

# AI Hardware Systems for All

## An Intro for Software Developers

Martin Raumann





# AI Silicon Roadmap

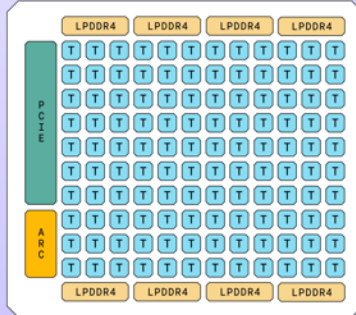
2022



High Perf AI ASIC

## Grayskull

AI Processor



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

GEN 1

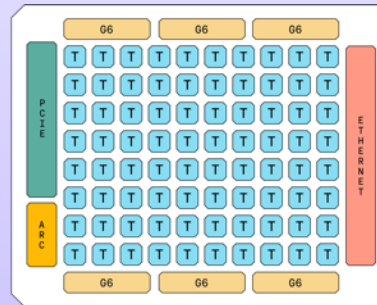
2023



Scalability

## Wormhole

Networked AI Processor



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

GEN 1

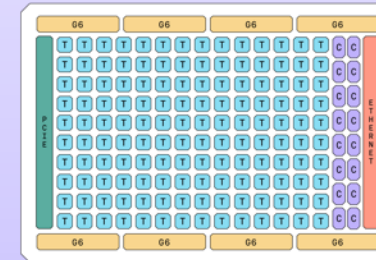
2025



Heterogeny

## Blackhole

Standalone AI Computer

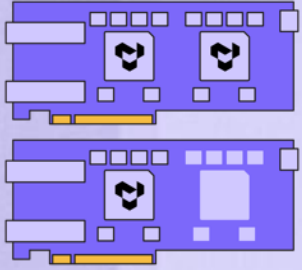


- 140 Tensix++ Cores
- 6nm
- 745 TOPS (FP8)
- 512 GB/s GDDR6
- Gen5x16
- 10x400 Gbps Ethernet
- 16 RISC-V CPU cores

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



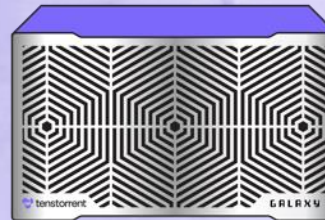
- Air-cooled 4U server for datacenter deployments
- Four n300s cards (8 Wormhole™ ASICs)
  - 512 Tensix Cores
  - 96GB GDDR6
  - 192MB SRAM

## 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

## Grayskull®

High Performance AI ASIC

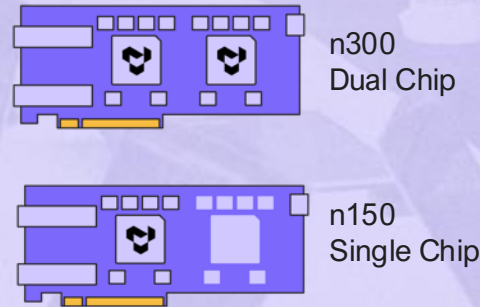


- First generation Tensix Processor
- Up to 120 Tensix Cores
- PCIe Gen 4
- 8GB 256-bit LPDDR4

2024

## Wormhole™

Scalability

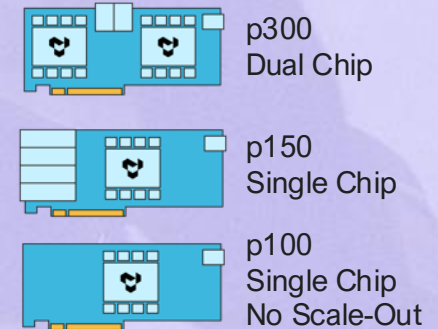


- Tensix Cores updated: 50% more SRAM per Tensix Core™, improved BLOCKFP8 performance, expanded precision format support
- Moves to 12GB 192-bit GDDR6
- 100GbE connectivity
- Inter-card and inter-system expansion

2025

## Blackhole™

RISC-V & AI Generation



- 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

## 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



\$15,000 | ADD TO CART



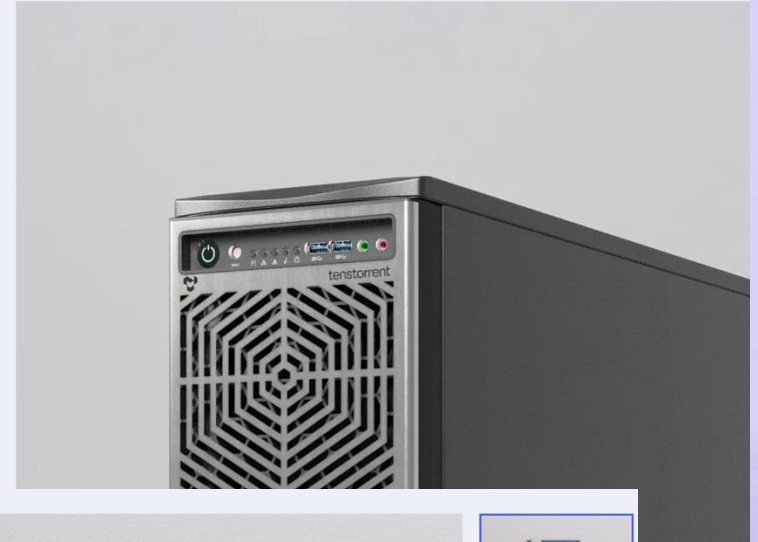
## TT CloudBox

\$12,000

TT CloudBox 4U/Desktop Workstation offers superior performance per dollar for developers looking to run, test, and develop models or port and develop libraries for HPC.

With four Tenstorrent Wormhole™ n300s cards for a total of eight Tensix Processors, this flexible, Ethernet-based workstation can expand to achieve a 96GB memory pool. This empowers TT CloudBox 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 CloudBox is supported by two open-source SDKs for either high-level (TT-Buda™) or 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



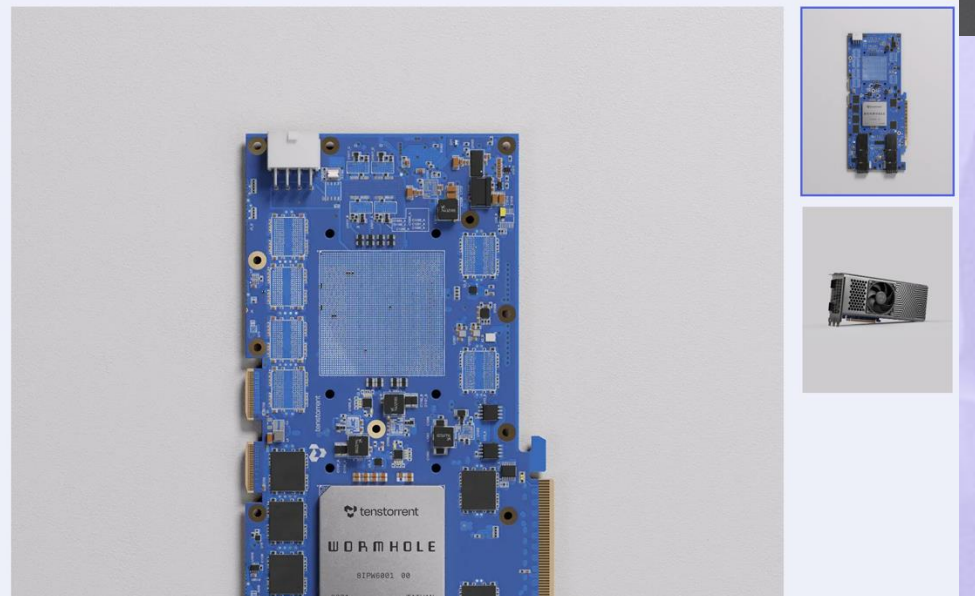
### SYSTEM REQUIREMENTS



n150d



\$1,099 | ADD TO CART



# Creating Topologies

## Warp 100 Bridge

\$84

The Warp 100 Bridge is an internal interconnect between Tenstorrent Wormhole™ Tensix Processor add-in boards. Available in dual-slot (TX-01002) and triple-slot (TX-01004) versions.

SUPPORT +

Triple Slot  \$84

In stock and ready to ship



tenstorrent

Products Support Vision Careers

## QSFP-DD 400G Cable

\$68

This QSFP-DD cable has been validated for connectivity between add-in boards and systems built around Tenstorrent Wormhole™ Tensix Processors. Cable length is 0.6m / 2ft.

\$68

In stock and ready to ship

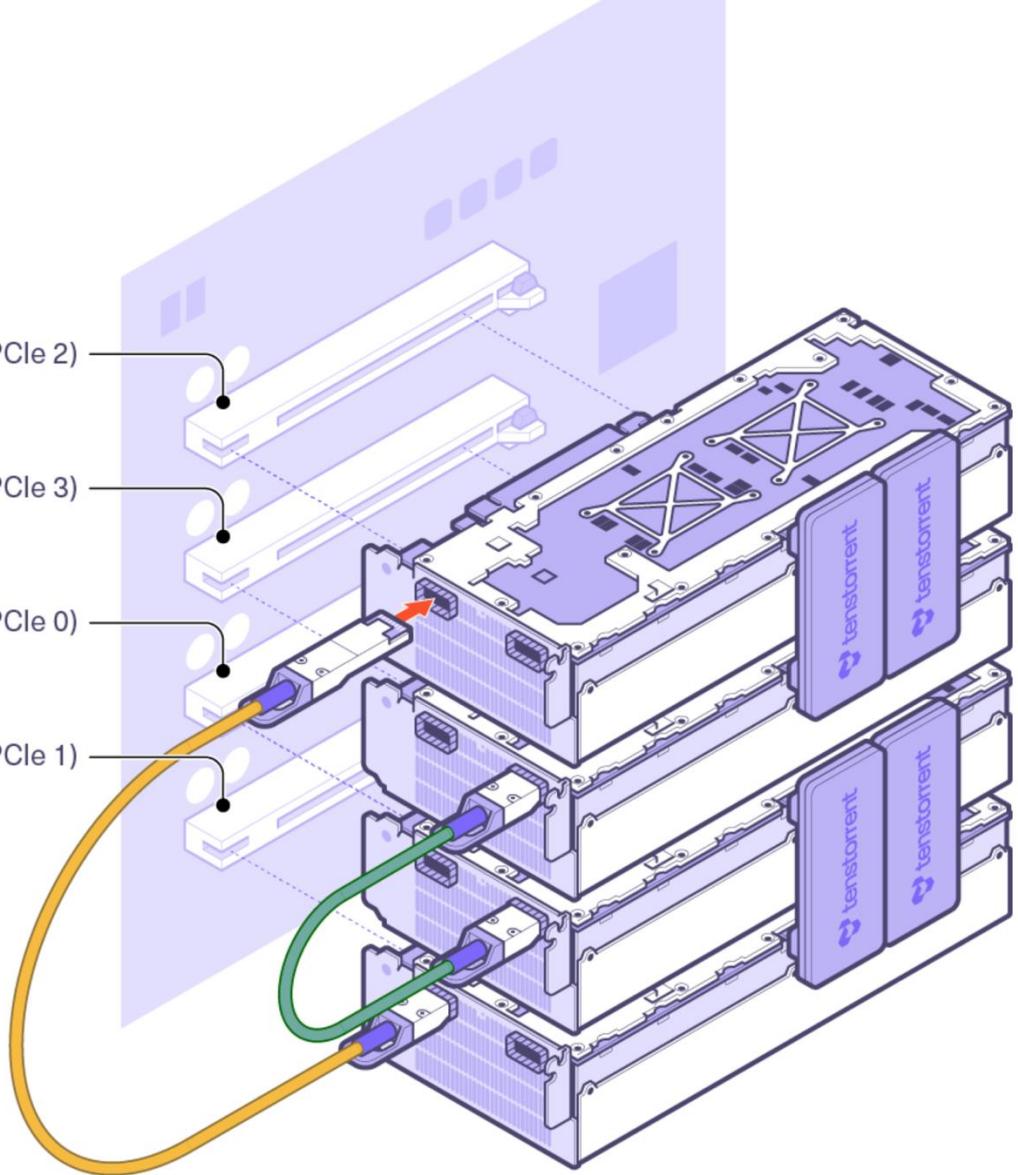


Slot 3 (PCIe 2)

Slot 4 (PCIe 3)

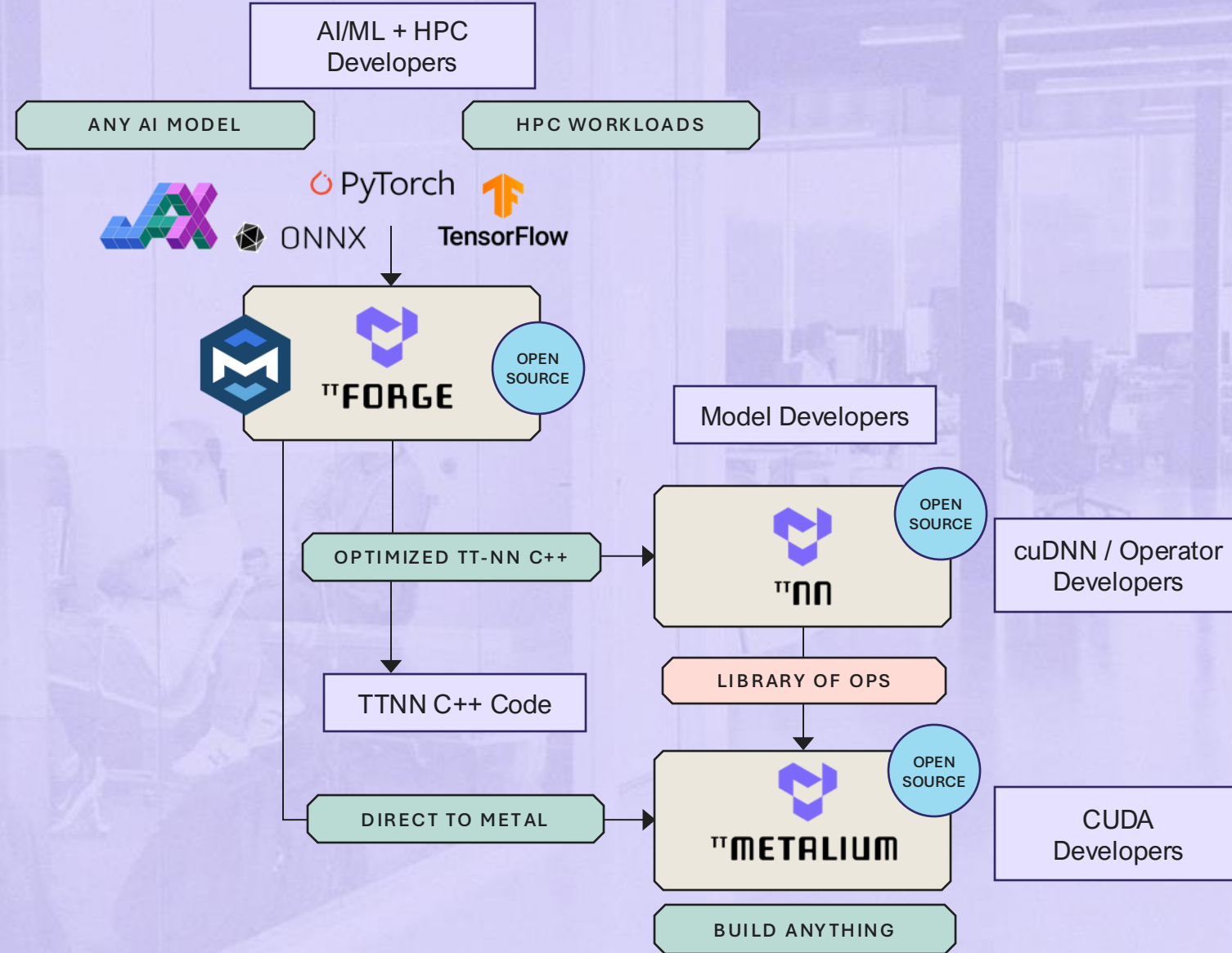
Slot 1 (PCIe 0)

Slot 2 (PCIe 1)



# 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

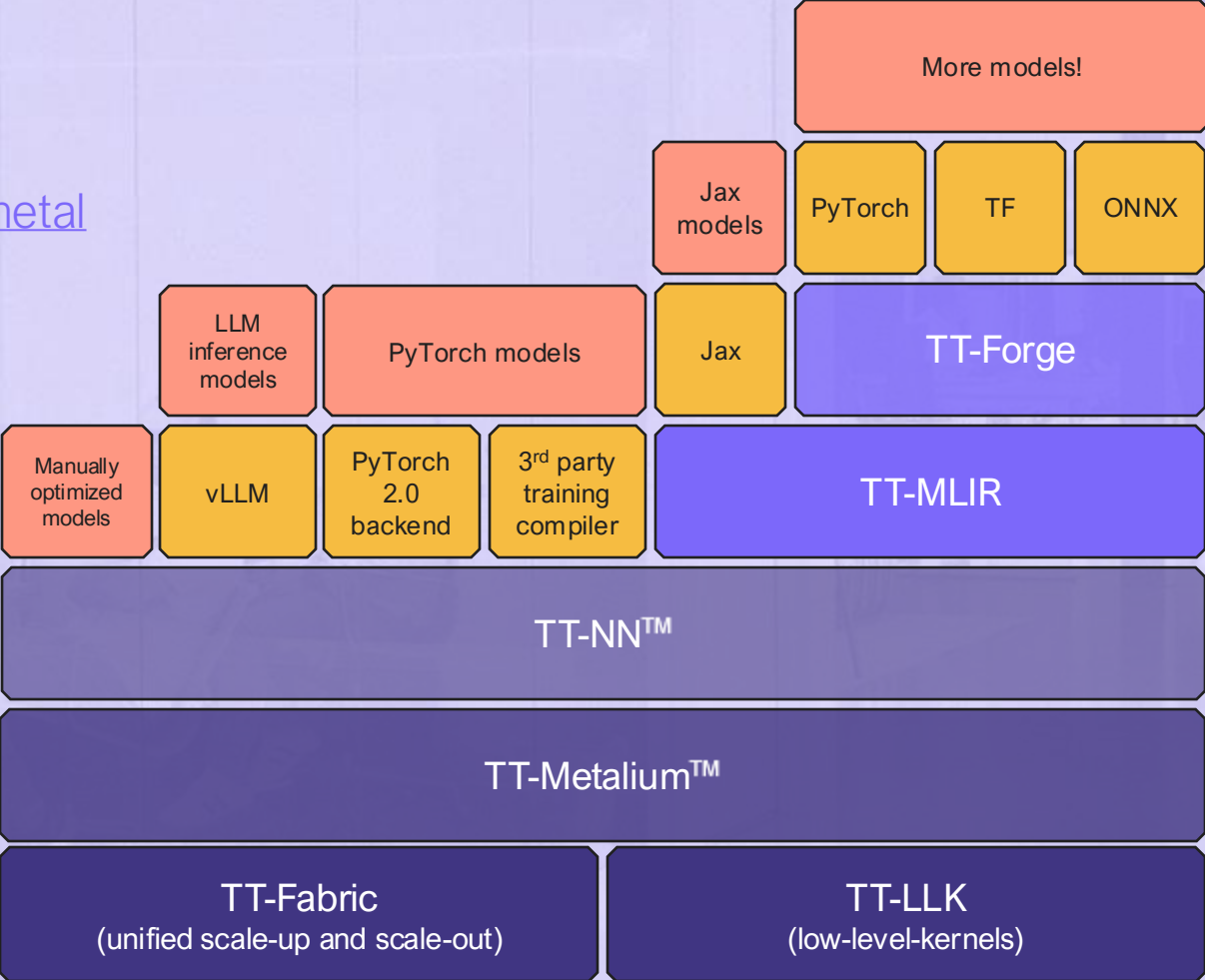




# Software Ecosystem and Integrations



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

```
Processing accelerators: Device 1e
Subsystem: Device 1e52:0018
Physical Slot: 2
Control: I/O- Mem+ BusMaster+ Spec
Status: Cap+ 66MHz- UDF- FastB2B-
Latency: 0, Cache Line Size: 32 by
Interrupt: pin A routed to IRQ 261
NUMA node: 0
Region 0: Memory at 203fc0000000 (
Region 2: Memory at e6600000 (32-b
Region 4: Memory at 203fe0000000 (
Capabilities: <access denied>
Kernel driver in use: tenstorrent
Kernel modules: tenstorrent
```

```
(.venv) mraumann@sl-L1000:~/t1/t1-flash/t1-firmware$ tt-flash --fw-pack-08.15.0.0.fwbundle --force
Stage: SETUP
Searching for default sys-config path
Checking /etc/tenstorrent/config.json: not found
Checking ~/config/tenstorrent/config.json: not found
Could not find config in default search locations. If you need it, either pass it in explicitly or generate one
Warning: continuing without sys-config, galaxy systems will not be reset.
Stage: DETECT
Stage: FLASH
Sub Stage: VERIFY
Verifying fw-package can be flashed: complete
Verifying wormhole@0 can be flashed
Stage: FLASH
Sub Stage FLASH Step 1: Wormhole@0
ROM version is: (00, 15, 0, 0), tt-flash version is: (00, 15, 0, 0)
Forced ROM update requested, ROM will now be updated.
Board will require reset to complete update, checking if an automatic reset is possible
Success: Board can be auto reset; will be triggered if the flash is successful
Sub Stage FLASH Step 2: Wormhole@1 (1518)
Writing new firmware... SUCCESS
Firmware verification... SUCCESS
Stage: RESET
Starting PCI link reset on WH devices at PCI indices: 0
Finishing PCI link reset on WH devices at PCI indices: 0
FLASH SUCCESS
```

Version 3.8.0 TT-SMI Jan 30 2025 01:53:08 PM

Host Info (Config Manager)

- OS: Linux 5.15.0-91
- Uptime: 10h:04:16.175
- Kernel: 5.15.0-91-generic
- Platform: Y12-140-04
- Memory: 1.52 TB
- Power: 2320W
- Driver: TT-008 3.27.1

Device Information

#	Box ID	Board Type	Board ID	Coords	DRAM-Trained	DRAM-Speed	Link-Speed	Link-Width
0	00000100-0	475	1000011122010	0, 0, 0, 0	Y	2000	Gen4 / Gen5	x8 / x16
1	00000100-0	475	1000011122010	1, 0, 0, 0	Y	120	Gen4 / Gen5	x16 / x16
2	00000100-0	475	1000011122010	1, 0, 0, 0	Y	120	Gen4 / Gen5	x16 / x16

Latest SW Versions

- tt-flash: 3.2.1
- tt-umd: 3.2.1
- firmware-bundle: 08.15.0.0
- tt-008: 3.26.0
- tt-008-1: 3.21.0
- tt-008-2: 3.21.0

```
(.venv) mraumann@sl-L1000:~/t1/t1-burnin$ tt-burnin
Detected Chips: 1
Detecting AICLK: /
Detecting DRAM: /
[] [16/16] ETH: /

Starting TT-Burnin workload on all boards. WARNING: Opening SMI might cause unexpected behavior

ID Core Voltage (V) Core Current (A) AICLK (MHz) Power (W) Core Temp (°C)
0 0.78 / 1.00 125.0 / 249.0 857 / 1000 98.0 / 100.0 38.0 / 75.0

Press Enter to STOP TT-Burnin on all boards...
```

## 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!

