From mW to MW: Scalable RISC-V Processors for AI Everywhere

Wei-han Lien
Chief CPU Architect

May 2023
Agenda

- Introduction
- RISC-V based AI
- RISC-V processor family
- Chiplets
Tenstorrent

- Founded in 2016 to build the best ML training/inference chips
- $230M 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

2021
Grayskull
ML Processor

• 12nm, 276 TFLOP (FP8)

Wormhole
Networked ML Processor

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

2022
Black Hole
Standalone ML Computer

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

2023
Quasar
Low Power, Low Cost ML Chiplet

• ML Chiplet

2024
Grendel
Highly Configurable and Performant ML Chiplet

• CPU + ML chiplets
Scalable Tensix Element

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

Grayskull: 120 Tensix cores
Wormhole Products (2\textsuperscript{nd} 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
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
RISC-V CPU
Ascalon 0-o-0 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 0-o-0 Processor Family

One Design and 5 IPs in a year

Performance

Open & Free

Higher Performance

4-Wide Decode
Sonic Boom with Vector

2-Wide Decode

3-Wide Decode

4-Wide Decode
Client and Edge

6-Wide Decode
Server, Laptop, and HPC

8-Wide Decode

Decode Width

Confidential
CPU in AI

Host CPU
- X86 replacement
- Virtualization
- Security
- System Management
- Computation kernel scheduling/setup
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
CPU for Network Packet Processing

• Scale out for large computation
  • Smart NIC
  • DPU
Confidential

AI↔RISC-V Collaboration

Original: Embedded simple RISC-V processors for AI

Now: Integrated general purpose X280 RISC-V

Future: Heterogenous high-performance RISC-V + AI chiplets
Chiplet
AEGIS Chiplet System Architecture

16 CPU-cluster system

- Companion CPU cluster for AI
- Inter-cluster coherency
- Directory-base coherency system
- Large memory cache per memory channel
- 4 cc-NUMA 32-core quadrants with hierarchical interconnection
- Ample coherent/non-coherent bandwidth for system scalability
Heterogenous ML Processor
AI Everywhere
Tenstorrent: Open Business Model

- Tenstorrent works with partners to design, create, modify, optimize heterogeneous designs

- Key technology providers for wide spectrum of products for our strategy partners
  - AI
  - CPU
AI Everywhere

CPU Family

Higher Performance

One Design and 5 IPs in a year

• CPU family
• Scalable ML processor
• Chiplets
• Tenstorrent RISC-V CPUs and ML technology unique position

Chiplets

Higher Performance

AI

One Design and 5 IPs in a year