# Powering Intelligent General-Purpose Computing with RISC-V

Charlie Su, Ph.D.
President and CTO
Andes Technology

2025/11/11

Taking RISC-V® Mainstream

# Agenda

- The Rise of Intelligent General Computing (IGC)
- RISC-V Ecosystem for AI/ML and General Computing
- Andes RISC-V Processor and Software Solutions for IGC
- Closing Remarks

# Andes Technology Corporation

## Background

CPU Pure-play for 20+ years, public since 2017 TW

Products in Endpoints, Edge, Auto and Cloud

Leading RISC-V IP market share since 2023 (SHD Group)


## Leading Roles in RISC-V

RISE RISC-V Software Ecosystem

- Founding & Premier Member
- Director of the Board
- Technical Steering Committee
- Chair/Co-Chair of Task/Work Groups
  - Early Days: Fast Interrupts, TEE
  - Active: IOPMP, RVP, System Libraries, IME

## Quick Facts

- **30+** RISC-V Processor IP's
- **500˜** employees
- **18 Billions** Andes-Powered SoC
- **First RISC-V vendor** to deliver
  - SIMD/DSP processor IPs
  - Vector processor IPs
  - Fully-compliant 26262 processor IPs

Taking RISC-V® Mainstream

# The Rise of Intelligent General Computing

- **Advancement in AI/ML brings intelligence to General Computing!**
- **AI PC/Phones, AI IPC/Edge Servers**
  - Personal use; factory automation, surveillance, drone




- **Software-Defined Vehicles:** ADAS Level 0-4, Autonomous Vehicles
- **Robot** (of any form) **as a Platform**: features enabled by APP

  ➔ expected to surpass the smartphone market !





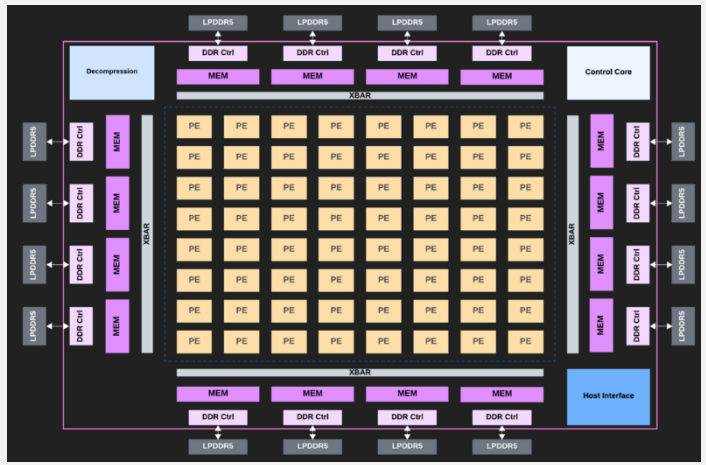
- **Intelligent General Computing** requires
  - Rich ecosystem for general-purpose computing
  - Advanced ecosystem for large-scale AI/ML

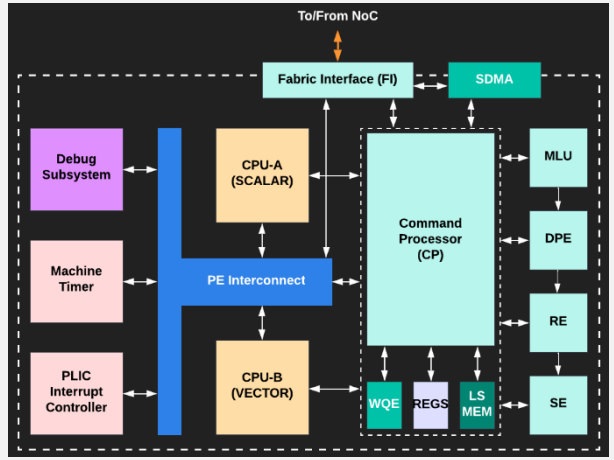
  → Both SW and HW

# RISC-V Enabled Innovations in Large-Scale AI/ML

➔**Most use Andes vector/scalar processors with automated custom extensions (ACE)**

- **Meta MTIA**:
  - Meta Training and Inference Accelerator
  - Using Andes vector/scalar cores and ACE to connect to their accelerators (ISCA 2023)
  - Two generations deployed in Meta datacenter since 2023




- **AI Accelerators Using SRAM-based CIM (Compute-In-Memory:)**

TetraMem Accelerate the World®

AXELERA ARTIFICIAL INTELLIGENCE

RAIN

FRACTILE

后摩智能 HOUMO.AI

- **AI SoC for Servers** RiVOS
- **AI Accelerator for Cloud Service** SAPEON
- **AI Accelerator Using Photonics** LIGHTELLIGENCE
- **AI Accelerator for ADAS**

Powered by AX46MPV/AX45MPV/NX27V/AX65 and more

Taking RISC-V® Mainstream

# RISC-V SW Ecosystem Fast Maturing with RVI and RISE

## Mission of RISE

- **Accelerate the development** of open-source SW for RISC-V
- **Raise the quality** of RISC-V Platform implementation
- **Align efforts to** push the RISC-V SW Ecosystem forward

| | |
|---|---|
| **Compilers & Toolchains** | LLVM, GCC, GLIBC |
| **System Libraries** | FFmpeg, OpenBLAS, OneDAL, XNNpack |
| **Kernel & Virtualization** | Linux, Android, Performance Profiles, DynamoRIO, Valgrind |
| **Language Runtimes** | Python, Java/OpenJDK, Go, Javascript, WebAssembly, Rust |
| **Linux Distro Integration** | Ubuntu, Debian, RedHat, Fedora, Alpine |
| **Simulators/Emulators** | QEMU, SPIKE |
| **System Firmware** | UEFI, U-Boot, Coreboot, TF-M |
| **Developer Infrastructure** | Build Farm, Board Farm, Developer Tools |
| **Security Software** | Secure Root-of-Trust, Confidential Compute |
| **AI/Machine Learning** | PyTorch, Llama.cpp |

**Premier Members**

ANDES, Google, intel, MEDIATEK, NVIDIA, QUALCOMM, Red Hat, Rivos, SAMSUNG, RISCSTAR SOLUTIONS, VENTANA

**General Members**

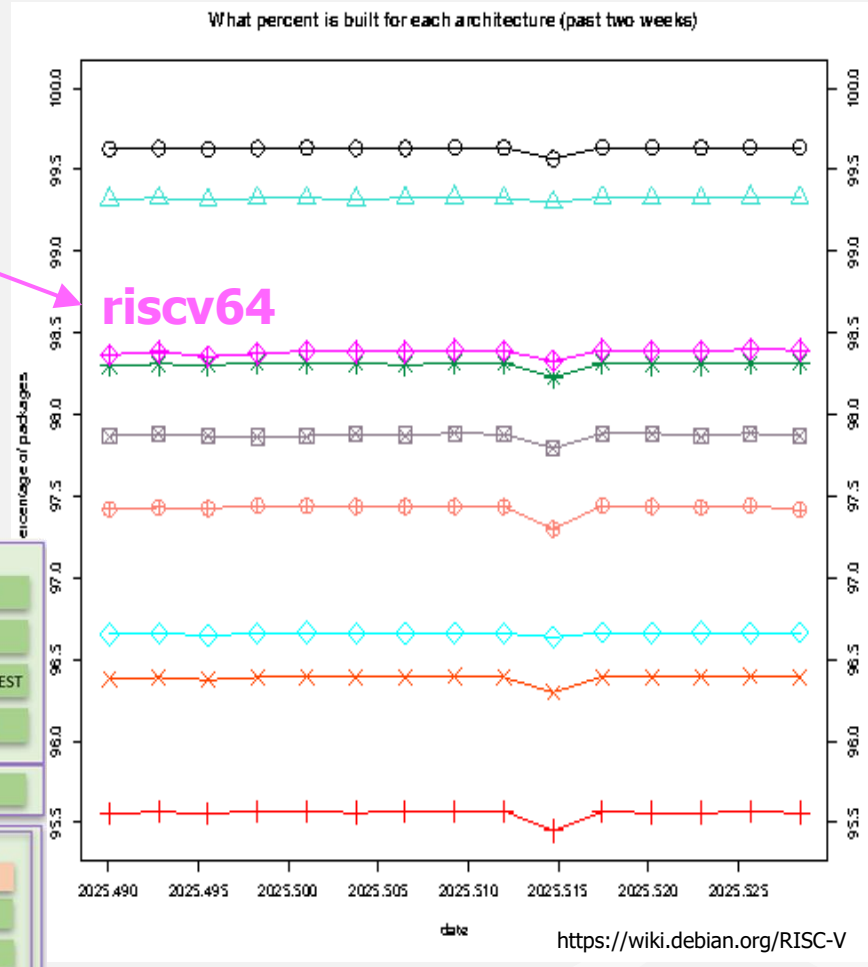
AMD, AKEANA, ESWIN, Canonical, ByteDance, ISCAS, MICROCHIP, SPACEMIT, tenstorrent, ZTE

Taking RISC-V® Mainstream

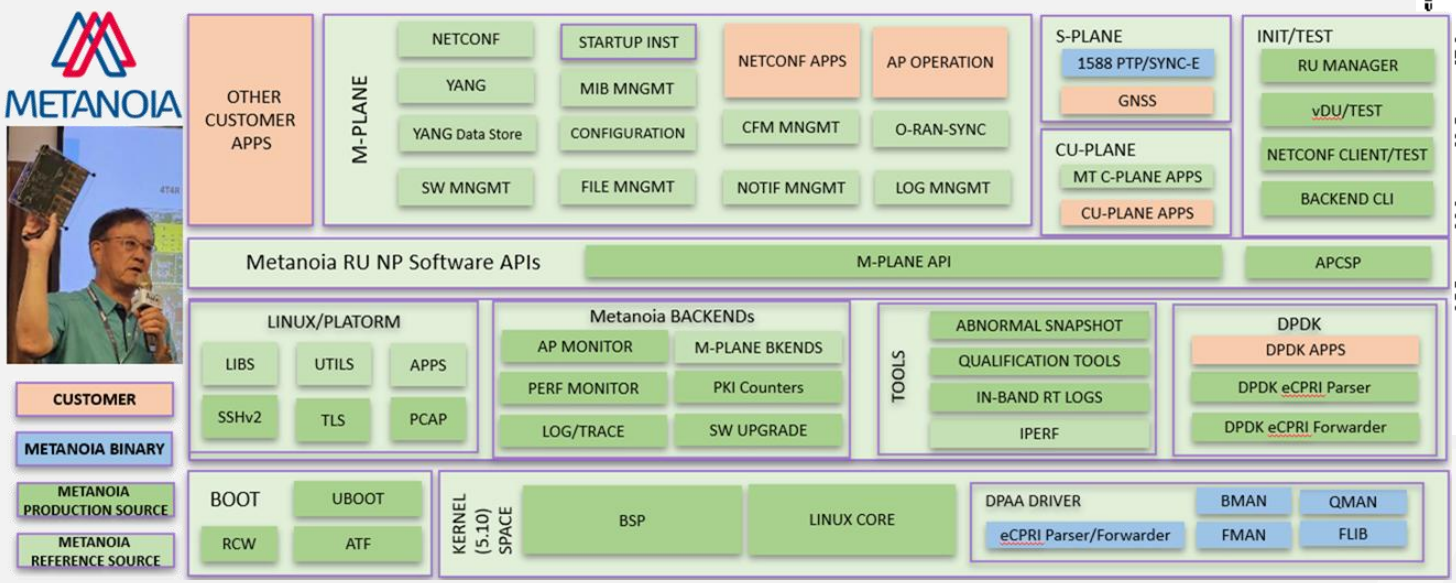
➔Java 22/21 supported: used in Cloud and IOT devices.

# RISC-V Maturity on Open-Source

- **Debian** regularly builds **> 64K packages**
  - **RISC-V**'s successful build rate: 98.4% (#3)

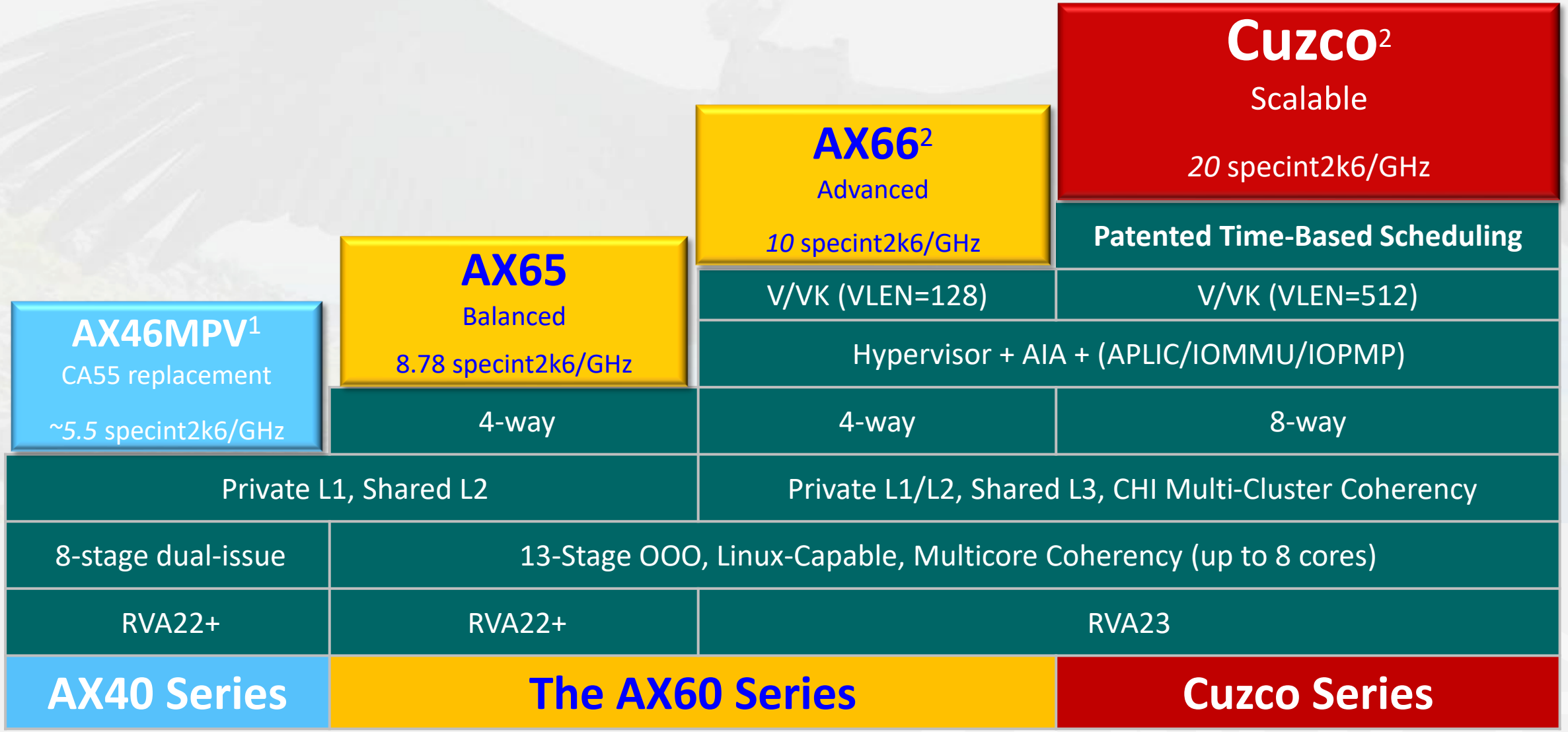



- **Metanoia's** 5G O-RAN SW Architecture
  - Modularized & Structured Semi-Turnkey with Full Open Source Code Release




Taking RISC-V® Mainstream

# Andes Processors for Intelligent General Computing

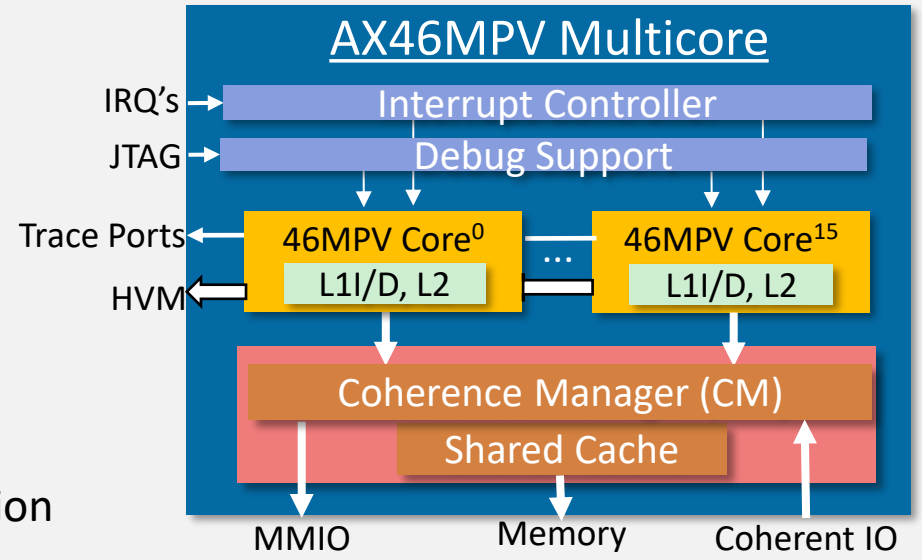


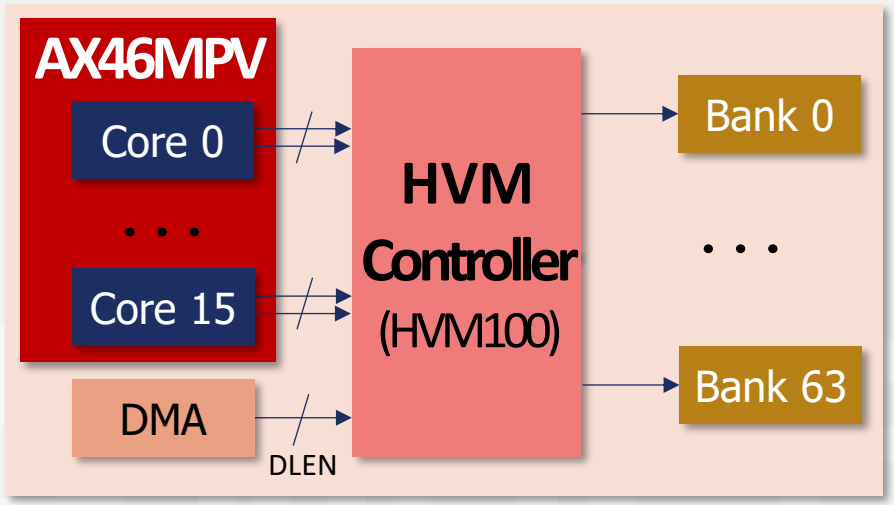

1: recently released; 2: to be released soon

# AX46MPV: Powerful Compute and Efficient Control

- RVA22+ with AIA and SV38/48/57
- **Dual issue for most vector/scalar instructions**
- Vector Processing Unit (VPU):
  - VLEN/DLEN: 128~1024, data formats: int4-int64, bf16/fp16-64
  - Enhanced ReductionSum
- Multicore complex: Up to 16 cores
- **Boosted memory performance:**
  - **Dual-issue load/store**, any load/store and vector-scalar combination
  - **Strong load/store outstanding** capability
  - **HVM** (High-speed Vector Memory) interface:
    - Capable of multiple outstanding requests with OOO return
    - HVM controller: up to 16 cores and 64 banks

- **Performance over AX45MPV:**
  - Specint2006: + ~18% (5.65)[1]
  - Vector libraries[2]: > 2x for key functions, +20% on average
  - Bandwidth[3]: +40%

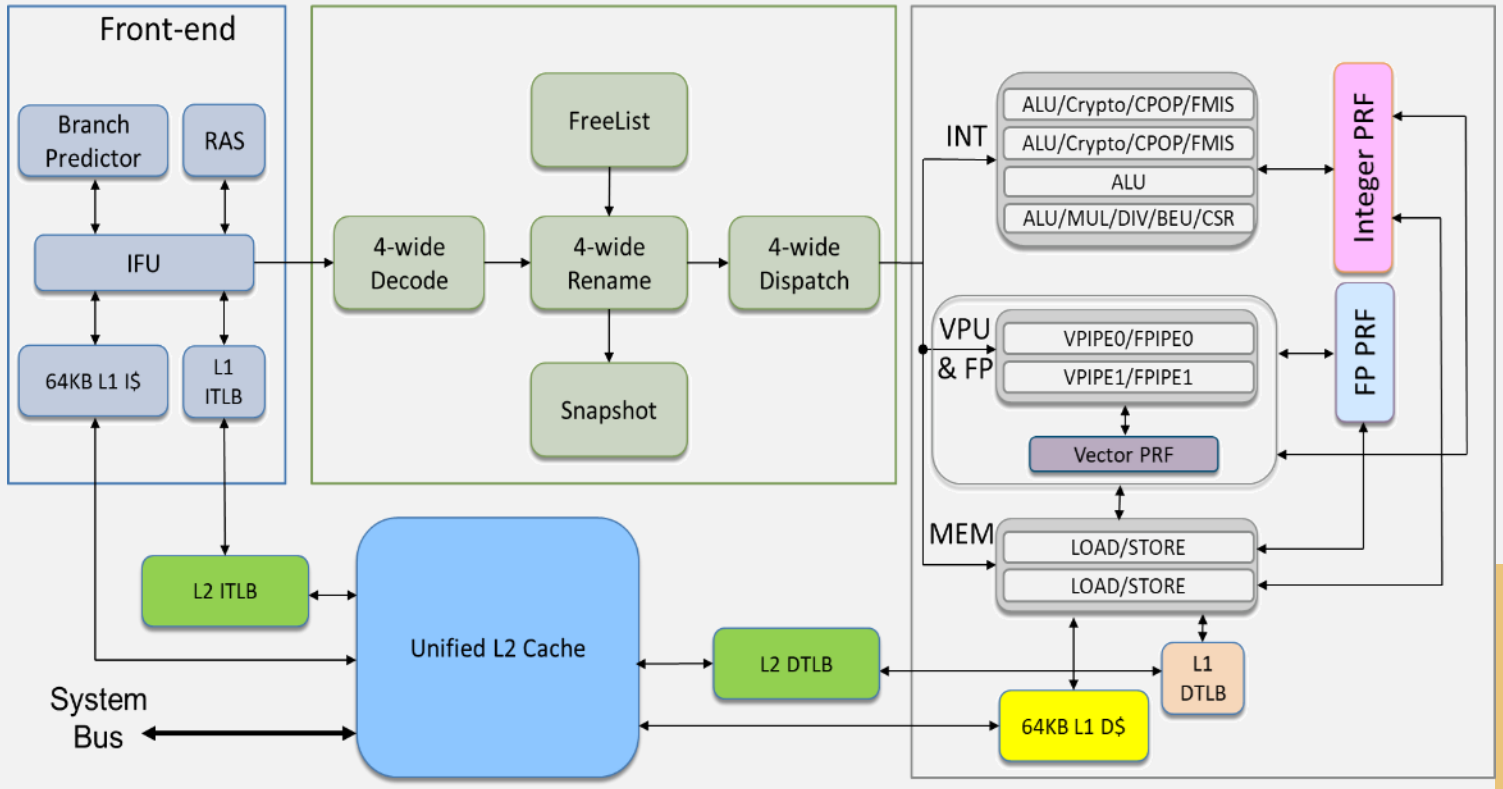



*Note: 1) VLEN 512b, 8 MB shared cache; 2) inc. libvec and libnn; 3) measured by tinymembench*

# AX66: Mid-Range Application Processor

- Built on the success of AX65
- **RVA23 compliant**
- **Dual vector pipes** with **VLEN=128**
- Frontend Decode: 4-Wide
- Reorder Buffer (ROB): 128-Entry
- Execution Pipelines: 8
- Branch Predictor: TAGE-L
- Multiprocessor: Up to 8 Cores
- **Shared L3 cache**: Up to 32 MB
  - Mostly exclusive with L1/L2
- Memory Port, IOCP, MMIO:
  - 128/256-bit AXI4
- **IOMMU, APLIC** and *CHI* support

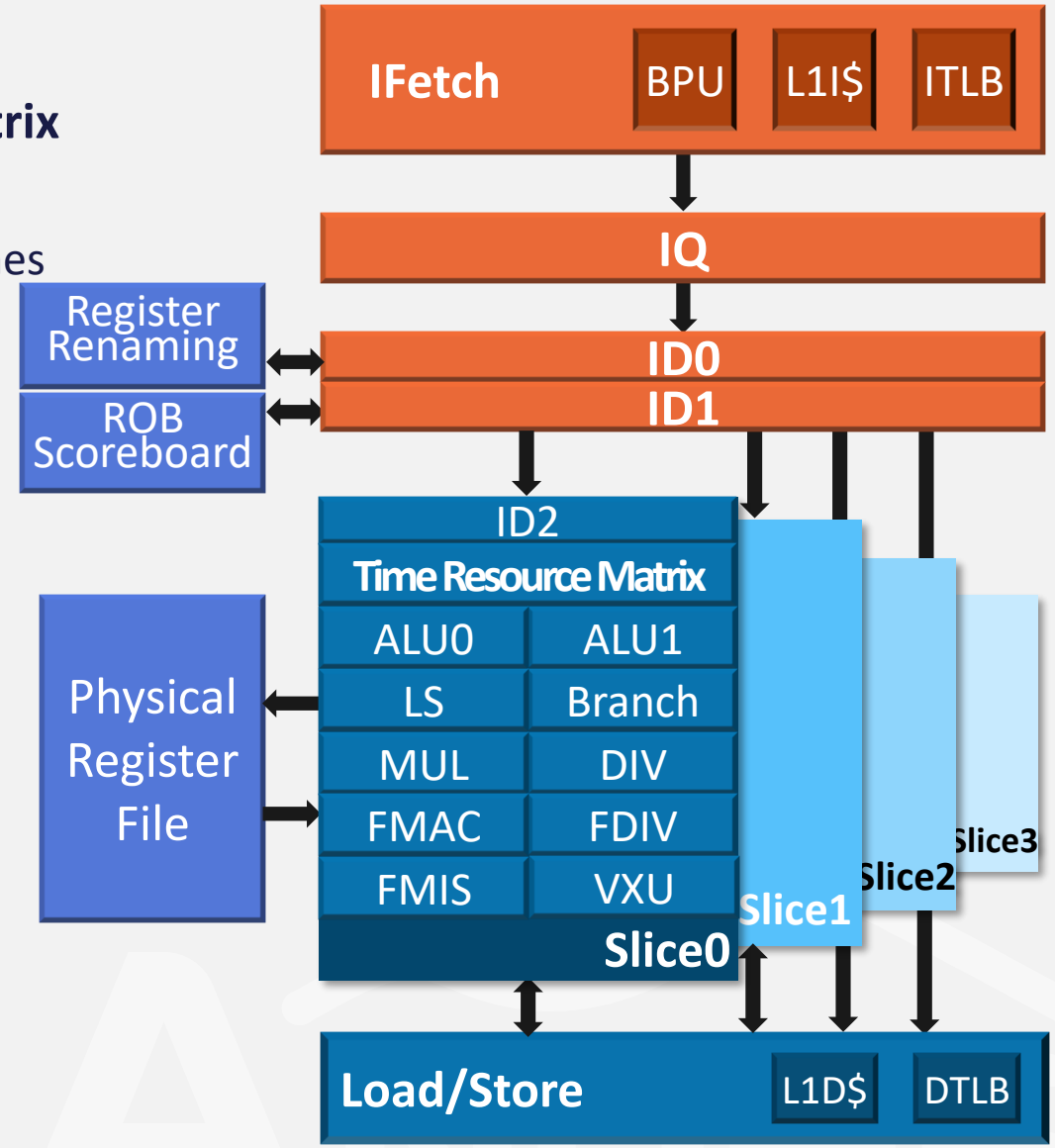



- **Vector performance**: (vs. scalar ISA)
  - Libnn: 9.6x speedup on average
  - Libvec: 3.6x speedup on average
  - Vector Crypto: 4.7/10.5/6.4x for SHA-256/AES-128/SM4
  - Bandwidth (tinymembench): + 25%

Taking RISC-V® Mainstream

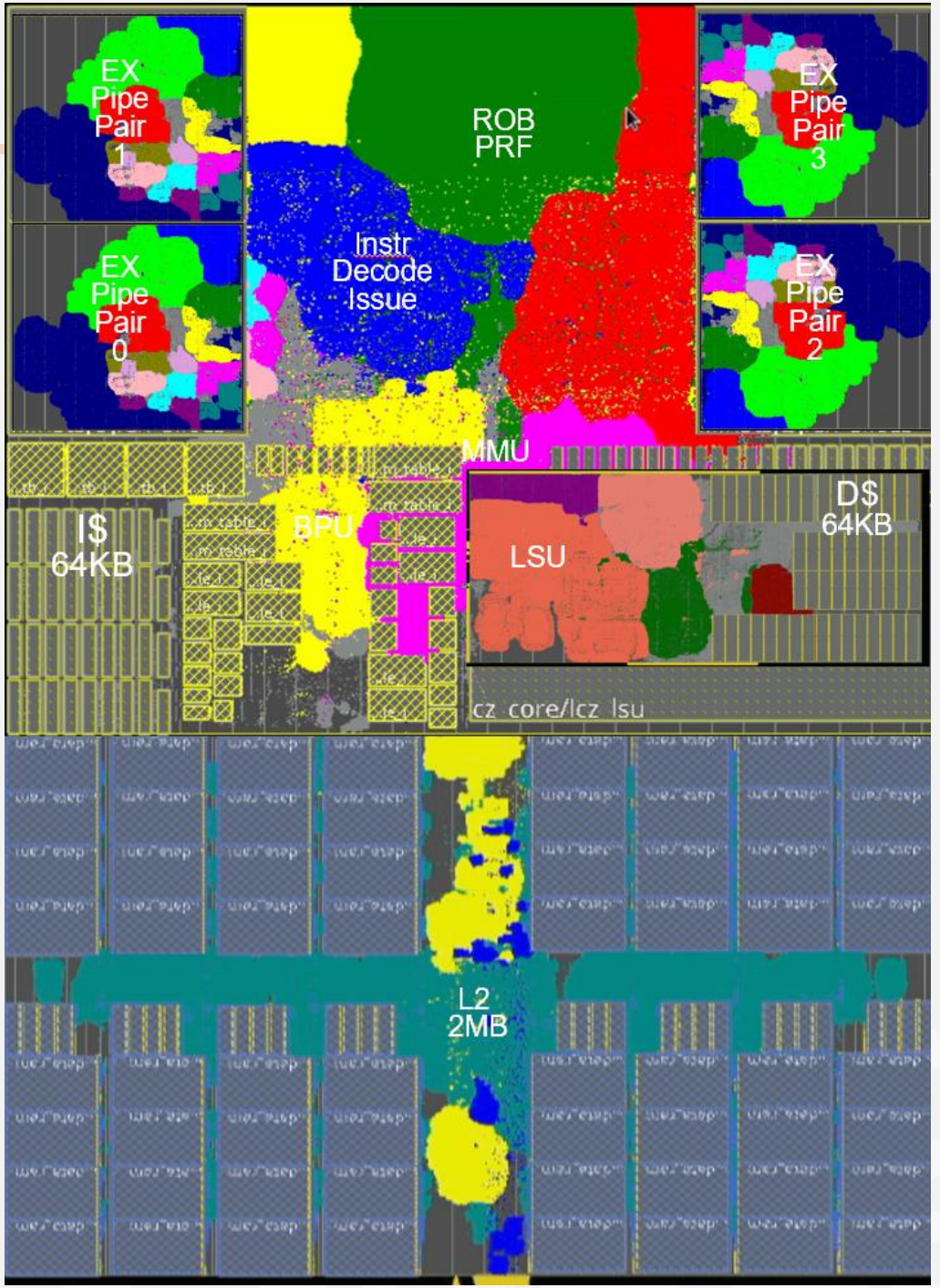
# Cuzco: High-End Application Processor

- RVA23 compliant
- Patented **time-based scheduling** using **Time Resource Matrix**
  - Schedule instruction issue cycles after decoding
  - Reduce logic complexity and dynamic power for wide machines
- Decode: **8-Wide** (and **6-Wide**)
- Reorder Buffer (ROB): 256 Entries
- **Execution Pipelines: 8 (2 per slice)**
- Branch Prediction: TAGE-L & Tournament, 2-level BTB
- L2 TLB: 1K/2K/4K, 4-Way
- Private L1$: I/D, 64KB, 4-way
- Private L2$: Up to 8MB
- Shared L3$: Up to 256MB (mostly exclusive with L2/L1)
- Multiprocessor: up to **8 Cores**
- Memory Port and IOCP: **CHI**, 256/512 bits
- MMIO: 256 bits



# Cuzco Early Implementation

- At 5nm
- CPU configuration
  - 8 execution pipelines
  - 7M gates
- L2 configuration
  - 4.5M gates (for 2MB config)
- Target: 2.5GHz
- Current specint2006[1]: ~18/GHz




*Note 1: RVV isn't used yet.*

# AndesAIRE™ "AI Runs Everywhere" end-to-end solutions

**NN models** — PyTorch, ONNX, TensorFlow Lite, TensorFlow

## AndeSight™ IDE

- GCC/LLVM Toolchains
- AndeSoft™ Vector/DSP Lib
- Build, debug, deploy, profile
- Analy
- RTOS
- Devic
- Samp
- Simu
- Documentation

generates files for LLVM to recognize new instructions

**AndesClarity™**
Pipeline Analyzer/Visualize

**ACE/COPILOT**

## AndesAIRE™ Software

### AndesAIRE™ NN SDK

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

**C Code Template**
- AnDLA command image
- AnDLA driver/runtime
- NN Library API

**TFL Models**
- .tflite file

AndesAIRE™ **NN Library** (>250 RVV functions)

AndesAIRE™ **XNNPACK** (>200 RVV functions)

## AI Compilers

OpenXLA, MLIR, tvm

**Andes LLVM Enablement**

LLVM Compiler Infrastructure

## AI Interpreters

TensorFlow Lite, TensorFlow Lite

PyTorch ExecuTorch

ONNX RUNTIME

LLaMA

| Host Processor | Compute Acceleration | Domain Extension | NPU Accelerator | SoC Hardwired Engines |
|---|---|---|---|---|
| Cuzco/AX60/AX40 series | Vector: 27V, 45V, 46V<br>DSP/SIMD: D23, D25F, D45 | ACE-scalar<br>ACE-RVV | AnDLA™ I350/370 | |

Note 1: RVV with VLEN 512 over without RVV

: Andes-enhanced open source

Taking RISC-V® Mainstream

# AndesAIRE™ "AI Runs Everywhere" end-to-end solutions

**NN models** — PyTorch, ONNX, TensorFlow Lite, TensorFlow

## AndeSight™ IDE

- GCC/LLVM Toolchains
- AndeSoft™ Vector/DSP Lib
- Build, debug, deploy, profile
- Analysis and tuning
- RTOS & Linux
- Device drivers
- Sample codes
- Simulator
- Documentation

**AndesClarity™**
Pipeline Analyzer/Visualizer

**ACE/COPILOT**

## AndesAIRE™ Software

### AndesAIRE™ NN SDK

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

| Models (Q4, running tg128) | Speedup[1] on AX45MPV |
|---|---|
| TinyLLaMA 1.1B | 26.70 x |
| Gemma 3 1B | 31.44 x |
| DeepSeek R1 distill Qwen 1.5B | 32.41 x |

AndesAIRE™ **NN Library** (>250 RVV functions)

AndesAIRE™ **XNNPACK** (>200 RVV functions)

## AI Compilers

OpenXLA, MLIR, tvm

**Andes LLVM Enablement**

LLVM Compiler Infrastructure

## AI Interpreters

TensorFlow Lite, PyTorch ExecuTorch, ONNX Runtime, LLaMA.cpp

Host Processor
Cuzco/AX60/AX40 series

Compute Acceleration
Vector: 27V, 45V, 46V
DSP/SIMD: D23, D25F, D45

Domain Extension
ACE-scalar
ACE-RVV

NPU Accelerator
AnDLA™ I350/370

SoC
Hardwired
Engines

Taking RISC-V® Mainstream



Note 1: RVV with VLEN 512 over without RVV

: Andes-enhanced open source

# General Computing: Rich OS Support

## From RISC-V community –

- **RISC-V spec:** RVA22/23 Profiles and SoC HW/SW Platforms
- **Linux Distro/Build Systems for RISC-V**
  (*: Andes verified)


## Additional support by Andes –

- **Linux kernel: upstream compatible**
  - Device drivers for Andes AE350 platform
  - **LTS** (Long-Term Support) kernels verified with **LTP** (Linux Test Projects)
  - Andes-enhanced/created features
    - **Upstreamed**: strace/ftrace, Perf, Kernel module, HIGHMEM, CPU hotplug
    - Ongoing: Suspend-to-RAM, Andes PowerBrake support
- **Bootloaders:** U-Boot, U-Boot-SPL, OpenSBI and BBL
- **RTOS:**
  - FreeRTOS (RV32/64 UP) aws qualified device Amazon FreeRTOS, Zephyr (RV32/64 UP/SMP)
  - Several commercial RTOS'es (Thread-X, RT-Thread, etc.)


Taking RISC-V® Mainstream

# Closing Remarks

- **Andes has been leading the RISC-V IP shipment with**
  - Rich portfolios from processors to software and development tools
- **Andes latest processors offer**
  - Powerful compute and efficient control in AX46MPV
  - A wide performance range of RVA23 application processors from AX66 to Cuzco
- **RISC-V ecosystem is ready** for the emerging market of **Intelligent General Computing**



- **Andes is well-positioned** with our latest offerings.

Taking RISC-V® Mainstream

# Thank You !!

Taking RISC-V® Mainstream