



MAX and Mojo: The best way to extend PyTorch

Chris Lattner

EVP, Advanced AI Software and Platforms

Qualcomm Technologies, Inc.



Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries. Qualcomm patents are licensed by Qualcomm Incorporated.

Tremendous innovation in compute ... but fragmented system software underlying it

01

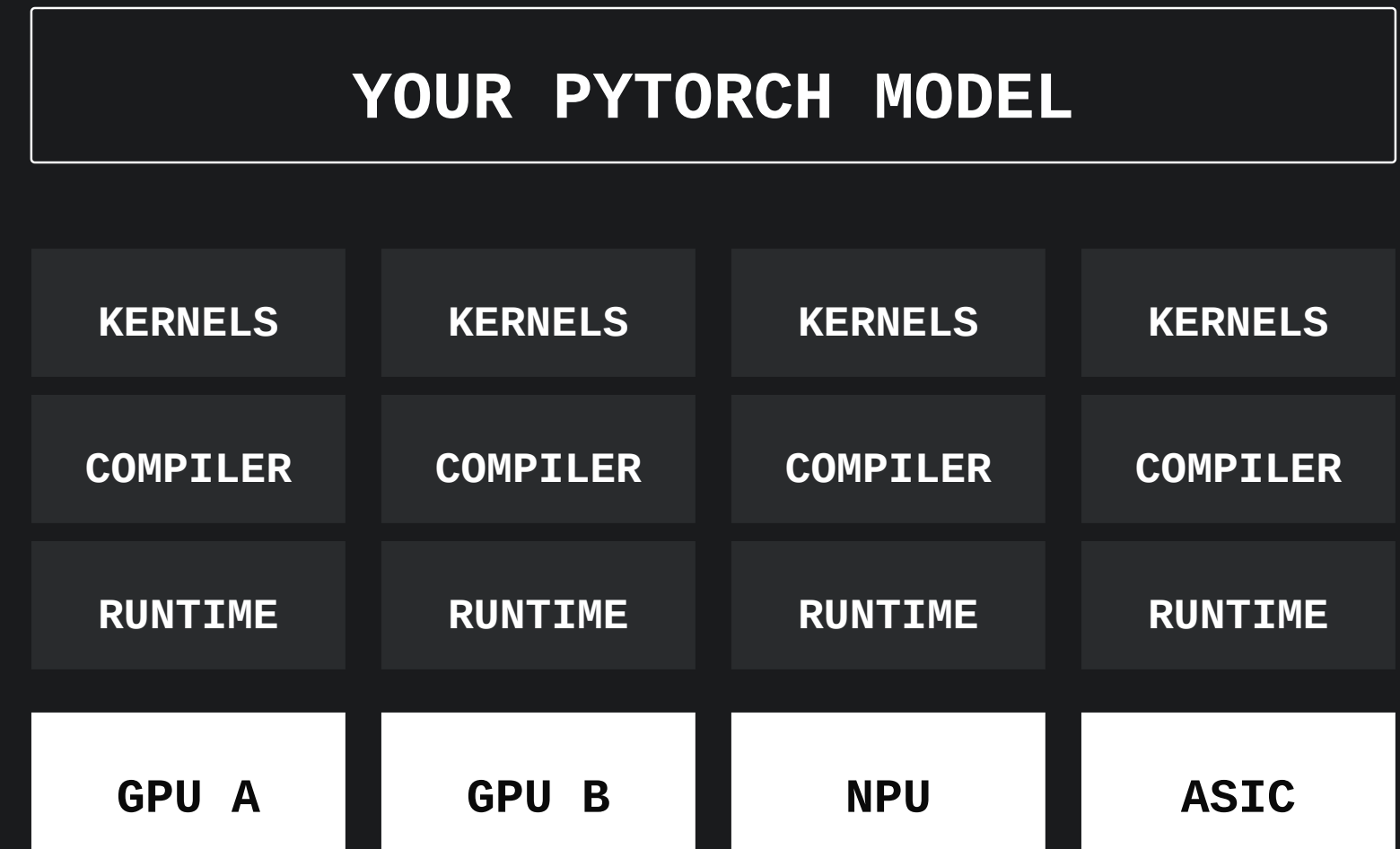
Every hardware team builds its own stack. GPUs, NPUs, and custom ASICs from more vendors than ever.

02

Kernels aren't portable. Every new backend means redoing performance work.

03

AI needs a shared foundation, as compilers did with LLVM.



Write code once, run it across all of that hardware.

OUT NOW

Mojo 1.x, open source under Apache 2.0

COMING SOON

Mojo on Windows, with Microsoft

NEXT WEEK AT LLVM DEVELOPERS' MEETING

MAX becomes source available

 PyTorch

The code you already have

```
import torch
```

MAX

Model authoring and a graph compiler that targets hardware from many vendors

Mojo

Pythonic code with the speed and low-level control of a systems language

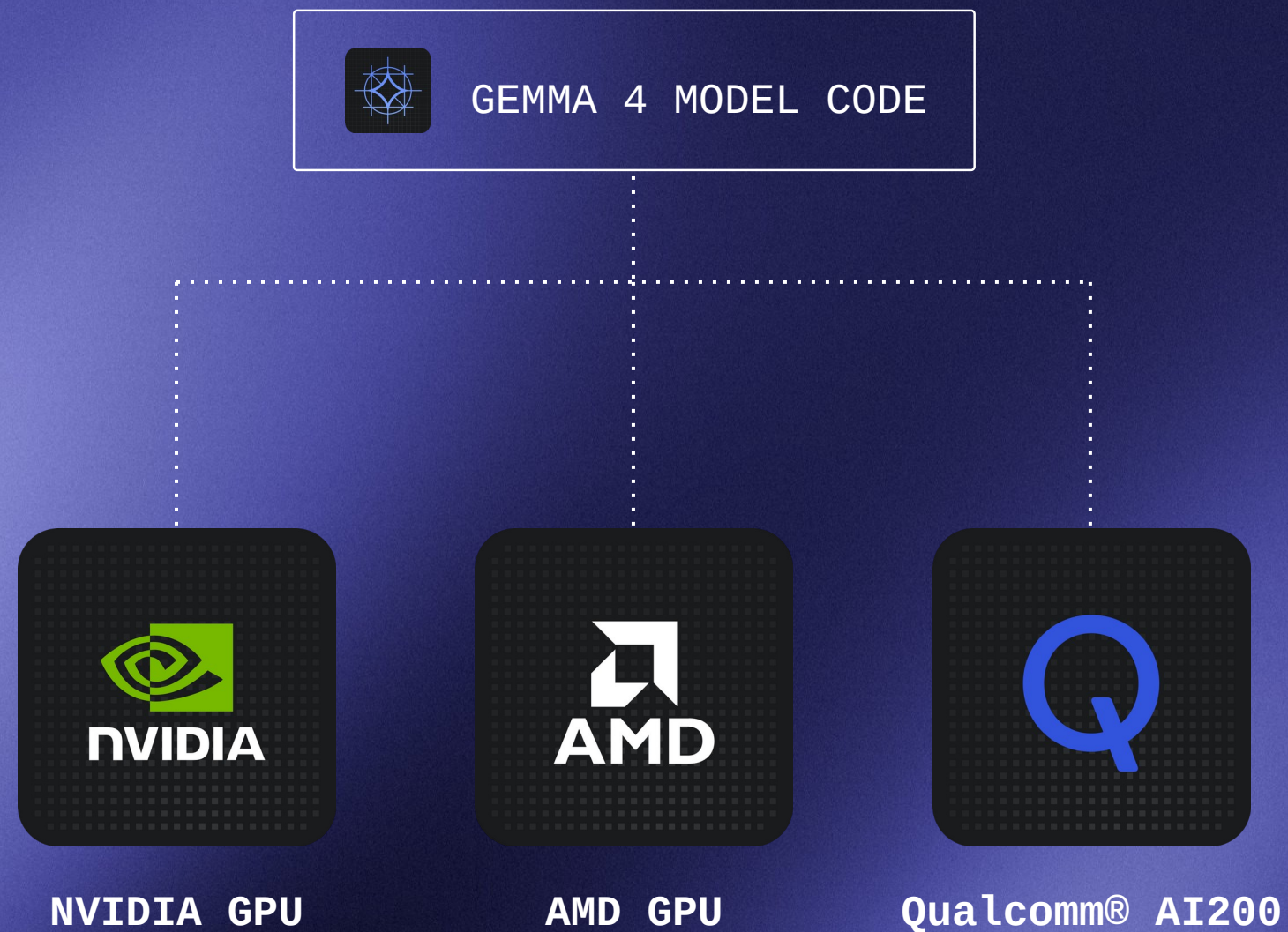
GPUS

NPUS

CPUS

ASICS

This runs in production today.
Same code, new silicon.



Trillions of tokens

Served every few days, across 6 architectures from 6 hardware vendors

Under a quarter

Gemma 4 31B on AI100, from first kernel to a live serving endpoint

Scaling PyTorch across edge AI

Qualcomm Technologies continues to advance PyTorch-native AI deployment through ExecuTorch enablement, helping developers bring optimized AI experiences to the edge

1 Across devices

Built with ExecuTorch and deploy across a broad range of devices powered by Qualcomm® technologies



IoT



XR



Mobile



Compute

2 Hardware enablement

Take advantage of purpose-built acceleration for increased performance and power efficiency.



CPU



GPU



NPU



LPAI

3 Operating systems

Preserve application portability across device ecosystem and deployment environments.



Windows



Android



Linux

Coming soon: open heterogeneous compute alliance

ONE NEUTRAL SOFTWARE LAYER · CLOUD TO EDGE

01

Silicon providers
contribute backends
and kernels

02

Deployers
bring scale across
mixed fleets

03

AI labs
share early
checkpoints

04

Academia and
research
advance the state of
the art

REACH OUT TO LEARN MORE

alliance@modular.com

Developers, start today

Run your PyTorch model across hardware with MAX,
without rewrites.



Get involved

- 01 Stop by the Qualcomm booth (#D6)
- 02 Stop by the Modular booth (#3)
- 03 "Ask the Experts" at Qualcomm booth

STEF LINDALL, PRINCIPAL ENGINEER
JIT Kernel Compilation: How Modular
writes fast kernels for any hardware

Oct 20 • 4:20 PM-4:45 PM • 210B (Concourse Level)



developer.qualcomm.com





Thank you

Nothing in these materials is an offer to sell any of the components or devices referenced herein.

© Qualcomm Technologies, Inc. and/or its affiliated companies. All Rights Reserved.

Qualcomm and Snapdragon are trademarks or registered trademarks of Qualcomm Incorporated. Other products and brand names may be trademarks or registered trademarks of their respective owners.

References in this presentation to “Qualcomm” may mean Qualcomm Incorporated, Qualcomm Technologies, Inc., and/or other subsidiaries or business units within the Qualcomm corporate structure, as applicable. Qualcomm Incorporated includes our licensing business, QTL, and the vast majority of our patent portfolio. Qualcomm Technologies, Inc., a subsidiary of Qualcomm Incorporated, operates, along with its subsidiaries, substantially all of our engineering, research and development functions, and substantially all of our products and services businesses, including our QCT semiconductor business.

Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries. Qualcomm patented technologies are licensed by Qualcomm Incorporated.

Follow us on: [in](#) [X](#) [@](#) [▶](#) [f](#)

For more information, visit us at qualcomm.com & qualcomm.com/blog