The Future of RiscV and RiscV AI

tenstorrent

Agenda

Opensource

RiscV CPU

RiscV AI

Software

Hardware

Opportunities

More people are designing RiscV than any architecture

- RiscV vs ARM vs Intel – new designs
  - 20+ > 5 > 2

- You can change it
- You can do what you want

- This is how Innovation happens

Ecosystem diversity in RISC-V

tenstorrent
Imagination
SiFive
ANDES TECHNOLOGY
Codasip
MIPS TECHNOLOGIES
VENTANA MICRO
Rivos
cortus
semidynamics
CONDOR
T-HEAD
微核芯
vcore
StarFive
赛昉科技
KRAKATOA TECHNOLOGIES

..and many more

Fundamental Technology - RISC-V

Open

PARTNER owns it: Open-Source, no-centralized control or roadblocks, no-IP limiting issues

Accelerated

PARTNER can adapt quickly: Modern architecture with strong industry support + OPEN and FLEXIBLE

Flexible

PARTNER can change it: Add/Remove functionality or change architecture to better suit your target

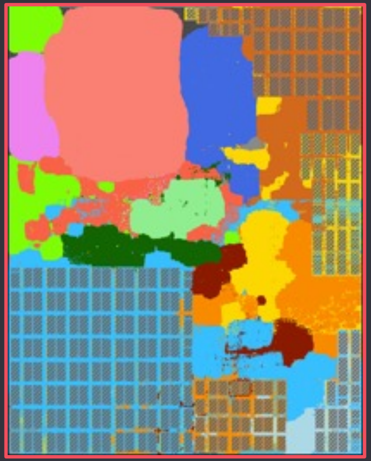
Future Proof

AI will generate future software and hardware – Need an architecture that can change to adapt

Tenstorrent’s mission is to bring high performance RiscV to general purpose computing and AI

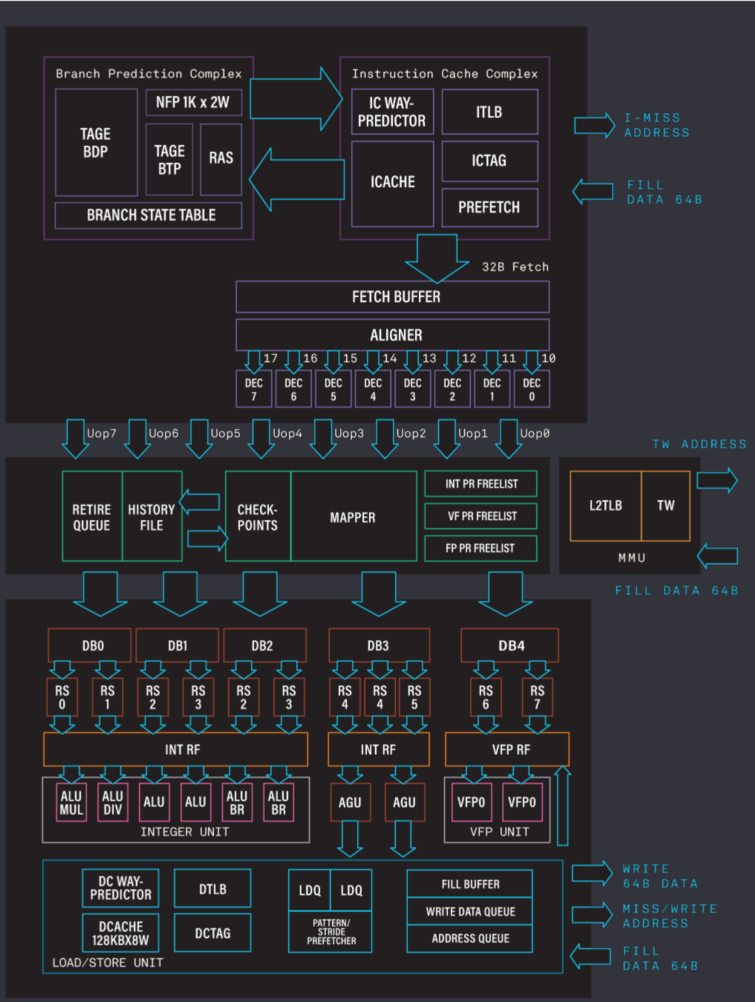
Ascalon O-o-O Superscalar Processor

Disruptive high-performance RISC-V processor for AI and server



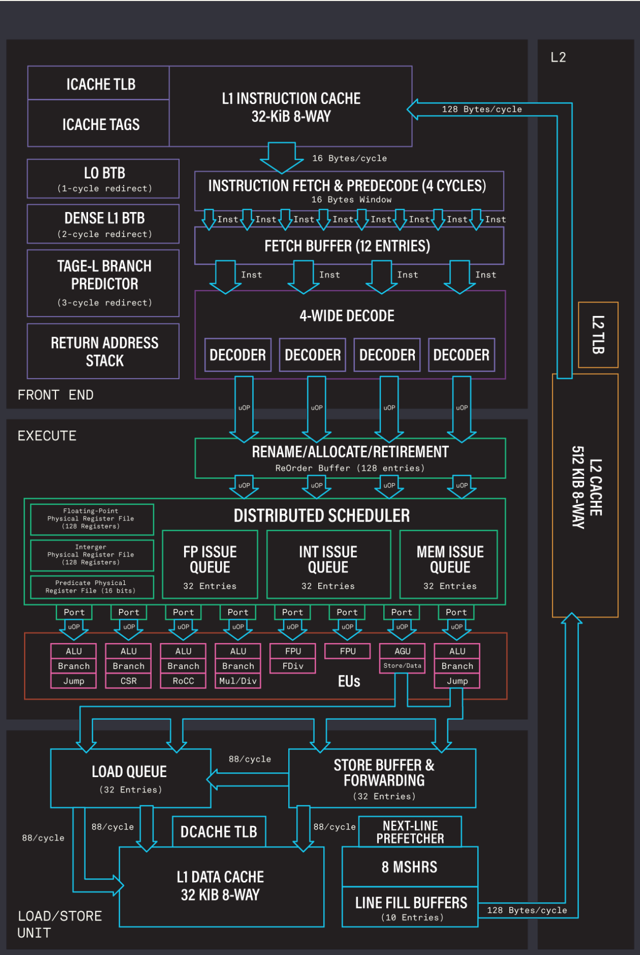
RVA-23

- Advanced branch predictions
- Up to 8-wide decode
- 3 LD/ST with large load/store queues
- 6 ALU/2 BR
- 2 256-bit vector units
- 2 FPU units



tenstorrent Confidential

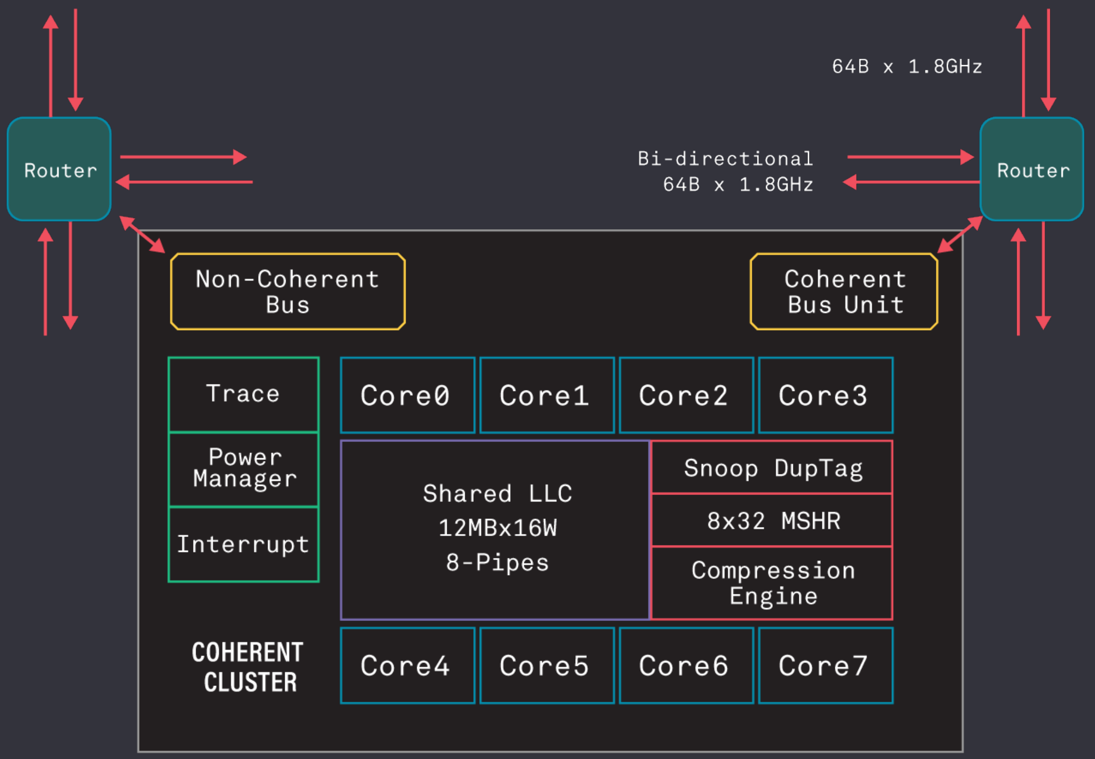
Ascalon-D4 Core Configuration



- Compact and power-efficient
- RV64ACDHFMV

	Features	Size/Width
Frontend	Direction Predictor	45 KB
	Indirect Target Predictor	26 KB
	I-Cache	32KB (8-ways)
	Decode Width	4
Backend	Integer/Branch pipes	2
	Integer pipes	2
	LS pipes	2
	FP pipes	2
	Vector pipes (256-bit)	1
	ROB	160
LSU	LDQ Entries	24
	STQ Entries	36
	DTLB	256 (4-ways)
	D-Cache	64KB (4-ways)
	L2 TLB	1K entries

Ascalon Clusters



Router

Router

64B x 1.8GHz

Bi-directional
64B x 1.8GHz

Non-Coherent Bus

Coherent Bus Unit

Trace

Power Manager

Interrupt

Core0

Core1

Core2

Core3

Shared LLC
12MBx16W
8-Pipes

Snoop DupTag

8x32 MSHR

Compression Engine

Core4

Core5

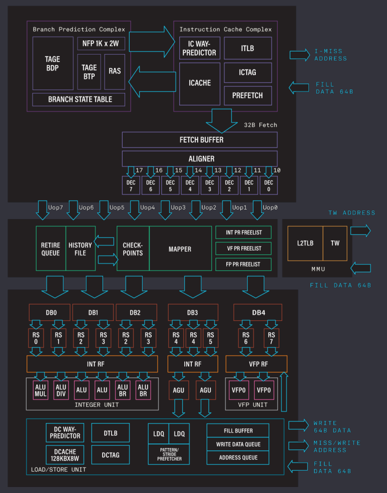
Core6

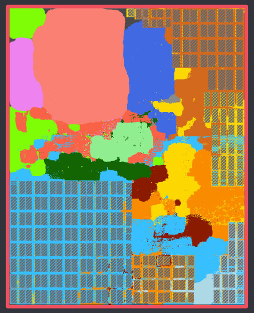
Core7

COHERENT CLUSTER

Cluster Architecture

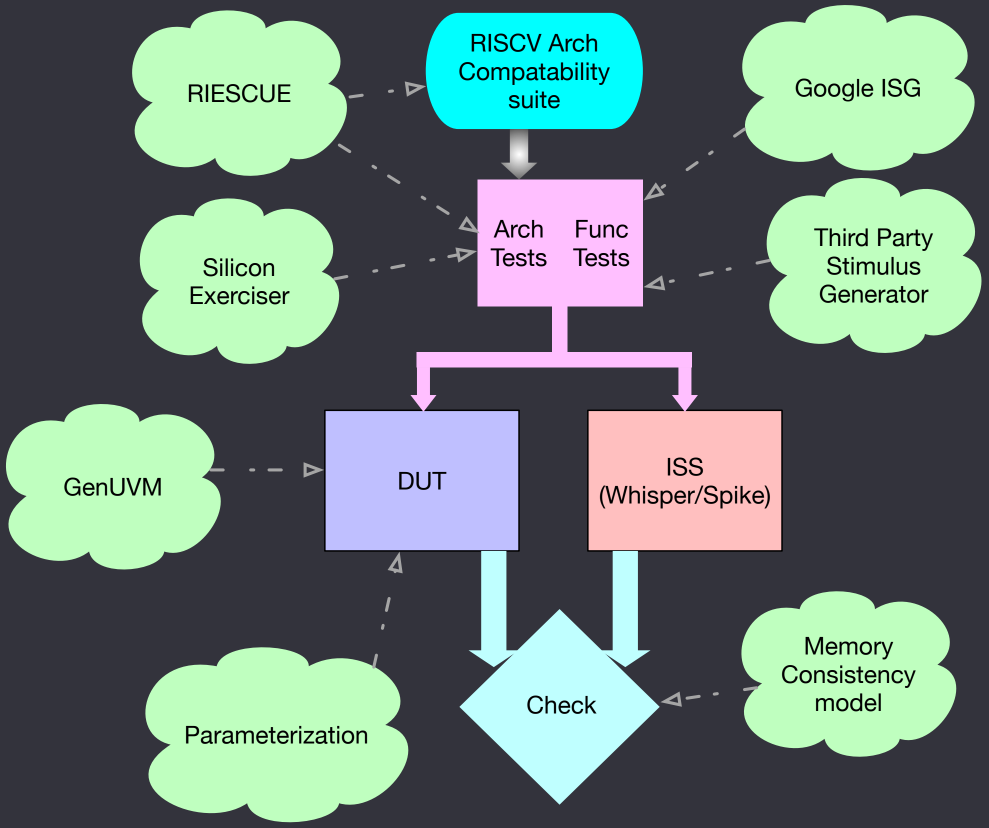
- Up to 8 Cores/per Cluster
- 230GB/S CHI coherency bus
- 230GB/S AXI message passing bus
- 12MB shared cluster cache





RiscV Methodology

- Whisper/Spike: Instruction set simulator
- GenUVM: Infrastructure to automatically generate testbench environments
- RIESCUE: Stimulus and workload development framework to create ISA and microarchitecture tests.
- Parameterization: flows in place to seamlessly re-configure/generate RTL and DV code for different views
- RISCV compliance suite
- Memory consistency model + stimulus suite
- Configurable core/cluster level testbench in-place to check architectural results, compatible with simulation and emulation



RISC-V Software Progress

Upstream projects adding RISC-V support:

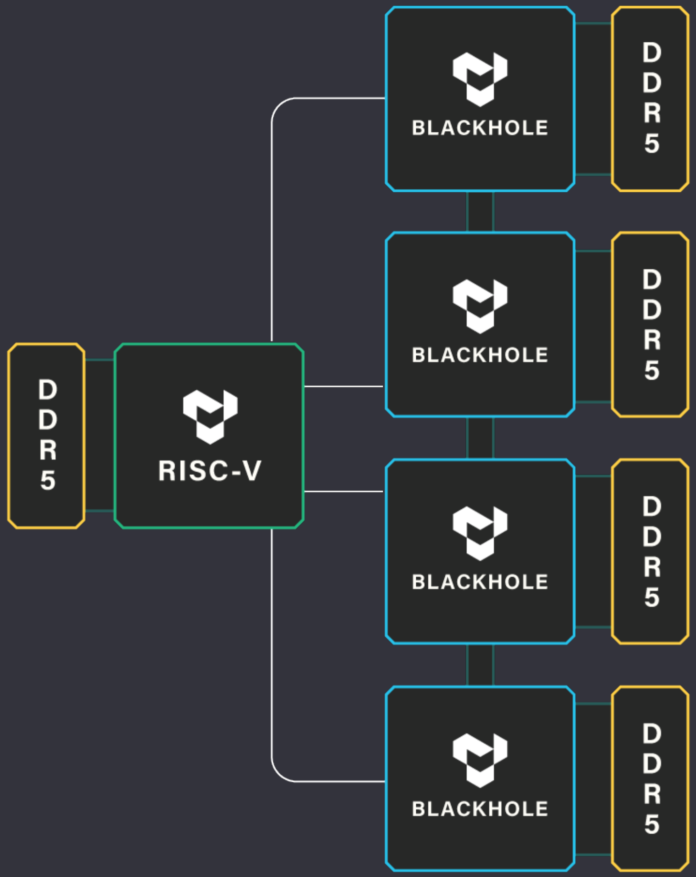
RISE
RISC-V Software Ecosystem

2017: GCC 7.0 includes RISC-V targets
2017: Linux 4.15 includes RISC-V architecture
2018: U-boot v2019.03 adds RISC-V support
2019: LLVM 9.0 supports RISC-V targets
2020: Rust and Go add RISC-V targets
2021: KVM support in upstream repo (Type 2 hypervisor)
2021: Node v.17 supports RISC-V
2022: Chromium V8 (Javascript) support merged
2023: Tianocore EDK2 adds RISC-V support for qemu-virt
2023: Android official port started by Google
2023: Xen platform support actively developed

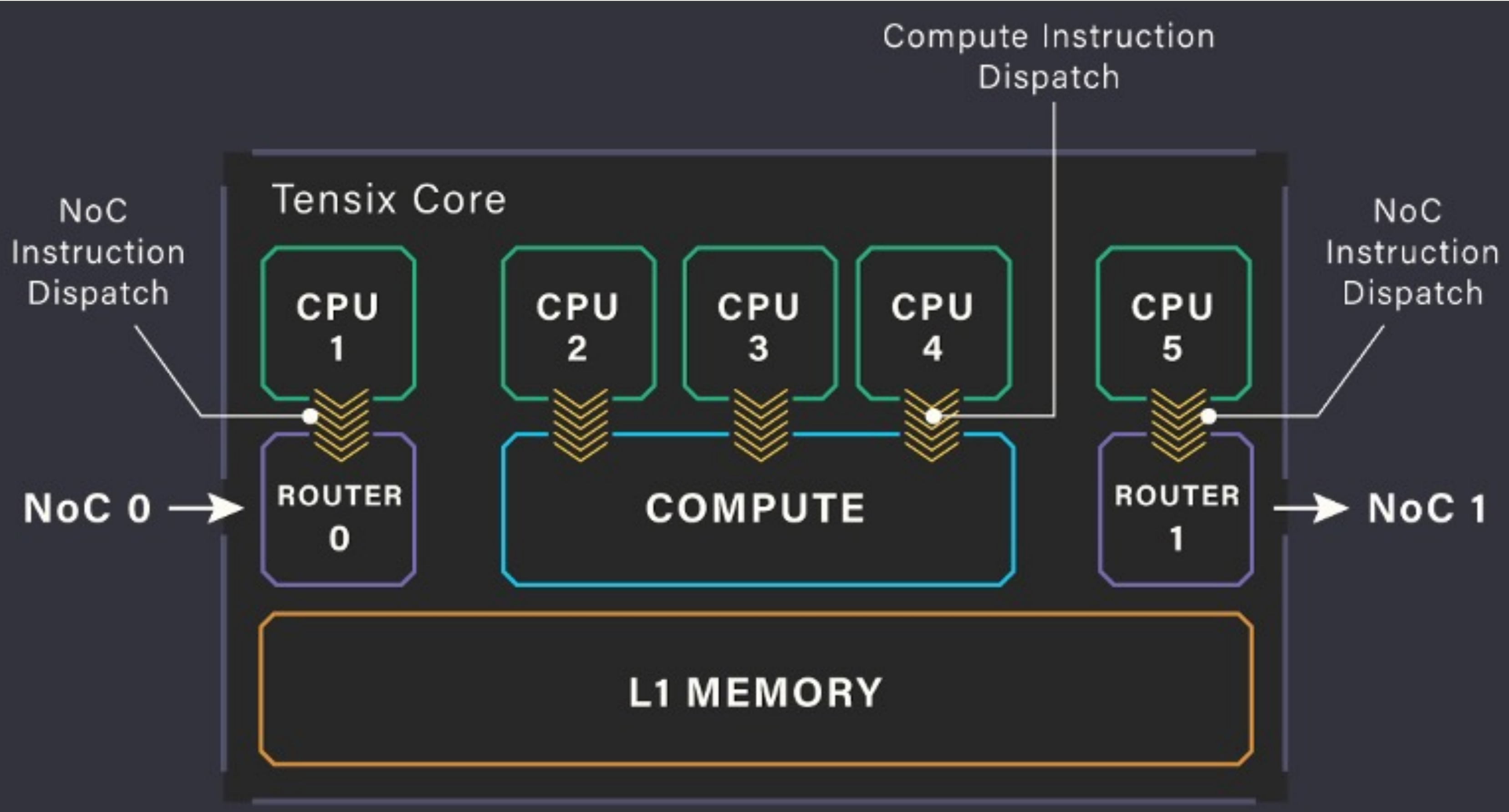
General purpose CPU and AI work closely together

Host CPU

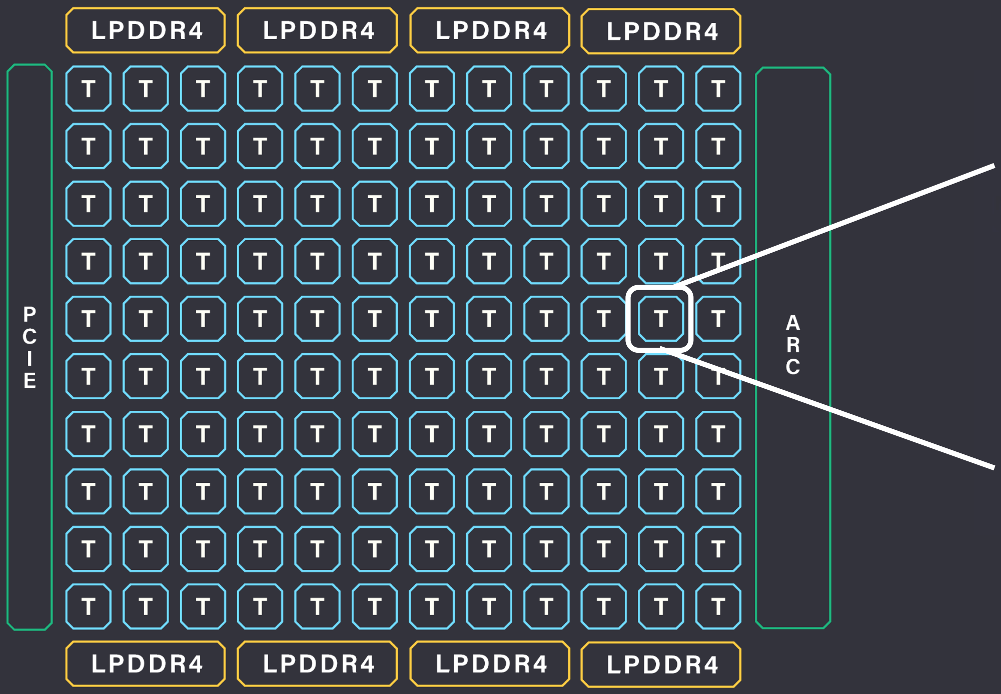
- Top level program
- Computation kernel scheduling/setup
- Virtualization
- Security
- System Management



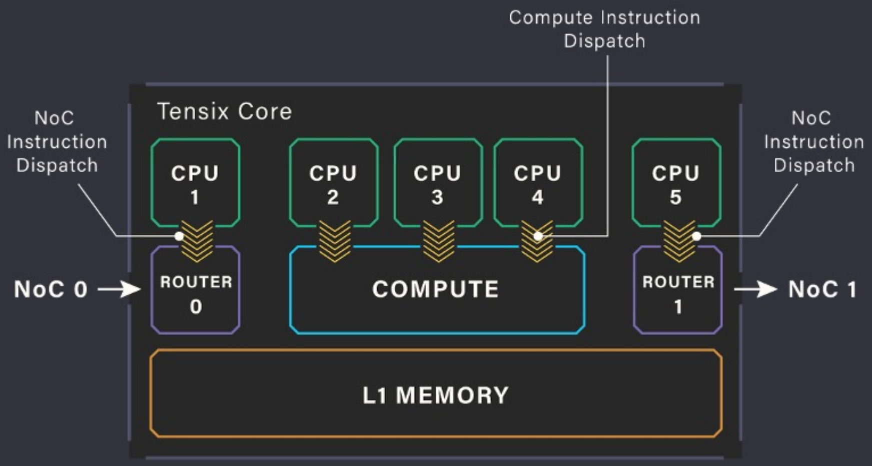
RiscV AI


Compute Instruction Dispatch
Tensix Core
NoC Instruction Dispatch
NoC Instruction Dispatch
CPU 1
CPU 2
CPU 3
CPU 4
CPU 5
NoC 0
ROUTER 0
COMPUTE
ROUTER 1
NoC 1
L1 MEMORY

Scalable Tensix Element

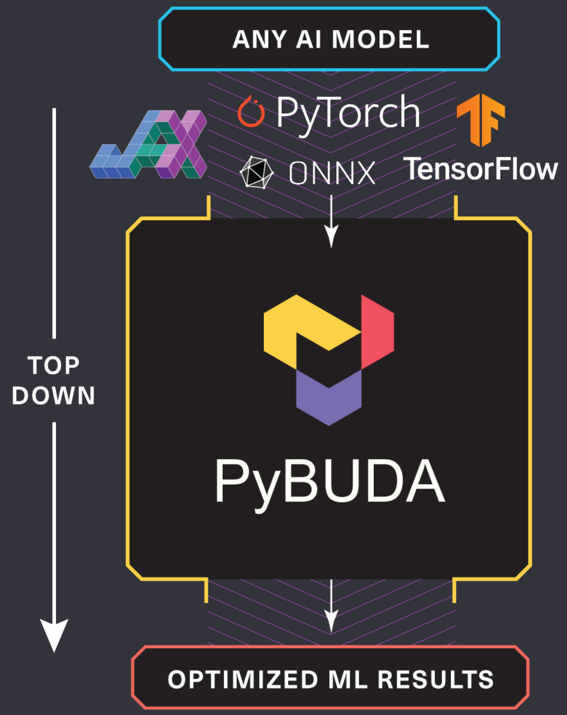


Grayskull: 120 Tensix cores



- Tensix core
- Embedded RISC-V processors
  - 1 Transmit
  - 1 Receive
  - 3 Compute
- Licensable IP elements for scalable AI

Tenstorrent Software – Two Distinct Approaches



ANY AI MODEL

PyTorch

ONNX

TensorFlow

TOP DOWN

PyBUDA

OPTIMIZED ML RESULTS

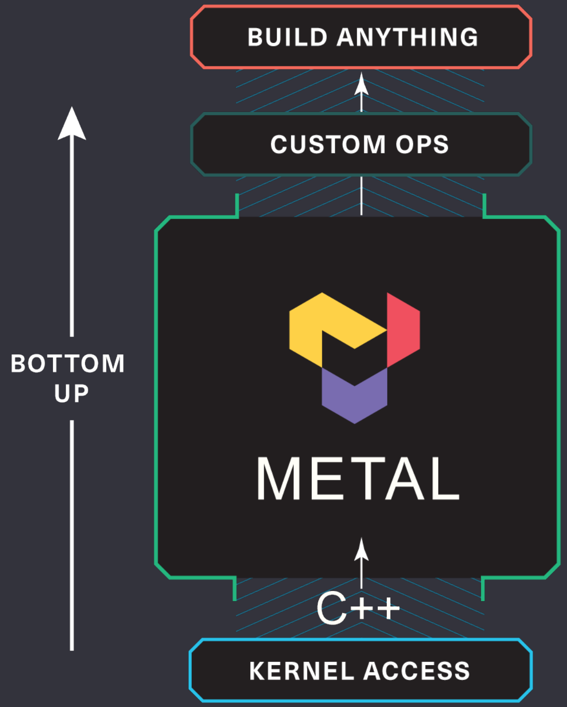
High level language compiler

Compiles source code

Builds graphs

Lowers graph

Automatically allocates resources



BUILD ANYTHING

CUSTOM OPS

BOTTOM UP

METAL

C++

KERNEL ACCESS

Low level coding environment

Code to the metal

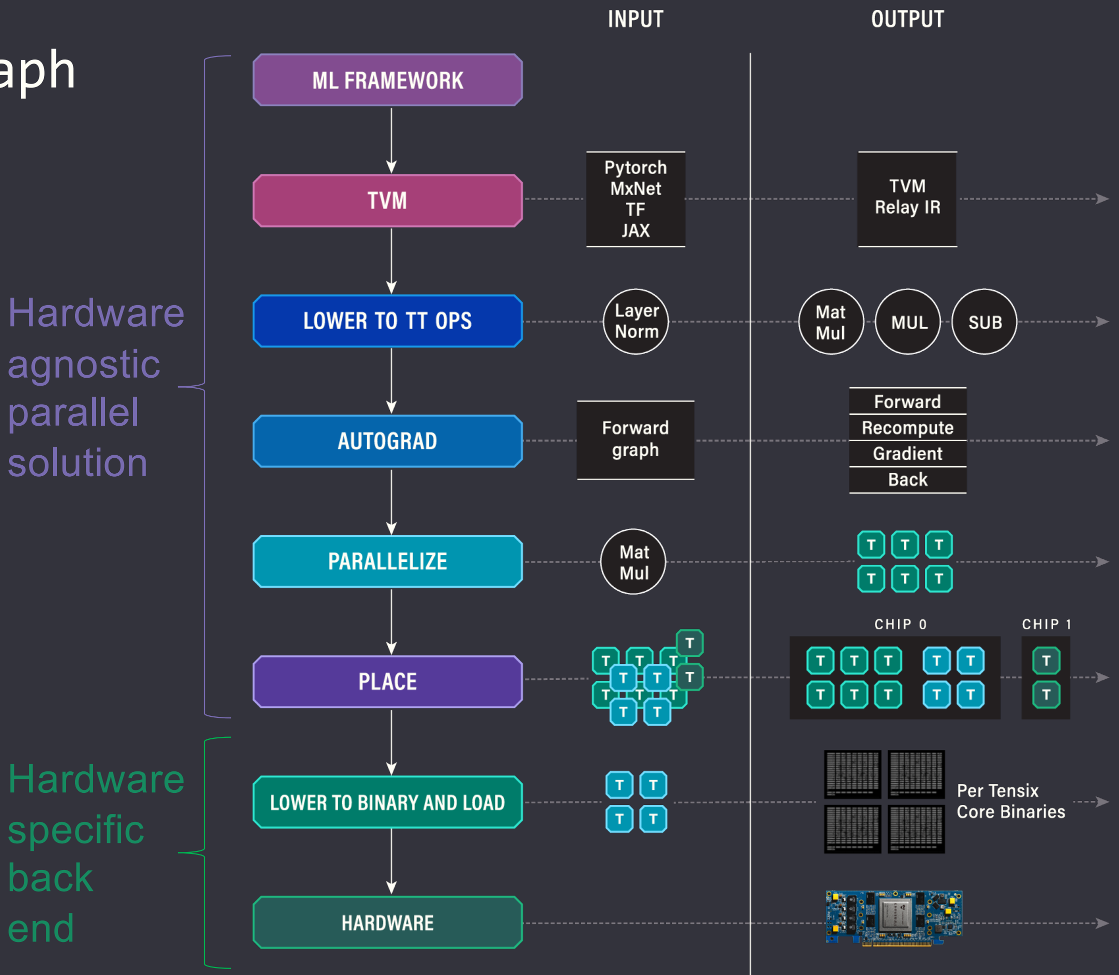
Low level libraries

Direct hardware access

Manually allocate resources

BUDA – Top Down
high level program to graph

- Fully automated path from all popular ML framework to optimized implementation
- High quality results with no manual effort
- Same compiler targets one chip or many thousands of chips



INPUT

OUTPUT

Hardware agnostic parallel solution

ML FRAMEWORK

TVM — Pytorch MxNet TF JAX → TVM Relay IR

LOWER TO TT OPS — Layer Norm → Mat Mul, MUL, SUB

AUTOGRAD — Forward graph → Forward, Recompute, Gradient, Back

PARALLELIZE — Mat Mul → T T T T T T

PLACE — CHIP 0, CHIP 1

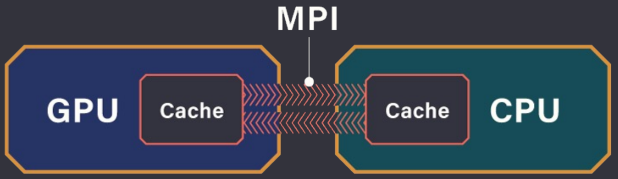
Hardware specific back end

LOWER TO BINARY AND LOAD — T T T T → Per Tensix Core Binaries

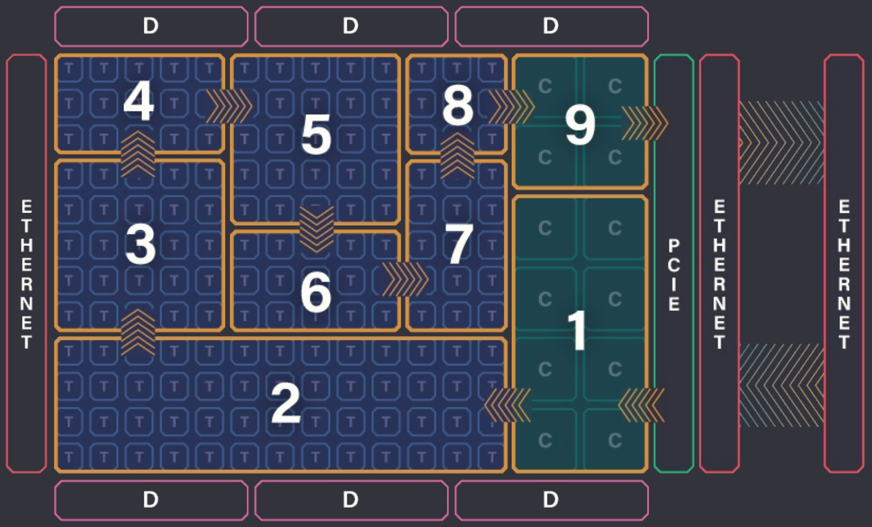
HARDWARE

Map the Graph

- AI computations map to processors
- Optimize data flow and limit use of memory
- Local CPUs manage and participle in the graph
- Opens up new software algorythms

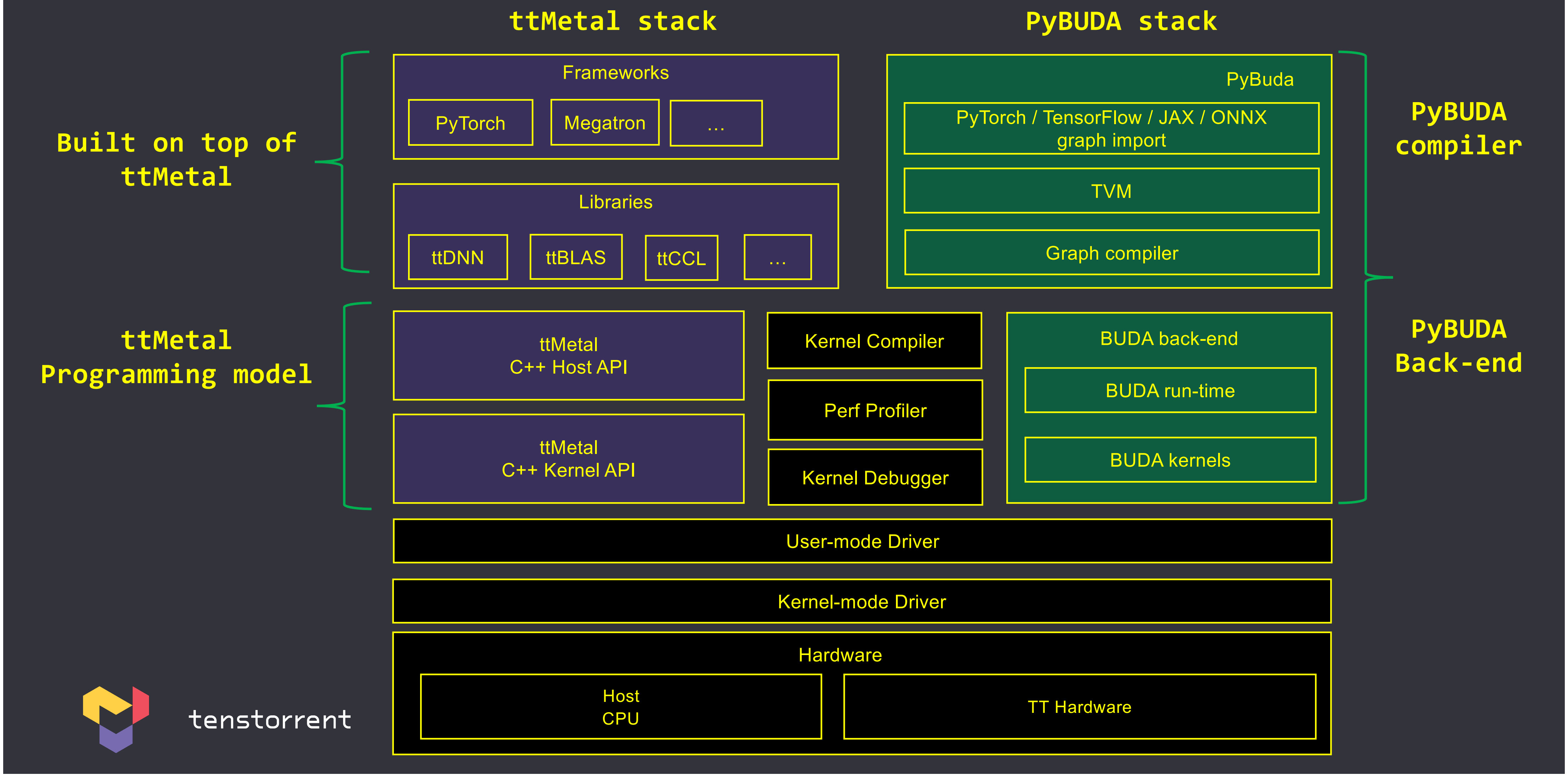


Dataflow Graph Mapping



T	Compute Core	E	Ethernet (400G)
C	RISC-V Cluster	D	DRAM Interface

Software at high-level



ttMetal stack

PyBUDA stack

Built on top of ttMetal

Frameworks

PyTorch

Megatron

…

Libraries

ttDNN

ttBLAS

ttCCL

…

ttMetal Programming model

ttMetal C++ Host API

ttMetal C++ Kernel API

Kernel Compiler

Perf Profiler

Kernel Debugger

PyBuda

PyTorch / TensorFlow / JAX / ONNX graph import

TVM

Graph compiler

PyBUDA compiler

BUDA back-end

BUDA run-time

BUDA kernels

PyBUDA Back-end

User-mode Driver

Kernel-mode Driver

Hardware

Host CPU

TT Hardware

tenstorrent

Hardware Roadmap

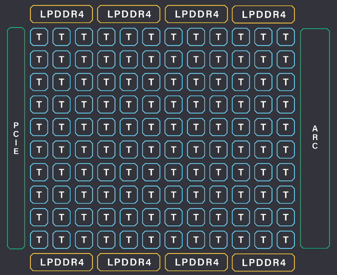
Heterogenous

Chiplet

2021

Grayskull

ML Processor

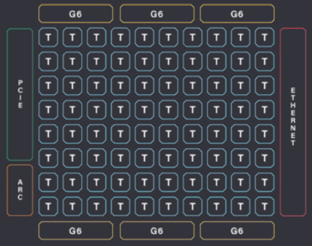


- 12nm, 276 TFLOP (FP8)

2022

Wormhole

Networked ML Processor

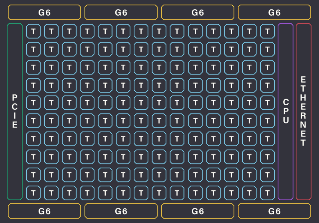


- 12nm, 328 TFLOP (FP8)
- 200 GB/S Scale-out Ethernet

2023

Black Hole

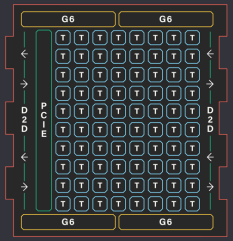
Standalone ML Computer



- 6nm
- SiFive RISC-V X-280
- Heterogenous compute

Quasar

Low Power, Low Cost ML Chiplet

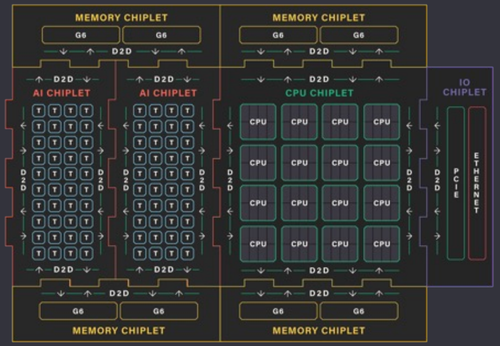


- ML Chiplet

2024

Grendel

Highly Configurable and Performant ML Chiplet



- CPU + ML chiplets

Wormhole Products (2nd Gen device for AI at scale)

Galaxy Card

- Modular device with 1.6TB onboard ethernet
- Natively scalable to an arbitrary number of devices
- High performance at low cost

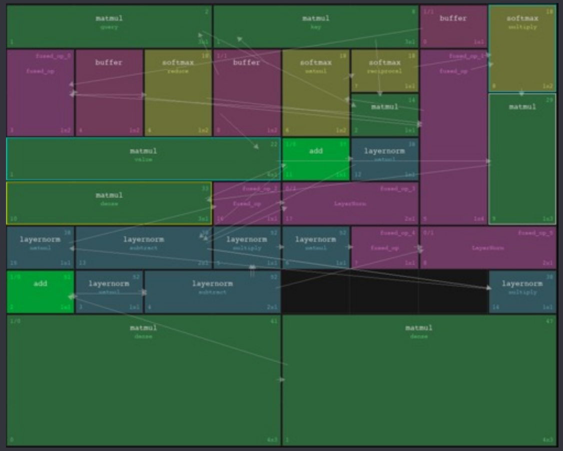
Galaxy Server

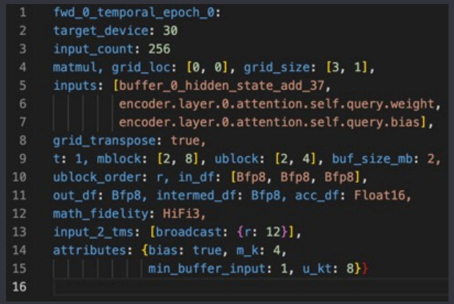
- High-density AI servers in 4U enclosures for rack systems
- Comprised of 32 devices
- Includes backplane interconnect, active cooling units and SDK
- 12 PFLOP at 6KW

BERT-large on Galaxy

- PyBuda in pipeline mode: suitable for LLM Models
- PyBuda places 1 encoder per chip on Galaxy, other placements possible
- Outputs from chip N/layer N flow directly over ethernet to chip N+1/layer N+1

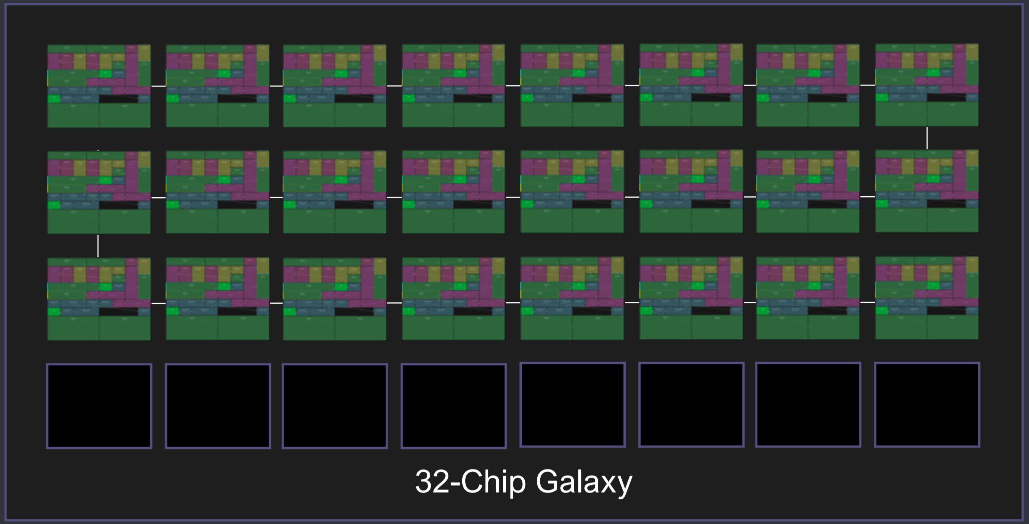
Single BERT-large encoder running on 1 Wormhole Chip





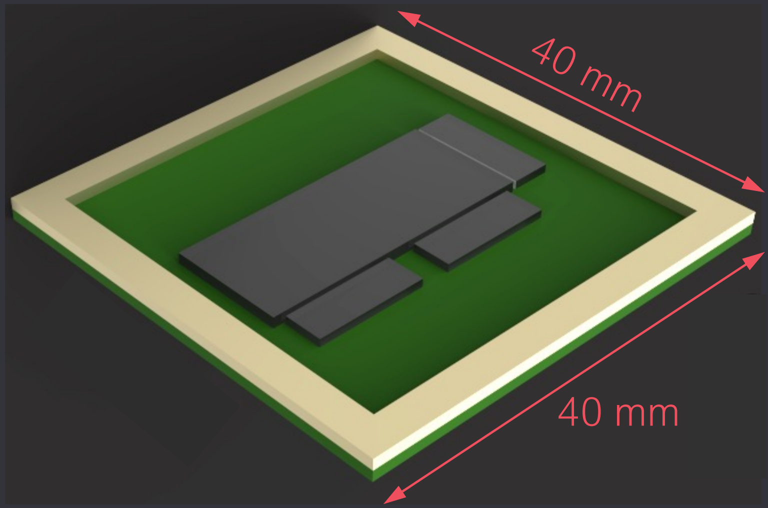
```
fwd_0_temporal_epoch_0:
target_device: 30
input_count: 256
matmul, grid_loc: [0, 0], grid_size: [3, 1],
inputs: [buffer_0_hidden_state_add_37,
         encoder.layer.0.attention.self.query.weight,
         encoder.layer.0.attention.self.query.bias],
grid_transpose: true,
t: 1, mblock: [2, 8], ublock: [2, 4], buf_size_mb: 2,
ublock_order: r, in_df: [Bfp8, Bfp8, Bfp8],
out_df: Bfp8, intermed_df: Bfp8, acc_df: Float16,
math_fidelity: HiFi3,
input_2_tms: [broadcast: {r: 12}],
attributes: {bias: true, m_k: 4,
             min_buffer_input: 1, u_kt: 8}}
```

24 BERT-large encoders running on 24 Wormhole Chips



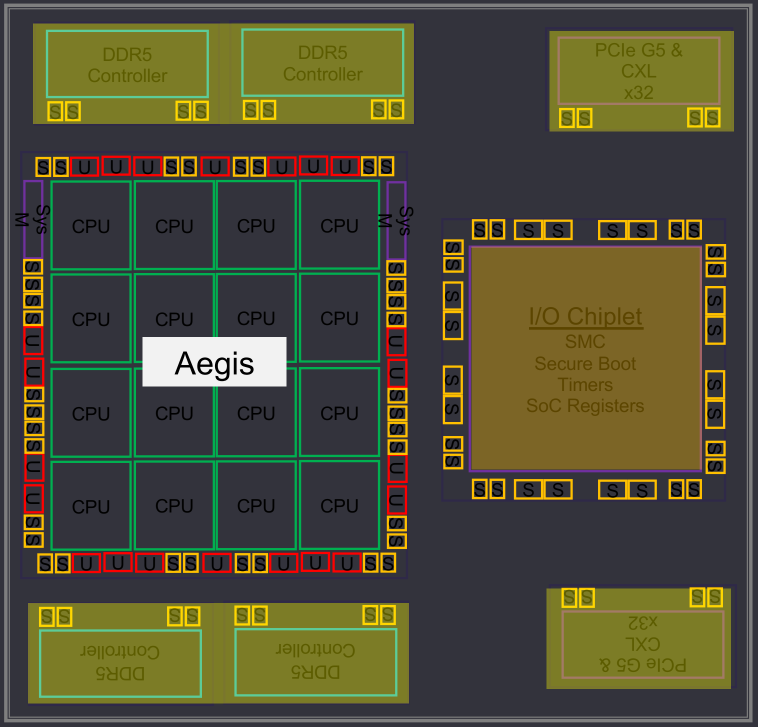
Next Gen Chiplet Solutions

- Organic substrate
- Range of package sizes
- 0.8 mm BGA pitch
- uCIE and BOW D2D phys make organic package work
- Building block approach
- AI, CPU, Memory and IO Chiplets in design



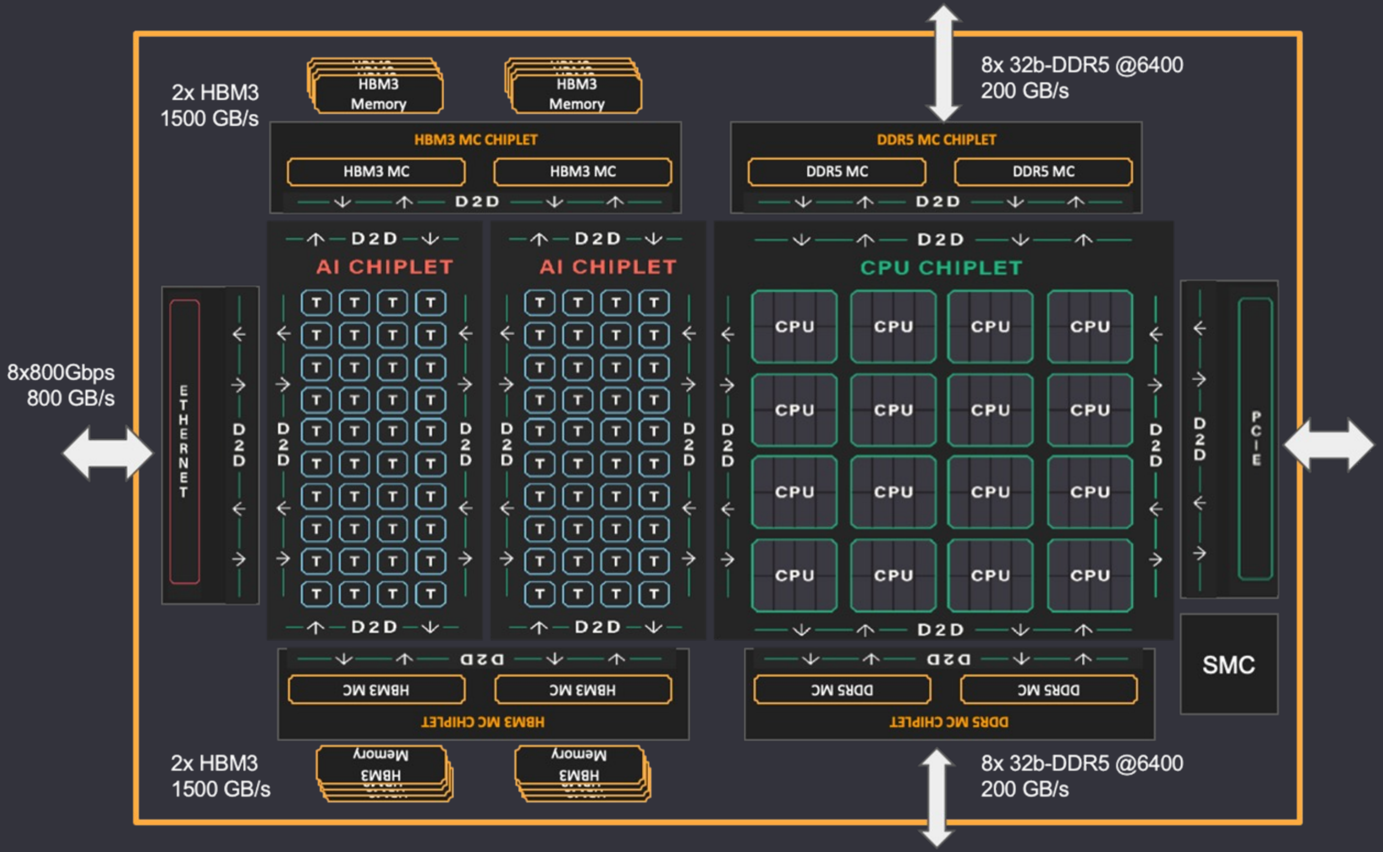
Lidless package is shown for demonstration

Same building blocks - edge RISC-V Server



tenstorrent Confidential

Next gen chiplet plans – RiscV CPU and RiscV AI



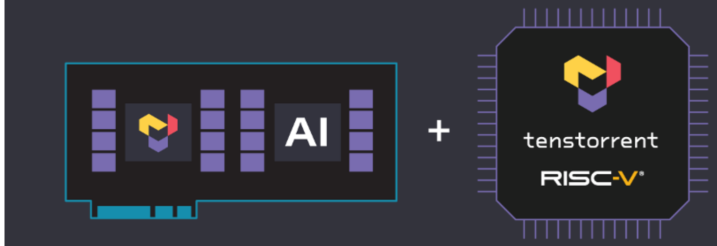
RISC-V enables Oppertunities

RISC-V IP or Chiplet

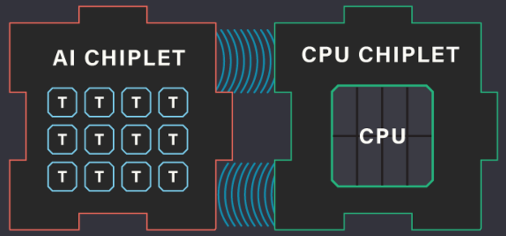
Ecosystem Partnerships

Market specific solutions

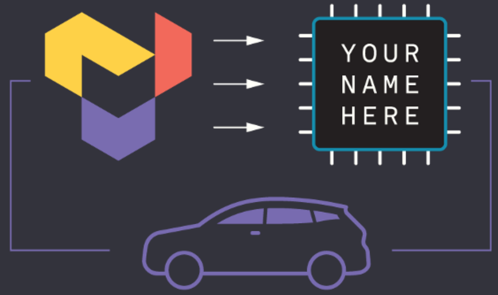
Tenstorrent Example Vertical: Automotive



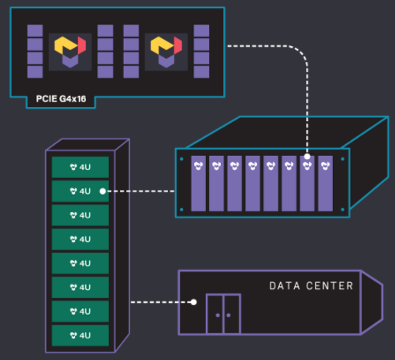
Tenstorrent AI and RISC-V IP deliver the compute power that ADAS and IVI require



Chiplet approach reduces cost while accelerating design and production schedules.



Automotive companies can own their own silicon working with Tenstorrent



Power Consumption is critical: Tenstorrent technology scales from MW to mW

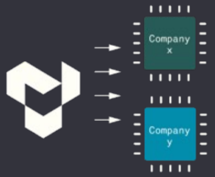
Tenstorrent: Open Business Model

- Tenstorrent works with partners to design, create, modify, optimize heterogenous designs
- Key technology providers for wide spectrum of products for our strategy partners
  - AI
  - CPU

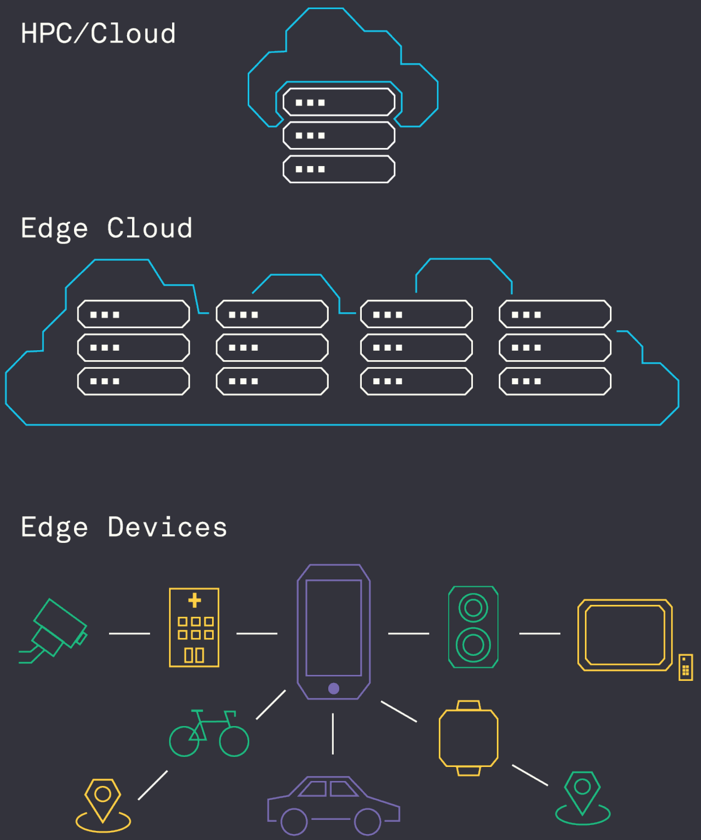
CPU

Chiplet

IP



Whitebox



HPC/Cloud

Edge Cloud

Edge Devices

Summary

Scalable platforms

Scalable AI

Powered by RISC-V