Scaling HPC with Andes Technology: Optimizing Performance at Every Level

Jason Lin

Senior Technical Manager

Andes Technology

International workshop on RISC-V for HPC
at HPC Asia 2025, Feb. 19th 2025

Andes Technology Corporation

Who We Are

CPU Pure-play CPU **IP Vendor**

19-year-old Public Company

Active Open-Source Contributor/Maintainer

Hsinchu, TW (HQ)

Active Roles in RISC-V Community

RISC-V® Founding & Premier Member

- Director of the Board
- Technical Steering Committee
- Chair/Co-Chair of Task Groups

RISE RISC-V Software Ecosystem Founding and Premier Member

- Governing Board
- Technical Steering Committee

Quick Facts

5th gen architecture
AndeStar™ V5, RISC-V adopted

500+
Employees, 80% R&D

350+
Worldwide Licensees

~100K
AndeSight IDE installations

~16Bn+
Total shipment of Andes-Embedded™ SoC

Andes and AI, HPC

- **Driving High Performance Computing with Scalability and Customization**
  - Scalable & Optimized Architectures
  - Customizable for HPC Applications
  - Comprehensive Ecosystem Support
  - Proven in Complex Applications
- **Some of Our Customers**

AndesCore™ RISC-V Families & Roadmap

32b / 64b General Purpose Perf. Apps

32b MCU

Vector & SMP Available

6x Series
High-Performance
OoO, 13-stage, 4-issue

65

4x Series
Mainstream
8-stage, dual-issue

45

2x Series
Perf. MCU/MPU
5-stage

27

25

Compact Series
Small MCU
2, 3-stage

23

225

Next-Gen OoO HPC

AX66
- RVA23 Profile
- CHI for Muli-cluster w. Cache Coherence

NEW!

AX46MPV
- 16-core cluster
- 2048b vector
- Matrix Extension
- Boosted Mem Perf

NEW!

Functional Safety Family ISO 26262

D45-SE ASIL-D

NEW!

D25-SE

N25-SE

Auto HSM

D23-SE ASIL-D

NEW!

Maximizing Scalar Performance with AX60-Series

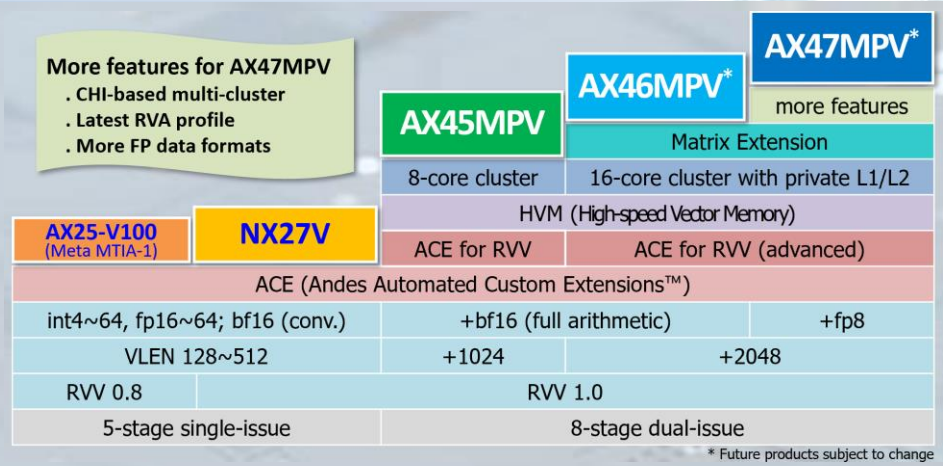
AX66 VPU:
- **1 or 2 pipes, shared with FP**
- **execute 2 ALU/load/store**

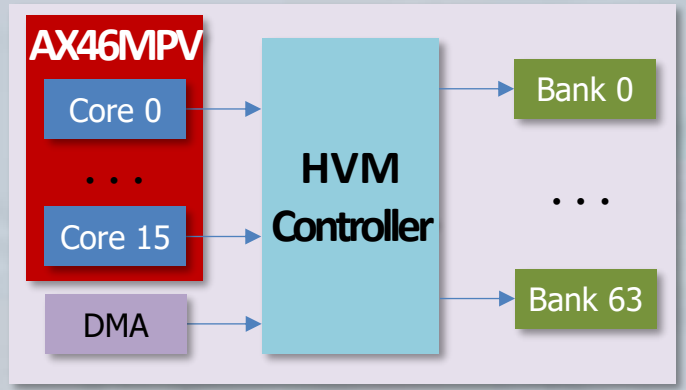
AX63 customer-driven **Power-optimized** *>7.0* specint2k6/GHz	**AX65** **Balanced** 8.78 specint2k6/GHz	**AX66*** **Advanced** *10* specint2k6/GHz	**AX67*** **Performant** *11* specint2k6/GHz	**Cuzco*** **Scalable** *15~20* specint2k6/GHz
	Android Base			
		RVA23	further perf boost	Private L1/L2, Shared L3
		V/VK (VLEN=128)	V/VK (VLEN to 512)	Vector/Vector Crypto
		Hypervisor + AIA + (IOMMU + IOPMP)		8-core Cluster with CHI
RVA22+		Private L1/L2, CHI Multi-Cluster Coherency	RVA23+	RVA23+
13-Stage, 4-way OOO, Linux-Capable, Multicore Coherency, Up to 8 Cores/Cluster				16-stage 6/8-way OOO with patented Time-Based Scheduling
The AX60 Series				**Cuzco Series**

** Future products subject to change*

Vector Processors and the New AX46MPV

- **Built on the success of AX45MPV**
  - 8-core cluster with 1024 VLEN/DLEN
  - Dual issue for vector and scalar
- **Doubled cores and VLEN**[1]
- **Enhanced ACE for scalar and RVV**
- **Andes Matrix Extension**
- **Boosted memory performance:**
  - **Dual loads,** or one load/one store
  - **Private L2$** (64KB~512KB, 8-way) for flat memory programming
  - **HVM** (High-speed Vector Memory) interface: multiple outstanding requests with OOO return
    - HVM controller: 16 cores + DMA, 64 banks
- **Other variants: AX46MP, A46MP(V)**

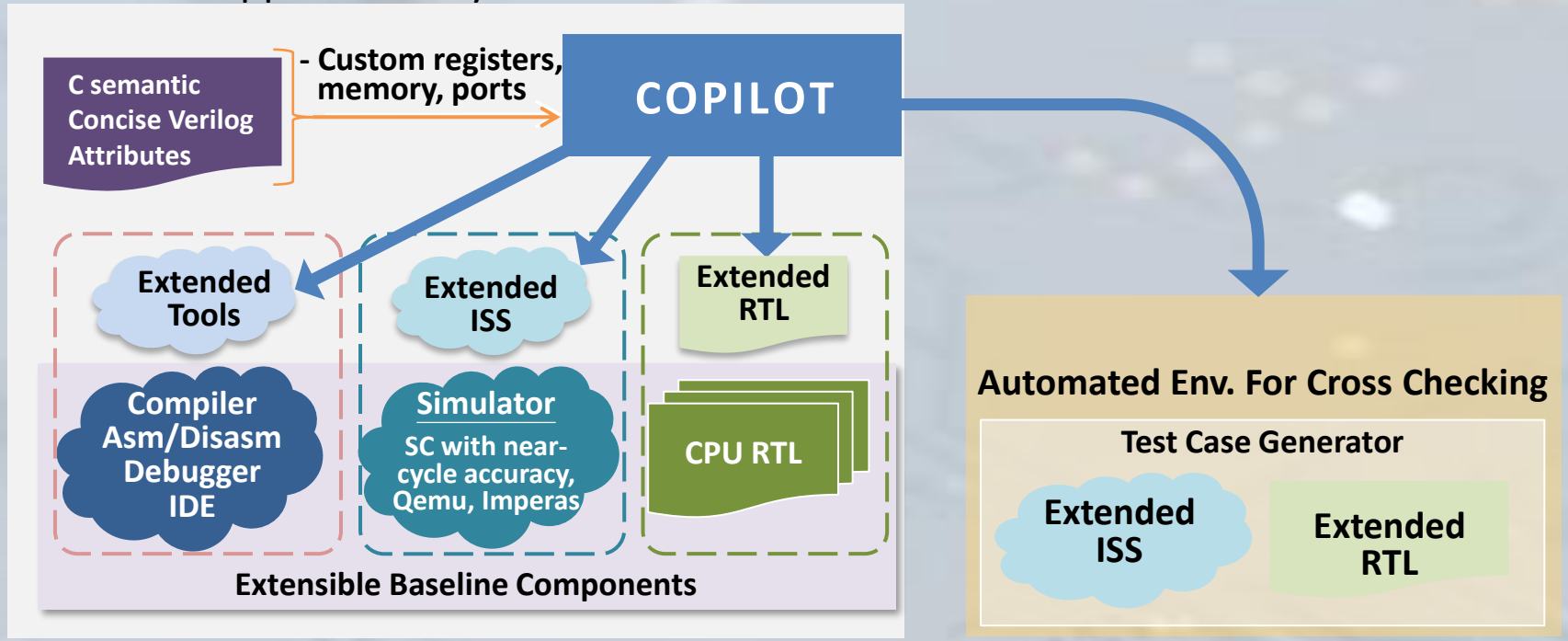




Note 1: VLEN/DLEN from 128/128 to 2048/1024

The Andes Custom Extension™ (ACE)

- **Andes Tool for Custom Instructions – COPILOT, Enables:**
  - ACE: Create new instructions to speed up computations and controls
  - ACE-RVV: support RVV-style instructions



AndesAIRE™ AI Software Stack

NN models — PyTorch, ONNX, TensorFlowLite, TensorFlow

AndesAIRE™ Software

AndesAIRE™ NNPilot™

- Graph-level optimization (Pruning/Quantization)
- Backend-aware optimization (Fusion/Tensor Allocation)

Generated TFL Models

TensorFlow Lite — TensorFlow Lite

Generated C Template

- **NN Library API**
- AnDLA driver and runtime
- AnDLA command image

AndesAIRE™ **XNNPACK**

AndesAIRE™ **NN Library**

AI Compilers

OpenXLA, IREE, MLIR, tvm, LLVM

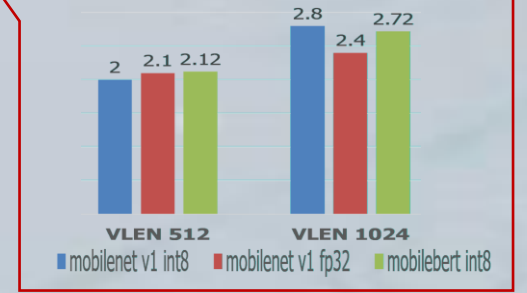
~200 high-level functions for PyTorch/TFL

220 functions for NNPILOT
(Plus, 420 in RVV library and 380 in RVP library)

LLVM auto-vectorizes code generated by IREE/TVM

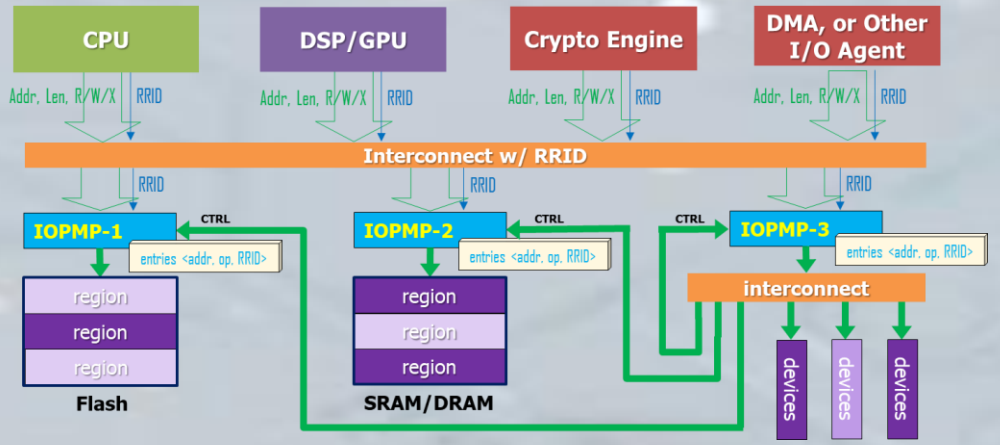
IREE can invoke **µKernel** optimized with RVV/ACE-RVV instructions

Auto-IREE: find a better tiling scheme efficiently; **>2x** speedup (over IREE)



Platform Level Security with IOPMP

- **Complete Protection on the Cybersecurity Vulnerabilities**
  - RISC-V PMP/ePMP covers CPU transactions, by
    - CPU privilege modes, memory regions, and operations
  - IOPMP mitigates threat from the other I/O agents
    - Andes chairs the IOPMP Task Group, and Vice-chair TEE TG in RISC-V International
    - Trusted Execution Environment
      - Checker with a set of ordered entries
      - Check transaction by
        - Address/Length
        - Operation (R/W/X)
        - Initiator (RRID)
      - Mechanisms in respond to violations



Meta Training and Inference Accelerator (MTIA)

Enabling Advanced AI & ML Solutions

- Sea of PEs (Processing Elements)
  - Core of the computations – the PEs
  - Mesh-connected (like Meta MTIA) and multi-clustered
- Three tiers of accelerations in PE's
  - Matrix Multiplications:
    - Hardwired solutions
    - Matrix instructions: RISC-V IME and AME extensions
  - Non-linear OP's: softmax, sigmoid, GeLu/SiLu/SwiGlu
    - Andes Automated Custom Extensions (ACE)
  - General compute: Catch all and future-proof
    - RISC-V Vector Extension (RVV)
- "MTIA: First Generation Silicon Targeting Meta's Recommendation Systems"
  - ISCA 2023 paper
  - Andes AX25-V100: an early version of the widely adopted NX27V for proc-A/B
  - Andes Custom Extension™: new interfaces, instructions and registers

Figure 3: High-level architecture of the accelerator

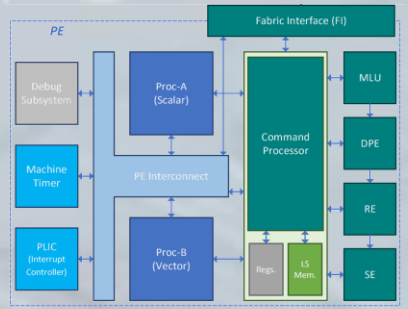


Figure 4: PE's internal organization

Takeaways

- RISC-V is Taking its Place in the Future of HPC
- Andes is Contributing to HPC with RISC-V Innovations
- Many Fields Yet to be Explored to Unleash its Full Potential

Thank You !!

ANDES TECHNOLOGY