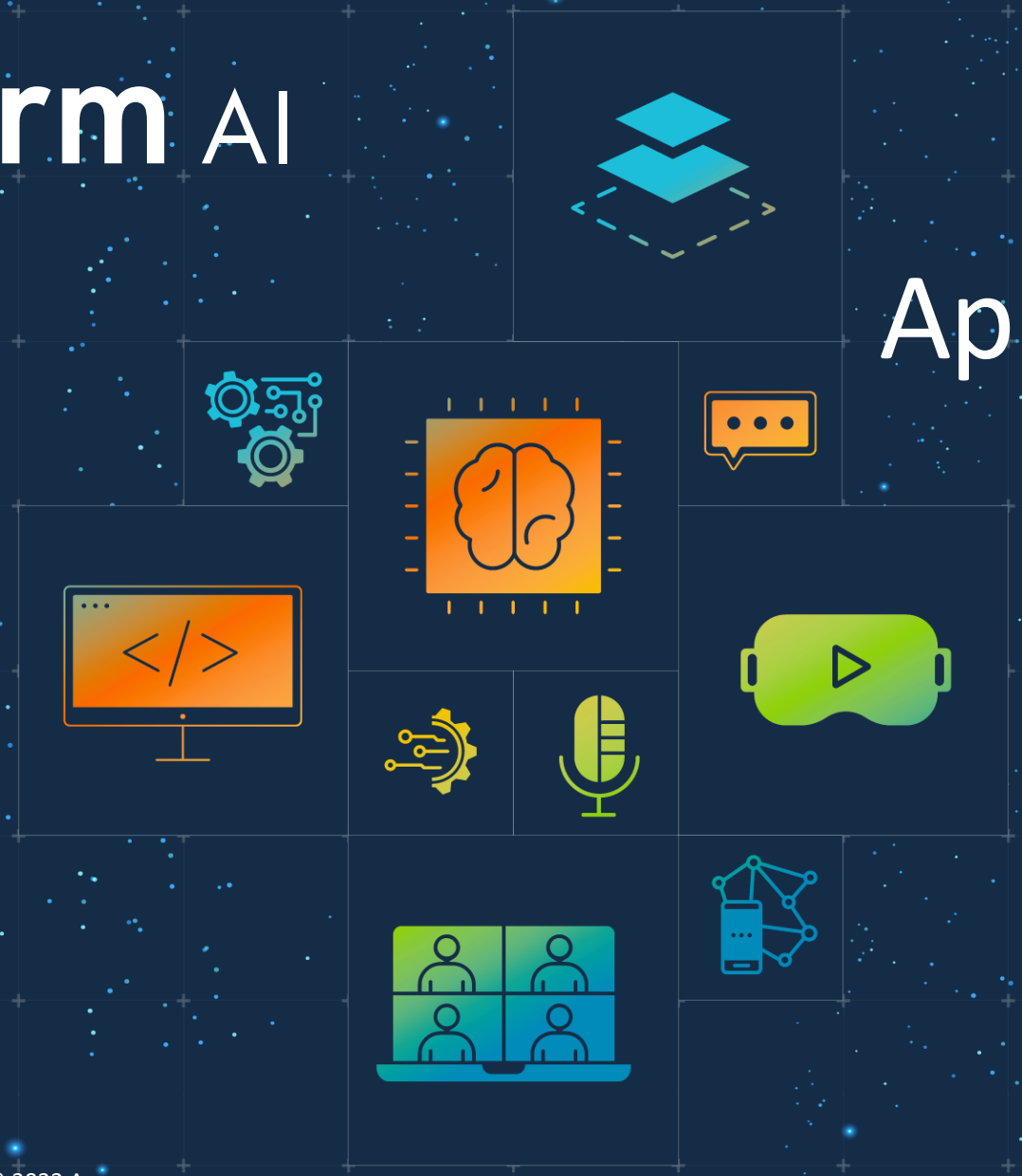
arm AI



Arm AI Tech Talks

A Hardware-aware Approach for Designing Neural Models on Arm Ethos-U65

Nota AI

Shinkook Choi, Tech Lead, Nota AI

28 June 2022

arm AI

Welcome!

Tweet us: @ArmSoftwareDev -> #AIVTT

Check out our Arm Software Developers YouTube channel

Signup now for our next AI Virtual Tech Talk: www.arm.com/techtalks

Our upcoming Arm AI Tech Talks

Date	Title	Host
28th June	Nota AI: A Hardware-aware Approach for Designing Neural Models	Nota
19th July	Talk on AI on Raspberry Pi (title TBD)	Raspberry Pi

Visit: www.arm.com/techtalks

Presenters

Shinkook Choi is a tech lead in Nota AI.

His research interests are improving the performance of AI model compression technology in various tasks such as image classification, object detection, and super-resolution.

Nota AI

- AI Model Optimization Company

History

2015 Founded

2021 Series B ($22M)

Strategic Investors

SAMSUNG VENTURE INVESTMENT

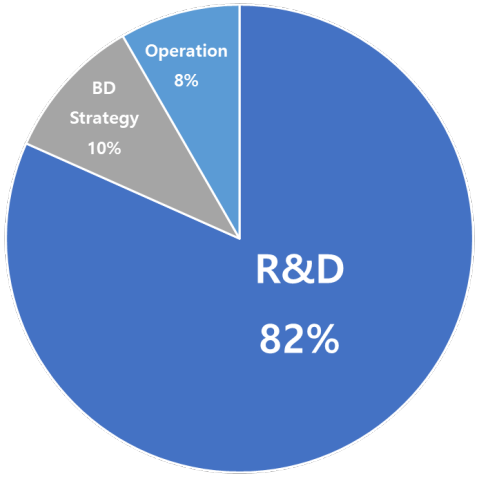
LG CNS

NAVER

kakaoinvestment

Location

Team



~70 Teammates, > 80% in R&D

arm AI

Nota AI

- **Our Technology**

1. AI Model Optimization

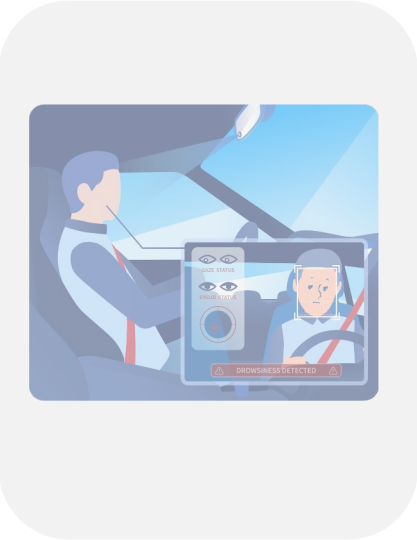
NetsPresso

Automated SW platform NetsPresso minimizes human resource input by automatically developing lightweight AI models

Optimized for devices NetsPresso creates optimized AI models for target devices and provides a wide range of devices as options

2. Edge AI Development and Optimization Services

Intelligent
Transportation System



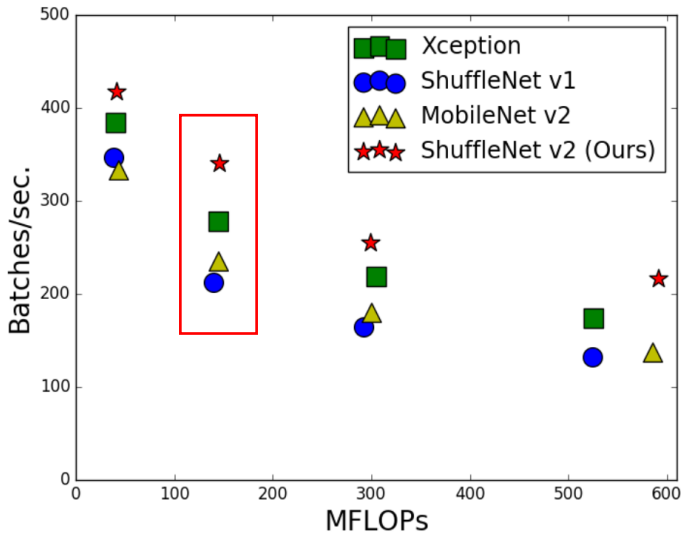
Driver
Monitoring System

Outline

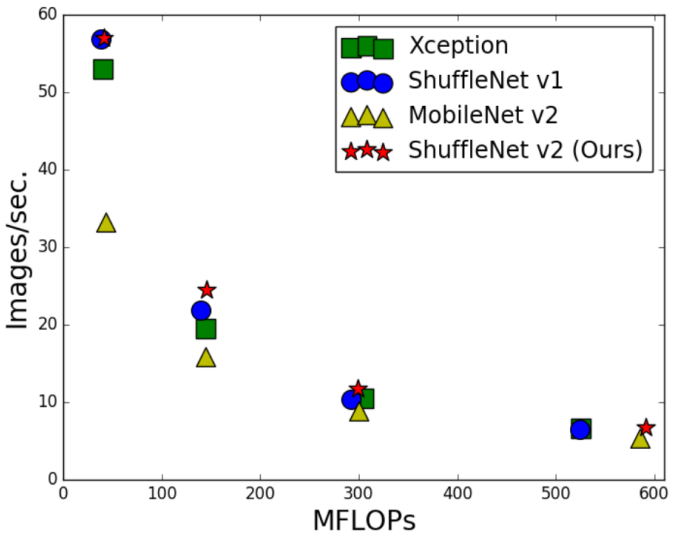
- Problem statement
- Arm Ethos-U65
- NetsPresso
- Results

Problem statement

- The number of float-point operations (FLOPs) is a widely used metric to measure computation complexity. However, FLOPs is not the same as the direct metric such as speed or latency.
- The figure below shows that networks with similar FLOPs have different speeds.



GPU

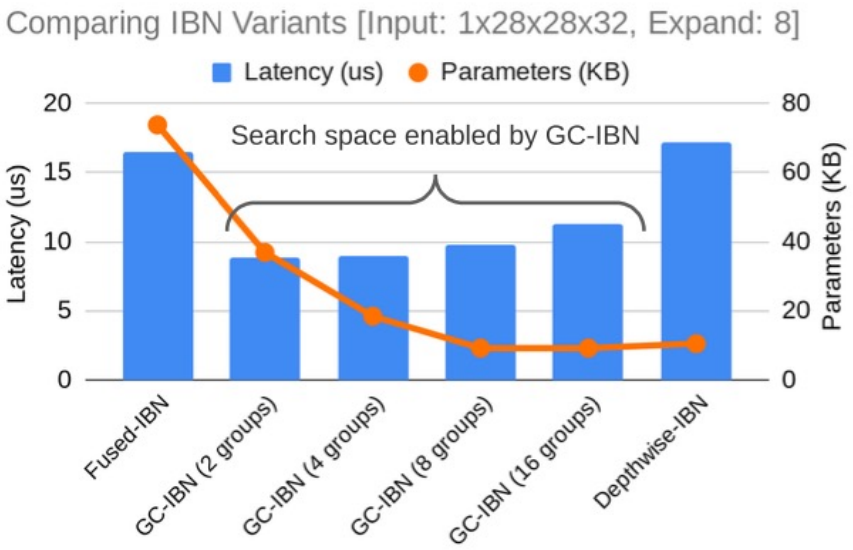


CPU

Ma, Ningning, et al. "Shufflenet v2: Practical guidelines for efficient cnn architecture design." *ECCV*. 2018.

Problem statement

- On various machine learning accelerators, not all FLOPs and the number of trainable parameters have the same efficiency.
- Fused-IBN may run the same as fast as a depthwise-IBN even with 7× as many Parameters.



- It is important to analyze the hardware and design the neural network accordingly.

Akin, Berkin, et al. "Searching for Efficient Neural Architectures for On-Device ML on Edge TPUs." *arXiv:2204.14007* (2022).

arm AI

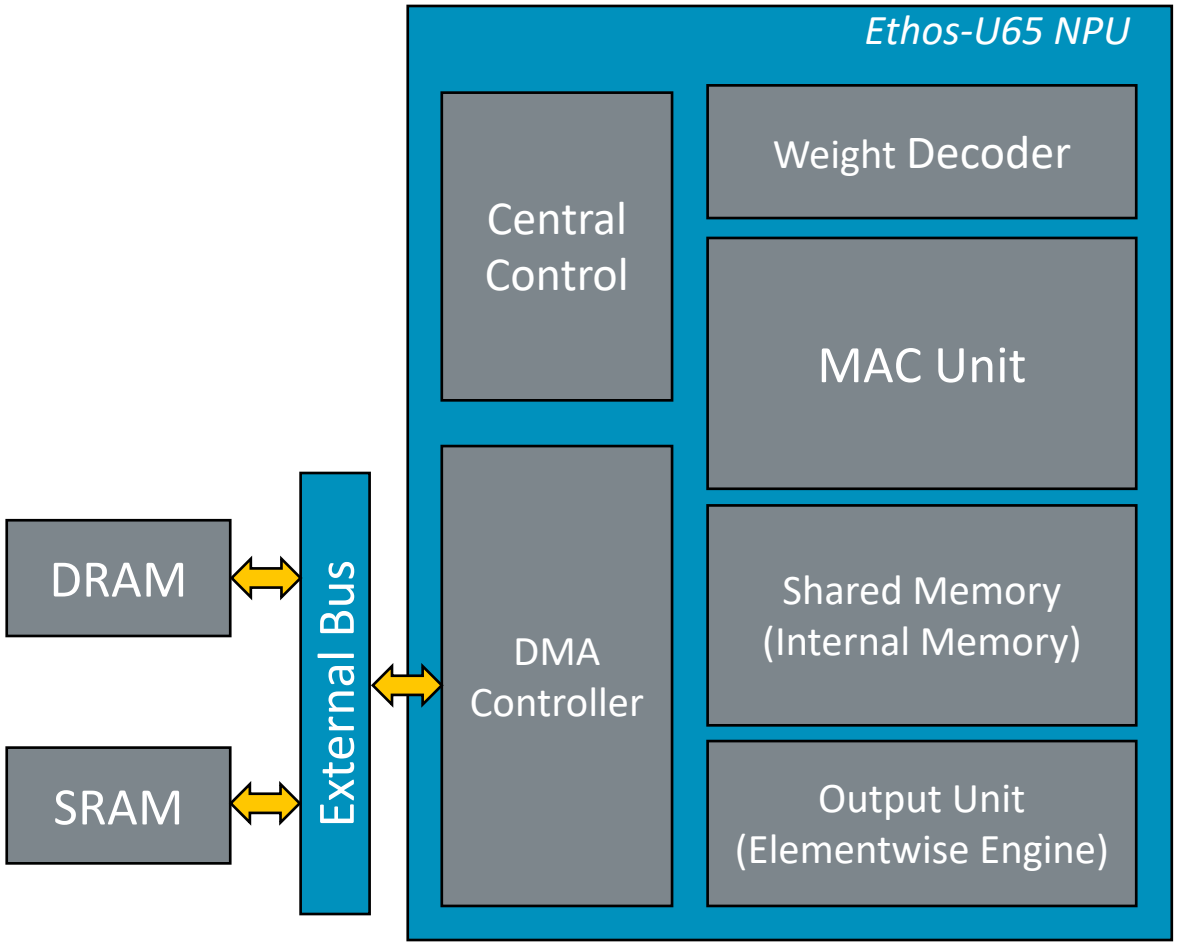
Arm Virtual Hardware (AVH)

- Arm Virtual Hardware (AVH) is an evolution of Arm's modeling technology delivering models of Arm-based processors, systems, third party hardware for application developers and SoC designers to build and test software before silicon and hardware availability.
- Fixed Virtual Platform (FVP)
  - digital twin of a development board with Ethos-U65 & Cortex-M55
- Corstone-300 (sse-300), available as part of Arm Virtual Hardware
- Device setting
  - Arm Ethos-U65 NPU
  - Optimize: Performance
  - 256 MAC units
  - Memory mode: Dedicated SRAM
  - System: High End

Arm Ethos-U65 NPU

- Device dedicated for Neural processing

- Consists of:
  - Computation Units
    - MAC Unit
    - Elementwise Engine
  - Memory
    - Internal Memory
    - External Memory: SRAM, DRAM

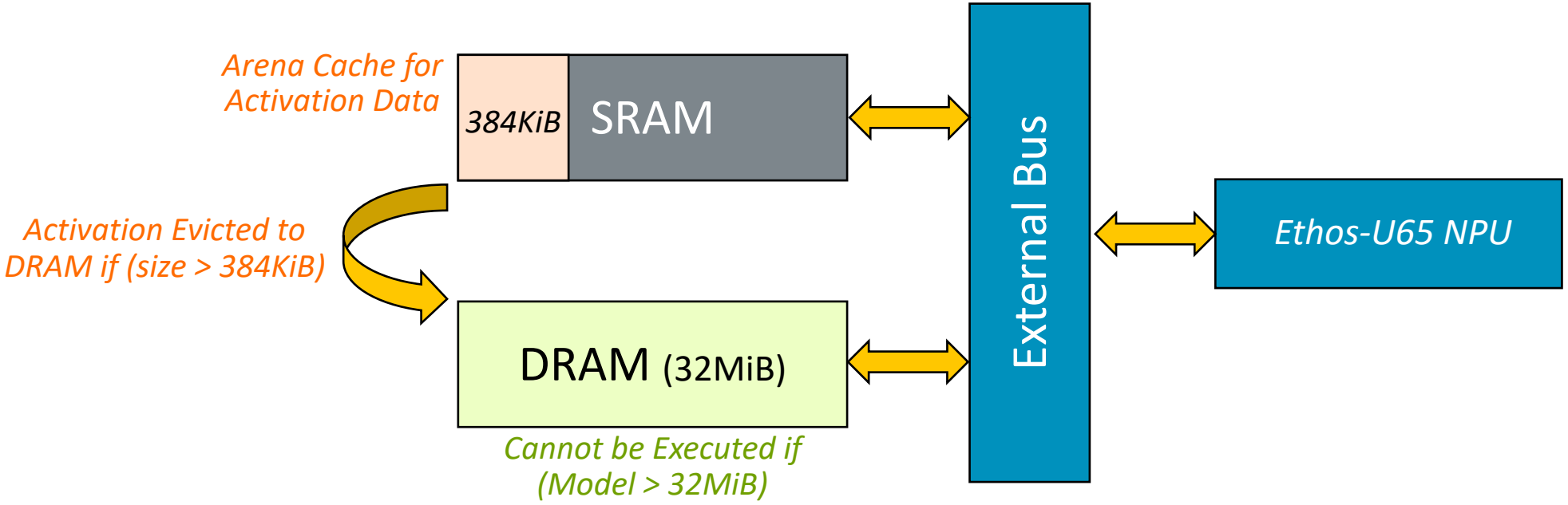


https://developer.arm.com/documentation/102023

* KiB is kibibyte which equals 2^{10} bytes
* MiB is mebibyte which equals 2^{20} bytes

Arm Ethos-U65 NPU

- When memory mode is dedicated SRAM, the arena cache size is 384 KiB
- If the activation buffer size is larger than 384 KiB, DRAM is used and latency increases dramatically.
- Since the memory size is 32 MiB, it cannot be executed if the sum of the model size and the activation buffer size is greater than 32 MiB.



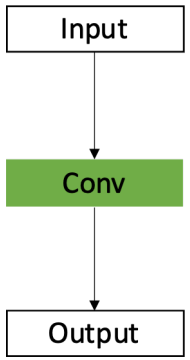
https://developer.arm.com/documentation/102023

PW: Pointwise Convolution

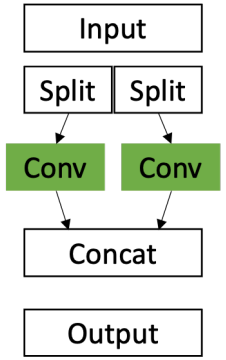
DW: Depthwise Convolution

Neural Network Blocks

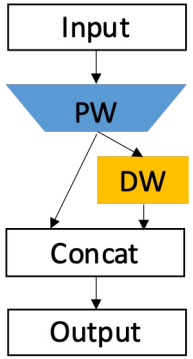
Convolution types



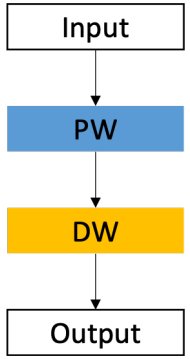
Conv



Group Conv

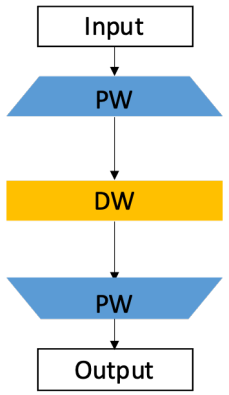


Ghost Conv[1]

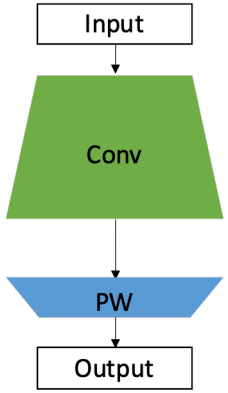


Blueprint Separable Conv (BS Conv)[2]

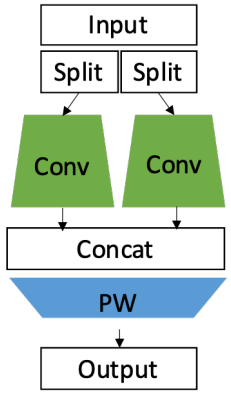
Building Blocks



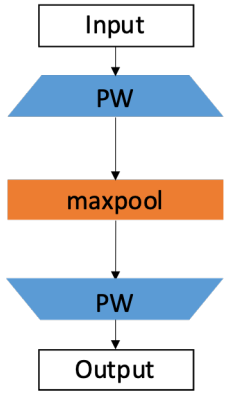
Inverted Bottleneck(IBN)



Fused-IBN[3]



Fused group convolution IBN (Fused-GC-IBN)[3]



maxpool-IBN[4]

[1] Han, et al. "Ghostnet: More features from cheap operations." *CVPR*. 2020.
[2] Haase, et al. "Rethinking depthwise separable convolutions: How intra-kernel correlations lead to improved mobilenets." *CVPR*. 2020.
[3] Akin, Berkin, et al. "Searching for Efficient Neural Architectures for On-Device ML on Edge TPUs." *arXiv:2204.14007* (2022).
[4] Han, Dongyoon, et al. "Learning Features with Parameter-Free Layers." *arXiv:2202.02777* (2022).

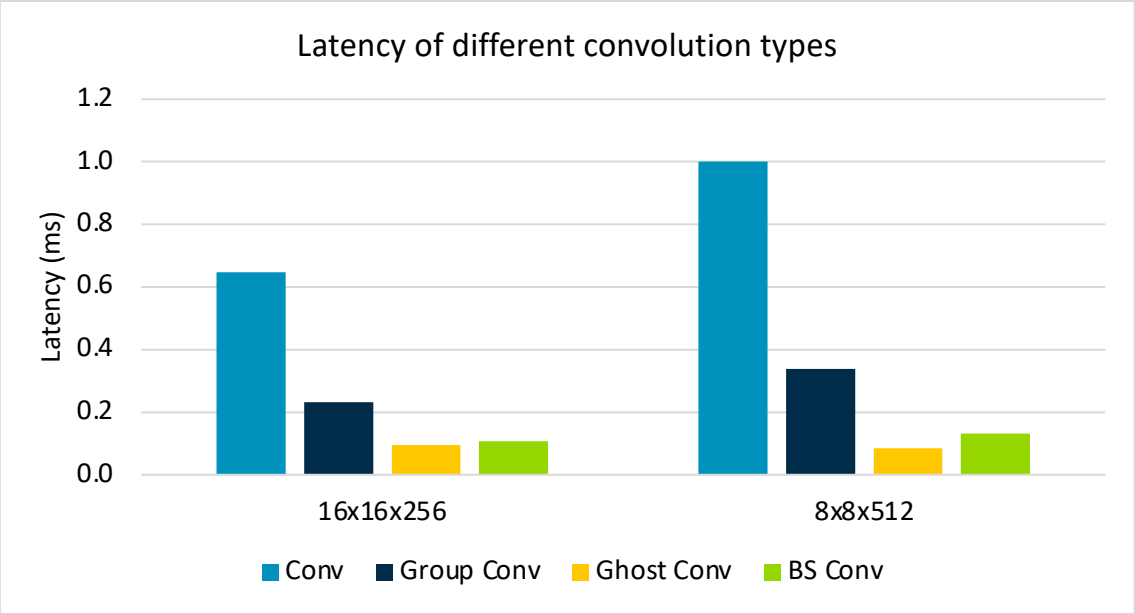
arm AI

* groups = 4 for Fused-GC-IBN

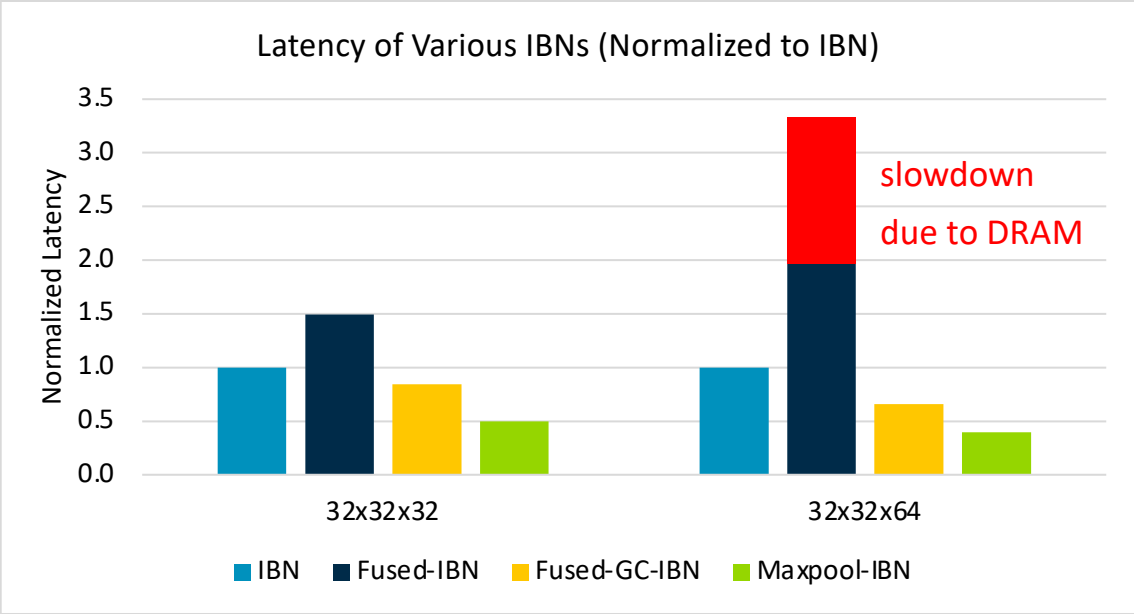
Neural Network Blocks

- Latency characteristics of blocks vary on the configuration

Convolution types



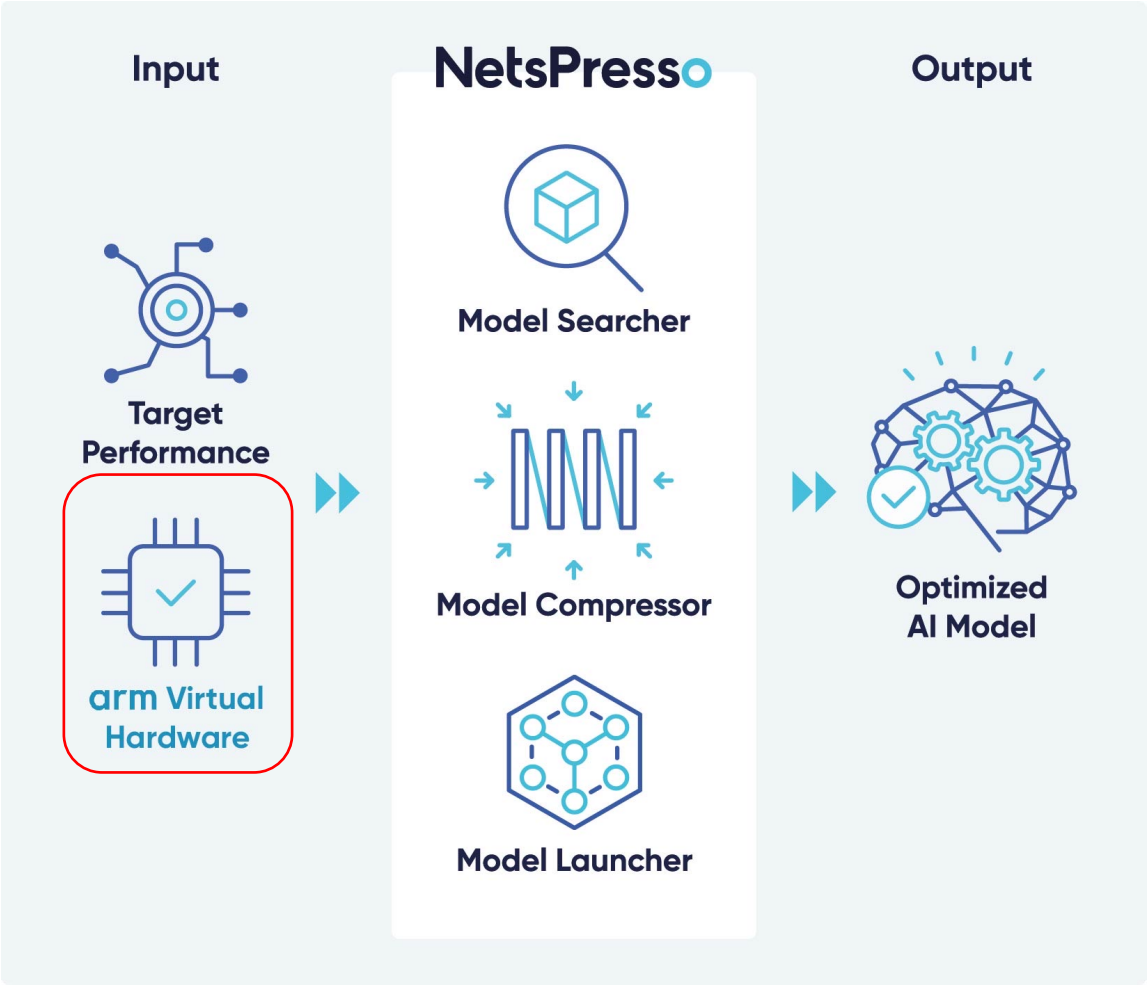
Building Blocks



* The value of the horizontal axis represents the value of $height \times width \times channels$ when the size of the input/output feature map is the same.

NetsPresso

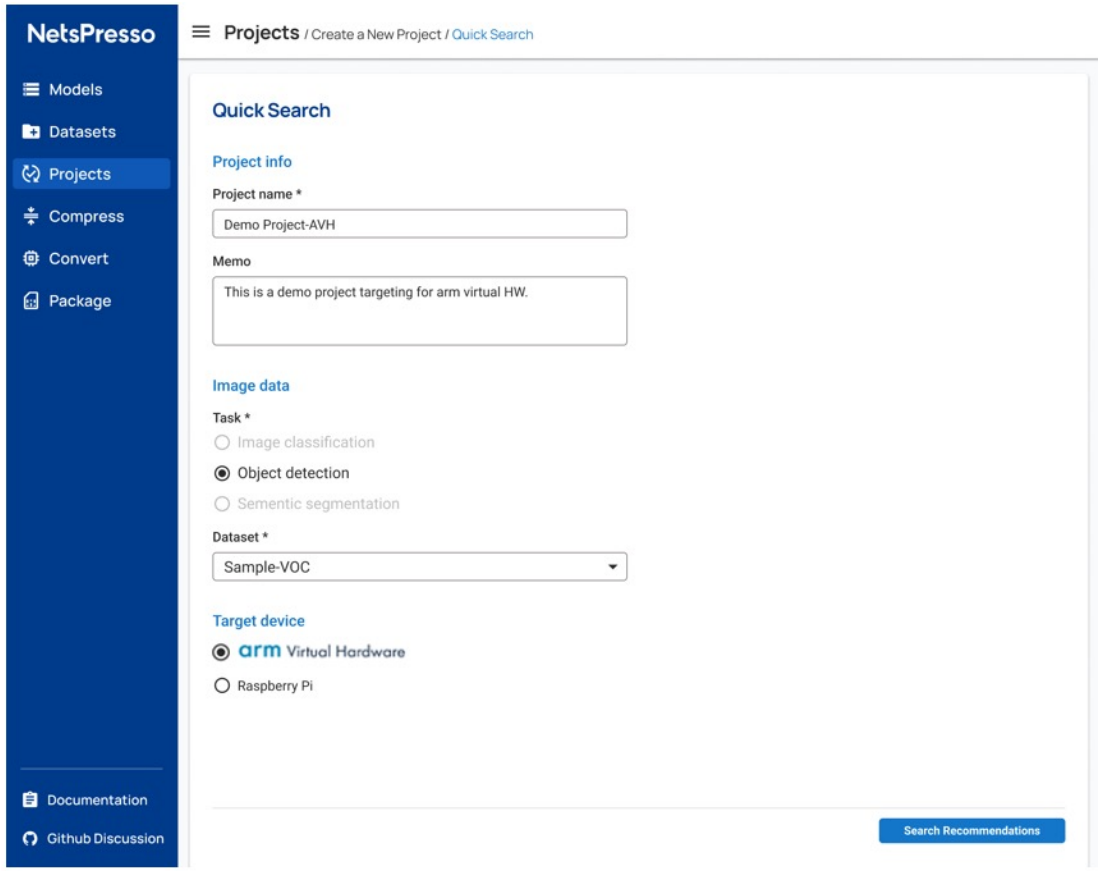
- Hardware-Aware AI Model Optimization

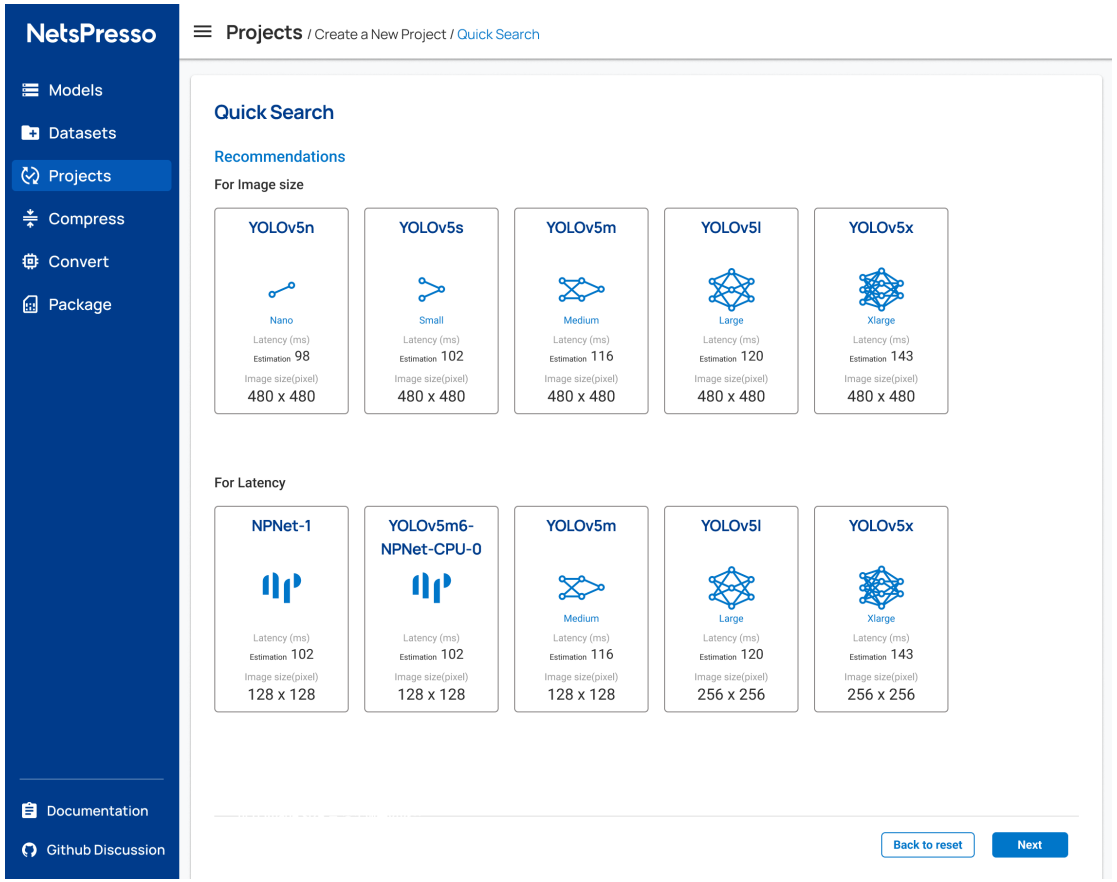


NetsPresso

- Arm Virtual Hardware in NetsPresso

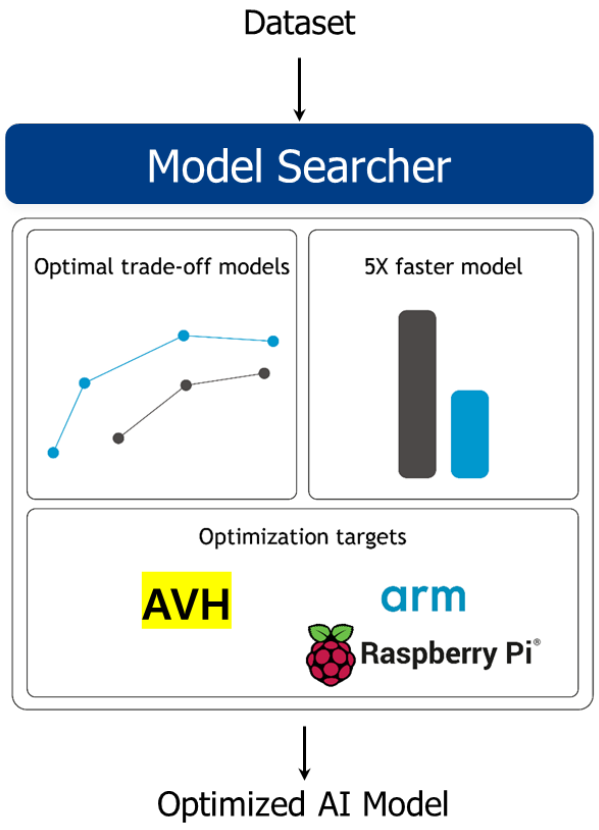
Snapshots of NetsPresso





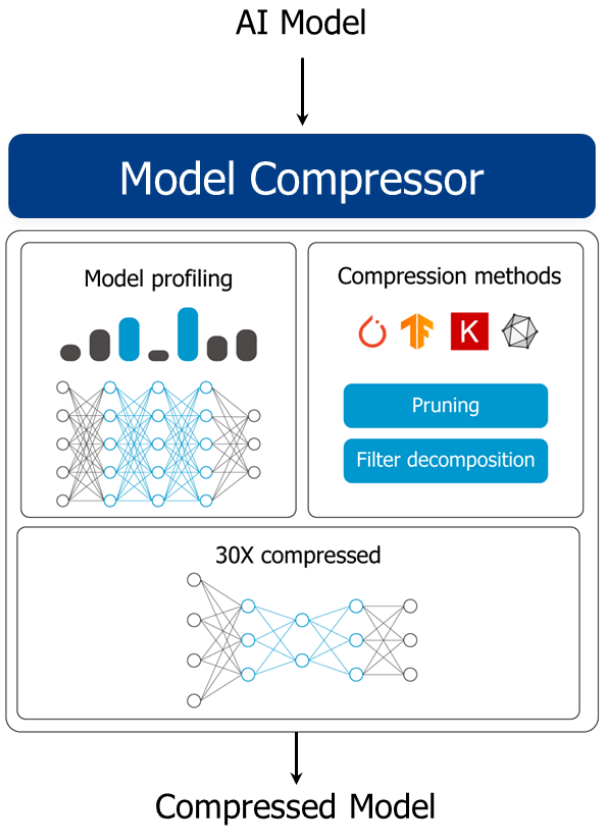
NetsPresso

- Model Searcher
  - Shorter AI model development time (months → weeks)
  - Better performance (latency, power consumption, etc.)
  - More options to choose from (performance/Hardware, etc.)
  - Near production-ready AI models (based on Hardware Validation)



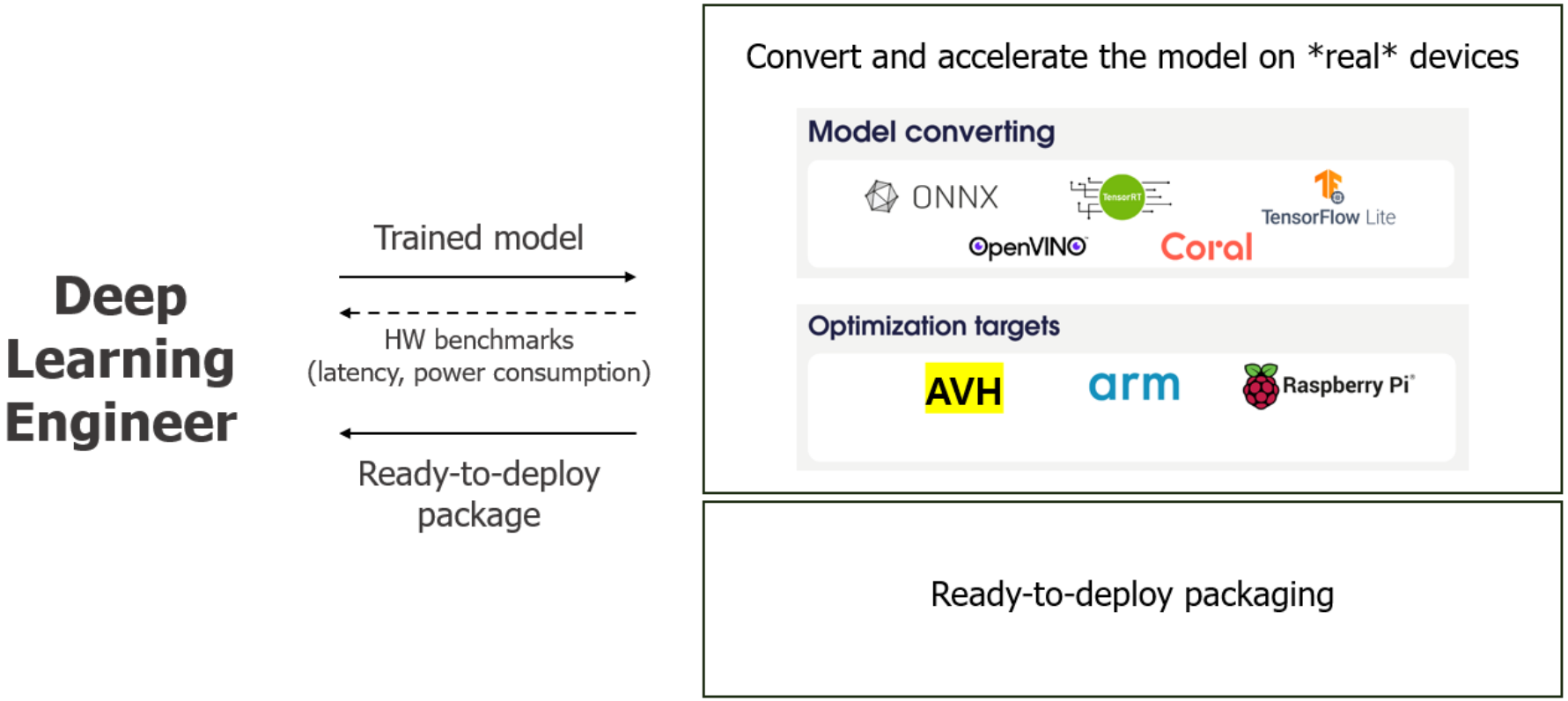
NetsPresso

- Model Compressor
  - Supports all CNN architectures
  - Optimal compression ratio is recommended
  - Eliminates months of paper implementation time
  - Minimal loss of information



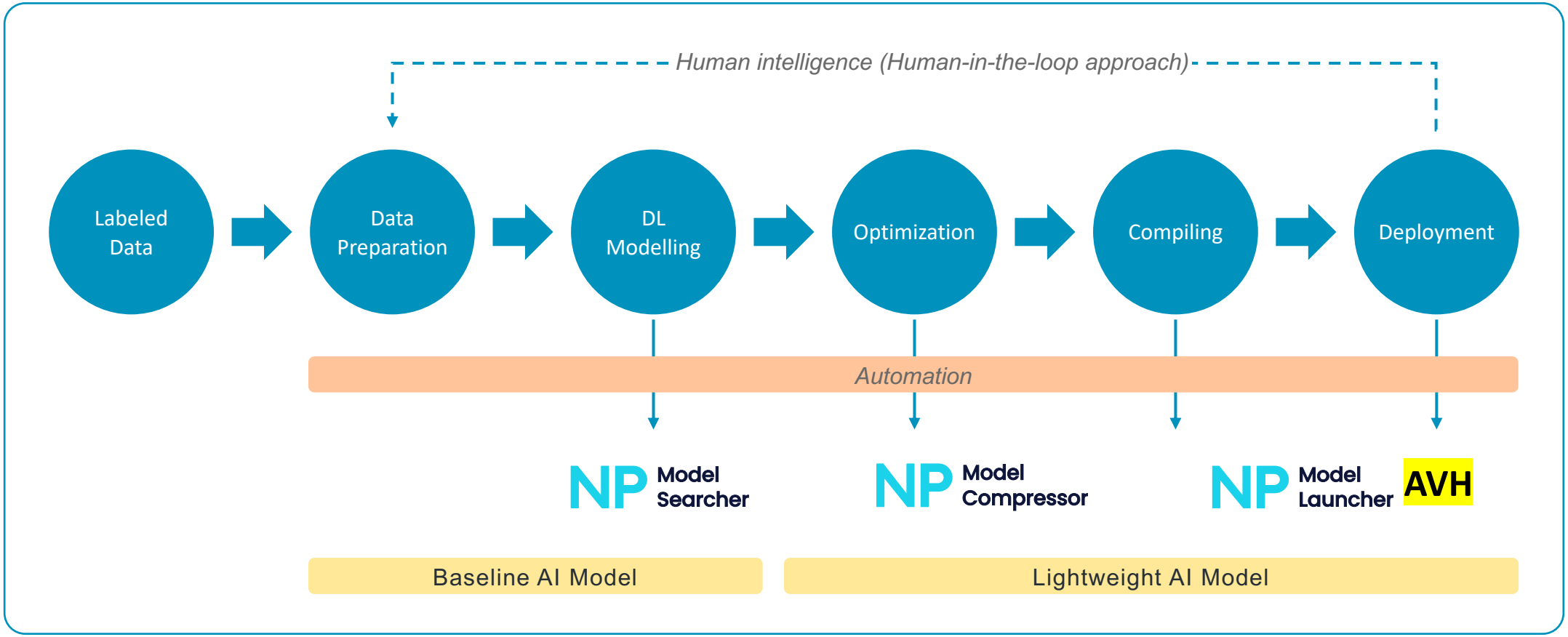
NetsPresso

- Model Launcher



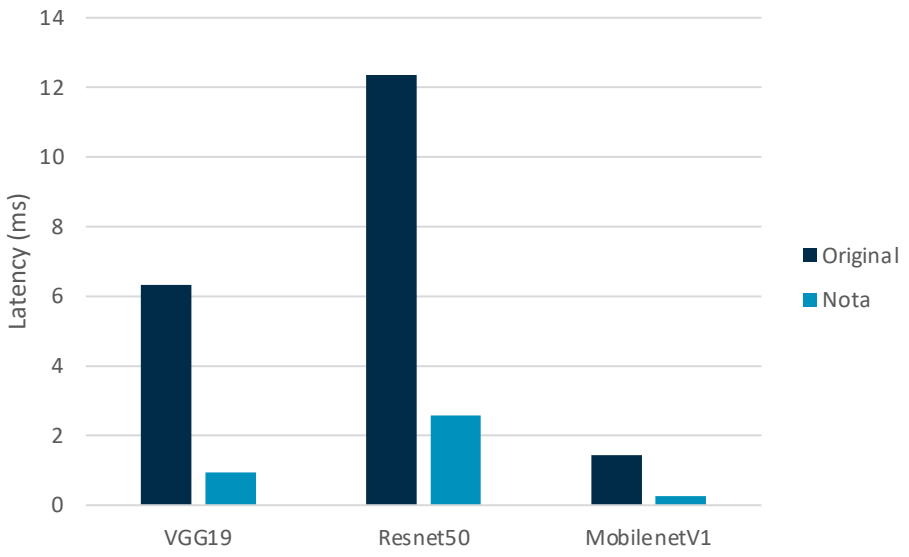
NetsPresso

- Pipeline
  - One-stop shop to build, optimize, and deploy optimized hardware-aware AI models with AVH



Results

- Image classification
  - On CIFAR-100 (image size = 32 × 32)
  - VGG19: up to 7x compression with ~1.5% drop in accuracy
  - MobileNetV1: up to 5x compression with marginally improved accuracy



Network	Type	Accuracy (%)	Macs (M)	Params (M)	Model size (MiB)	Latency (ms)
VGG19	Original	73.66	398.3	20.1	19.3	6.32
	Nota	72.12	61.5	0.7	1.4	0.68
ResNet50	Original	78.58	1,298.0	23.8	23.5	12.35
	Nota	76.84	132.1	2.4	2.9	2.58
MobileNet V1	Original	66.36	46.5	3.3	3.5	1.43
	Nota	66.64	9.7	0.4	0.5	0.27

Results

- Image classification
  - On Imagewoof (image size = 224 × 224)
  - Because the sum of model size and activation buffer size of the baseline VGG19 is larger than 32MiB, Ethos-U65 could not run the model.
  - Nota's NetsPresso reduced the size of VGG19 to ~50% so that Ethos-U65 can run the model.

Network	Type	Accuracy (%)	Macs (M)	Params (M)	Model size (MiB)	Latency (ms)
VGG19	Original	88.39	19527.4	32.9	32.0	X
	Nota	87.99	6914.7	18.4	18.0	57.41

Where to try

NetsPresso.ai

arm AI

AI Virtual Tech Talks Series

NotaAI

Thank You

Danke

Merci

谢谢

ありがとう

Gracias

Kiitos

감사합니다

धन्यवाद

شكرًا

תודה

arm AI

Thank you!

Tweet us: @ArmSoftwareDev -> #AIVTT

Check out our Arm Software Developers YouTube channel

Signup now for our next AI Virtual Tech Talk: www.arm.com/techtalks