AI Hardware Systems for All

An Intro for Software Developers

Martin Raumann

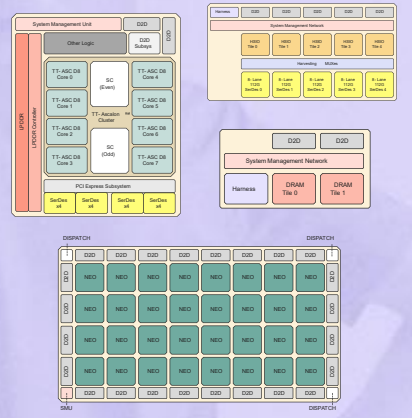
tenstorrent

Tenstorrent Product Summary

IP (TT-Ascalon™ / Tensix NEO)

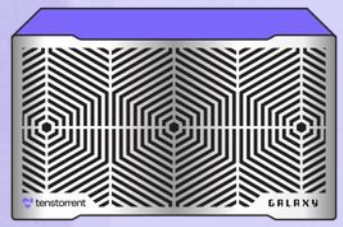
- Scales from mW to MW for efficiency and performance
- IP available for licensing
- Industry-leading performance
- Modular design available in varied configurations

Chips & Chiplets



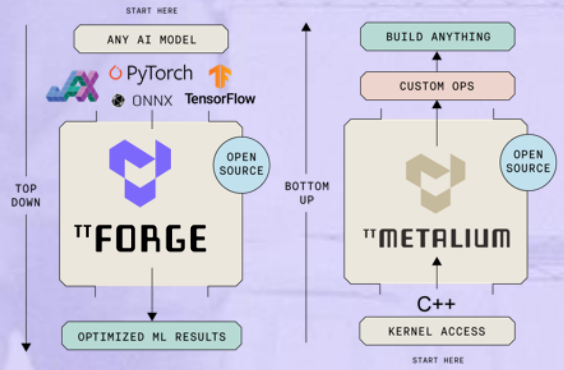
- Portfolio of products powered by scalable Tensix AI cores
- Inference and training, CNN and NLP, recommendation engines, all on the same silicon
- Hardware available for purchase as well as IP available for licensing
- Multi-component modular chiplets

Servers (Tenstorrent Galaxy™)



- Galaxy Server – 32 high performance ASICs in a custom chassis
- Easily combine servers into a Galaxy Rack with high bandwidth chip-to-chip connectivity

Software



- ML compilers that scale from one chip to thousands
- TT-Forge™ - MLIR-based compiler
- TT-Metalium™ - Bare metal software stack

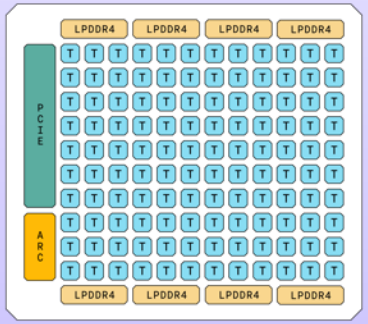
AI Silicon Roadmap

2022 → **2023** → **2025**

High Perf AI ASIC	Scalability	Heterogeny

Grayskull

AI Processor

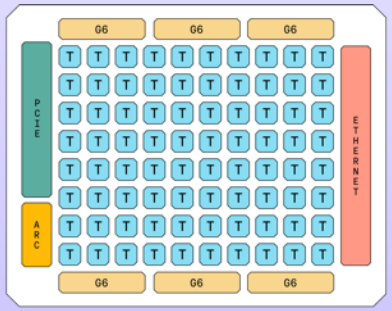


- 120 Tensix Cores
- 12nm
- 276 TOPS (FP8)
- 100 GB/s LPDDR4
- Gen4x16

GEN 1

Wormhole

Networked AI Processor

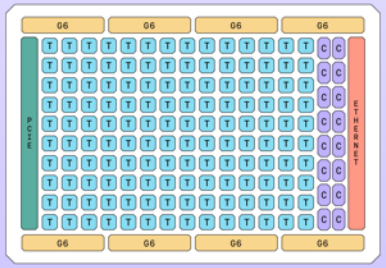


- 80 **Tensix+** Cores
- 12nm
- 328 TOPS (FP8)
- 336 GB/s GDDR6
- Gen4x16
- **16x100 Gbps Ethernet**

GEN 1

Blackhole

Standalone AI Computer

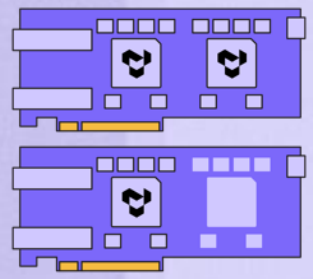


- 140 **Tensix++** Cores
- 6nm
- 745 TOPS (FP8)
- 512 GB/s GDDR6
- Gen5x16
- **10x400 Gbps Ethernet**
- **16 RISC-V CPU cores**

GEN 2

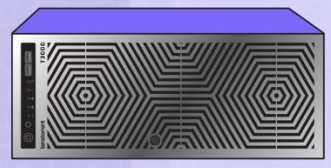
Wormhole Product Portfolio

PCIe Cards



- **n300d:** Two Wormhole™ ASICs operating at up to 300W, active axial fan cooler
- **n300s:** Two Wormhole™ ASICs operating at up to 300W, passive cooler
- **n150d:** One Wormhole™ ASIC operating at up to 160W, active axial fan cooler
- **n150s:** One Wormhole™ ASIC operating at up to 160W, passive cooler

TT-LoudBox

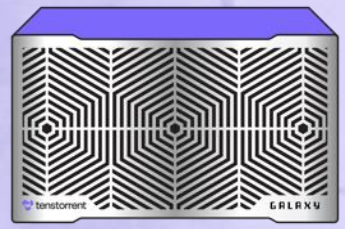


- Air-cooled 4U server for datacenter deployments
- Four n300s cards (8 Wormhole™ ASICs)
  - 512 Tensix Cores
  - 96GB GDDR6
  - 192MB SRAM

TT-QuietBox

- Liquid-cooled desktop workstation
- Four n300 cards (8 Wormhole™ ASICs)
  - 512 Tensix Cores
  - 96GB GDDR6
  - 192MB SRAM

Tenstorrent Galaxy™ Wormhole Server



- 6U UBB design for enterprise use
- 32 Wormhole™ ASICs for ultra-dense/high-performance data center deployment
- DGX level inference with higher efficiency and lower cost

Add-In Board Overview

2023 → **2024** → **2025**

Grayskull®

High Performance AI ASIC

e150

e75

- First generation Tensix Processor
- Up to 120 Tensix Cores
- PCIe Gen 4
- 8GB 256-bit LPDDR4

Wormhole™

Scalability

n300
Dual Chip

n150
Single Chip

- Tensix Cores updated: 50% more SRAM per Tensix Core™, improved BLOCKFP8 performance, expanded precision format support
- Moves to 12GB 192-bit GDDR6
- 100GbE connectivity
- Inter-card and inter-system expansion

Blackhole™

RISC-V & AI Generation

p300
Dual Chip

p150
Single Chip

p100
Single Chip
No Scale-Out

- Move to 6nm manufacturing
- Updates NoC on Tensix Cores and adds 16 x280 RISC-V cores
- Increases to 32GB 256-bit GDDR6 at faster speed
- Moves to PCIe Gen 5
- Upgrades to 400GbE connectivity

Product Availability

tenstorrent

Products Support Vision

TT-QuietBox

$15,000

The TT-QuietBox Liquid-Cooled Desktop Workstation is a great solution for developers running or testing AI models, or port and develop libraries for HPC. TT-QuietBox is equipped with four Tenstorrent Wormhole™ cards for a total of eight Wormhole™ Tensix Processors.

These processors are connected with a flexible, Ethernet-based mesh topology that can expand to achieve a 96GB memory pool. This empowers TT-QuietBox to run single user/single models up to approximately 80 billion parameters and single/multiple user, multiple models up to approximately 20 billion parameters.

TT-QuietBox is supported by two open-source SDKs for either high-level (TT-Buda™) or low-level (TT-Metalium™) development.

TECHNICAL SPECS +

OPERATING SYSTEM REQUIREMENTS +

− 1 + $15,000 | ADD TO CART

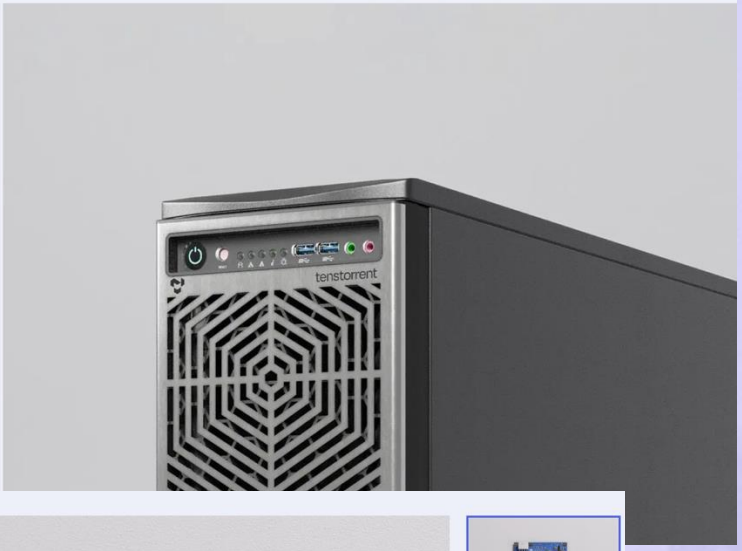
TT-LoudBox

$12,000

Box 4U/Desktop Workstation offers superior
per dollar for developers looking to run, test, and
odels or port and develop libraries for HPC.

h four Tenstorrent Wormhole™ n300s cards for a total of
ole™ Tensix Processors, this flexible, Ethernet-based
gy can expand to achieve a 96GB memory pool. This
oudBox to run single user/single models up to
y 80 billion parameters and single/multiple user,
dels up to approximately 20 billion parameters. TT-
upported by two open-source SDKs for either high-level
r low-level (TT-Metalium™) development.

ECS +



Wormhole™ n150d

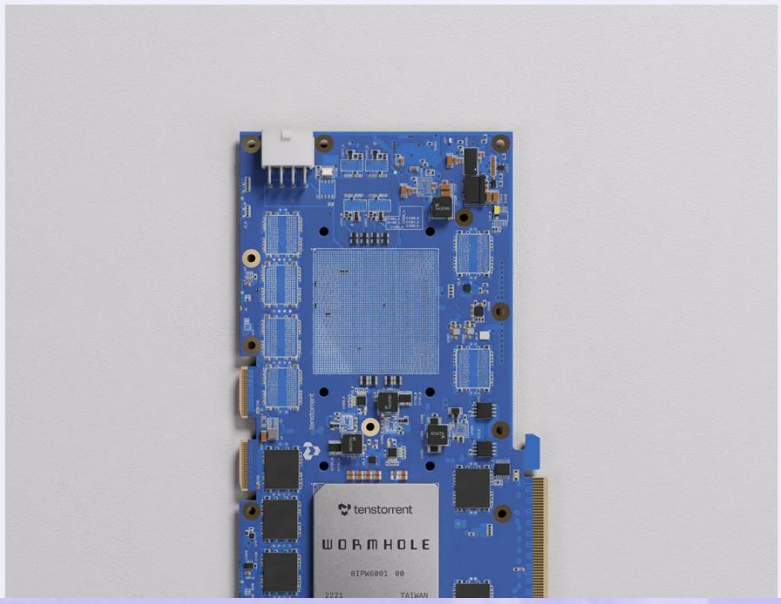
$1,099

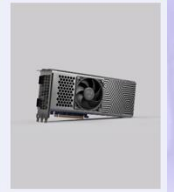
Wormhole™ n150d features Tenstorrent's flexible, scalable Wormhole™ Tensix Processor operating at up to 160W, offering superior performance for cost compared to traditional GPUs and broad data precision format support. The processor can network into a multichip mesh for workstations and servers (such as Galaxy) and is supported by two open-source SDKs for either high-level (TT-Buda™) or low-level (TT-Metalium™) development. n150d includes a 2.5-slot active cooling solution.

SOFTWARE CAPABILITIES +

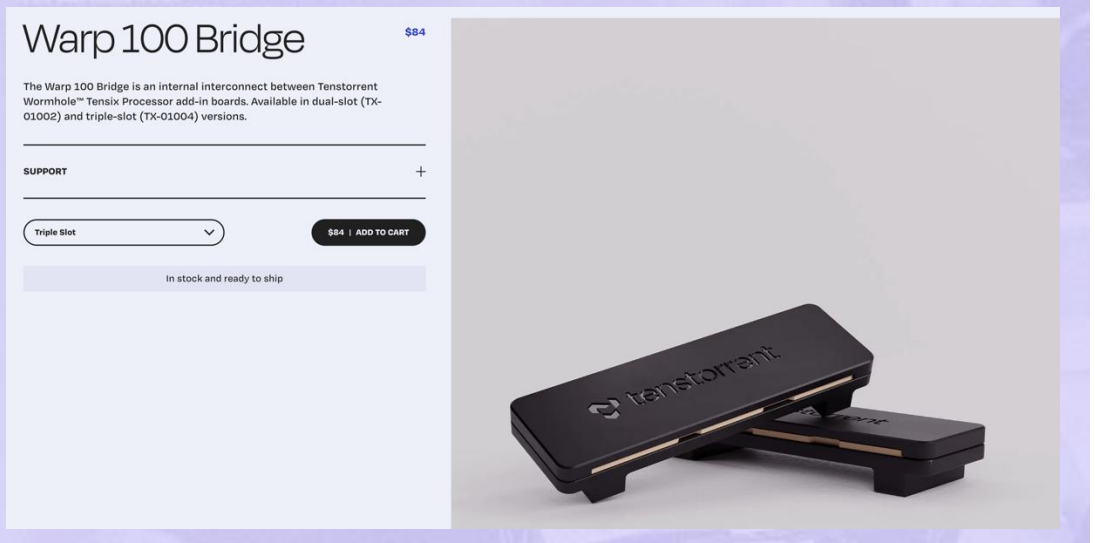
SYSTEM REQUIREMENTS +

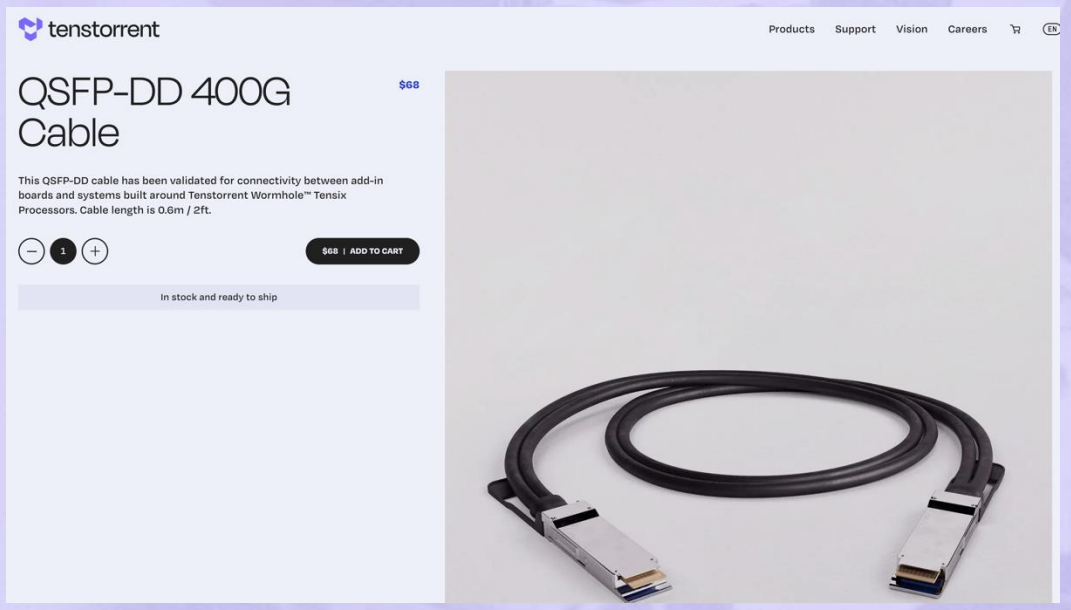
n150d ⌄ $1,099 | ADD TO CART

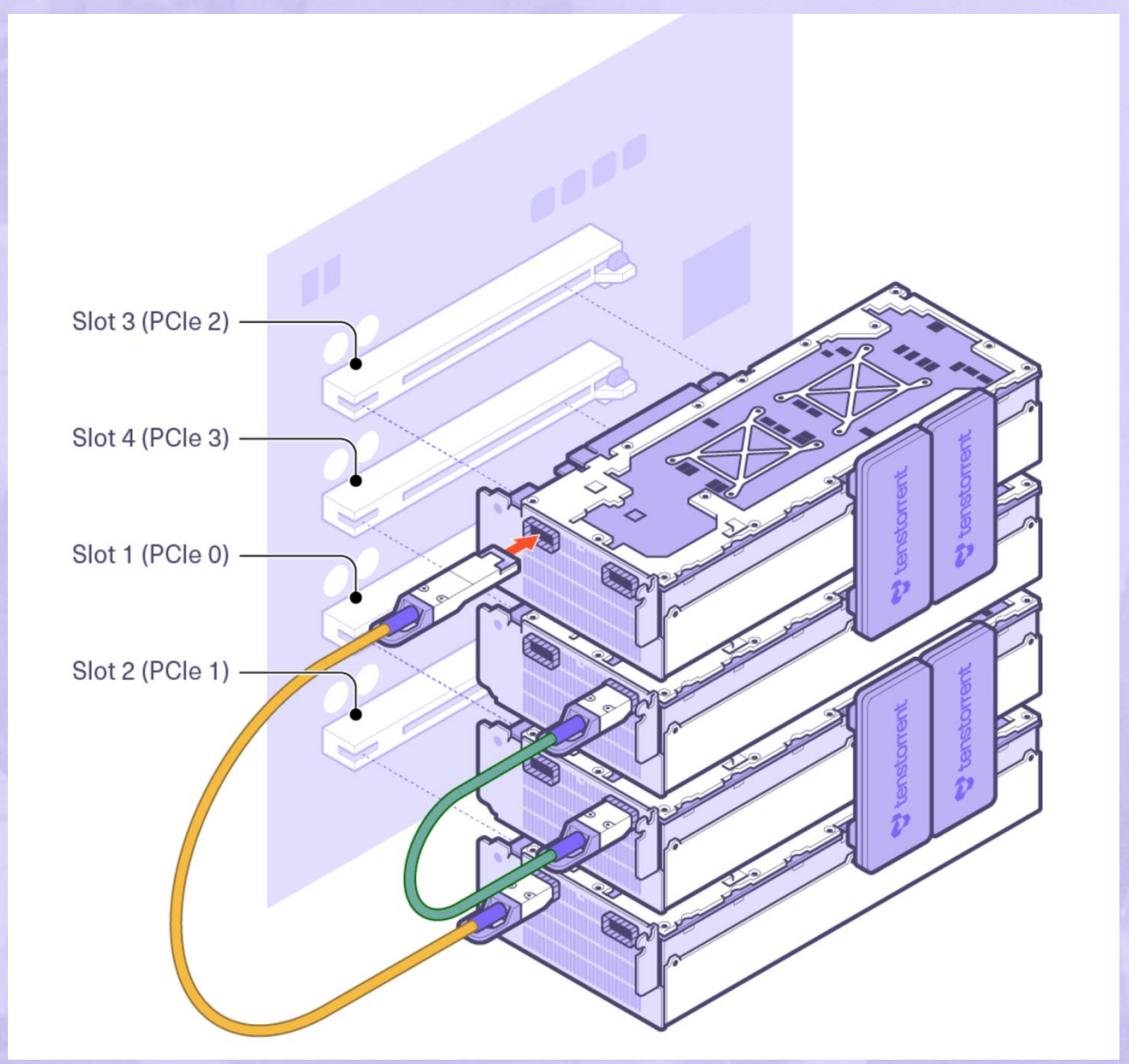




Creating Topologies





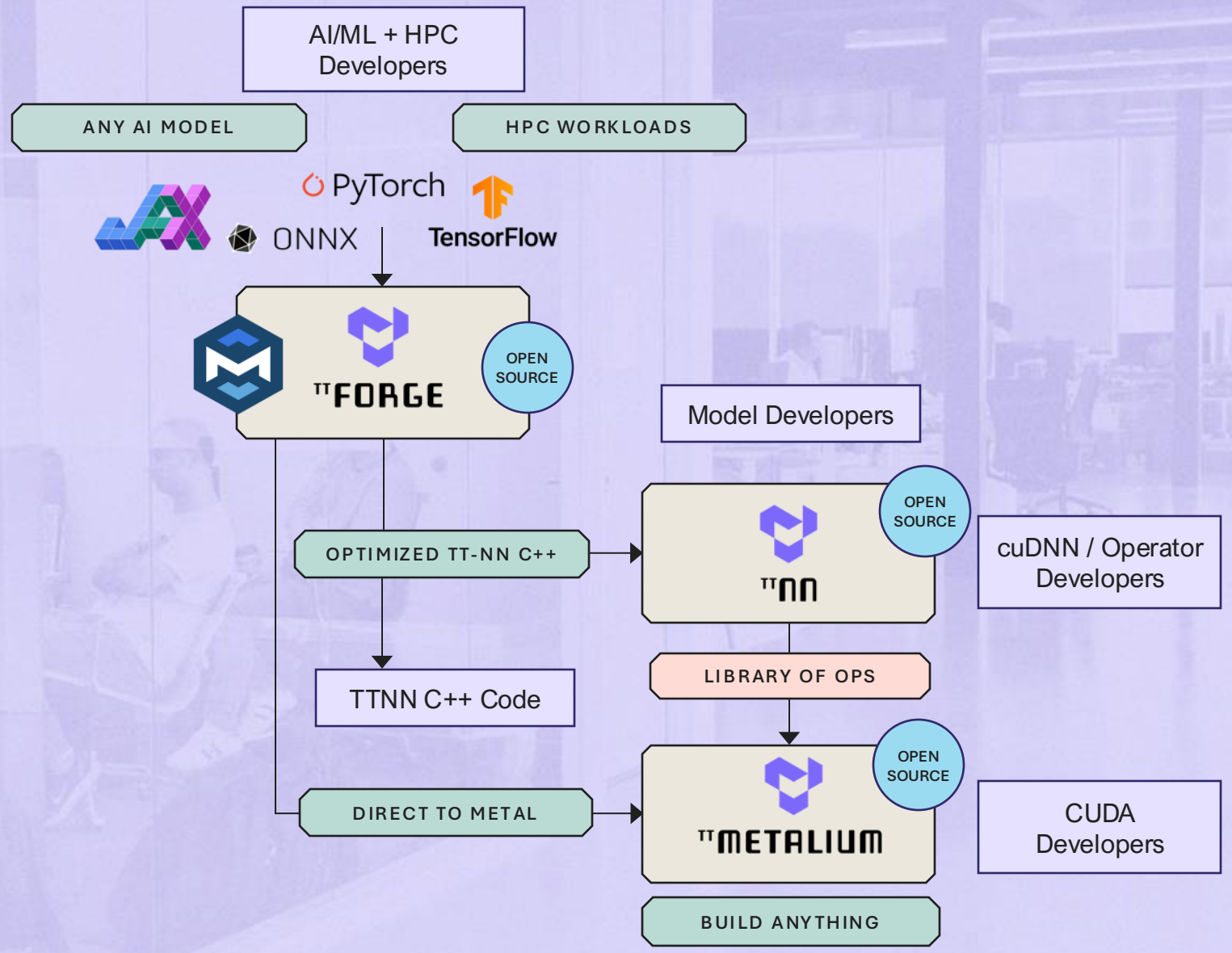


Tenstorrent Open Source Software

- **TT-Forge** – MLIR-based compiler integrated into various frameworks; AI/ML models from domain-specific compilers to custom kernel generation

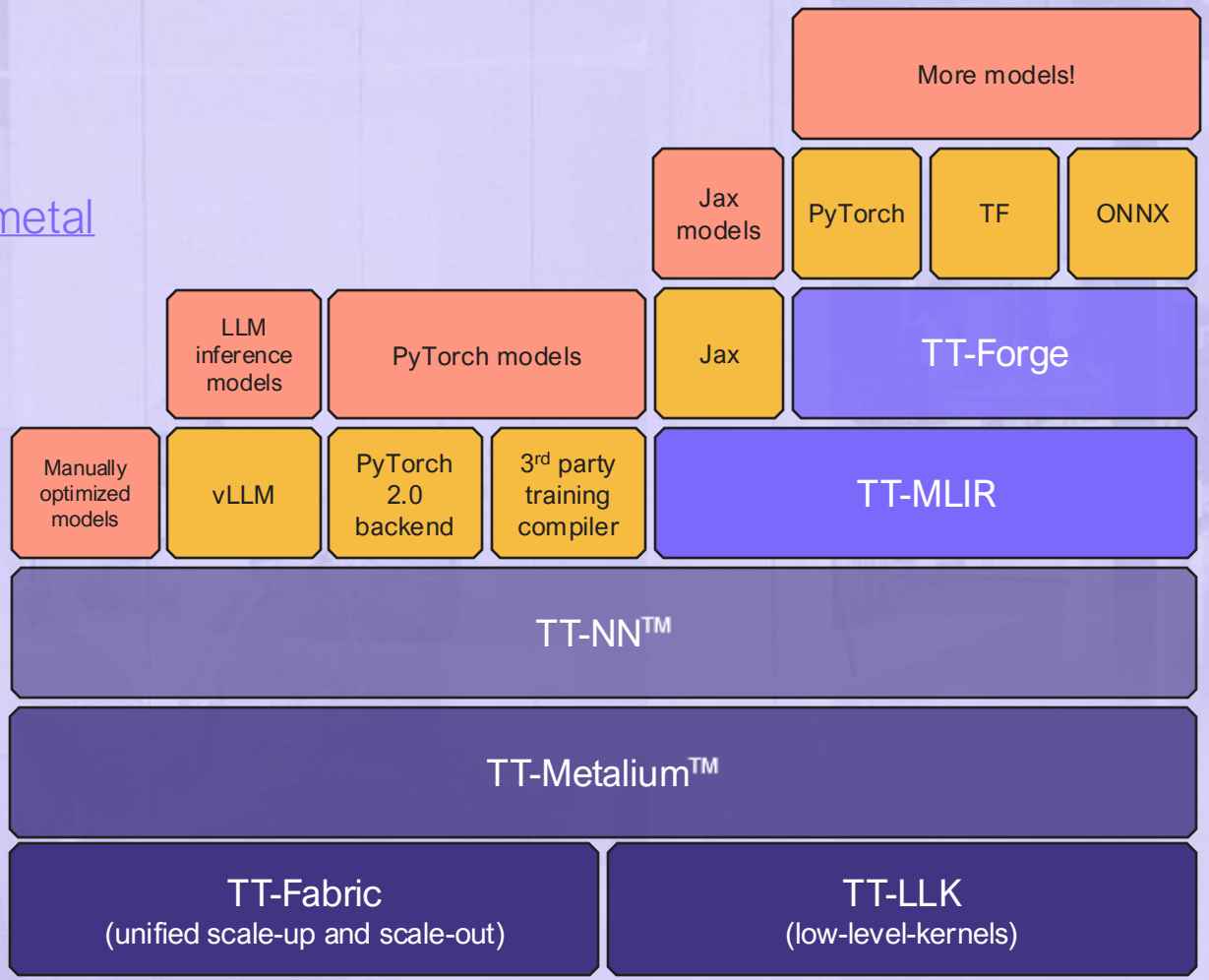
- **TT-NN™** – Library of optimized operators
  - ATen coverage
  - PyTorch-like API

- **TT-Metalium™** – Low-level programming model and entry point
  - Build your own kernels
  - User-facing host API

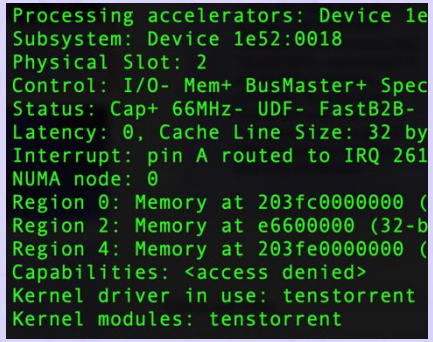


Software Ecosystem and Integrations

General: https://github.com/tenstorrent
TT-Metalium™: https://github.com/tenstorrent/tt-metal
TT-MLIR: https://github.com/tenstorrent/tt-mlir

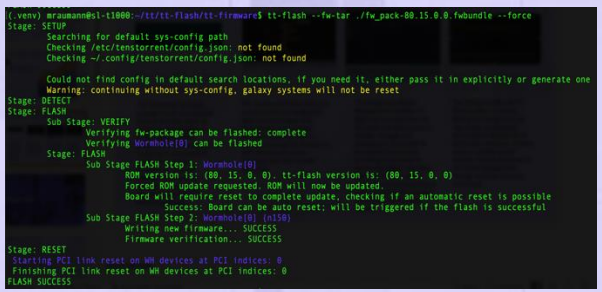


Simple, practical and intuitive tooling and debug utilities



TT-KMD, TT-UMD

The Tenstorrent AI drivers consists of both Kernel-Mode (tt-kmd) and User-Mode (tt-umd) that work together to enable communication between Linux systems and Tenstorrent's specialized AI accelerator hardware, providing the essential interfaces and memory management required for efficient AI and machine learning workloads.

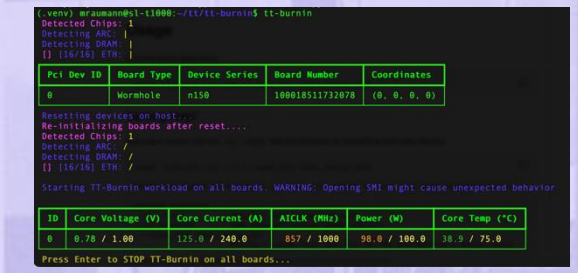


TT-Flash

The Tenstorrent tt-flash utility is a command-line tool designed to flash firmware images onto Tenstorrent's AI accelerator hardware, enabling users to update or replace the firmware on their devices with proper packages from the firmware repository.

TT-SMI

The Tenstorrent System Management Interface (TT-SMI) is a command-line and GUI utility that allows users to monitor, manage, and troubleshoot Tenstorrent AI accelerator hardware by providing comprehensive device information, telemetry data, and firmware details, while also enabling board-level reset capabilities for Grayskull, Wormhole, and Blackhole devices.



TT-Burnin

The Tenstorrent Burnin (TT-Burnin) utility is a stress-testing tool designed to run high power consumption workloads on Tenstorrent AI accelerator hardware, allowing users to validate system stability, thermal performance, and power management through real-time telemetry monitoring.


Thank You!