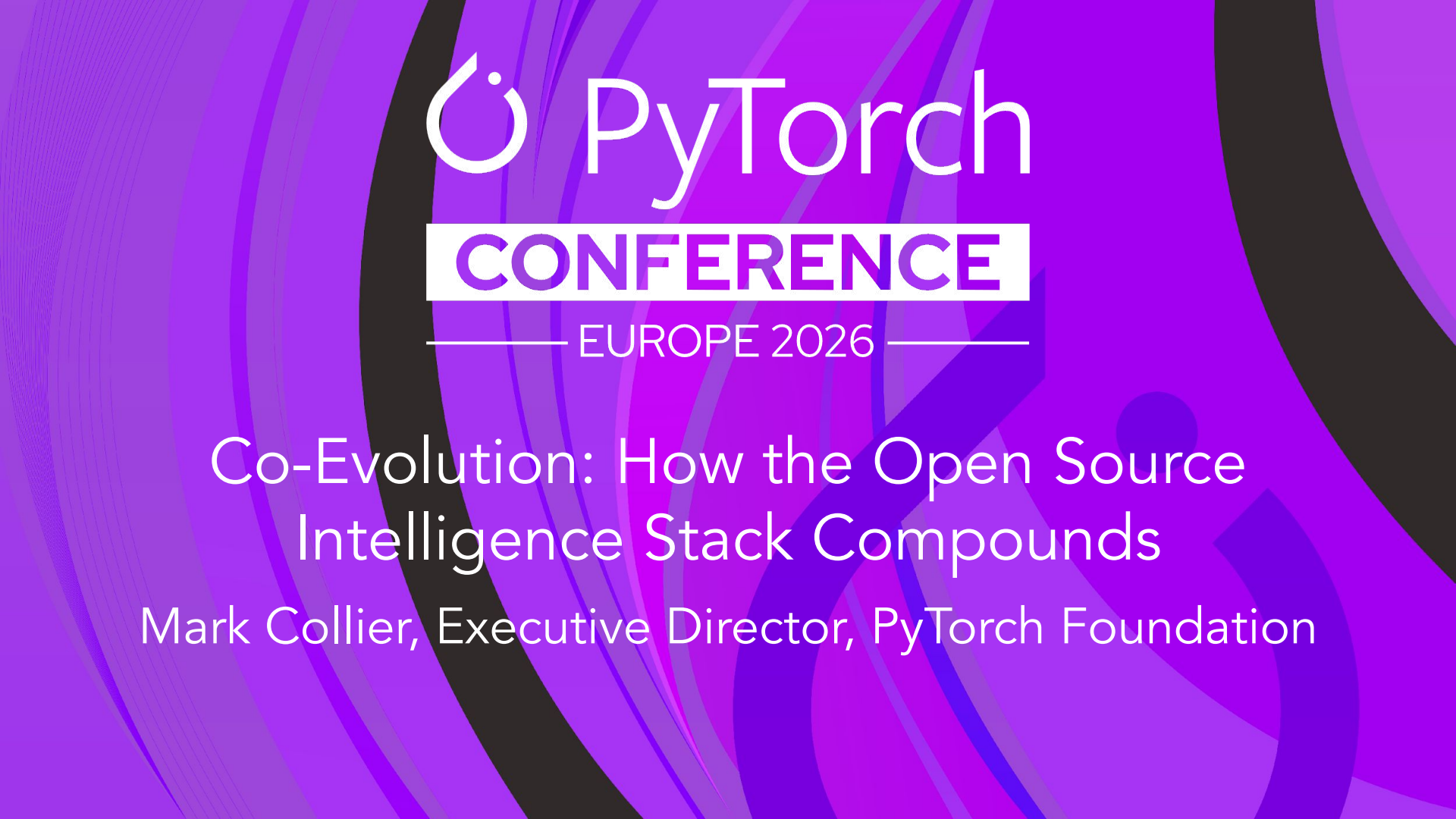


Co-Evolution: How the Open Source Intelligence Stack Compounds

Mark Collier, Executive Director, PyTorch Foundation

PyTorch Members

Platinum

AMD, aws, arm, Google, HUAWEI, intel, IBM, Lightning AI (Creators of PyTorch Lightning), Meta, NVIDIA, snowflake

Gold

DELL Technologies, Hugging Face, Microsoft, QUALCOMM

Silver

CLOCKWORK.io, Emmi AI, GRAPHCORE, nipa 정보통신산업진흥원, Nota AI, rebellions_, yasp

Associate

BAAI 智源研究院, Carnegie Mellon University, Columbia University, COMMON CRAWL, CommonAI, Georgia Tech, MESSIAH UNIVERSITY, MONASH University, onefact.org, Rensselaer, SKOLTECH, TEXAS A&M UNIVERSITY, ISAAC NEWTON, UCI, University of Leicester, William Carey University, WEDF

PYTORCH FOUNDATION EXISTS TO SERVE THIS COMMUNITY

Today PyTorch Foundation hosts 4 critical open source projects:

PyTorch  deepspeed vLLM RAY

To meet the moment, we need to do three things

1. Add capabilities to existing PyTorch Foundation Projects
2. Bring on new foundation hosted projects
3. Work well with the wider open source ecosystem to solve hard problems

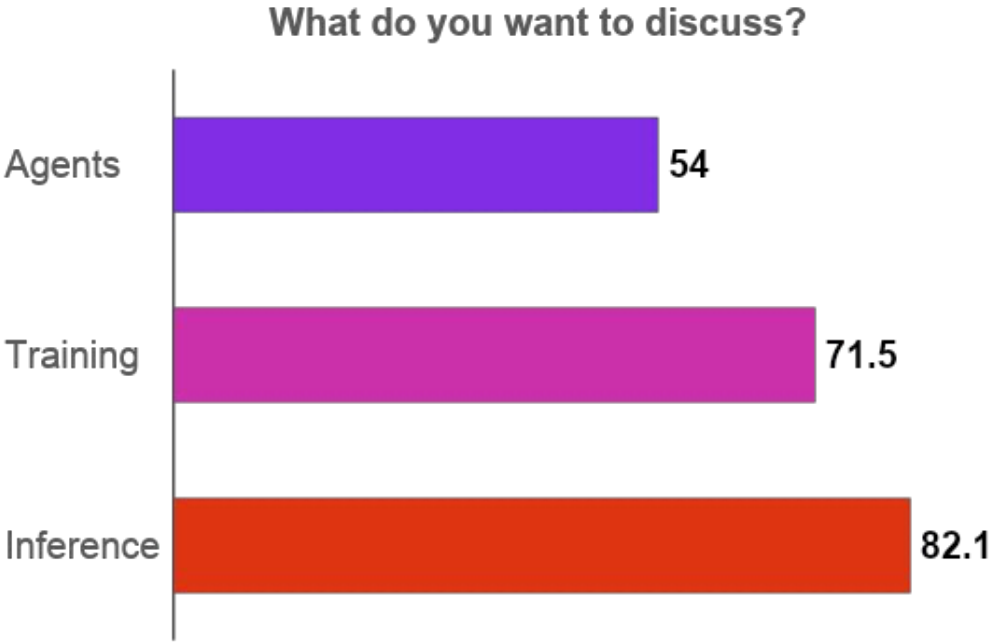
THE COMMUNITY HERE TODAY

250+

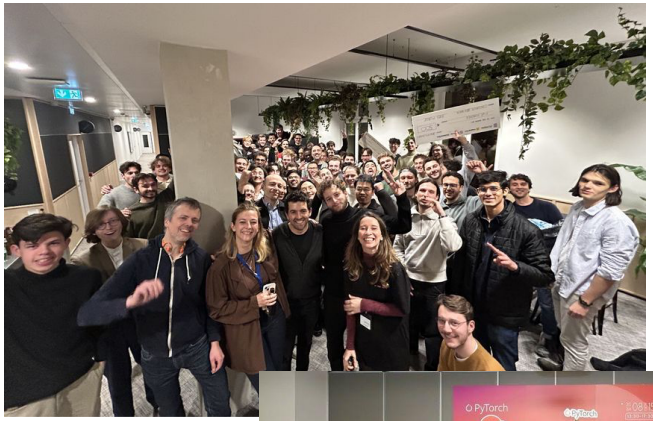
Organizations

45+

Countries

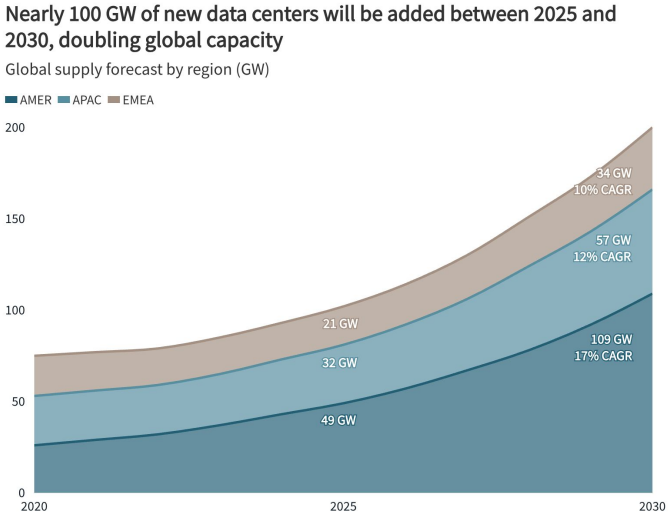


THIS WEEK IS ALL ABOUT COMMUNITY



HUMANITY IS MAKING AN ENORMOUS BET

100 gigawatts of data centers filled with accelerators hoping to create and deliver machine intelligence.



Hardware alone creates more complexity than intelligence

For the first time we're changing hardware and software architectures in fundamental ways at the same time

Hardware is diversifying

GPUs, TPUs, custom ASICs, inference chips, and CPUs are all competing to train and serve useful intelligence.

Model architectures are changing fast

Inference engines have to keep up with a software moving target while hardware is also moving underneath them.

Software creation is changing

Coding agents are altering how software gets written, evaluated, and improved.

Hardware creation is changing too

Automated layout, simulation, and design systems are beginning to reshape how chips themselves get built.

These domains are not evolving independently. They are co-evolving.

When Everything Co-Evolves, Coordination Becomes the Bottleneck

Models change fast

New architectures arrive monthly. Each one changes what the inference and training stack needs to support.

Hardware diversifies

GPUs, TPUs, custom ASICs, and inference chips are all competing. Each needs a software path to market.

Agents multiply calls

Agentic workflows call inference thousands of times. Cost and reliability become existential.

Coordination is the bottleneck

Without shared layers that keep training, inference, and hardware in sync, the whole system fragments.

The only proven way to coordinate work at this scale is open source

MODELS ARE PROLIFERATING AND SPECIALIZING FAST

- Kimi cadence: K2.5 → K2.6 → K3
- Qwen3.5-27B: near 397B on agent coding benchmarks
- Cursor: product layer absorbing rapid model iteration
- Uber: training and running thousands of specialized models in production

PYTORCH OSS COMMUNITY IS THE EPICENTER OF THIS WORK

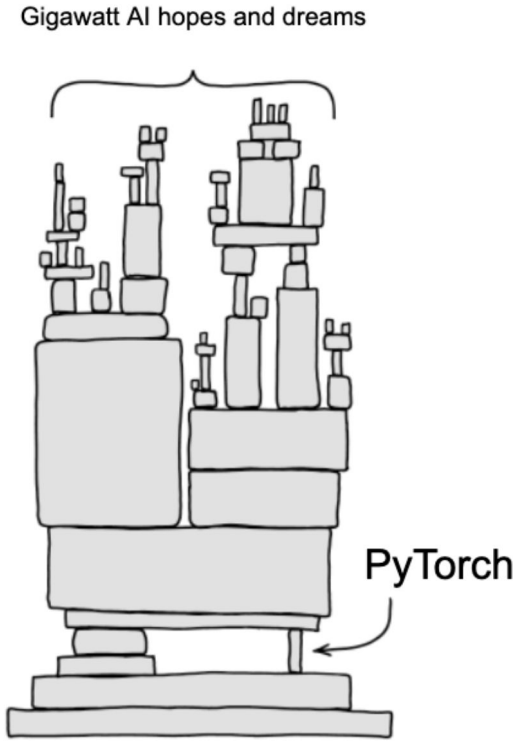
PyTorch

12,000+ Contributors

2,000+ Organizations

The AI accelerator supply chain is contributing to PyTorch:

Intel · NVIDIA · AMD · Google · Apple · Huawei · Arm · Qualcomm · Amazon · Microsoft · Habana · Graphcore · Cerebras · Groq · d-Matrix · FuriosaAI · Cambricon · Iluvatar CoreX · Enflame · Blaize · Syntiant · NeuReality · Recogni · SiMa.ai · Ampere · Rivos · Xilinx · MediaTek · Broadcom · Marvell · Micron · Infineon · Texas Instruments · Synaptics · Lightmatter · Lightelligence · Phytium · ChangXin Memory Technologies · Fujitsu Monaka Research



Source:https://insights.linuxfoundation.org/project/ptproject/contributors

vLLM OSS COMMUNITY ENABLES INFERENCE EVERY DAY

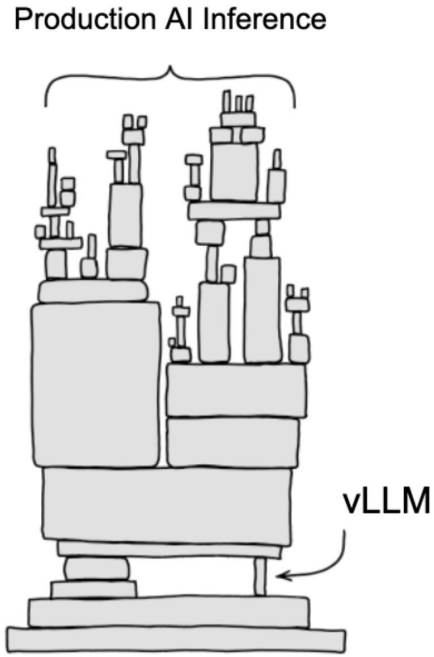
vLLM

12,000+

Contributors

1,300+

Organizations



The AI accelerator supply chain is contributing to vLLM:

Huawei · Intel · Google · AMD · Habana · NVIDIA · Microsoft · Arm · Samsung · Iluvatar CoreX · Tenstorrent · Graphcore · Moore Threads · FuriosaAI · Groq · Moreh · Enflame · Sanechips · NeuReality · Cambricon · Cerebras · d-Matrix · SAPEON

Source: https://insights.linuxfoundation.org/project/vllm/contributors

RAY POWERING RL SPECIALIZED MODEL REVOLUTION

RAY

4000+

Contributors

- UBER uses RAY to create thousands of specialized models
- CURSOR uses RAY to create a new model every 5 hours with real time RL

Source: https://insights.linuxfoundation.org/project/vllm/contributors

OPTIMIZATION DRIVES ECONOMICS

Example: DeepSeek-R1 on GB300:

- 2.77x higher token throughput in 6 months
- Same hardware, >60% lower token cost
- A model is born once, but optimized many times

- Once a model runs, software tuning and systems optimization can unlock massive gains in performance, efficiency, and cost.

An Agent Is Just a Loop Calling the Stack

Train & improve

PyTorch · DeepSpeed

Agents are only as capable as the models behind them. Training, fine-tuning, and adaptation remain foundational.

Serve

vLLM

Agents call inference not once, but maybe 100,000 times. Inference cost and reliability become central.

Execute & scale

Ray · Helion

Agentic systems create orchestration pressure, parallelism, new abstractions, and weird infrastructure behavior.

Agents are not replacing the stack. They are stress-testing every layer of it.

PyTorch Foundation: Now Hosts Multiple Critical OSS for AI

RAY

deepspeed

PyTorch

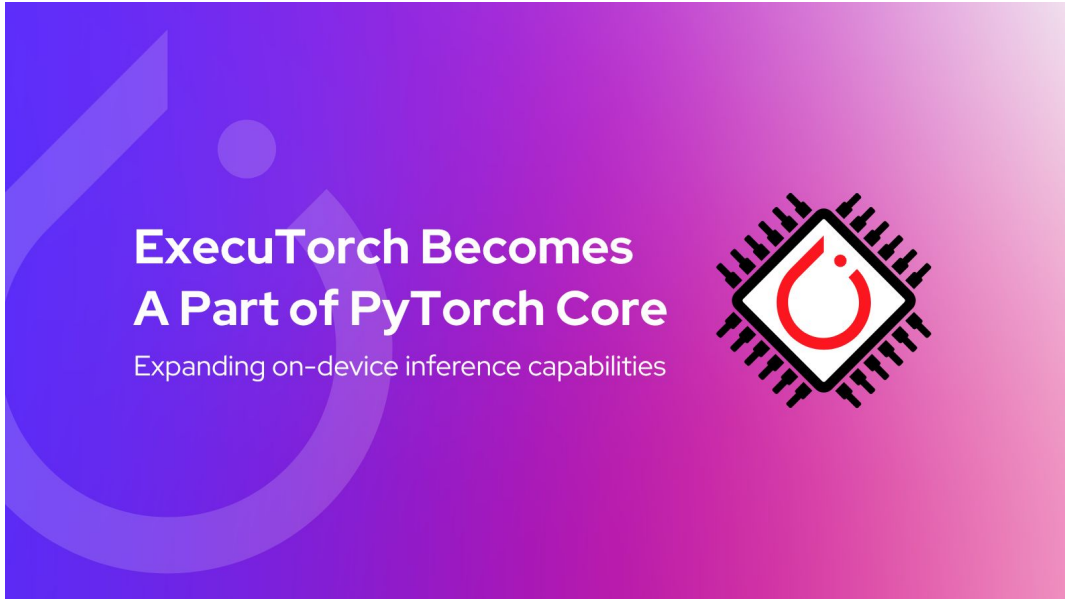
TBA

Coming soon!

TBA

ExecuTorch becomes a part of PyTorch Core

ExecuTorch extends PyTorch functionality for efficient AI inference on edge devices, from desktop/laptop to mobile phones and embedded systems



And the stack keeps expanding.

If coordination layers are where the leverage is, and the stack keeps adding new layers, then the Foundation should keep widening to meet them.

Today, we welcome **Helion** as the newest PyTorch Foundation project.

Helion joins the PyTorch Foundation as a hosted project

Helion eliminates bottlenecks associated with model architectures and execution, providing developers with radically simpler kernels, automated ahead-of-time autotuning, and greater hardware performance portability.

The Mission: Accelerate the Open Flywheel

Community compounds the stack

12,000+ contributors do not just maintain code. They are how breakthroughs become tested, optimized, portable infrastructure that the next breakthrough can build on.

Widen the coordination surface

As new layers become critical, the Foundation gives them a durable open home. Each new project extends the surface where the community can coordinate and co-design.

Keep the flywheel open

Proprietary stacks can move fast in one direction. Open stacks compound in every direction at once, because the community is doing the integration work across the full surface.

The Foundation’s job is to accelerate and widen that open flywheel.

The question is not whether this system will evolve. It already is.

The question is whether it keeps compounding in the open, and how fast.

Thank you

 @PyTorch

 PyTorch

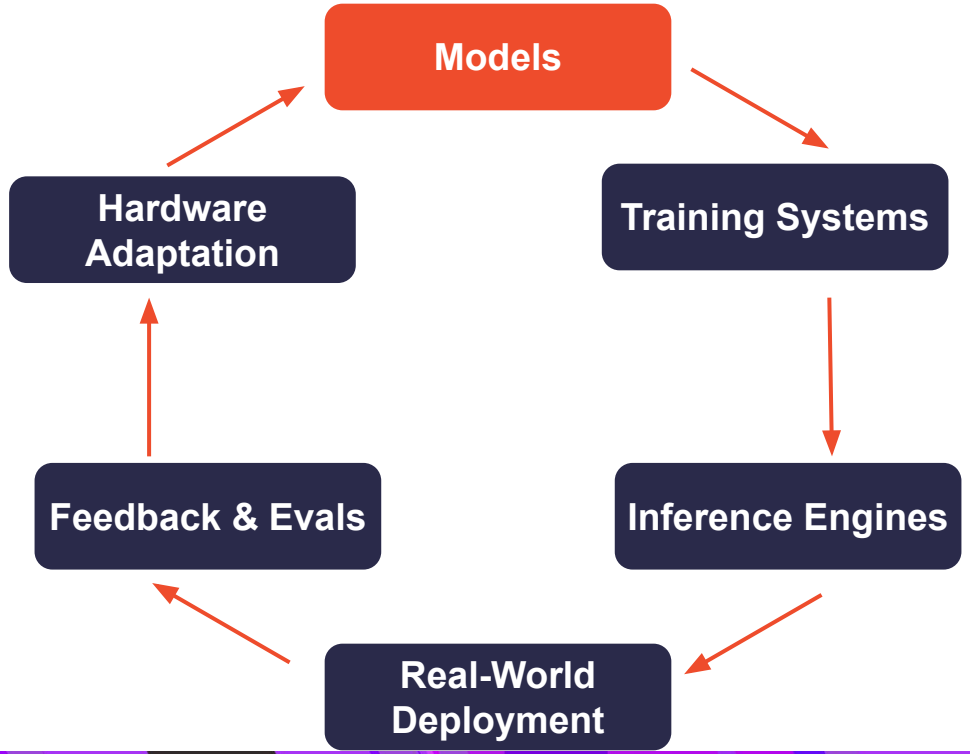
 PyTorch

PyTorch CONFERENCE EUROPE 2026

BACK UP SLIDES - HIDDEN

 @PyTorch  PyTorch  PyTorch

Co-Evolution Is a Loop, Not a Pipeline



The breakthrough is not just better models. It is a live syst
where each layer improves the others.

Why PyTorch and vLLM Matter So Much Right Now

PyTorch is the path to market for training hardware

If a new accelerator wants to matter for AI training, it has to work with PyTorch. That is why you see participation from NVIDIA, AMD, IBM, Google, and many others.

vLLM is becoming the path to market for inference

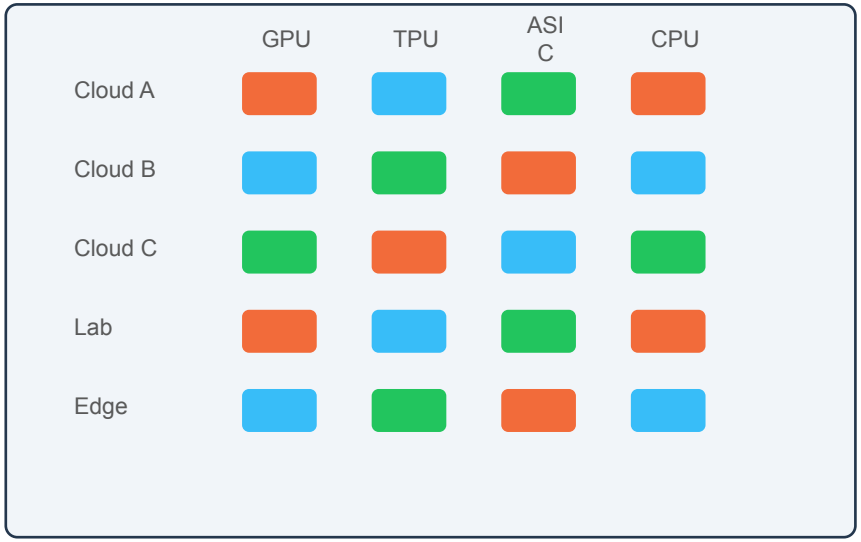
Serving modern models means handling rapidly changing model architectures on top of rapidly changing hardware. vLLM sits in the middle of that pressure.

When hardware diversifies and model architectures churn, coordination layers become strategic.

Community CI Is How Open Source Actually Compounds

Daily reality

- Community is how hardware diversity becomes usable
- Community is how the stack actually co-evolves in practice
- Community is how regressions get caught before they ship
- Community is how open source keeps pace when everything accelerates



The PyTorch CI is community managed and provides a critical role in ensuring accelerators can actually do useful things