Introduction to Tenstorrent

tenstorrent

Tenstorrent at a Glance

Founded
March 2016

Capital Raised
$1B+

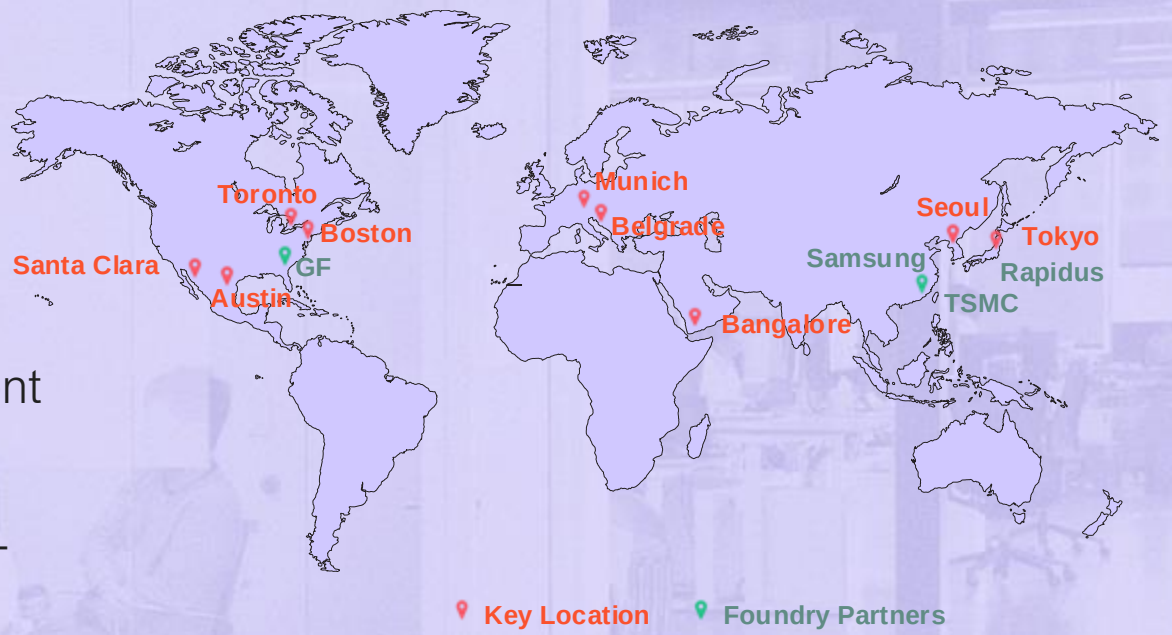
Investors

Fidelity INVESTMENTS
ECLIPSE
HYUNDAI
EPIQ
SAMSUNG
MESH

Target Markets

Data Center
Client
Automotive
IoT

Global Footprint



Key Performance Indicators

~630	Employees	~$150M		$40M+	2024E Revenue
300+	R&D Professionals		IP Bookings In last 8 months*	~$200M	2025E Revenue

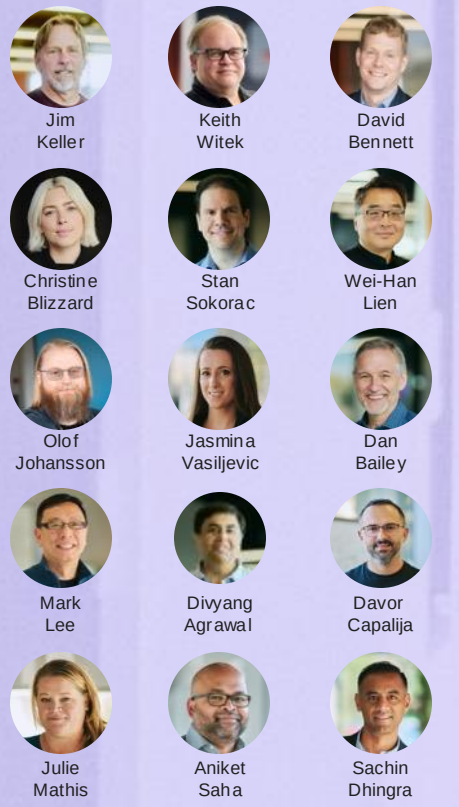
Business Model

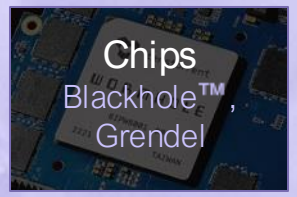
Tenstorrent monetizes its AI and RISC-V Technology in 3 ways:

1. Licenses AI and RISC-V IP to customers building silicon
2. Sells AI chiplets, RISC-V chiplets, & AI chips
3. Sells AI boards, servers, and systems

Software, Silicon, and Systems to Run AI, ML, and Compute Cheaper and Faster than Anyone Else

World Class Team → Leading Technology → Innovative Products → Large Target Markets → Blue-Chip Partner Ecosystem





Data Center
$41bn TAM

Automotive
$15bn TAM

Client & IoT
$70bn TAM

TAM figures represent the silicon TAM.

Chiplet Ecosystem

Customers: ALPHAWAVE SEMI, HYUNDAI MOTOR GROUP, SYNOPSYS, OLA

ODM Partners: Supermicro, Quanta, baya Systems, BLUE CHEETAH ANALOG DESIGN

Strategic Alliances: SAMSUNG, METI, LG, ThunderSoft, HYUNDAI MOTOR GROUP, Codasip

AI Software: MOREH, multicoreware

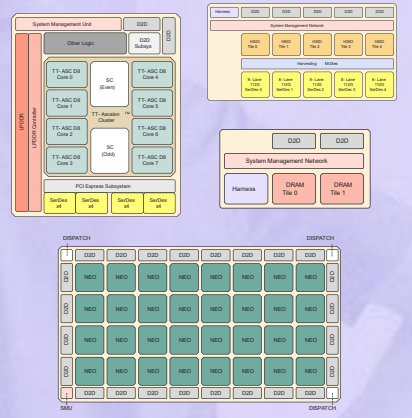
IP: Encore Semi, iar, WNDRVR, ARTERIS IP, NOSCHIP, KRAKATOA TECHNOLOGIES, ignitarium, Imagination, imperas, nimble, ETAS, SEMIFIVE, Green Hills Software, TESSOLVE, BOSCH, BLOOMING RICH, RESILTECH, cast, IRON MOUNTAIN, Rambus

Tenstorrent Product Summary

IP (TT-Ascalon™ / Tensix NEO)

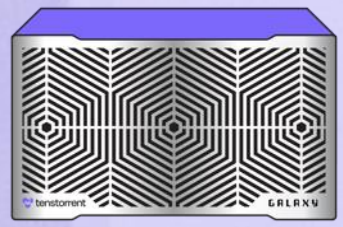
- Scales from mW to MW for efficiency and performance
- IP available for licensing
- Industry-leading performance
- Modular design available in varied configurations

Chips & Chiplets



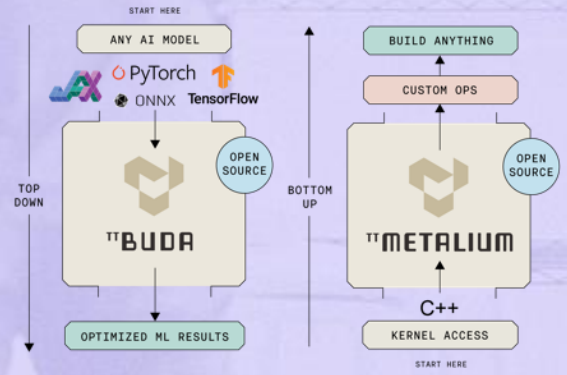
- Portfolio of products powered by scalable Tensix AI cores
- Inference and training, CNN and NLP, recommendation engines, all on the same silicon
- Hardware available for purchase as well as IP available for licensing
- Multi-component modular chiplets

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-Buda™ - Automated AI/ML Compiler
- TT-Metalium™ - Bare metal software stack

Core Silicon Roadmap

2021 Tapeout
2023 Product

High Perf AI ASIC

2022 Tapeout
2024 Product

Scalability

2023 Tapeout
2025 Product

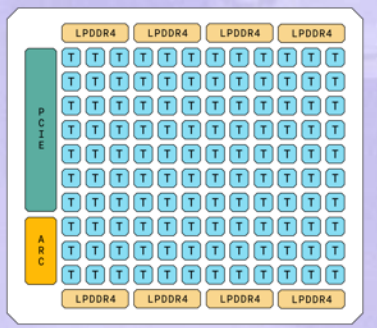
Heterogeny

2025 Tapeout
2026 Product

Chiplets

Grayskull®

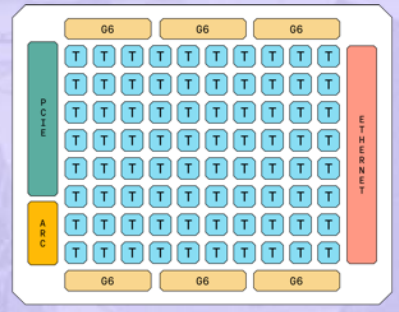
AI Processor



- 120 Tensix Cores
- 12nm
- 332 TFLOPS (FP8)
- 83 TFLOPS (BLOCKFP8)
- 16 lanes of PCIe Gen 4
- 8 channels LPDDR4

Wormhole™

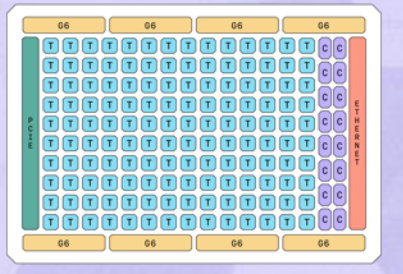
Networked AI Processor



- 80 Tensix+ Cores
- 12nm
- 292 TFLOPS (FP8)
- 164 TFLOPS (BLOCKFP8)
- 16 lanes of PCIe Gen 4
- 16x100 Gbps Ethernet
- 6 channels GDDR6

Blackhole™

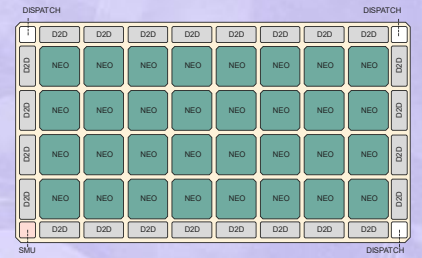
Standalone AI Computer



- 140 Tensix++ Cores
- 6nm
- 774 TFLOPS (FP8)
- 387 FLOPS (BLOCKFP8)
- 12x400 Gbps Ethernet
- 48 lanes of SerDes
- 8 channels of GDDR6
- 16 RISC-V CPU cores

Quasar

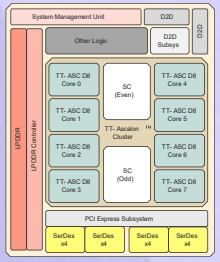
Low Power AI Chiplet



- 32 Tensix NEO Cores
- 4nm Chiplet
- Features incl. SMC with self-boot/Reset
- Non-blocking D2D interfaces
- Easily stack Quasar or combine to choose your own compute

Athena

High Performance RISC-V CPU Chiplet



- 4nm Chiplet
- Feature support incl. SMC, IOMMU, AIA
- Non-blocking D2D Interfaces
- Composable IO, MEM, CPU compute
- Details TBD

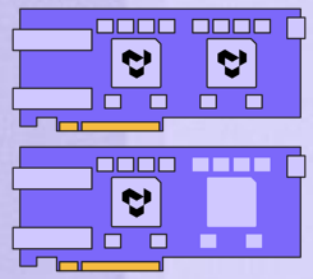
GEN 1

GEN 2

GEN 3

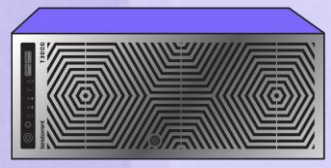
Wormhole Product Portfolio

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

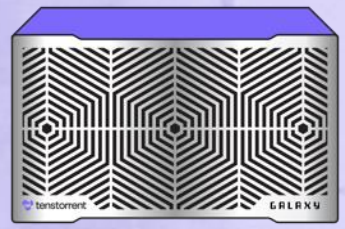


- Air-cooled 4U server for datacenter deployments
- Four n300s cards (8 Wormhole™ ASICs)
  - 512 Tensix Cores
  - 96GB GDDR6
  - 192MB SRAM

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

Tenstorrent Galaxy Roadmap

2023

Wormhole™

Prototype

- 32 Wormhole™ cards in a highly optimized, highly dense, custom 4U chassis
- 5.2 PetaFLOPS at BLOCKFP8
- 384GB of globally accessible GDDR6 memory
- 3.8GB SRAM
- Expansion beyond a single server accomplished via standard network cabling

2024

Wormhole™

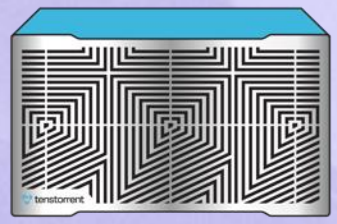
ODM Redesign

- 32 Wormhole™ cards in a highly optimized, highly dense, custom 6U chassis
- Internal head node for built-in host capability
- Increased PetaFLOPS and GDDR6 memory bandwidth
- UBB module for flexible customer infrastructure
- Expansion beyond a single server accomplished via standard network cabling

2025

Blackhole™

RISC-V & AI Generation



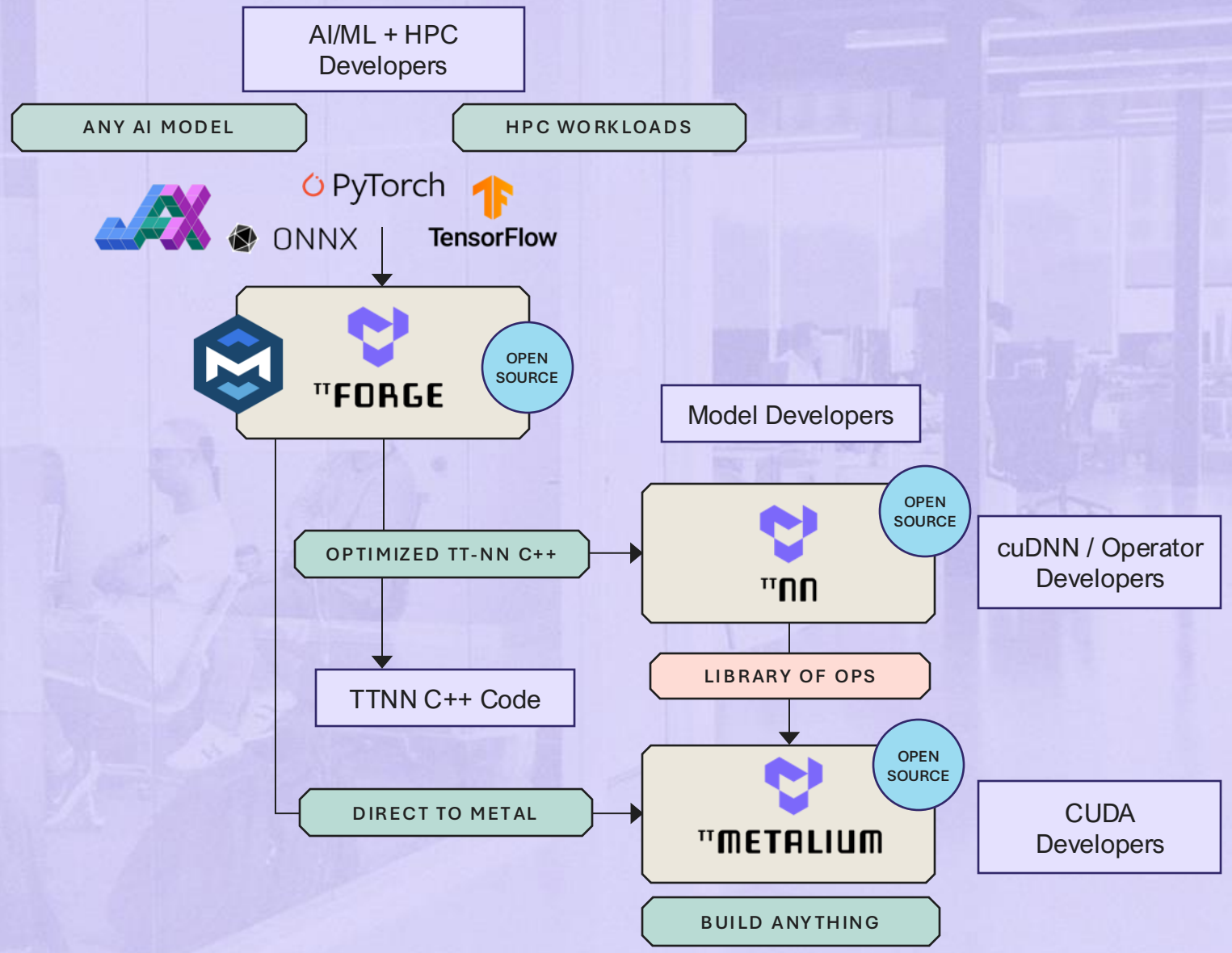
- Utilize 6U Wormhole UBB design for easy customer transitions
- Mesh connections and topologies still being determined
- Training vs. inference focus being considered

Tenstorrent Open Source Software

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

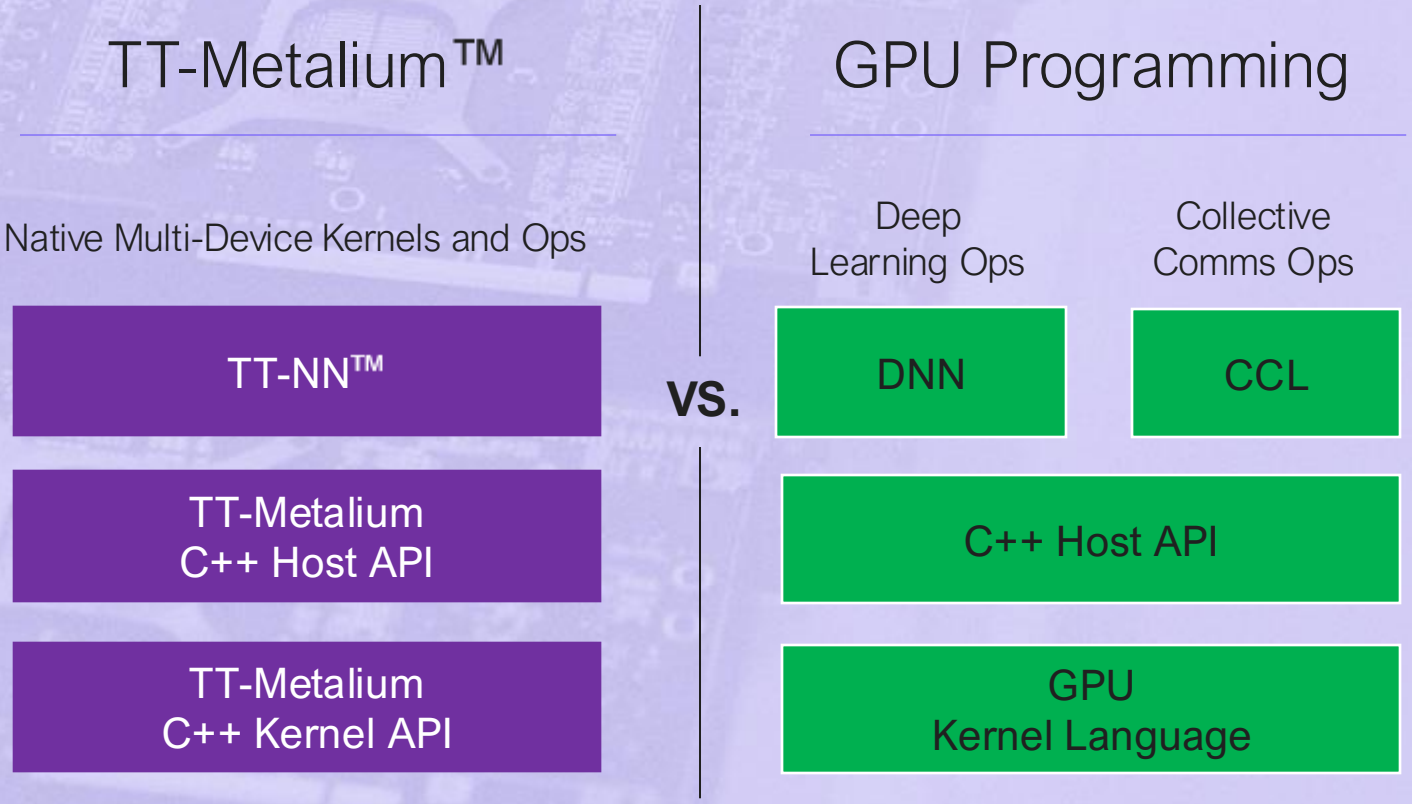
- **TT-NN™** – Library of optimized operators
  - ATen coverage
  - PyTorch-like API

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



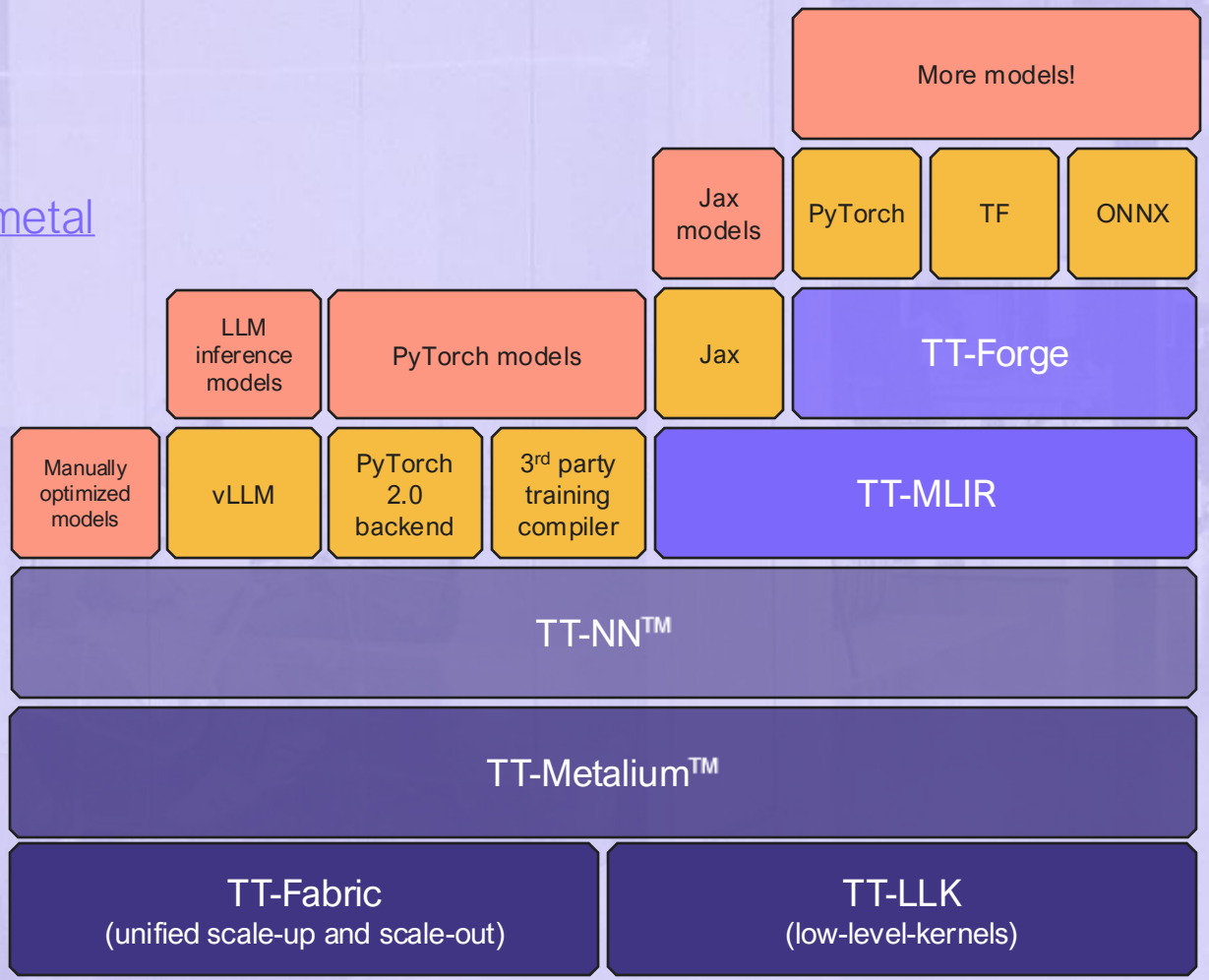
TT-Metalium™: Built for AI and Scale-Out

- Kernels are plain C++ with APIs
- Dedicated data movement and compute kernels
  - Optimize data movement and compute overlap directly
- Any core can read/write/sync to any core or chip directly
- Full control of data layout and persistency in SRAM and DRAM
- Different cores can run different kernels and flow data directly between them
- Native multi-device kernels
  - Fused and overlapped compute and inter-chip communication within the kernels



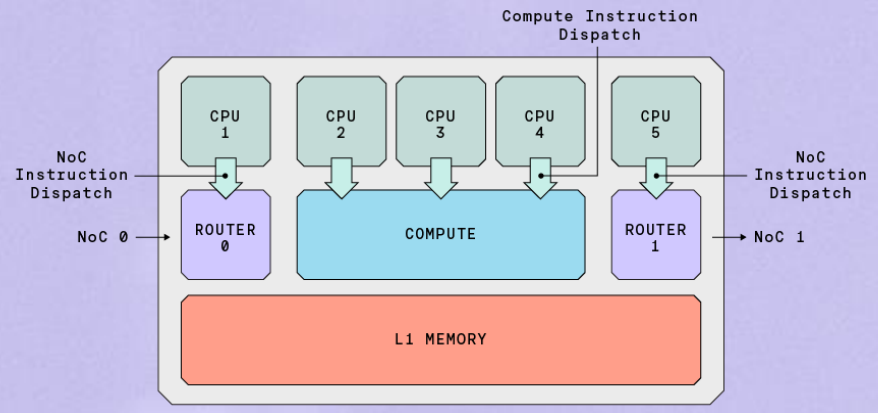
Software Ecosystem and Integrations

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

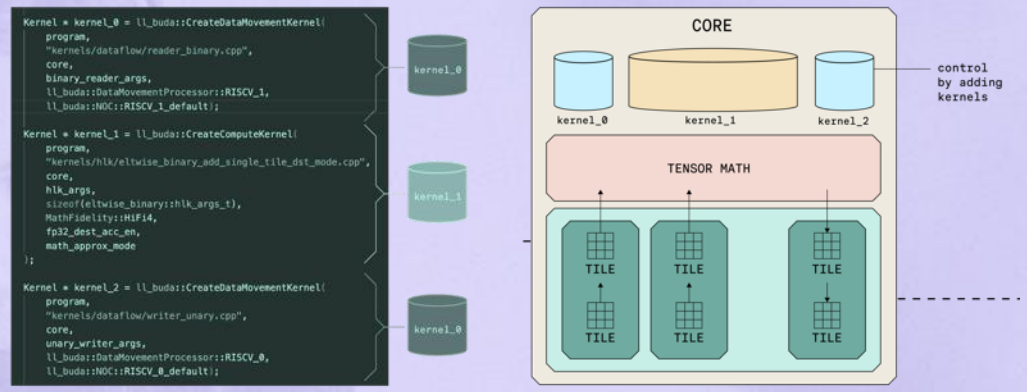


TT-Metalium™: Tensix Core to Multi-Chip Scale-Out

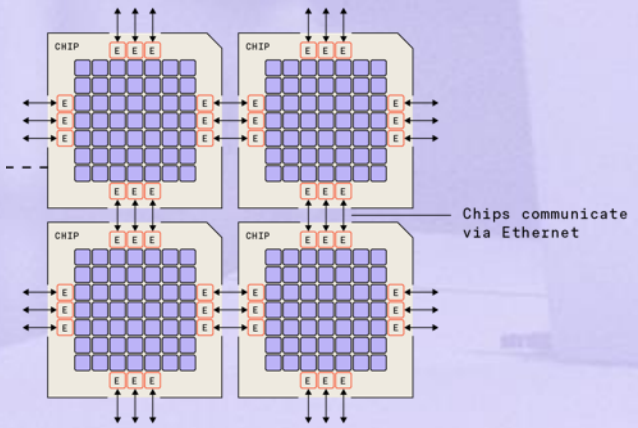
Tensix Core (Direct Access)



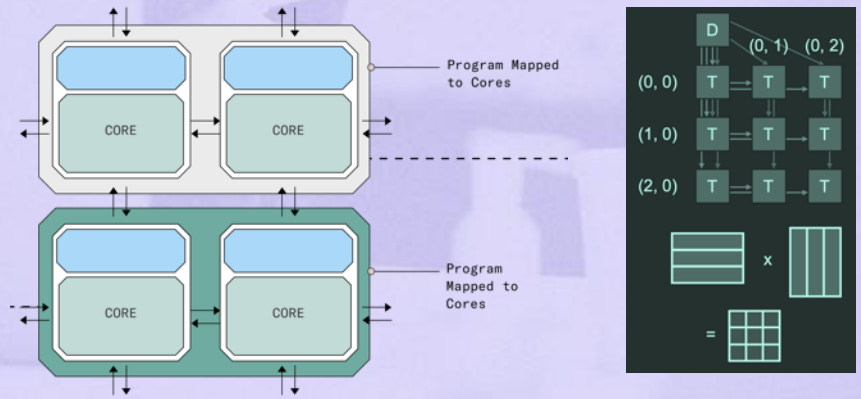
Compute & Data Movement Kernels (Decoupled)



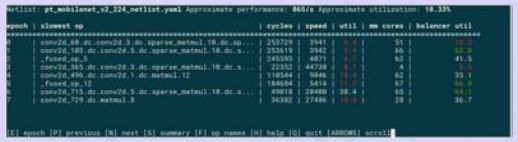
Multi-Chip Scale-Out (Sea of Cores)



SIMD/MIMD & Multi-Cast (Across Chip)



Simple, practical and intuitive tooling and debug utilities



Perf Analyzer

Our Perf Analyzer utility shows the sequence of on-chip operations at runtime, exposing individual op performance.

This helps users identify operations in a process which are in a waiting state so that suboptimal workflows can be identified and addressed.

Reportify

Tenstorrent AI Hardware processes graph-type applications (such as Deep Learning).

To facilitate the understanding of the graph architecture, we have included a visualizer to help expose the connections between the layers of graphs and help refine the application workflow.

Human Readable IR/Netlist

Tenstorrent software compiles a human readable netlist describing how operations will be mapped to Tensix cores, just prior to compiling the machine readable binary.

This provides expert developers with the flexibility to modify the placement and routing on the chip for fine-grain optimizations.

RouteUI

To help developers visualize the spatial mapping of operations on Tensix cores, we offer the RouteUI utility that display on-chip operations.

This enhances understanding of our dataflow approach and enables bottleneck identification for performance improvement.

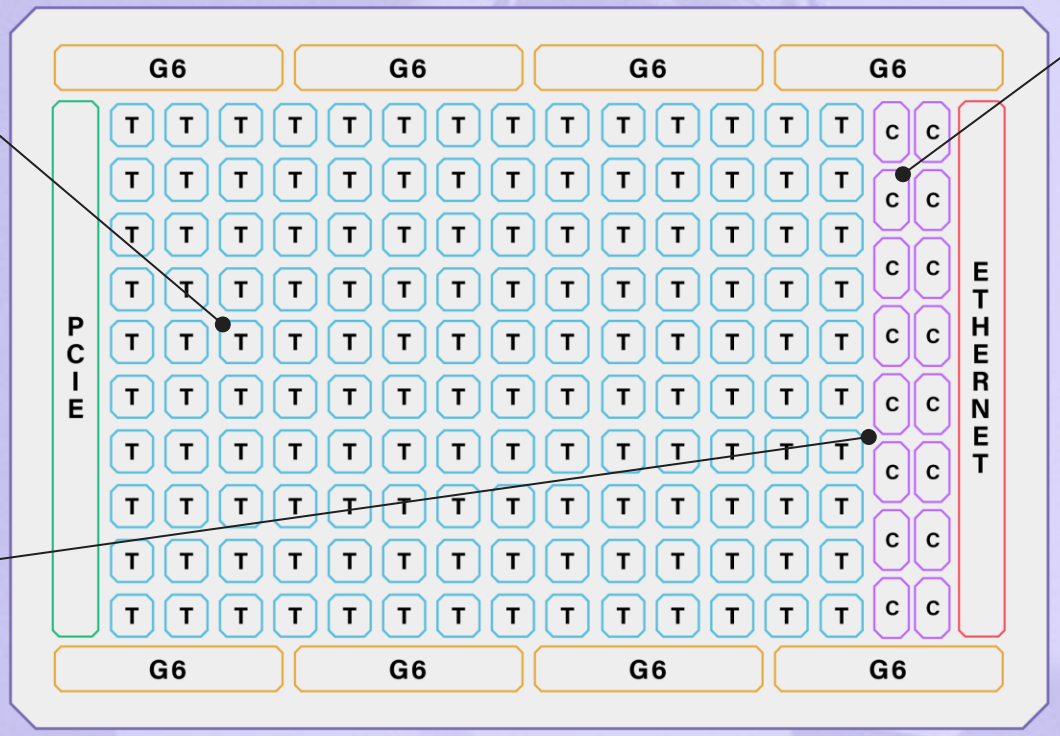
Why AI Needs Both RISC-V Cores and AI Accelerators

Tensix cores are ideal for big math operations:

- Vector calculations
- Matrix arithmetic
- Large data sets

Merging Tensix cores and CPU cores on the same die:

- Lowers latency
- Boosts utilization
- Increases ML performance

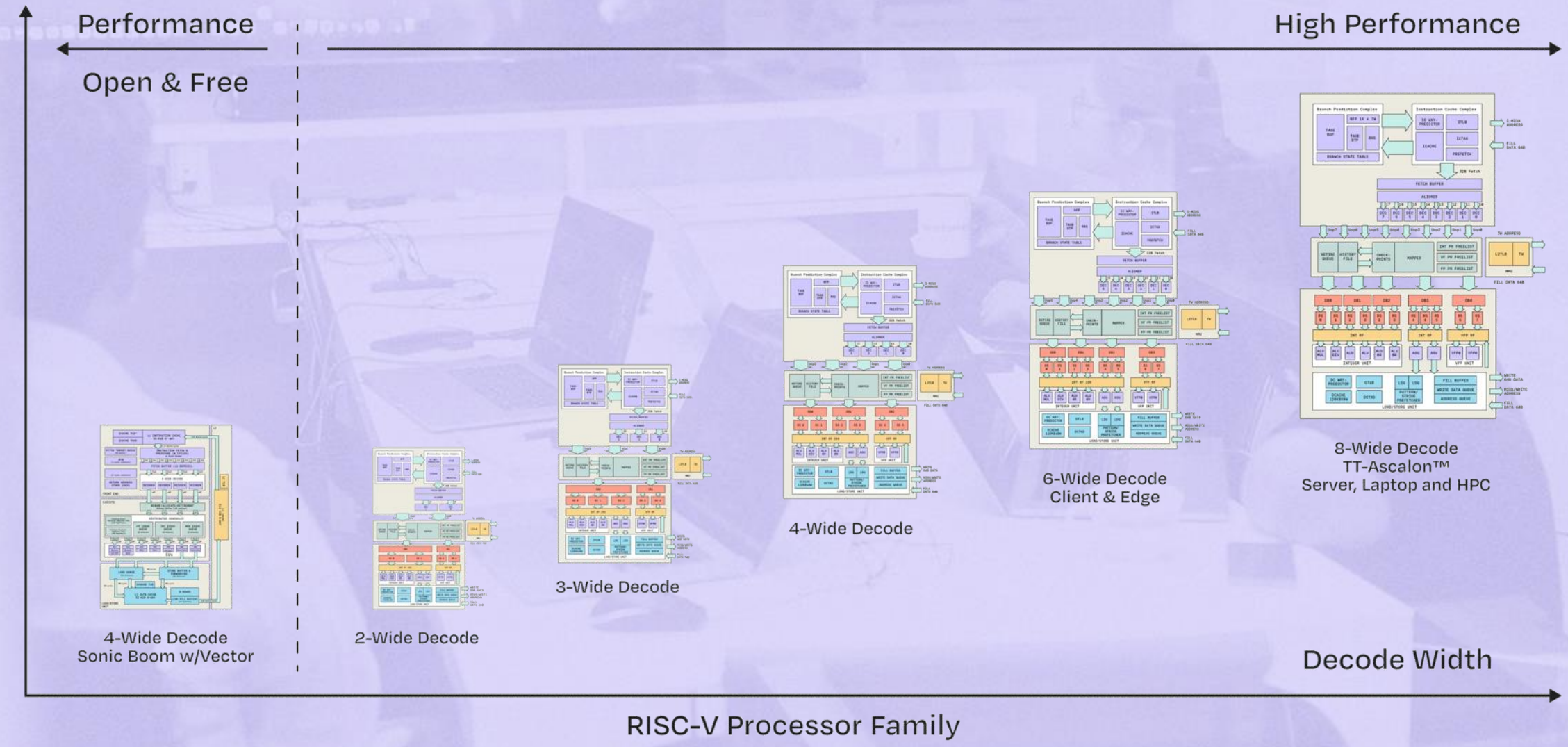


CPU cores are ideal for:

- Conditionality
- Traditional math
- High performance
- Robust programmability

ML Developers need both CPU and AI cores to build dynamic models of the future that are not possible today due to latency and utilization problems of using the host CPU.

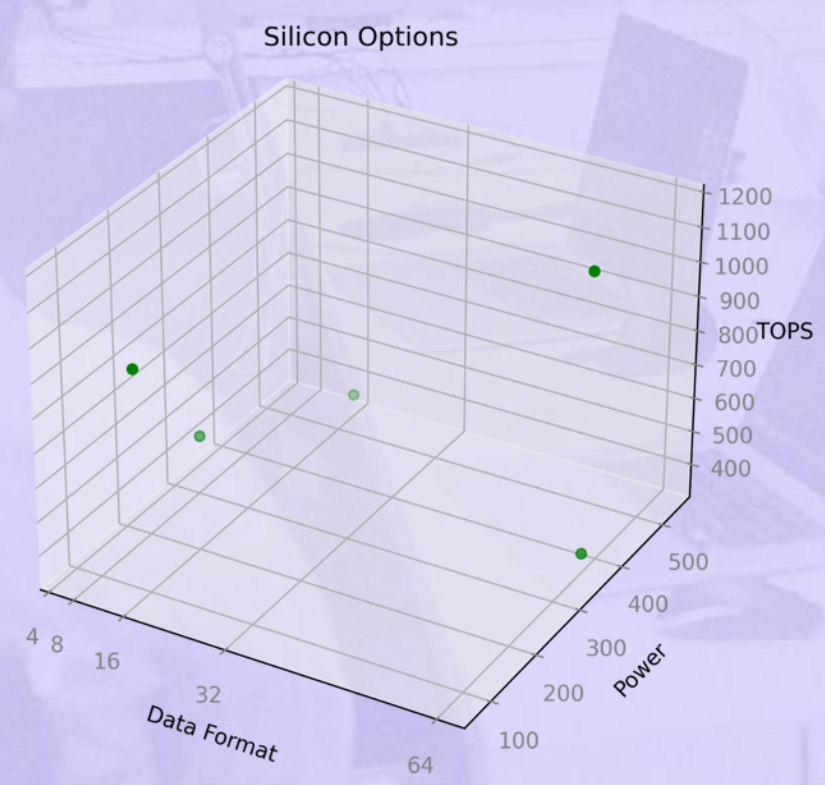
Tenstorrent RISC-V O-o-O Processor Family



IP Customization Advantage

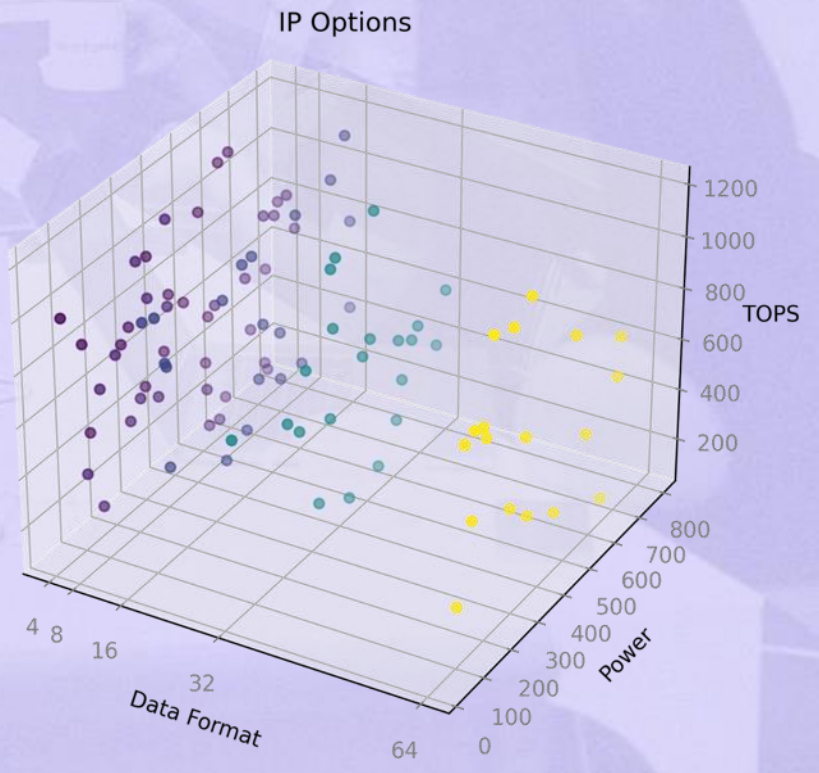
Silicon Providers

Choose from a small set of available options



Tenstorrent

Get exactly what you need



- ✓ FP4
- ✓ FP8
- ✓ F16
- ✓ FP32
- ✓ FP64

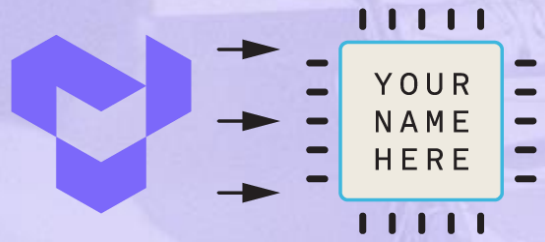
Tenstorrent CPU IP Licenses

Tenstorrent offers **two** CPU IP licensing options

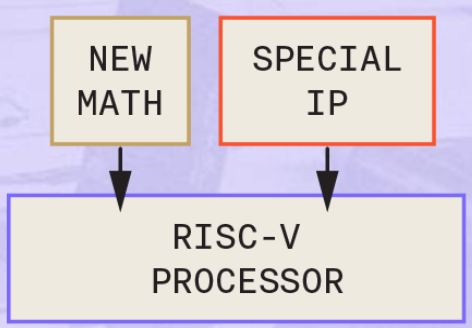
Innovation CPU IP License	**Standard** CPU IP License
• Fully modifiable license to Tenstorrent CPU IP enabling faster time to market and differentiated CPU products • Source RTL provided for faster time-to-market	• License Tenstorrent CPU IP without modification rights (limited rights are negotiable) • Encrypted RTL provided
• Access to infrastructure IP required for modifying and extending Tenstorrent CPU IP design • Custom instructions could be possible	• Licensees can configure IP with various parameters but cannot extend beyond allowed design space • Any unauthorized modification voids support and maintenance; warranties; indemnification, etc.
• Licensable option for branding/naming rights	• Licensable option for branding/naming rights

Innovation License

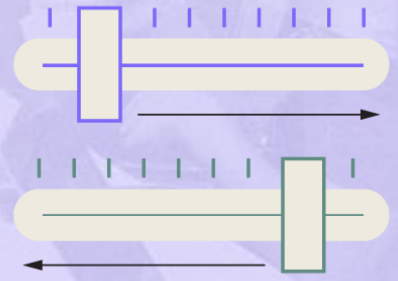
RISC-V



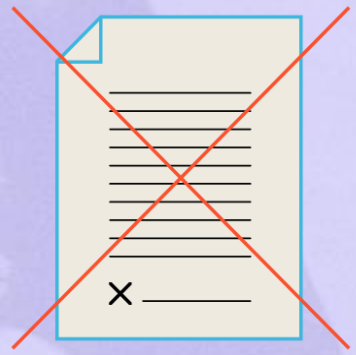
Fully customizable
Complete ownership
Source RTL for faster TTM



Change the ISA
(Do what x86 and ARM cannot)



Optimize performance for
your specific workloads



No crazy license
restrictions


Thank You