CLOUD NATIVE
COMPUTING FOUNDATION

Open Source Infrastructure for the AI Native Era

Jonathan Bryce
Executive Director, CNCF

Cloud Native Today: A Global Community

In the past 10 years,

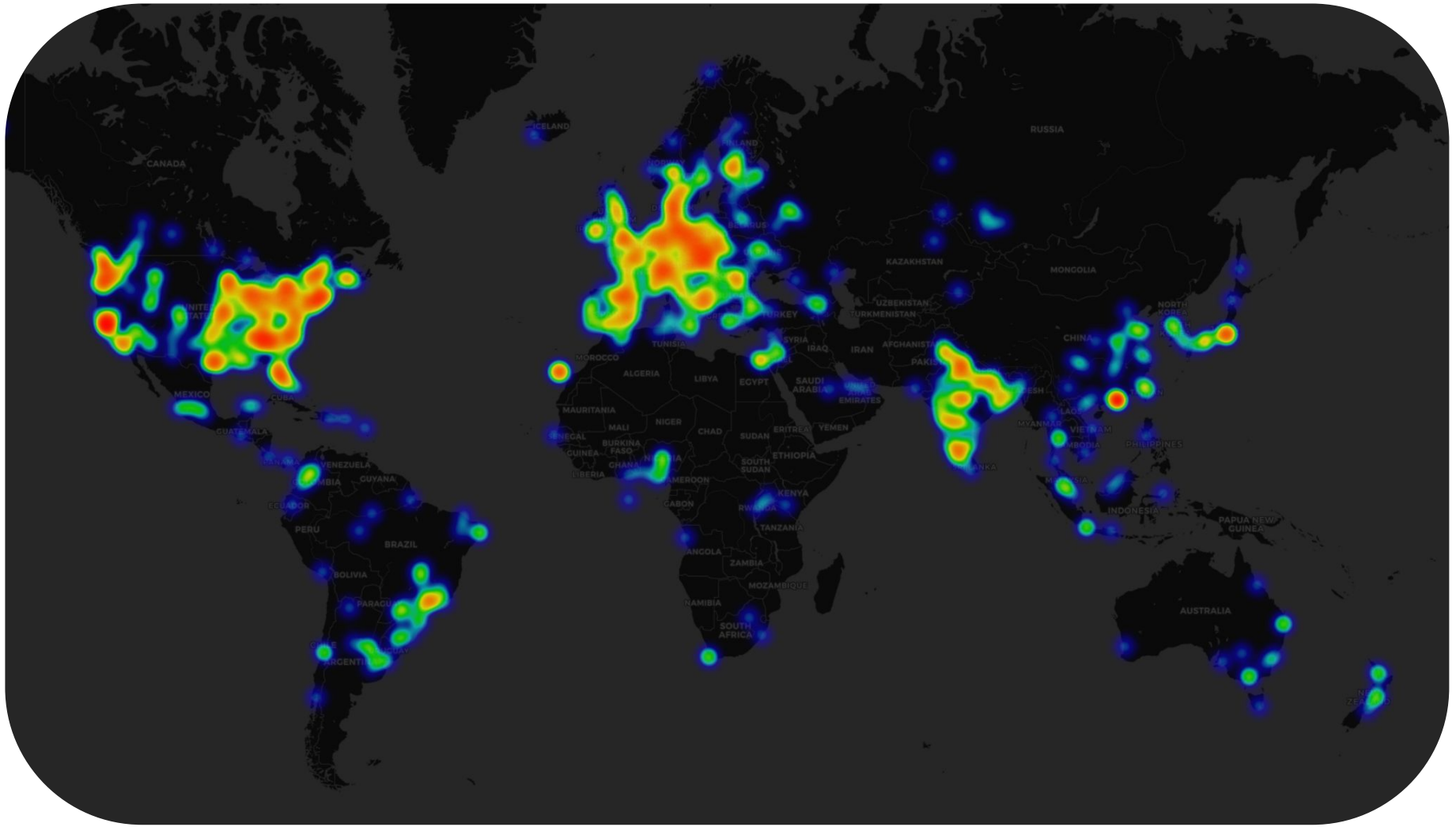
300,000+ contributors made **3,350,000+ code commits**

1,200,000+ pull requests and **18,800,000+ contributions**

Across 223+ Projects and **190+ countries**

Kubernetes trails Linux (33 years old) as the highest velocity project
github.com/cncf/velocity

CNCF 300,000+ Contributor Heatmap



3 Pillars of Open Source AI Today

Agents are getting hyped, *but we can't have agents without* ***models*** *and a* ***massive footprint of inference*** *computing*

	Training	Inference	Agents
	Model Development	*Model serving*	*Connect models to humans, tools, apps, data, other agents*
Key Projects	PyTorch, DeepSpeed, Volcano, kubeflow... (dozens more)	vllm, ONNX Runtime, Kserve, LMCache, Triton Inference, AIBrix, llm-d, SGLang, (dozens more)	MCP Agent2Agent Agent Gateway (MANY more every day)
Market Trends	PyTorch has **80% share** of training on Huggingface	**1.3 quadrillion** tokens/month (google) Cost to serve down **33X/year**	**<5%** Production Agent Adoption

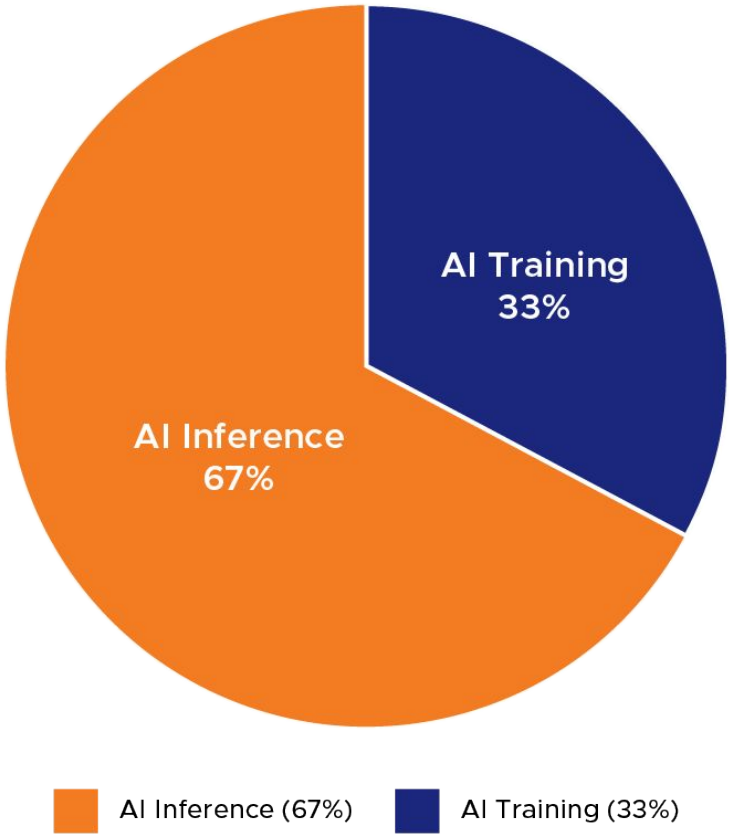
From Massive Training to Mainstream Inference

Waiting for Tokens

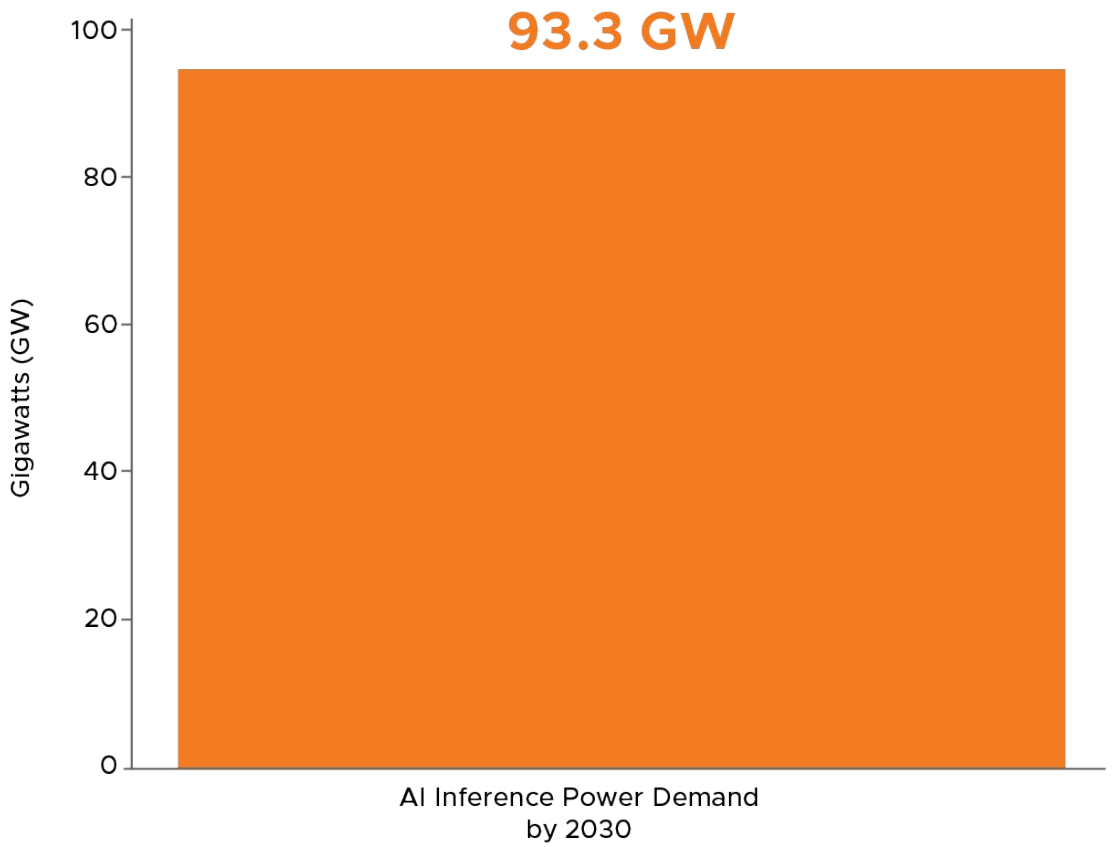
- 2 ways to deliver more tokens: more inference, faster answers
- Enterprises will take both approaches
- More inference means someone has to deploy, scale, secure and observe hundreds of models deployed by dozens of different teams

A Shift in AI Workloads

AI Compute Distribution by 2026



AI Inference Power Demand Projection



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

Specialized Models Will Drive Additional ROI

- Uber: “Currently, approximately 400 active ML projects are managed on Michelangelo, with over 20K model training jobs monthly. There are more than 5K models in production, serving 10 million real-time predictions per second at peak.”
- Most enterprises won't use one giant model. They'll use hundreds of smaller, fine-tuned, open-source models for specific tasks (sentiment analysis, code gen, contract review).

Why Will Specialized Models Matter?

- **Cost-Effective:** Vastly cheaper to run and fine-tune
- **Performance:** Faster and often more accurate for a specific domain
- **Access to hardware:** Do not require the largest, latest and most scarce GPUs for inference
- **Security & Privacy:** Can be self-hosted, on-prem, or in a cloud

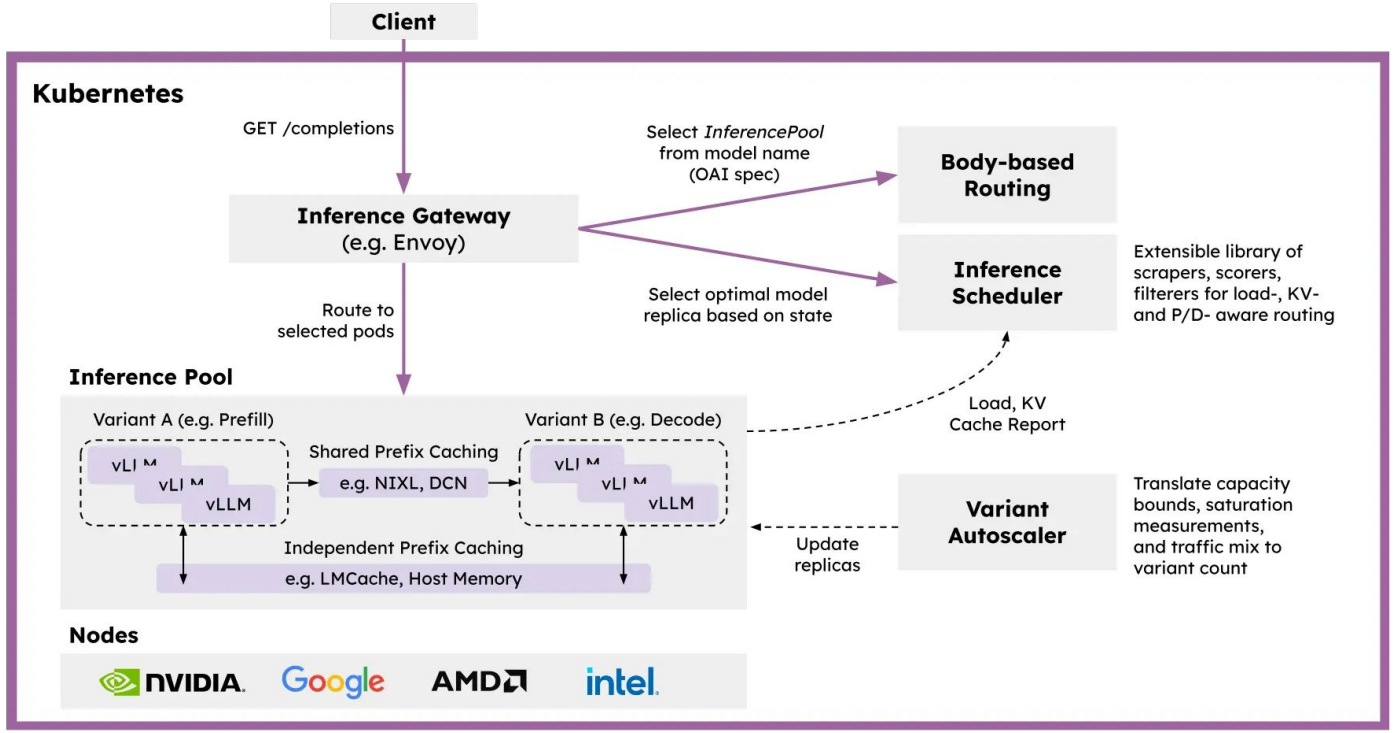
AI Inference: Next Big Cloud Native Workload

- Cloud Native Inference systems will be added to the list of other apps, microservices, databases and systems Platform Engineering teams are responsible for because it needs...
  - Standardized deployment
  - Auto-scaling (including scaling to zero)
  - Security policies and enforcement
  - Observability (cost, performance, accuracy)

Cloud Native Inference Platforms

- **New project: llm-d**
  KV-Cache Centric Inference: Building a State-Aware Serving Platform With Llm-d and VLLM

  11:05 in Central Room



We want agents, we need inference

- AI infrastructure is moving from a few "Training Supercomputers" to widespread "Enterprise Inference"
- We see many organizations adding AI Native systems to their existing Cloud Native footprints
- Open Source will win if we build strong communities, working on these challenges together