Tensix AI processor IP

IP Business Unit

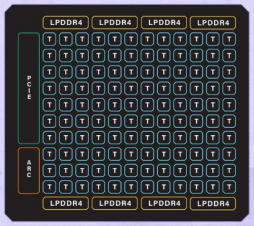
tenstorrent

Silicon Roadmap

2021	2022	2023	2025
HIGH PERF AI ASIC	SCALABILITY	HETEROGENY	CHIPLETS

Grayskull

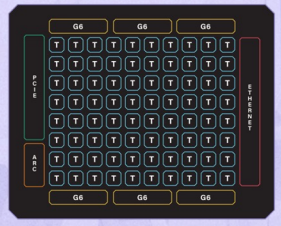
AI Processor



- 120 Tensix Cores
- 12nm
- 276 TOPS (FP8)
- 16 lanes of PCIE Gen4
- 8 channels LPDDR4

Wormhole

Networked AI Processor



- 80 Tensix+ Cores
- 12nm
- 292 TOPS (FP8)
- 16x100 Gbps Ethernet
- 6 channels GDDR6
- 16 lanes of PCIE Gen4

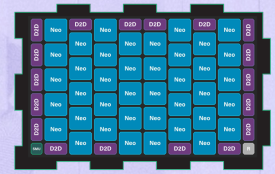
Blackhole

Standalone AI Computer

- 140 Tensix++ Cores
- 6nm
- 790 TOPS (FP8)
- 12x400 Gbps Ethernet
- 48 lanes of SERDES
- 8 channels of GDDR6
- 16 RISC-V CPU cores

Quasar

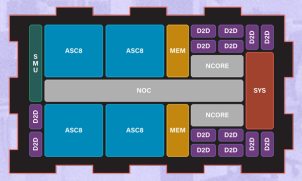
Low Power
AI Chiplet



- 160 Tensix Neo Cores on 4nm Chiplet
- Features incl SMC with Self-boot/Reset
- Non-blocking D2D Interfaces
- Easily stack Quasar or combine to choose your own compute

Aegis

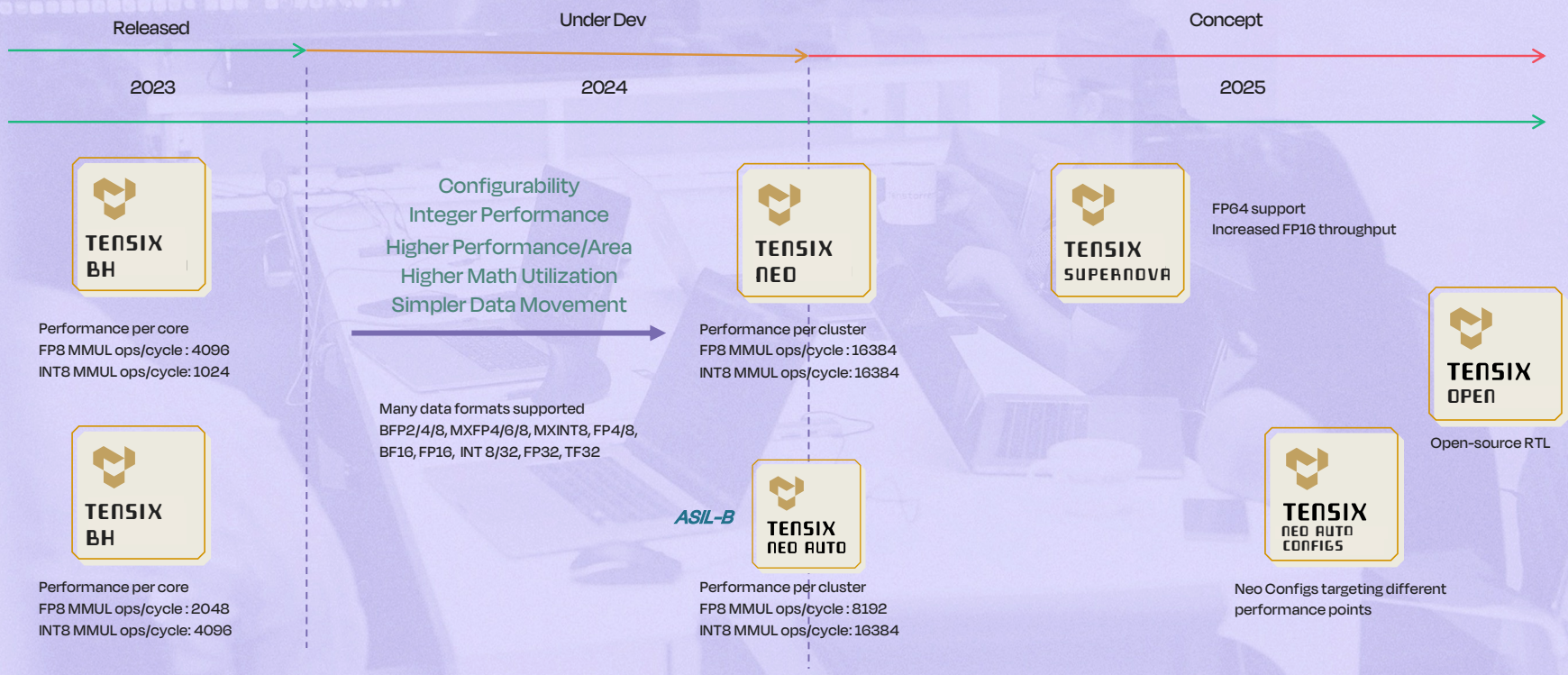
Highly Performance
RISC-V CPU Chiplet



- 4nm 32 RISC-V Ascalon CPU Cores
- Scalable up to 128 Cores / 4 Chiplets
- Feature support incl SMC, IOMMU, AIA
- Non-blocking D2D Interfaces
- Composable IO, MEM, CPU compute

GEN #1 → GEN #2 → GEN #3

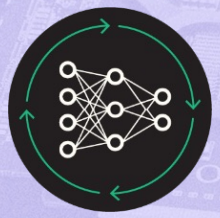
Tensix AI IP Roadmap



Disclaimer: Parts in 'Concept' phase are subject to change

Tensix: IP for your IP Workload

Efficient : Industry leading Perf/W and perf/$$



Training & Inference

Silicon proven

Scalable & Configurable

Easy to Use & Mature Software stack(s)

Tensix-BH Features

Near-memory compute

- High-bandwidth L1 <-> Register interface
- Vector int/float accumulation in L1(1.5MB) memory

Many data formats supported

- FP8, BFP2/4/8, INT8, INT32, BF16, FP16, TF32, FP32

General purpose SIMD engine (SFPU)

- Fast transcendental instructions (Gelu, Relu, Exp)
- SFPU C++ compiler

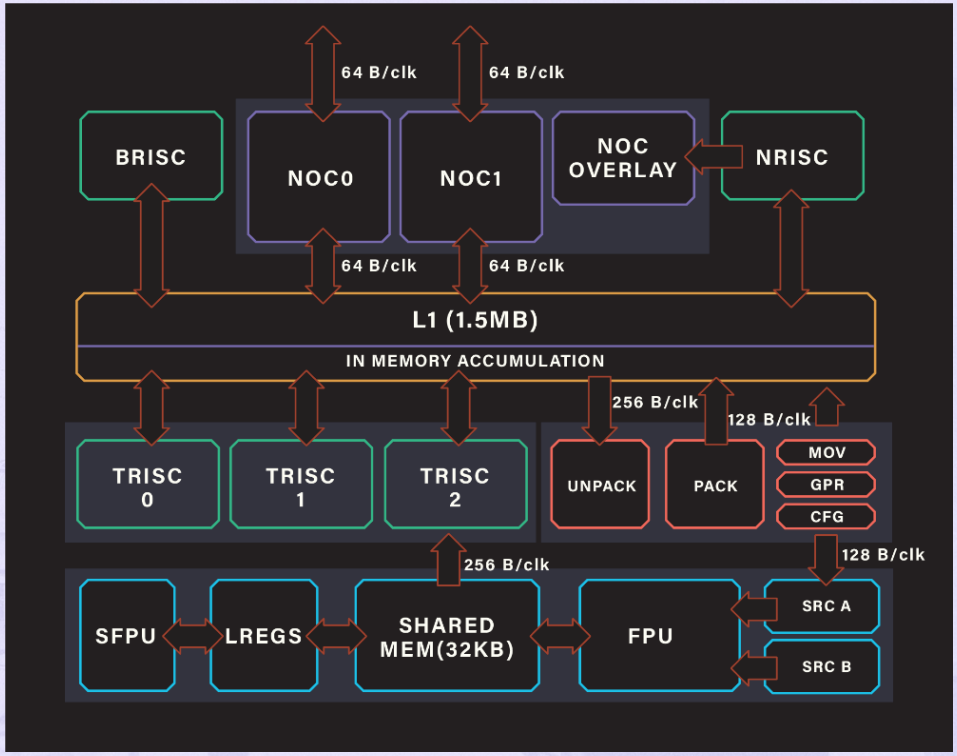
5 C++-programmable RiscV cores per Tensix

- RISC-V cores manage overhead of computation and NoC transfers
- Icache, data cache, data local memory, instruction prefetcher, branch predictor, floating-point and atomics support
- Support for interrupts

Sparsity

- HW support for fine-grained structured sparsity

Silicon Proven with Grayskull, Wormhole, and Blackhole



Data Formats

Format	Spec.
BFP8/BFP4/BFP2 (A/B)	Shared exp, 1-bit sign, 7/3/1-bit mantissa
FP8	4-bit exponent, 3-bit mantissa 5-bit exponent, 2-bit mantissa
Integer	Int8: signed/unsigned Int32: signed
FP16 (A/B)	A: 1-bit sign, 5-bit exp, 10-bit man B: 1-bit sign, 8-bit exp, 7-bit man
TF32	1-bit sign, 8-bit exp, 10-bit man
FP32	1-bit sign, 8-bit exp, 23-bit man

FPU multiplier operates on 5b x 7b input and uses phases for higher fidelity

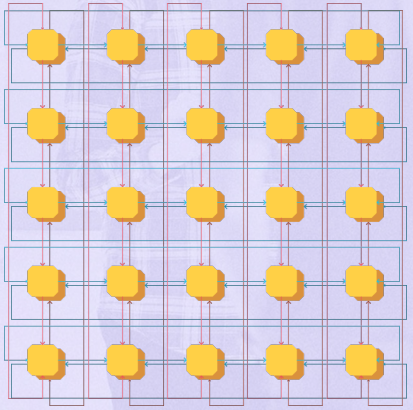
- BFP4 -> 1 phase (2K MACs/cycle/core)
- BFP8 -> 2 phases
- FP16/TF32 -> 4 phases
- BFLOAT can be done in 2 phases with loss of 1 LSB

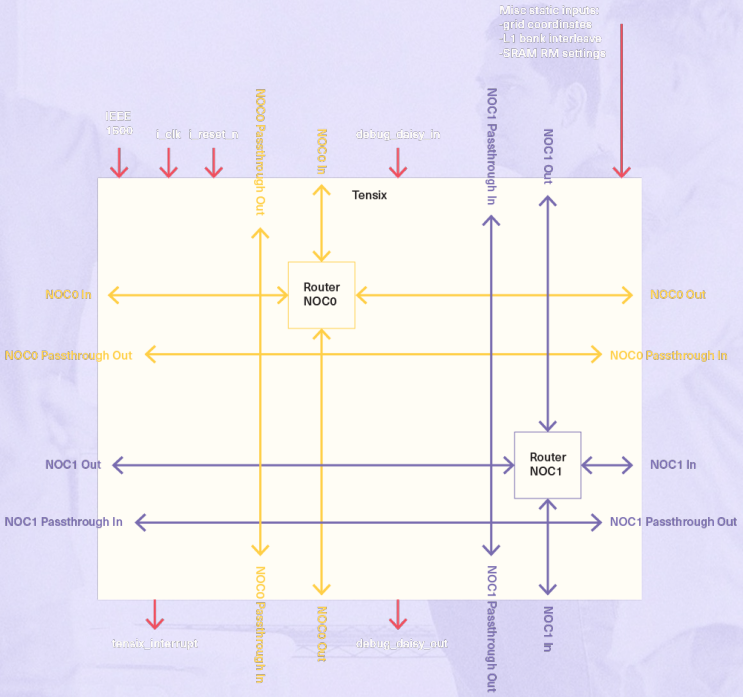
Tensix I/O

4 ring interconnects

- N/S: 1 ring CW and 1 ring CCW
- E/W: 1 ring CW and 1 ring CCW

Data width: 256b (WH), 512b (BH)






Tensix NEO Architecture

Tensix NEO

High performance / mm²

- Reduced total SRAM due to better utilization
- Fewer L2 banks with simpler crossbar
- Implementation improvements with tiled design
- Many micro-architectural optimizations

Improved SFPU and FPU utilization

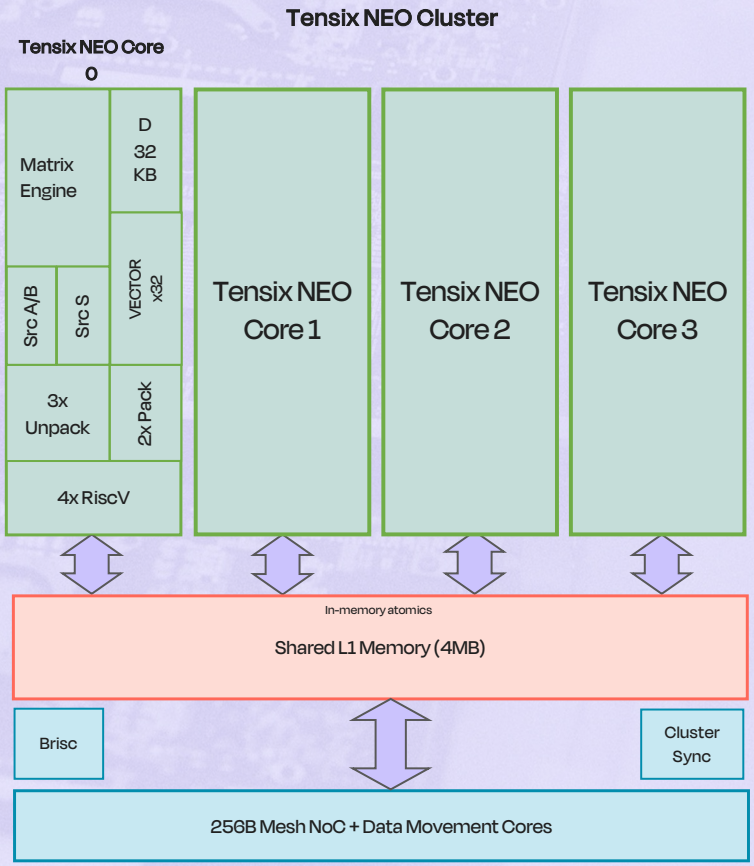
- Higher L1 bandwidth to more easily saturate FPU
- Optimized ISA for complex data patterns
- Dedicated RiscV core and registers for SFPU

Simpler data movement

- More powerful RISC-V data movement cores
- 64-bit address space
- Significantly smaller DMA engines
- Wide bi-directional mesh NOC

Even more data formats

- Support for new Microscaling formats
- Added FP4



Tensix Neo Data Formats

- Formats
  - FP16
  - BF16
  - UINT8/INT8
  - MXFP8
  - MXFP6P
  - MXFP6R
  - *MXINT8*
  - *MXINT4*
  - *MXINT2*
- Vector Unit with FP32 support

AI IP Configurations

Data-types

- Floating-point + integer
- Integer only

Throughput

- Configurable matrix and vector engine throughputs

RISC-V Cores

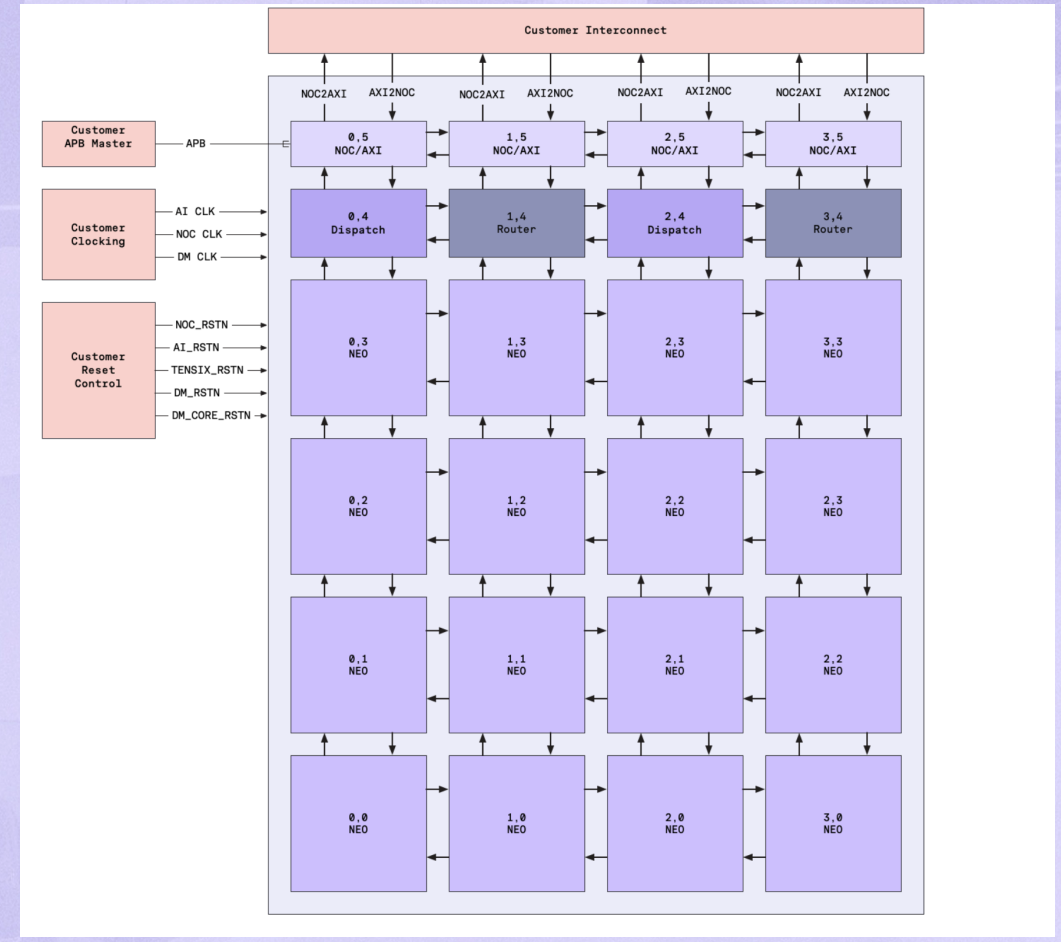
- Configurable cache sizes
- Optional vector unit with configurable vector width

L1 Memory size

- Number of banks 16/32
- Configurable space per bank

Neo Config-1

Feature	Neo Config-1
Number of Tensix	4*4*4 = 64
SRAM/Tensix	0.75MB
Tensix/Dispatch	8
Dispatch SRAM	4MB
NOC BW (Aggregate)	256GB/s
I/O BW	256GB/s
DRAM BW (Per Endpoint)	15GB/s
DRAM Endpoints	8
Aggregate DRAM BW	120GB/s
Memory Locations	North Edge



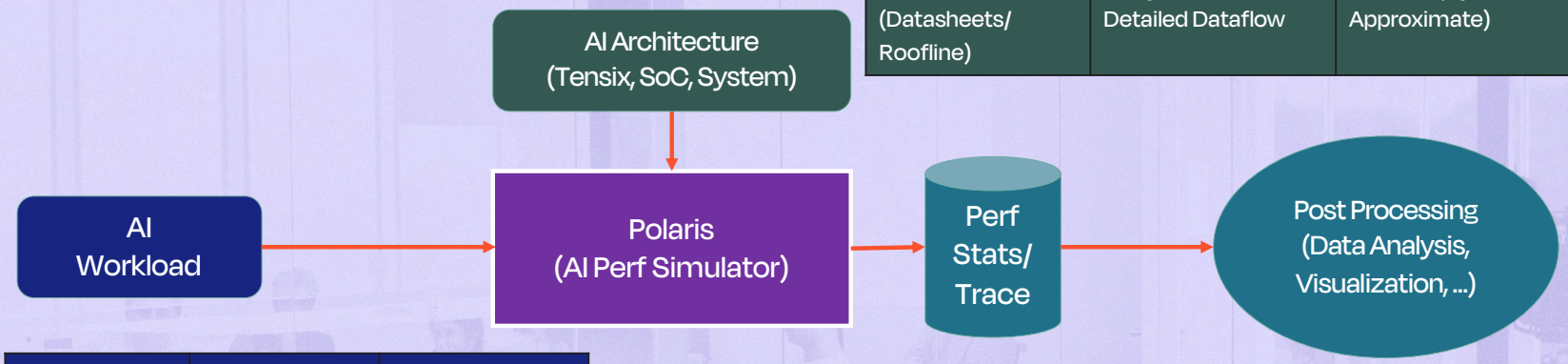
Tensix IP Configuration Overview

	BH	BH+	NEO Auto	NEO
Matrix Multiply Throughput				
FP4 MACs	N/A	N/A	4096	16384
BFP4/FP8 MACs	2048	1024	4096	8192
BFP8 MACs	1024	2048	4096	8192
BFLOAT16 MACs	1024	2048	4096	8192
FP16 MACs	512	512	1024	2048
INT8 MACs	512	2048	8192	8192
FP32 accumulator	Yes	Yes	Yes	Yes
Vector Multiply Throughput				
FP32 MACs	32	32	128	128
SRAM				
L1 size (MB)	1.5	1.5	3-4	4
NoC				
Topology	2D Torus	2D Torus	2D Mesh	2D Mesh
Link BW	64B/cycle	64B/cycle	256B/cycle	256B/cycle
ASIL	No	No	Yes	No
Physical				
Area	2.5	TBD	6.5mm²	6.3mm²
Power	2.11	TBD	2.0	3.0
Node	N6	N6	S5A	S4
Clock Speed	1.45	1.45	1	1.15


Simulation Modelling

Introduction

Abstract	Hybrid	Concrete
Speeds & Feeds (Datasheets/ Roofline)	Analytical Compute, Detailed Dataflow	Detailed (Cycle Approximate)



Abstract	Hybrid	Concrete
Analytical (Research Papers)	PyTorch 2.x Inductor / ONNX / MLIR	BUDA/Metal LLKs/Runtime

Different Abstraction Levels: Answer Specific Questions/Demands

- Abstract Models : Product/Roadmap Planning, Competitive Analysis, TCO Models
- Hybrid Models : Architecture/Design Space Exploration, HW/SW Codesign
- Concrete Models : Run real kernels on a single core

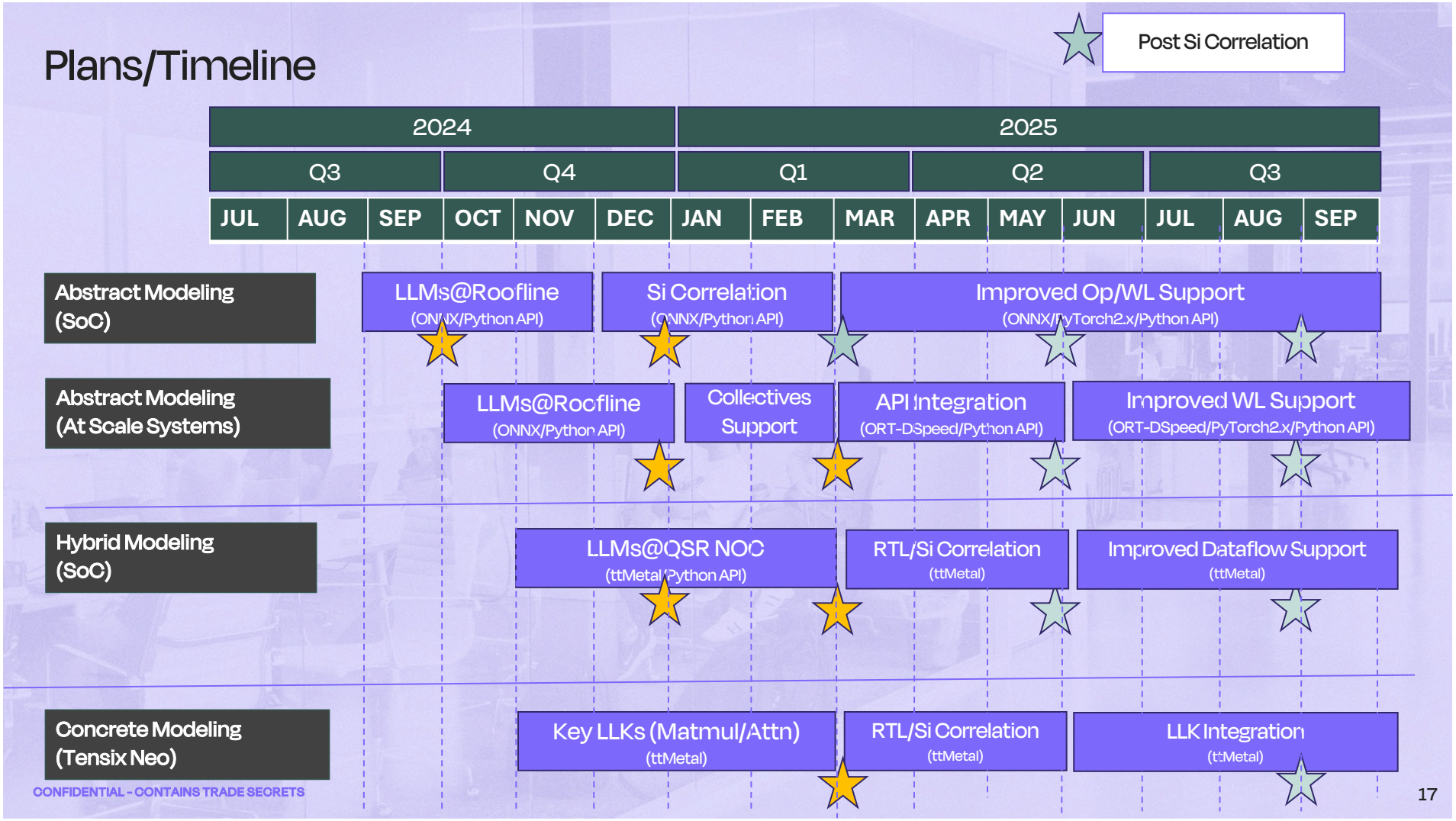
Architecture Performance Modelling Goals

- **Concrete Modeling**: Create A Configurable, Cycle Approximate, Microarchitectural Simulation Model for Tensix Core, with an ability to run real kernels (Binary/Source Compatible, Extensible ISA) on a single Core (or a small grid of Cores) in an idealized SoC setting

- **Hybrid Modeling:** Create A Configurable, High Level Architecture Simulation Model for TT SoC, to do performance projection of real workloads (Buda/Metal) on a single SoC (e.g. Quasar) or at scale (e.g., multiple Quasars + Ethernet Interconnect)

- **Abstract Modeling:** Create A Configurable, High Level Analytical Model for TT System Configuration TCO Analysis, Roadmap Planning, and Competitive Projections

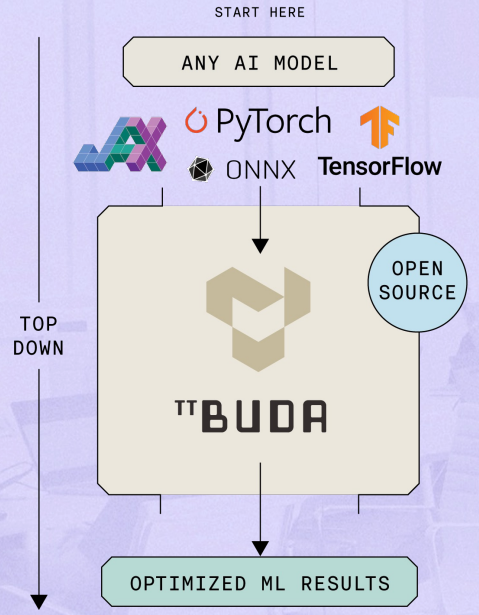
Plans/Timeline




Software

Tenstorrent Software – Two Different Approaches

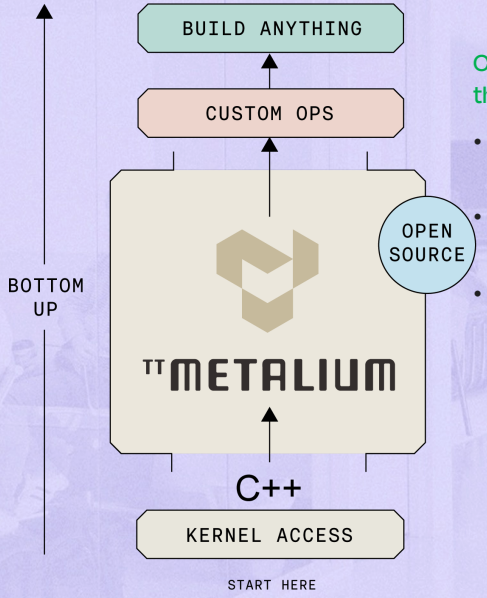
GENERALITY > PERFORMANCE



Quick deployment and improving performance

- Out of the box, run many models, small overhead incurred
- Performance continues to increase on monthly release cadence
- Support for all major frameworks including PyTorch and Tensorflow

PERFORMANCE > GENERALITY

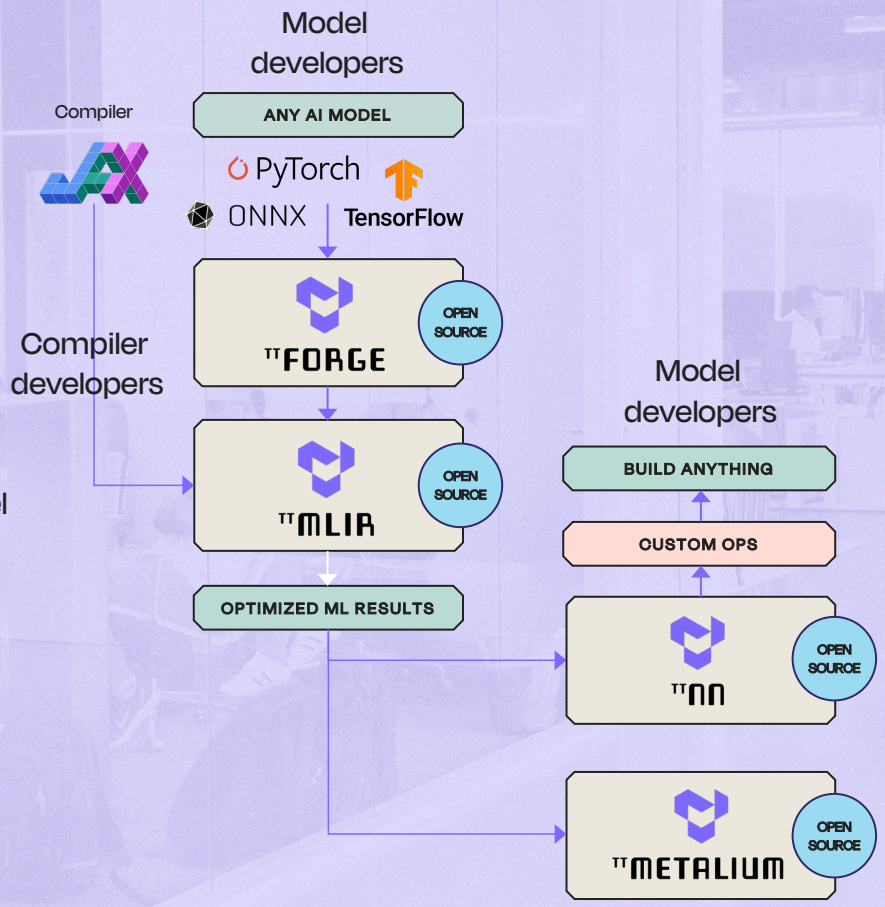


Complete low-level control of the hardware

- Targeted models running higher performance at higher utilization
- Completely open hardware allows you to program non-ML applications.
- Useful for HPC, C++ environments and low-level model development.

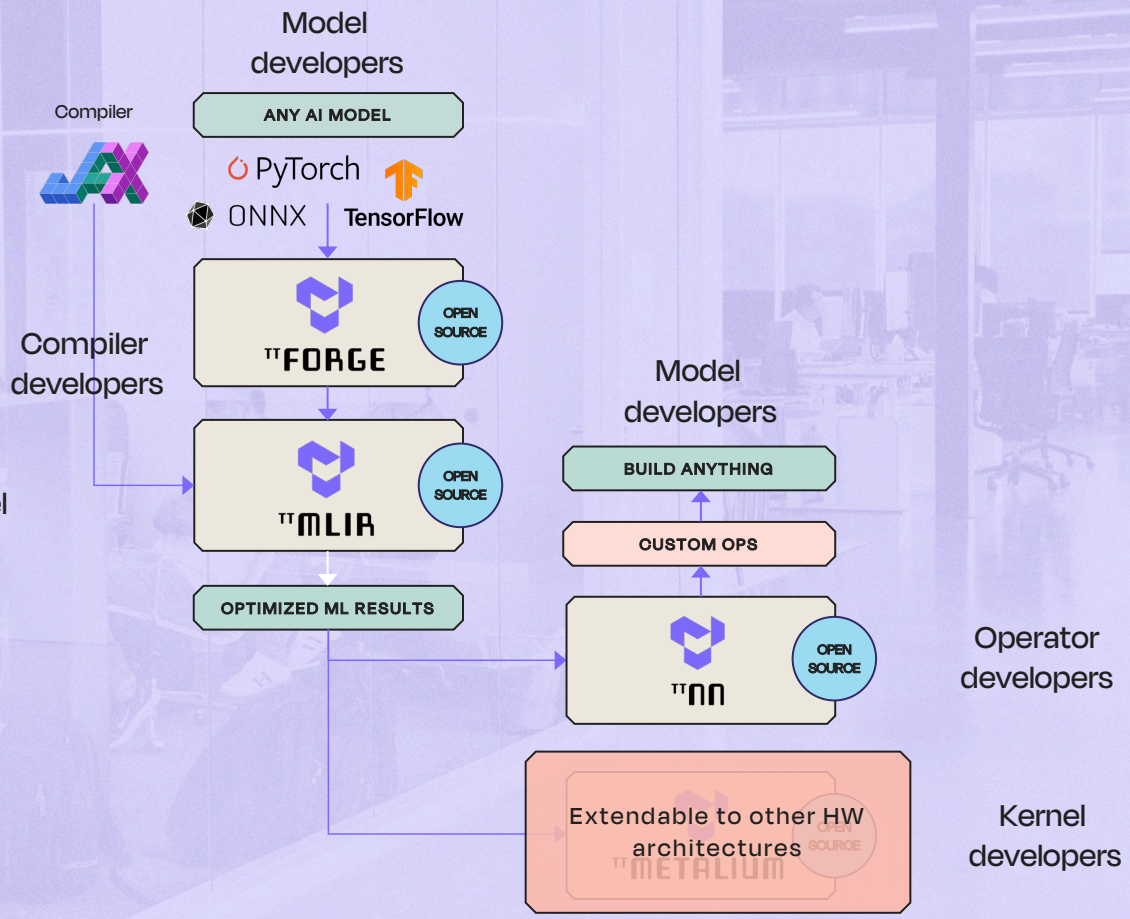
Open Tenstorrent Software

- **TT-Forge** - Integrated into various frameworks for native model ingest
- **TT-MLIR** - new MLIR-based compiler
- **TT-NN** – a library of optimized operators
  - ATen coverage
  - PyTorch-like API
- **TT-Metalium** – low level programming model & entry point



Open Tenstorrent Software

- **TT-Forge** - Integrated into various frameworks for native model ingest
- **TT-MLIR** - new MLIR-based compiler
- **TT-NN** – a library of optimized operators
  - ATen coverage
  - PyTorch-like API
- **TT-Metalium** – low level programming model & entry point



LLMs

Model	Batch	Hardware	ttft (ms)	t/s/u	Target t/s/u	t/s	Release
Falcon7B-decode	32	e150		4.2	4.4	134.4	
Falcon7B	32	n150	75	17.1	26	547.2	v0.53.0-rc33
Mistral-7B	32	n150		9.9	25	316.8	v0.51.0-rc28
Mamba-2.8B	32	n150	48	12.3	41	393.6	v0.51.0-rc26
LLaMA-3.1-8B	1	n150	291	22.9	23	22.9	v0.53.0-rc16
Falcon7B (DP=8)	256	QuietBox	101	14.4	26	3686.4	v0.53.0-rc33
LLaMA-3.1-70B (TP=8)	32	QuietBox	190	15.1	20	483.2	v0.53.0-rc33
Falcon40B (TP=8)	32	QuietBox		5.3	36	169.6	v0.53.0-rc33
Mixtral7Bx8 (TP=8)	32	QuietBox	235	14.2	33	454.4	v0.53.0-rc33
Falcon7B (DP=32)	1024	Galaxy	242	4.4	26	4505.6	v0.53.0-rc33
LLaMA-3.1-70B (DP=4, TP=8)	128	Galaxy	190	14.3	20	1835.5	v0.52.0-rc31

> **Last Update:** November 4, 2024

https://github.com/tenstorrent/tt-metal

CNNs

Model	Batch	Hardware	fps	Target fps	Release
ResNet-50 (224x224)	20	e150	5,100	10,000	
ResNet-50 (224x224)	16	n150	4,100	7,000	
ResNet-50 (224x224) (DP=2)	32	n300	8,200	14,000	
ResNet-50 (224x224) (DP=8)	128	QuietBox	32,250	56,000	
ResNet-50 (224x224) (DP=32)	512	Galaxy	95,900	224,000	
ResNet-50 (224x224) (DP=64)	1024	Two Galaxies	145,000	448,000	
ViT	9	e150	1,360	2,000	
ViT	8	n150	912	1,600	
Stable Diffusion 1.4 (512x512)	1	n150	0.167	0.3	
Yolo V4 (320x320)	1	n150	95	300	

NLPs

Model	Batch	Hardware	sen/sec	Target sen/sec	Release
BERT-Large	12	e150	370	410	
BERT-Large	8	n150	270	400	
T5 small		e150	140		
Bloom		e150	70		

https://github.com/tenstorrent/tt-metal


Thank You!