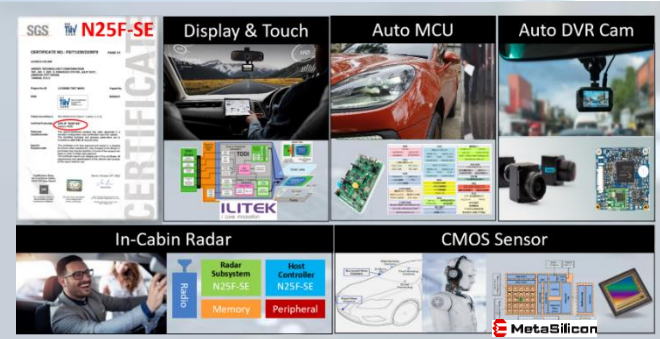
RISC-V for Intelligent Edge Application Processing

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

RISC-V Market Research by SHD Group

Total RISC-V SoC Market

- $6.1B in 2023: 276.8% growth over 2022
- $92.7B by 2030: a CAGR of 47.4%



N25F-SE, AX45MPV

Andes has >30% of RISC-V IP Market

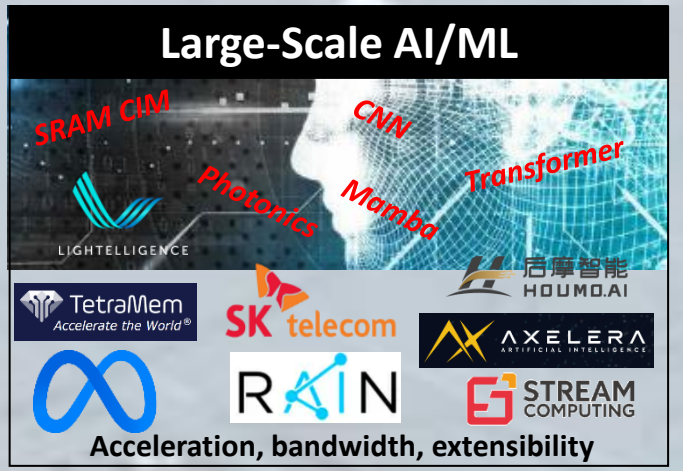
- Applications: Endpoint, Edge, Cloud

N25F, N45

D25F, D45, AX25MP, AX45MP

D23, N25F, N45, AX45MP

N25F, A25, A45MP, AX45MP

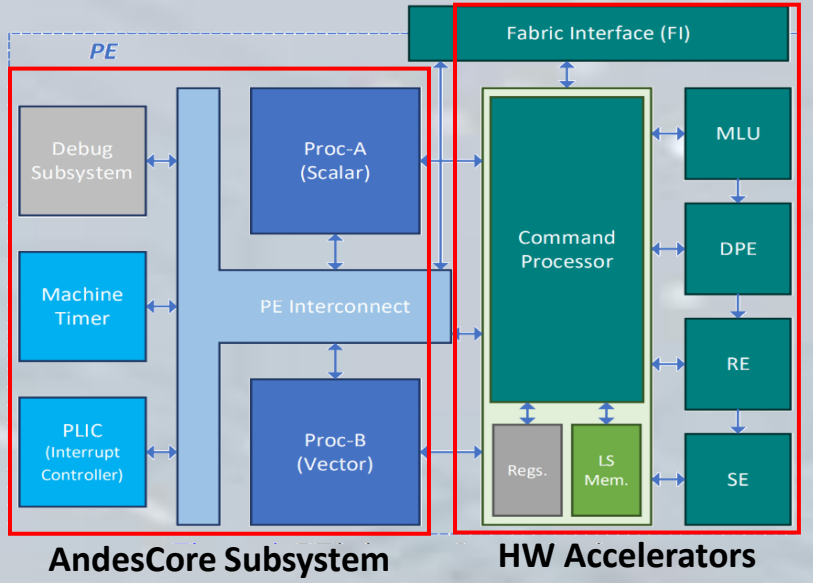


NX27V, AX25, AX45MPV, AX65

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 the popular **NX27V vector processor**
- **Custom extensions (ACE)**: create new interfaces/instructions/registers
- *Next generation MTIA (Meta blog): Serving models in production*

All photos: courtesy of ACM



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

Based on NX27V/AX45MPV (RVV), AX65 (OOO), and AX25

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

- **AI Accelerators Using SRAM-based Compute-In-Memory:**

AXELERA ARTIFICIAL INTELLIGENCE

后摩智能 HOUMO.AI

RAIN

TetraMem Accelerate the World®

- **AI Accelerator Using Photonics**

LIGHTELLIGENCE

- **AI Accelerator for Cloud Service**

SAPEON

- **AI Accelerator for ADAS**

Intelligent Edge Application Processing (IEAP)

- **Market projection: Edge AI will be bigger than Cloud AI**
- **Definition of IEAP:**
  - Not Cloud AI/ML, Not AI PC
  - Including application handling and AI/ML accelerations
- **AI/ML for Edge Applications: On-device Intelligence**
  - Intelligent Edge Servers with domain knowledge (e.g. legal, medical, or enterprise)
  - Smart Factory/Healthcare/Transportation: monitoring and predictive maintenance
  - Network Systems: performance optimization, security enhancement, traffic management
  - Automotive: ADAS, Self Driving Vehicles
  - Robots, Surveillance, and more

Required Architecture Support

- **AI/ML speedup:**
  - Matrix Multiplication:
    - Hardwired solution (APU/NPU/TPU): most efficient and powerful
    - Matrix Instructions: most flexible
      - IME (Integrated Matrix Extension) Task Group: leveraging VRF/MAC's in VPU
  - Non-linear functions: softmax, sigmoid, sin/cos, GeLu/SiLu
  - General compute instruction extensions: Vector Extension (RVV), SIMD/DSP (RVP)
- **Rich application enablement:**
  - Profiles: RVA20/22/23

debian

openSUSE

yocto PROJECT

fedora

OpenWrt Wireless Freedom

RISE RISC-V Software Ecosystem

AndesAIRE™ End-to-End AI/ML Solutions

Andes AI Runs Everywhere

NN models — PyTorch, ONNX, TensorFlow Lite, TensorFlow

AndeSight™ IDE

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

AndesClarity™
Pipeline Optimizer

ACE/COPILOT

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

SoC

Linux/RTOS Host
AX45MP, AX65

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

Domain Extensions
ACE

Accelerator
AnDLA™ I350

Andes Vector Processor Families

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

			AX46MPV*	AX47MPV*
		AX45MPV		more features
		8-core cluster	16-core cluster with private L1/L2	Integrated Matrix Ext. (IME)
AX25-V100 (Meta MTIA-1)	NX27V	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

AX45MPV and ACE: RAIN AI's Adoption*

Leveraging Andes Custom Extensions (ACE) automated

In-memory Compute Integration — ACE Streaming Port

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

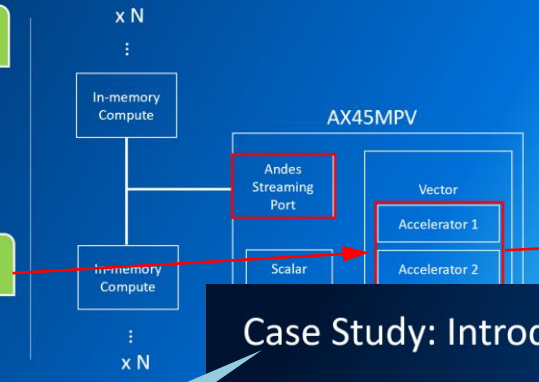
Leverage custom addressing and address control registers to control a plurality of in-memory compute blocks with different data sizes

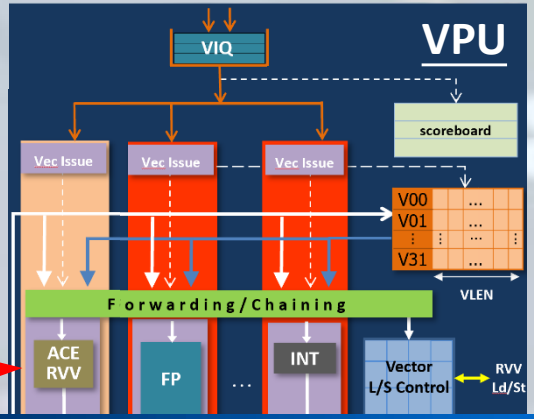
Non-linear Operator Integration — ACE RVV

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





Case Study: Introducing a Sigmoid Instruction

```
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
};
```

.ace

.v

XXXXXXXXXXXX

COPILOT

- Updated core RTL with new vector unit
- AndeSysC model with support for new instruction
- C Wrappers (Intrinsics)
- Compiler collateral

```
static inline vfloat16m4_t __attribute__ ((always_inline))
ace_sigmoid_f16m4(vfloat16m4_t x, size_t vl){
  vfloat16m4_t result_RetVar;
  vsetvl_e16m4(vl);
  __asm__ __volatile__ (
    "ace_sigmoid %[result], %[x], %[vm]"
    : [result] "=&vr" (result_RetVar)
    : [x] "vr" (x), [vm] "i" (1)
    : "memory"
  );
  return result_RetVar;
}
static inline vfloat16m8_t __attribute__ ((always_inline))
ace_sigmoid_f16m8(vfloat16m8_t x, size_t vl){
  vfloat16m8_t result_RetVar;
  vsetvl_e16m8(vl);
  __asm__ __volatile__ (
    "ace_sigmoid %[result], %[x], %[vm]"
    : [result] "=&vr" (result_RetVar)
    : [x] "vr" (x), [vm] "i" (1)
    : "memory"
  );
  return result_RetVar;
}
```

Sample wrappers for our sigmoid instruction

Also added **Softmax**, **SiLu**, and other instructions using ACE

* Presented in Andes RISC-V CON San Jose, June 2024

https://www.bilibili.com/video/BV1eDWpe6EcZ/?vd_source=b34de578c60c056a381185399b7b80ad

https://youtube.com/live/TpGFCTu_OFw

Andes Application Processor Families

AX66 VPU:
- **has 2 pipes shared by FP**
- **execute 2 ALU/load/store**

Android Base

AX63 customer-driven	AX65	AX66*	AX67*
Power-optimized	**Best-Balanced**	**Advanced**	**Most performant**
>7.0 specint2k6/GHz	8.25 specint2k6/GHz	*>10* specint2k6/GHz	*>11* specint2k6/GHz
	RVA22+	RVA23	RVA24
		further perf boost	
		V/VK (VLEN=128)	V/VK (VLEN up to 512)
		Hypervisor + AIA + (IOMMU + IOPMP)	
		Private L1/L2 Caches, CHI Multi-Cluster Coherency	

13-Stage, Out-of-Order Execution, Multicore Coherency, Linux-Capable, Up to 8 Cores/Cluster

Roadmap for AX60 Series

** Roadmap/features: subject to change. Future performance: estimated.*

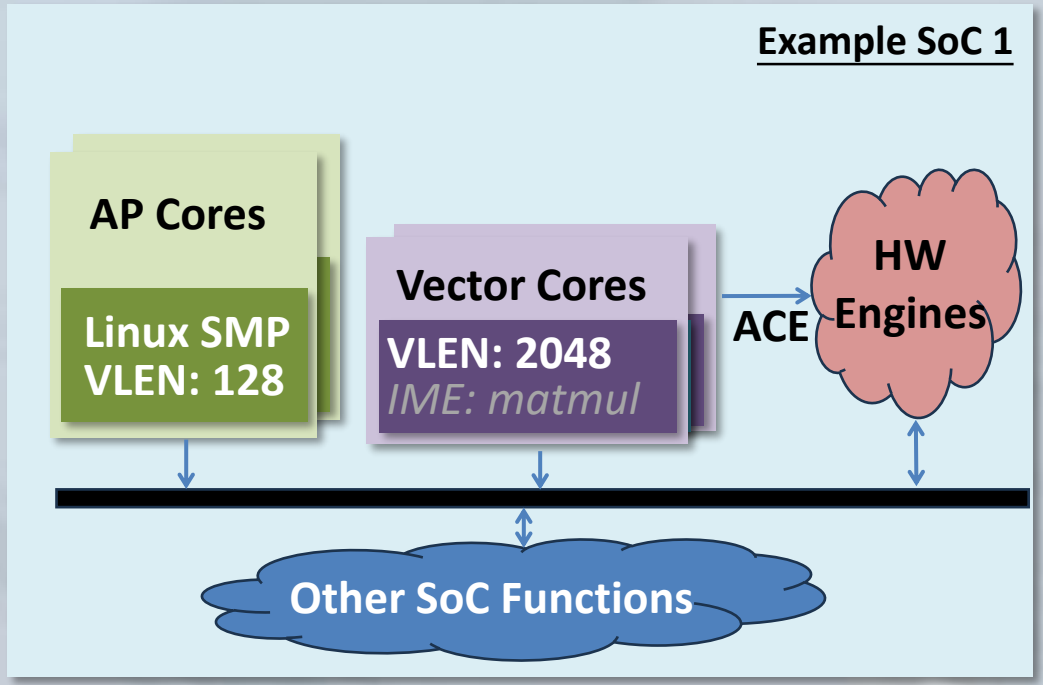
Andes Application Processor Families

AX63 customer-driven Power-optimized *>7.0* specint2k6/GHz	AX65 Best-Balanced 8.25 specint2k6/GHz	AX66 Advanced *>10* specint2k6/GHz	AX67 Most performant *>11* specint2k6/GHz	Cuzco Scalable performance 15~20 specint2k6/GHz
		RVA23	further perf boost	Private L1/L2, Shared L3
		V/VK (VLEN=128)	V/VK (VLEN up to 512)	Vector/Vector Crypto
		Hypervisor + AIA + (IOMMU + IOPMP)		8-Core Cluster with CHI
	RVA22+	Private L1/L2, CHI Multi-Cluster Coherency		14-stage 8-way/6-way OOO with Patented Time-Based Scheduling
13-Stage, Out-of-Order Execution, Multicore Coherency, Linux-Capable, Up to 8 Cores/Cluster				
Roadmap for AX60 Series				**Cuzco Series**

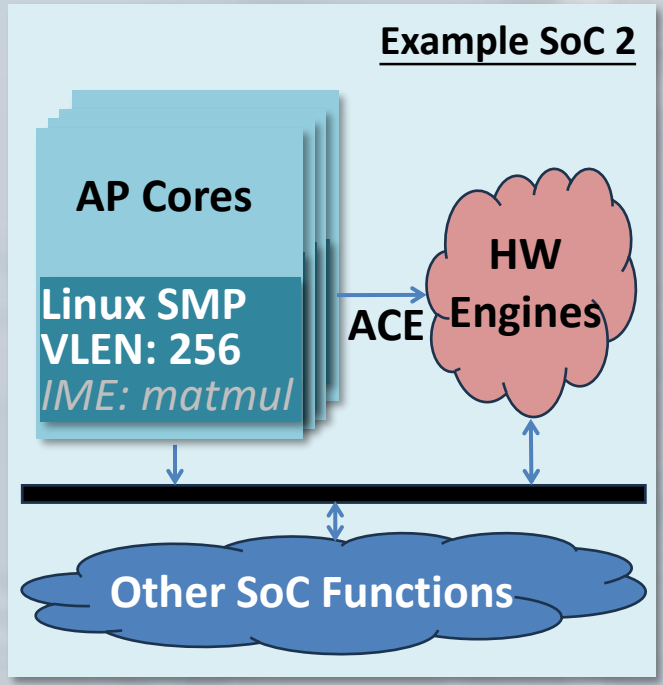
Roadmap/features: subject to change. Future performance: estimated.

Choices of SoC Architecture

- **Comparison: <u>perf</u>ormance, area efficiency (AE), ease of <u>programming</u>/optimization**



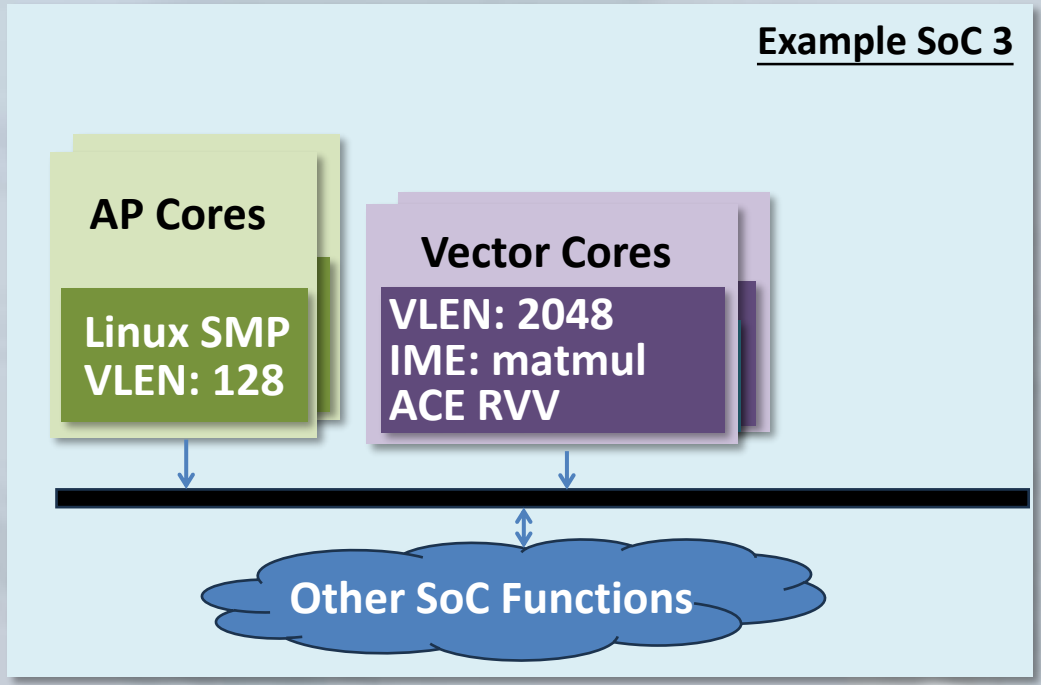
(perf, AE, programming)= (high, mid, low)



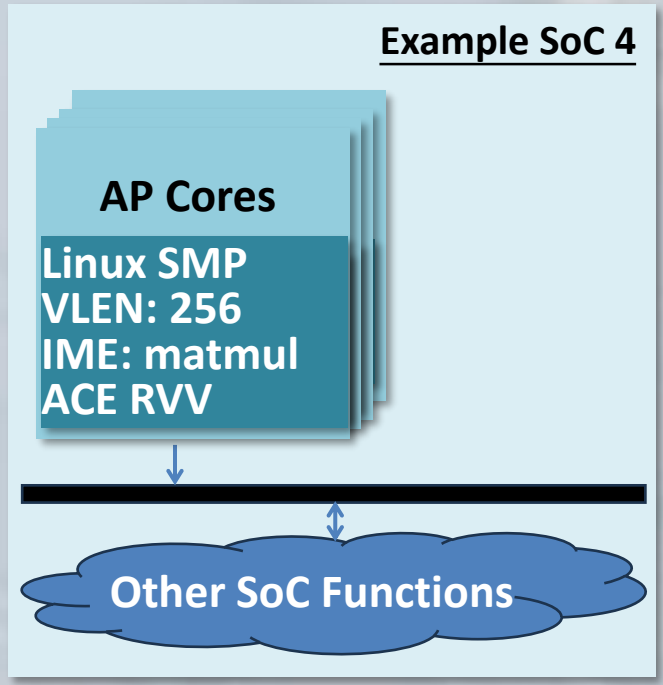
(perf, AE, programming)= (high, low, mid)

Choices of SoC Architecture

- **Comparison: perf**ormance**,** area efficiency (**AE**)**,** ease of **programming**/optimization



(perf, AE, programming)= (mid, high, mid)



(perf, AE, programming)= (low, mid, high)

Concluding Remarks

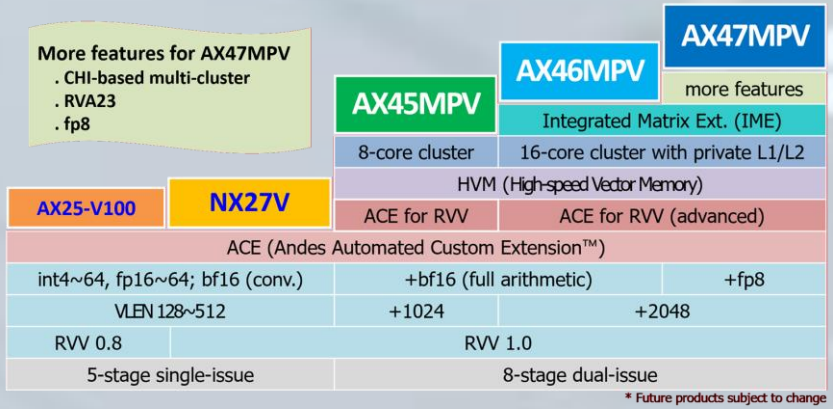
■ Intelligent Edge Applications

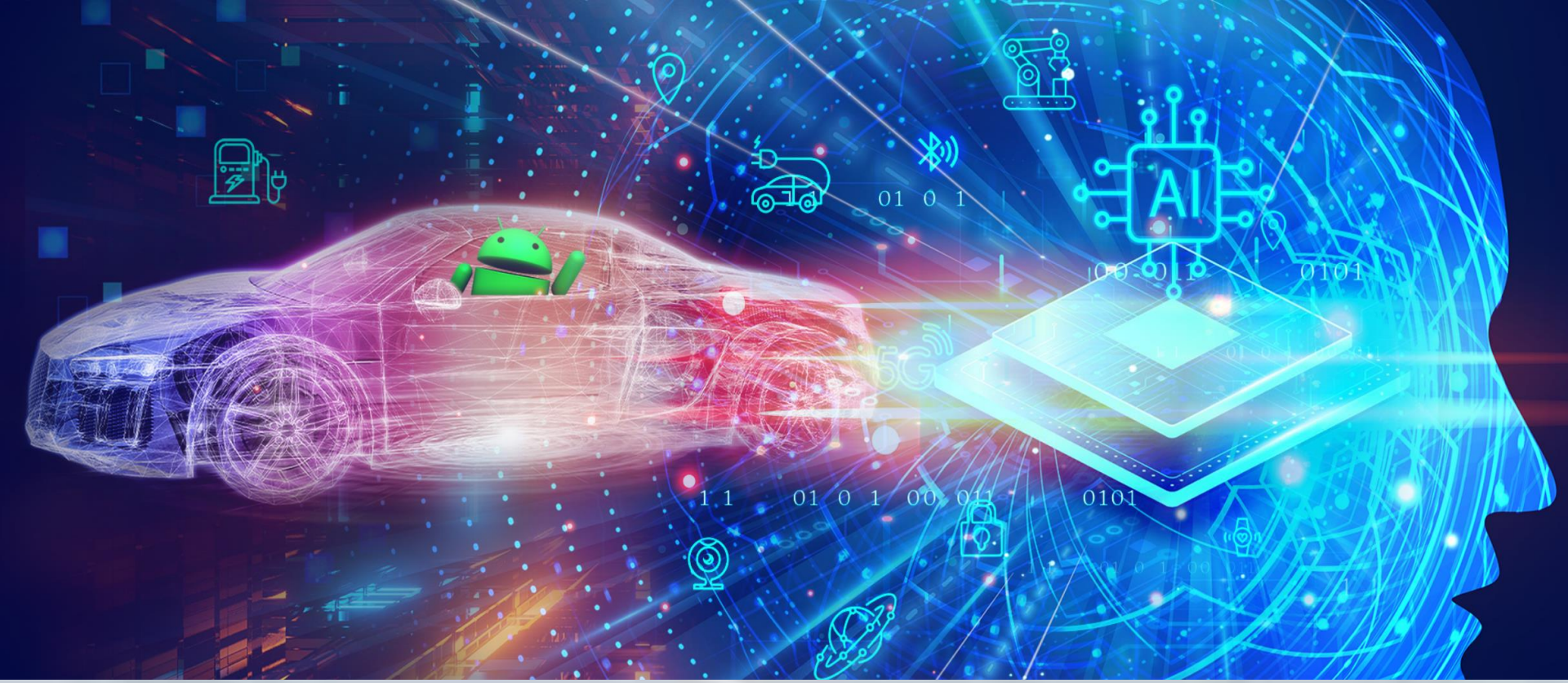
■ RISC-V Architecture Advantages

- Vector/Matrix processing: RVV, IME
- Custom Extensions
- RVA20/22/23: Linux/Android
- *Capacity/Bandwidth Management*

■ Andes Offerings

- Vector cores: 40-Series
- Linux cores: 60-Series, and Cuzco
- *AnDLA I350: hardwired engine*





Thank You !!

ANDES TECHNOLOGY

Taking RISC-V Mainstream