next gen agentic, reasoning, and video generation applications.

2.52 PFLOPS
MXFP8 Compute

144 GB
HBM3E Capacity

4.9 TB/s
HBM3E Bandwidth

Amazon EC2 Trn3 UltraServers

vs Trn2 UltraServers

5x

OUTPUT TOKENS PER MEGAWATT

PROJECT RAINIER

The world's largest AI compute cluster

PRIME DAY 2025

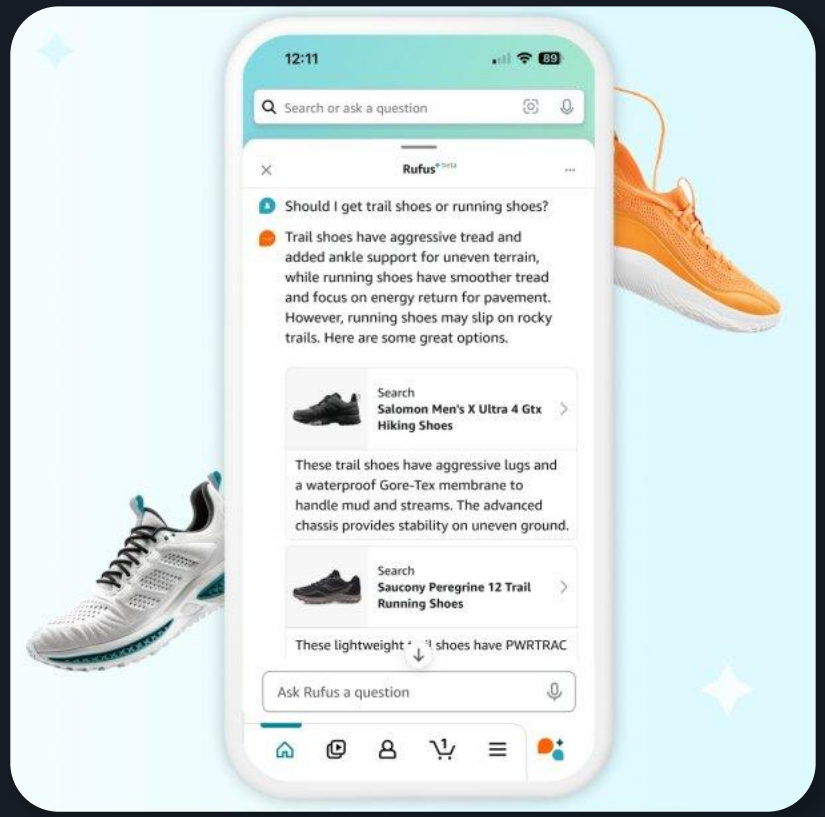
Powering Rufus

120k+

TRAINIUM & INFERENTIA CHIPS DEPLOYED

3 MILLION

TOKENS PER MINUTE WHILE MAINTAINING P99 <1 SECOND LATENCY



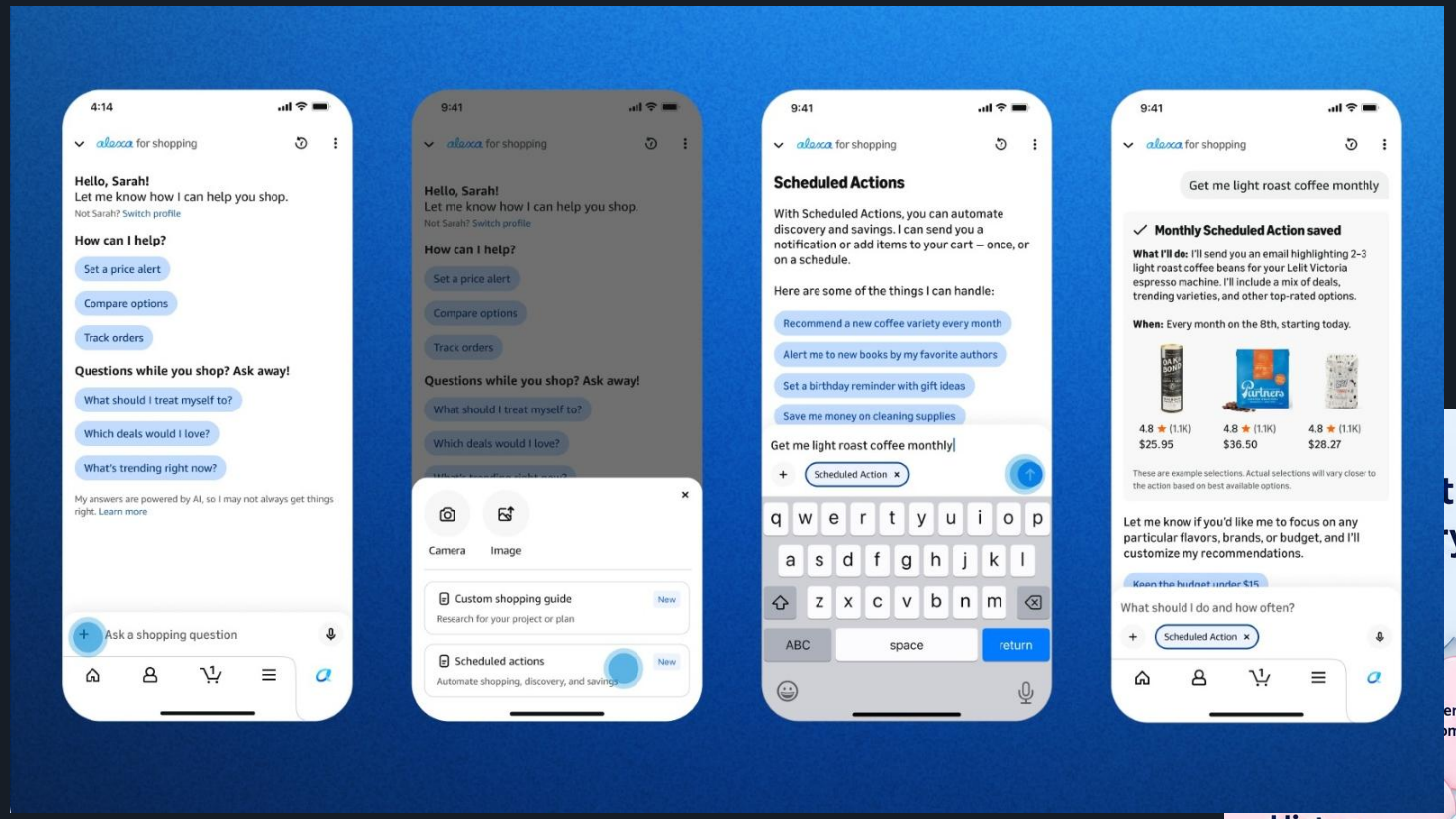
4.5X

COST REDUCTION VS GPU-BASED SOLUTIONS

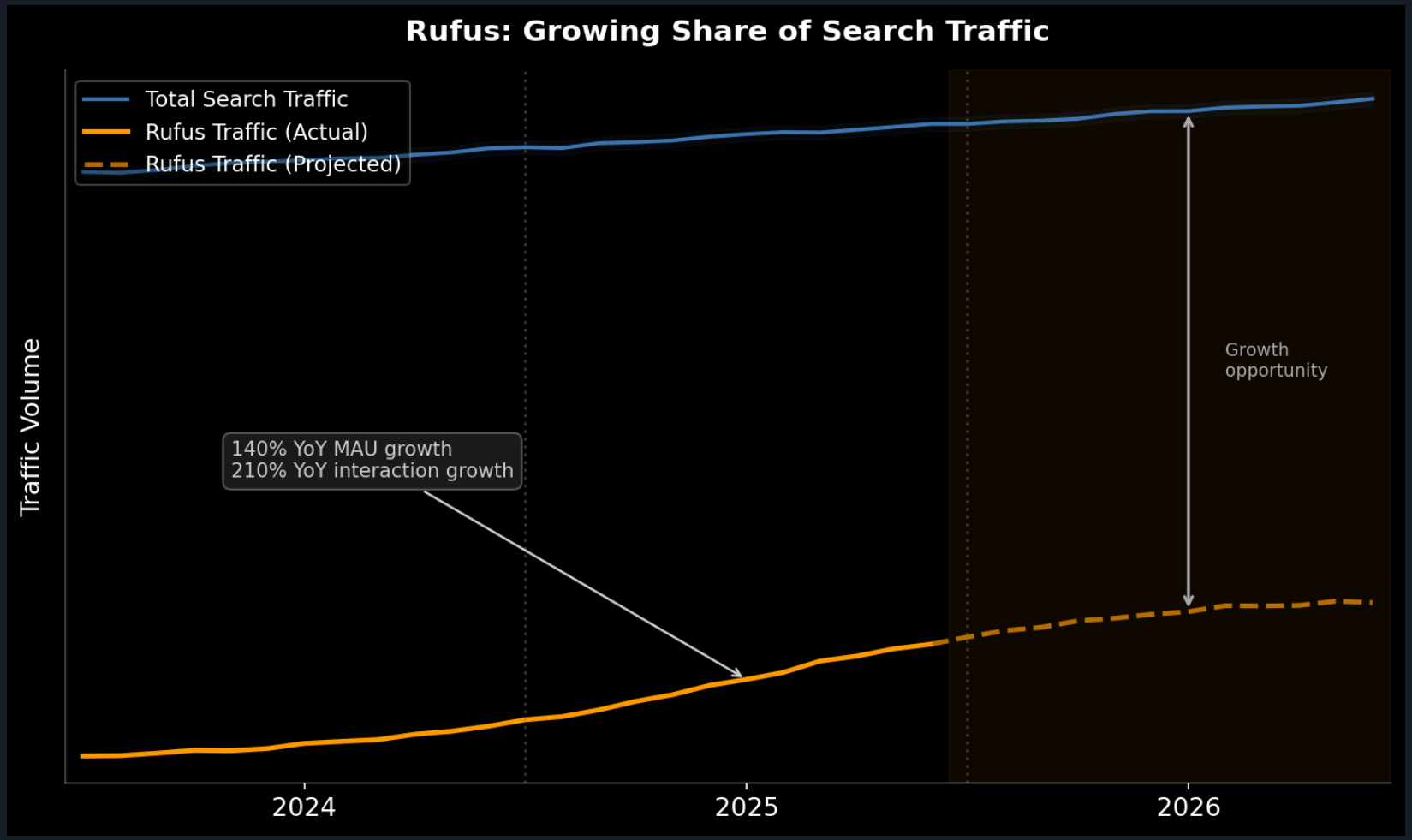
54%

BETTER PERFORMANCE PER WATT

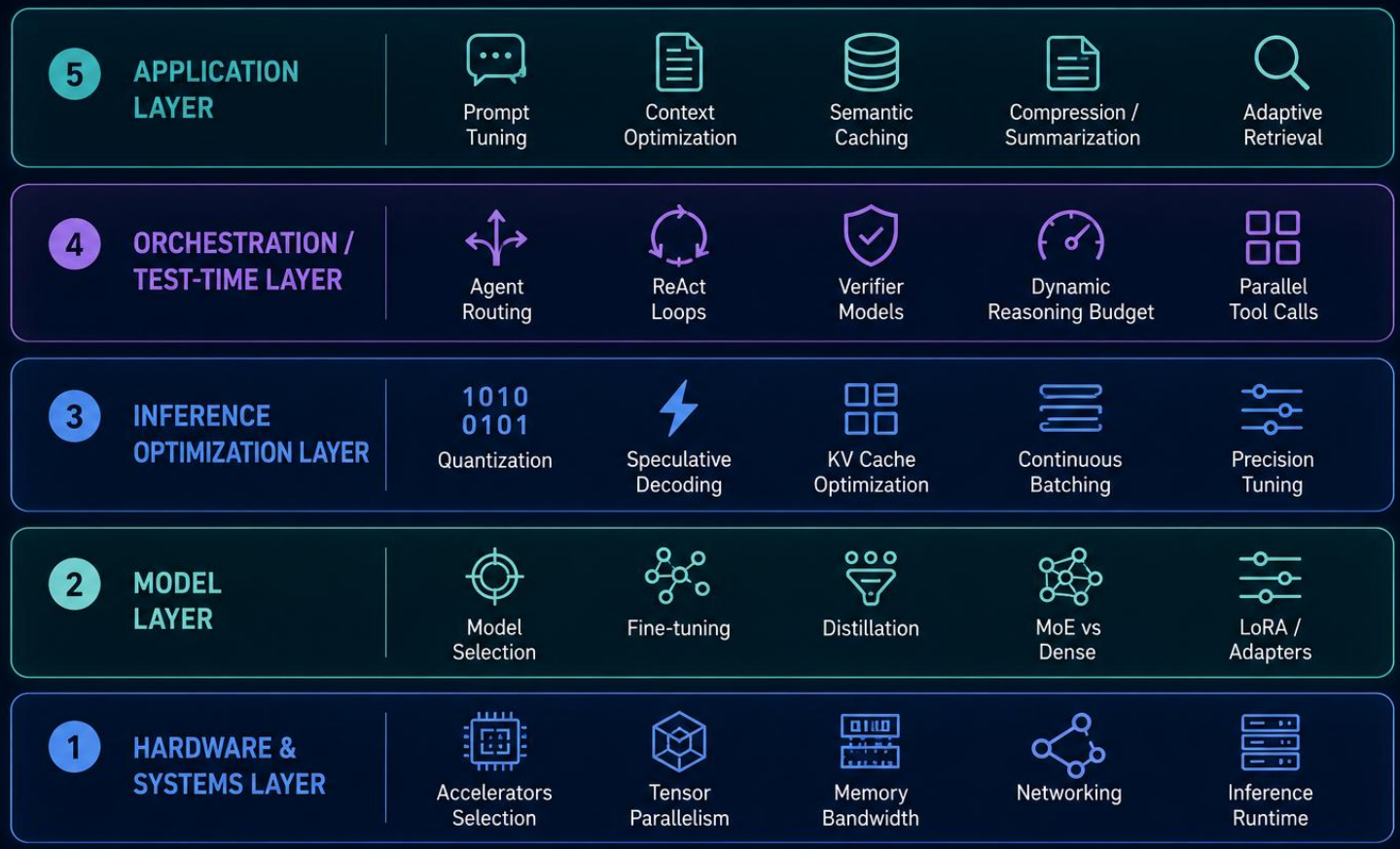
Alexa for Shopping



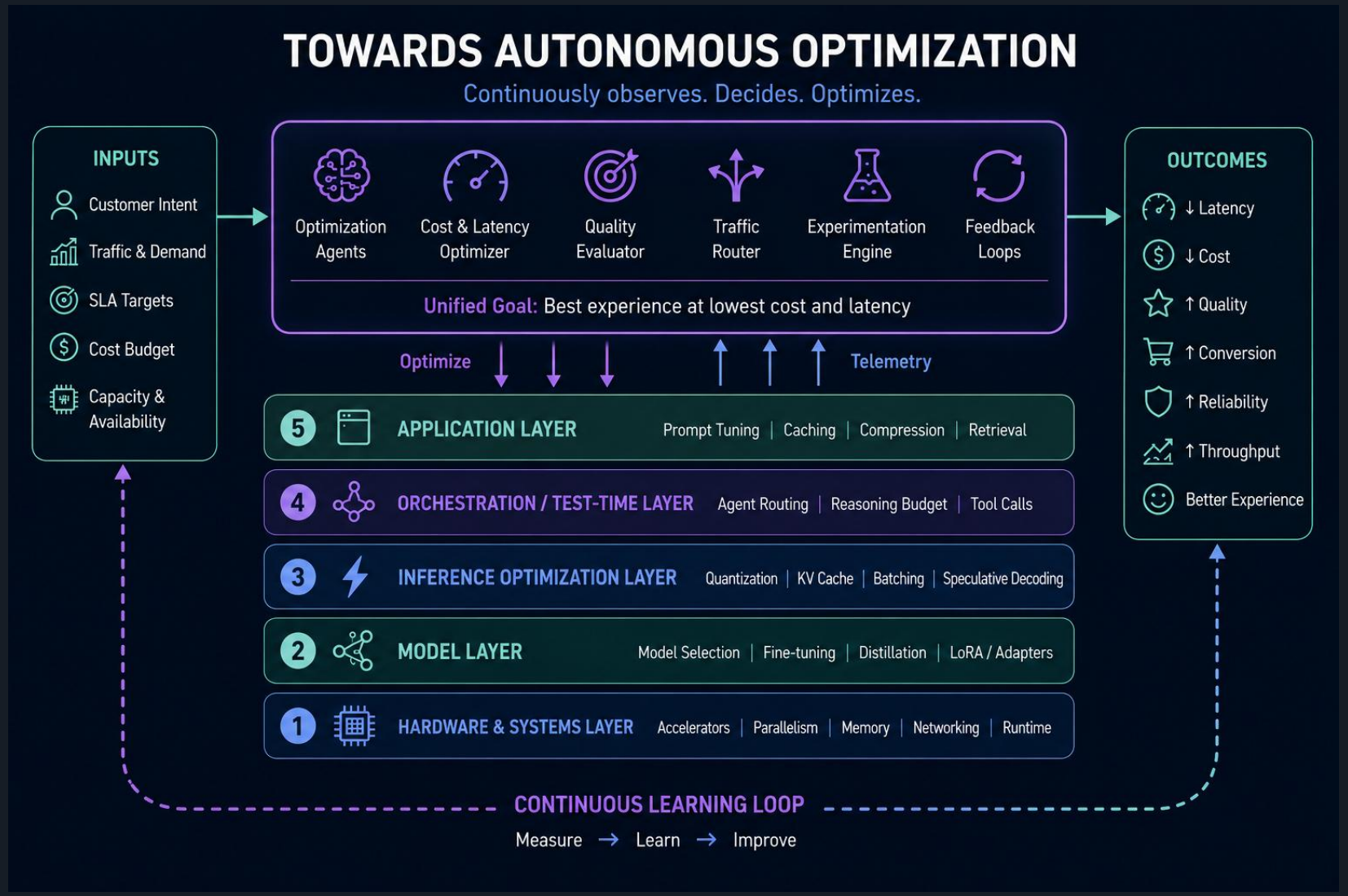
Alexa for Shopping Going Big



OPTIMIZATION LEVERS ACROSS THE LLM STACK



Different layers. Different levers. One optimization goal.



Amazon at MLSys 2026

- **AccelOpt: A Self-Improving LLM Agentic System for AI Accelerator Kernel Optimization**
- **CRAFT: Fine-Grained Cost-Aware Expert Replication For Efficient Mixture-of-Experts Serving**
- **GUARD: Scalable Straggler Detection and Node Health Management for Large-Scale Training**
- LEANN: A Low-Storage Overhead Vector Index
- FlexiCache: Leveraging Temporal Stability of Attention Heads for Efficient KV Cache Management
- ProTrain: Efficient LLM Training via Automatic Memory Management
- Zero redundancy distributed learning with differential privacy
- HetRL: Efficient Reinforcement Learning for LLMs in Heterogeneous Environments

AWS Trainium 2/3 MOE Kernel Challenge

The challenge

- Maximize performance without sacrificing accuracy on baseline Qwen3 30B-A3B MoE
- Round one: single chip Trn2 instance
- Round two: top 15 teams received single chip Trn3
- The top teams developed NKI kernels reduced latency by ~75% and improved throughput by an average of 4x on each.

Join the session on Friday, May 22 11am-1pm to learn more!

NKI MOE Project

Come build with us!

AWS Trainium and Neuron SDK

Internship opportunities

SDE & Applied Science Opportunities

Thank You !

Internship opportunities

SDE & Applied Science Opportunities