AI for Good

Rethinking Infrastructure for Agentic AI:
From Large-Scale AI Operations to Distributed Agent Execution

Joongi Kim CTO | Lablup Inc.

The future of AI native communication networks

AI for Good Summit 2026

lablup

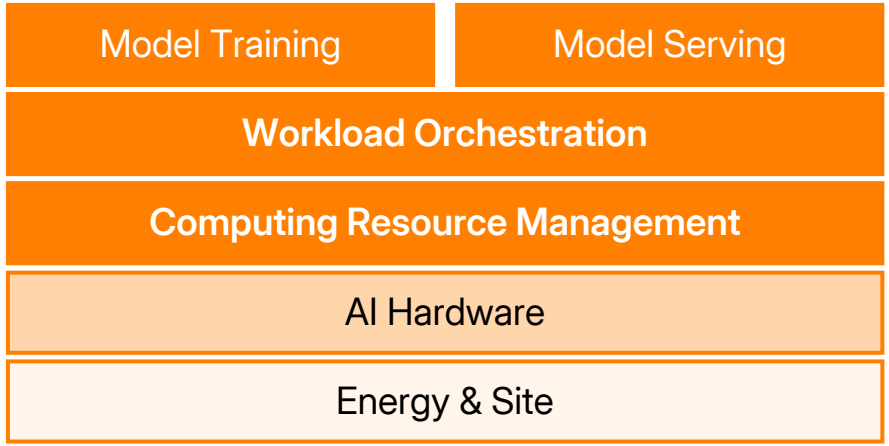
AI Native Communication Networks

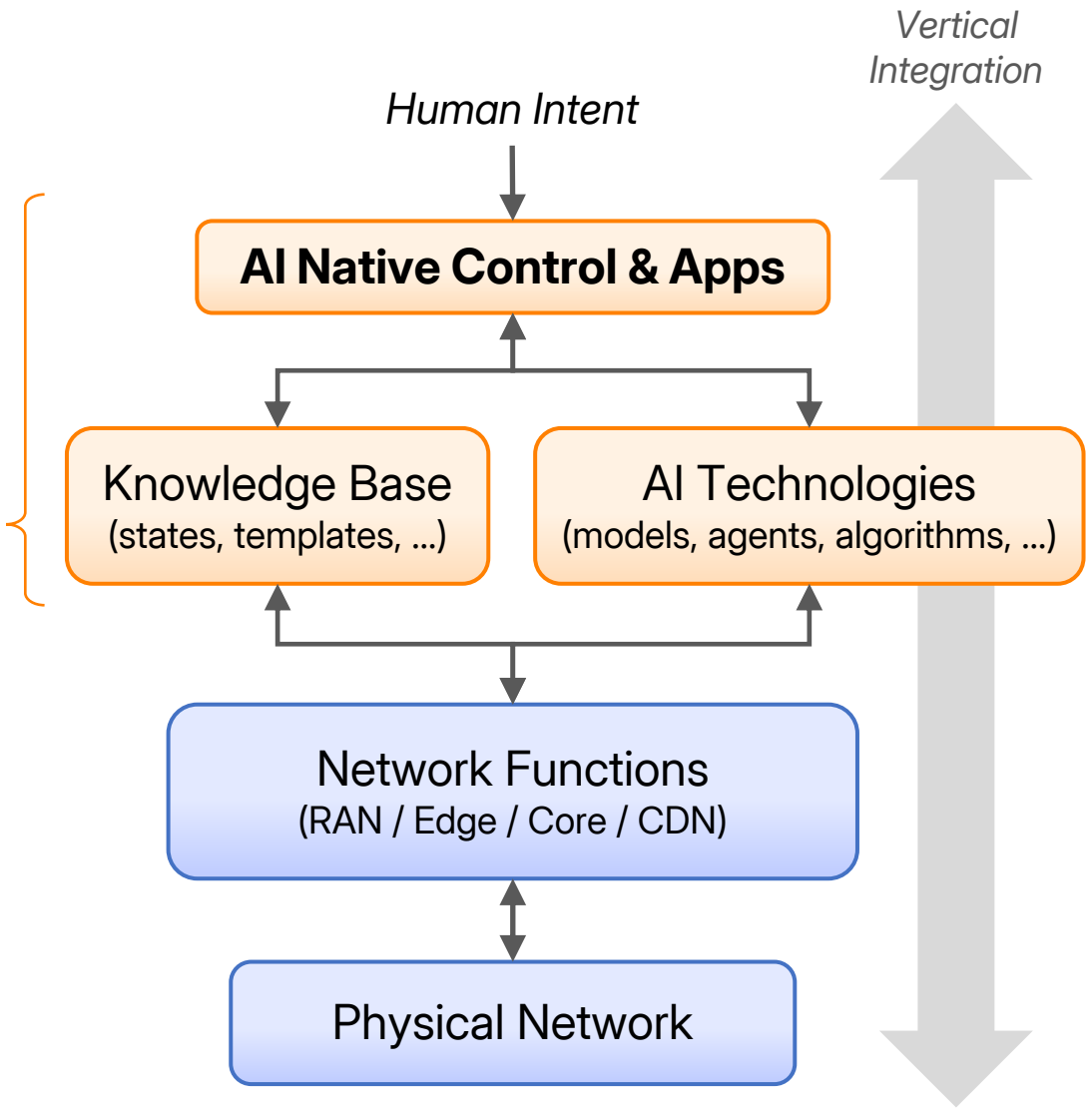
The future of AI native communication networks

AI Native Communication Networks

> Telecommunication networks integrating **AI as a core component**, enabling novel *value-added services*, and enhancing *network and service performance*.

"AI Infrastructure" as the foundation for this



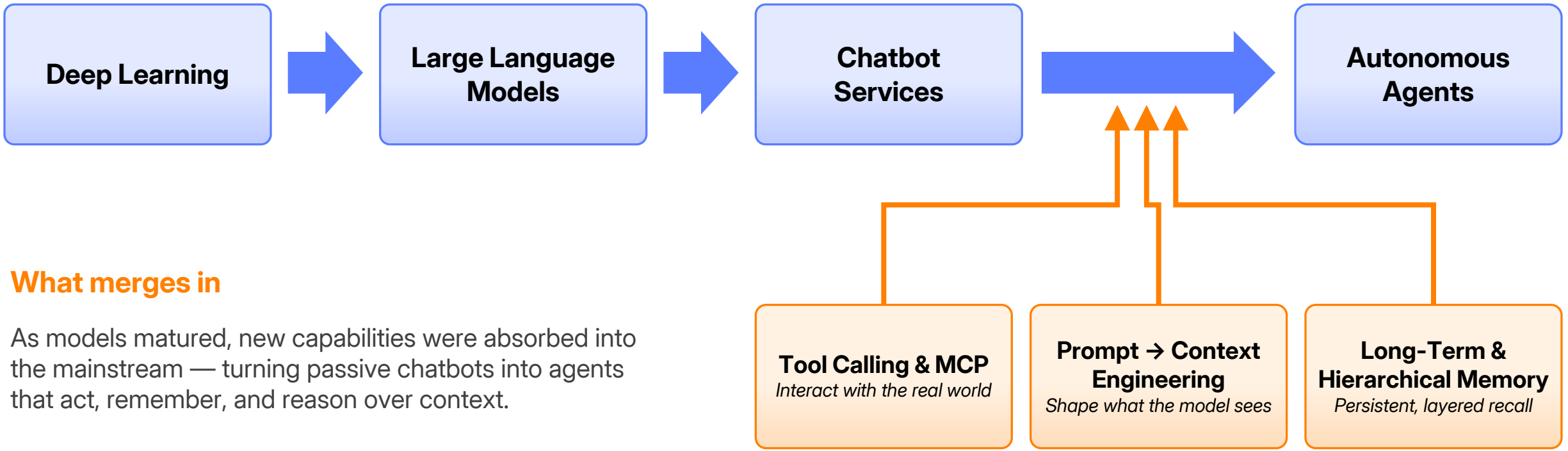


From MLOps to Agentic AIOps

The future of AI native communication networks

Era of Agentic AI

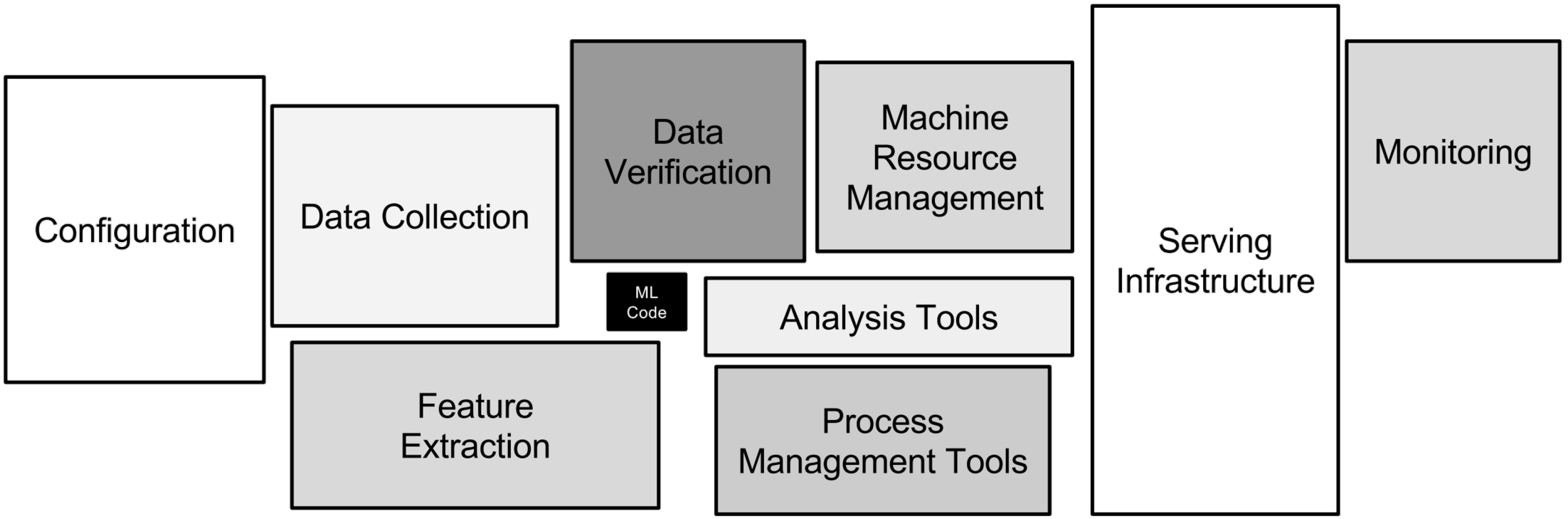
MLOps *LLMOps* *AIOps & AgentOps*



What merges in

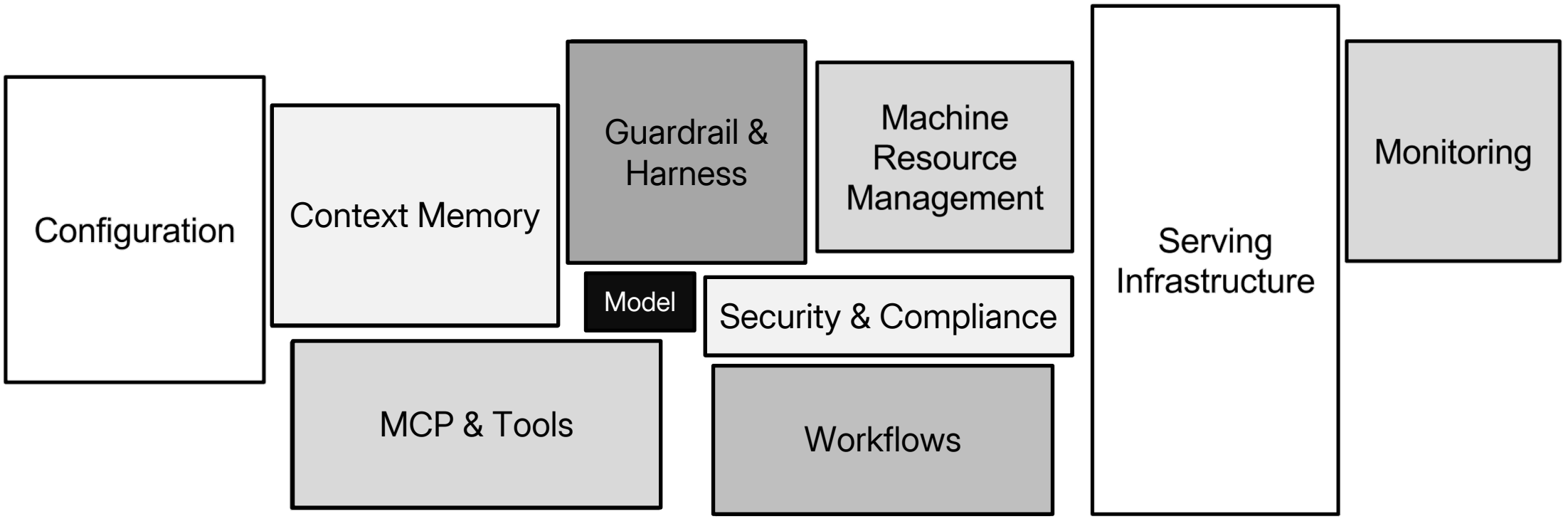
As models matured, new capabilities were absorbed into the mainstream — turning passive chatbots into agents that act, remember, and reason over context.

MLOps Challenges



Courtesy of NeurIPS 2015 Paper "Hidden Technical Debt in Machine Learning Systems" (D. Sculley et al.)

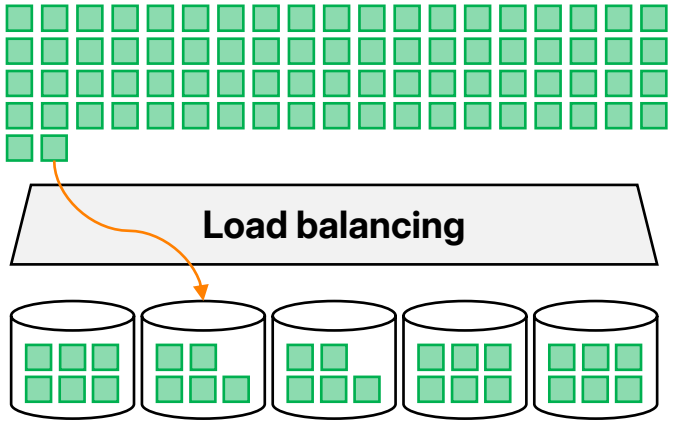
Agentic AIOps Challenges

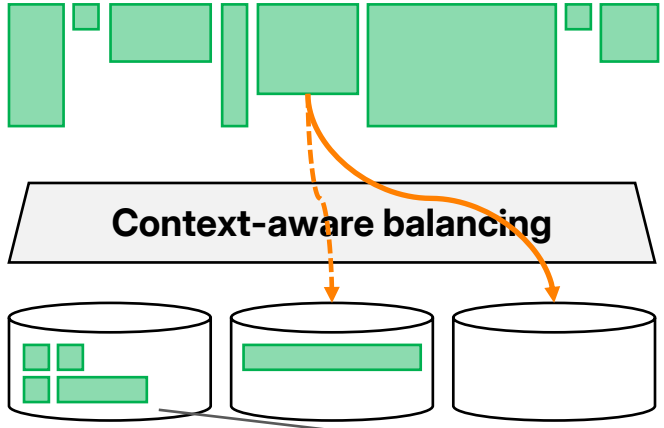


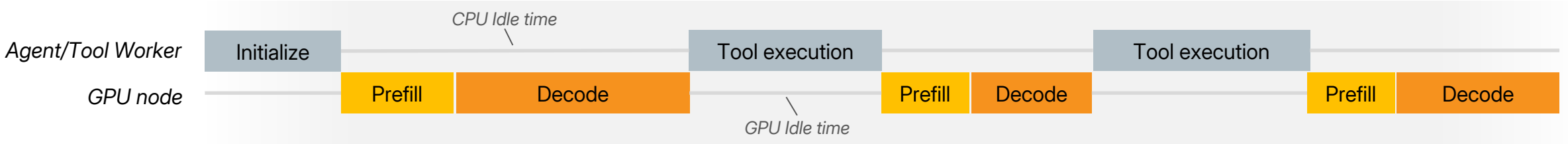
"Agentic AIOps" in 2026

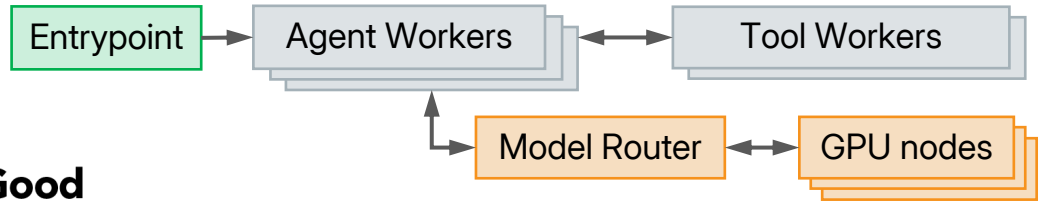
AI Workload Difference

- Web Applications & Microservices
  - Requests: short-lived, homogeneous
  - *Preemptible*: easy to scale-in & out
- AI Inference & Agents
  - Requests: long-running, multi-turn, interleaved
  - *Context-senstive*: requires smart routing & kvcache









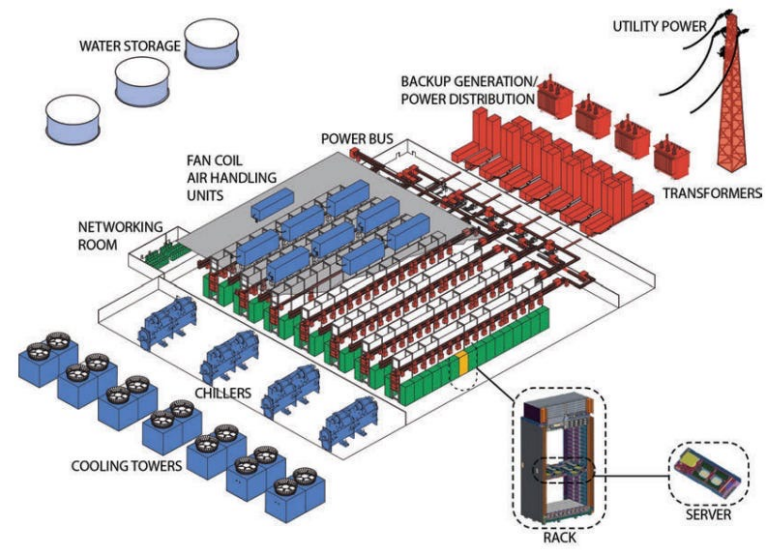
- Fixed initial contexts (system prompts, souls, skills, ...)
- Long input contexts (source codes, knowledge base, ...)
- Generated contexts (analysis, reasoning, ...)

AI Infrastructure Challenges

The future of AI native communication networks

HW · SW challenges to catch up the growth speed

Growing computing needs:
~4x/yr



[1] https://www.construction-physics.com/p/how-to-build-an-ai-data-center

Interconnect Bottleneck

30%

bandwidth lost

Poor data placement wastes inter-node bandwidth in multi-vendor clusters

Scheduler Scalability

100+

GPU ceiling

Multi-vendor scheduling collapses past clusters of ~100 mixed GPUs

Utilization Efficiency

30%

industry avg.

Average GPU utilization sits at 30%. The rest is paid-for idle silicon.

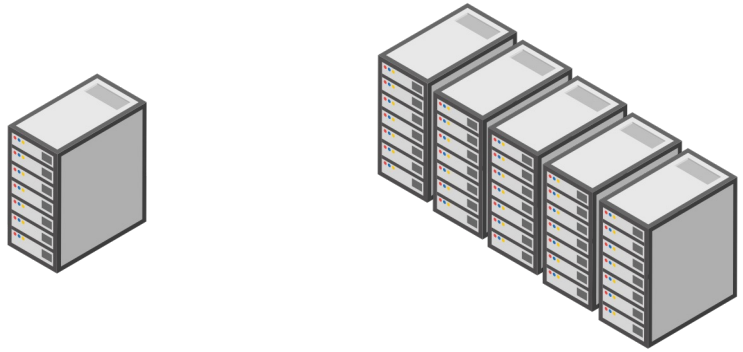
Scaling Challenge Example

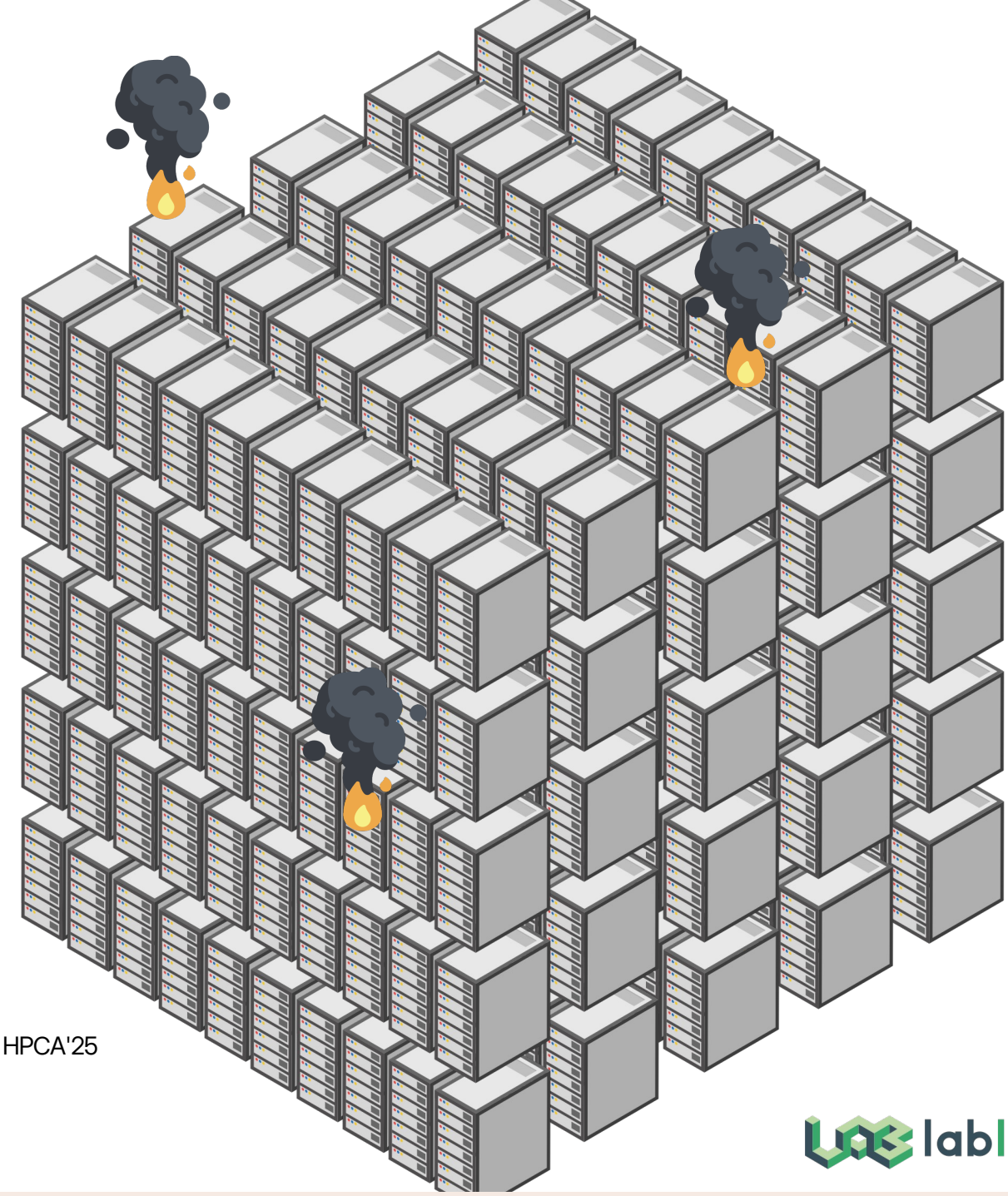
Suppose a single node experiences a fault once per 3 years (1,000 days), or has a 0.1% chance of a failure per day.

What is the probability of having a failure in a day when there are 1,000 nodes?

Approx. 63%

$$P(X \geq 1) = 1 - P(X = 0) = 1 - (1 - p)^{1000}$$

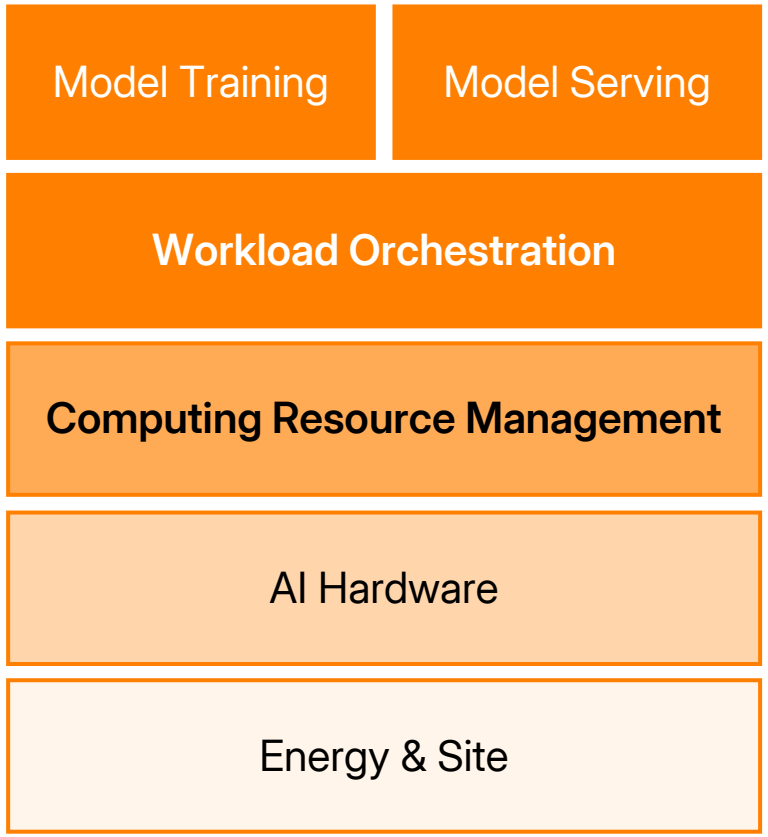




- On 1,024 GPUs, MTTF is **7.9 hours**
- On 16,384 GPUs, MTTF is **1.8 hours**
- On 131,072 GPUs, MTTF is **14 minutes**

† Kokolis et al., Revisiting Reliability in Large-Scale Machine Learning Research Clusters, IEEE HPCA'25

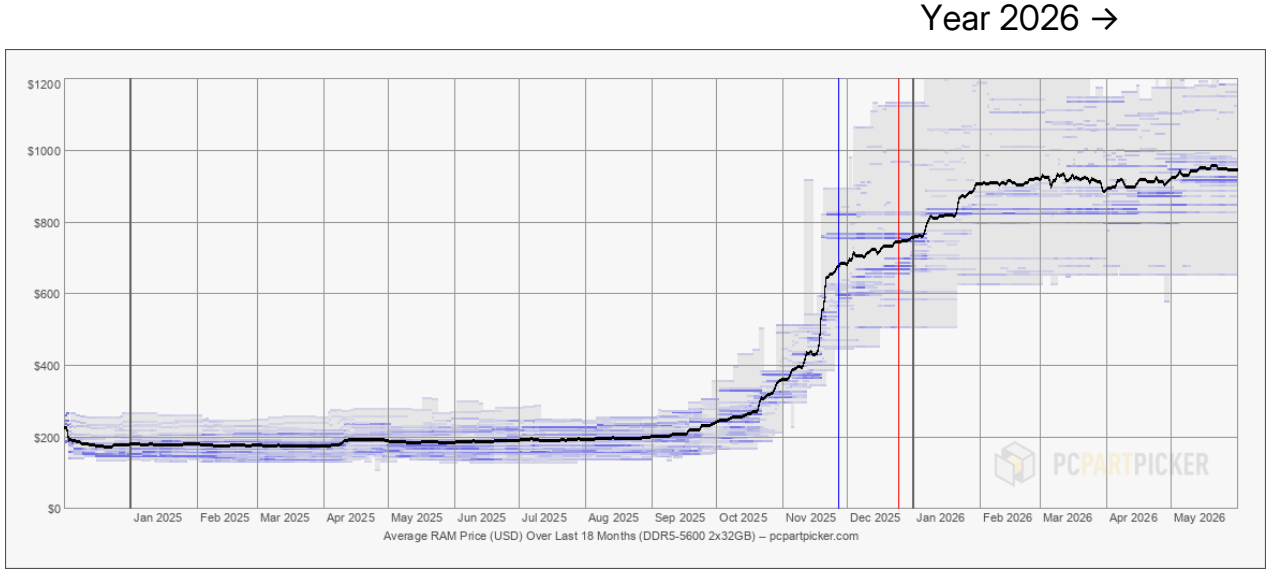
Scalability Strategies by Layers



- Models & AI Hardware
  - Quantization, Fusion Ops
  - Compiler-level optimization + Batching + Pipelining
- Workload Orchestration
  - Scheduling latency ∝ {# jobs, # users, # nodes, # GPUs}
  - Partitioning + **Batching** + Pipelining
  - **Reconciliation loops** to handle partial success/failures
  - Queue mgmt.: priority / fairness / HoL blocking avoidance
  - Awareness/control of **interconnect networks (east-west)**
    - ✓ *Conventional deployment schemes won't fit!*
- Resource Management
  - Topology-aware placements
    - ✓ Rack-scale / node-scale / NUMA / SR-IOV / …
  - Binpacking for shared resourecs
  - Continuous metric monitoring to predict/respond HW failures

Continuously Moving Bottlenecks

- We need to keep GPUs (and CPUs) busy!
- Memory
  - HBM solves throughput but is pricey.
  - Shortage of DDR5 RAM due to market skew
- Networking & Storage
  - Model weight (MoE) + KVCache sharing
  - Disaggregated serving
- CPU
  - Tool calling and MCPs are becoming new hotspot for large-scale, highly concurrent agent deployments.
  - NVIDIA announces the Vera CPU rack solution.



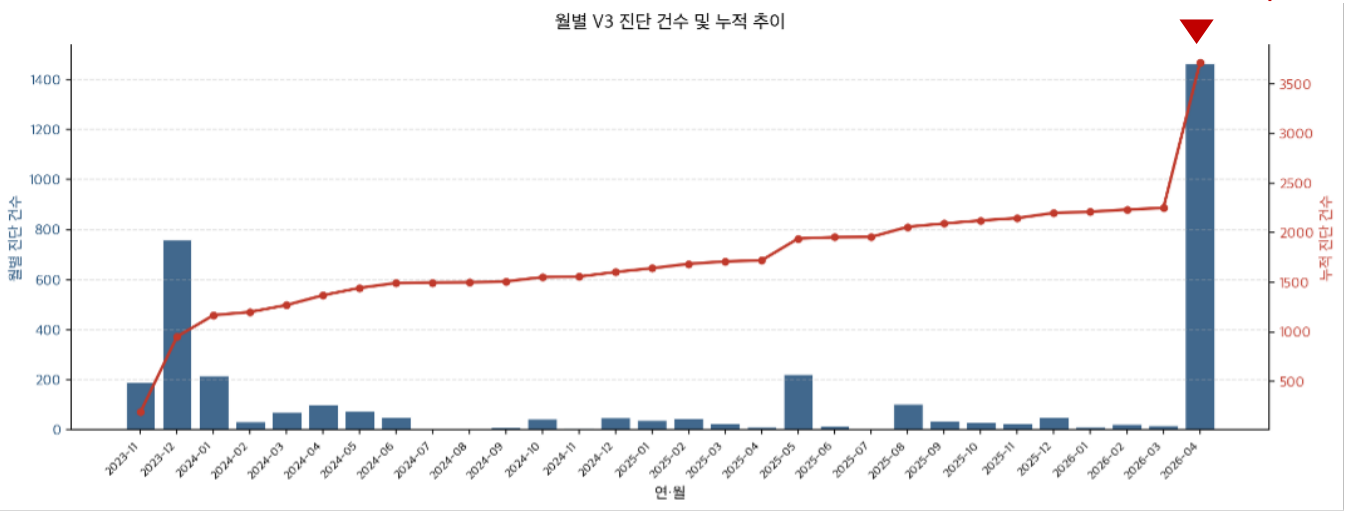
DDR5-5600 2x32GB Price (https://pcpartpicker.com/trends/price/memory/)

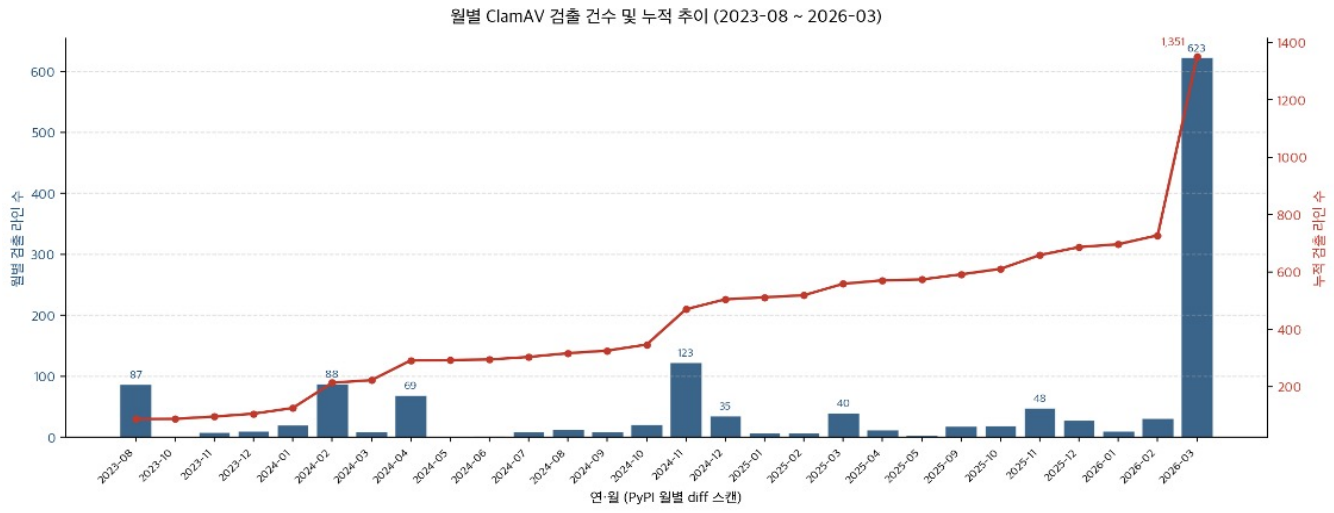
Intel Stock Price (Apple Stocks)

AI Supply Chain Security

- Experience from Backend.AI Reservoir
  - Lablup's add-on product to mirror open source registries in air-gapped clusters
  - Monthly package sync & malware scan
    - ✓ PyPI scanned using V3, ClamAV
- Vulnerabilities uncovered with AI assistance
  - copy-fail, dirty-frag, fragnesia, nginx-rift, nginx-poolslip, Google Big Sleep, React2Shell, ...
- Increasing supply chain attacks
  - Shai-Hulud, LiteLLM, TanStack, ...
- Limited accessibility of frontier AI models
  - Anthropic's Project “Glasswing”
  - US export control of Mythos and Fable 5

2026 March/April





From Large Scale to Distributed Agents

The future of AI native communication networks

Core Elements for AI Infra

Model Training / R&D

- Topology-aware
- Gang Scheduling
- Job Admission Control
- Occupation-based Metering

Model Serving & Agentic Workloads

- **Model Router**
- MCP and Tool Hosting
- Auto-scaling
- **KVCache / Disaggregation**
- Token or Task-based Metering

Common

- Quota
- RBAC
- Workload Scheduler
- **Observability & Telemetry (Cluster · User · Workload)**
- Ops Library
- Compiler
- Device Driver
- Accelerator
- Interconnect
- Storage

Korea's Foudational AI Model Training

Postmortem Tech Report

- Training a Korean-optimized 102B LLM from scratch

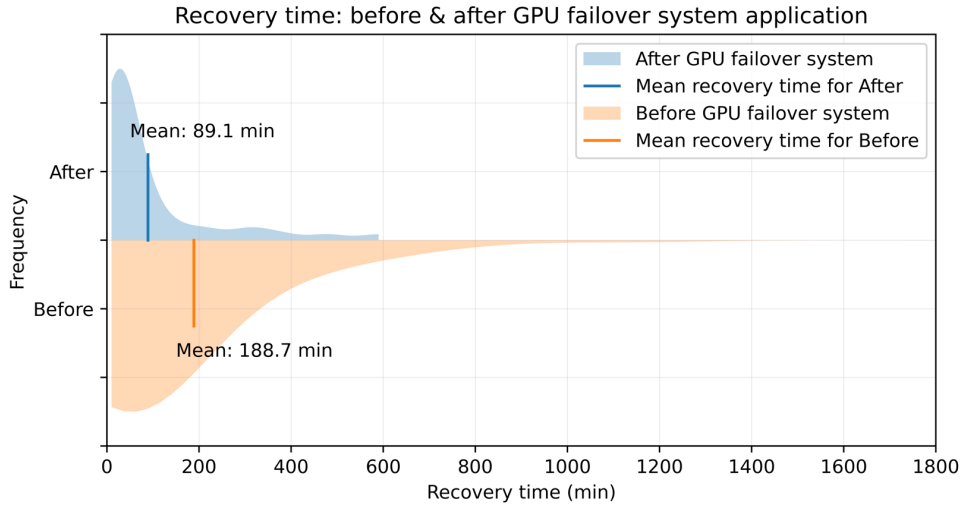
upstage
Model Development & Training

backend AI
Workload Scheduler / Trainer

VAST
Data Storage

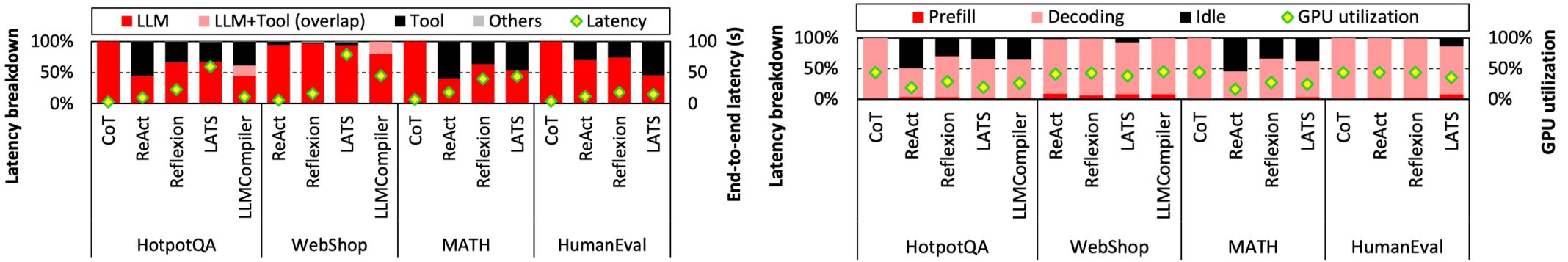
NVIDIA
GPU & Interconnect

SKT 'HAEIN' Datacenter



- Lablup and our team had run a 504x NVIDIA B200 (63 nodes) cluster for 73 days.
- Technical challenges solved via **cross-stack multi-org collaboration**
  - ✓ NFS driver bug: missing readahead batching → Reduced data loading latency from 8+ hours to 30 minutes
  - ✓ Automatic GPU failure detection and job restarts → Reduced recovery time by 47%
- **Postmortem tech report:** https://arxiv.org/html/2605.09370v2
  - ✓ Failure prediction should involve multiple signals while we could observe strong ones like ECC error rates.

Agentic Workload & Infra Utilization



- Mixed CPU-GPU workloads and switching
  - GPU idle time (~54.5%) due to tool calling waits
  - 74.1% of GPU active time spent in memory-bound decode stages
- Dircetions to explore
  - KVCache sharing/offloading to minimize end-to-end latency
  - Heterogeneous hardware routing
    - ✓ e.g., prefill/decode disaggregation
  - Speculative decoding for structured outputs (e.g., JSON)
  - Workload/step-aware scheduling

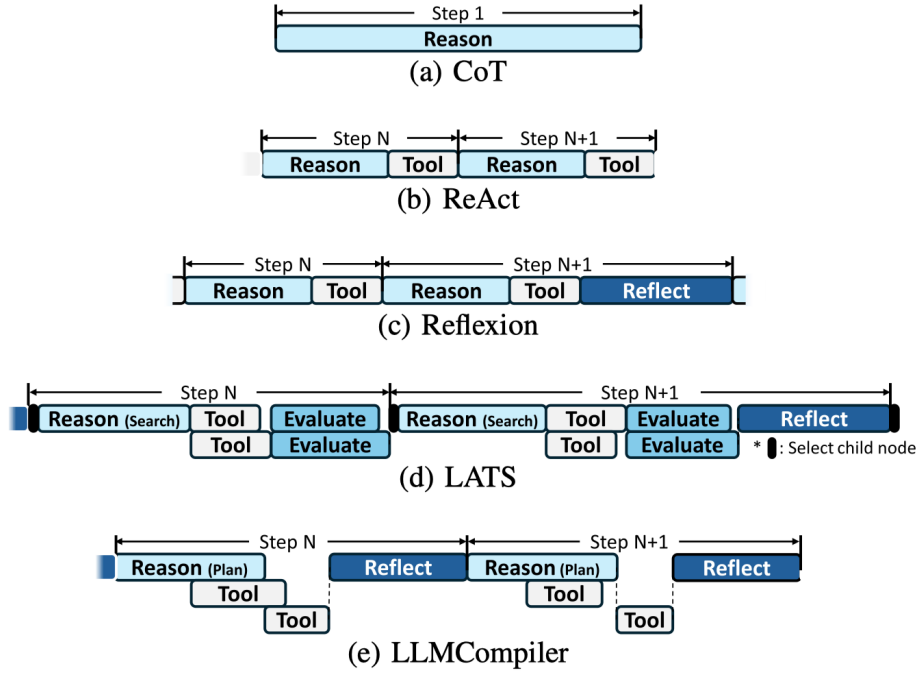
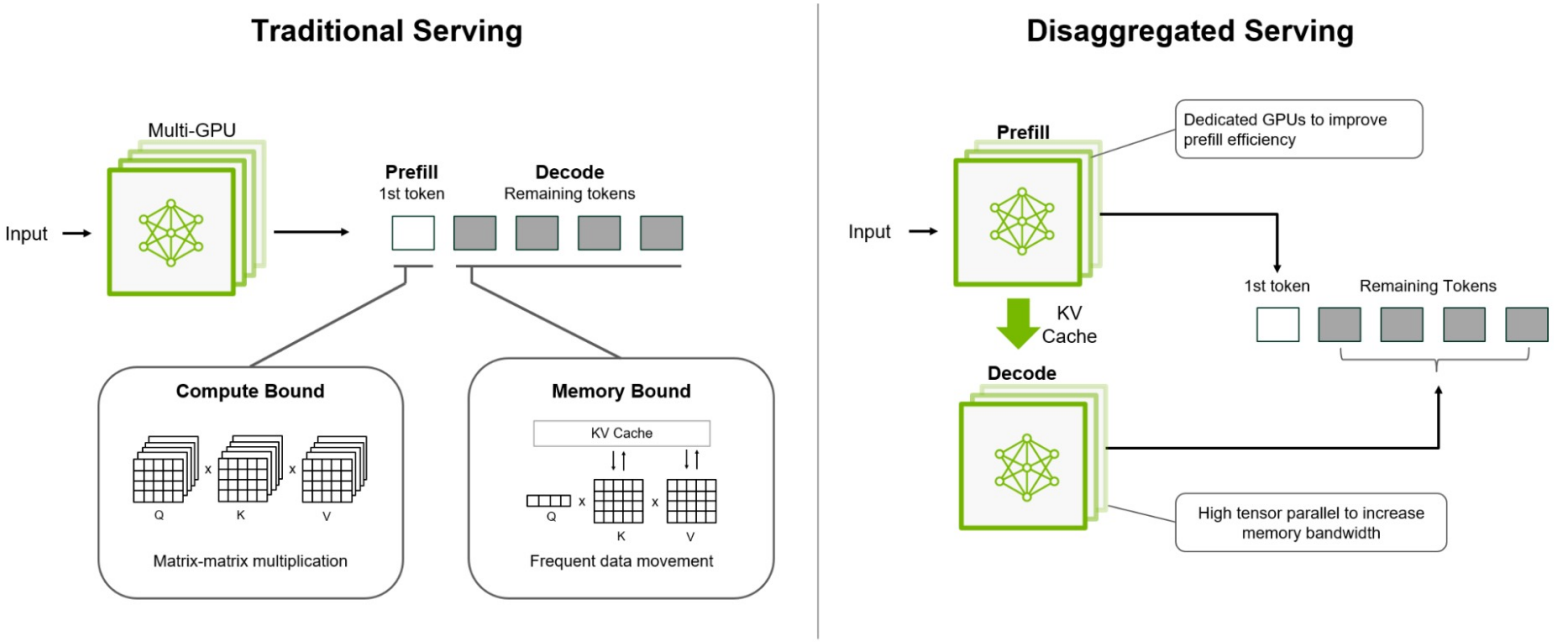


Fig. 3. Execution timeline of each AI agent.

Jiin Kim et al., The Cost of Dynamic Reasoning: Demystifying AI Agents and Test-Time Scaling from an AI Infrastructure Perspective, IEEE HPCA'26

Disaggregated Serving: Concept



[1] NVIDIA Technical Blog

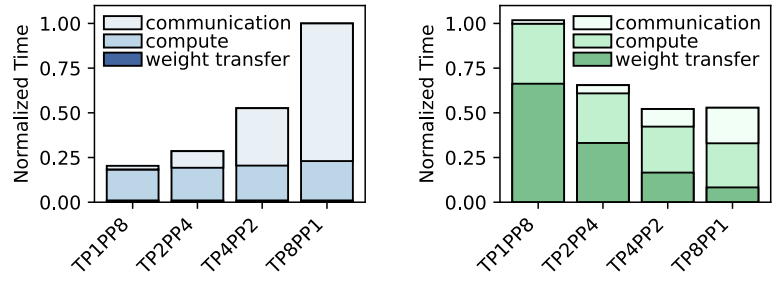


Figure 1. Breakdown of execution time for the prefill and decode stages for LLaMA2-13B inference on 8×L4 GPUs connected with PCIe. Communication overhead increases with the degree of tensor parallelism. (The global batch size is 16. Pipeline parallelism further divides the data into micro-batches of size 16/PP to fully utilize pipelining).

[2] Qidong Su, et al. "SEESAW: HIGH-THROUGHPUT LLM INFERENCE VIA MODEL RE-SHARDING.", Proceedings of 8th MLSys Conference, 2025

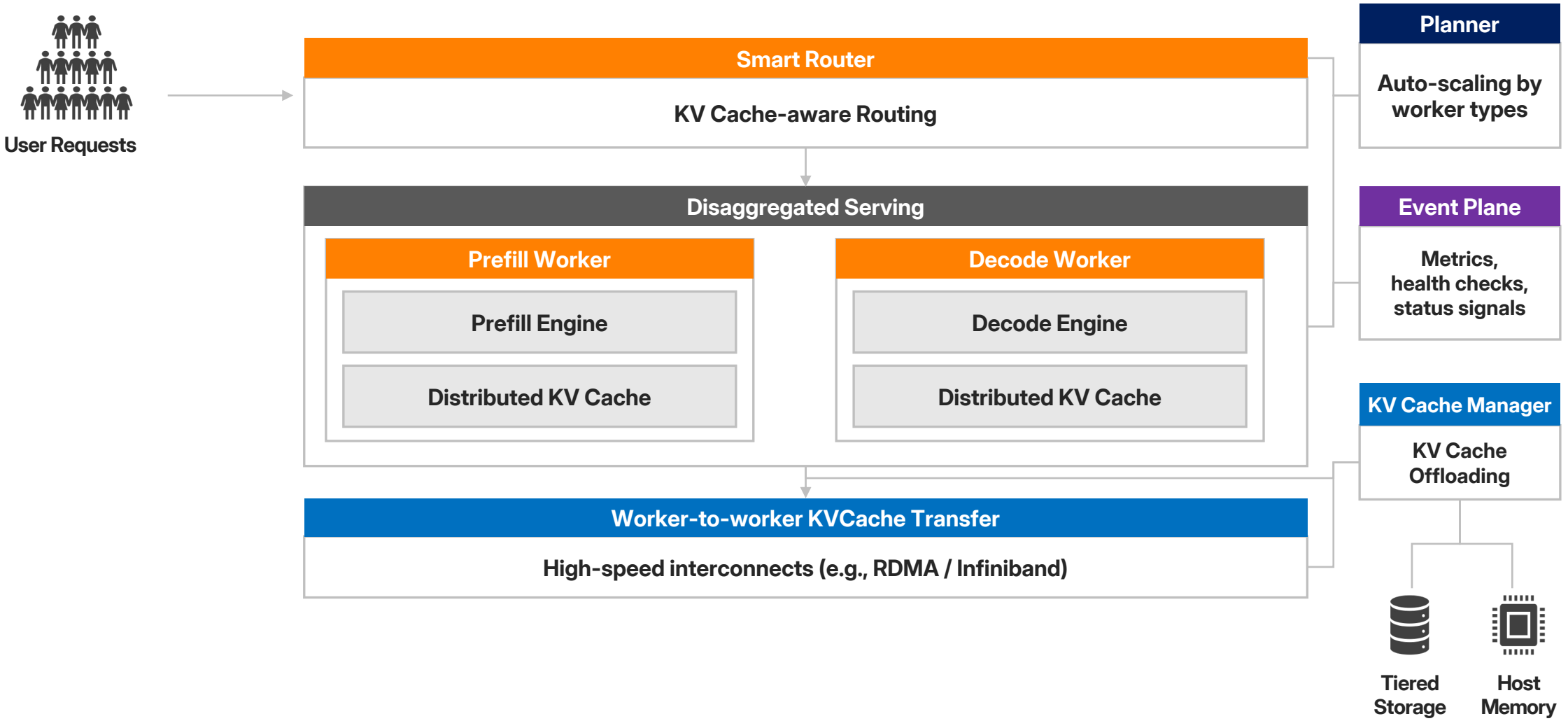
- Two-phases of transformer-based LLM: Prefill + Decode
  - Prefill: processing input tokens
    - ✓ Compute-bound / TTFT (time to first token)
  - Decode: generating output tokens
    - ✓ Memory-bound / TPOT (time per output token)

Running both phases in the same GPUs results in idle chip resources

Let's use dedicated chips and scale differently for each phase!

AI for Good

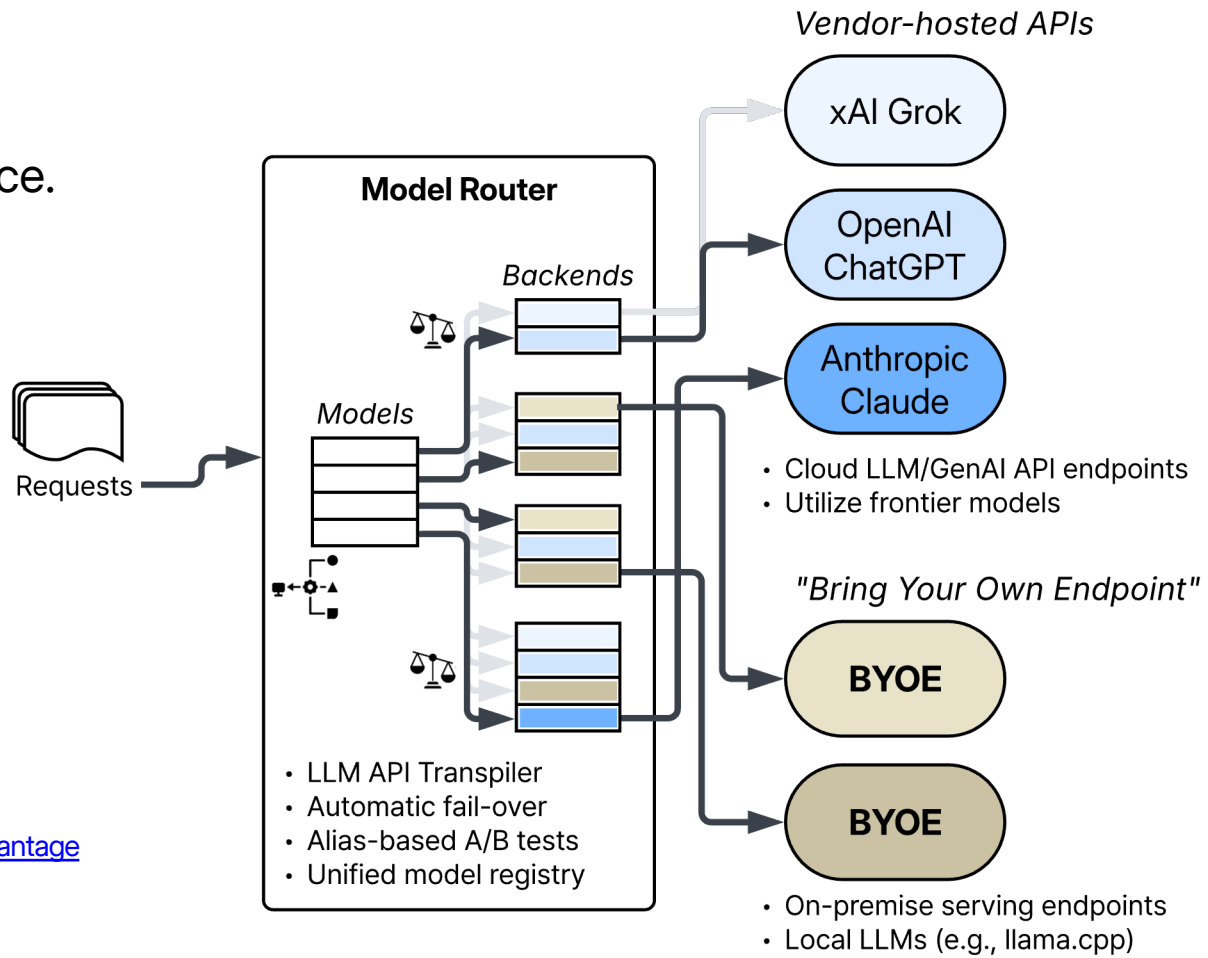
Disaggregated Serving: Architecture



Model Router

- Rising as an essential component of AI agent platforms
- Motivation
  - No single model fits all needs.
  - Agent is costly — running a continous loop of inference.
  - Service providers need to solve production reliability.
  - Enterprise requires security & compliance.
    - ✓ PII filtering, human escalation, etc.
- Roles
  - Model selection incl. fail-over and balancing
  - LLM API transpiler (universal endpoint)
  - Cost-quality optimization
  - Agent-aware routing
  - Observability and auditing

† https://www.mckinsey.com/capabilities/quantumblack/our-insights/seizing-the-agentic-ai-advantage
‡ Hu et al., RouterBench: A Benchmark for Multi-LLM Routing System [arxiv 2403.12031]
Ong et al., RouteLLM: Learning to Route LLMs with Preference Data [arxiv 2406.18665]



Recap

The future of AI native communication networks

Recap

- **AI native communication networks integrated with AI agents**
  - It needs a strong foundation of AI infrastructure.
- **AI infrastructure challenges**
  - Scaling strategies per layer
  - Continuously moving bottlenecks
  - Software supply chain management
- **AI infrastructure evolution**
  - Model training → Model serving → ***Agentic workloads*** → *Distributed (physical) agents*
    - ✓ Scaling experiences from Korea's sovereign AI foundation model project
    - ✓ New elements of cloud AI infra: disaggregated serving, model router
    - ✓ Physical AI: Communication networks as the fabric and interconnect itself
  - Continuously arising challenge: **Cross-layer co-optimization and problem solving**

AI Enterprise

AI Cloud

AI Open Source

AI MLOps

Thank you!