Tenstorrent

High Performance Computers for HPC & AI

@SC24

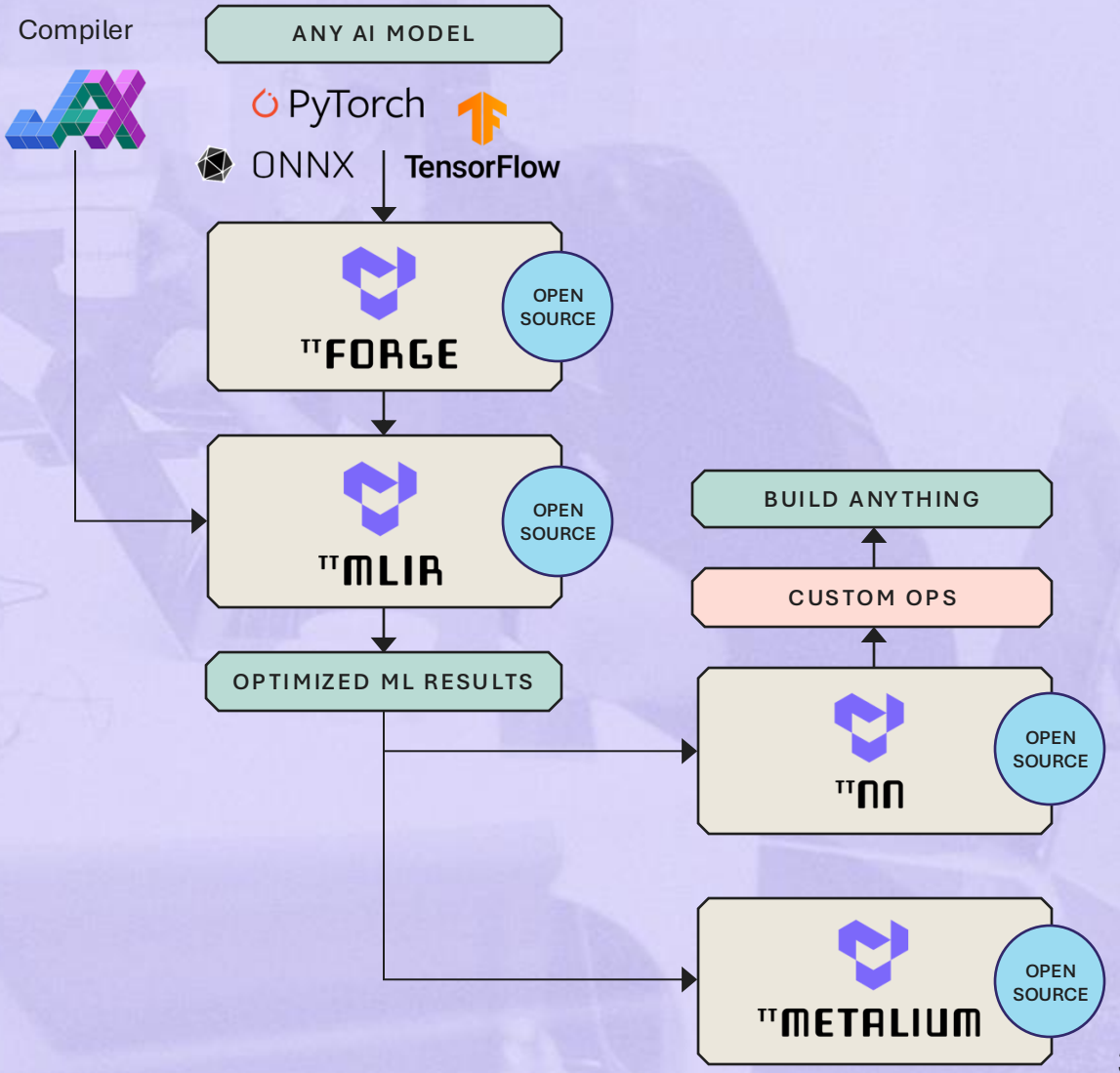
Nov 20, 2024

tenstorrent

Highlights:

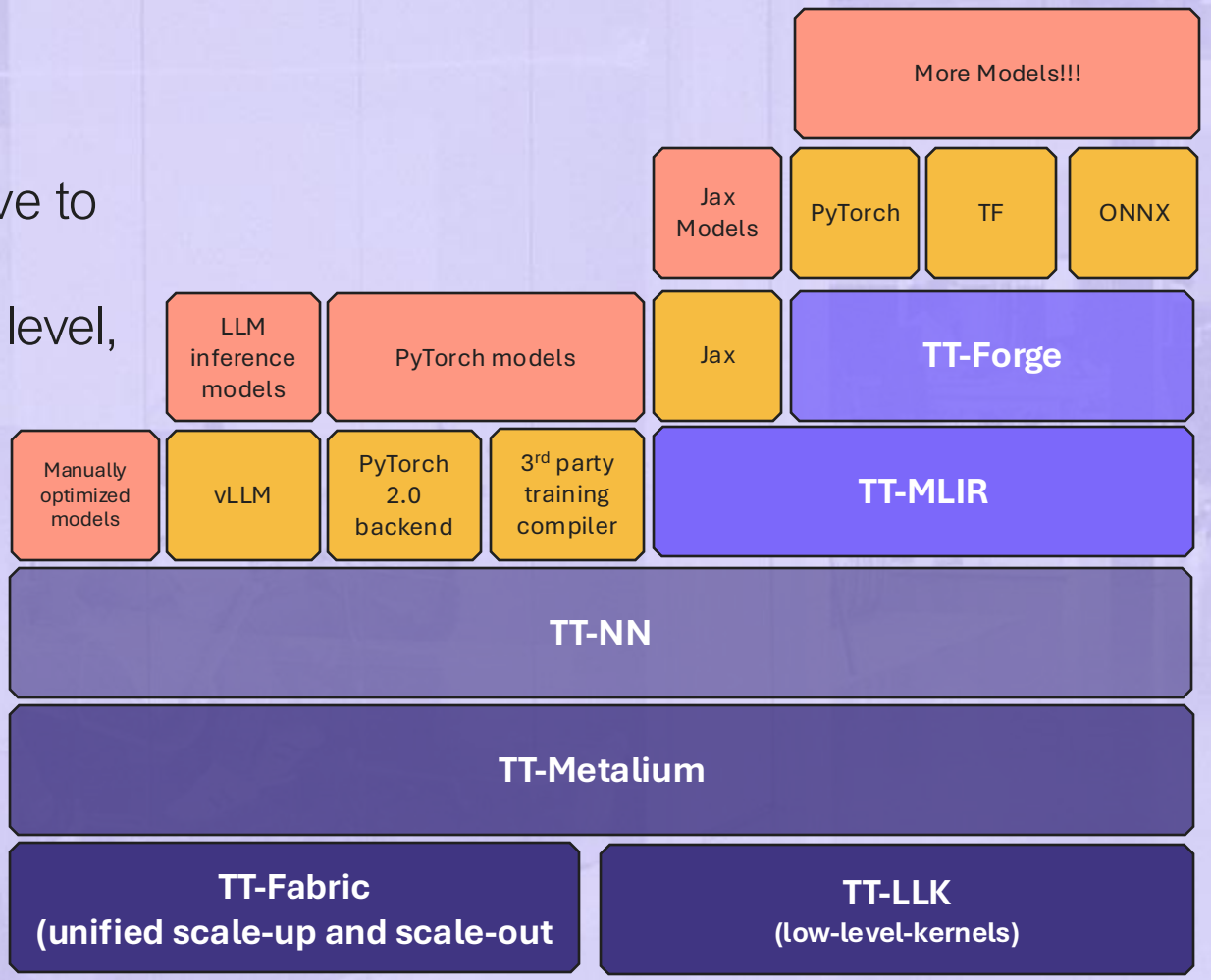
HPC Software roadmap – AI – Enable every level of Developers

- AI HPC
  - Leverage work from "Pure" AI stack
  - Flexible entry points for high level model and python developers

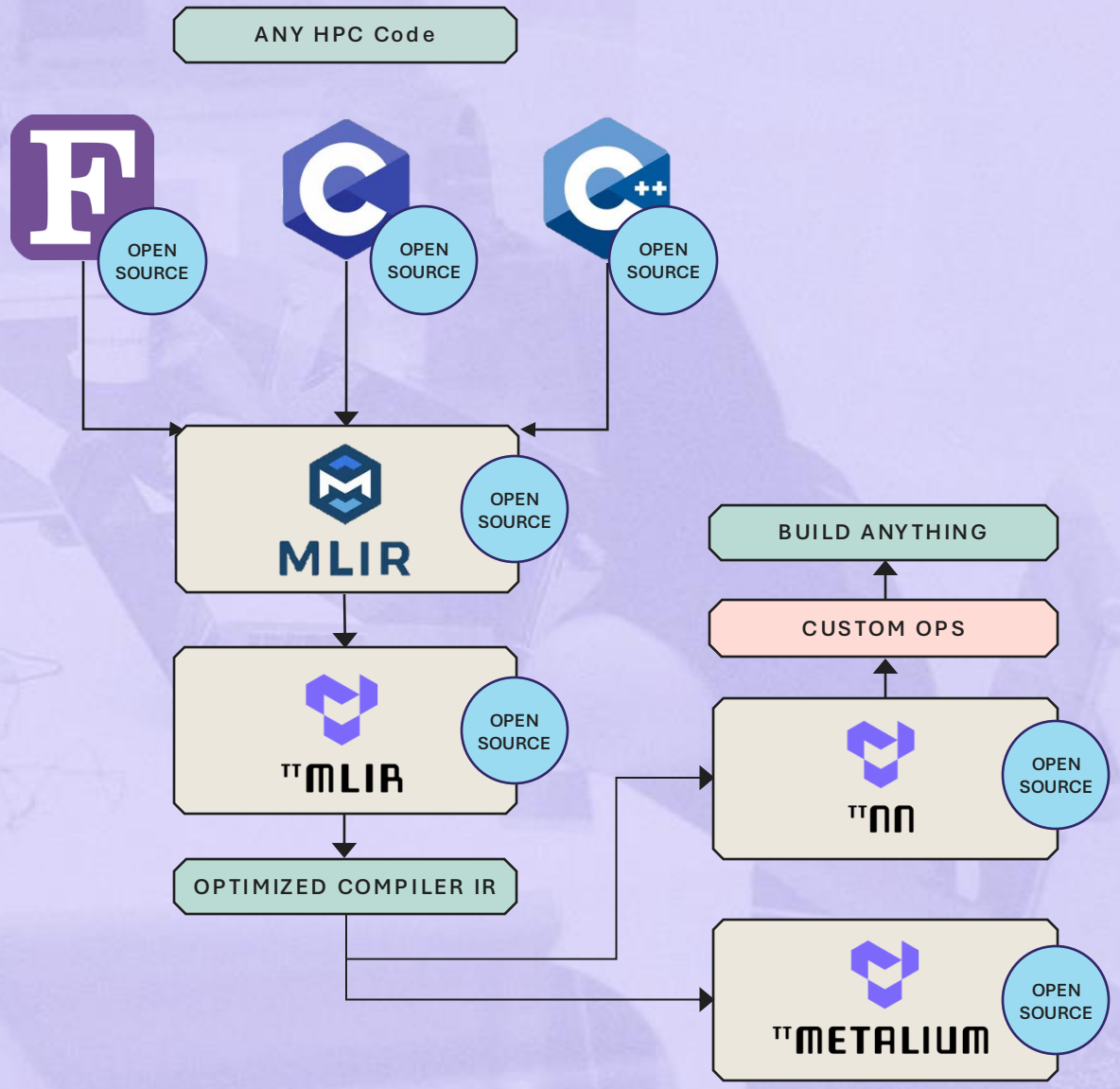


HPC Software roadmap – AI HPC

- Developers are tired of needing to understand everything before they get started
- Enable users to use the models they have to get reasonable performance
- Focus on enabling perfomance at every level, and meet developers where they are



- Classical HPC
  - Leverage work from AI stack - MLIR
  - Flexible entry points for developers
  - Leverage and contribute to the Open Source community
  - Provide users a friendly, familiar baseline to start, enable tooling to go further



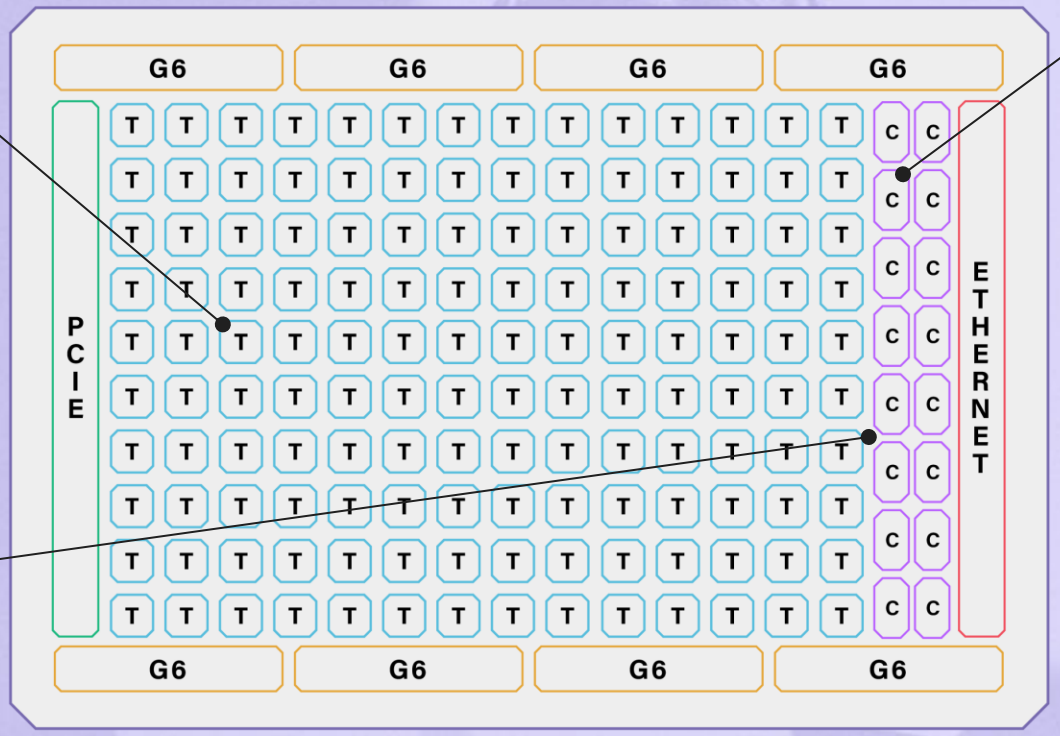
Why AI Needs Both RISC-V Cores and AI Accelerators

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

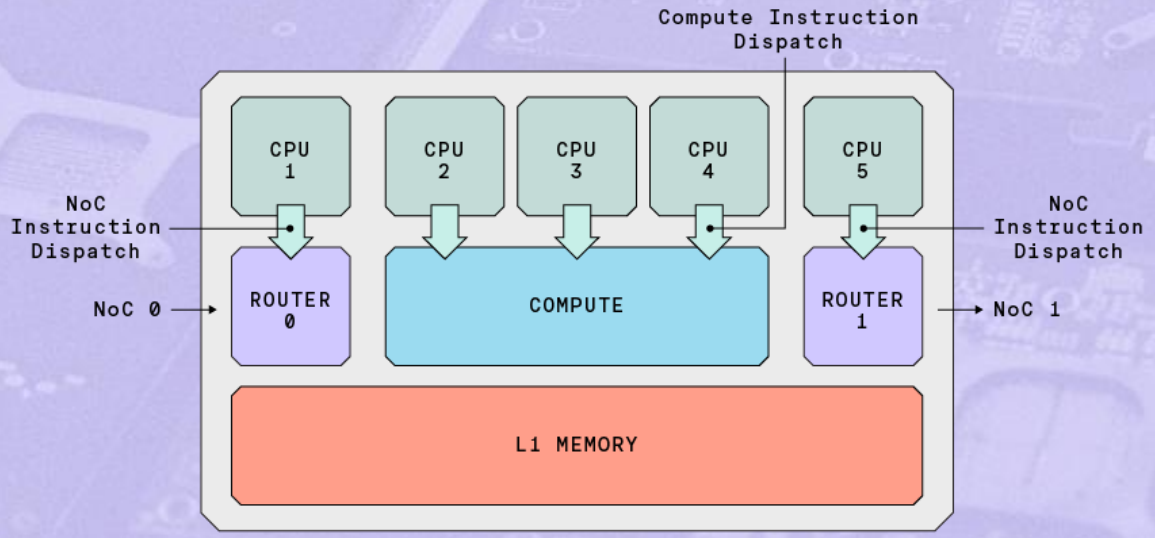


CPU cores are ideal for:

- 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.

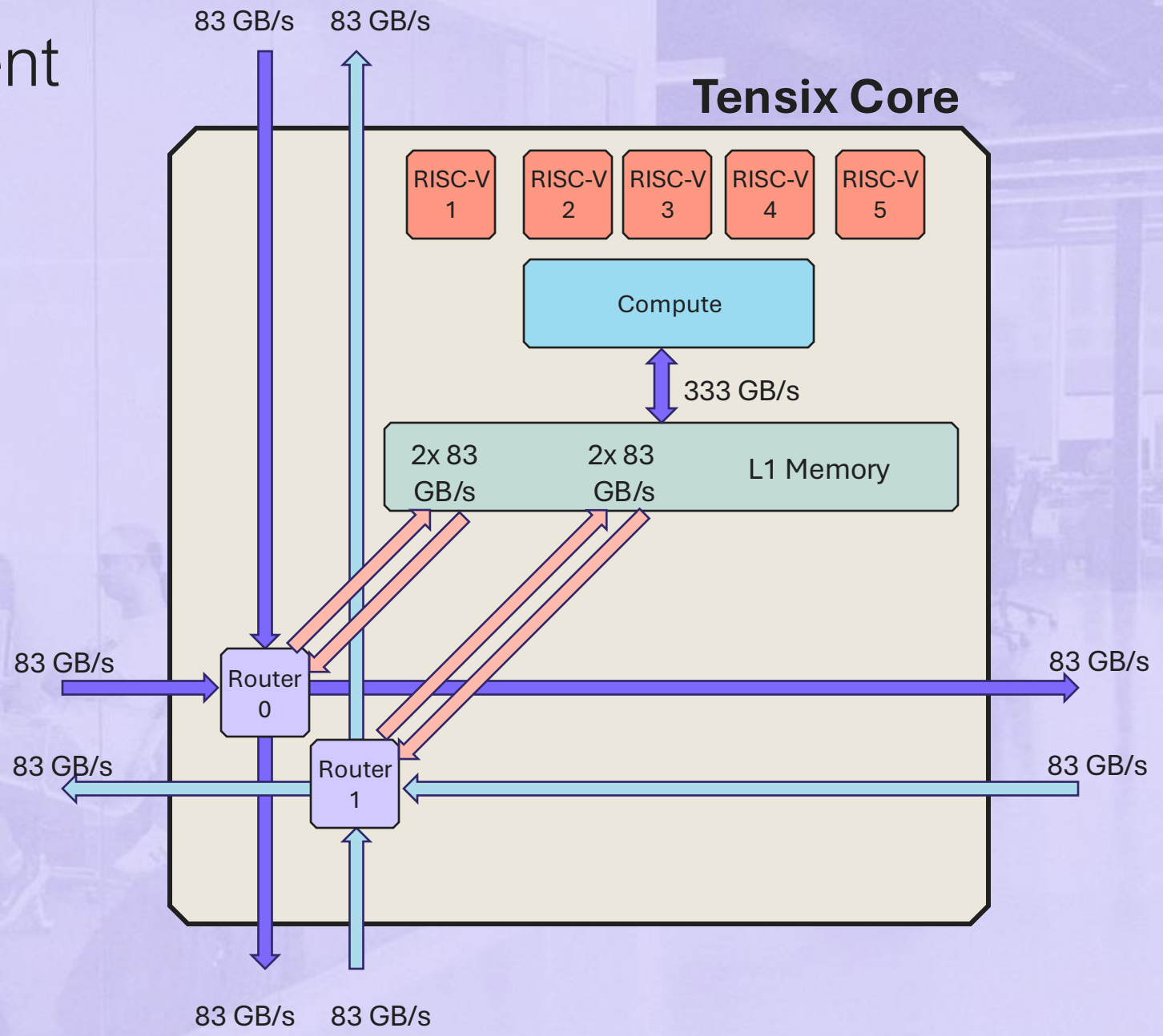
TT-Metalium™: Tenstorrent's Low-Level Programming Framework



- TT-Metalium™ is a groundbreaking low-level programming framework designed to harness the power of Tenstorrent's parallel processing cores
- Each processing core in our Tensix architecture contains 5 "baby RISC-V" cores: 3 programmable cores and 2 cores for Network-on-Chip (NoC) management
- Our framework aims to maximize performance, efficiency, and flexibility

Tensix Core – Data Movement

Feature	Spec
Independent NoCs	2
NoC type	2-dimensional torus
NoC link width	64 Bytes
NoC link BW	83 GB/s
Tensix -> NoC I/O BW	665 GB/s
SRAM <-> NoCs	333 GB/s
SRAM <-> NoC aggregate BW	47 TB/s



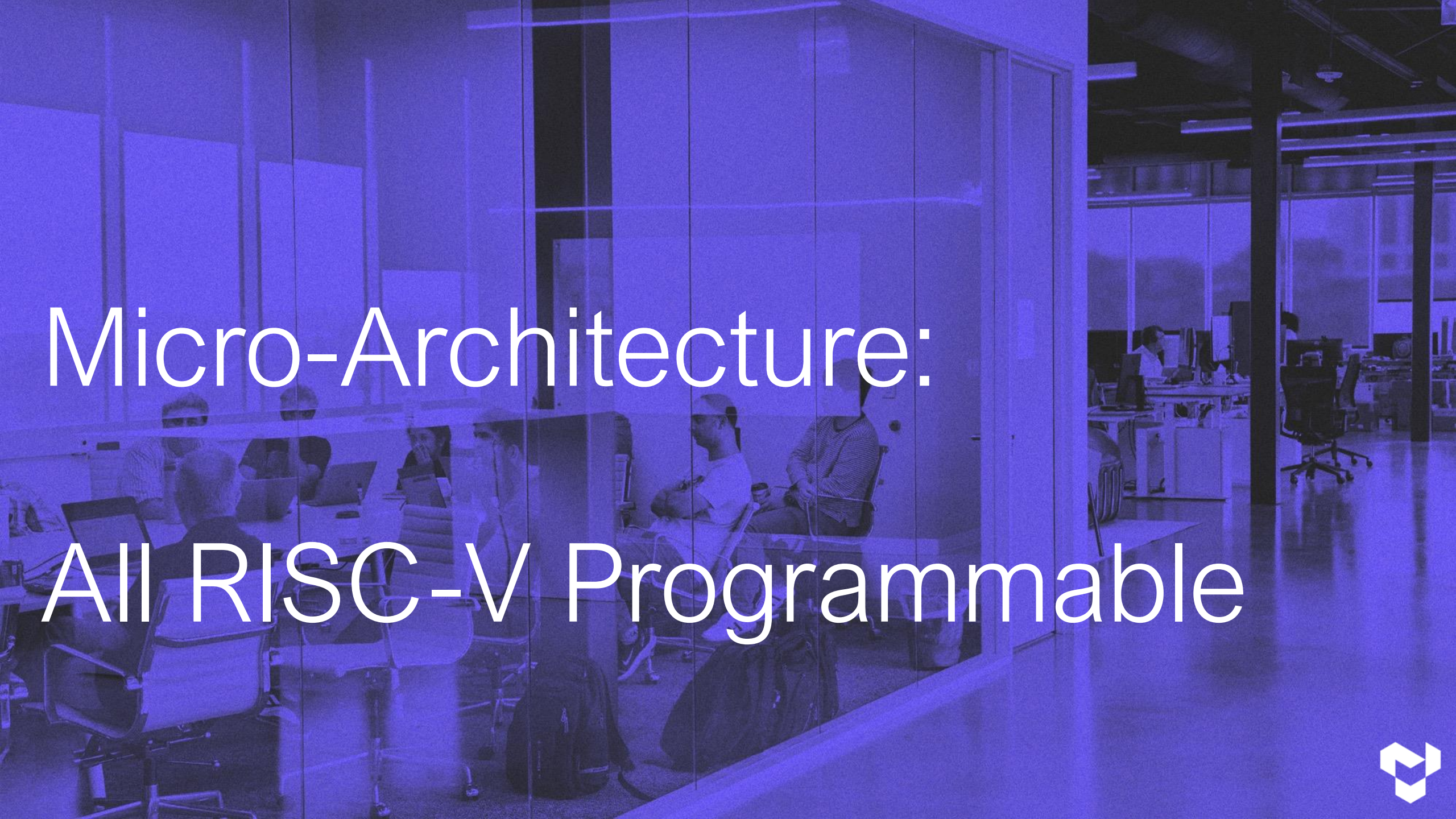
Performance

TT Tensix HPC Software Plan

Programming Model	Compiler	Communication Library	Math		Profiler	Debugger
TT-Metalium™ (C++ Based)	C / C++ (GCC for RISC-V) Kernels	TT-CCL	TT-BLAS	TT-NN™	Tracy Host Profiler	TT-Metalium GDB
FLANG-MLIR Tensix Fortran	GCC for x86 Host	Multi-Node Scale-Out	Sparse	Solver	Tracy Device Profiler	Watcher
REDACTED	GCC for ARM Host	OpenMP/MPI on Host Only	FFT	Rand Gen	Memory Allocation	Kernel PRINT

Today

2025-2026

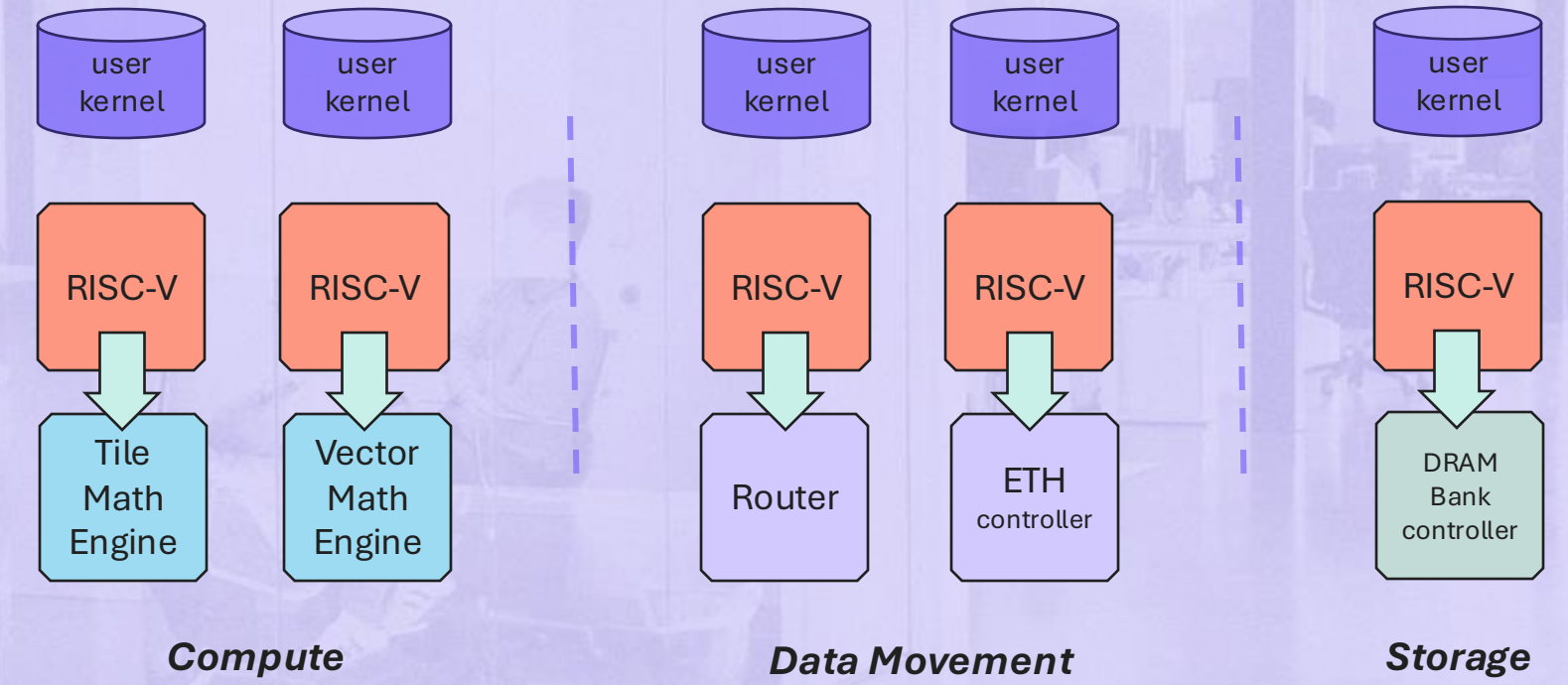


Micro-Architecture:

All RISC-V Programmable

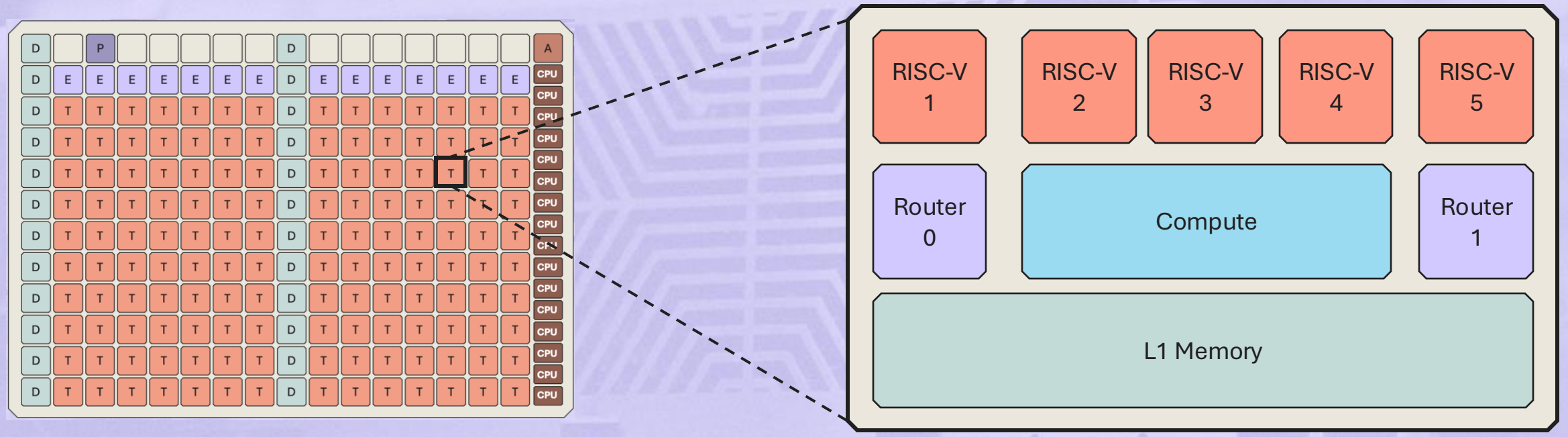
All RISC-V Programmable Baby RISC-Vs

Feature	Spec
Total Baby RISC-Vs	752
Compute	32-bit Int multiplier / divider Floating point (FP32 / BFLOAT16) 128-bit vector (1 per Tensix)
I-cache	4 KB
D-scratch	8 KB



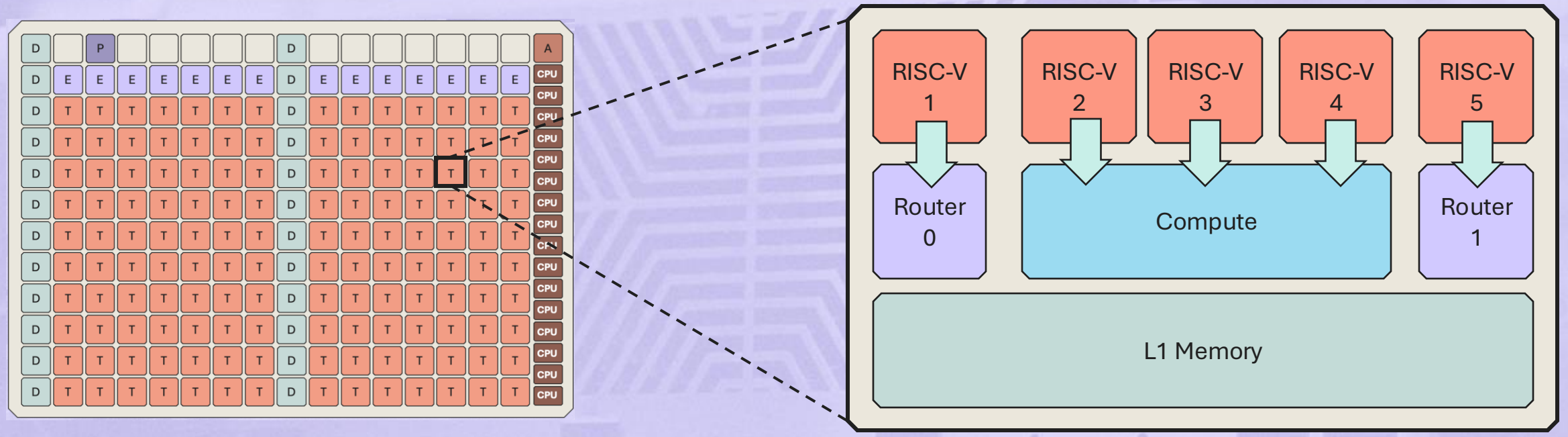
All RISC-V Programmable
Within the Tensix Core

- 5 baby RISC-Vs
- 32-bit RISC-V ISA



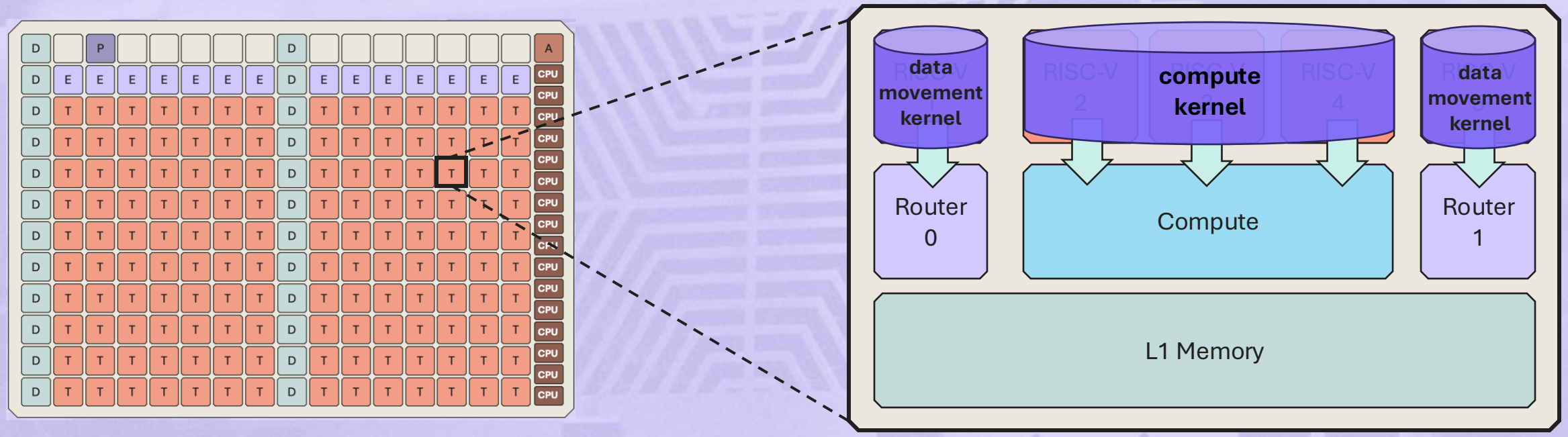
All RISC-V Programmable
Within the Tensix Core

- 5 baby RISC-Vs
- 32-bit RISC-V ISA



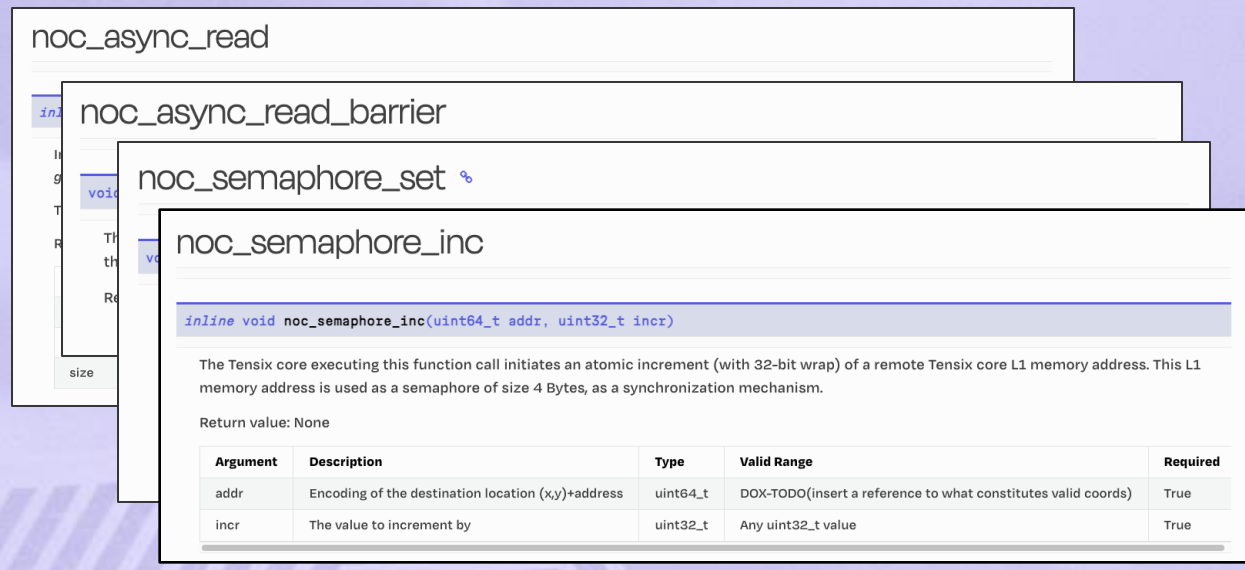
All RISC-V Programmable
Within the Tensix Core

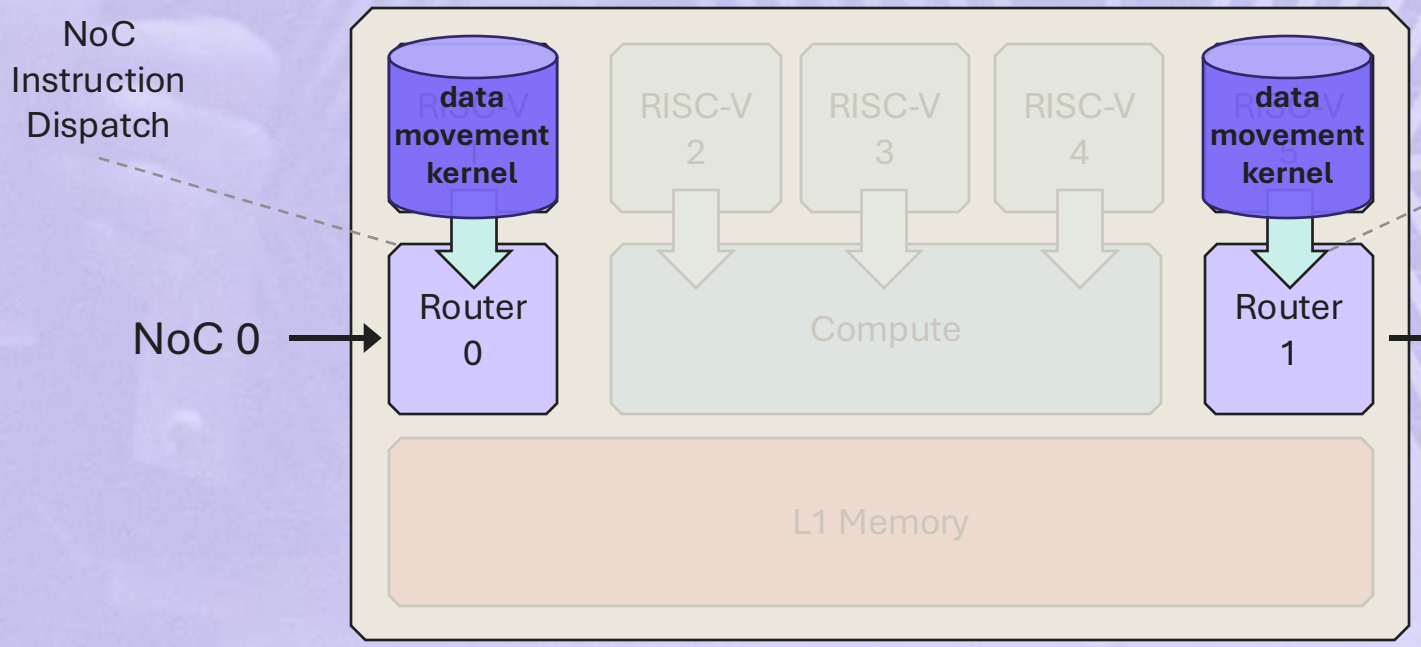
- 3 user C kernels program a single Tensix core
  - 1 compute kernel
  - 2 data movement kernels



Tensix Core – Data Movement

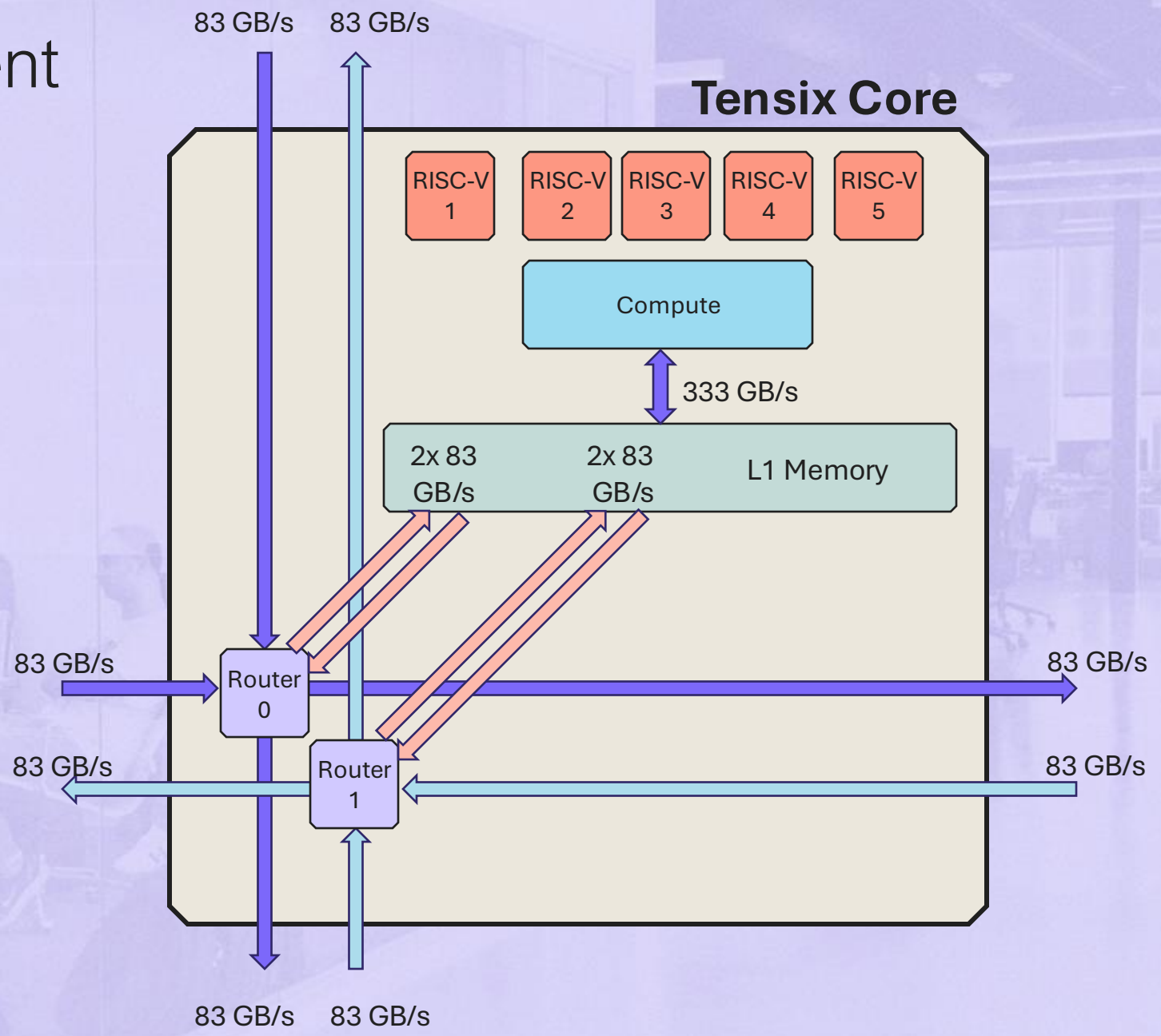
- 2 data movement kernels
- Asynchronous reads & writes
- Access to all SRAM & DRAM banks
- Memory barriers
- Atomic semaphores





Tensix Core – Data Movement

Feature	Spec
Independent NoCs	2
NoC type	2-dimensional torus
NoC link width	64 Bytes
NoC link BW	83 GB/s
Tensix -> NoC I/O BW	665 GB/s
SRAM <-> NoCs	333 GB/s
SRAM <-> NoC aggregate BW	47 TB/s



Blackhole: Built for AI Data Movement Patterns

- Data patterns in MatMuls, Convolutions, and Sharded Data Layouts are regular.
- They have a great mapping to Mesh Architecture

Memory & I/O	Data Movement Pattern	Bandwidth
SRAM	Local / Sharded	94 TB/s
SRAM	Neighbor (Halo)	47 TB/s
SRAM	Row / Column / Mesh Multicast	24 TB/s
SRAM	Gather / Scatter (3 hops)	16 TB/s
SRAM	Gather / Scatter (10 hops)	5 TB/s
DRAM	Row	512 GB/s
Ethernet	Column	1 TB/s