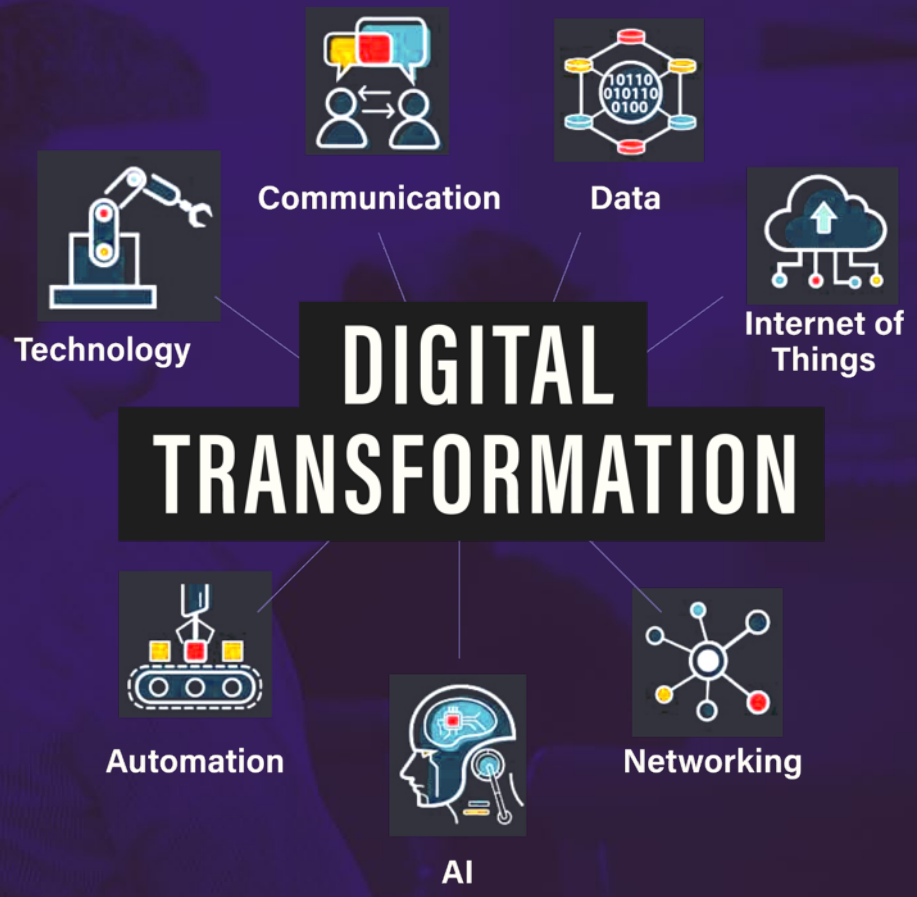
Scalable and Open-source Edge AI Architecture

練維漢
Wei-han Lien
Chief Architect and Senior Fellow

tenstorrent

July 2024

Digital Transformation

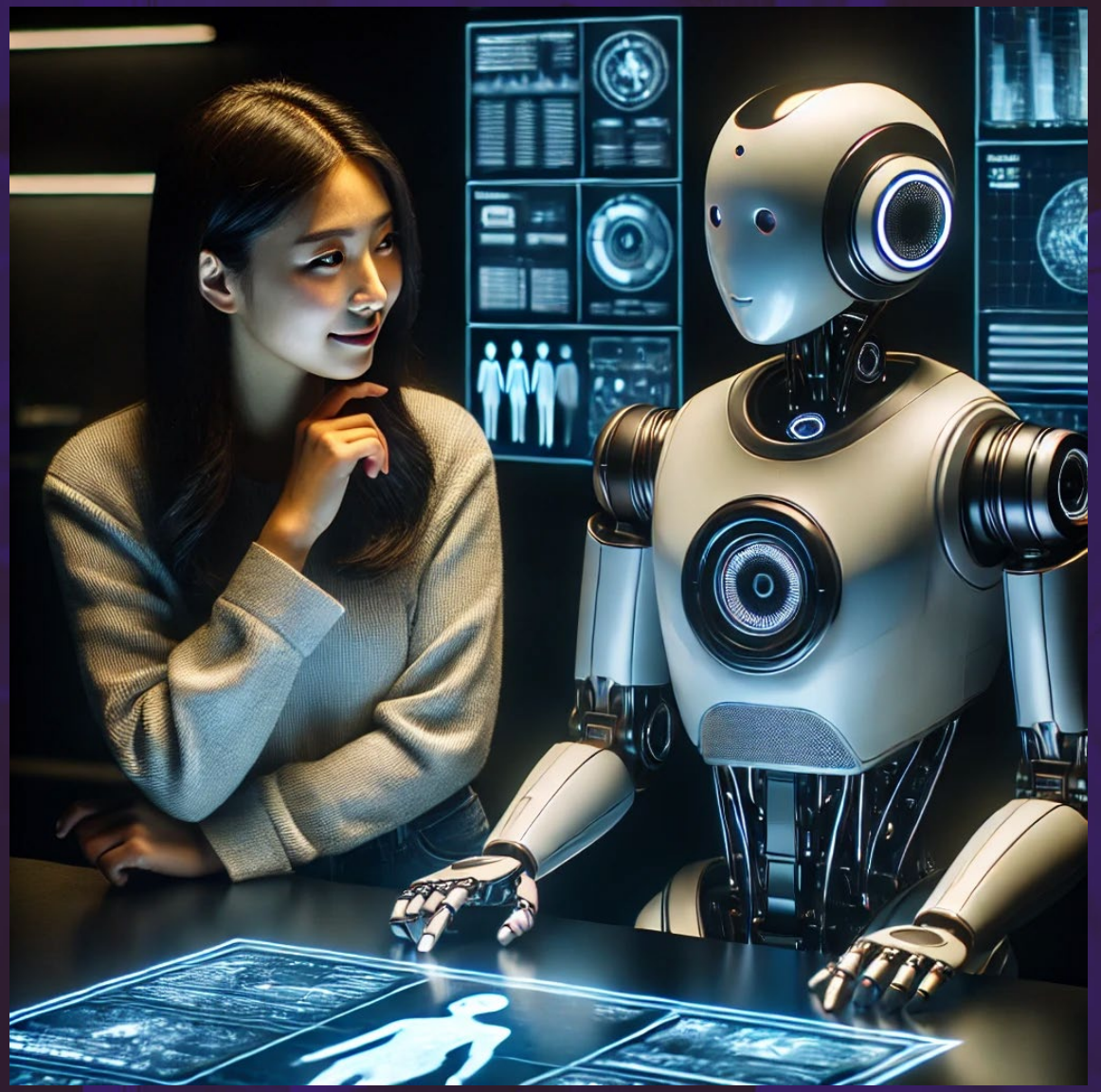


Human race is entering Digital AI Transformation

- AI revolution: Machine intelligence replaces human Intellect
- Reshape business models, practices, and cultures for competitiveness
- 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

AI Personalization

- ChatGPT3 significantly improves the AI usability
- Intimacy of AI
  - Tailored experiences
  - Real-time processing of nuanced inputs
  - Human like interactions
  - Privacy and performance
  - Adaptable and Evolving

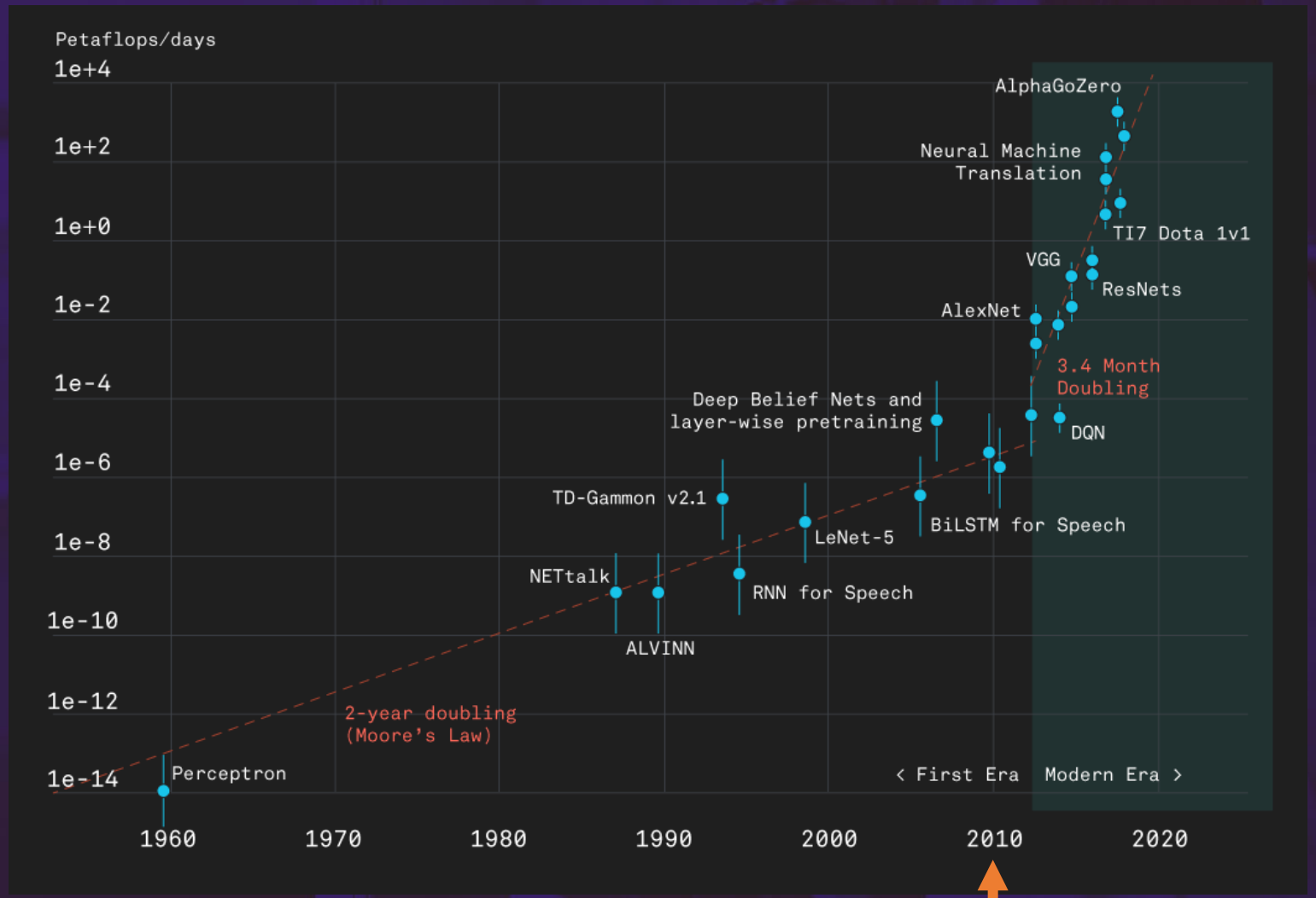


Digital Transformation Compute Everywhere

- Exponential AI model size grow since 2012
- RISC-V started in 2010
- ChatGPT4 = 2-trillion parameters
- 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



RISC-V Start

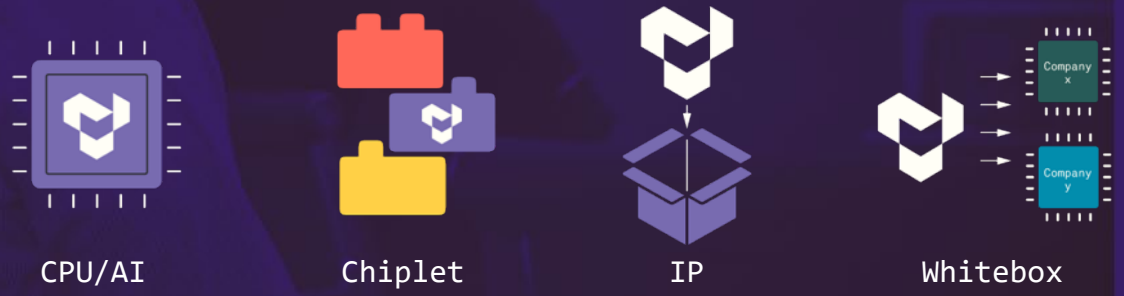
Distribute Compute for AI

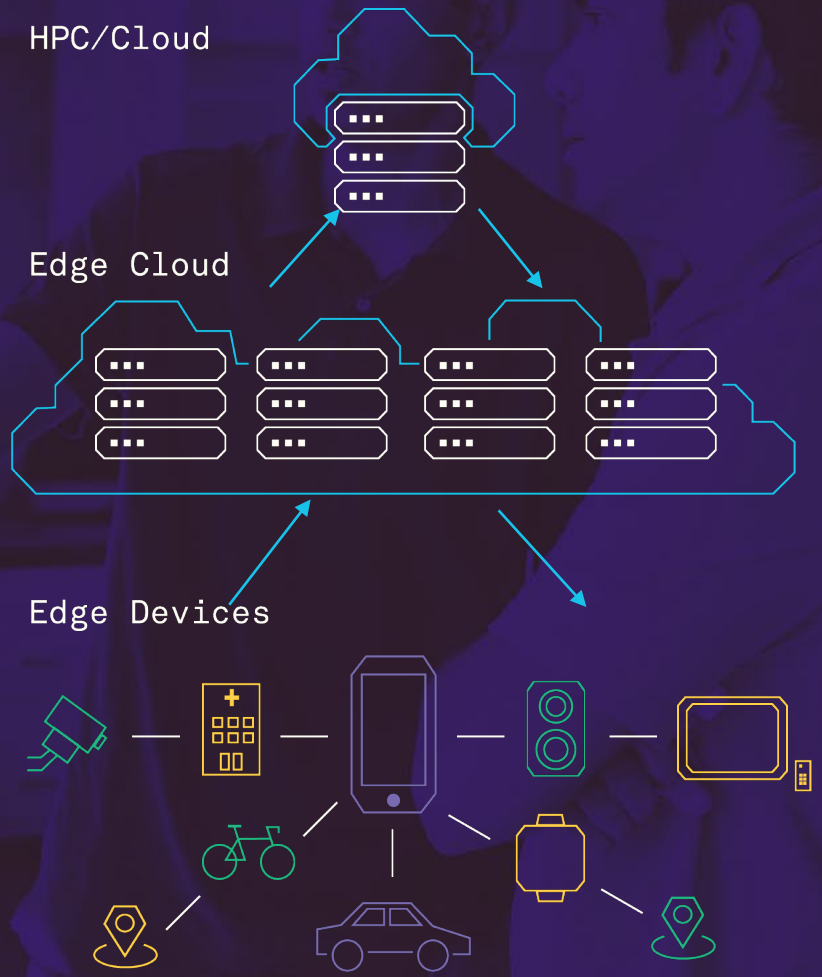
- Localized Processing
- Adaptive Resource Allocation
- Hierarchical Data Processing
- Efficient Pathway for Communication
- Intelligent Scaling and Dormancy
- Redundancy for Fault Tolerance and Recovery
- Real-time Energy

Unified AI Architecture

- AI pervasive computing from mw to MW
  - Client devices
  - Edge device
  - Data centers

- Tesntorrent provides key scalable AI enablement technologies
  - CPU
  - AI
  - Chiplets

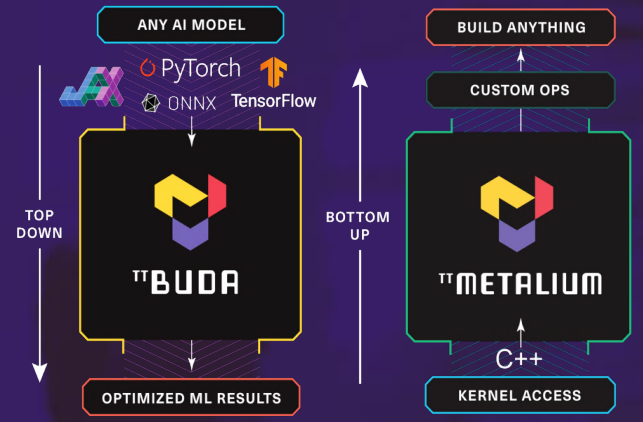


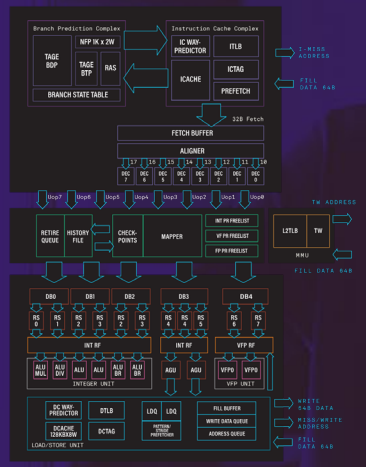


tenstorrent

Benefits of Open-source

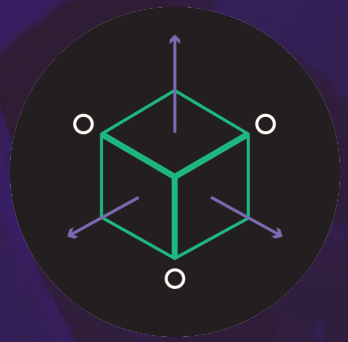
AI Software Stacks



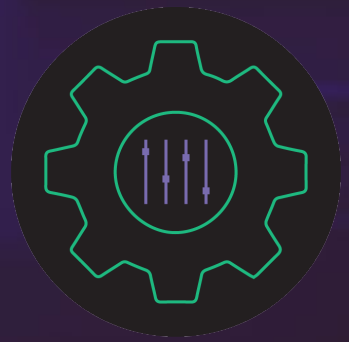


RISC-V

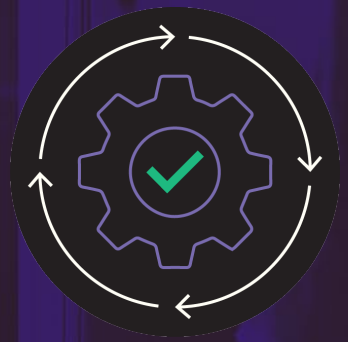
Open Standard CPU



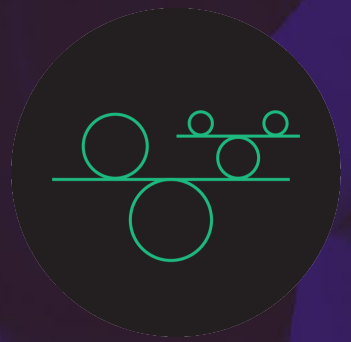
Scalable



Extensible



Efficient

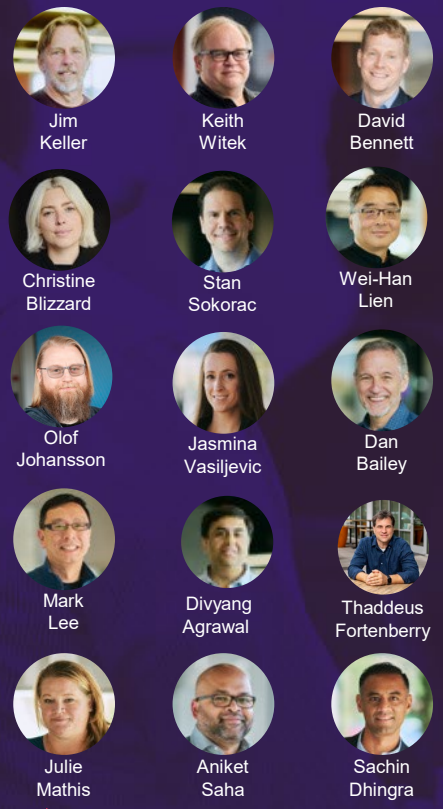


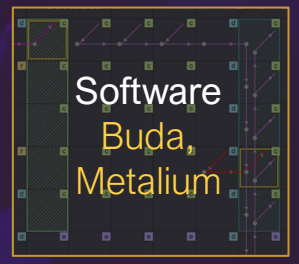
Stable

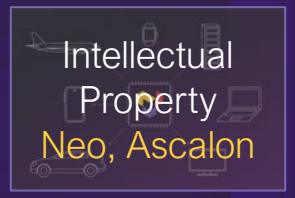
tenstorrent

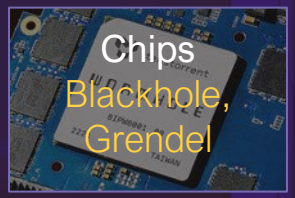
Software, Silicon and Systems to Run AI and ML Fast

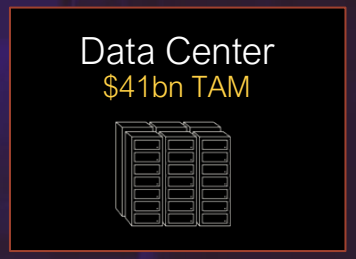
World Class Team → **Leading Technology** → **Innovative Products** → **Large Target Markets** → **Blue-Chip Partner Ecosystem**

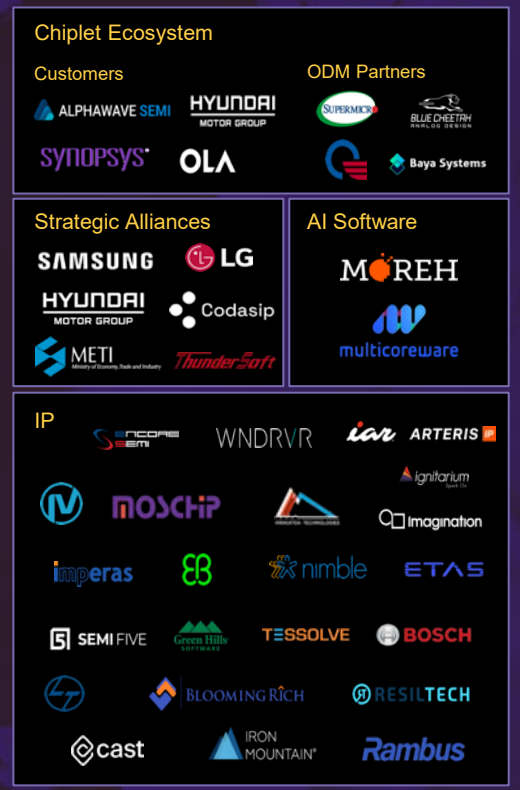












tenstorrent

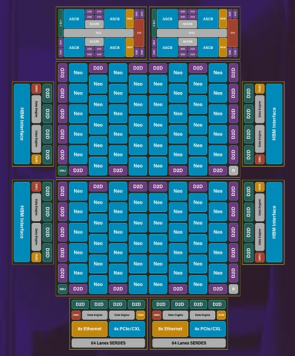
TAM figures represent the silicon TAM.

Our Technology

IP (Ascalon / Tensix-Neo)

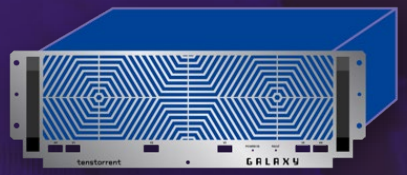
- Scales from mW to MW for efficiency and performance
- IP available for licensing
- Industry-leading performance
- Modular design available in varied configurations

Chips & Chiplets



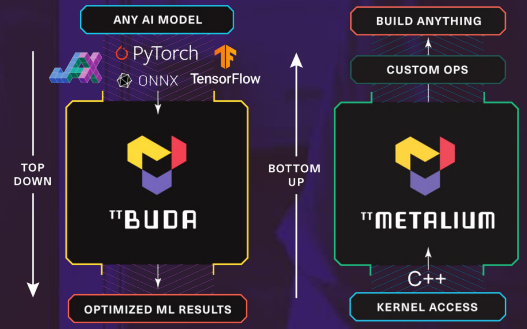
- Portfolio of cards powered by scalable Tensix AI cores
- Inference and Training, CNN and NLP, Recommendation Engines, all on the same silicon
- Hardware available for purchase, as well as IP available for licensing
- Multi-component modular chiplets

Servers (Galaxy)



- Galaxy Server – 32 high performance cards in a custom chassis - starts shipping in 2024
- Servers are easily combined into a Galaxy Rack for high bandwidth chip-to-chip connectivity

Software



- ML compilers that scale from one chip to thousands
- Buda - Automated AI/ML Compiler
- Metalium - Bare Metal Software Stack

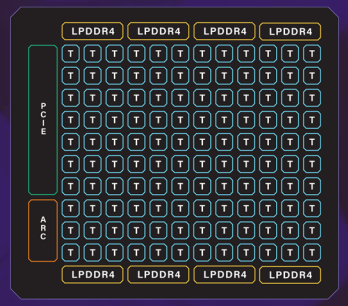
tenstorrent

Silicon Roadmap

2021	2022	2023	2024
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

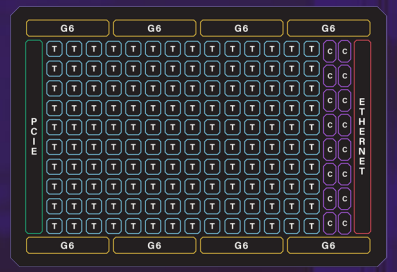
Wormhole

Networked AI Processor

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

Blackhole

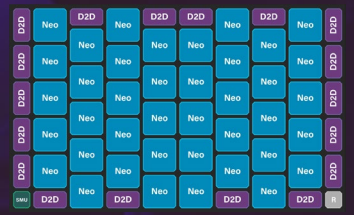
Standalone AI Computer



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

Quasar

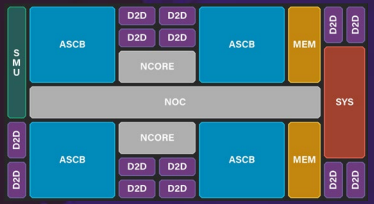
Low Power AI Chiplet



- 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

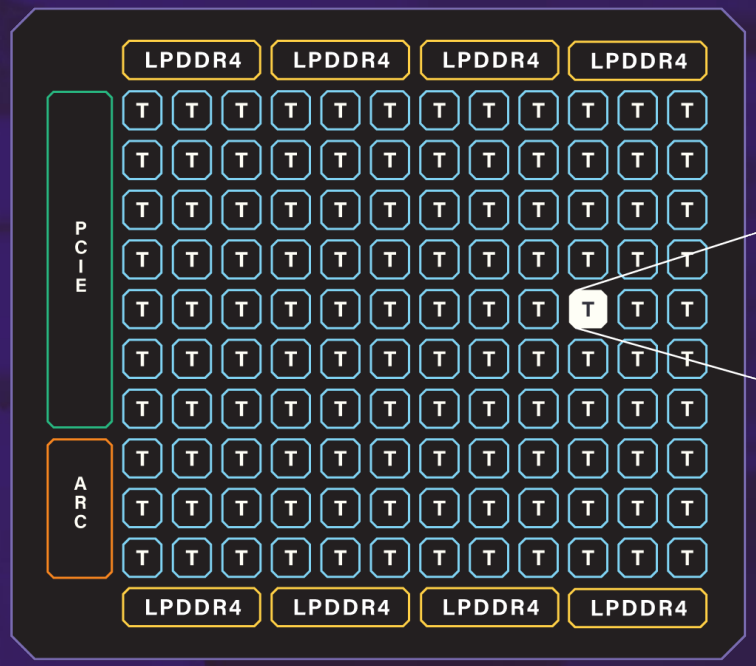


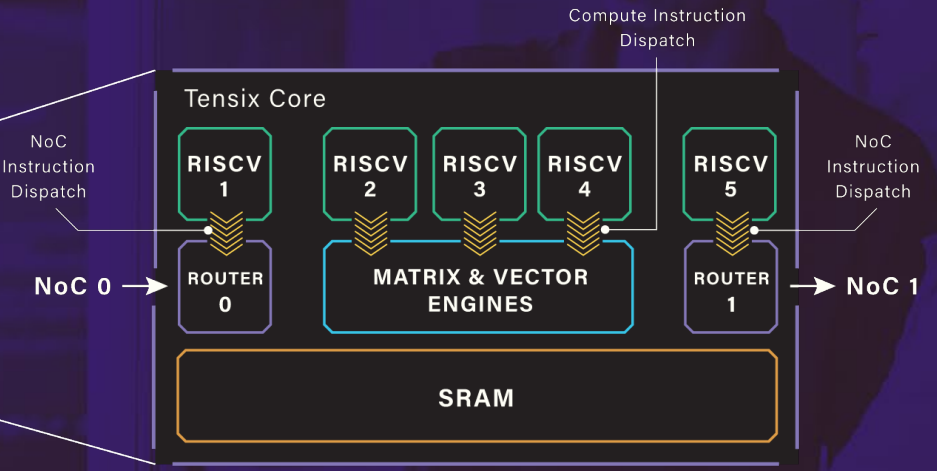
- 4nm 32 RISC-V Ascalon CPU Cores
- Scalable up to 128 Cores / 4 Chiplets
- Feature support incl SMC, IOMMU, AIA
- Non-blocking D2D Interfaces
- Composable IO, MEM, CPU compute

GEN #1	GEN #2	GEN #3

tenstorrent

Scalable Tensix Element





Grayskull: 120 Tensix cores

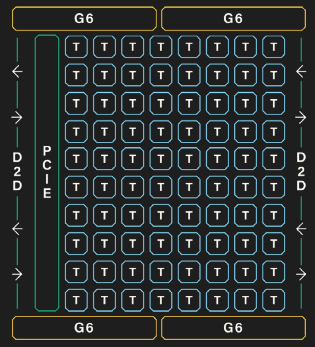
- Communication Subsystem
  - RISC-V controlled NoC subsystems
- Computation subsystem
  - General computation: “Baby” RISC-V
  - RISC-V controlled Matrix and vector engines
- Collaboration Mechanisms
  - Hardware supported through RISC-V

tenstorrent

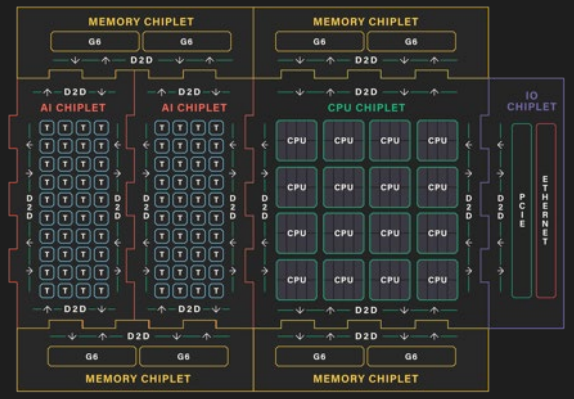
Scalable AI Architecture

IPs

Chiplet



System Chiplets



Scale-out System

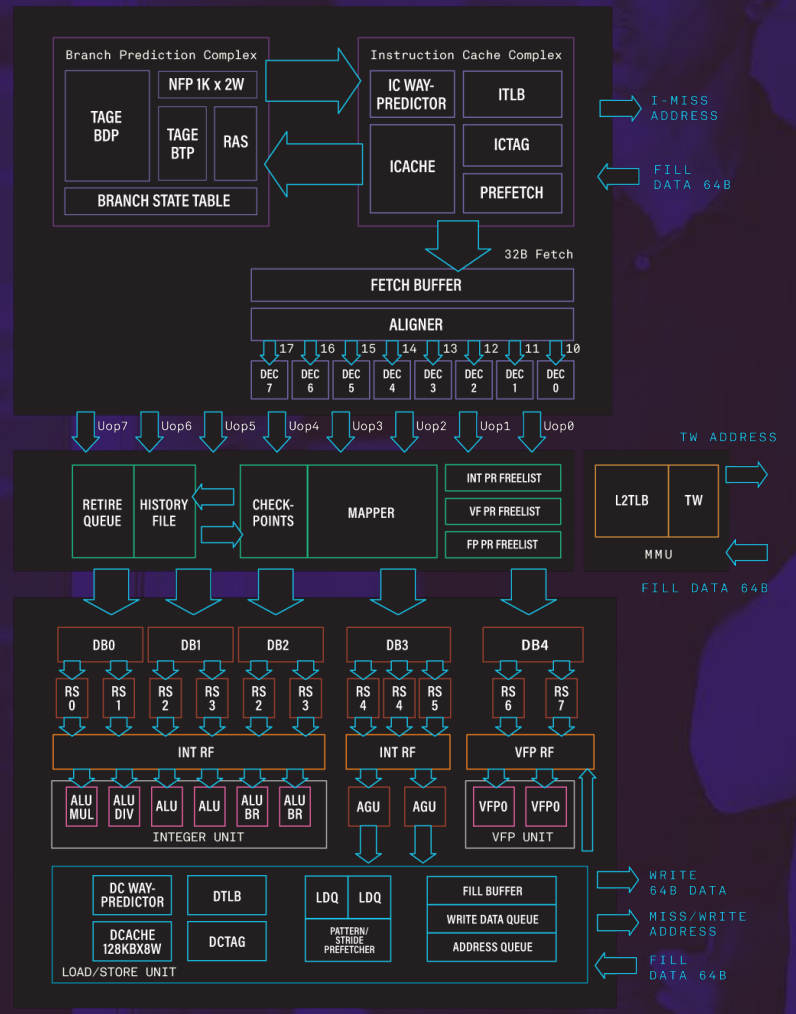
AI scalability from 1 Tensix core to thousands of chips

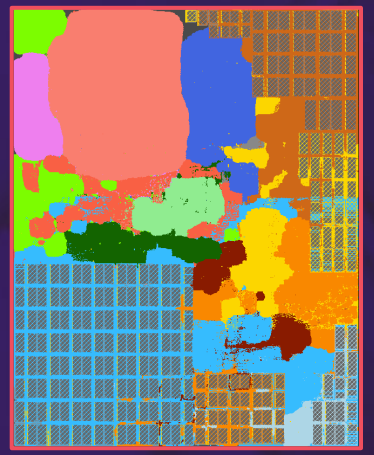
Ascalon O-o-O Superscalar Processor

- Disruptive high-performance RISC-V processor for AI and server
- Best performance & power efficiency

RVA23

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





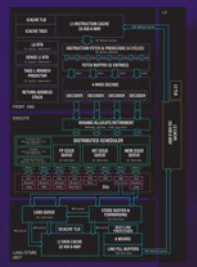
tenstorrent

Tenstorrent RISC-V O-o-O Processor Family

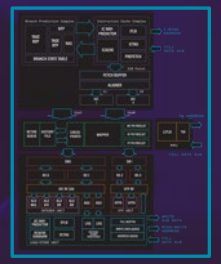
Open & Free

Higher Performance

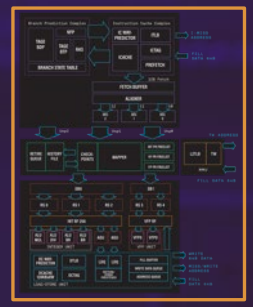
Performance



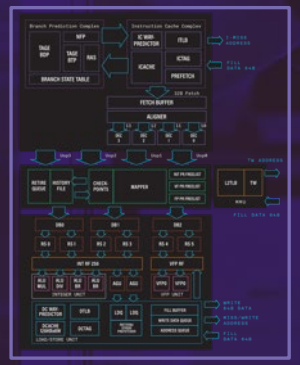
4-Wide Decode
Sonic Boom with Vector



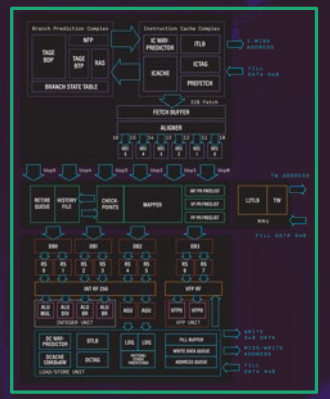
2-Wide Decode



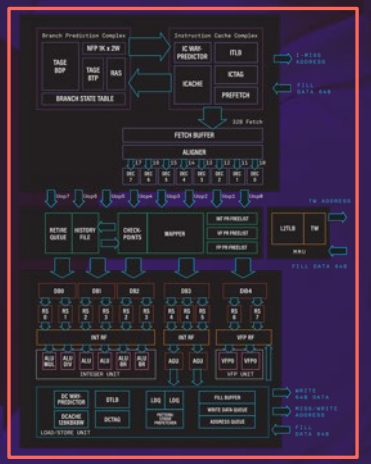
3-Wide Decode



4-Wide Decode



6-Wide Decode
Alastor
Client and Edge

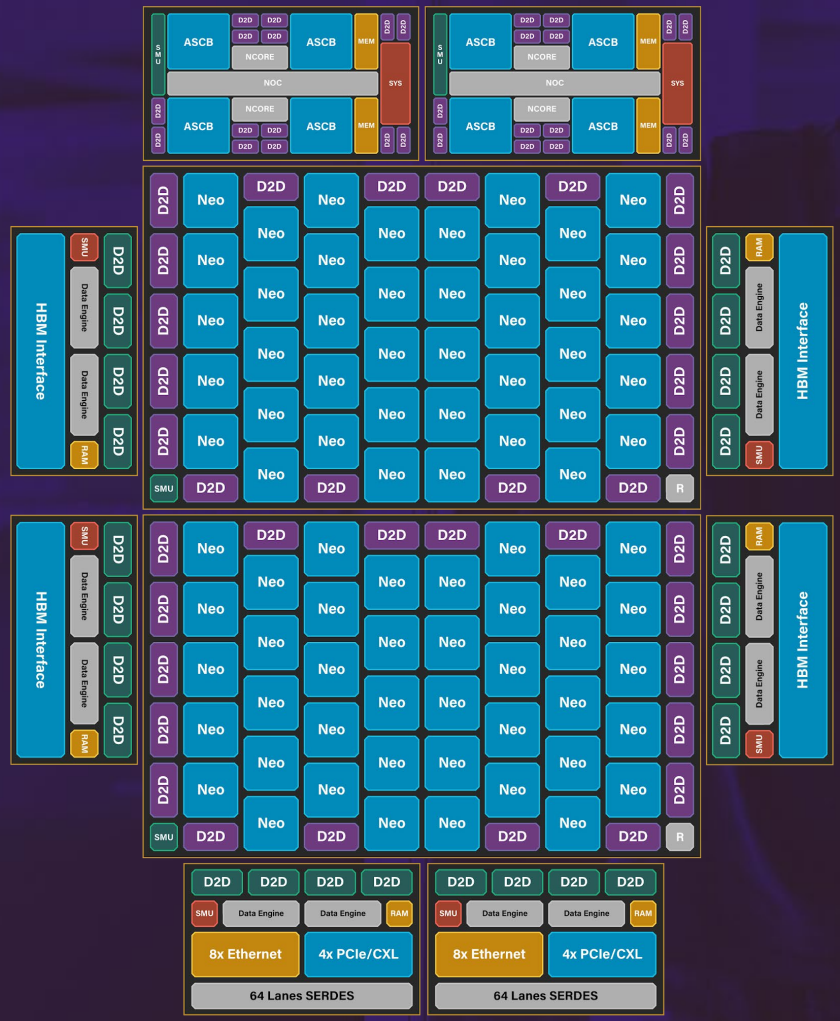


8-Wide Decode
Ascalon
Server, Laptop, and HPC

Decode Width

Chiplet

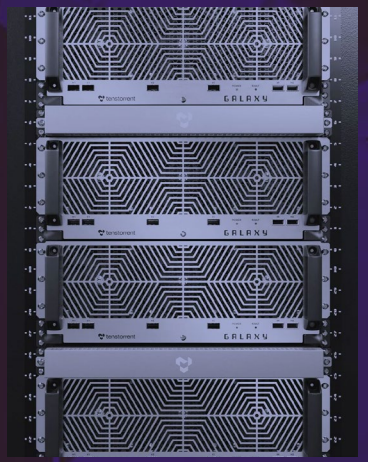
- Design Reuse
- Compossibility
- Scalability

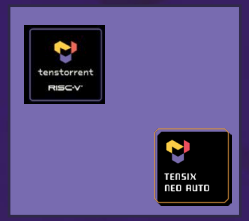


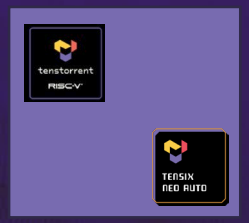
tenstorrent

Auto IP potential user cases in ADAS/ADS

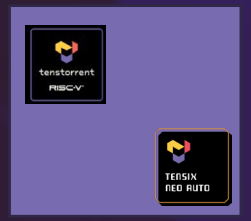
Level 2+/3 → Level 3/4 → Level 4/5 → Data Center

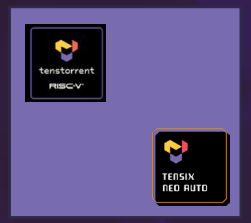






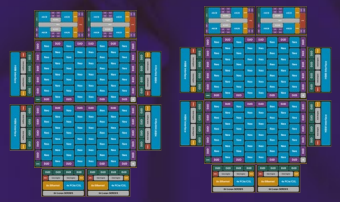
Dual SoCs





Dual SoCs + AI Co Processors

Central Compute Unit - Heterogenous



tenstorrent

Summary

- AI compute is pervasive
- Unified scalable architecture
  - Scalable AI
  - Scalable RISC-V
  - Chiplet
  - Open-source for innovation
- Edge AI
  - Necessary for tailored user experiences
  - Deployment constraints
    - Power and Thermal
    - Confidentiality
    - Safety