Transforming Modern Computing With RISC-V

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

Andes Technology Corporation

Who We Are

- CPU — Pure-play CPU IP Vendor
- 20-year-old Public Company
- Products Used in Endpoints, Edge and Cloud

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

30+ V5 (RISC-V) AndesCore™

400+ Licensees

500~ Employees

100K+ AndeSight™ IDE users

16Bn+ Units of Andes-Embedded™ SoC

ANDES TECHNOLOGY Subject to change without notice
copyright © 2021-2024 Andes Technology

AndesCore™ Lineup

C					X3/X4
AX60 Series OOO Linux MP	AX63	AX65	AX66/AX67	AX60-SE	A72~A76
Categories	*Power-efficient*	*Balanced*	*Extended*	*FUSA*	
40 Series 8-stage Superscalar	N45, NX45	**AX45MPV**, A*46MP(V), AX47MPV		**D45-SE**	A53/55, R52/ R82, M7
		D45	A45(MP), AX45(MP)		
27 Series 5-stage MemBoost		NX27V	A27(L2), AX27(L2)		A5/7/35
25 Series 5-stage Fast&Compact	N25F, NX25F	D25F	A25(MP), AX25(MP)	**D25F-SE** N25F-SE	A5/7/35, R4/5, M4/33
Compact Series	N225	**D23**, D23V		D23-SE	M0/0+/3/33/4
Categories	*Embedded Control*	*Compute Acc.*	*Linux AP*	*FUSA*	*References*

Black: available now. Red: under development. Blue: future roadmap.

Note: roadmap subject to change without notice

Subject to change without notice
copyright © 2021-2024 Andes Technology

RISC-V Market Research by SHD Group

- SoC revenue: 2.8x in 2023, 15x from 2023 to 2030
- Andes has >30% RISC-V IP Market in 2023

Mobile

N25F
D25F
N45

Performance, code size

NOVATEK MEDIATEK

MPU/MCU/AIoT

D25F
D45
AX25MP
AX45MP

Storage

D23
N25F
N45
AX45MP

PHISON

Performance, bandwidth, real-time

5G Networks

N25F
A25
A45MP
AX45MP

N25F-SE, D25F-SE, AX45MPV

Large-Scale AI/ML

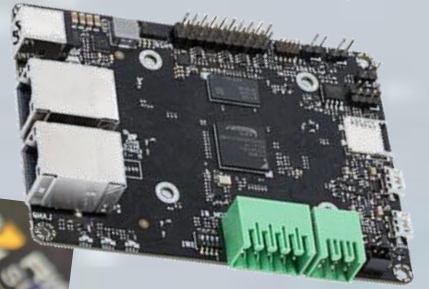
Acceleration, bandwidth, extensions

NX27V, AX25, NX45, AX45MP, AX45MPV, AX65

ANDES TECHNOLOGY
Subject to change without notice
copyright © 2021-2024 Andes Technology

AIOT and Wearable

- **Renesas Voice-Control ASSP Solution**
  - **R9A06G150** ASSP with Andes DSP-capable **D25F**, which speeds up the application by over 50%.
- **ASUS IoT Tinker V Single-Board Computer:**
  - Based on **Renesas** RZ/Five 1 GHz SoC with **Andes AX45MP**
  - Linux Debian and Yocto distro for for Industrial IoT and gateway



- **Smart watch** with long battery life:
  - D25 as main processor
  - Battery life:
    - 4 weeks for normal use
    - 2 weeks with 20km running

Powered by Andes D25F/AX45MP

Subject to change without notice
copyright © 2021-2024 Andes Technology

Enterprise Storage and Spherical Image Processor

- **Phison X1** (PS5020-E20) **Enterprise Storage**
  - Leveraging N25F with Andes Custom Extension™
    - "The ACE automation tool is very powerful in creating customized instructions that fits our exact needs", Vincent Cheng, VP of R&D of Phison
  - For AI, HPC, and Hyperscale Datacenters

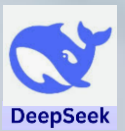
- **Aspeed AST1230** Multi-Cam Panorama Image Processor
  - Using N25F for 8K2K real-time 360° cameras with rich audio processing
  - For immersive applications such as video conferencing, virtual factory inspection/audit, shopping, touring, house showing, etc.

Powered by Andes N25F and ACE

Subject to change without notice
copyright © 2021-2024 Andes Technology

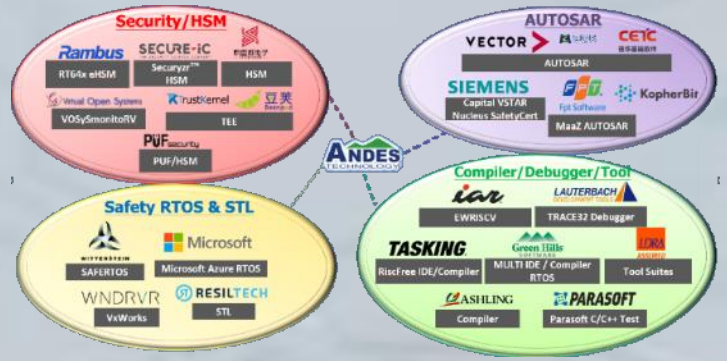
Key Computing Technologies Enabled by RISC-V

- **AI/ML Accelerations**
  - Vector processing, custom extensions, NPU, and E2E SW Stack



- **Embedded & Real-Time Processing**
  - Code size, device control, interrupts

- **Functional Safety and Security**
  - ISO 26262 fully compliant
  - Secure boot, TEE, control flow integrity



- **General-Purpose Application Processing**
  - Linux, Android, rich applications

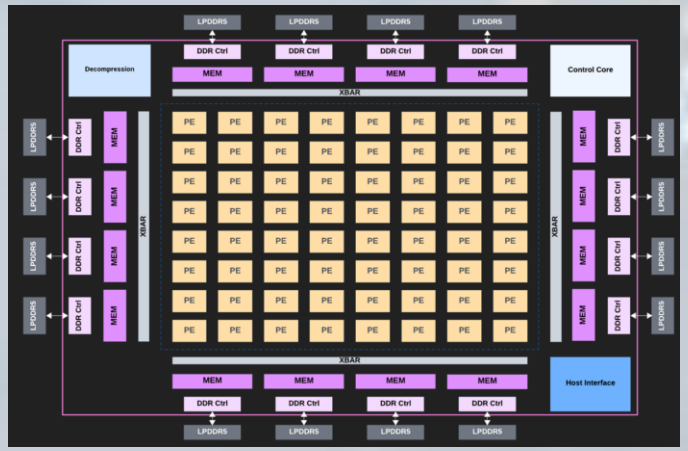
ANDES TECHNOLOGY
Subject to change without notice
copyright © 2021-2024 Andes Technology

Large-Scale AI/ML Accelerations

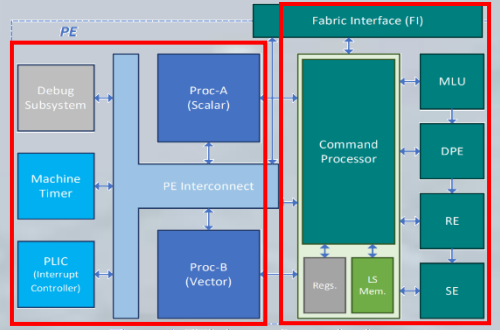
Subject to change without notice
copyright © 2021-2024 Andes Technology

Large-Scale AI/ML from Edge to Cloud

- **Popular SoC architectures – Sea of PE's**
  - Mesh-connected (like **Meta MTIA**) or multi-clustered
  - Core of the computations – Processing Elements (PE's)
- **Three tiers of accelerations in PE's**
  - Matrix Multiplications:
    - Hardwired solutions: NPU
    - 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)
- **Fast growing AI SW Stack**



Meta MTIA 2



Subject to change without notice
copyright © 2021-2024 Andes Technology

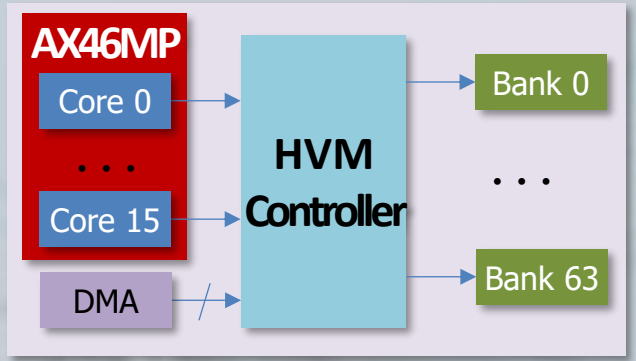
Andes New Vector Processor: AX46MPV

- **Built on the success of AX45MPV**
  - 8-core cluster with 1024 VLEN/DLEN
  - Dual issue for vector and scalar
- **Doubled core count 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: up to 16 cores and 64 banks
- **Four incarnations: AX46MP(V), A46MP(V)**

More features for AX47MPV
. CHI-based multi-cluster
. Latest RVA profile
. More FP data formats

	AX25-V100 (Meta MTIA-1)	NX27V	AX45MPV	AX46MPV*	AX47MPV*
					more features
				Matrix Extension	
			8-core cluster	16-core cluster with private L1/L2	
			HVM (High-speed Vector Memory)		
			ACE for RVV	ACE for RVV (advanced)	
ACE (Andes Automated Custom Extensions™)					
int4~64, fp16~64; bf16 (conv.)			+bf16 (full arithmetic)		+fp8
VLEN 128~512			+1024	+2048	
RVV 0.8		RVV 1.0			
5-stage single-issue			8-stage dual-issue		

* Future products subject to change



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

Subject to change without notice
copyright © 2021-2024 Andes Technology

AndesAIRE™ Andes AI Runs Everywhere SW 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**

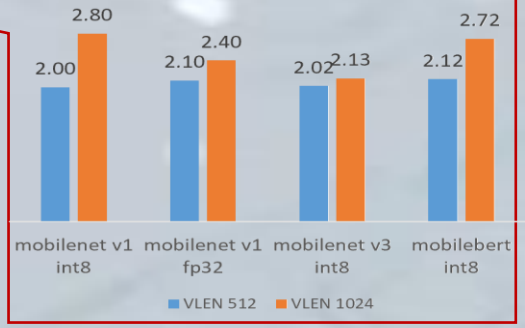
AndesAIRE™ **NN Library**

AI Compilers

OpenXLA IREE MLIR tvm LLVM

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

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



200 high-level functions for PyTorch/TFL

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

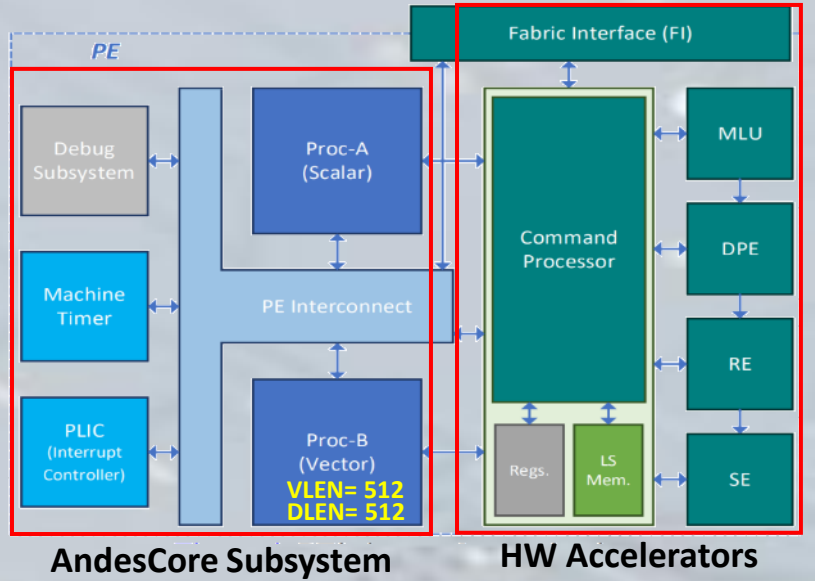
LLVM auto-vectorizes code generated by IREE/TVM/etc.

Subject to change without notice
copyright © 2021-2024 Andes Technology

MTIA: Meta Training/Inference Accelerator

- ISCA 2023 paper, "MTIA: First Generation Silicon Targeting Meta's Recommendation Systems"
- Proc-A/B: AX25-V100, an early version of Andes popular **NX27V vector processor**
- **Automated Custom Extensions (ACE)**: create new interfaces/registers/instructions
  - And auto-generate all the files necessary to enable LLVM

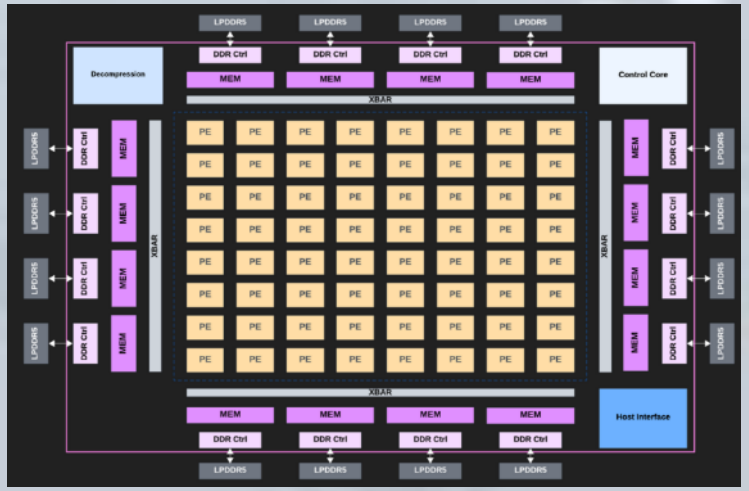
All photos: courtesy of ACM and Hot Chips

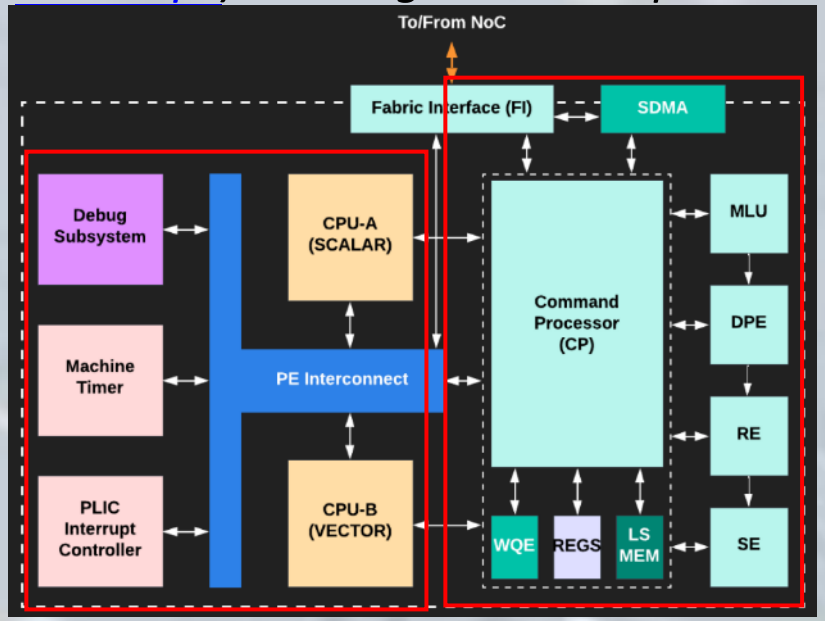


Subject to change without notice
copyright © 2021-2024 Andes Technology

MTIA: Meta Training/Inference Accelerator

- ISCA 2023 paper, "MTIA: First Generation Silicon Targeting Meta's Recommendation Systems"
- Proc-A/B: AX25-V100, an early version of Andes popular **NX27V vector processor**
- **Automated Custom Extensions (ACE)**: create new interfaces/registers/instructions
  - And all the files necessary to enable LLVM
- *Next generation MTIA (Meta blog and Hot Chips): Serving models in production*





All photos: courtesy of ACM and Hot Chips

Subject to change without notice
copyright © 2021-2024 Andes Technology

AX45MPV and ACE: RAIN AI's Adoption*

Leveraging Andes Custom Extensions (ACE) automated

RAIN

In-memory Compute Integration — ACE Streaming Port

Introduce 2nd vector load/store unit; direct access to VRF

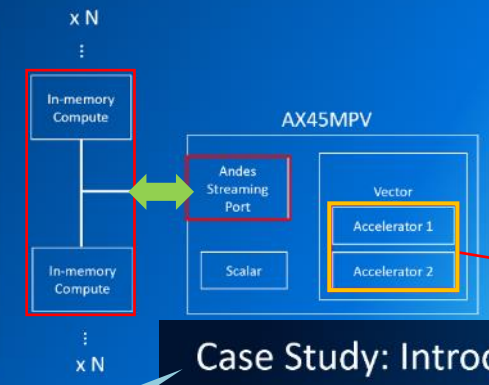
Leverage custom addressing and address control registers to control a plurality of in-memory compute blocks with different data sizes

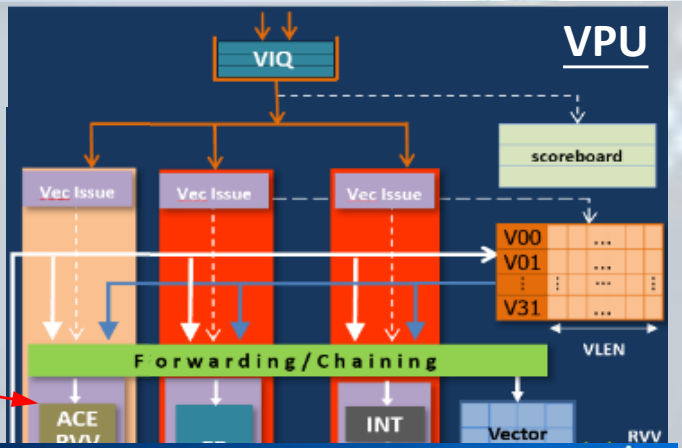
Non-linear Operator Integration — ACE RVV

Integrate non-linear approximations as a functional unit in vector pipeline with direct access to VRF

Introduce vector instructions that look and feel familiar

Support for LMUL, pipelined multi cycle operations





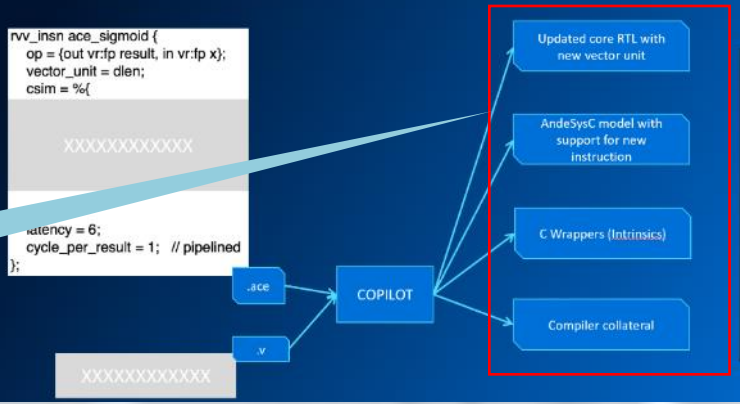
Plus **Softmax**, **SiLu**, and other ACE instructions.

COPILOT auto-updates RTL, compiler, debugger and SC simulator

Case Study: Introducing a Sigmoid Instruction

RAIN

```
rvv_insn ace_sigmoid {
  op = {out vr:fp result, in vr:fp x};
  vector_unit = dlen;
  csim = %{
XXXXXXXXXXXX
latency = 6;
  cycle_per_result = 1;  // pipelined
};
```



Sample wrappers for our sigmoid instruction

- Presented in Andes RISC-V CON San Jose, June 2024
  https://youtube.com/live/TpGFCTu_OFw

ANDES TECHNOLOGY
Subject to change without notice
copyright © 2021-2024 Andes Technology

RISC-V Enabled Innovations for Large-Scale AI/ML

Based on 27V/45MPV/46MPV (RVV) and AX65/NX45/AX25

With **Andes Automated Custom Extensions™ (ACE)**

- **AI Accelerators Using SRAM-based Compute-In-Memory:**
  - AXELERA ARTIFICIAL INTELLIGENCE
  - 后摩智 HOUMO
  - RAIN
  - TetraMem Accelerate the World®
  - FRACTILE
  - Company-X
- **AI Accelerator Using Photonics**
  - LIGHTELLIGENCE
- **AI Accelerator for Cloud Service**
  - SAPEON
- **AI SoC for Servers**
  - Rivos
- **AI Accelerator for ADAS**

ANDES TECHNOLOGY
Subject to change without notice
copyright © 2021-2024 Andes Technology

Deepseek Running on the 45-Series Platform

```
main: interactive mode on.
sampler seed: 0
sampler params:
        repeat_last_n = 64, repeat_penalty = 1.200, frequency_penalty = 0.000, presence_penalty = 0.000
        dry_multiplier = 0.000, dry_base = 1.750, dry_allowed_length = 2, dry_penalty_last_n = 2048
        top_k = 40, top_p = 0.950, min_p = 0.050, xtc_probability = 0.000, xtc_threshold = 0.100, typical_p = 1.000, temp = 0.100
        mirostat = 0, mirostat_lr = 0.100, mirostat_ent = 5.000
sampler chain: logits -> logit-bias -> penalties -> dry -> top-k -> typical -> top-p -> min-p -> xtc -> temp-ext -> dist
generate: n_ctx = 2048, n_batch = 512, n_predict = -1, n_keep = 6

== Running in interactive mode. ==
 - Press Ctrl+C to interject at any time.
 - Press Return to return control to the AI.
 - To return control without starting a new line, end your input with '/'.
 - If you want to submit another line, end your input with '\'.

< | User | > Please write a 200-word short article that Andes technology has successfully deploy the DeepSeek R1 distill model on the
AX45MPV vector CPU processor platform by Andes Technology for embedded AI applications < | Assistant | >


>
```

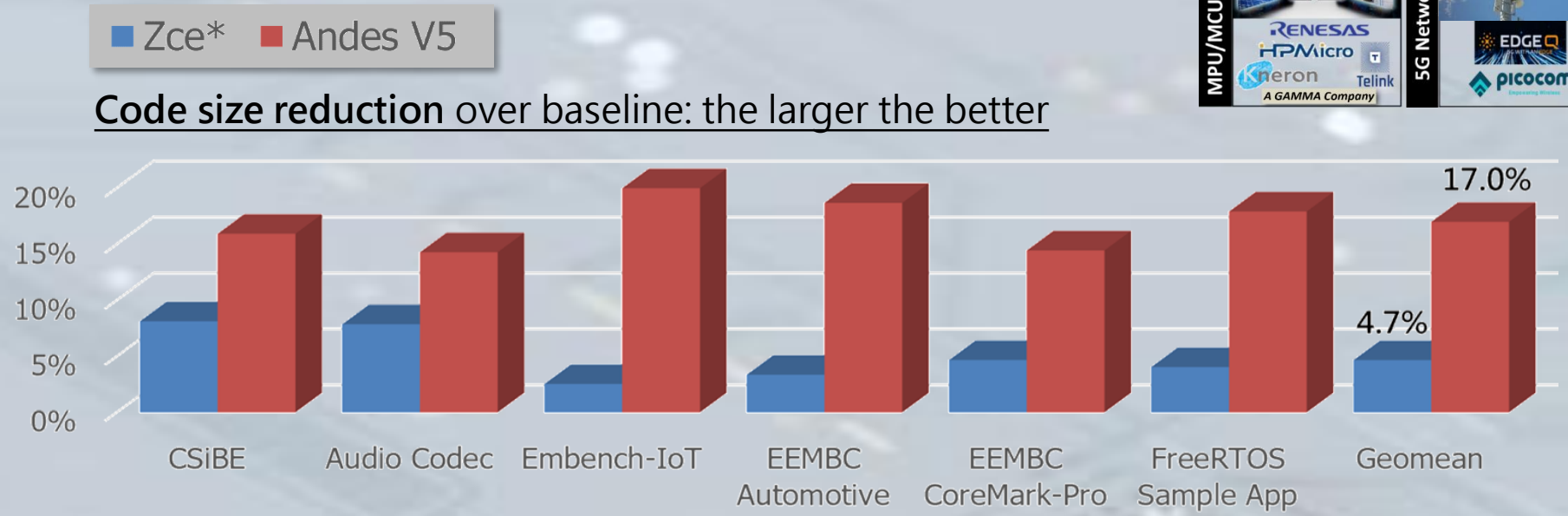
Subject to change without notice
copyright © 2021-2024 Andes Technology

Embedded and Real-Time Processing

copyright © 2021-2024 Andes Technology

Embedded and Real-Time Processing

- **Compact code size:**
  - **Zce** brings 4.7% reduction over IMC.
  - New Task Group **Scalar Efficiency** is looking for more reduction
  - Andes V5 CoDense™: reduce another 11-15% over Zce

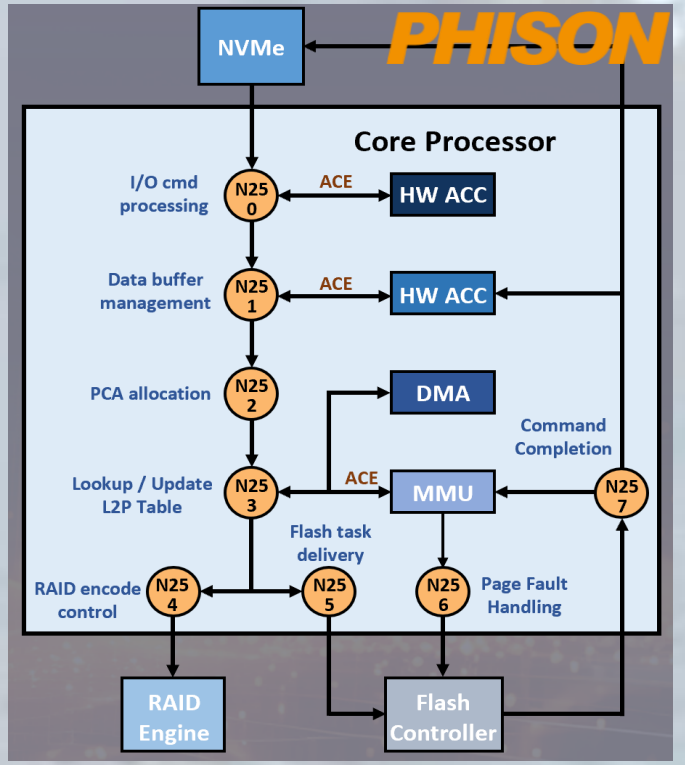


Subject to change without notice
copyright © 2021-2024 Andes Technology

Embedded and Real-Time Processing

- **Efficient control and accesses of HW engines**
  - **Multiple outstanding accesses** for uncached and device space
  - **ACE instructions** with custom ports for direct accesses and control

  ➔Supported by AndesCore families
- **Interrupt handling:**
  - Single cores: **CLIC**
  - Multicores: **PLIC + Andes vectoring/priority preemption**
  - Shadow registers: D23
- **Quality of Service:**
  - **CBQRI**: Capacity/Bandwidth QoS Register Interface
  - 4 cores with 16-way shared cache: 30% boost in copy bandwidth



Subject to change without notice
copyright © 2021-2024 Andes Technology

Safety and Security

(will focus on Security in this talk)

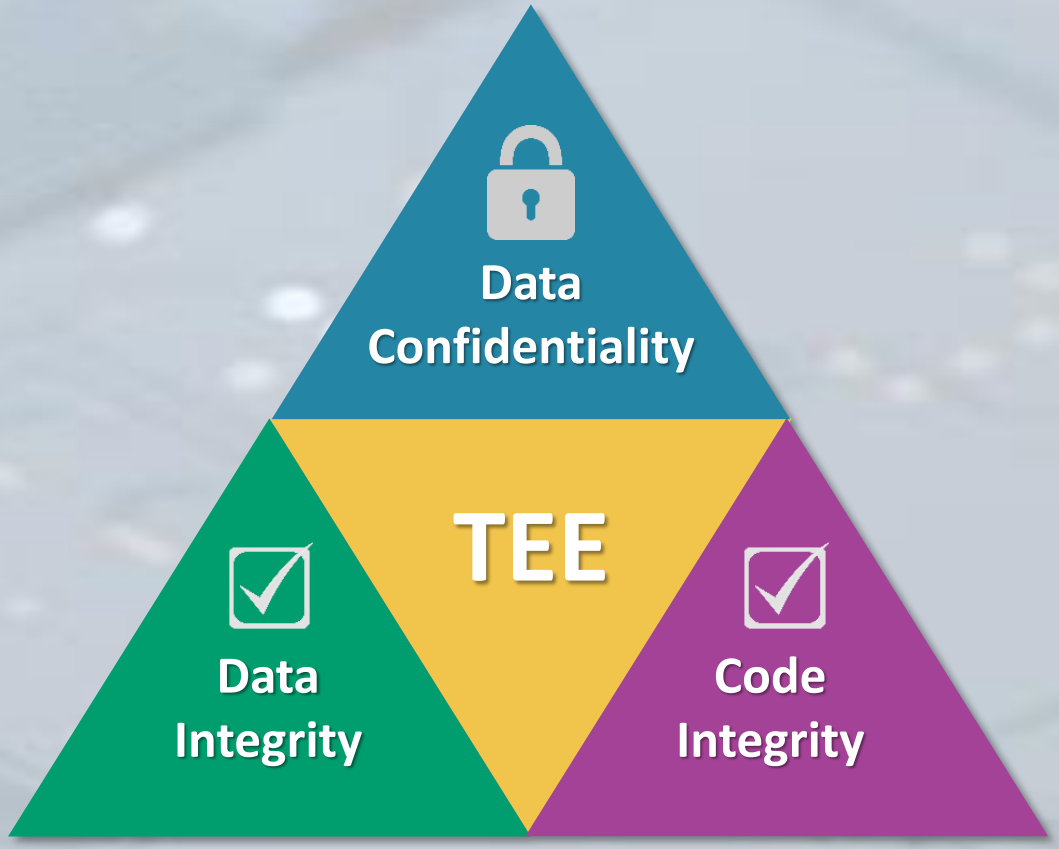
Subject to change without notice
copyright © 2021-2024 Andes Technology

Security TG/SIG in RVI

- **CFI (Control Flow Integrity, shadow stacks, landing pads)**
- **Memory Tagging (Pointer Masking, Memory Tagging, CHERI)**
- **Table-based Protection/Attributes**
  - Supervisor Domain Access Protection
  - SPMP/hgPMP/vSPMP
  - PMP-based Memory Types
  - IOPMP
- **Countermeasure for Side-Channel Attacks:**
  - Address-Independent Latency of User-Mode Faults to Supervisor Addresses
  - Timing Fences
- **MISC:**
  - External Debug Security
  - HFI (Hardware Fault Isolation)
- **More**

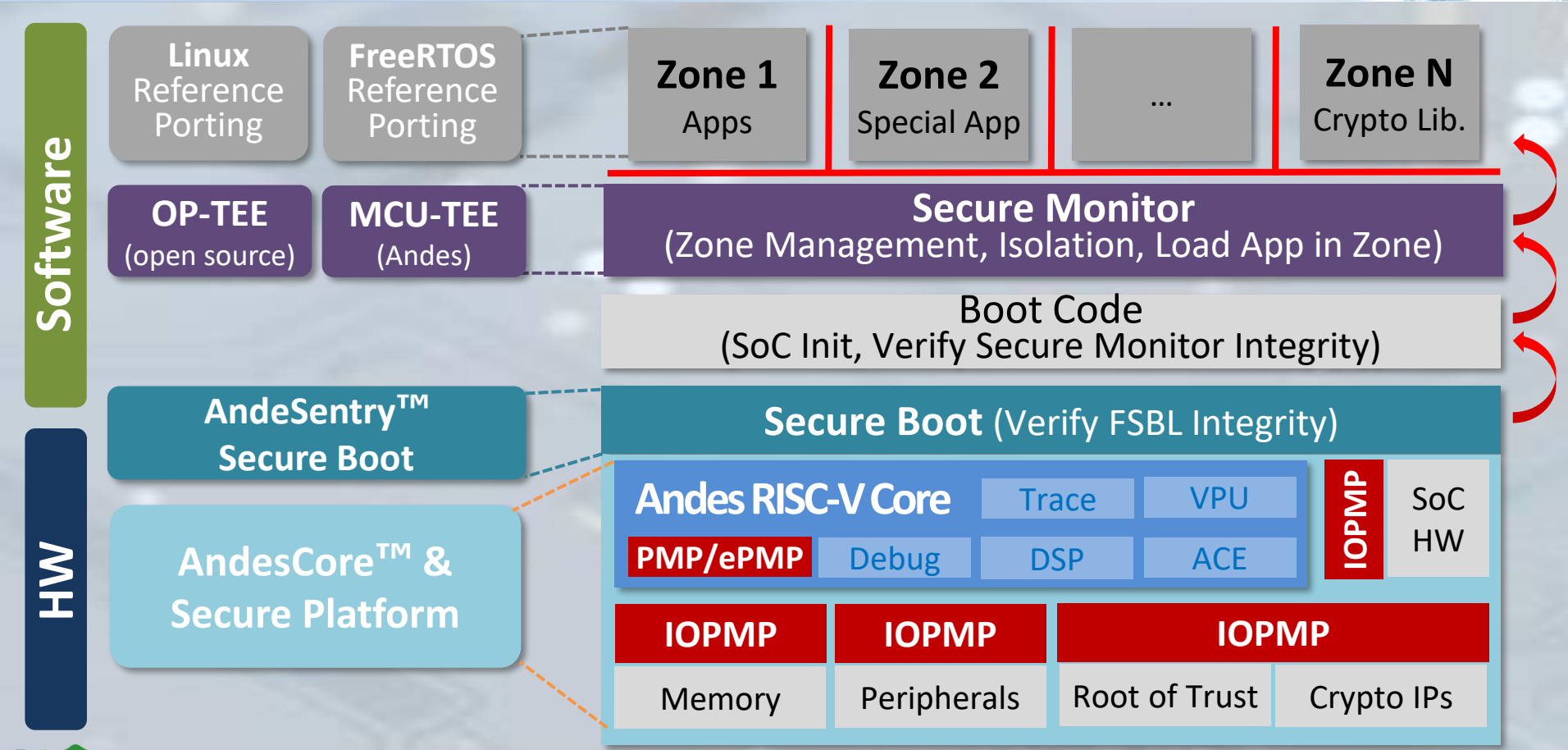
Subject to change without notice
copyright © 2021-2024 Andes Technology

Trusted Execution Environment



Subject to change without notice
copyright © 2021-2024 Andes Technology

AndeSentry™ TEE Solutions



Subject to change without notice
copyright © 2021-2024 Andes Technology

GP Application Processing

copyright © 2021-2024 Andes Technology

General-Purpose Application Processing

- **Applications: control/main processors for**
  - Personal computing to Servers
  - AI/ML, Multimedia, Networking, Industrial
- **Requirements:**
  - High-performance Linux processors
  - Ecosystem support:
    - RVA22/RVA23 profile
    - Linux distros
    - Android AOSP for Multimedia
    - OpenWrt for Networking
    - Debian for Industrial
  - ➔ RISE projects

debian

openSUSE

yocto PROJECT

fedora

OpenWrt
Wireless Freedom

RISE
RISC-V Software Ecosystem

Subject to change without notice
copyright © 2021-2024 Andes Technology

Roadmap for Andes Application Processors

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

Android Base

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
	RVA22+	RVA23	RVA23+	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
		Private L1/L2, CHI Multi-Cluster Coherency		16-stage 6/8-way OOO with patented Time-Based Scheduling
13-Stage, 4-way OOO, Linux-Capable, Multicore Coherency, Up to 8 Cores/Cluster				
The AX60 Series				**Cuzco Series**

** Future products subject to change*

Subject to change without notice
copyright © 2021-2024 Andes Technology

Concluding Remarks

copyright © 2021-2024 Andes Technology

RISC-V for All Computing Devices

- **30+ AndesCores: tiny to RVA23-capable**
  - Support RISC-V standard extensions
  - Automat RISC-V custom extensions with **ACE**
- **Upcoming Andes Products:**
  - Vector processor: **AX46MPV**
  - Application processors: **AX66 and Cuzco (RVA23)**
  - ASIL-D core: **D23-SE**
  - **IOPMP-based Security Platform and AndeSentry**
  - **AndeSight** IDE **and AndesAIRE** NNSDK
- **Enabling key computing technologies:**
  - AI/ML Accelerations
  - Embedded and Real-time
  - Functional Safety and Security
  - GP Application Processing

AX60 Series OOO Linux MP	AX63	AX65	AX66/AX67	AX60-SE
Categories	*Power-efficient*	*Balanced*	*Extended*	*FUSA*
40 Series 8-stage Superscalar	N45, NX45	**AX45MPV**, A*46MP(V), AX47MPV		**D45-SE**
		D45	A45(MP), AX45(MP)	
27 Series 5-stage MemBoost		NX27V	A27(L2), AX27(L2)	
25 Series 5-stage Fast&Compact	N25F, NX25F	D25F	A25(MP), AX25(MP)	**D25F-SE** N25F-SE
Compact Series	N225	**D23**, D23V		D23-SE
Categories	*Embedded Control*	*Compute Acc.*	*Linux AP*	*FUSA*

Black: available now. Red: under development. Blue: future roadmap.

Andes Quick Facts

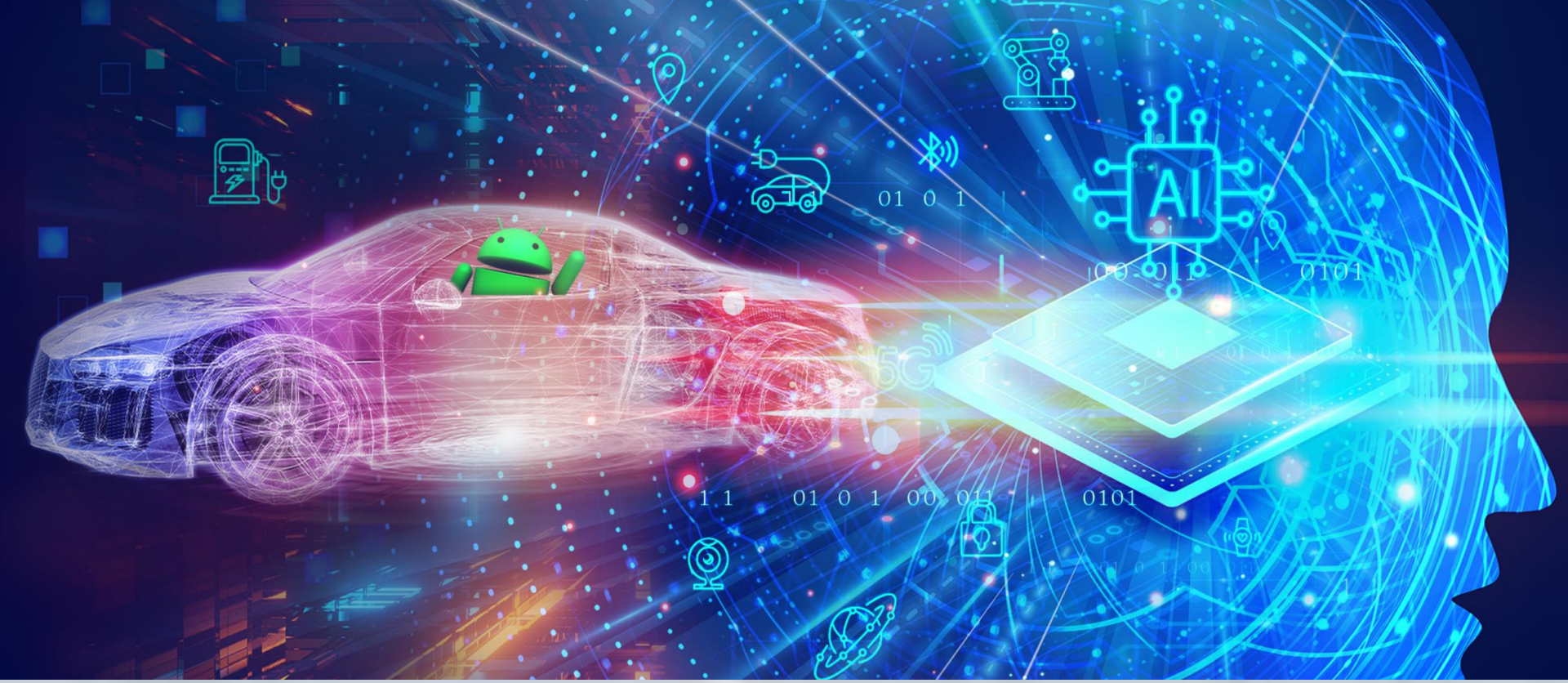
~20 Years/Public	**30+** RISC-V Cores	**400+** Licensees	**500+** Employees

100K+ AndeSight™ IDE Users

16 Bn+ Andes-Embedded™ SoC

- **Andes helps bring the RISC-V vision to reality, billion SoC's at a time !**

Subject to change without notice
copyright © 2021-2024 Andes Technology



Thank You !!

ANDES TECHNOLOGY

Taking RISC-V Mainstream

ANDES TECHNOLOGY
Subject to change without notice
copyright © 2021-2024 Andes Technology