AutoIREE: Automatic Performance Tuning for AI models on RISC-V Vector Architectures

洪培翔,張元銘,吳奕緯

Andes Technology Corporation

RISC-V Summit China, 2024/08/22

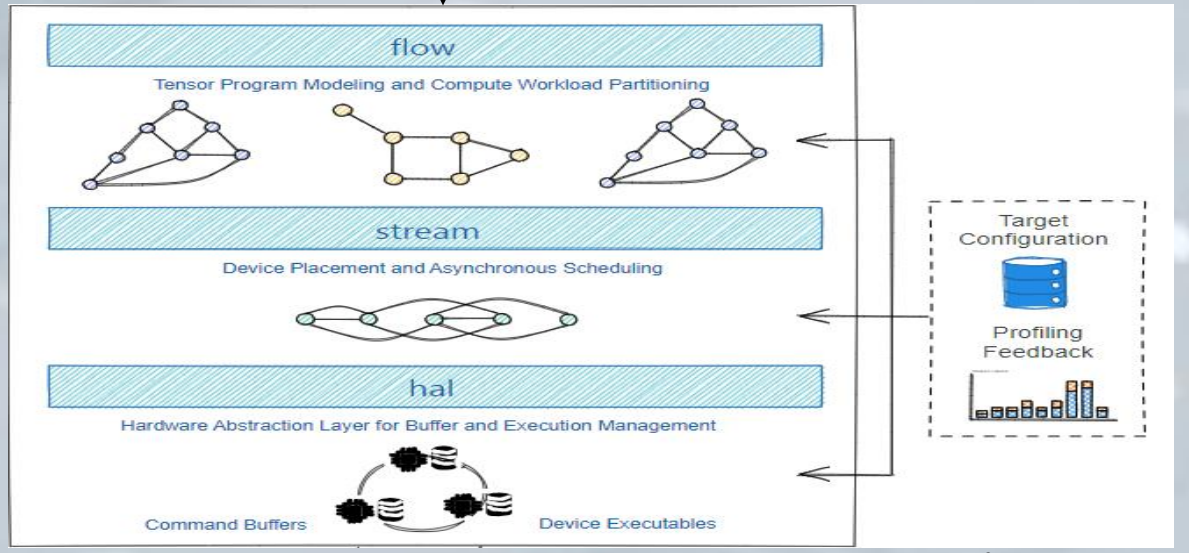
Outline

- Background Introduction
  - IREE
  - Andes Vector Processor Families
- Implementation and Innovation
- Experiment Results
- Future Works

IREE, MLIR-based Compiler + Runtime

Model input

Prebuilt AI Model (Pytorch, TF, TFL)



Courtesy of IREE website

Targets

Andes Cores

GPUs

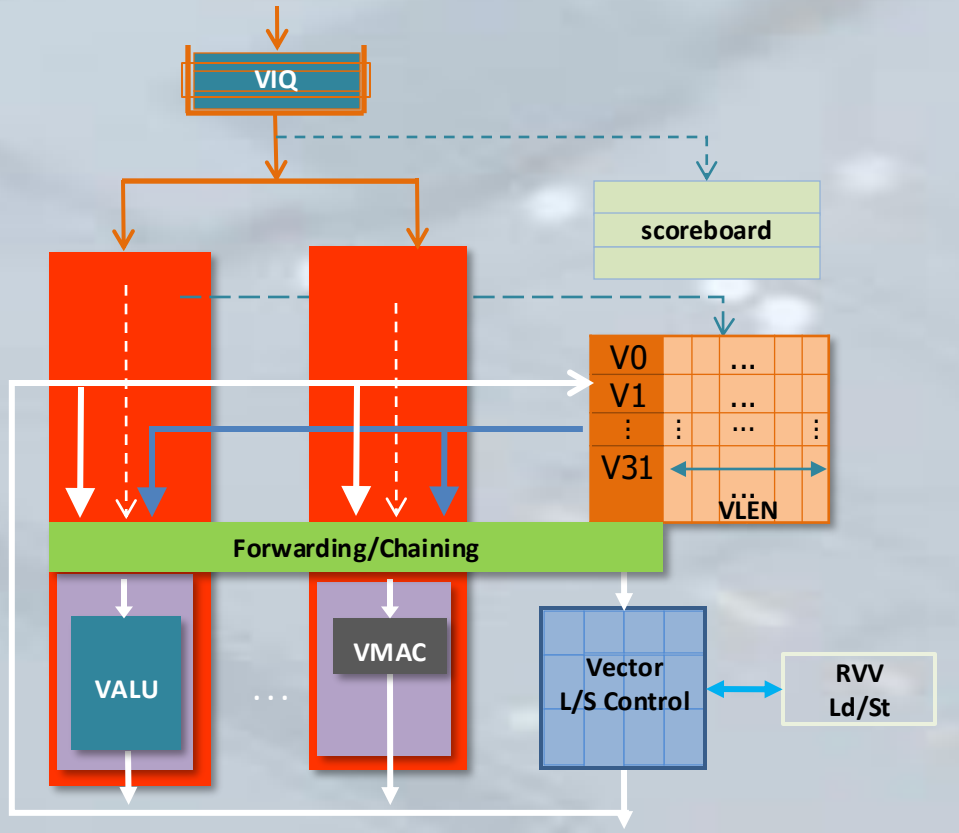
Andes Vector Processor Families

AX25-V100 is adopted in Meta's training and Inference Accelerator (MTIA) v1.

			AX47MPV*
		AX46MPV*	more features
	AX45MPV	Integrated Matrix Ext. (IME)	
	8-core cluster	16-core cluster with private L1/L2	
	HVM (High-speed Vector Memory) Interface		
AX25-V100	NX27V	ACE for RVV	
ACE (Andes Automated Custom Extension™)			
int4~64, fp16~64; bf16 (conv.)	+bf16 (full arithmetic)	+fp8	
VLEN: 128, 256, 512	VLEN: 128, 256, 512,1024	+2048	
RVV 0.8	RVV 1.0		
5-stage single-issue	8-stage dual-issue with shared cache for multicore		

* Future products subject to change

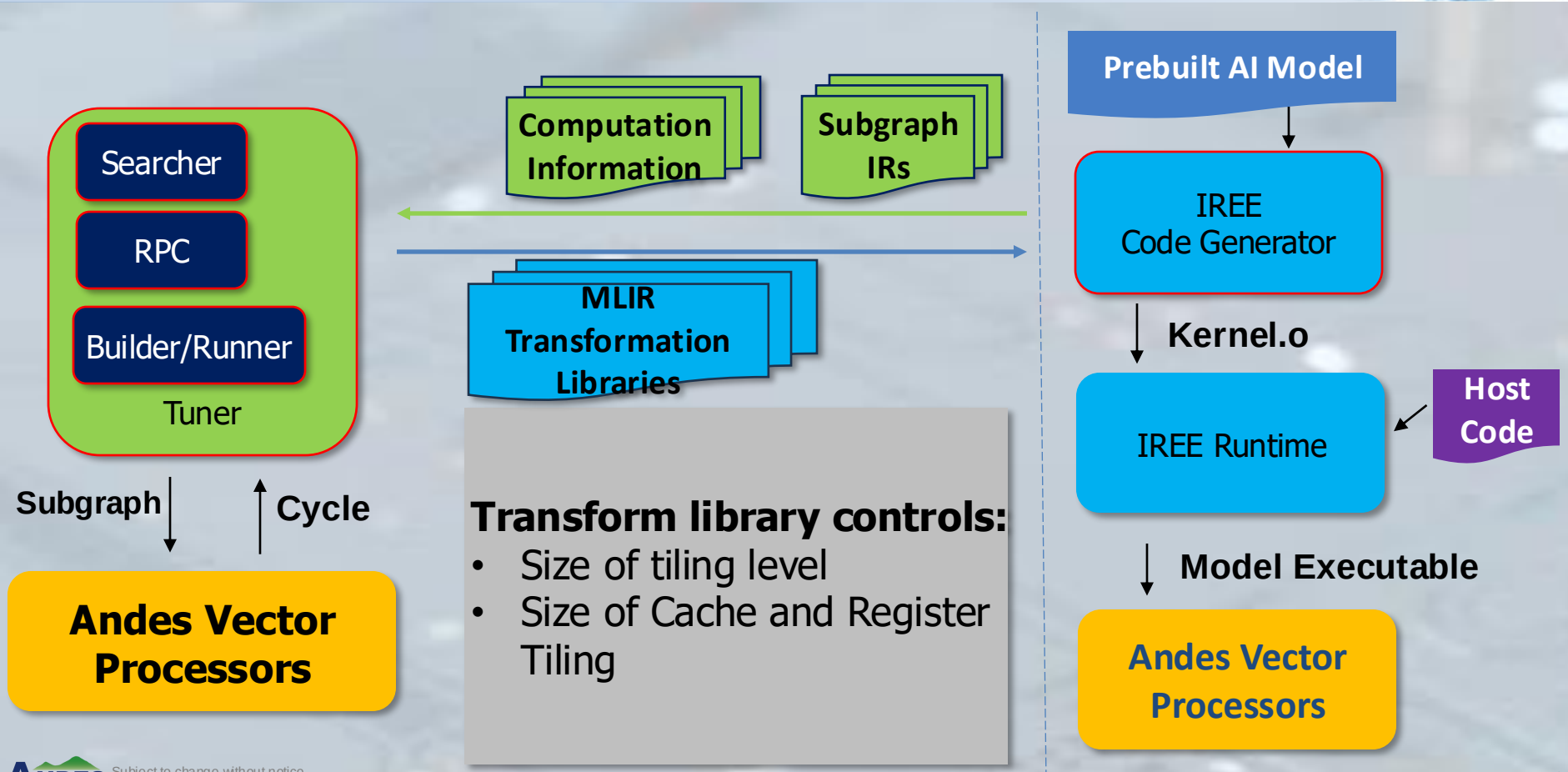
AX45MPV VPU Features



- Dual Issue/Dispatch, Out-of-order execution
- Multiple Vector Functional Units (VFUs) operating independently and simultaneously
- Up to 5 DLEN results are generated per cycle
- Support precise exception
- (optional) ACE-RVV

VLEN/DLEN/BIU Combinations		
VLEN	**SIMD (DLEN)**	**BIU(AXI)**
1024	1024	512/256/128
1024	512	512/256/128
512	512	512/256/128
512	256	256/128
256	256	256/128
256	128	128
128	128	128

Implementation



Innovation: Subgraph Similarity Analysis

Problem

- The number of tuning iterations usually must increase with the number of subgraphs.
- However, increasing the number of tuning iteration is time-consuming.

model		Number of subgraphs	Speedup with 20000 iterations
Mobilenet fp32		Around 30	2.1~2.24
Mobilebert	W/O similarity analysis	Around 600	1.2
	W/ similarity analysis	Around 60 categories	2.1~2.7

Observation:

- Similar subgraphs can share the same transformation library.

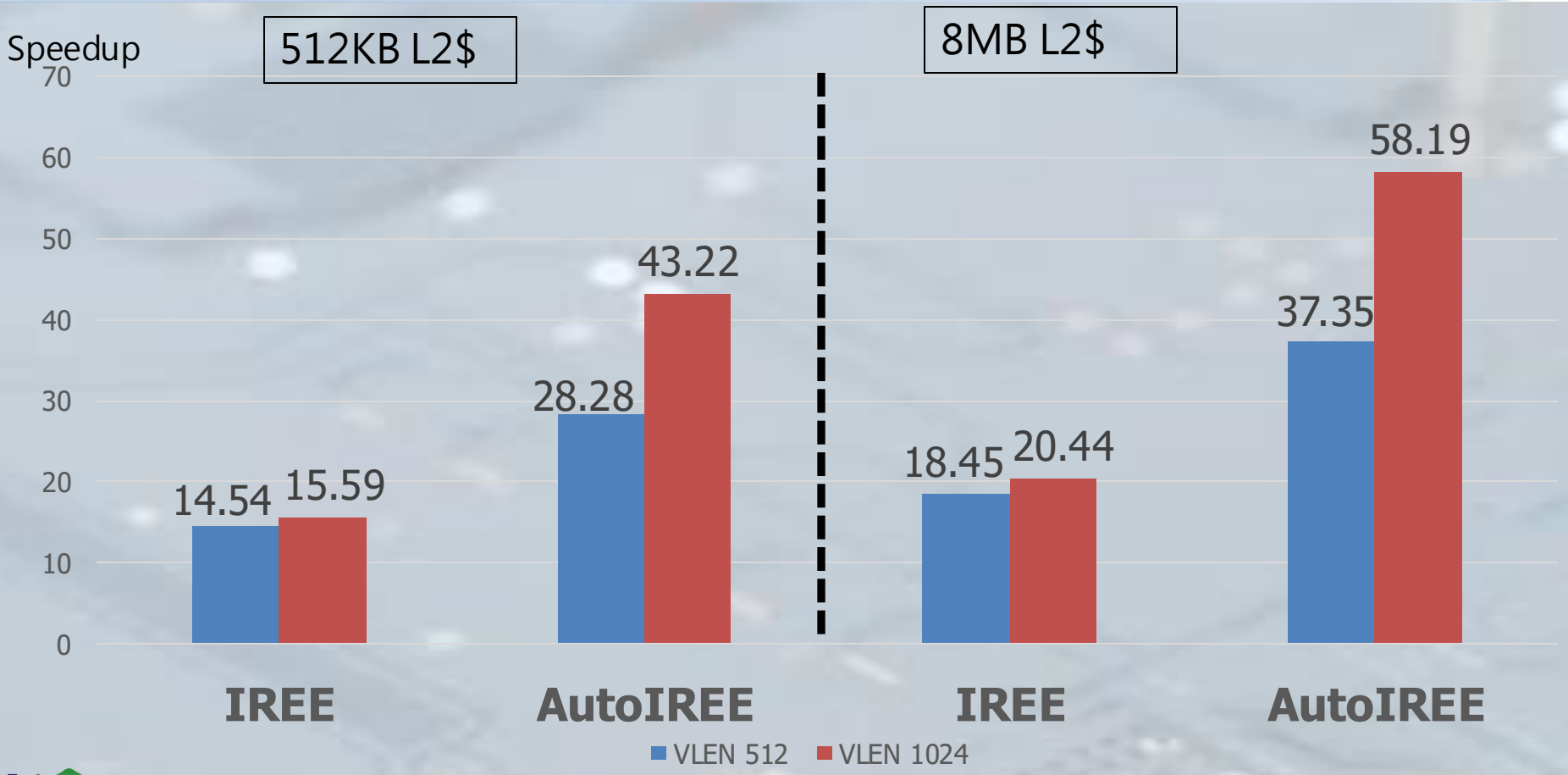
Solution:

- Using runtime profiling and static computation information to group 600 subgraphs into around 60 categories so that subgraphs in the same category can be tuned together.

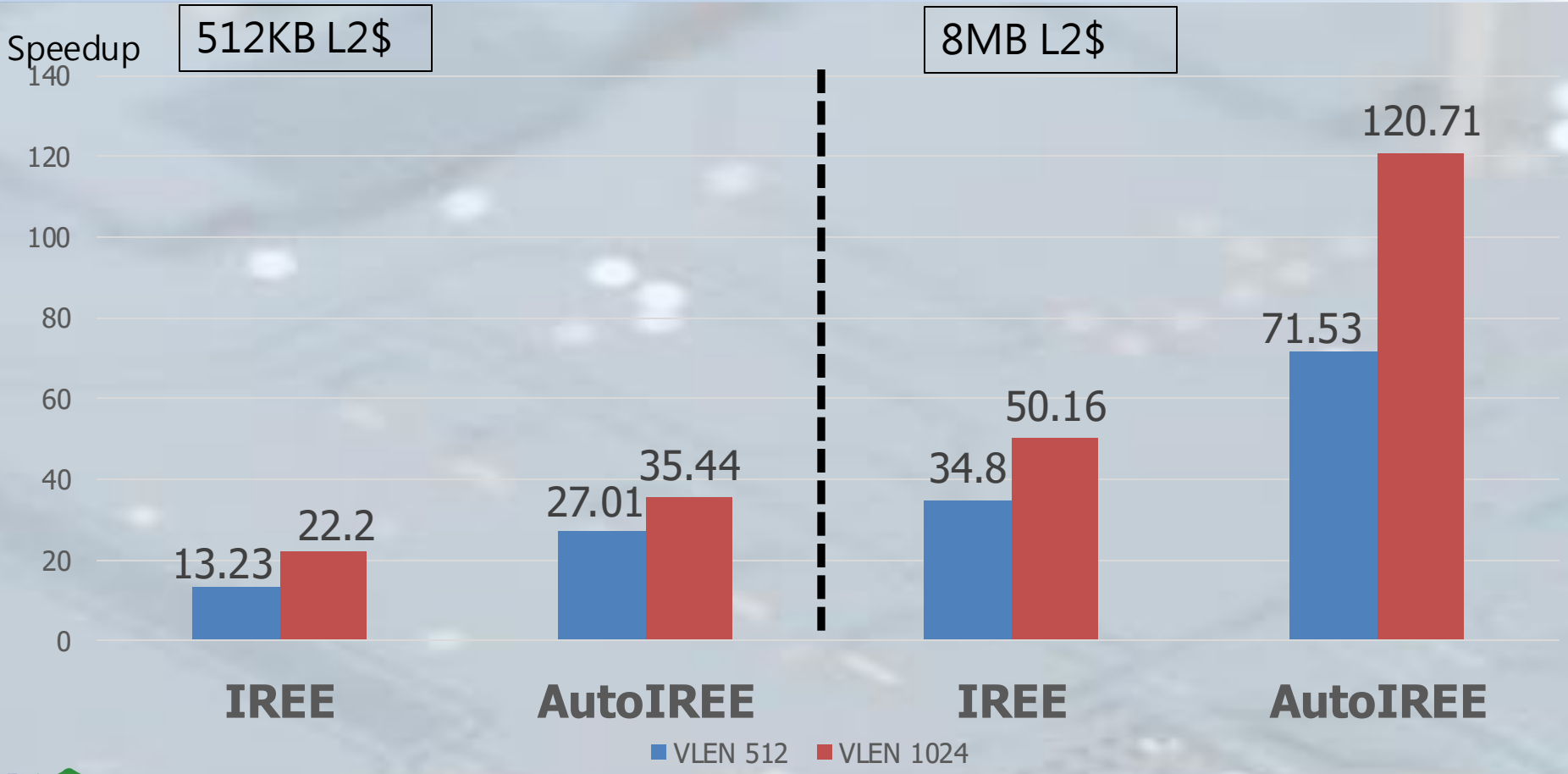
Experiment Environment

- **IREE commit:** ‘afd7cab’ with Andes patches.
  - With upstream LLVM package.
- **Models**
  - Mobilenet v1 int8 and fp32 (Image classification for mobile device)
  - Mobilebert int8 (Language model for mobile device)
- **Search algorithm**
  - Genetic evolution, 20000 iterations.
- **RISC-V Vector Processors on FPGA**
  - Andes AX45MPV (one core, DLEN/SIMD width = VLEN, BIU = 512 bits)
    - VLEN: 512 and 1024
    - L2$: 512KB and 8MB

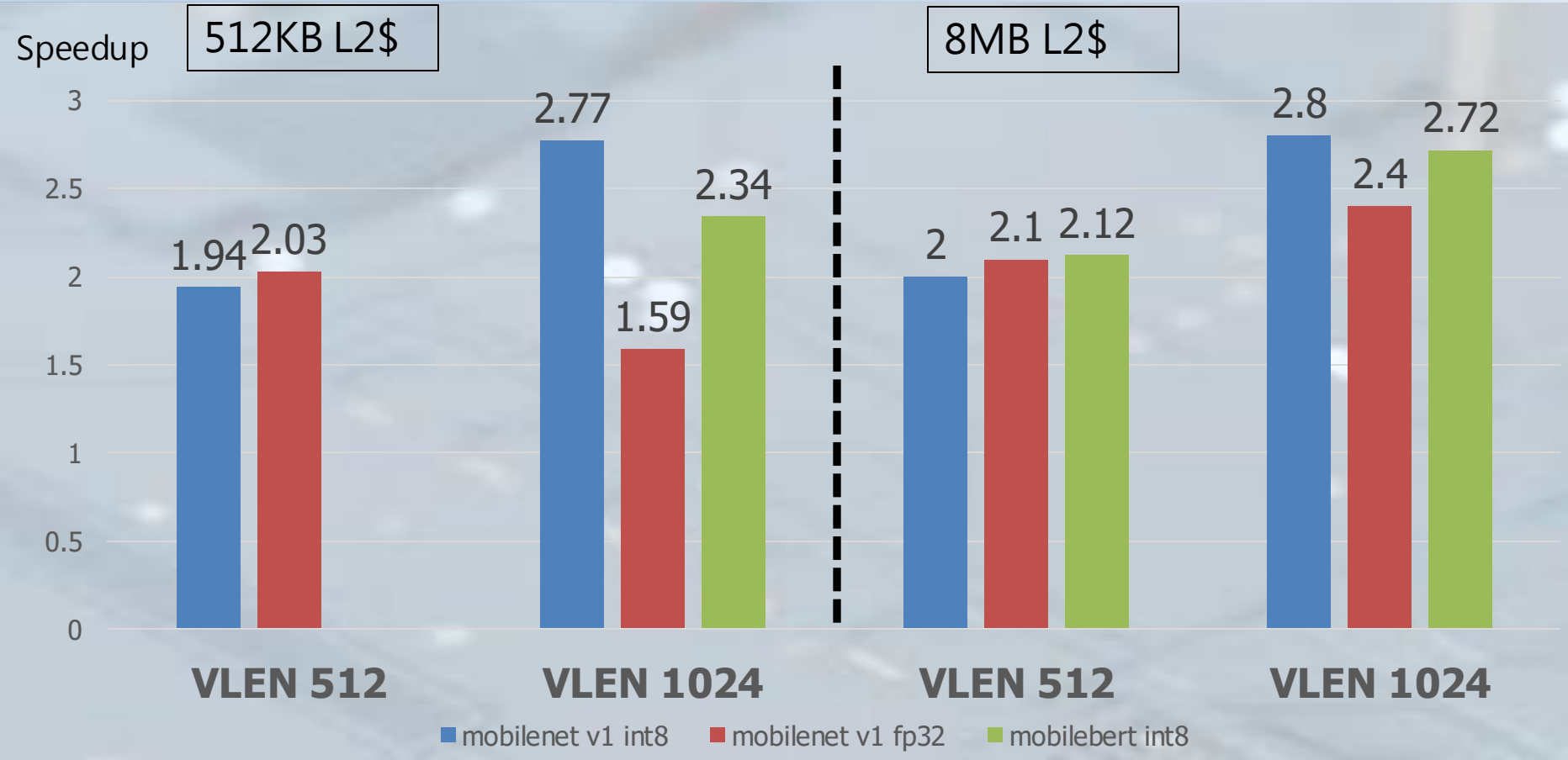
Mobilenet v1 INT8, Speedup of RVV over Pure Scalar



Mobilenet v1 FP32, Speedup of RVV over Pure Scalar



Speedup of AutoIREE W/ RVV over IREE W/ RVV



Future Works

- **Complex targets**
  - Multicore (AX45MPV)
  - Heterogeneous computing (Andes QiLai SoC, AX45MP + NX27V)
- **Reduce Compilation/Tuning Time**
  - Heuristic early stop
  - AI-based cost model to predict cycle

ANDES
TECHNOLOGY