tinyML® Talks

Enabling Ultra-low Power Machine Learning at the Edge

"Minimizing resource usage in microcontrollers for cost effective solutions"

Ilya Gozman – Senior Fellow, Chief AI Architect, Grovety

September 26, 2023

TINY ML

www.tinyML.org

Thank you, **tinyML Strategic Partners**,

for committing to take tinyML to the next Level, together

Executive Strategic Partners

Qualcomm AI research

Advancing AI research to make efficient AI ubiquitous

Power efficiency

Model design, compression, quantization, algorithms, efficient hardware, software tool

Personalization

Continuous learning, contextual, always-on, privacy-preserved, distributed learning

Efficient learning

Robust learning through minimal data, unsupervised learning, on-device learning

A platform to scale AI across the industry

Perception

Object detection, speech recognition, contextual fusion

Reasoning

Scene understanding, language understanding, behavior prediction

Action

Reinforcement learning for decision making

Edge cloud

Cloud

IoT/IIoT

Automotive

Mobile

Accelerate Your Edge Compute

SYNTIANT

Making Edge AI A Reality

www.syntiant.com

Platinum Strategic Partners

AITRIOS

DEPLOY VISION AI AT THE EDGE AT SCALE

SONY

Gold Strategic Partners

ANALOG DEVICES
AHEAD OF WHAT'S POSSIBLE™

ANALOG DEVICES
AHEAD OF WHAT'S POSSIBLE™

Where what if becomes what is.

Witness potential made possible at analog.com.

Build the Future of tinyML

on arm

TINY ML TALKS webcast

EDGE IMPULSE

The Leading Development Platform for Edge ML

edgeimpulse.com

Driving decarbonization and digitalization. Together.

Infineon serving all target markets as
Leader in Power Systems and IoT

www.infineon.com

infineon

innatera

NEUROMORPHIC INTELLIGENCE FOR THE SENSOR-EDGE

www.innatera.com

Microsoft

Renesas is enabling the next generation of AI-powered solutions that will revolutionize every industry sector.

renesas.com

RENESAS
BIG IDEAS FOR EVERY SPACE

STMicroelectronics provides extensive solutions to make tiny Machine Learning easy

life.augmented

www.st.com/ai

synaptics

ENGINEERING EXCEPTIONAL EXPERIENCES

We engineer exceptional experiences for consumers in the home, at work, in the car, or on the go.

www.synaptics.com

Silver Strategic Partners

Join Growing tinyML Communities:

16.9k members in
49 Groups in 41 Countries

tinyML - Enabling ultra-low Power ML at the Edge

https://www.meetup.com/tinyML-Enabling-ultra-low-Power-ML-at-the-Edge/

4k members
&
13k followers

The tinyML Community

https://www.linkedin.com/groups/13694488/

TINY ML TALKS webcast

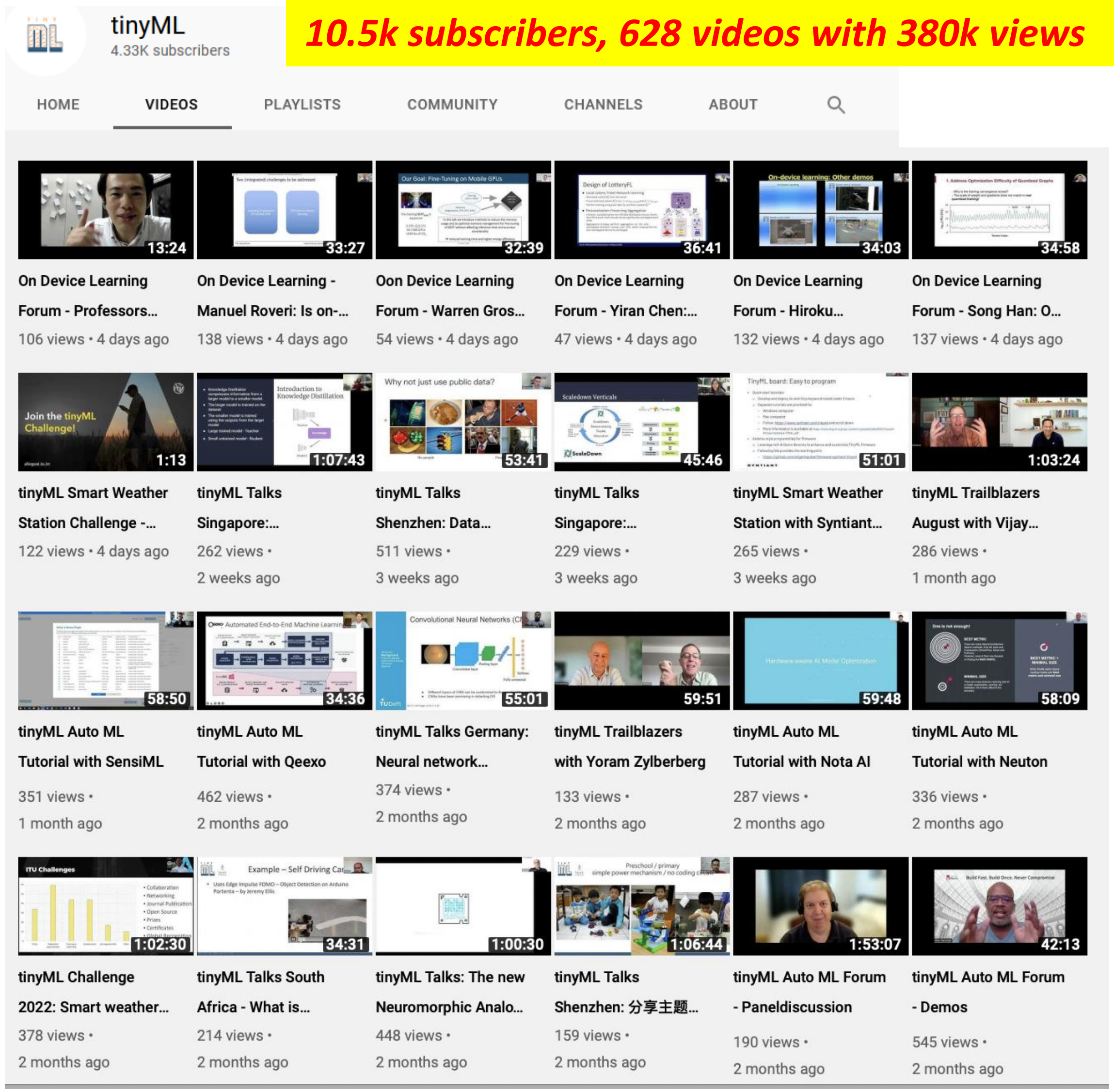
Subscribe to tinyML YouTube Channel for updates and notifications

(including this video)

www.youtube.com/tinyML

YouTube

10.5k subscribers, 628 videos with 380k views



TINY ML FOUNDATION

Register now
https://www.tinyml.org/event/asia-2023/

https://www.wevolver.com/article/2023-edge-ai-technology-report

Reminders

Slides & Videos will be posted tomorrow

tinyml.org/forums

youtube.com/tinyml

Please use the Q&A window for your questions

Ilya Gozman

Ilya is a Senior Fellow and a Chief AI Architect at Grovety, where he worked out his way from a rising talent developer to a veteran expert in AI, a frontline and prospective trend in IT-industry in recent years. He acquired extensive experience in developing general and AI compilers, and chip architectures both in LLVM and TVM backend optimizations; he also led teams working on compiling-related projects, video processing, and protocols support for IP cameras (C/C++). Ilya received Master degree in Applied Mathematics and Computer Science in 2007. Wide range of projects and profound research activity makes Ilya's experience valuable and demanded.

TinyML - growing interest

Edge AI allows business to improve the AI applications' overall cost-effectiveness by optimal use of NNs, computing resources and power consumption reduction.

At the same time, the numerous potential benefits of Edge AI face several challenges associated with its implementation and usability. [1]

Fine-Tuning Strategies

Model modification:
- Compress off-the-shelf networks by pruning and quantization
- Simplify unsupported operations to primitive blocks
- Transform and merge network layers
- Optimize resource-intensive layers

Inference time optimization:
- Use hardware-specific acceleration instructions

Memory requirements optimization:
- Optimize schedule of the operation flow
- Store weights on external storage

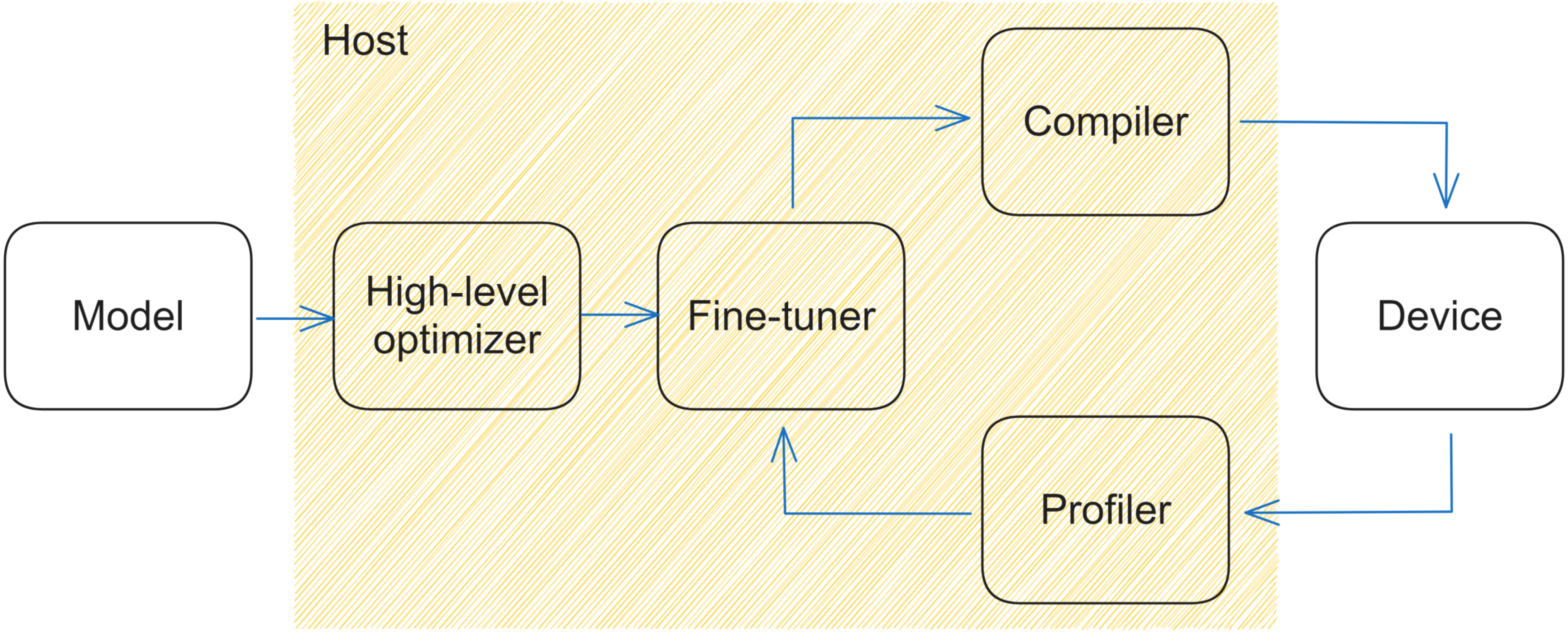
Energy Efficiency
- Throttling MCU/NPU operating frequencies
- Use advantages of heterogenous systems
- Intelligent power management

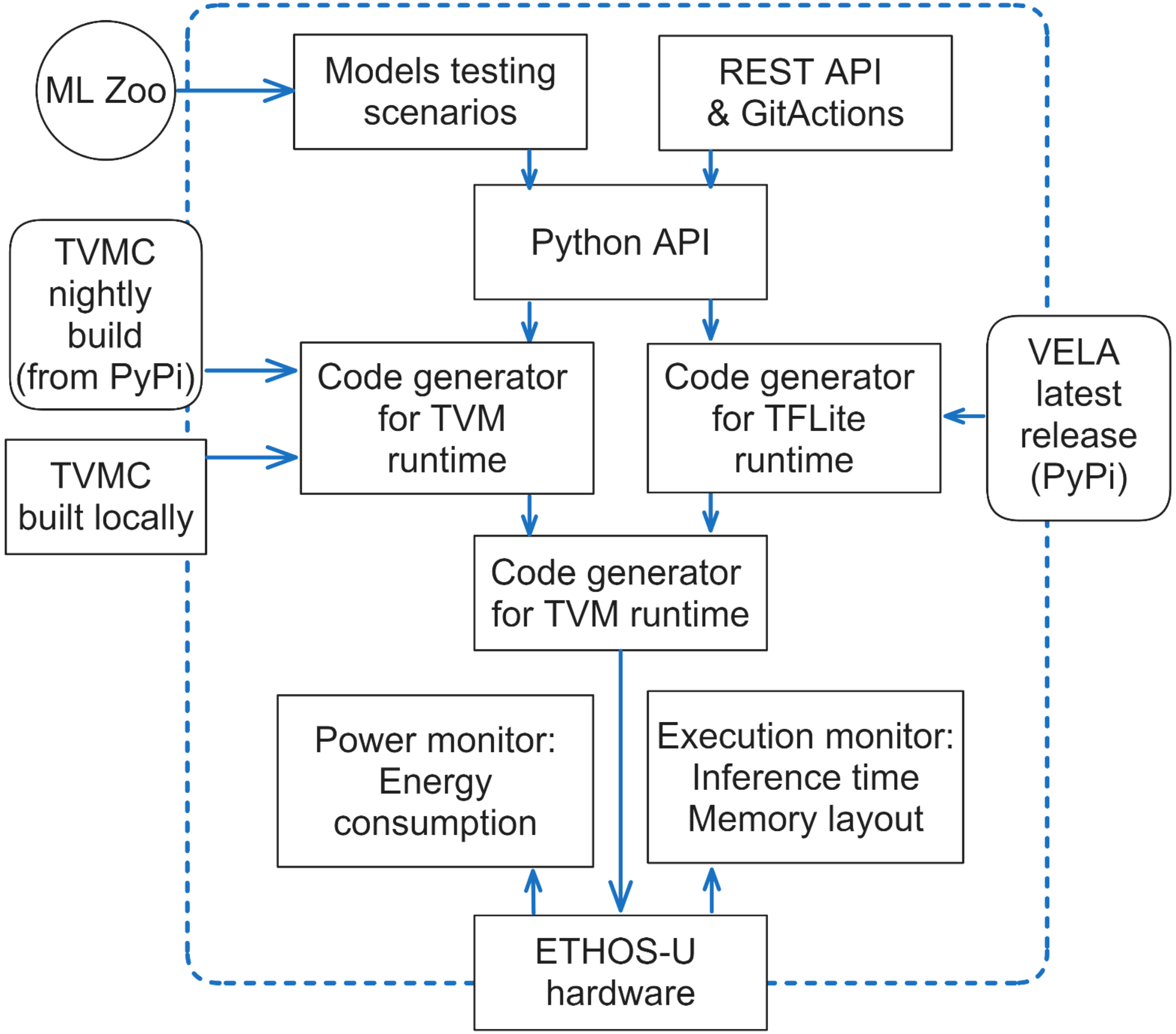
Approaches to handle Cost Challenges in TinyML

Minimizing Development costs and time of device

Reducing Device cost and its power consumption

Fine-Tuning Strategy





Fine tuning platform for CI tests and experiments on target HW

- NN inference on Alif hardware and FVP simulator
- Run on TFLiteMicro and TVM runtimes
- Support of any TVM commit
- Unified API for running NN inferences, various architectures and runtimes
- Actual inference time and power consumption measurements
- Model bottlenecks analysis and numerical mismatches

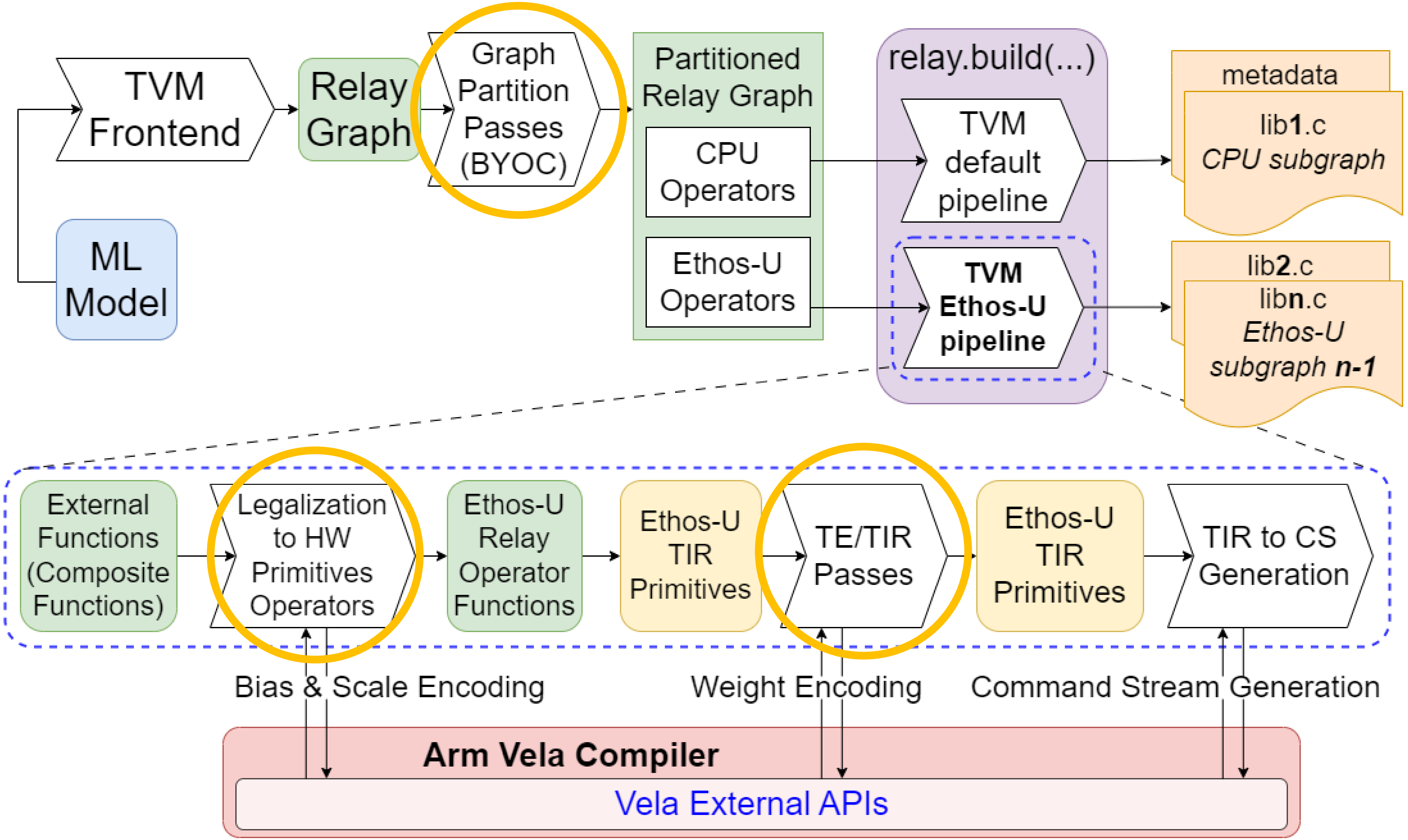
TVM: ML Compiler Framework

Wide range of ML frameworks and deployment targets

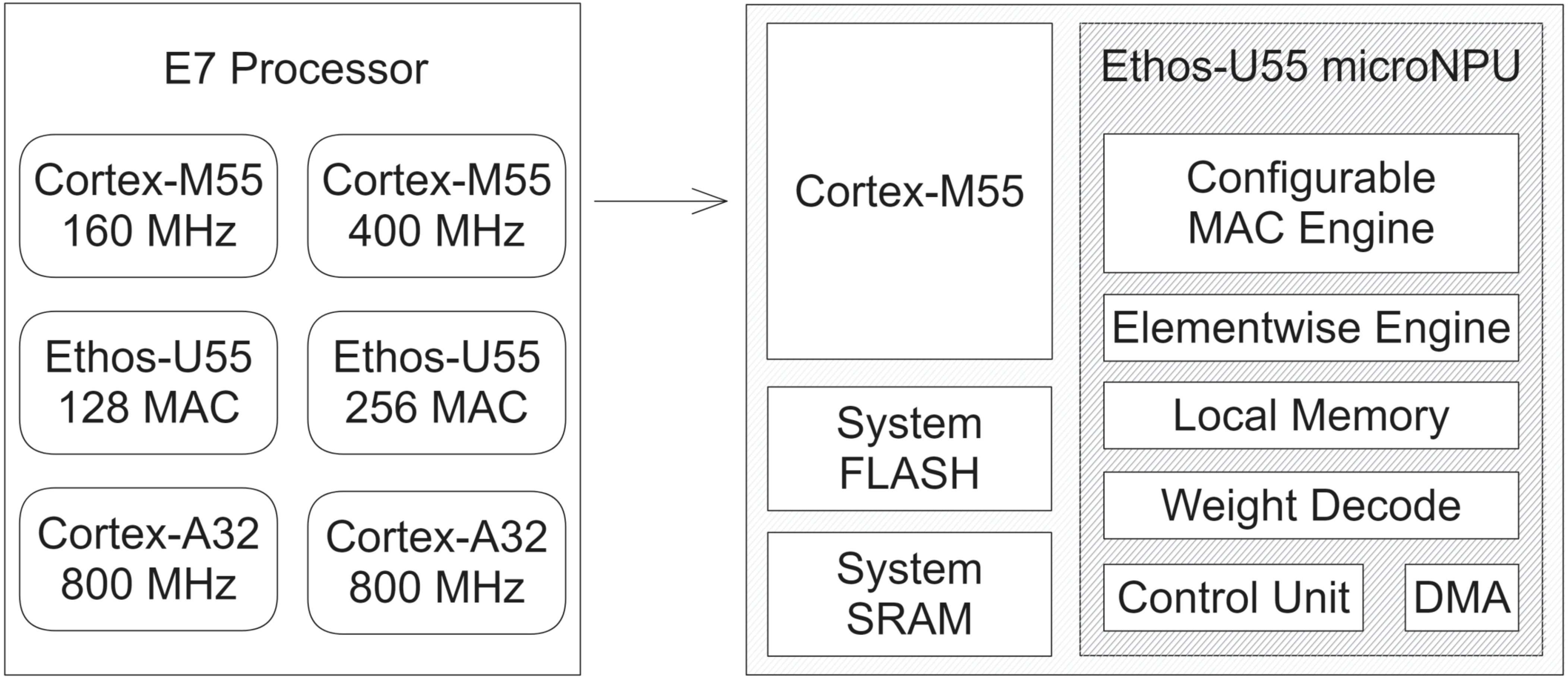
Open-source project, large community

Integrated with ARM® Vela Compiler for acceleration on Ethos™-U55 NPU

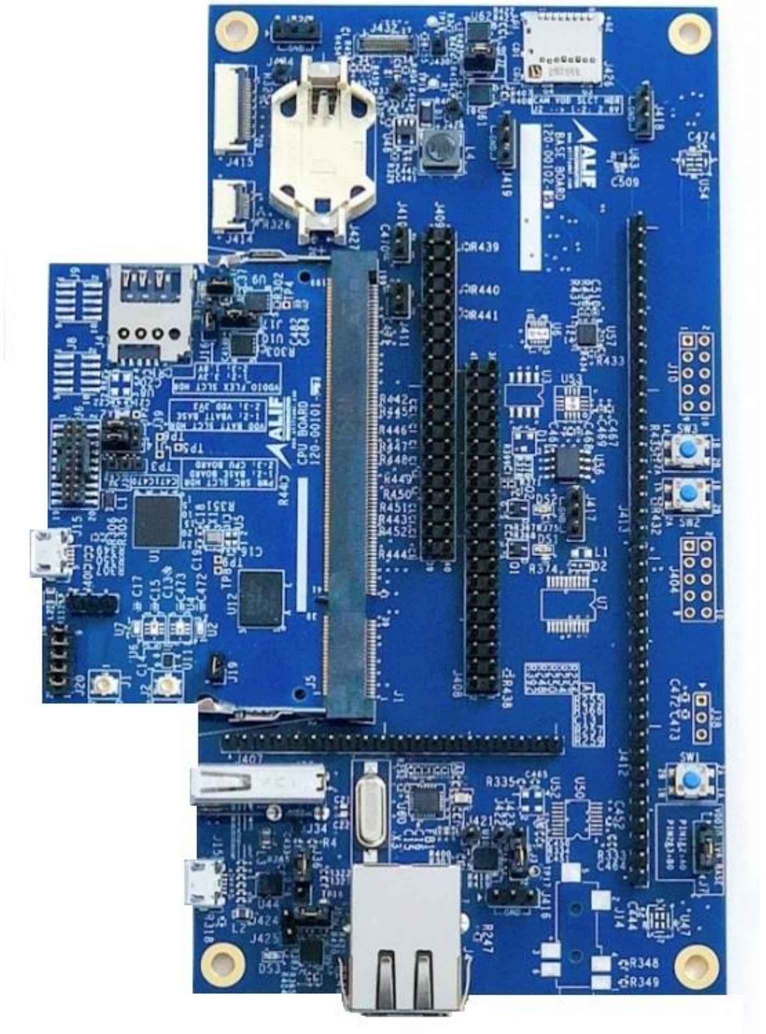
Fine-grained control over model compilation, deployment and execution

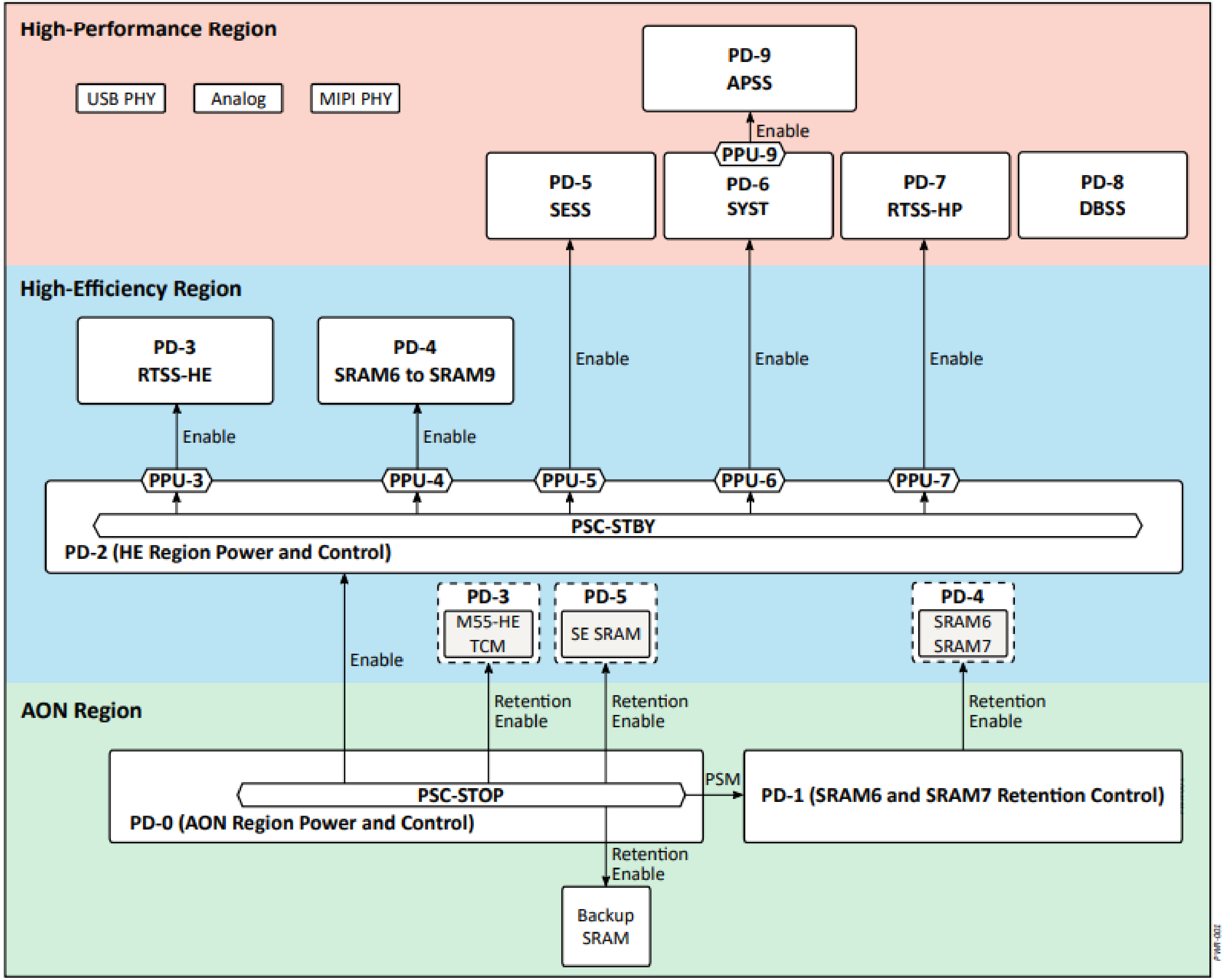

TVM Frontend
Relay Graph
Graph Partition Passes (BYOC)
Partitioned Relay Graph
CPU Operators
Ethos-U Operators
relay.build(...)
TVM default pipeline
TVM Ethos-U pipeline
metadata
lib1.c
CPU subgraph
lib2.c
libn.c
Ethos-U subgraph n-1
ML Model
External Functions (Composite Functions)
Legalization to HW Primitives Operators
Ethos-U Relay Operator Functions
Ethos-U TIR Primitives
TE/TIR Passes
Ethos-U TIR Primitives
TIR to CS Generation
Bias & Scale Encoding
Weight Encoding
Command Stream Generation
Arm Vela Compiler
Vela External APIs

Why Ethos-U and Alif Ensemble SoC



Our Experience with Alif E5





Pooling with high strides

Benefits:

Inference speed-up: +25%

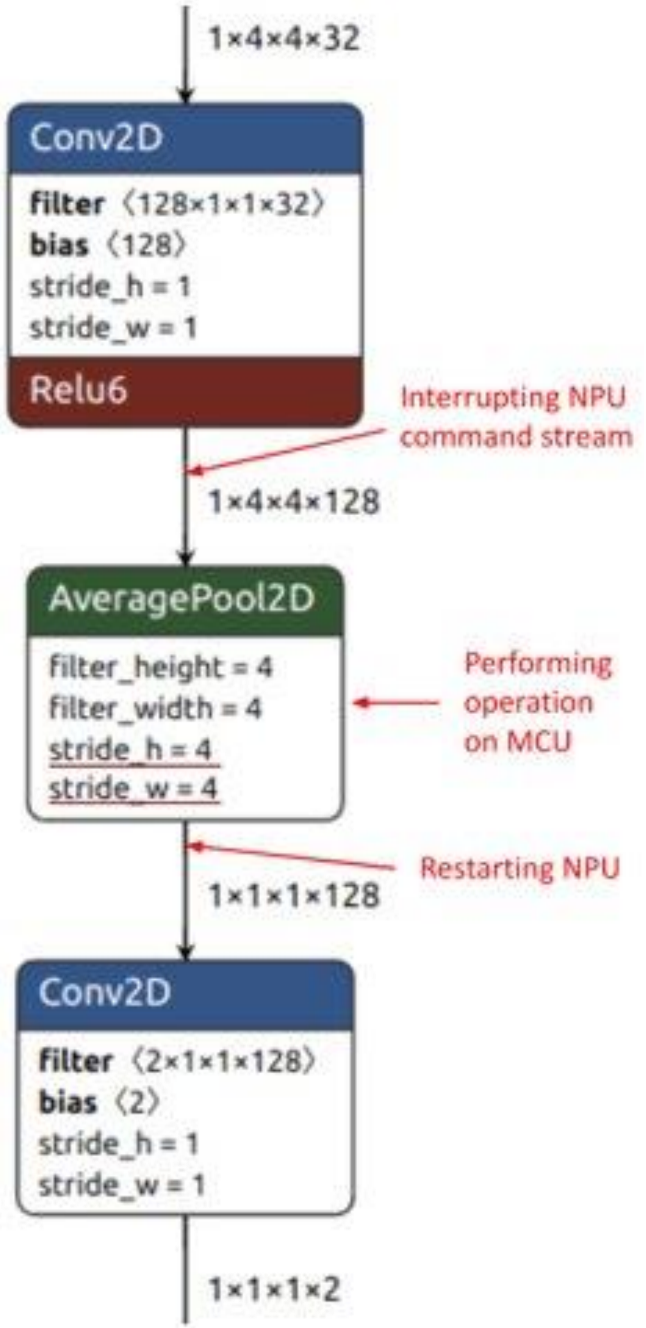
ARM ML Zoo Models affected: ~17%

```
cmsis-nn        <-      cmsis-nn.qnn_avg_pool2d
cmsis-nn        <-          %213 = cast(%212, dtype="int32")
cmsis-nn        <-          %214 = nn.avg_pool2d(%213, pool_size=[4, 4], strides=[4, 4], padding=[0, 0, 0, 0], layout="NHWC")
cmsis-nn        <-          %215 = cast(%214, dtype="int8")
```

```
strides = params.strides
if params.strides[0] > 3 or params.strides[1] > 3:
    strides = [1, 1]
```

Here we already know that IFM.shape == kernel.shape

```
ethos-u         <-      ethos-u.avgpool2d
ethos-u         <-          %213 = cast(%212, dtype="int32")
ethos-u         <-          %214 = nn.avg_pool2d(%213, pool_size=[4, 4], strides=[4, 4], padding=[0, 0, 0, 0], layout="NHWC")
ethos-u         <-          %215 = cast(%214, dtype="int8")
```



Padding over channel axis

Benefits:
Inference speed-up: 250% - 400%
ARM ML Zoo Models affected: ~10%

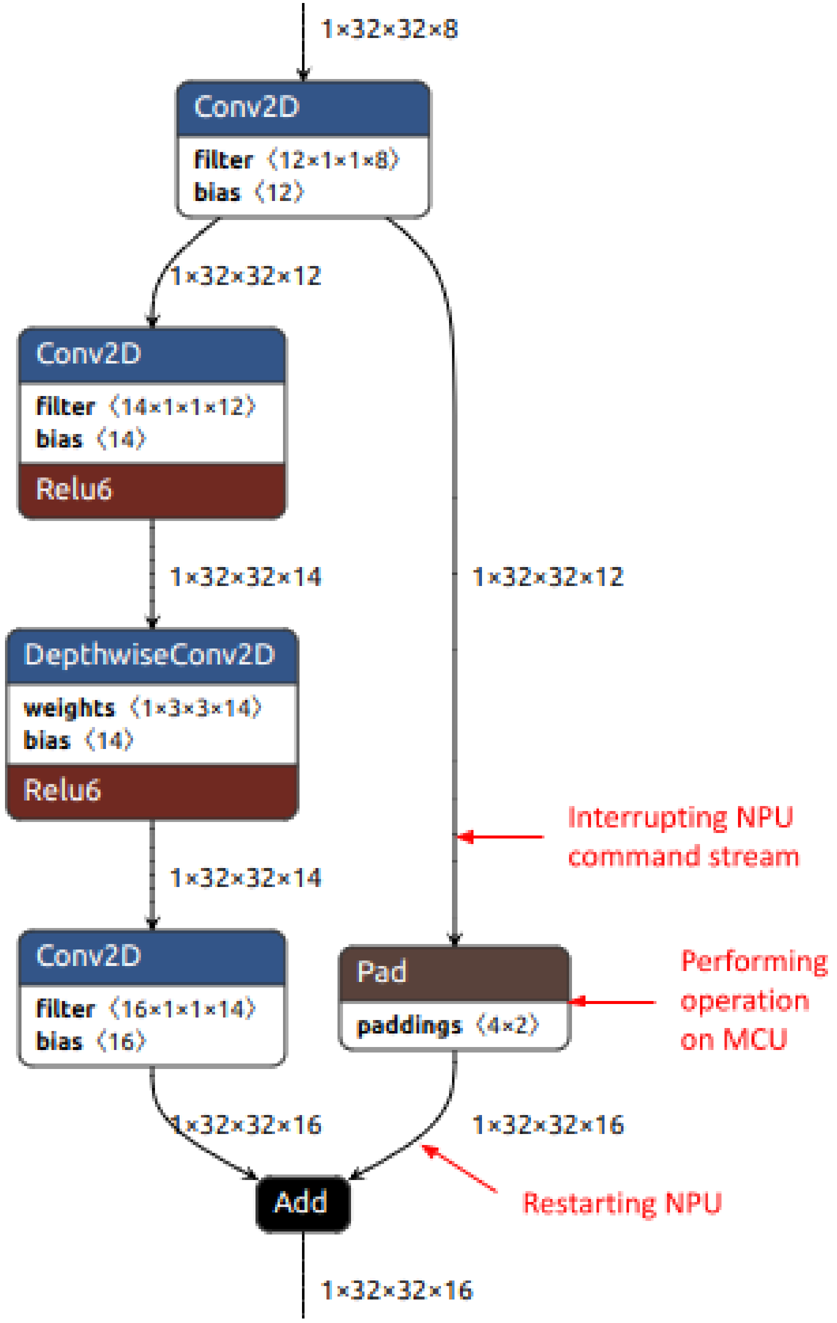
```
# pad channels before
if params.ch_padding[0] > 0:
    identity1 = ethosu_ops.ethosu_identity(pad_values, ...)
    concat_args.append(identity1)

identity2 = ethosu_ops.ethosu_identity(ifm.tensor, ...)
concat_args.append(identity2)

# pad channels after
if params.ch_padding[1] > 0:
    identity3 = ethosu_ops.ethosu_identity(pad_values, ...)
    concat_args.append(identity3)

relay.op.concatenate(relay.Tuple(concat_args), axis=3)
```

Create a memory area with padding values "before"

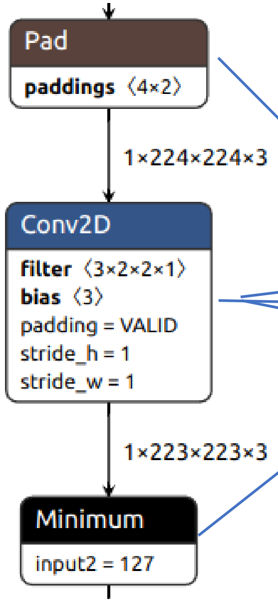
Our actual channel data

Create a memory area with padding values "after"

Concatenate everything together over channel axis



Understanding TVM's patterns

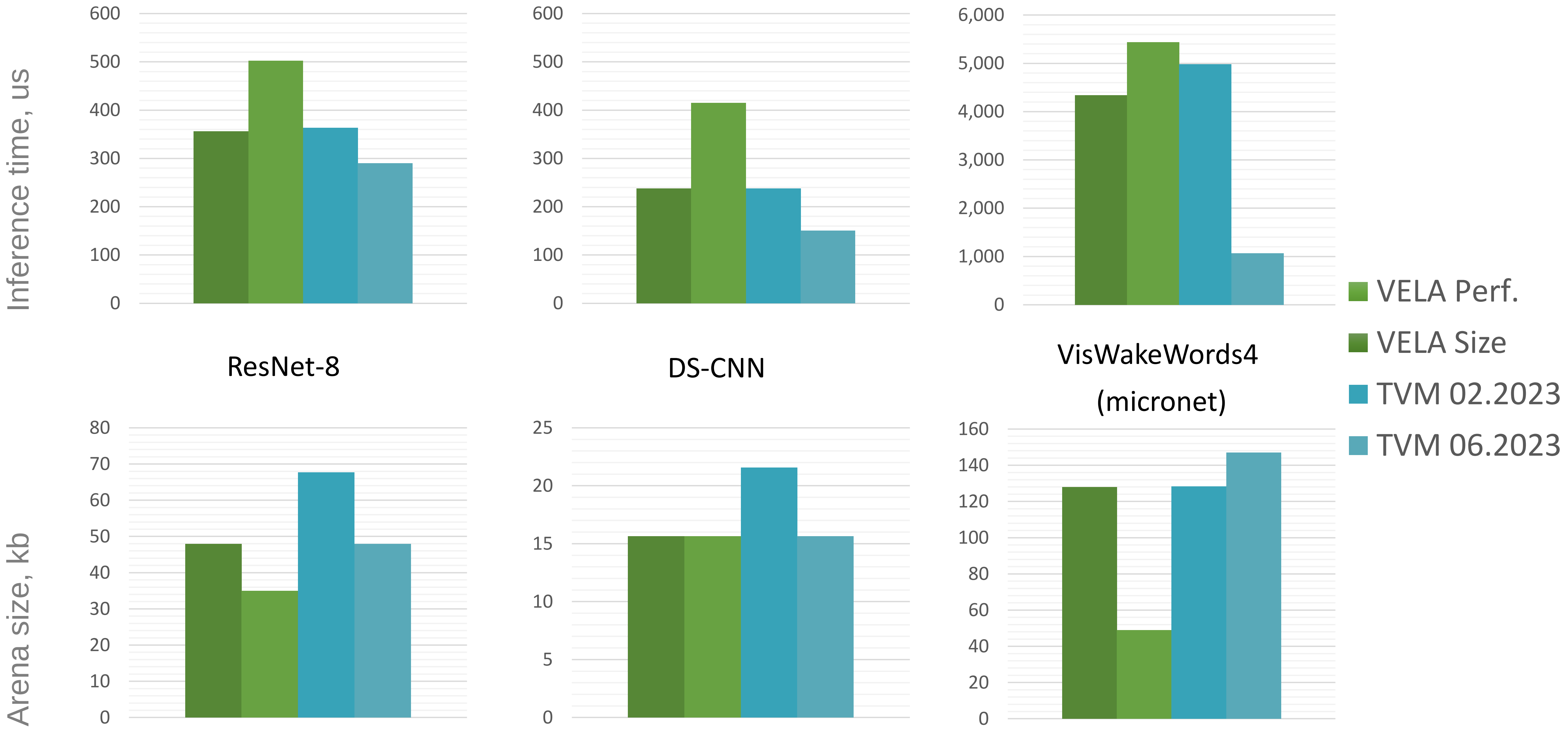



```
nn.pad(...)
qnn.conv2d(...)
nn.bias_add(...)
qnn.requantize(...)
clip(...)
```

```
optional_pad = is_op("nn.pad")(wildcard(), is_constant())
qnn_conv2d = is_op("qnn.conv2d")(
    optional_pad | wildcard(),
    is_constant(),
    is_constant(),
    is_constant(),
    is_constant(),
    is_constant(),
).has_attr({"kernel_layout": "HWIO"})
bias_add = is_op("nn.bias_add")(qnn_conv2d, is_constant())
req = is_op("qnn.requantize")(
    bias_add, is_constant(), is_constant(), is_constant(), is_constant()
)
clip_or_req = req.optional(is_op("clip"))
```

```
ethos-u        <-        %0 = nn.pad(%data, -128
ethos-u        <-        %1 = qnn.conv2d(%0, %v_
ethos-u        <-        %2 = nn.bias_add(%1, %v
ethos-u        <-        %3 = qnn.requantize(%2,
ethos-u        <-        %4 = clip(%3, a_min=-12
```

```
Conv2D name=None:
    IFM: h=1,w=296,c=39, region=3, NHWC, INT8, size=11544, scale: 0.17129258811473846, zero: 4
      Stride y/x/c: 1/39/1, tiles: w0=296, h0=1, h1=0, base=['0x0', '0x0', '0x0', '0x0']
      name=None
    OFM: h=1,w=148,c=250, region=1, NHCWB16, INT8, size=37000, scale: 0.24689388275146484, zero: 3
      Stride y/x/c: 1/16/2368, tiles: w0=148, h0=1, h1=0, base=['0x0', '0x0', '0x0', '0x0']
      name=None
    Kernel: w=48, h=1, stride=(2, 1), dilation=(1, 1)
    NpuPadding(top=0, left=23, bottom=0, right=23)
    Weights: (region=0, address=0x1f6090, length=271472)
    Scales: (region=0, address=0x1b810, length=2512)
    Activation: TABLE_LOOKUP, min=None, max=None, lut index=0
    NpuBlockTraversal.PART_KERNEL_FIRST
    Block config: h=2,w=60,c=64, NpuResamplingMode.NONE, NpuRoundingMode.TFL
```



Inference time, us
600
500
400
300
200
100
0
6,000
5,000
4,000
3,000
2,000
1,000
ResNet-8
DS-CNN
VisWakeWords4
(micronet)
Arena size, kb
80
70
60
50
40
30
20
10
25
15
5
160
140
120
80
VELA Perf.
VELA Size
TVM 02.2023
TVM 06.2023

Open Challenges

- Per-layer Analysis: computational and memory usage
- The memory scheduling according to the overall network topology [2]
- Transitioning Network Weights to External Storage [5]
- On Device Learning
- Inference of multiple NNs on heterogeneous computing architectures
- Dig in Ethos-U Platforms specific
- More practice and experience on real applications

References

[1]: 2023 Edge AI Technology Report
https://www.wevolver.com/article/2023-edge-ai-technology-report

[2]: MCUNet: Tiny Deep Learning on IoT Devices
http://tinyml.mit.edu

[3]: Tiny Reservoir Computing for Extreme Learning of Motor Control
https://www.researchgate.net/publication/354752261_Tiny_Reservoir_Computing_for_Extreme_Learning_of_Motor_Control

[4]: Partha Pratim Ray, 2021. A review on TinyML: State-of-the-art and prospects
https://www.sciencedirect.com/science/article/pii/S1319157821003335

[5]: Miao, H. and Lin, F.X., 2021. Enabling Large Neural Networks on Tiny Microcontrollers with Swapping. arXiv preprint arXiv:2101.08744.
https://arxiv.org/abs/2101.08744

[6]: Work With microTVM
https://tvm.apache.org/docs/how_to/work_with_microtvm/index.html

[7]: Arm Ethos-N Processor Series, Product Brief
https://developer.arm.com/-/media/Arm%20Developer%20Community/PDF/AI-ML%20Datasheet%20and%20briefs/Arm%20Ethos-N%20Product%20Brief%20-%20May%202020.pdf

Grovety
AI Powered Devices
arm AI Partner

Thank you for your attention!

gozman@grovety.com

Copyright Notice

www.tinyml.org

Copyright Notice

www.tinyml.org