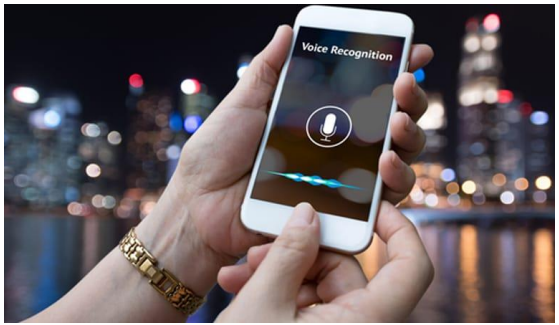
The Future of Edge Is About RISC-V and AI

Simon TC Wang
wangtc@andestech.com
Senior Technical Marketing Manager
Andes Technology
Feb. 22, 2023

The Diversity of AI Use-Cases

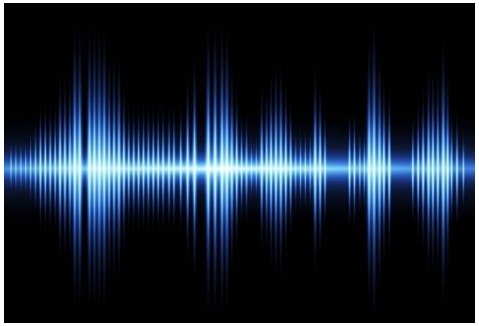
Vision

- Image classification
- Object detection
- Image segmentation
- Spoof detection
- Face unlock
- Eye tracking
- Avatar
- SLAM
- …



Voice and Speech

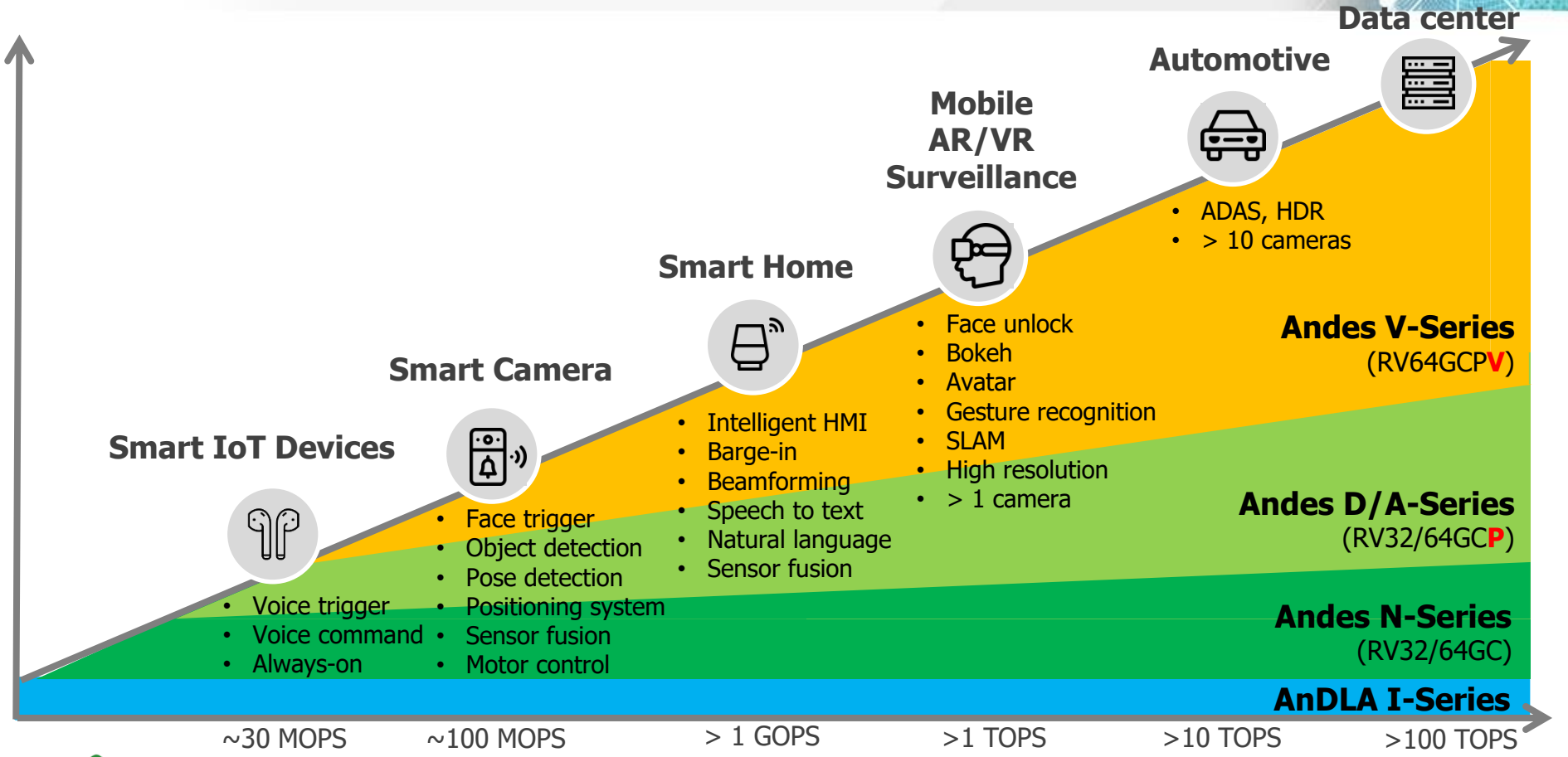
- Audio front-end processing
- Keyword spotting
- Voice command
- Speech to text
- Natural language processing
- Text to speech
- …



Any signal

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

Andes Processors to Fit Your AI



Andes RISC-V Processors Family

N-Series Baseline

RISC-V baseline 32/64-bit
✚ AndeStar™ V5 ext. (CoDense™, Performance)

2 to 13 stage pipelines
Single issue, superscalar
In-order, out-of-order

SMP, cache, local memory, ECC, …

D/A-Series DSP/SIMD

RISC-V baseline 32/64-bit
✚ AndeStar™ V5 ext.
✚ RISC-V Packed-SIMD ext. (RVP draft)

MMU (A-Series)

SIMD width: 32, 64

Data types: INT8/16/32

V-Series Vector

RISC-V baseline 64-bit
✚ AndeStar™ V5 ext.
✚ RISC-V Vector ext. (RVV)

VLEN/SIMD width: 128-1024

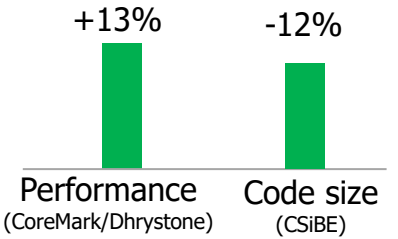
LMUL (Length Multiplier): 1-8

Data types: INT4/8/16/32/64, BF16, FP16/32/64

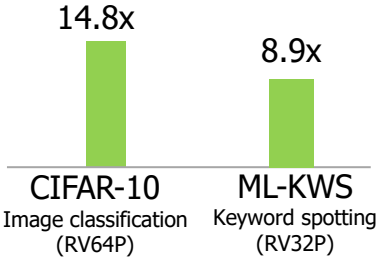
AnDLA

- Accelerated engine for GEMM, MAC accumulator, element-wise, pooling, etc
- Dedicated DMA and local shared SRAM

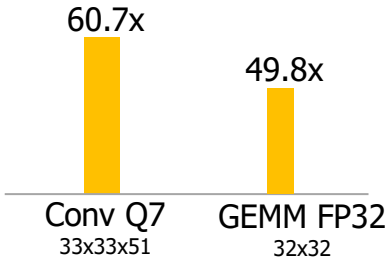
Speedup and code size reduction with AndeStar™ V5[1]



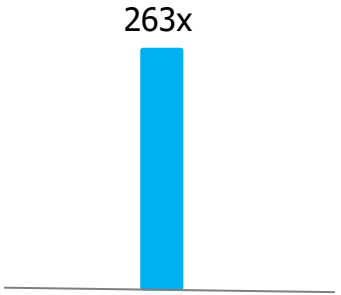
Speedup with RVP draft[2]



Speedup with RVV[3]



Speedup with AnDLA[4]



1: 25-Series, FPGA, Andes toolchain over open-source GCC v7.4

2: 25-Series, FPGA

3: NX27V, FPGA, (VLEN,SIMD)=(512,512), L1D 512 KB

4: I350, 64 MAC, FPGA

Andes RISC-V DSP/SIMD Processors

- AndesCore™ D25F
  - RV32GC/P*/B + AndeStar™ V5 ext.
  - 5-stage in-order single-issue
- AndesCore™ D45
  - RV32GC/P* + AndeStar™ V5 ext.
  - 8-stage in-order dual-issue
  - MemBoost
    - Instruction and data cache pre-fetch
    - Non-blocking loads/stores
    - Data cache write-around
    - Optional separated BIU I/D buses
- AndesCore™ common technologies
  - Hardware misaligned access, CoDense™, PowerBrake, QuickNap™, …
  - Andes Custom Extension (ACE)

Features	D25F	D45 (over D25F)
CoreMark/MHz[1]	3.57	5.67 (+59%)
DMIPS/MHz (no-inline)[1]	1.98	2.86 (+44%)
Gate count[2]	100%	140% (+40%)

D25F

- Note 1. DMIPS: AndeSight 5.0 toolchain, COPTS= -O3 - flto -fno-inline; CoreMark: AndeSight 5.0 toolchain, OPTFLAGS= -O3 -funroll-all-loops -finline-limit=600 -ftree-dominator-opts -fno-if-conversion2 -fselective-scheduling -fno-code-hoisting
- Note 2. TSMC N7 ULVT, LVT and SVT 240nm cell height library, TSMC '7nm Fin FET High Speed L1 Cache Memory Compiler', 256-entry BTB, 16-entry PMP and 32KB I/D$ (no Local Memory), without RVB, with I/O constraint, die area and power are core only without SRAM, 65% utilization. Frequency condition: SSGNP/0.675V/-40oC; Dynamic power condition: TT/0.75V/85oC

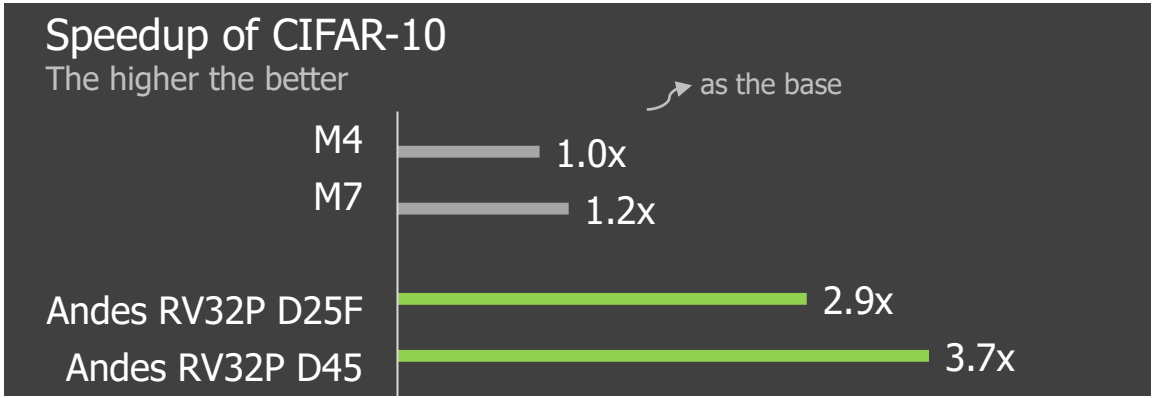
D45

- Note 1. DMIPS: AndeSight 5.0 toolchain, COPTS= -O3 - flto -fno-inline; CoreMark: AndeSight 5.0 toolchain, OPTFLAGS= -O3 -funroll-all-loops -finline-limit=600 -ftree-dominator-opts -fno-if-conversion2 -fselective-scheduling -fno-code-hoisting
- Note 2: TSMC 7nm FIN FET ULVT/LVT/SVT, cell height 240nm, High Speed L1 Cache Memory Compiler. Frequency condition: worst: : SSGNP/0.675V/-40°c, typical: TT/0.75v/+85°c. Power and area : typical corner. Configurations: 256-entry BTB, PMP&PMA 16-entry, 32KB I/D$ (no Local Memory), MemBoost, with I/O constraint; die area and power are core only, 65% utilization

*: RVP (draft)

Use Case: CIFAR-10 Image Classification

CIFAR-10	Operators	Andes NN Library APIs
Layer 1	Convolution Activation (ReLU)	riscv_nn_conv_HWC_s8_s8_s8_RGB_sft_bias_fast riscv_nn_relu_s8
	Pooling (maxpool)	riscv_nn_maxpool_HWC_s8
Layer 2	Convolution Activation (ReLU)	riscv_nn_conv_HWC_s8_s8_s8_sft_bias_fast riscv_nn_relu_s8
	Pooling (maxpool)	riscv_nn_maxpool_HWC_s8
Layer 3	Convolution Activation (ReLU)	riscv_nn_conv_HWC_s8_s8_s8_sft_bias_fast riscv_nn_relu_s8
	Pooling (maxpool)	riscv_nn_maxpool_HWC_s8
Layer 4	Fully-connected	riscv_nn_fc_s8_s8_s8_sft_bias_fast
Layer 5	Softmax	riscv_nn_softmax_s8_fast

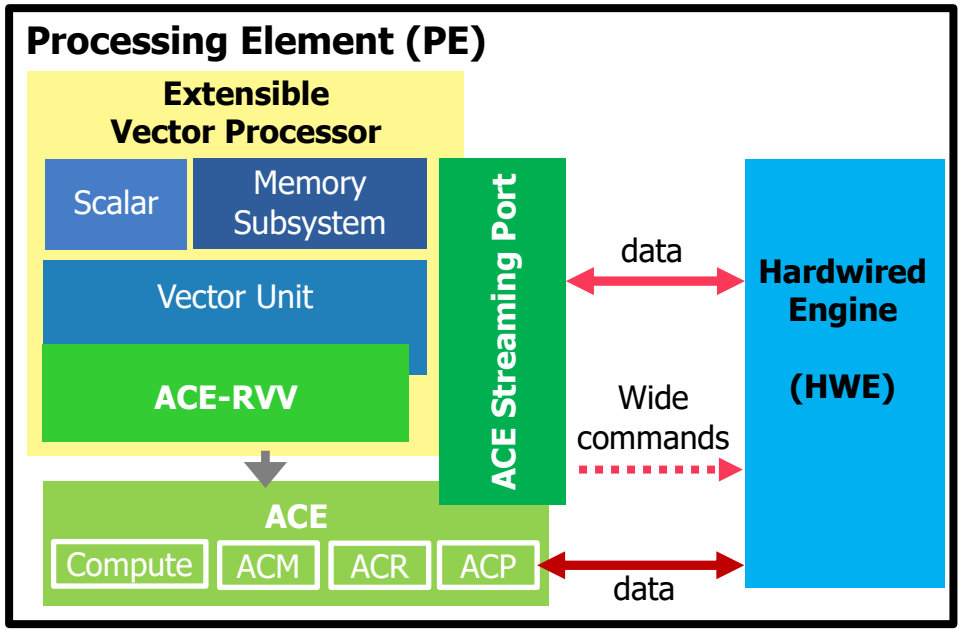


Classification

CAT

Andes RISC-V Vector Processors

- AndesCore™ NX27V
  - RV64GC/P[1]/V[2] + AndeStar™ V5 ext.
  - 5-stage in-order single-issue scalar unit
  - Vector Processing Unit (VPU)
    - INT4/8/16/32/64, FP16/32/64, BF16
    - VLEN & SIMD: 128/256/512, 1:1 or 2:1
- AndesCore™ AX45MPV
  - RV64GC/B/P[1]/V + AndeStar™ V5 ext.
  - 8-stage in-order dual-issue scalar unit
  - Vector Processing Unit (VPU)
    - Dual-issue vector pipeline
    - INT4/8/16/32/64, FP16/32/64, BF16
    - VLEN & SIMD: 128/256/512/1024, 1:1 or 2:1
- AndesCore™ common technologies
  - MemBoost memory subsystem
  - ACE (Andes Custom Extension)
  - ACE Streaming Port (ASP), ACE-RVV[3]



1: RVP draft
2: RVV v1.0 except segment load/store
3: 45V only

Use Case: MobileNet-v1 Person Detection

Inference latency of MobileNet-v1

The lower the better

MobileNet-v1	Data Type	ISA	VLEN (bit)	SIMD (bit)	Normalized latency (ms @1GHz, 1 core)
CA9 (Xilinx PYNQ)[1]	FP32	NEON	-	128	703.4
CA53 (Raspberry Pi-3B)[1]		NEON	-	128	438.5
CA72 (Firefly RK3399)[1]		NEON	-	128	210.3
CA73 (Kirin 970)[1]		NEON	-	128	292.4
Andes NX27V[2]	FP16	RVV	128	128	127.6
Andes NX27V[2]		RVV	256	128	100.5
Andes NX27V[2]		RVV	256	256	69.9
Andes NX27V[2]		RVV	512	256	56.9
Andes NX27V[2]		RVV	512	512	42.2

Table 1. MobileNet Body Architecture

Type / Stride	Filter Shape	Input Size
Conv / s2	3 × 3 × 3 × 32	224 × 224 × 3
Conv dw / s1	3 × 3 × 32 dw	112 × 112 × 32
Conv / s1	1 × 1 × 32 × 64	112 × 112 × 32
Conv dw / s2	3 × 3 × 64 dw	112 × 112 × 64
Conv / s1	1 × 1 × 64 × 128	56 × 56 × 64
Conv dw / s1	3 × 3 × 128 dw	56 × 56 × 128
Conv / s1	1 × 1 × 128 × 128	56 × 56 × 128
Conv dw / s2	3 × 3 × 128 dw	56 × 56 × 128
Conv / s1	1 × 1 × 128 × 256	28 × 28 × 128
Conv dw / s1	3 × 3 × 256 dw	28 × 28 × 256
Conv / s1	1 × 1 × 256 × 256	28 × 28 × 256
Conv dw / s2	3 × 3 × 256 dw	28 × 28 × 256
Conv / s1	1 × 1 × 256 × 512	14 × 14 × 256
5× Conv dw / s1	3 × 3 × 512 dw	14 × 14 × 512
5× Conv / s1	1 × 1 × 512 × 512	14 × 14 × 512
Conv dw / s2	3 × 3 × 512 dw	14 × 14 × 512
Conv / s1	1 × 1 × 512 × 1024	7 × 7 × 512
Conv dw / s2	3 × 3 × 1024 dw	7 × 7 × 1024
Conv / s1	1 × 1 × 1024 × 1024	7 × 7 × 1024
Avg Pool / s1	Pool 7 × 7	7 × 7 × 1024
FC / s1	1024 × 1000	1 × 1 × 1024
Softmax / s1	Classifier	1 × 1 × 1000

Table 2. Resource Per Layer Type

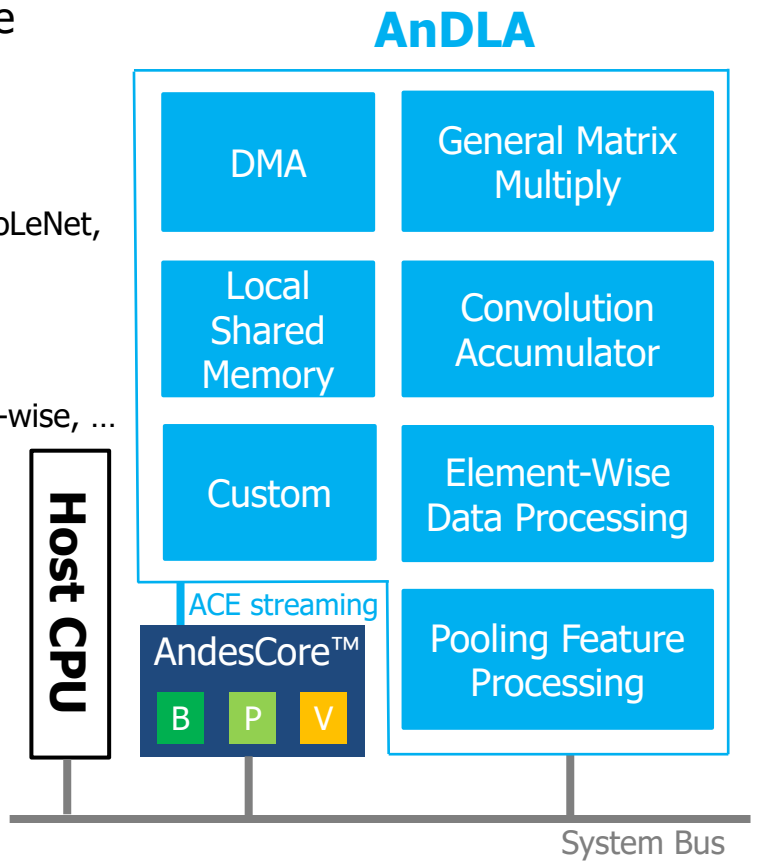
Type	Mult-Adds	Parameters
Conv 1 × 1	94.86%	74.59%
Conv DW 3 × 3	3.06%	1.06%
Conv 3 × 3	1.19%	0.02%
Fully Connected	0.18%	24.33%

1: TVM, https://github.com/apache/tvm/wiki/Benchmark#arm-cpu

2: Andes libnn, PyTorchCV imagenet-1k "MobileNet x1.0" (https://pypi.org/project/pytorchcv/), based on FPGA and scaling to 1.0 GHz. Real SoC performance will depend on memory subsystem

Andes Deep Learning Accelerator (AnDLA)[1]

- A standalone deep learning accelerator for edge inference
  - Scalable and multicore accelerator
  - Cooperate with AndesCore™ 25/27/45/60-Series
- Supported NN models
  - Video and image: VGG, Mobilenet, ResNet, YOLO, SSD, Inception, GooLeNet, DenseNet, …
  - Audio and voice: RNN, LSTM, GRU, …
- Accelerated NN operators
  - Convolution, fully-connected, activation, pooling, depth-wise, element-wise, …
  - Operator fusion
- Target performance
  - Configurable MACs: 32 to 2048 (INT8)
  - Performance: 64 GOPS to 4 TOPS (INT8 @1GHz)
  - Leading power efficiency >5 TOPS/W (@28nm)
- Integrated DMA and local shared memory

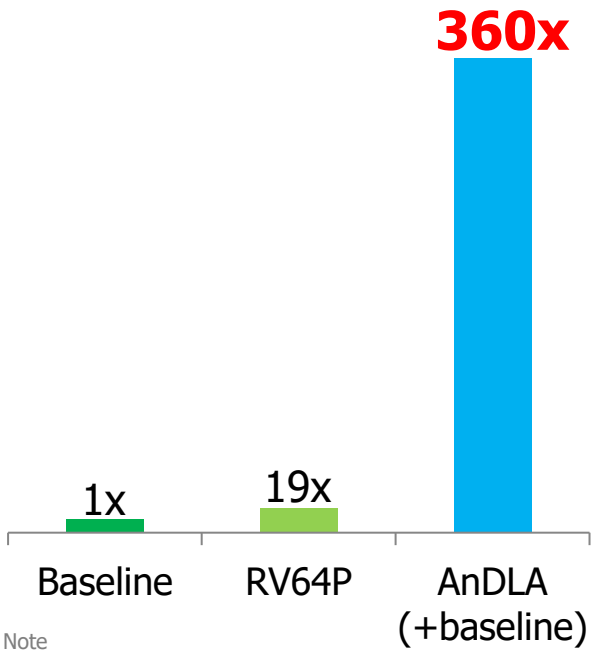


1: the preliminary target on current roadmap; VGG and fixed 64 MACs only in 2022

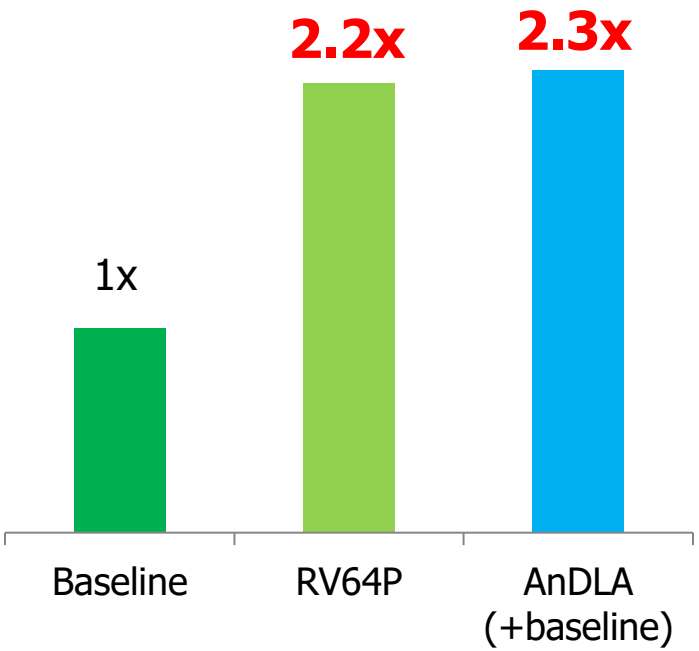
Efficient Engine for NN Computation

Convolution kernel

(from CIFAR-10 layer 2)



Gate counts

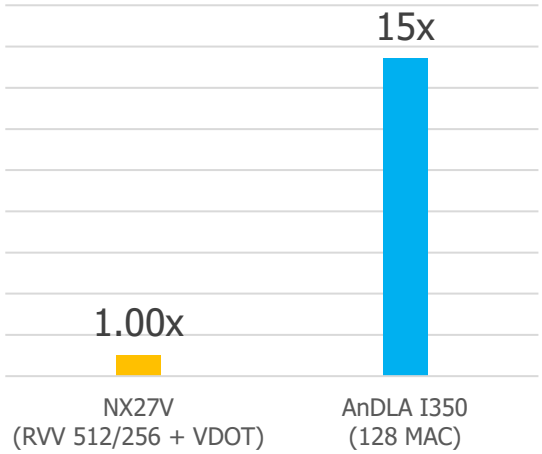


Note

- preliminary data based on 25-Series
- AnDLA with 64 MACs configuration
- Gate counts @28nm HPC+; AnDLA logic only and without shared memory; CPU logic only and without L1 cache

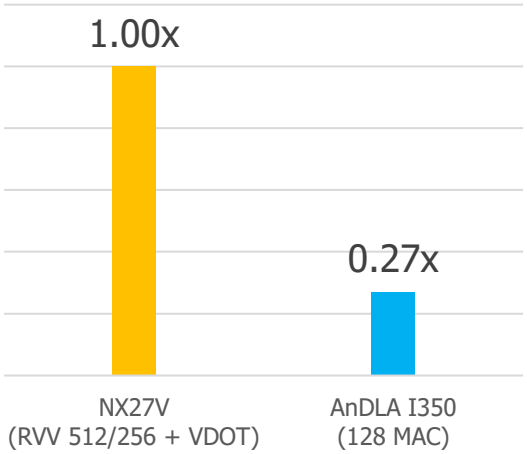
Efficient Engine for NN Computation

ResNet-50 Performance[1]



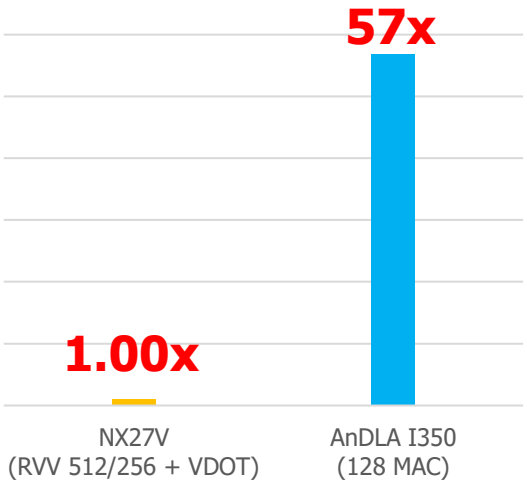
÷

Gate Counts[2]



=

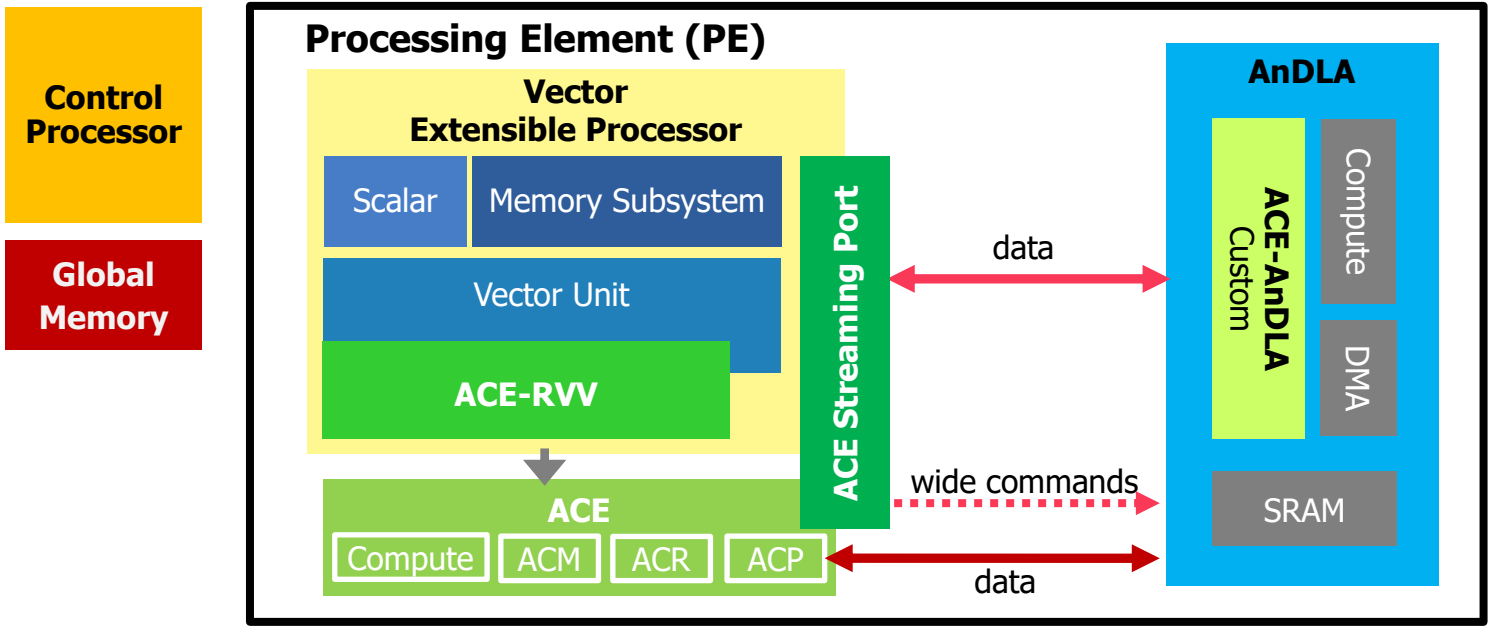
Performance Efficiency



1: AnDLA performance doesn't include softmax which is around 0.009% of total cycles and it could be ignored; AnDLA results is estimated from tools plus the estimated error from Tiny YOLO experience between FPGA and tools

2: TSMC N7 (SVT, LVT, ULVT) 240H library; AXI BUS with I/O constrain, synthesis with 30% of clock period margin; frequency condition: 0.675v/-40oc, SS; area condition: 65% utilization with scan; dynamic power condition: 0.75v/25oC, TT. Pre-layout simulation; **RVV, logic only, 128-entry BTB, 16-entry PMP/PMA and 32KB I $; AnDLA, logic only, 128 MAC**

Andes Extensible RISC-V AI Subsystem[1]

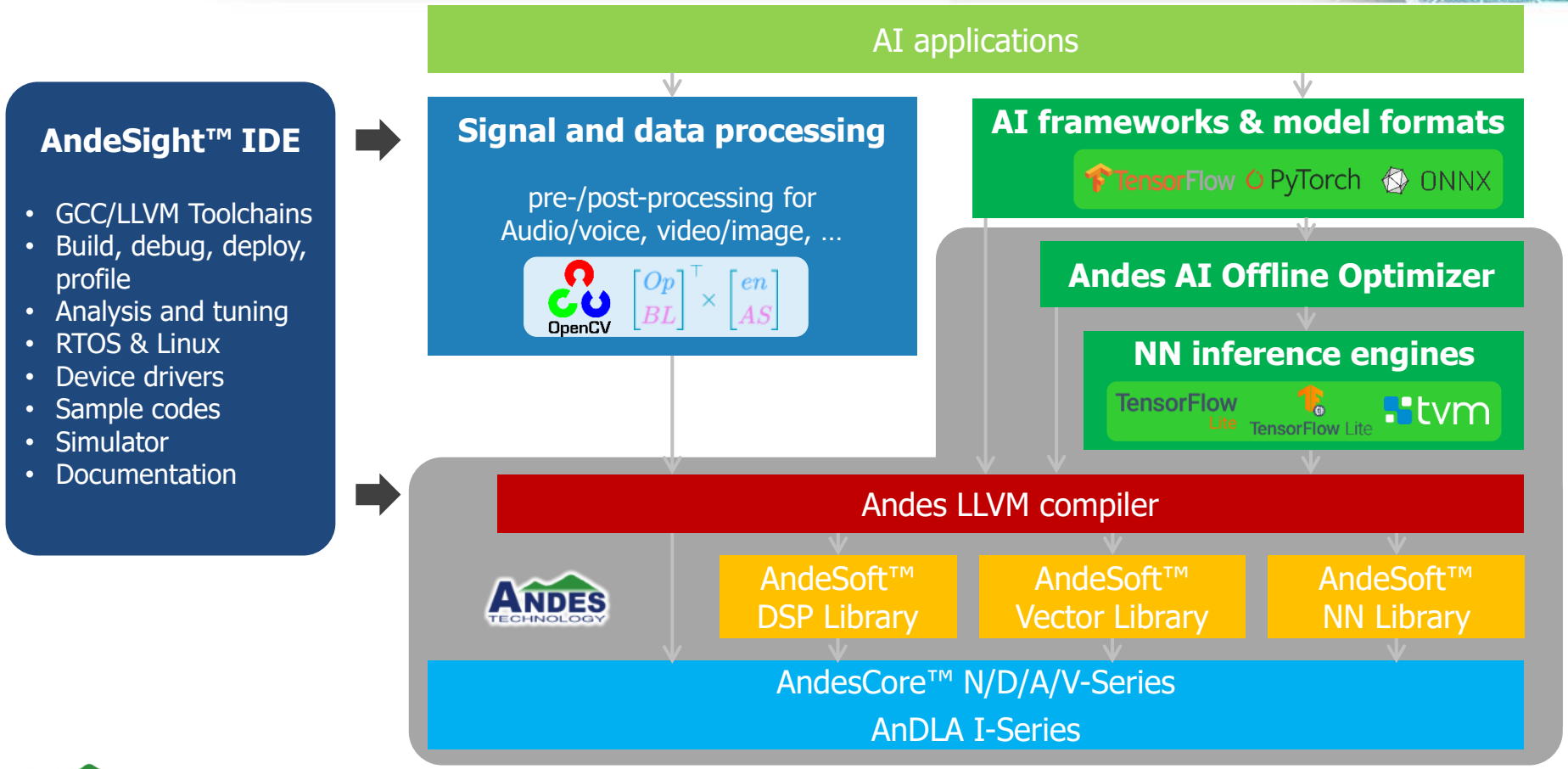


Extended Type 1. **Computing power**	Extended Type 2. **Data exchange**	Extended Type 3. **Control signal**

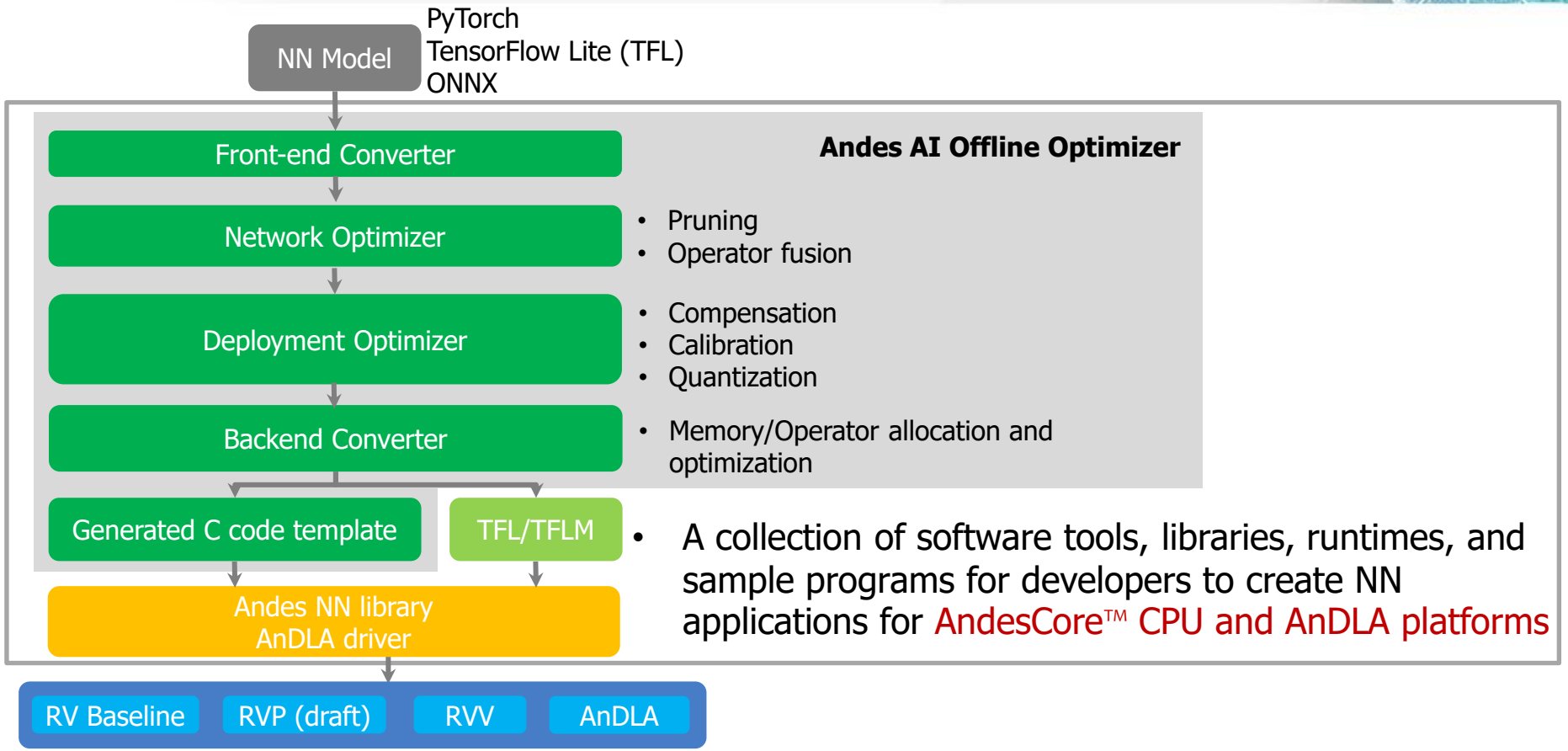
1: the preliminary target on current roadmap; ACE, ACE-RVV, ASP and ACE-AnDLA are separated add-on packages

AndeSight™ IDE
&
Andes NN SDK

Andes Tools and Software Stacks for AI

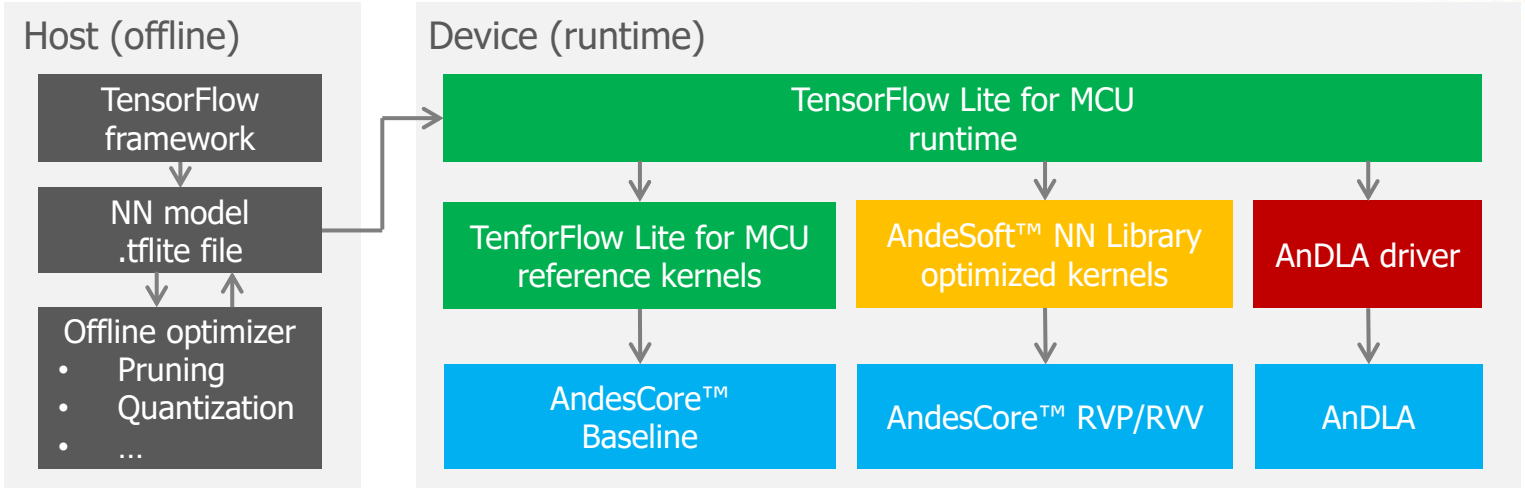


Andes NN SDK



- A collection of software tools, libraries, runtimes, and sample programs for developers to create NN applications for AndesCore™ CPU and AnDLA platforms

Inference Flow with TensorFlow Lite for MCU



- TFLM reference kernels (fallback; pure-C implementation) are always available
- Switch to hardware architecture with higher performance if possible (at compiling time)

TFLM kernel	TFLM reference implementation	Libnn for RVP/RVV	AnDLA driver
CONV_2D	✓	✓	✓
FULLY_CONNECTED	✓	✓	✓
SOFTMAX	✓	✓	
EXP	✓		

Andes Readiness for TFL Models

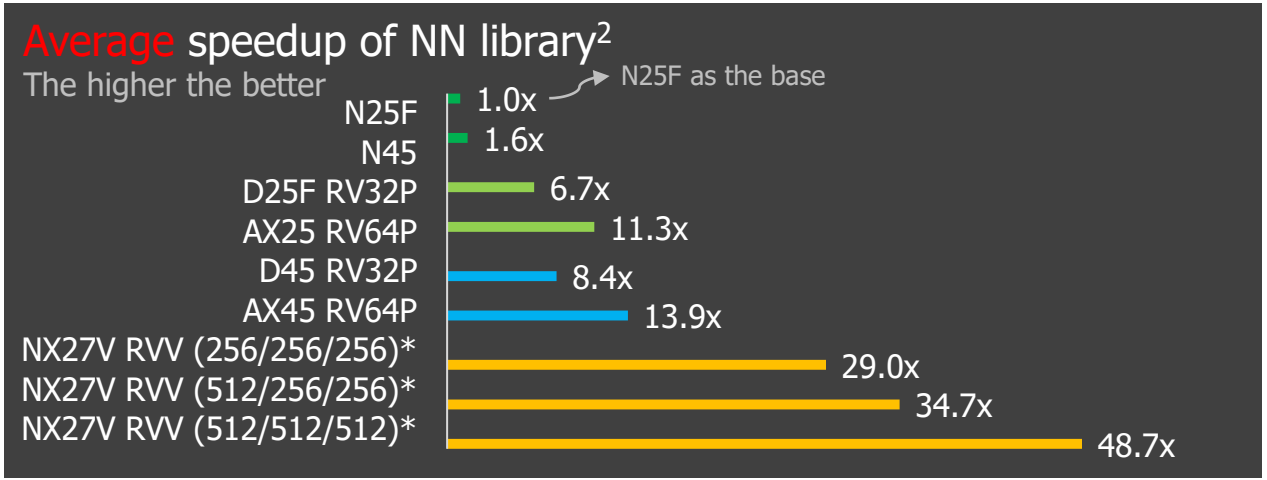
- TensorFlow Hub: runnable percentage > 93% in models of Image, Text, Video, Audio domains[1]
  - Using TF Flex operators[2] and proprietary custom operators are the most of reasons causing failure

Domain	Sub-Domain	Valid Models (excluded TPU)	Successfully Executed	Executable Rate
Image	Classification	107	107	100.0%
	Segmentation	18	18	100.0%
	Object Detection	33	30	90.9%
	Pose Detection	8	8	100.0%
	Super Resolution	5	5	100.0%
	Others	30	24	80.0%
Text		8	5	62.5%
Video		6	6	100.0%
Audio		16	12	75.0%
Total		**231**	**215**	**93.1%**

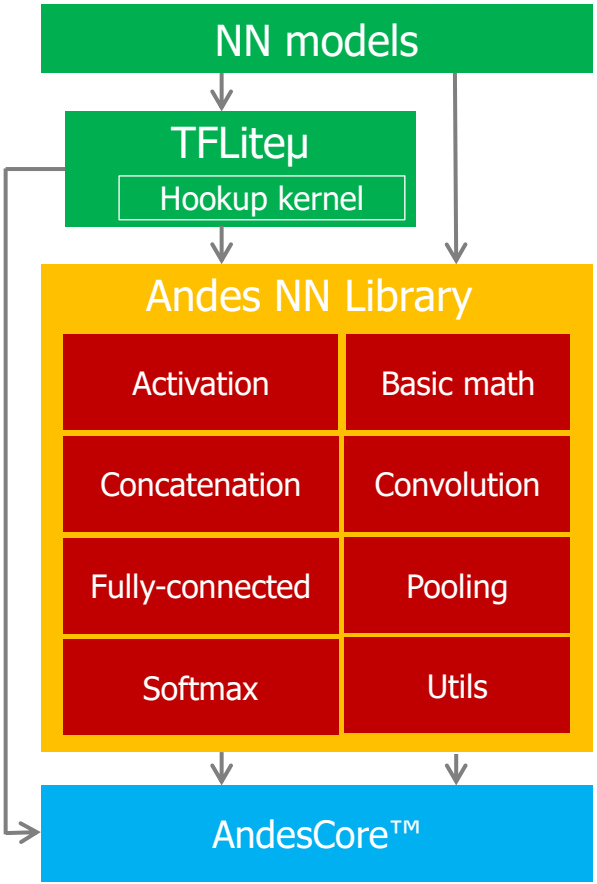
Note
1: Models from TensorFlow Hub, Nov. 2022, https://tfhub.dev/
2: A TF experiment features: need full TF runtime and not suitable for edge devices

AndeSoft™ NN Library[1]

- Optimized neural network functions for RVP (draft) and RVV processors
- Boost NN performance by using SIMD and Vector instructions
- >170 functions in 8 categories
- Superset and compatible with CMSIS-NN APIs



1: separated product packages needed additional licenses
2: compare with full APIs; data based on FPGA
*: (VLEN/SIMD/BIU)



Commit to Compute Library Evolution

- AndeSoft™ DSP Library
  - Optimized for RVP DSP/SIMD processors
  - >320 functions in 10 categories: basic, complex, controller, distance, filtering, matrix, sort, statistic, svm, transform, utils
  - Superset and compatible with CMSIS-DSP API
- AndeSoft™ Vector Library
  - Optimized for RVV Vector processors
  - > 200 functions in 5 categories: basic, filtering, image, matrix, and transform
  - Superset and compatible with NE10 library APIs

The total function number of
Andes NN / DSP/ Vector libraries

Before	After	Number of newly added functions
~560 (ASTv510 @2022/2/18)	~690 (ASTv520 @2023/1/13)	+130
~690 (ASTv520 @2023/1/13)	~870 (ASTv530 @2023/Q4)	+180

Summary

- Andes RISC-V DSP/SIMD and Vector processors provide highly efficient computing power for diversified AI applications and segments
- Andes is the leading pioneer to deliver the RISC-V extensible architecture with Andes Custom Extension (ACE), ACE-RVV, ACE Streaming port (ASP)
- AnDLA plays a key role to bring the power-efficiency computing to the "Andes Extensible RISC-V AI Subsystem"
- AndeSight™ IDE and Andes NN SDK brings the ultimate runtime performance and development efficiency to developers

ANDES
TECHNOLOGY