



Open Source for the AI Era

What We're Learning from the Next Generation of AI Infrastructure

Jonathan Bryce

Executive Director, Cloud & Infrastructure, Linux Foundation

Welcome to Agent Scale!



Matthew Prince 🌟🔵📷
@eastdakota · Follow



Welp, that happened faster than I predicted. Thought it would be end of 2027, then early 2027, but agentic traffic growing so fast that bots have now passed human traffic online for the first time in the Internet's history.



radar.cloudflare.com

Traffic Worldwide | Cloudflare Radar
Global Traffic trends and insights.

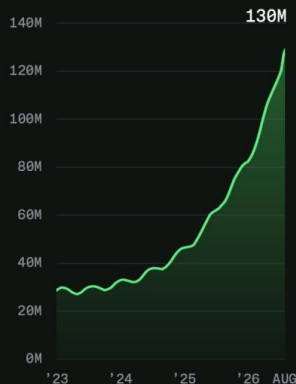
12:39 AM · Jun 4, 2026



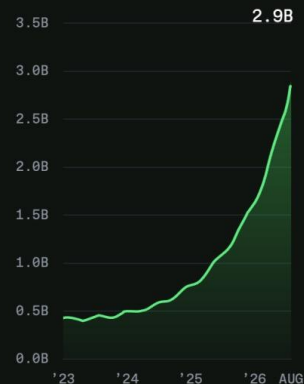
👍 8.2K 💬 Reply 🔗 Copy link

[Read 399 replies](#)

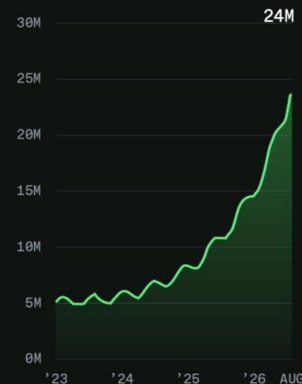
Merged pull requests per month



Commits per month



New repos per month



<https://github.blog/news-insights/company-news/the-august-17-outage-and-the-work-ahead>

Agent speed can work against—or for—a system

WIKIMEDIA

35%

of page views

65%

of resource-intensive core traffic

UNPLANNED MACHINE DEMAND

KLARNA

2/3

of customer-service chats

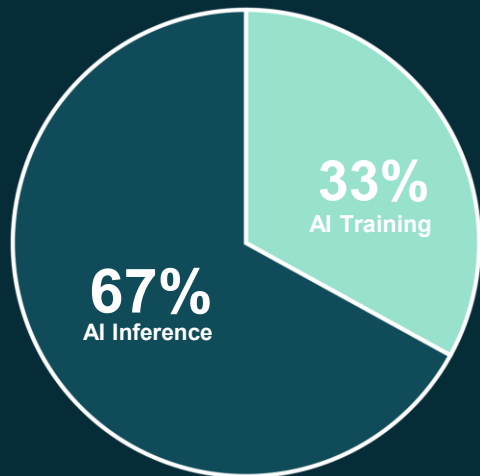
11 min → <2 min

resolution time

SYSTEM DESIGNED FOR AGENT LEVERAGE

A Shift in AI Workloads

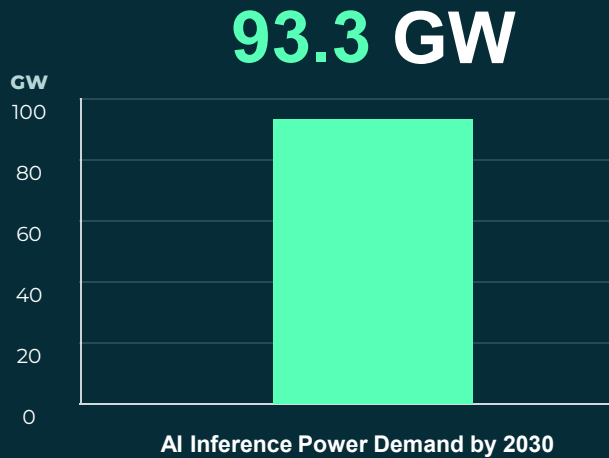
AI Compute Distribution by 2026



■ AI Inference (67%) ■ AI Training (33%)

Source: Slide data — McKinsey / CNCF KubeCon EU 2026 Keynote

AI Inference Power Demand Projection



Source: McKinsey Data Center Demand Model — KubeCon EU 2026 Keynote

What is Driving This?



TRAINING

Happens sporadically



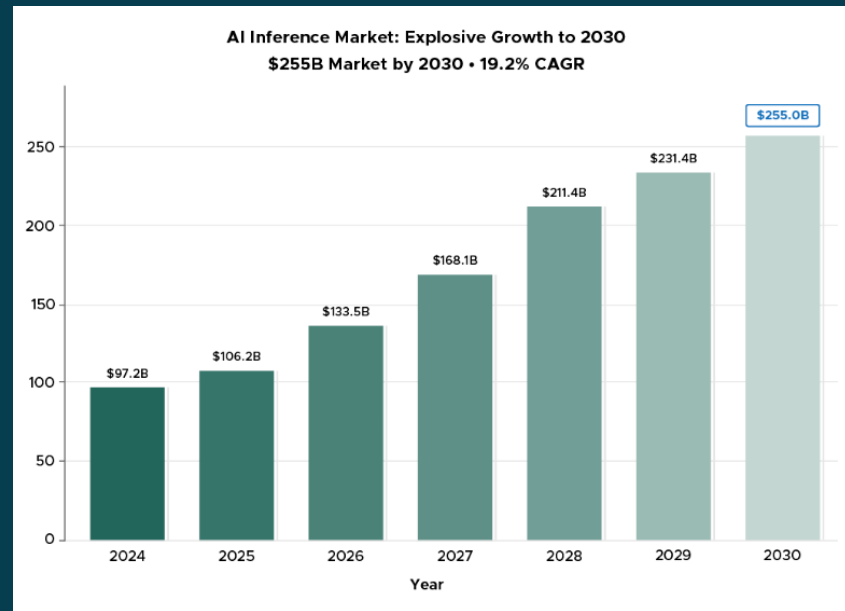
INFERENCE

happens on every prompt



AGENT WORKLOADS

are inference hungry



Source: McKinsey Data Center Demand Model — KubeCon EU 2026 Keynote

→ What's Happening?

AI is expanding from models into systems



Accelerator diversity

More hardware choices create more opportunities, but also more work to make each accelerator usable through familiar software.



Inference at Scale

AI is moving from experimentation into continuous production workloads where throughput, latency, cost and reliability matter.



Agents

Tool-using, multi-step applications multiply model calls and make routing, observability and operational resilience more important.

AI is Quickly Diversifying Compute – Software Must Not Fragment.



More models and architectures than ever

GPUs, NPUs, DCUs, edge silicon – new entrants every year; new open weights every week



One stack per chip does not scale

Every port re-done, every kernel re-tuned, every bug re-found



Portability is the prize

Models and tools that run across silicon, from data center to edge

These layers are not evolving independently. Open source is the only way to solve this problem at scale.

Source: "The Open Acceleration," Mark Collier, WAIC 2026, Shanghai, July 19, 2026 (slide 9).

OPEN SOURCE FOR THE AI ERA



AI APPLICATIONS & AGENTS

AGENTIC AI FOUNDATION



AGENTS



HARNESSES



APPLICATIONS



MCP

WHAT DEVELOPERS
AND ORGANIZATIONS
BUILD



BUILD & SERVE INTELLIGENCE

PYTORCH FOUNDATION



PyTorch



LLM



RAY

TRAIN, FINE-TUNE
AND SERVE MODELS
EFFICIENTLY



OPERATE & SCALE AI WORKLOADS

CLOUD NATIVE COMPUTING FOUNDATION



KUBERNETES



KSERVE



KUEVE



OPEN TELEMETRY



LLM-D

RUN, SERVE &
OBSERVE AT SCALE



INFRASTRUCTURE & ISOLATION

OPENINFRA FOUNDATION



OPENSTACK



KATA CONTAINERS

PROVIDE & CONTROL
INFRASTRUCTURE



HETEROGENEOUS COMPUTE



NPU



CPU



GPU



OTHER ACCELERATORS

THE HARDWARE
THAT POWERS AI

We did not build a stack We wired one **together**

Viettel Platform



L4 | Token-as-a-Service

One door in, tokens and a bill out.



Envoy AI Gateway



Envoy Gateway



Gateway API



Strimzi



L3 | Inference Serving

GPUhours become tokens



vLLM



PyTorch



KServe



ILM-d



Knative



SGLang



NVIDIA Dynamo



L2 | Orchestration

Cards become a shared pool



Kubernetes



HAMi



KEDA



Cilium



Prometheus



Argo CD



Harbor



Kubeflow



KAI Scheduler

Operating Frontier AI at Scale: Karmada Project Graduation



Case Study: Alibaba Cloud & Qwen

Alibaba Cloud deployed Karmada as a unified multi-cluster orchestration plane to manage its large-scale Qwen inference infrastructure.

Ecosystem: Karmada

CNCF Graduated Project: Multi-cluster, multi-cloud Kubernetes orchestration system proven at immense scale.

Production Footprint: Spans global telecom, public cloud, and AI platforms.

Scale: Frontier AI model operated as a production service for hundreds of millions of users.

Architecture: Distributed inference across tens of heterogeneous Kubernetes clusters and locations.

Performance: **+25–30% GPU utilization**, 99.99% availability, with sub-second latency through massive surges.

Cloud native technology is the critical foundation for reliability, efficiency, and scalability as AI workloads grow **beyond individual clusters** into massive distributed fleets.

SHOPIFY TURNS ITS AGENT'S PRODUCTION FAILURES INTO BETTER MODEL WEIGHTS EVERY DAY WITH PYTORCH & vLLM

Commerce for millions of merchants · PyTorch Foundation Platinum member

A merchant asks: “Which products are almost out of stock?”



Source: Shopify, “How Shopify built a continual learning loop with PyTorch” (forthcoming).

PYTORCH FOUNDATION MEMBERS BUILDING AND SERVING INTELLIGENCE

ANY MODEL ·

ANY CHIP ·

ANY CLOUD ·

ANY AGENT

AI LABS + BUILDERS



Creators of PyTorch Lightning



SILICON + SYSTEMS



寒武纪



CLOUDS + PLATFORMS



Creators of PyTorch Lightning



PRODUCTION + SERVICES

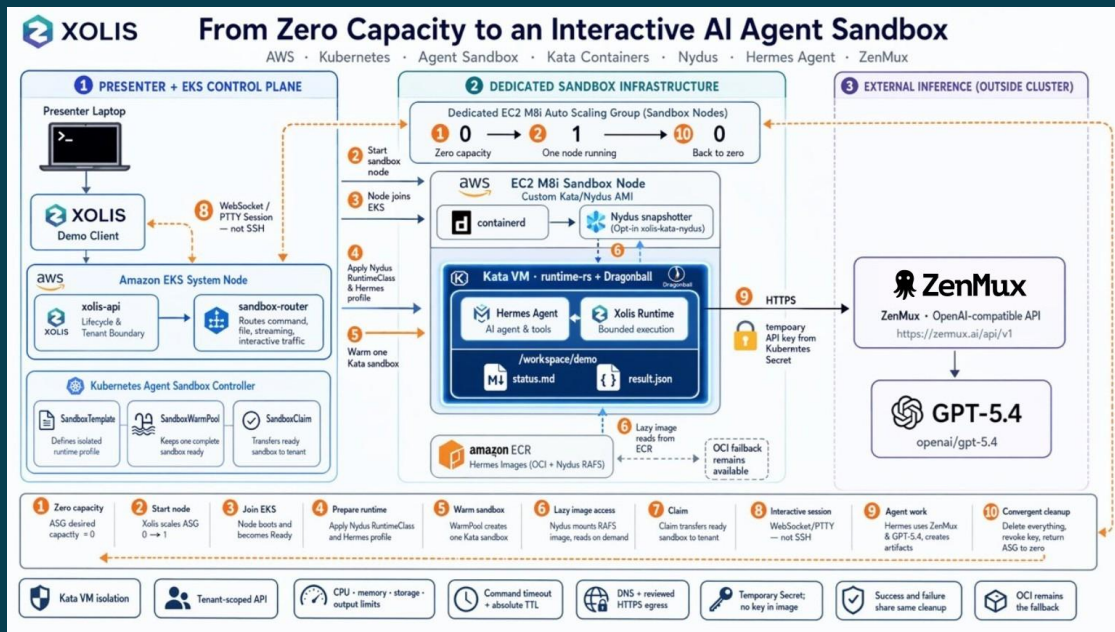


NO SINGLE ORGANIZATION BUILDS THE WHOLE SYSTEM. TOGETHER, OUR MEMBERS SPAN IT.

Platinum and Gold members shown · Repeated placement reflects work across layers · September 2026

Silver members: Graphcore · Nota AI · Rebellions Startup members: Clockwork · Emmi AI · Lablup · LightSeek · Yasp

Agent Sandbox with Open Infrastructure Stack



KubeCon Japan demo · July 2026

What the demo shows

PVM provides virtualization support for any cloud instance.

Kata Containers 4.0 with Dragonball runs the sandbox. The runtime stack converges on Rust.

Dragonfly distributes images peer to peer. Nydus accelerates startup with lazy loading.

K8s sig agent-sandbox for warm pool management.

Resources

Demo project: <https://github.com/gnawux/xolis>

Kata blog: <https://katacontainers.io/blog/kata-sandbox-demo-on-kubecon-jp-2026/>

OPEN SOURCE FOR THE AI ERA



AI APPLICATIONS & AGENTS

AGENTIC AI FOUNDATION



AGENTS



HARNESSES



APPLICATIONS



MCP

WHAT DEVELOPERS
AND ORGANIZATIONS
BUILD



BUILD & SERVE INTELLIGENCE

PYTORCH FOUNDATION



PyTorch



LLM



RAY

TRAIN, FINE-TUNE
AND SERVE MODELS
EFFICIENTLY



OPERATE & SCALE AI WORKLOADS

CLOUD NATIVE COMPUTING FOUNDATION



KUBERNETES



KSERVE



KUEVE



OPEN TELEMETRY



LLM-D

RUN, SERVE &
OBSERVE AT SCALE



INFRASTRUCTURE & ISOLATION

OPENINFRA FOUNDATION



OPENSTACK



KATA CONTAINERS

PROVIDE & CONTROL
INFRASTRUCTURE



HETEROGENEOUS COMPUTE



NPU



CPU



GPU



OTHER ACCELERATORS

THE HARDWARE
THAT POWERS AI