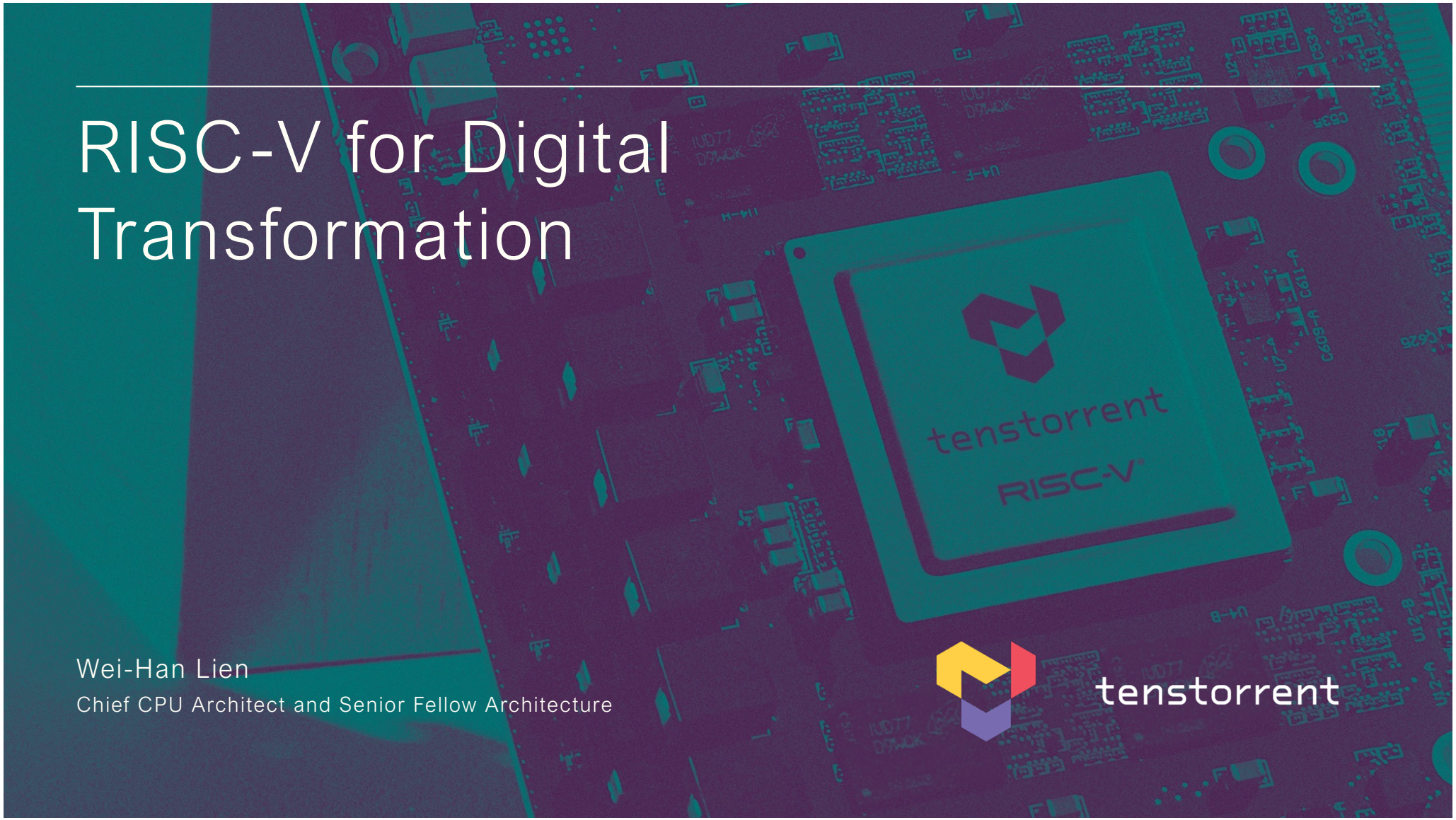
RISC-V for Digital Transformation



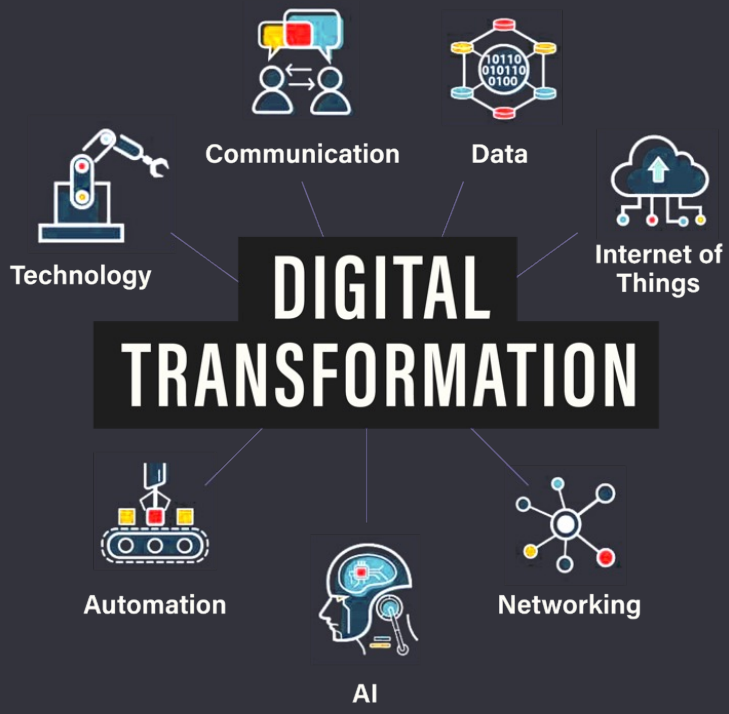
Wei-Han Lien
Chief CPU Architect and Senior Fellow Architecture

tenstorrent

Agenda

- Digital Transformation
- Scalable RISC-V based AI
- Scalable RISC-V processor family
- Chiplets

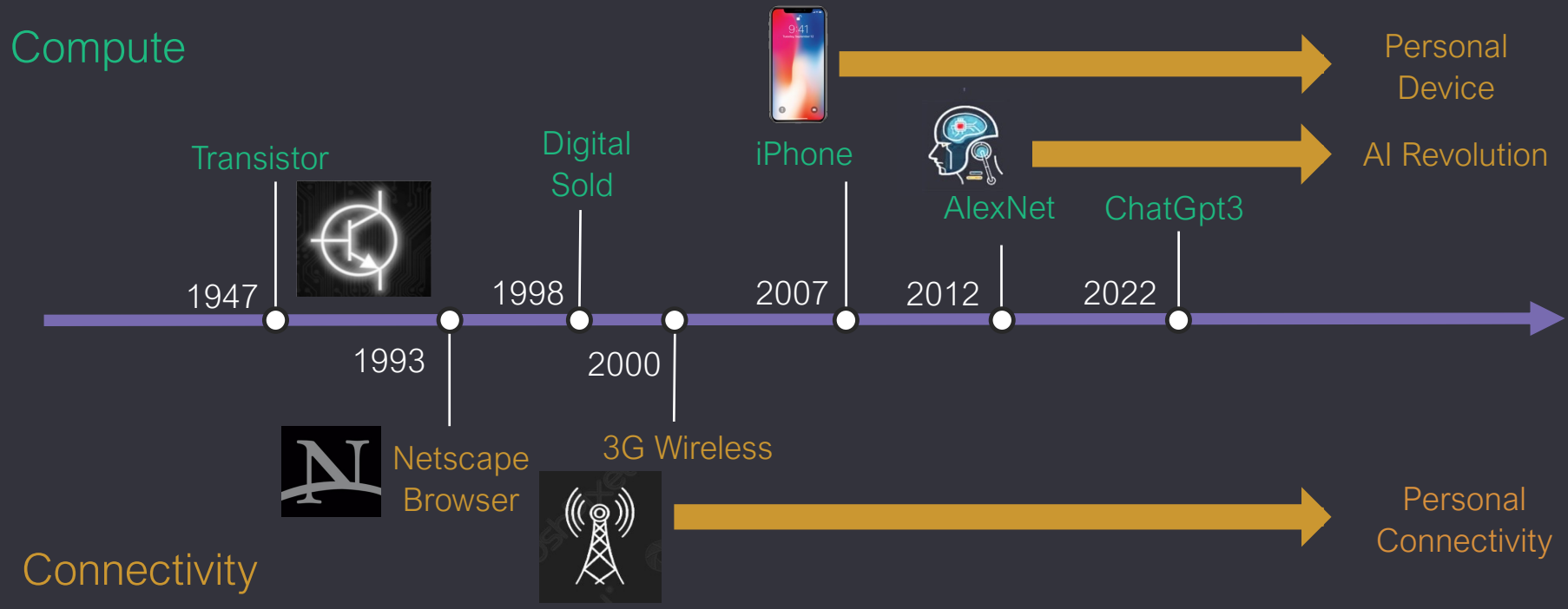
Digital Transformation



Human race is entering digital AI transformation

- AI revolution: Machine intelligence replaces human brain
- Reshape business models, practices, and cultures for competitiveness
- Cloud computing, AI, IoT, and data analytics are revolutionizing the digital landscape
- Real-time data and streamlined processing enables agile decision-making and strategy adjustments
- Digital insights allow personalized experiences and tailored solutions, fostering customer loyalty

Chronical of AI-driven Digital Transformation

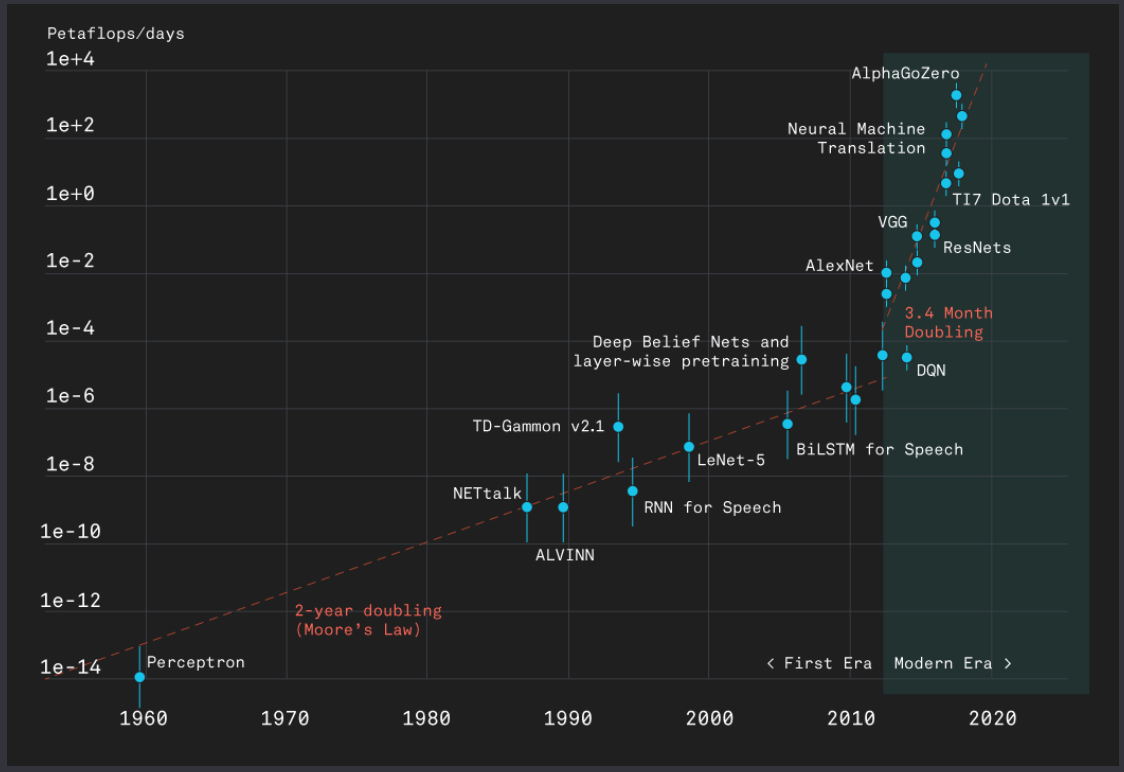


Digital Transformation Compute Everywhere

- ChatGPT4 = 2-trillion parameter
- Data Generation = 2.5 Quintillion Byte/per day
- Both still growing…..
- How about power and cost?

$2x10^{12}$ parameters X
$2.5x10^{18}$ Byte data per day

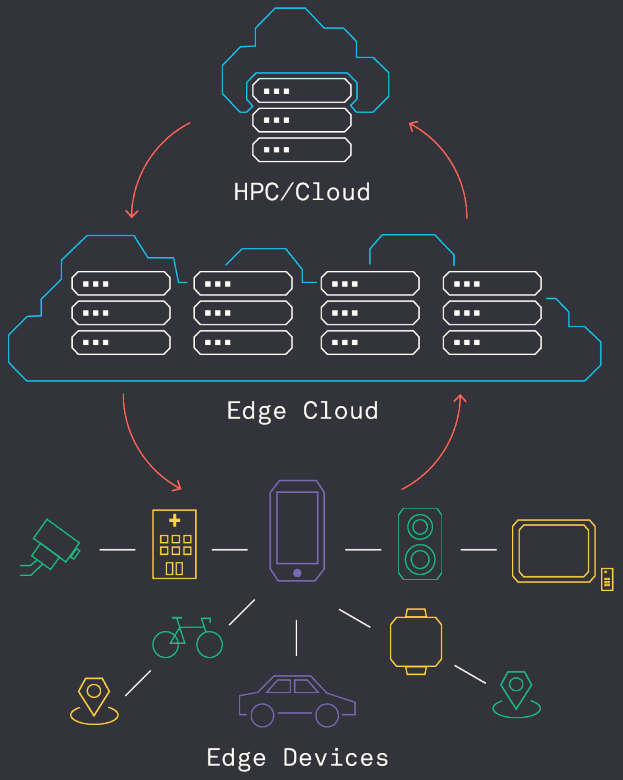
Compute everywhere

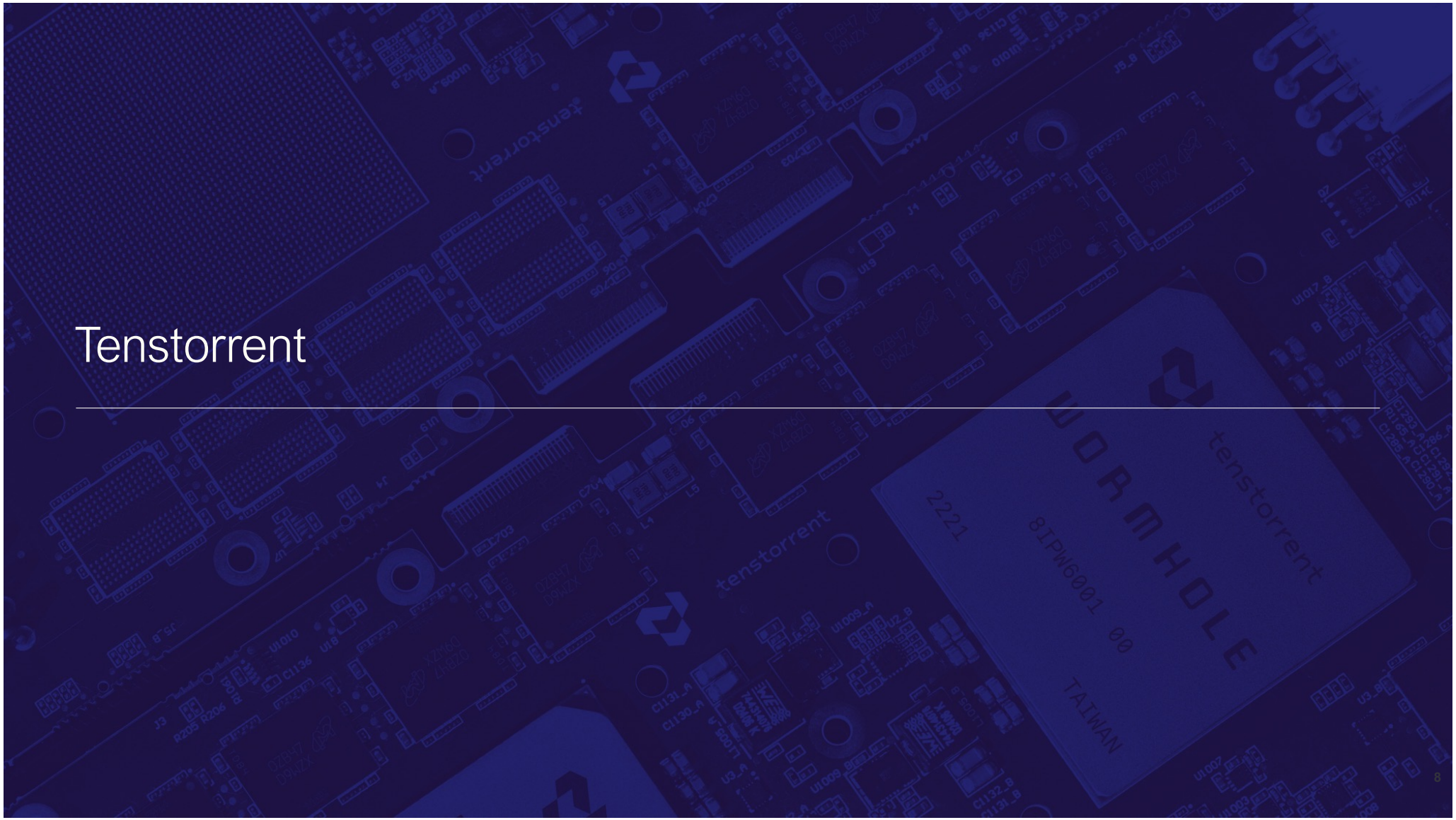


Pervasive Computing for Digital Transformation

- Massive data movement and compute everywhere
  - Devices, edge, and cloud computing
  - Improve compute/communication/energy costs
- Compute requirements
  - Heterogeneity
  - Scalable to meet wide-range PPA requirements
  - Uniform architecture specification reduces complexity
  - *Open-source fosters innovations and specializations*

RISC-V



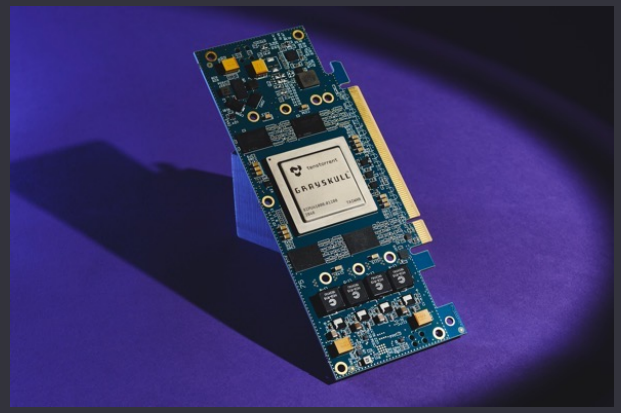

Tenstorrent
WORMHOLE
tenstorrent
2221
8IPW6001 00
TAIWAN

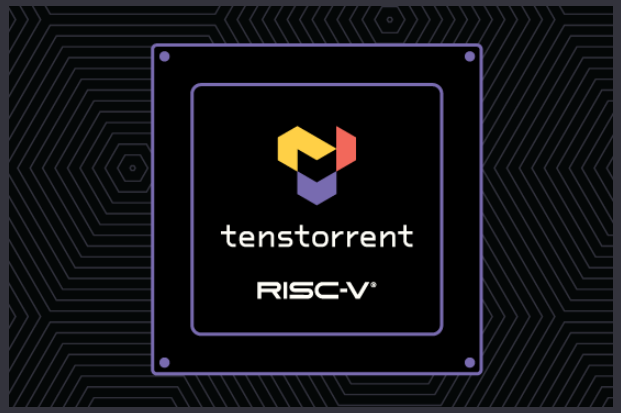
Tenstorrent

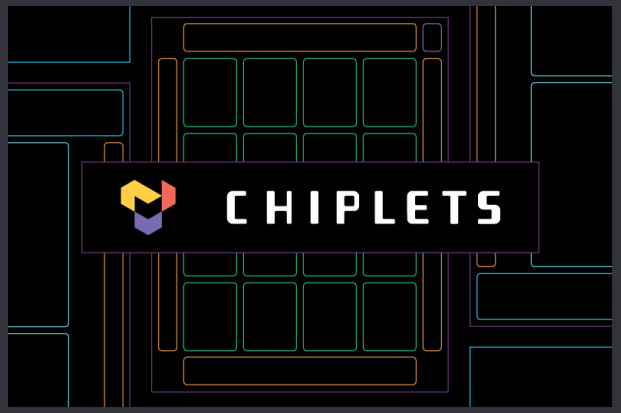
- Founded in 2016 to build the best ML training/inference chips
- $330M raised with 300 employees
- Two ML chips - Grayskull and Wormhole – in production, working on third
- Building a high-performance RISC-V processor
- Only company in the world with high-performance RISC-V and ML processors

Jim Keller

CEO, Digital Alpha processor, Apple A series, AMD Zen, Tesla Autonomous Driving system







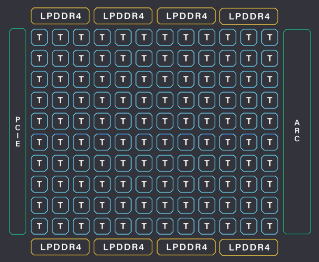
AI Chip Roadmap

Heterogenous

Chiplet

2021 — Grayskull

ML Processor



- 12nm, 276 TFLOP (FP8)

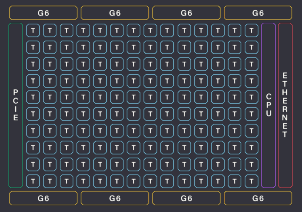
2022 — Wormhole

Networked ML Processor

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

2023 — Black Hole

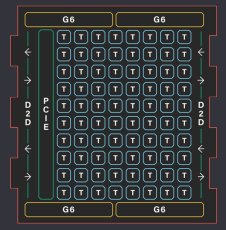
Standalone ML Computer



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

2024 — Quasar

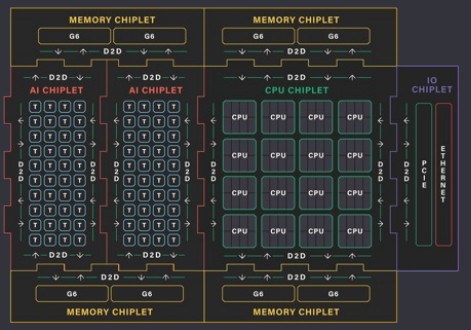
Low Power, Low Cost ML Chiplet



- ML Chiplet

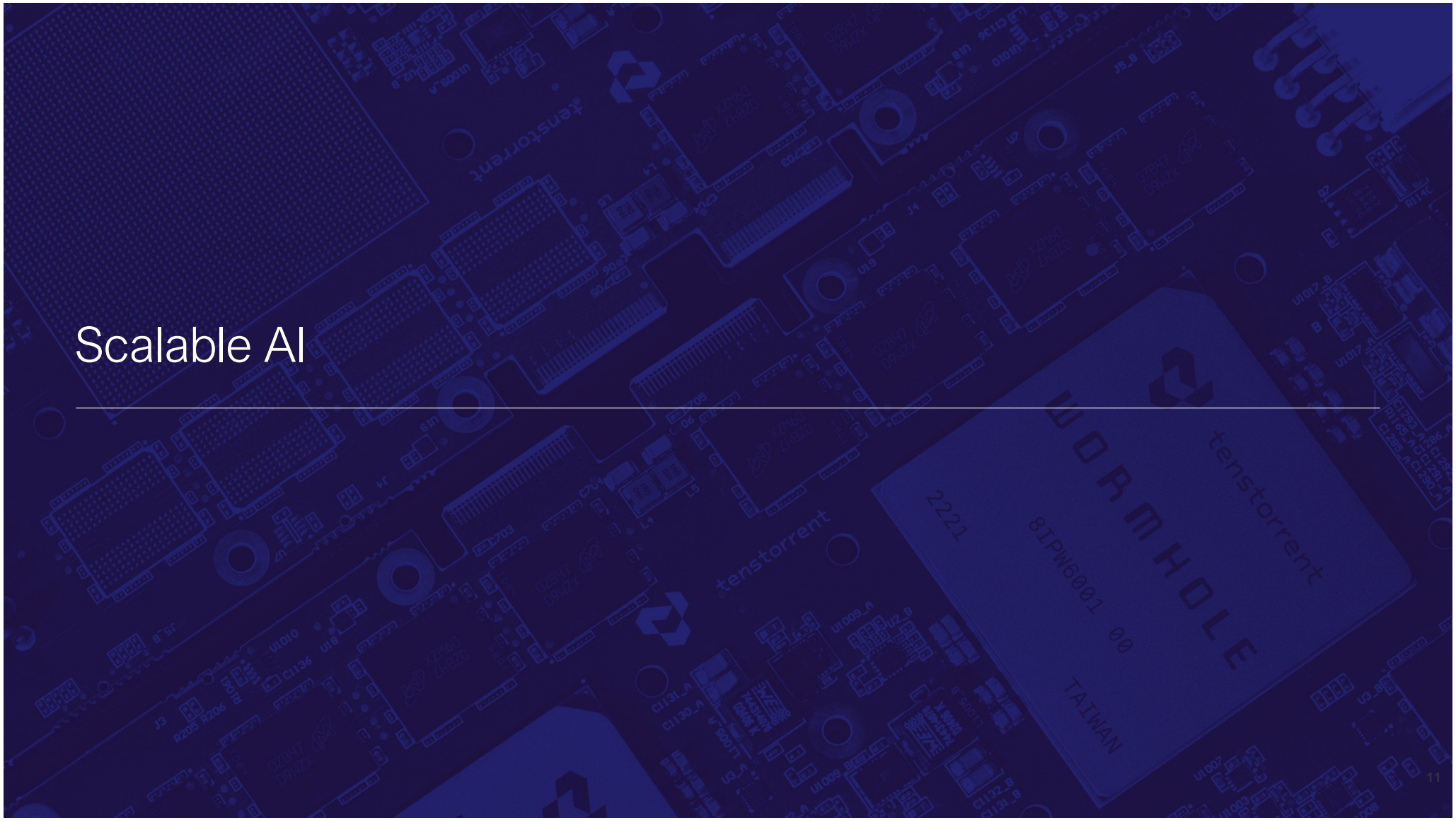
2024 — Grendel

Highly Configurable and Performant ML Chiplet

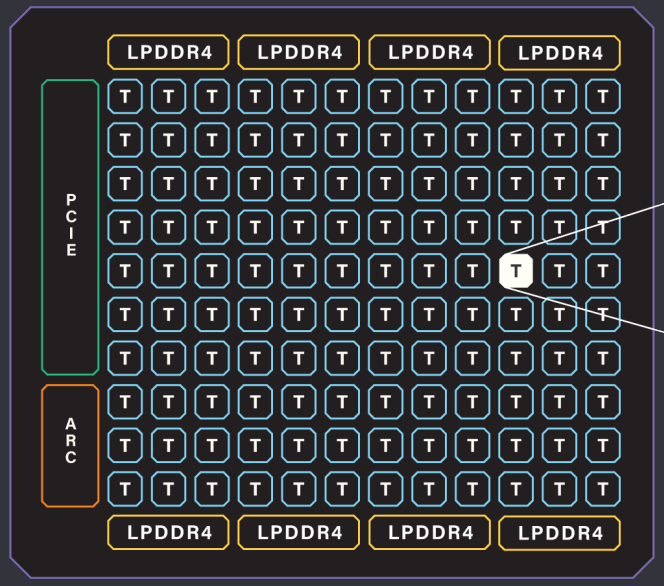


- CPU + ML chiplets

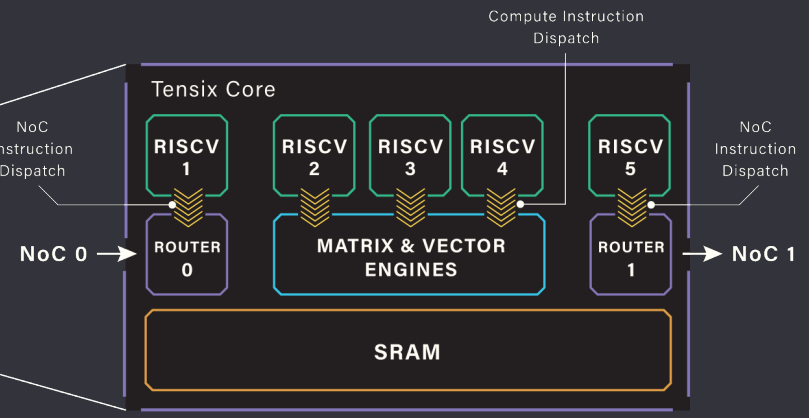
Scalable AI



Scalable Tensix Element



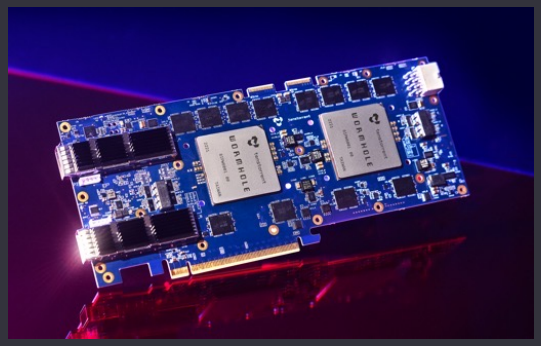
Grayskull: 120 Tensix cores



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

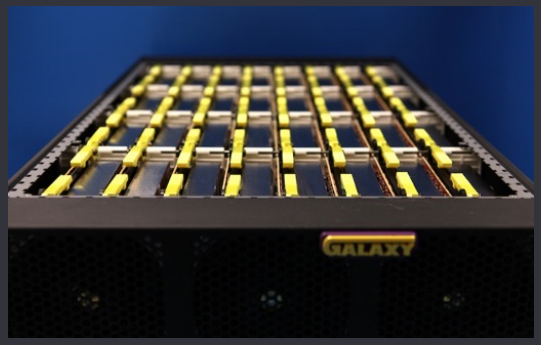
Wormhole Products (2nd Gen device for AI at scale)

12nm AI Accelerator on PCIe Gen 4



N300s/d (Nebula, single or dual chip config available)

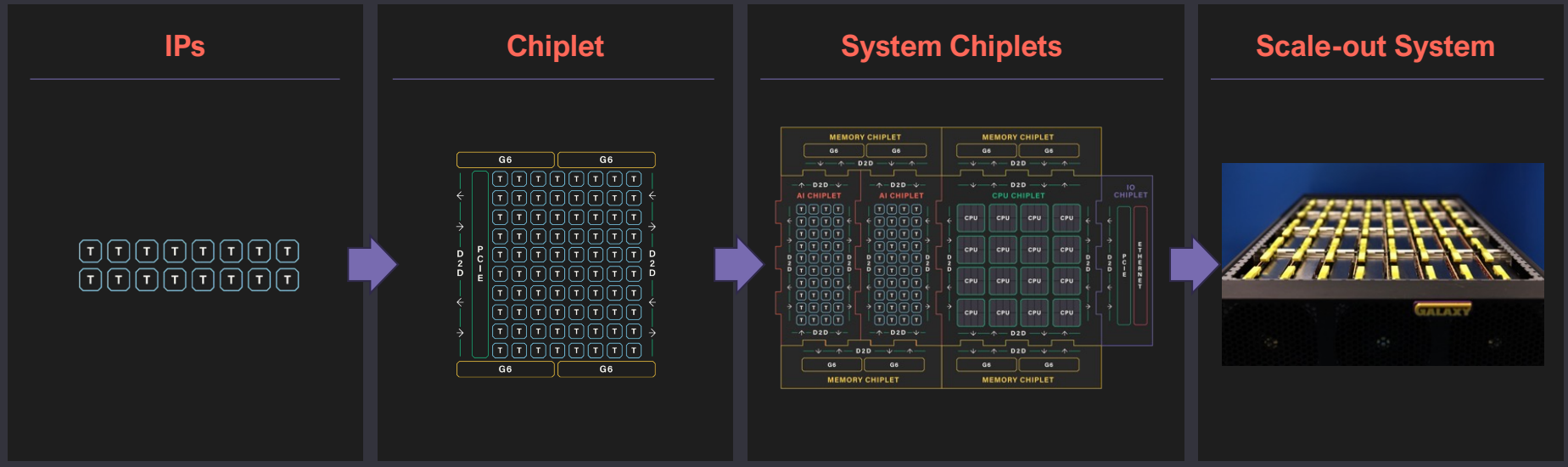
- Modular device with 1.6TB onboard ethernet
- Natively scalable to an arbitrary number of devices
- High performance at low cost



Nebula Server

- Pre-built, high-density AI servers in 4U enclosures for rack systems
- Comprised of 32 x n300s devices
- Includes backplane interconnect, active cooling units and SDK
- 12 PFLOP (BF8) at 6KW

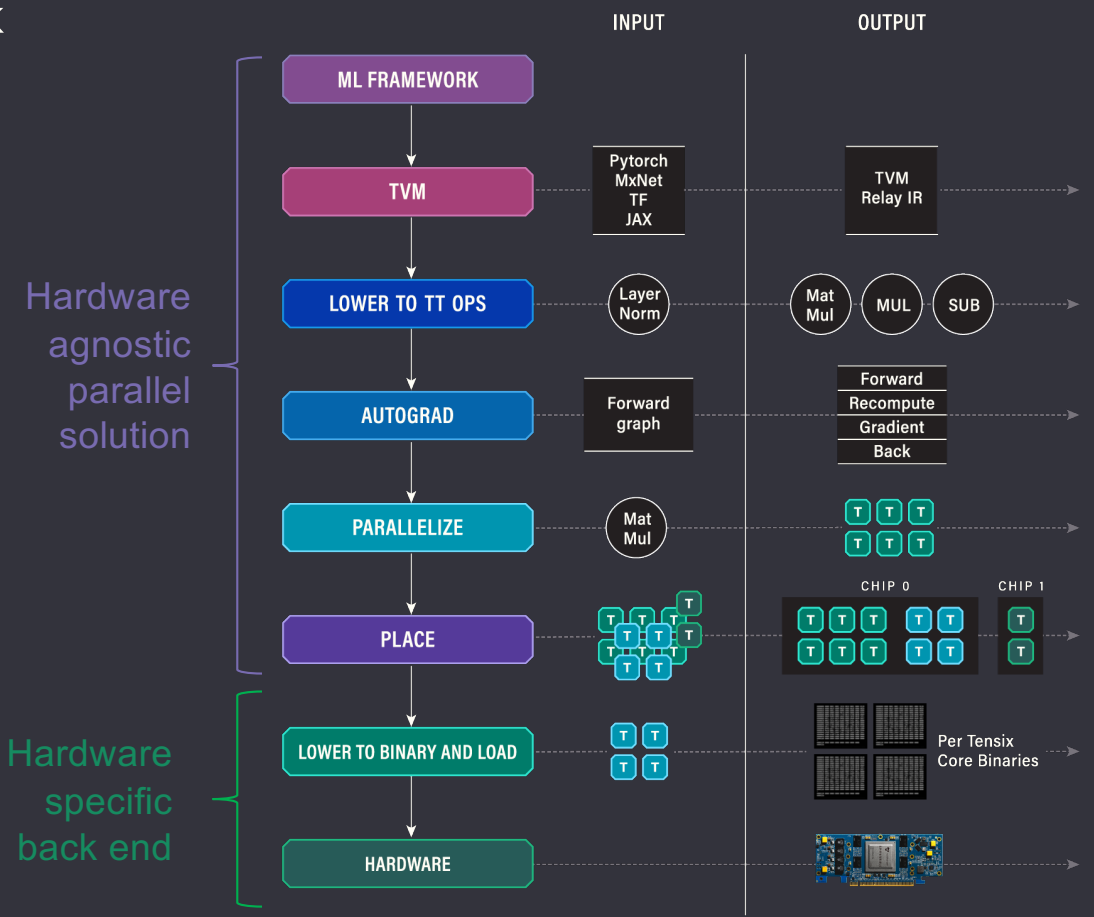
Scalable AI Architecture



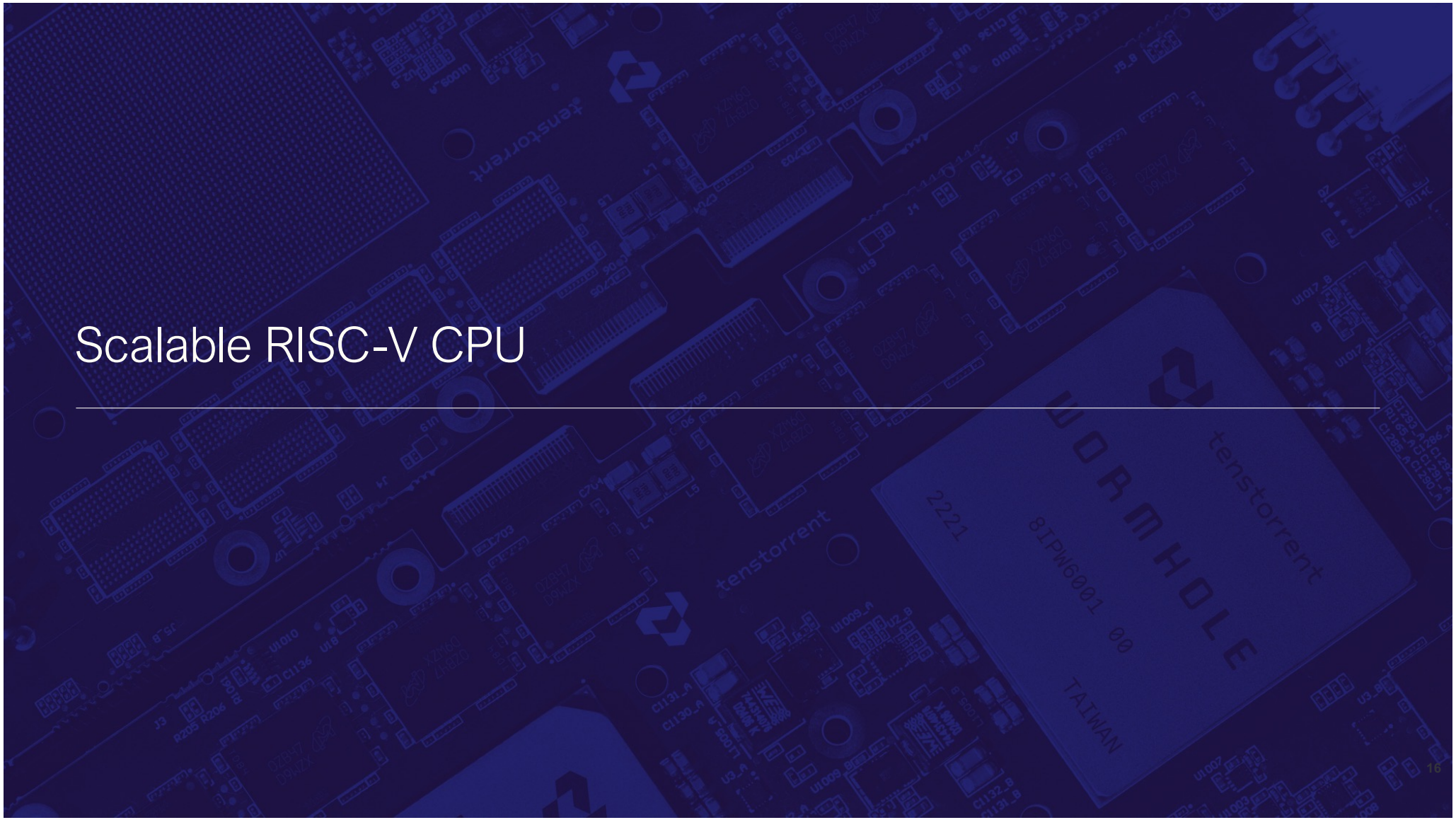
AI scalability from 1 Tensix core to thousands of chips

Scalable Software Stack

- 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



Scalable RISC-V CPU

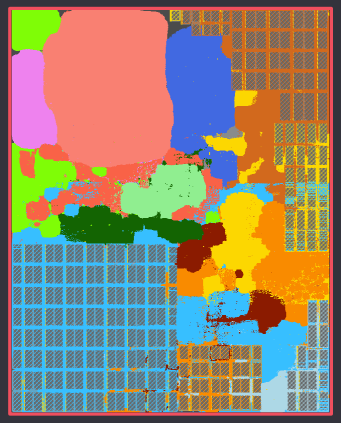


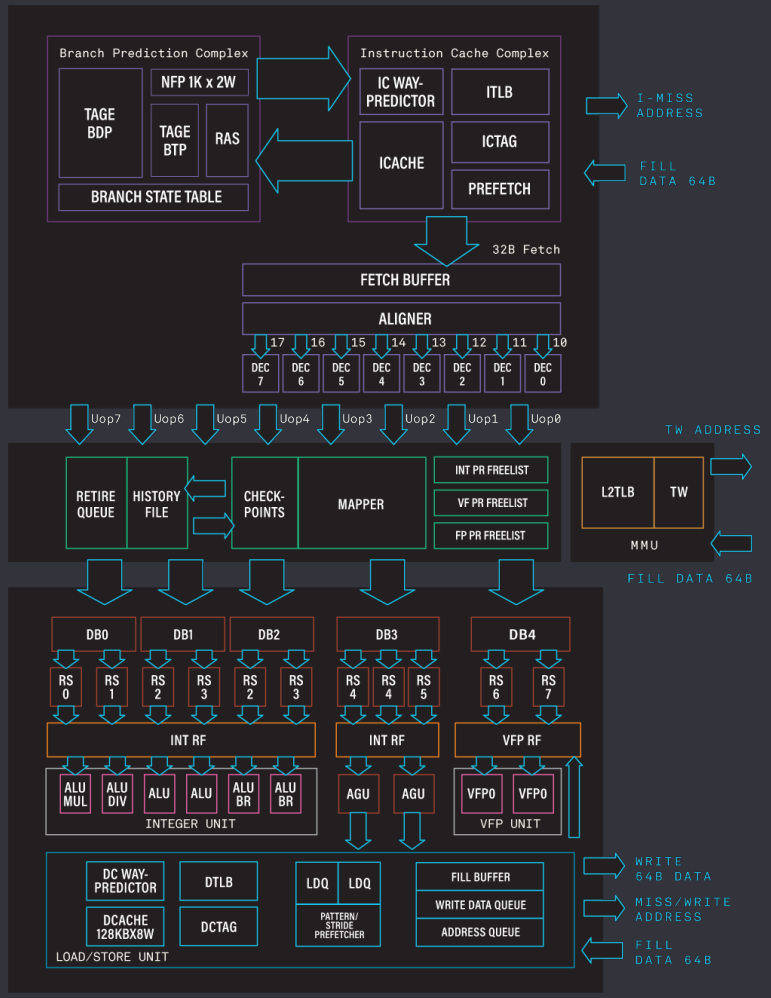
Ascalon O-o-O Superscalar Processor

- Disruptive high-performance RISC-V processor for AI and server
- Projected Zen5 performance in 2024

RVA-23

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





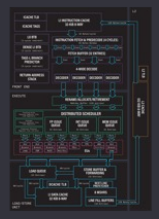
Tenstorrent RISC-V O-o-O Processor Family

Open & Free

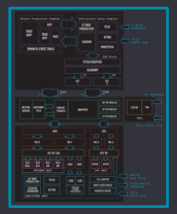
Higher Performance

Performance

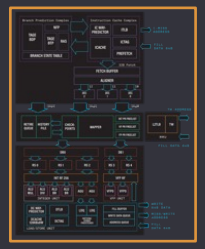
One Design and 5 IPs in a year



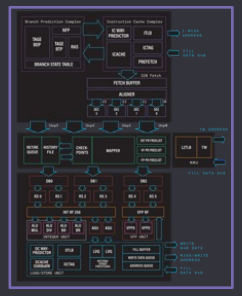
4-Wide Decode
Sonic Boom with Vector



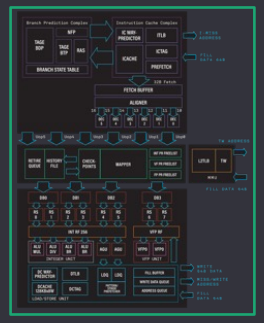
2-Wide Decode



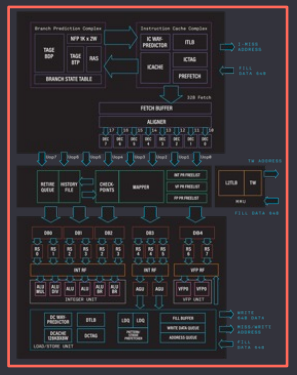
3-Wide Decode



4-Wide Decode



6-Wide Decode
Alastor
Client and Edge

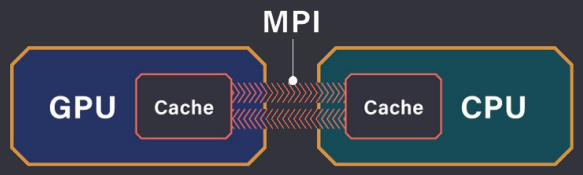


8-Wide Decode
Ascalon
Server, Laptop, and HPC

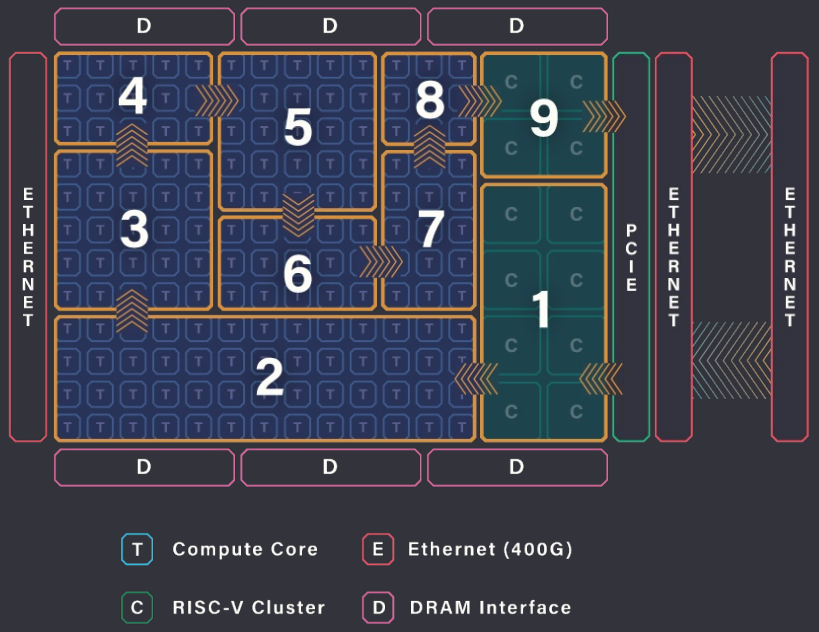
Decode Width

CPU for AI Computation

- AI computations
  - Data pre/post processing
  - Adaptive computing resources for future AI's algorithms
- CPU/GPU uniform node abstraction
  - Tenstorrent overlay technology
  - Same topological capability



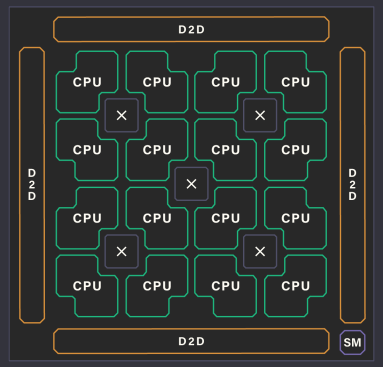
Dataflow Graph Mapping



AEGIS Chiplet System Architecture

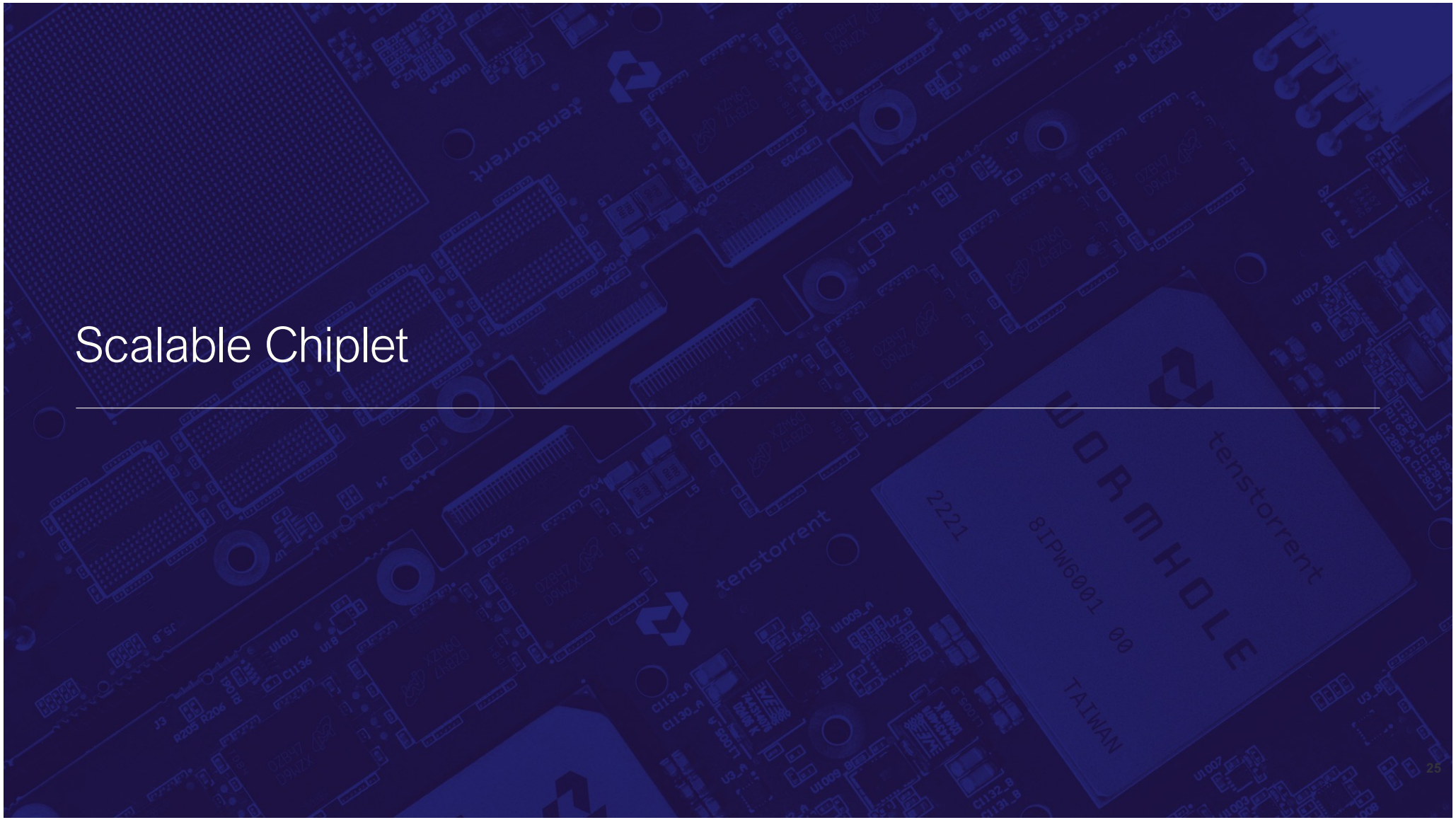
16 CPU-cluster system

- Companion CPU cluster for AI
- Inter-cluster coherency
- Directory-base coherency system
- Large memory cache per DDR5-6400 channel
- 4 cc-NUMA 32-core quadrants with hierarchical interconnection
- Ample coherent/non-coherent bandwidth for system scalability

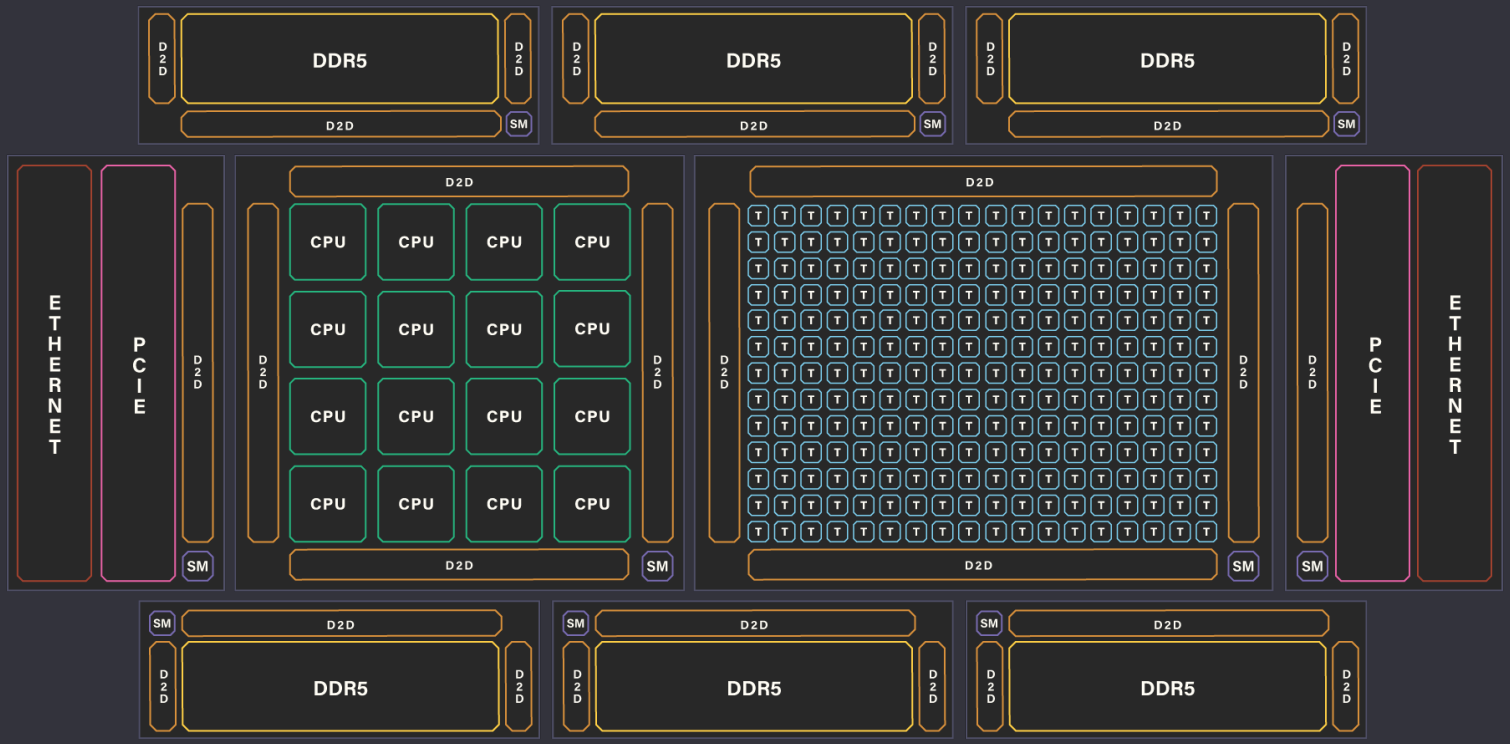


Fabric Chiplet Floorplan

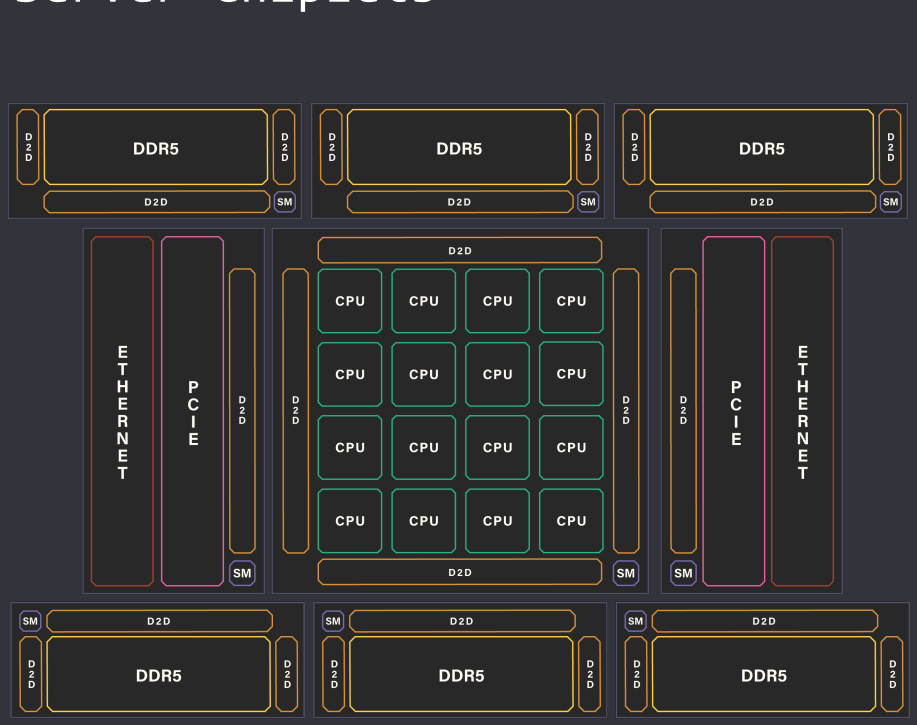
Scalable Chiplet

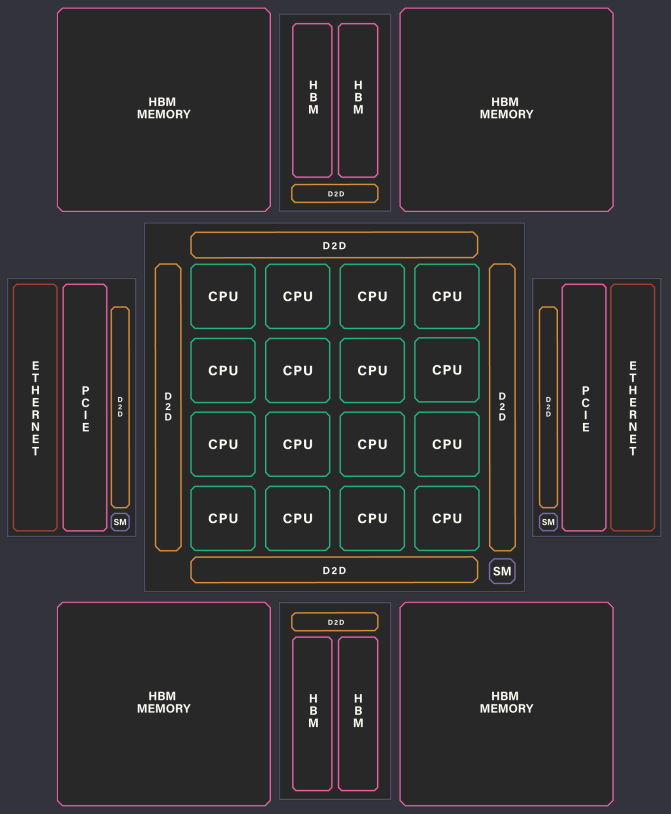


Heterogenous ML Processor

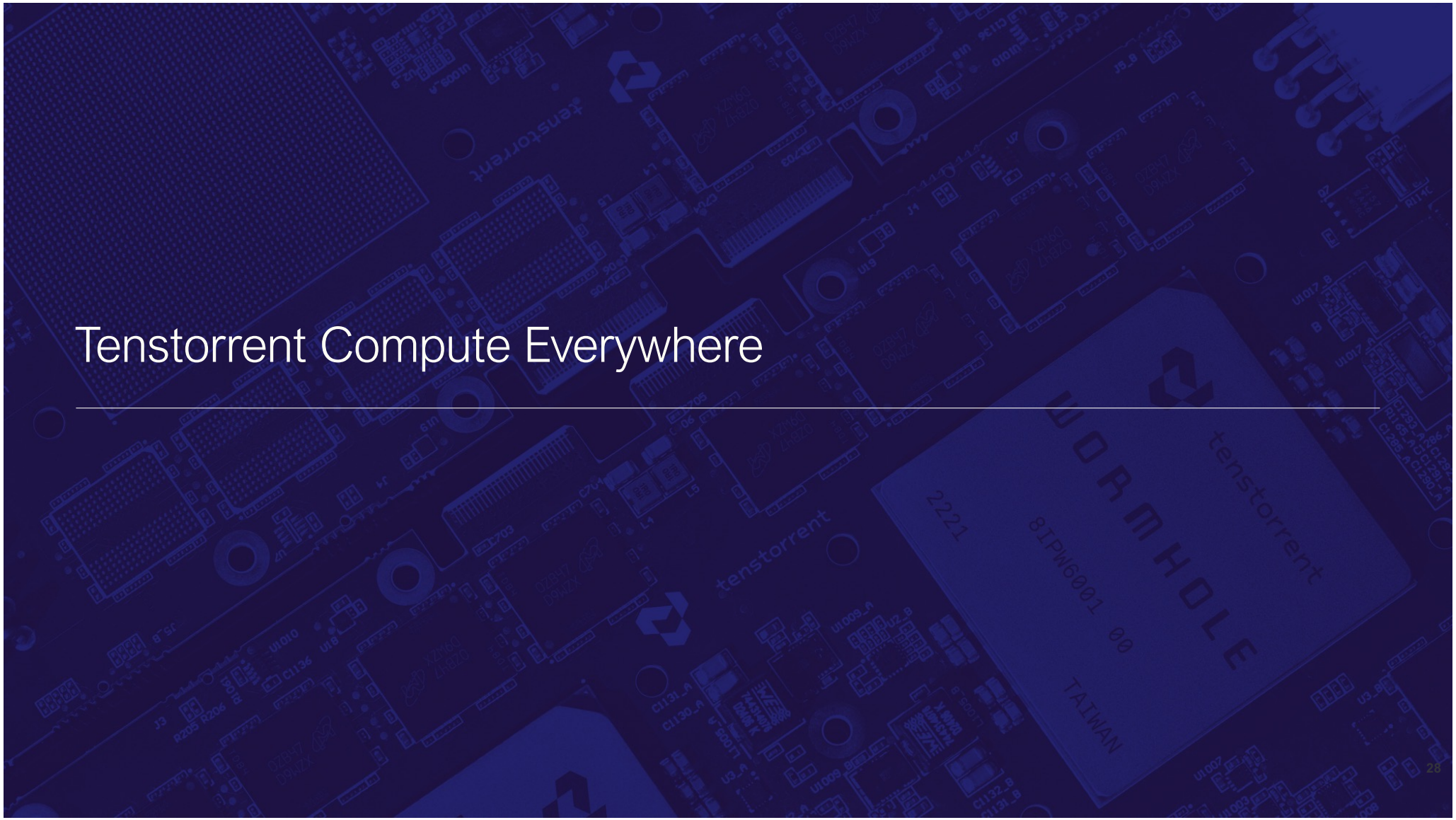


Server Chiplets





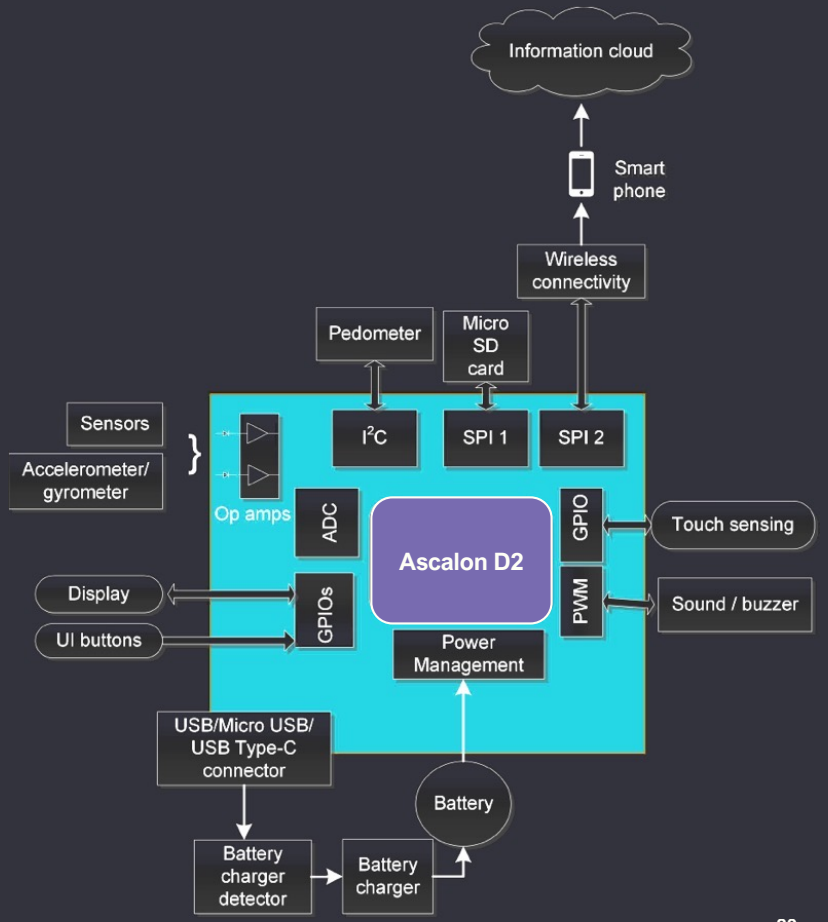
Tenstorrent Compute Everywhere



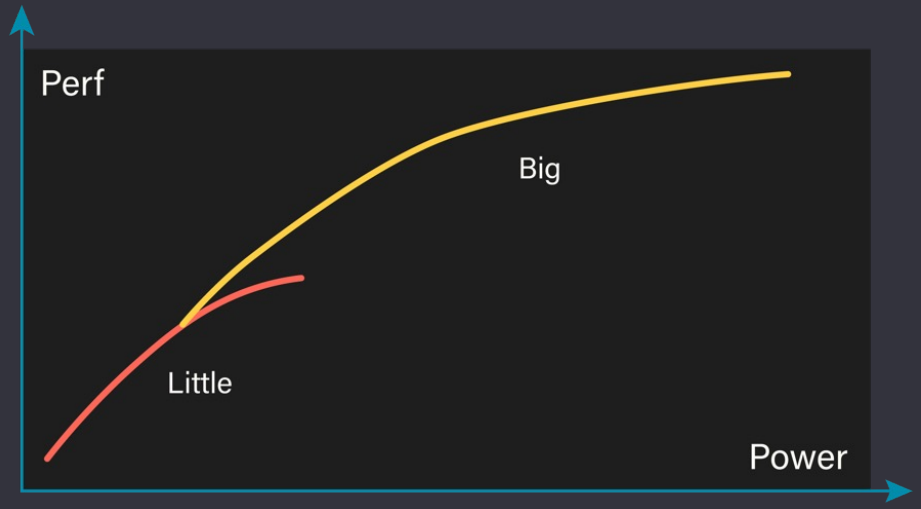
Wearable Computing

Wearable SoC with Ascalon-D2

- 10 mw–100 mw power consumption in advance node
- ARM A72 high-performance superscalar processor

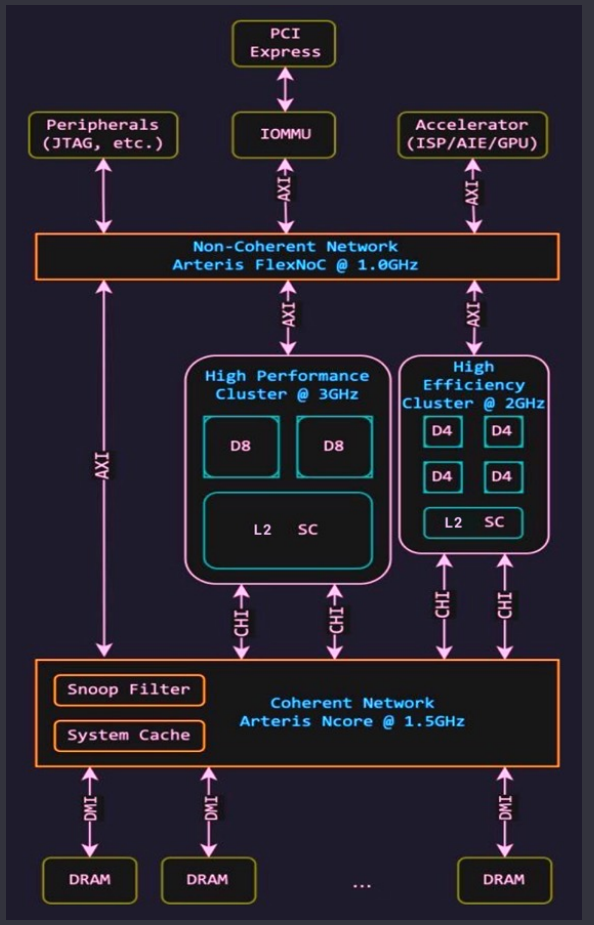


Mobile Computing



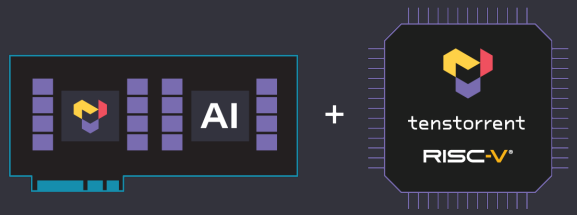
Big/Little Cores

- Big core Ascalon 8-wide decode
- Little core Ascalon 4-wide decode
- Implementation based on power-efficiency curves
- Complementary DVFS states cover wide range performance/power operating points

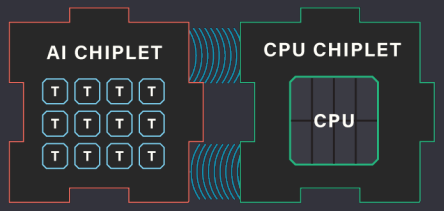


tenstorrent

Automotive

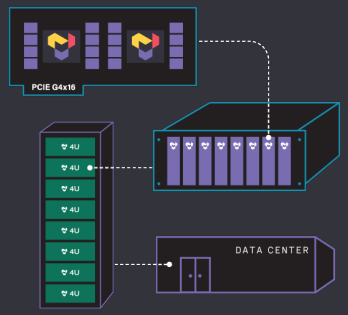


Tenstorrent AI and RISC-V IP deliver the compute power that ADAS and IVI require



Chiplet approach reduces cost while accelerating design and production schedules.

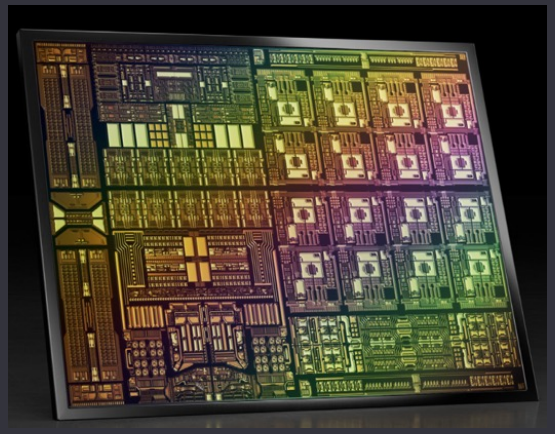
Automotive companies can own their own silicon working with Tenstorrent

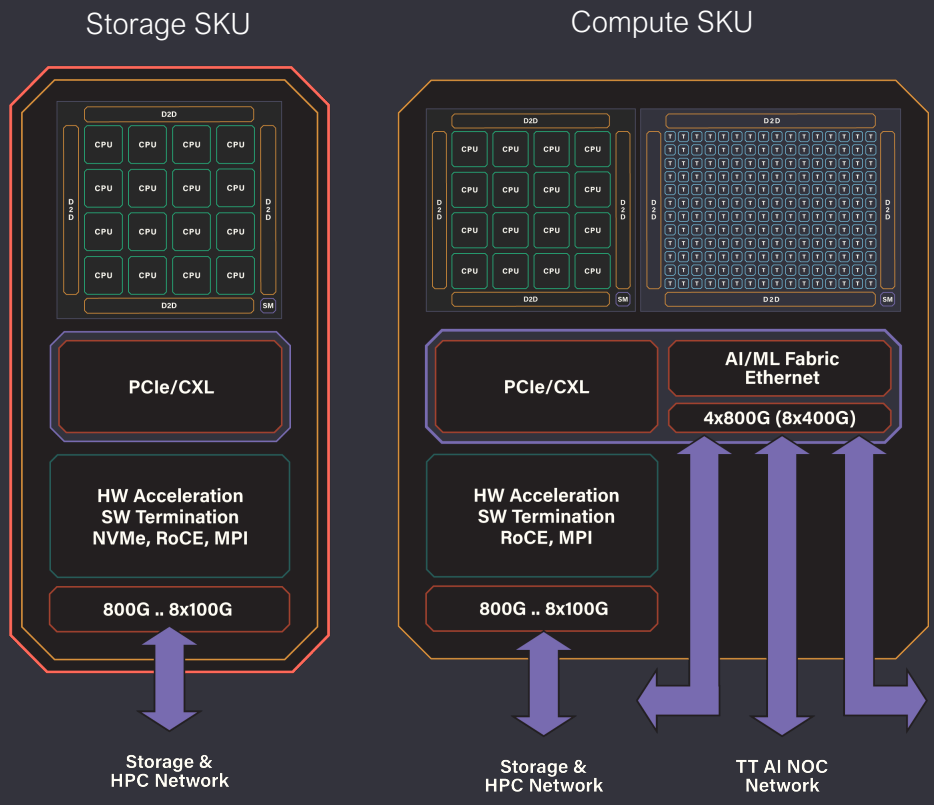


Power Consumption is critical: Tenstorrent technology scales from MW to mW

Network Packet Processing

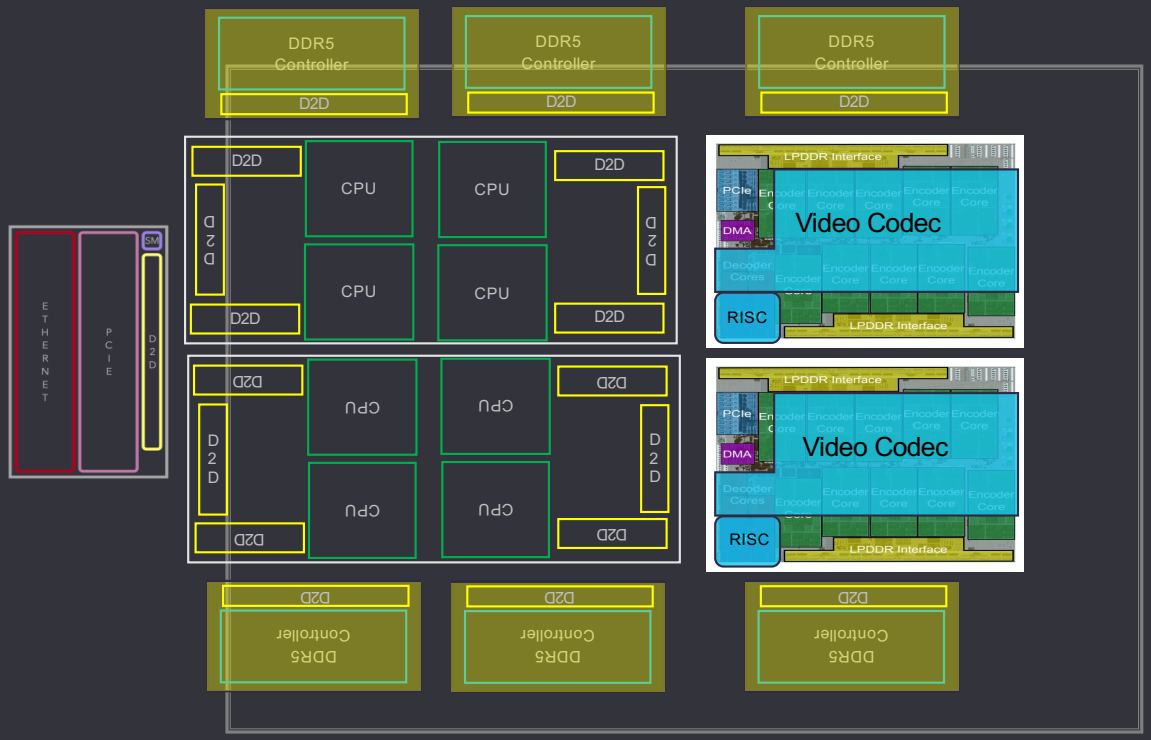
- Scale out for large computation
  - Smart NIC
  - DPU
  - Storage Server



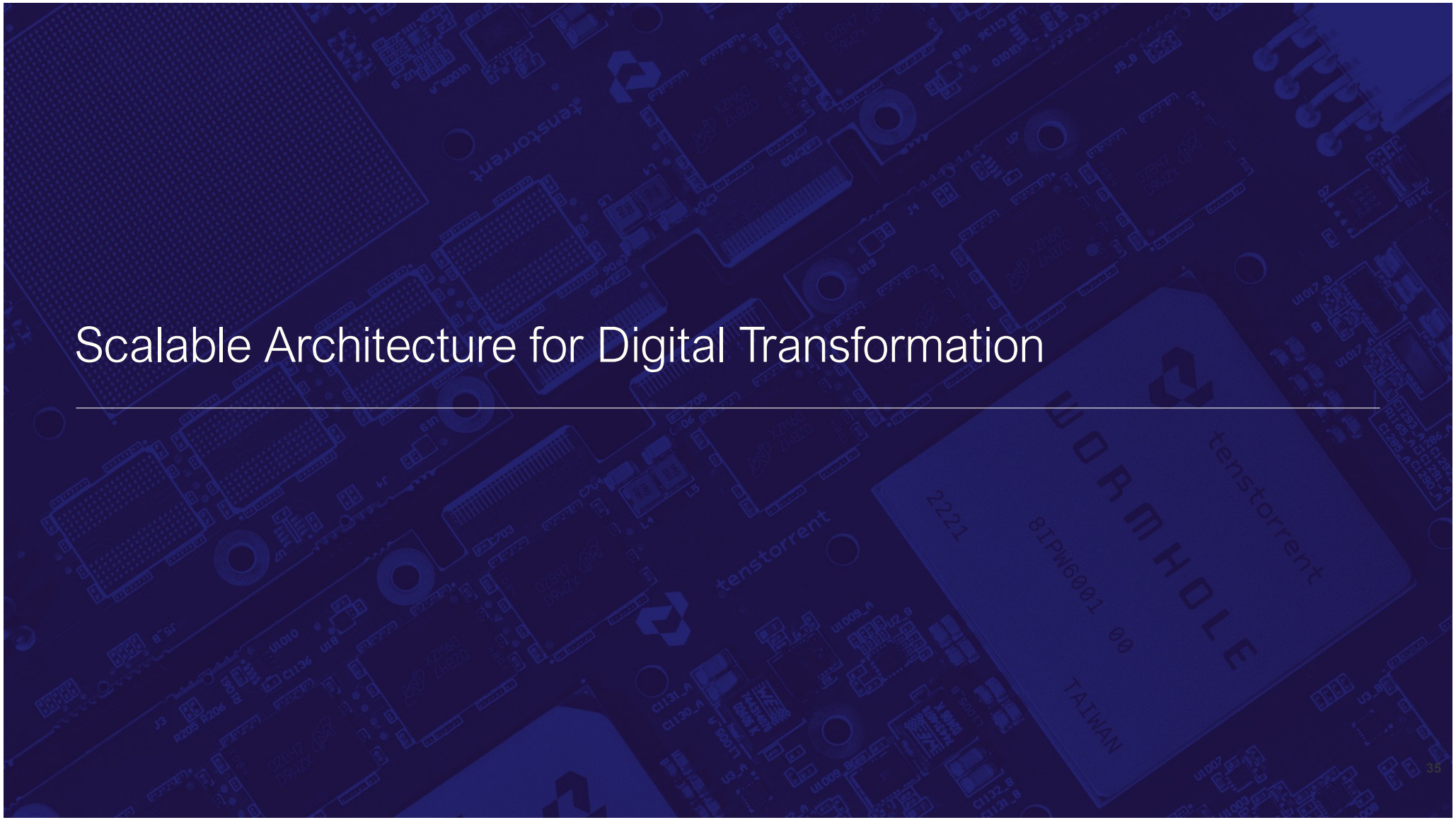


Tenstorrent CPU/AI-based Video Server

- Host
  - 2 x 32-core Aegis chiplet
- 2 x Video accelerator chiple
  - Video IP
    - 10 x 4Kp60 transcodes
  - Controller CPU
    - Ascalon D2 cores, or
    - TT Baby RISC cores

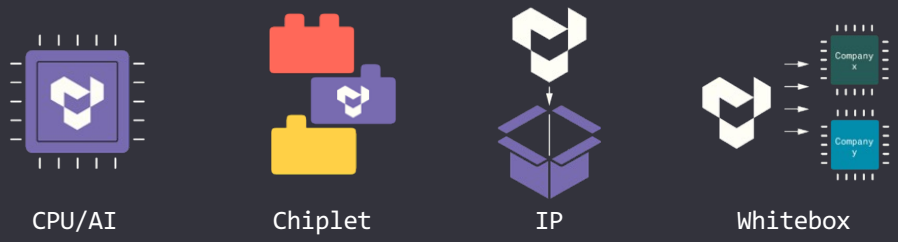


Scalable Architecture for Digital Transformation



Tenstorrent Scalable Architecture for Digital Transformation

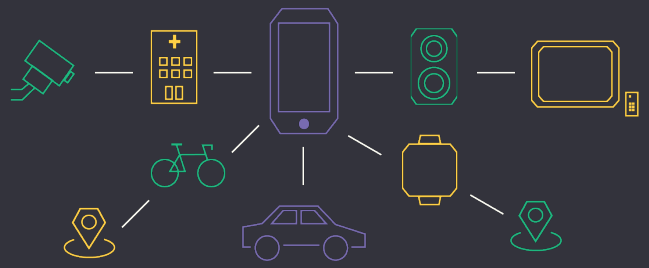
- Digital transformation requires CPU/AI computing everywhere
- Key technology providers for wide spectrum of products for our strategy partners
  - AI
  - CPU



HPC/Cloud

Edge Cloud

Edge Devices

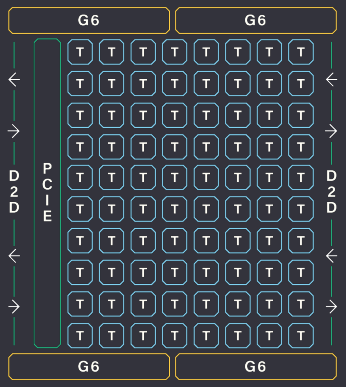


Compute Everywhere

Scalable CPU Family

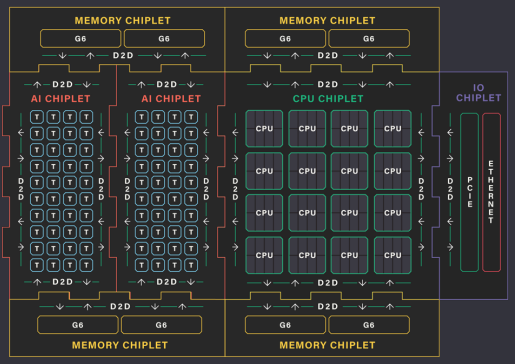
+

Scalable AI



+

Scalable Chiplet



Tenstorrent RISC-V CPUs and ML technology are in a unique position