Andes Infuses into Artificial Intelligence

High-Efficiency and High-Flexibility Processor IPs + NN SDK for AI

John Min
Director of Field Application Engineering
Andes Technology
July 9, 2020
Agenda

- The Diversity of AI Use-Cases
- Andes RISC-V Processors for AI
- Andes NN SDK for AI
- Summary
Andes at A Glance

Who We Are

CPU
Pure-play CPU IP Company

RISC-V Founding Premier Member

Taiwan Stock Exchange Listed

Hammer
Major Open-Source Contributor/Maintainer

Running Task Groups Vice Chair of TSC
Director of the Board
RISC-V Ambassador

Quick Facts

15 years old company

200+ Licensees Worldwide

80% R&D employees

5B+ accumulated Andes-embedded SoC shipped

17K+ AndeSight IDE installation

Taking RISC-V® Mainstream
The Diversity of AI Use-Cases

**Vision**
- Image classification
- Object detection
- Image segmentation
- Spoof detection
- Face unlock
- Eye tracking
- Avatar
- SLAM
- ...

**Voice and Speech**
- Audio front-end processing
- Keyword spotting
- Voice command
- Speech to text
- Natural language processing
- Text to speech
- ...

**Any signal**
- Sensor fusion with force, pressure, accelerometer, gyro, ampere meter, vibration, radar/lidar, sonar, temperature, ...
- Pattern recognition
- Predictive maintenance
- Healthcare
- ...

**Taking RISC-V® Mainstream**
Andes Processors to Fit Your AI

Smart IoT Devices
- Voice trigger
- Voice command
- Always-on

Smart Camera
- Face trigger
- Object detection
- Intelligent HMI
- Barge-in
- Beamforming
- Speech to text
- Natural language

Smart Home
- Face unlock
- Bokeh
- Avatar
- SLAM
- Gesture recognition
- High resolution
- > 1 camera

Mobile AR/VR Surveillance
- Face unlock
- ADAS, HDR
- > 10 cameras

Automotive
- ADAS, HDR
- > 10 cameras

Data center
- ADAS, HDR
- > 100 TOPS

V-Series CPUs
- ADAS, HDR
- > 10 TOPS

D/A-Series CPUs
- ADAS, HDR
- > 100 TOPS

N-Series CPUs
- ADAS, HDR
- > 100 TOPS

~30 MOPS
~100 MOPS
> 1 GOPS
>1 TOPS
>10 TOPS
>100 TOPS

Taking RISC-V® Mainstream
# Andes RISC-V Processors Family

## N-Series Baseline
- RISC-V baseline 32/64-bit SMP
  - Andes V5 instructions (RV-EIMACFD-XV5)
- FPU, cache, local memory, ECC
- 2-stage to 8-stage pipeline
- Frequency up to 1.2GHz @28nm worse case

- ✓ Leading PPA and high efficiency CPU
- ✓ Control logic and simple data computation

## D/A-Series DSP/SIMD
- RISC-V baseline 32/64-bit SMP
  - Andes V5 instructions
  - DSP/SIMD instructions (RVP)
- MMU (A-Series)
- SIMD width: 32, 64
- Data types: INT8, INT16, INT32

- ✓ Efficient SIMD for data computation
- ✓ Compact MCU AI and basic edge AI applications

## V-Series Vector
- RISC-V baseline 64-bit
  - Andes V5 instructions
  - Vector instructions (RVV)
  - VLEN/SIMD width: 128, 256, 512
  - LMUL(Length Multiplier): 1, 2, 4, 8
  - Data types: INT4/8/16/32/64, BF16, FP16/32/64

- ✓ High performance, efficiency and configurability
- ✓ Enable data intensive computing from edge to cloud

---

Taking RISC-V® Mainstream
RVP and RVV for Data Computation

Andes RISC-V Baseline
- Clean state
- Compact
- Modular
- Andes V5 ISA extension

Relative benchmark with Andes V5 instructions
- +20% Performance (CoreMark)
- -12% Code size (CSiBE)

RISC-V DSP/SIMD P-ext
- Andes contributed market-proven DSP/SIMD to RVP
- Use RV32 and RV64 XLEN-bit GPRs
- SIMD with 8b, 16b, 32b
- Complex DSP operating on 16/32/64-bit
- Saturation and rounding
- Min, max, shift, byte swap, bit reverse, pack, unpack, ...

Speedup with RVP
- 14.32x CIFAR-10 image classification (RV64P)
- 8.86x ML-KWS keyword spotting (RV32P)

RISC-V Vector V-ext
- Follow RVV latest standard
- >300 vector instructions
- Scalable vector registers
- 2x/4x data expansion arithmetic
- Load/store, integer, fixed-point/floating-point operations

Speedup with RVV
- 57x SGEMM

Note, Based on N25F, Andes/mainline GCC v7.4
Typical Andes CPU Usages for AI from Edge to Cloud

**Best-fitting control logic**
- RISC-V Compact and modular design
- Remove the components which not needed (e.g. FPU, multiplier)

**MCU edge AI**
- Single MCU with small data computation (e.g. voice/face trigger)
- Always-on, low power, and cost-sensitive devices (e.g. smart doorbell, ear pod)

**Performance edge AI**
- Application SoC for large data process of CV/ML (e.g. AR/VR, surveillance)

**Cloud AI**
- Heterogeneous and cluster computing for AI data center

---

**Baseline**
- e.g. Accelerator Connectivity ...

**Baseline + RVP**
- e.g. Connectivity

**Baseline + RVV**
- x n

---

**Taking RISC-V® Mainstream**
Efficiency Boost with Andes Custom Extension™

**Compute kernel functions**
- Extend instructions for kernel functions (e.g. CONV, GEMM)
- Typical case: implement few dedicated kernel functions which consumes heavy computing power
- Could fit in low power and cost-sensitive devices

**Control ports**
- Extend instructions to control ports (e.g. send command, ack, wait-for-result)
- Typical case: a very compact CPU as a powerful accelerator controller which can send 90-bit commands in one cycle

**Streaming ports** + control ports + compute kernel functions
- Extend instructions for high volume information transferring between vector processors and external compute units
- Typical case: increase data bandwidth and shorten data latency when using vector to offload hard-wired AI compute unit (e.g. sigmoid)

**Baseline + RVP + ACE**
- Baseline + RVP + ACE
  - e.g. Connectivity
  - Control port
  - e.g. Accelerator
  - x N

**Baseline + ACE**
- Baseline + ACE
  - e.g. Accelerator
  - x N

**Baseline + RVV**
- Baseline + RVV
  - e.g. Accelerator
  - streaming port
  - Compute unit
  - x N

Taking RISC-V® Mainstream
Voice-Based Human Machine Interface Use Case

Pre-Processing
- Echo cancellation
- Noise reduction
- Beamforming
- Auto gain control
- ...

Feature extraction
- FFT
- Mel-Frequency Cepstral Coefficients
- Filter bank
- ...

Voice trigger
- Keyword spotting (always-on)

Speech/Text transform
- Automatic Speech Recognition
- Speech synthesis

Simple neural network model

Baseline + RVP

Baseline
Baseline + RVP
Baseline + RVV

Baseline + RVP

Taking RISC-V® Mainstream
Voice-Based Human Machine Interface Use Case

**Pre-Processing**
- Echo cancellation
- Noise reduction
- Beamforming
- Auto gain control
- ...

**Feature extraction**
- FFT
- Mel-Frequency Cepstral Coefficients
- Filter bank
- ...

**Voice trigger**
- Keyword spotting (always-on)

**Speech/Text transform**
- Automatic Speech Recognition
- Speech synthesis

**Language processing**
- Natural Language Processing
- Natural Language Understanding
- Natural Language Generation
- Dialog State Tracking
- ...

**Simple neural network model**

**Simple data computation**

**Intensive data computation**

**Complex neural network model**

---

**Baseline + RVP**

**Baseline + RVV**

**Baseline**

**Baseline + RVP**

**Baseline + RVV**

**AMBA bus**

**edge**

**cloud**

**Taking RISC-V® Mainstream**
Deep Learning Chipset Global Market

- Deep learning chipset market growing at 42.2% CAGR from 2016 to 2025
- Largest growth coming from ASIC including:
  - CPU
  - DSP
  - Vector processing unit
  - Hard-wired engine
  - ...

Tractica, March, 2017
Andes NN SDK
Full ecosystem of AI software frameworks, compilers and libraries
Andes DSP Library

- Optimized low-level DSP functions for RISC-V baseline and RVP processors
- Boost signal processing performance
- >200 functions in 8 categories
- CMSIS-DSP like APIs

Speedup of RV64P over baseline\(^1\)
Andes NN Library and TensorFlow Lite Micro

**Andes NN library**
- An optimized low-level NN functions for RISC-V baseline, RVP and RVV processors
- Boost NN performance by using SIMD and vector instructions
- CMSIS-NN like API

**TensorFlow Lite for Microcontroller (TFLiteμ)**
- Create bare-metal binary with offline flow
- Major kernel functions hooked up with Andes NN library
RVP DSP/SIMD Processors Speedup

Speedup of RVP over baseline
The higher the better

MP3 decode
AMR voice codec
ML-KWS (Keyword Spotting)
PNET (90% of Face Detection)
CIFAR-10 (Image Classification)

CIFAR-10 image classification speedup
The higher the better

1: based on 25-Series, FPGA
2: based on 25-Series, FPGA with similar configurations

- Performance boost with Andes NN/DSP libraries
- Increase power efficiency
- Higher response time

Taking RISC-V® Mainstream
## RVV Vector Processors Speedup over Baseline

<table>
<thead>
<tr>
<th>Operation</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGEMM</td>
<td>57x</td>
</tr>
<tr>
<td>Q7 filtering</td>
<td>39x</td>
</tr>
<tr>
<td>F32 filtering</td>
<td>19x</td>
</tr>
<tr>
<td>Pointwise CNN</td>
<td>21x</td>
</tr>
<tr>
<td>Depthwise CNN</td>
<td>18x</td>
</tr>
<tr>
<td>RGB CNN</td>
<td>18x</td>
</tr>
<tr>
<td>F32 basic mathematics</td>
<td>19x</td>
</tr>
</tbody>
</table>

**Note**
- Compared to pure C scalar code compiled with high optimization
- Both vector and scalar code ran on the NX27V FPGA with 512-bit VLEN, 256-bit bus
### Voice trigger
- To wake up the system
- Consume lower power than ASR for always-on usage
- Reduce false alarms

### Voice command
- Hands-free solutions
- Simple and offline HMI
Andes KWS Solution

- **KWS software stack**
  - Andes NN/DSP library accelerated by Andes RISC-V DSP/SIMD P-ext

- **KWS application**
  - Feature extraction: MFCC
  - AI model: DNN, DS-CNN, GRU

- **KWS tools**
  - KWS TensorFlow training script
  - KWS quantization tool
  - KWS model code-gen to .c/.h

<table>
<thead>
<tr>
<th>Model</th>
<th>DS-CNN</th>
<th>DNN</th>
<th>GRU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>94.4%</td>
<td>84.6%</td>
<td>93.5%</td>
</tr>
<tr>
<td>Flash size</td>
<td>186 KB</td>
<td>243 KB</td>
<td>243 KB</td>
</tr>
<tr>
<td>(code+rodata +data)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SRAM size</td>
<td>35 KB</td>
<td>35 KB</td>
<td>36 KB</td>
</tr>
<tr>
<td>(data+bss)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cycles¹</td>
<td>3,498,638</td>
<td>179,136</td>
<td>5,055,417</td>
</tr>
</tbody>
</table>

1: collected only from one inference sample of WAV file on D25F FPGA
Andes Partners for AI

AI tools and IP
- DeepLite
- PEAKHILLS
- Kheron
- skyaizer

Open Source SW
- TVM
- ONNC
- ELVM
- TensorFlow
- OpenCL
- Zephyr

DSP and Vision
- Sensory
- Open AI LAB
- multicoreware
- RELAJET

Development tools
- SEGGER
- Imperas
- Ultrasonic
- IAR Systems
- Lauterbach

Taking RISC-V® Mainstream
Summary

- Andes RISC-V processors support the diversity of AI use-cases
  - **Baseline**: compact and modular control logic
  - **Baseline + RVP**: efficient DSP/SIMD for simple data computation
  - **Baseline + RVV**: high performance, efficiency and configurability to enable data intensive computing from edge to cloud

- Andes NN SDK targets to boost your SoC AI performance, achieve outstanding hardware utilization and most importantly, improve time-to-market
  - **Andes DSP library** for signal processing
  - **Andes NN library** for NN operators

- Ecosystem further advances your AI project developments
Thank you,
See you next webinar!