All Things Open

Open-Source powered and Developer-First AI Stack

Koyeb

Yann Léger @yann_eu / Co-Founder @gokoyeb / koyeb.com

NetActuate

Greg Wallace / netactuate.com

Yann Leger

Co-Founder & CEO at Koyeb

@yann_eu

14 YEARS BUILDING CLOUD INFRASTRUCTURE

Greg Wallace

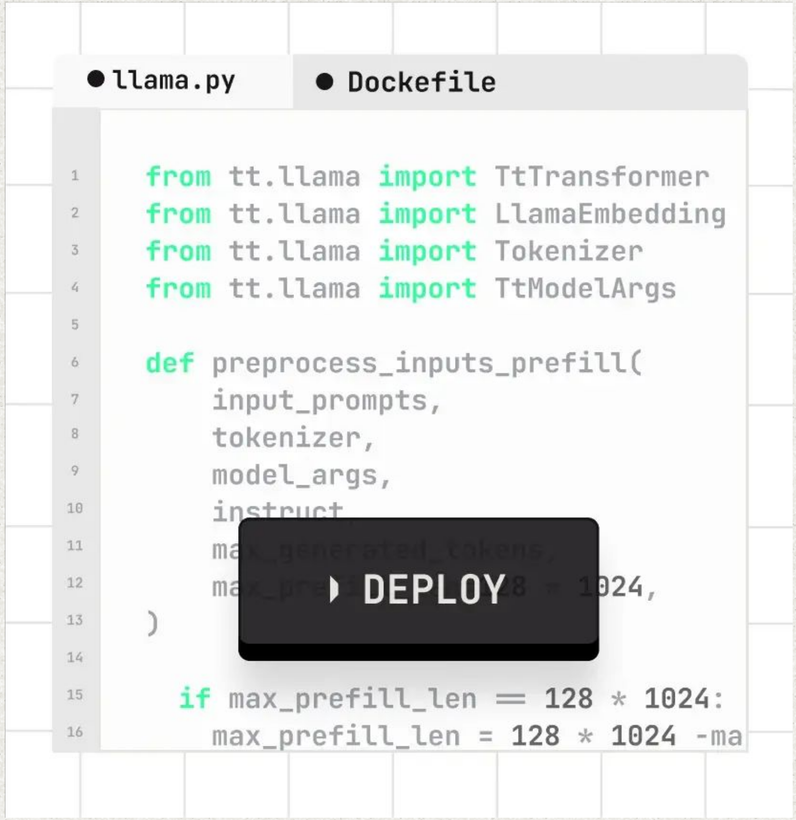
Partnerships at NetActuate

in/gtewallace

15 YEARS IN OPEN SOURCE

Developer First & Open Source Powered AI Infrastructure

Agenda



Industry context
The role of open source
Demo
Evolving infrastructure
Q&A

$1.0T by 2029 in global data center capex as AI accelerators drive spend

Dell'Oro Group

MIT report: 95% of generative AI pilots at companies are failing

Yann Léger @yann_eu / Co-Founder @gokoyeb / koyeb.com

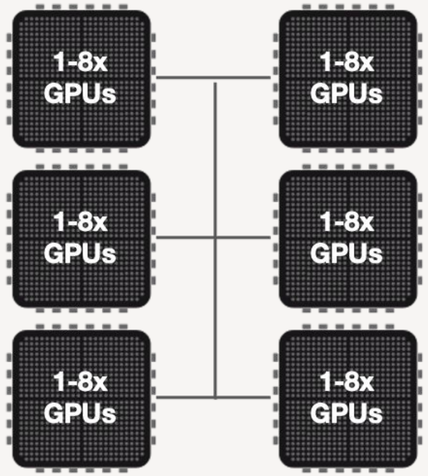
Most organizations miss AI cost forecasts, with nearly a quarter busting their budgets by more than 50% (CIO)

There’s room for improvement

What is AI Infrastructure?

NVIDIA

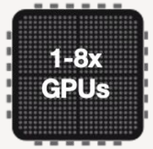
Large Training



Up to 100K GPUs clusters with high-speed interconnect

NVIDIA

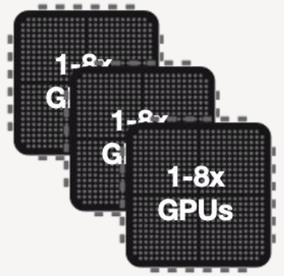
Fine-Tuning & Small Training



1x to 8x GPUs or Small Clusters

NVIDIA

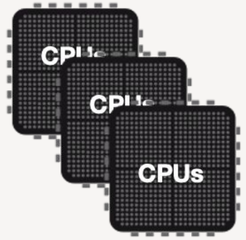
Inference



Fits one machine, needs more machines to handle parallel requests

AMD

Agents



High-volume, lightweight sandboxed code execution

AI Infra is a lot of GPUs

Koyeb NetActuate

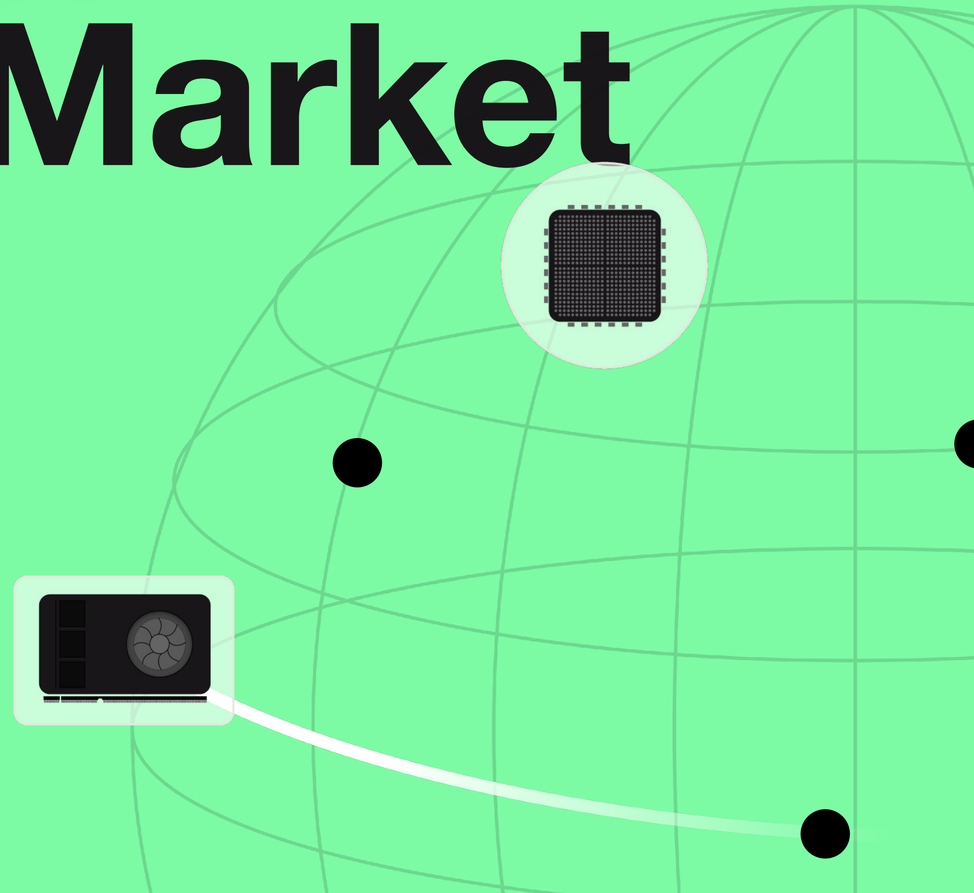
Yann Léger @yann_eu / Co-Founder @gokoyeb / koyeb.com

Maybe we need more choice?

New AI Accelerators are Emerging

Building for the Agentic Era

Introducing Open Accelerators to Market



Koyeb NetActuate

Yann Léger @yann_eu / Co-Founder @gokoyeb / koyeb.com

We build a Global Serverless Platform for Agents and Inference

Koyeb

Yann Léger @yann_eu / Co-Founder @gokoyeb / koyeb.com

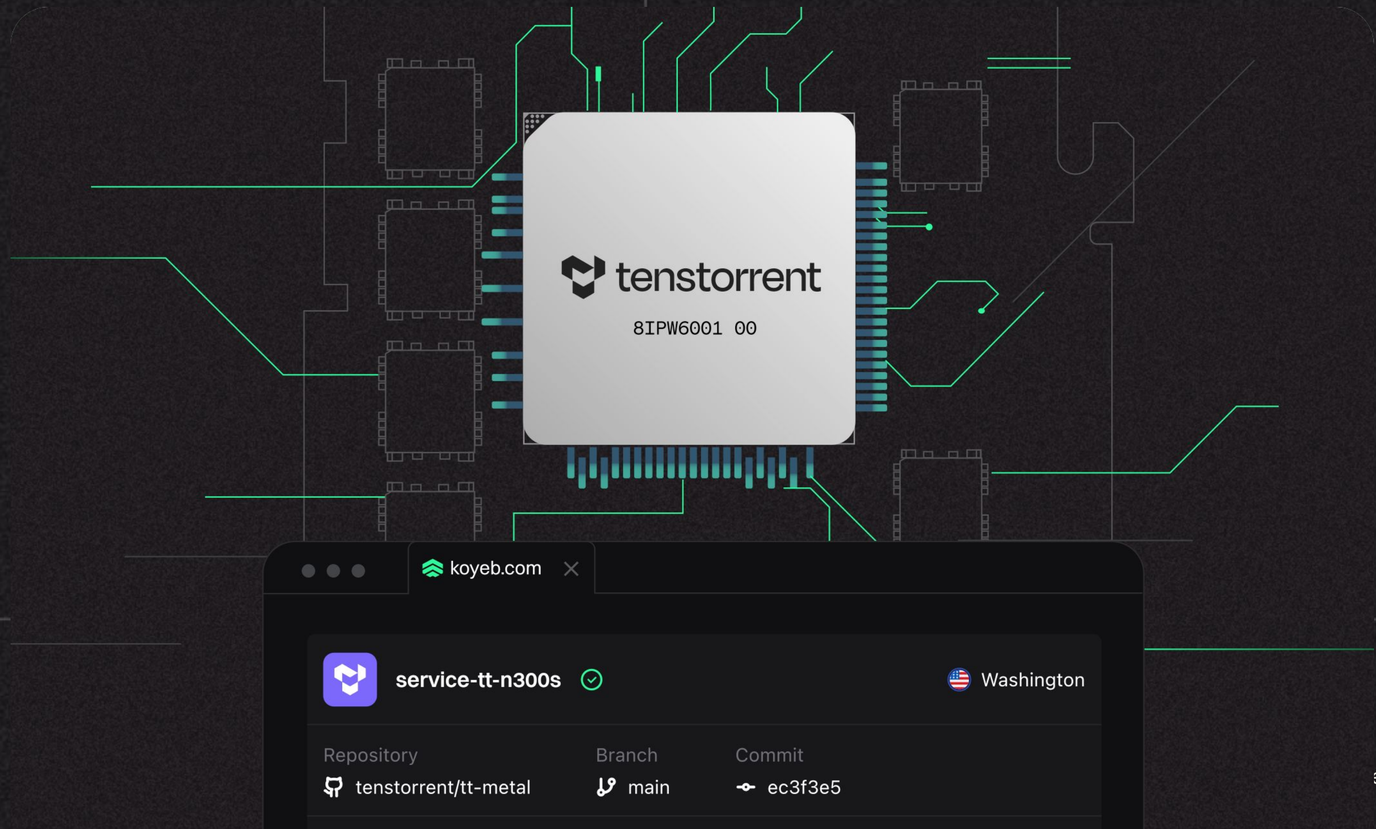
We Run Models and Agents

- **Model & Agents continuous deployment with Containers**
- Zero infrastructure management
- Low latency with global deployments
- **High-performance hardware**
- Sustainable costs with Serverless capabilities
- **Native resiliency and scalability**

Koyeb NetActuate

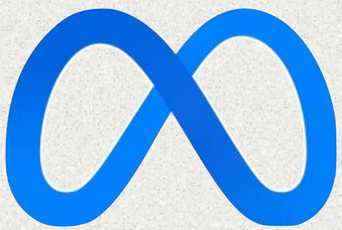
Yann Léger @yann_eu / Co-Founder @gokoyeb / koyeb.com

Available in Private Preview only on Koyeb



Koyeb

Deploy LLAMA 3.2 11B Vision On Tenstorrent



github.com/koyeb/tenstorrent-examples

Yann Léger @yann_eu / Co-Founder @gokoyeb / koyeb.com

Unified and global experience across CPUS, GPUS, and Accelerators

tenstorrent

AMD

NVIDIA

intel

Koyeb

Yann Léger @yann_eu / Co-Founder @gokoyeb / koyeb.com

THE STACK

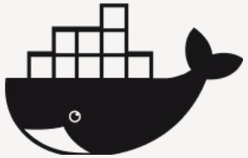
CONTAINERS IN VMS ON BARE METAL

Orchestration

Koyeb

Inference Engine

Container & Packaging



Networking

Virtualization

BareMetal & Accelerator

Koyeb

Software is key to Performance and Efficiency

Yann Léger @yann_eu / Co-Founder @gokoyeb / koyeb.com

New Accelerators Need New AI Stacks

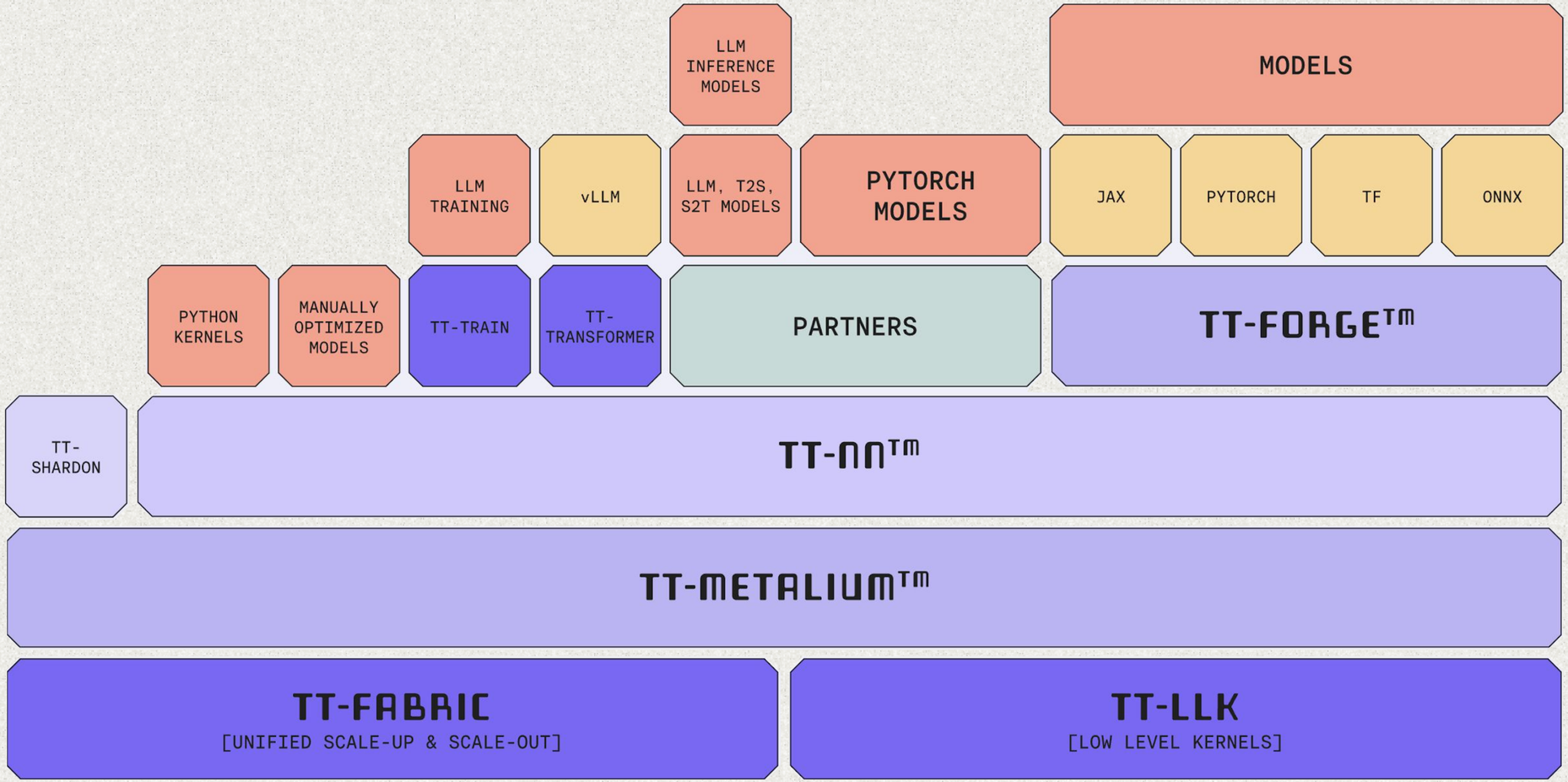
Proprietary

Open source on open hardware

Open source on closed hardware

Open standard

Tenstorrent provides a fully open stack



How can we build?

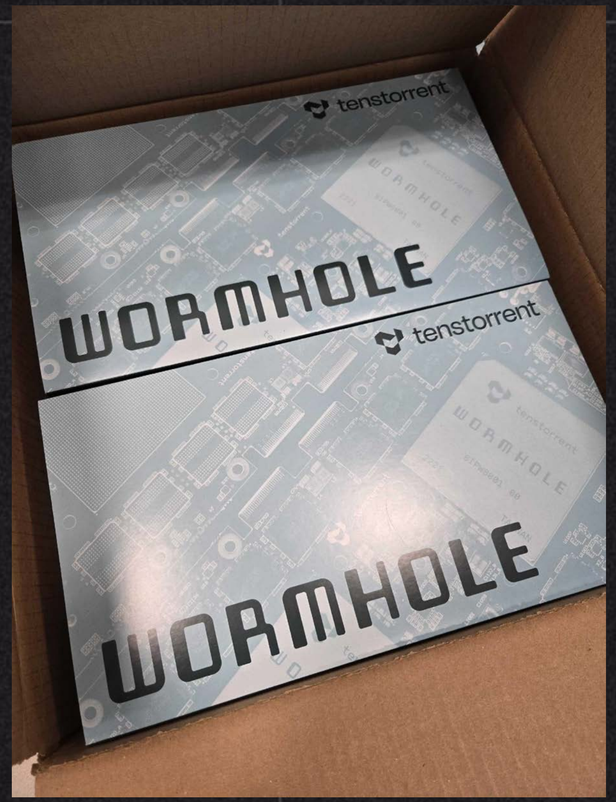
Instantly access development environments
with Tenstorrent open-source SDKs, including
TT-Metalium on n300s or TT-Loudbox

github.com/koyeb/tenstorrent-examples

Powered by Open Hardware

Koyeb NetActuate

Yann Léger @yann_eu / Co-Founder @gokoyeb / koyeb.com


tenstorrent
WORMHOLE
tenstorrent
WORMHOLE


BLACKHOLE
BLACKHOLE
BLACKHOLE
tenstorrent
BLACKHOLE
p150a
BLACKHOLE TENSIX PROCESSOR
BLACKHOLE TENSIX PROCESSOR


tenstorrent

Koyeb

NetActuate Global Footprint

NetActuate

Yann Léger @yann_eu / Co-Founder @gokoyeb / koyeb.com

GPU hardware is vastly underutilized

Use Serverless

Automatically adapt infrastructure to demand

- SCALE-TO-ZERO
- AUTOSCALING

Deploy on Tenstorrent in seconds

tenstorrent

Koyeb

Yann Léger @yann_eu / Co-Founder @gokoyeb / koyeb.com

Thank you.

GET STARTED WITH KOYEB

Chat with our Team

Koyeb

NetActuate

Yann Leger

yann@koyeb.com
/in/yannleger
@gokoyeb

$200 on us