Nota AI

PyNetsPresso and LaunchX

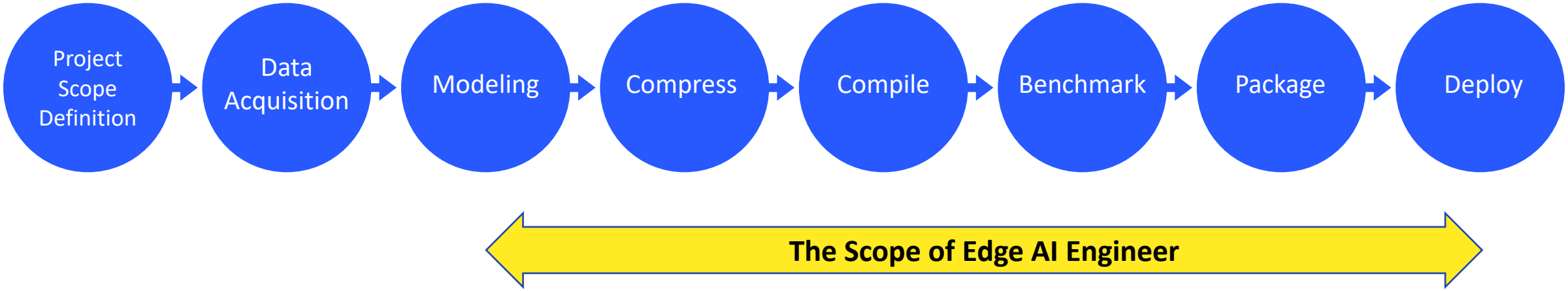
: An Integrated Toolchain for HW-Aware AI Model Optimization and Benchmarking

Jong Won Baek | NetsPresso Product Owner

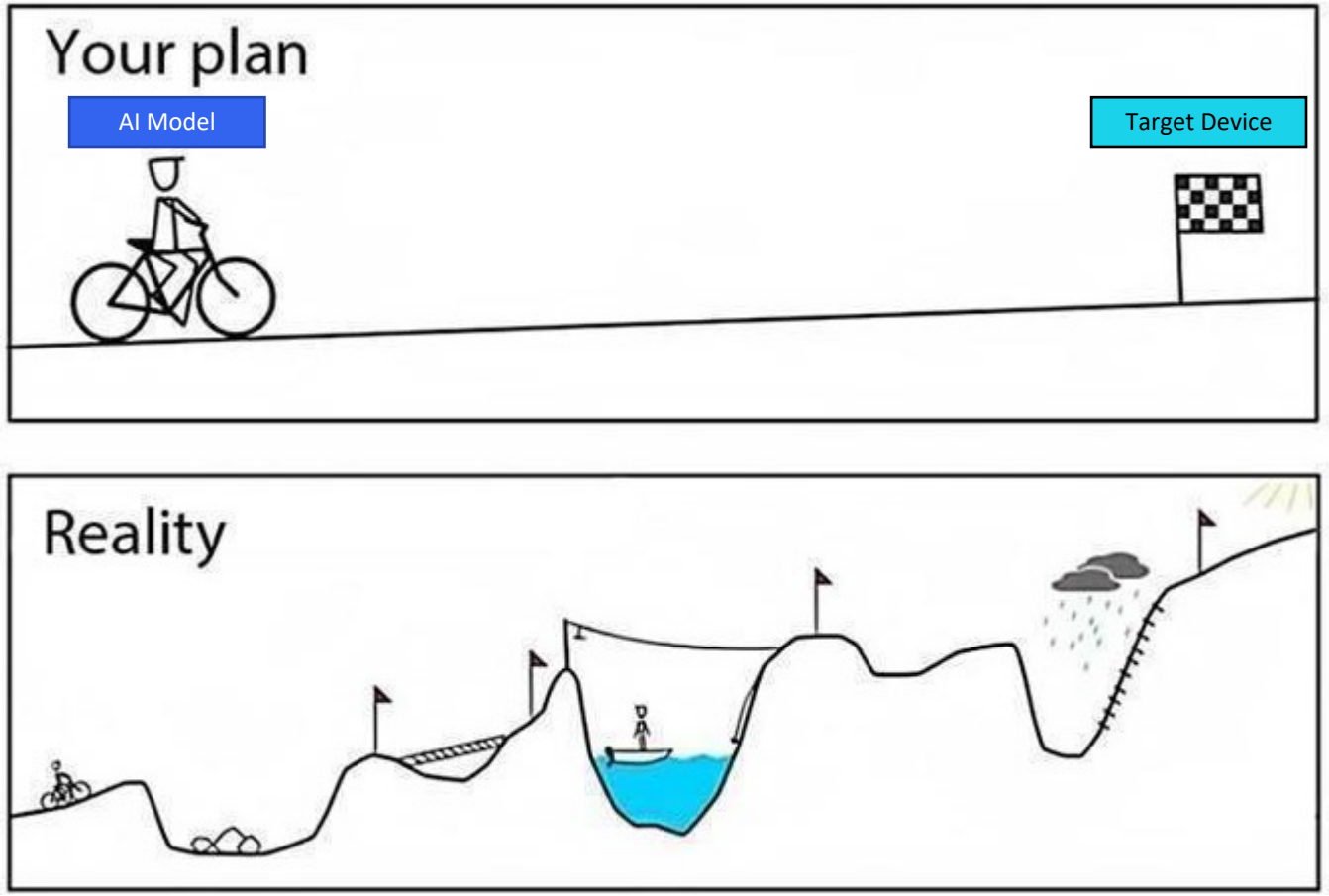
Table of Contents

Introduction: The Journey of Edge AI Engineer

It starts from AI model, to deployment

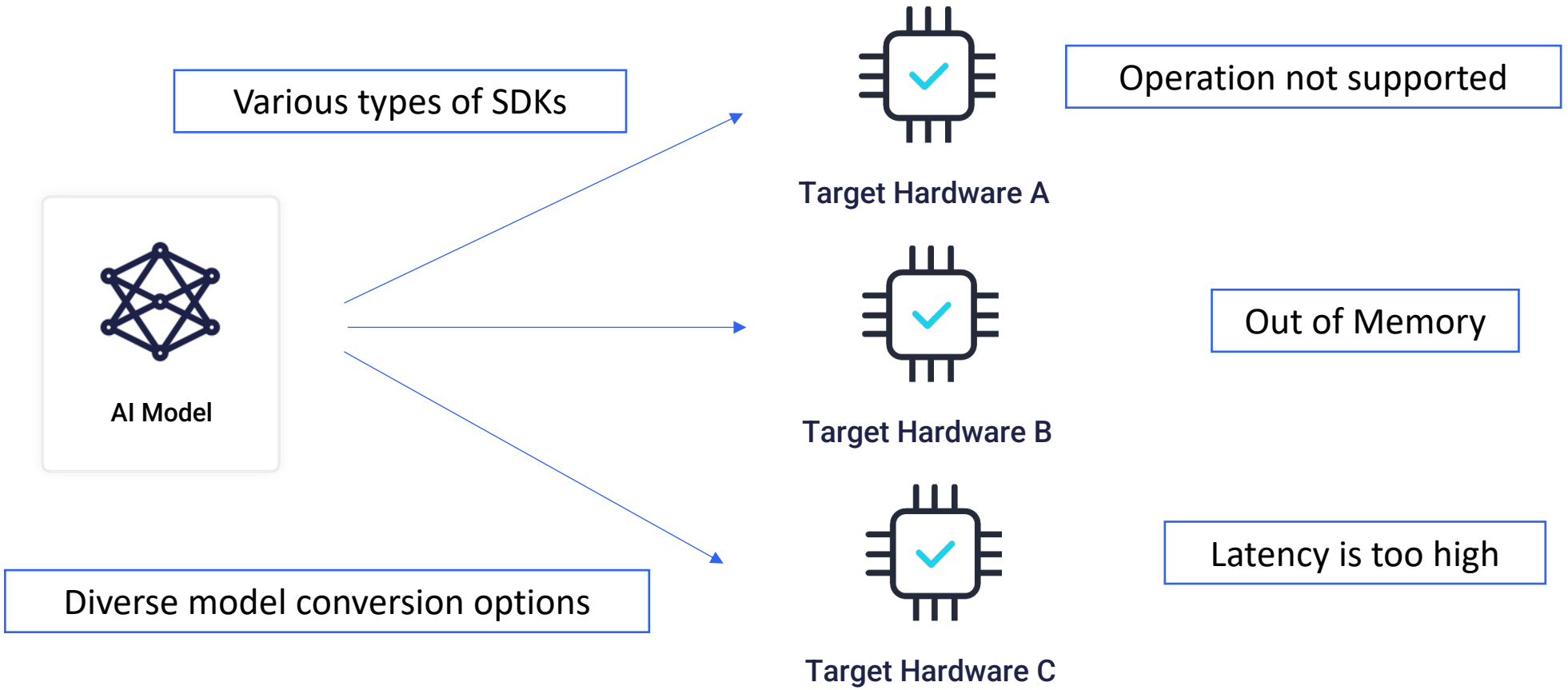


Things Don’t Always Go As Planned

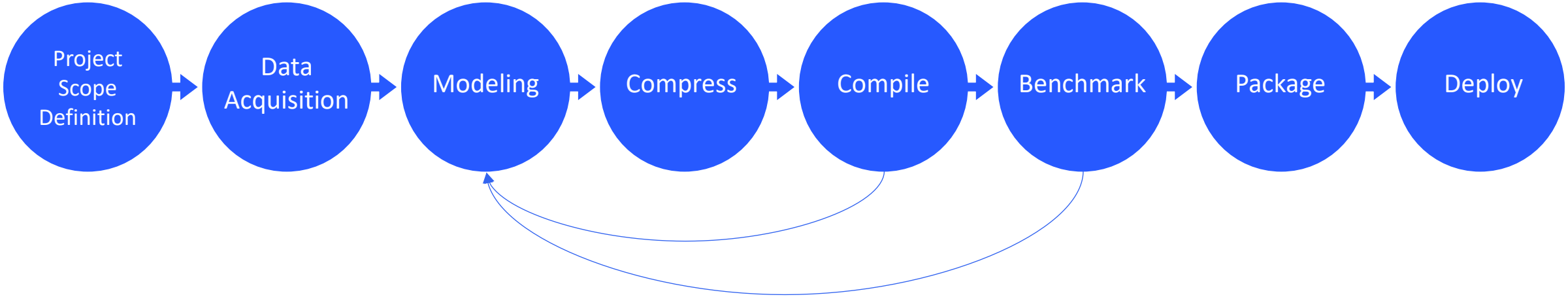


Source: lesswrong.com

Problems: What's happening on the journey?



What’s happening on the journey?



If the model does not work, you get back to square one

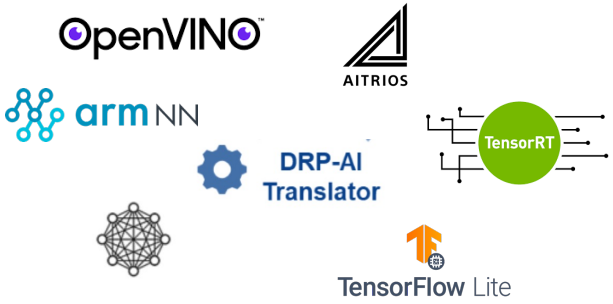
You need to go through the whole thing again to make your model work on the device.

Or, choose better solution

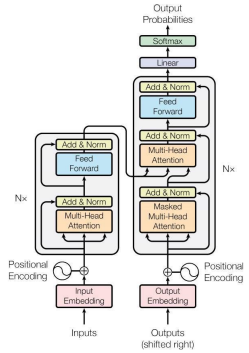
Solution: NetsPresso

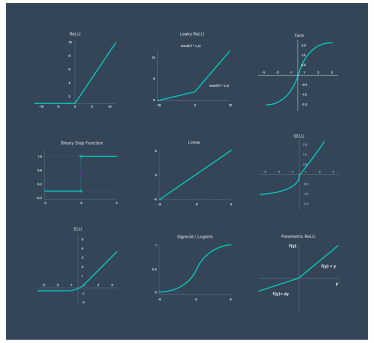
The easiest tool for Edge AI engineers

Challenges

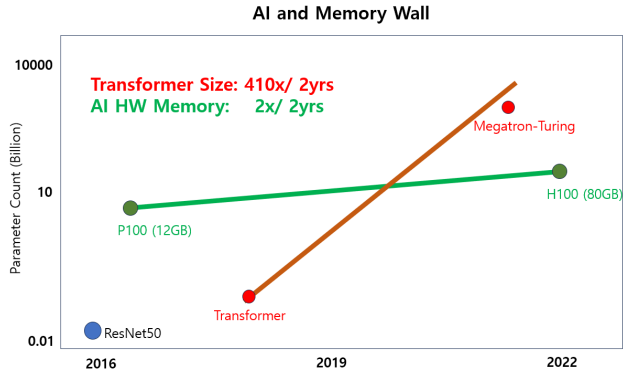


Diverse Edge AI SDK





Unsupported layers and operations



Limited memory size
Latency is too high

Solution

NetsPresso®

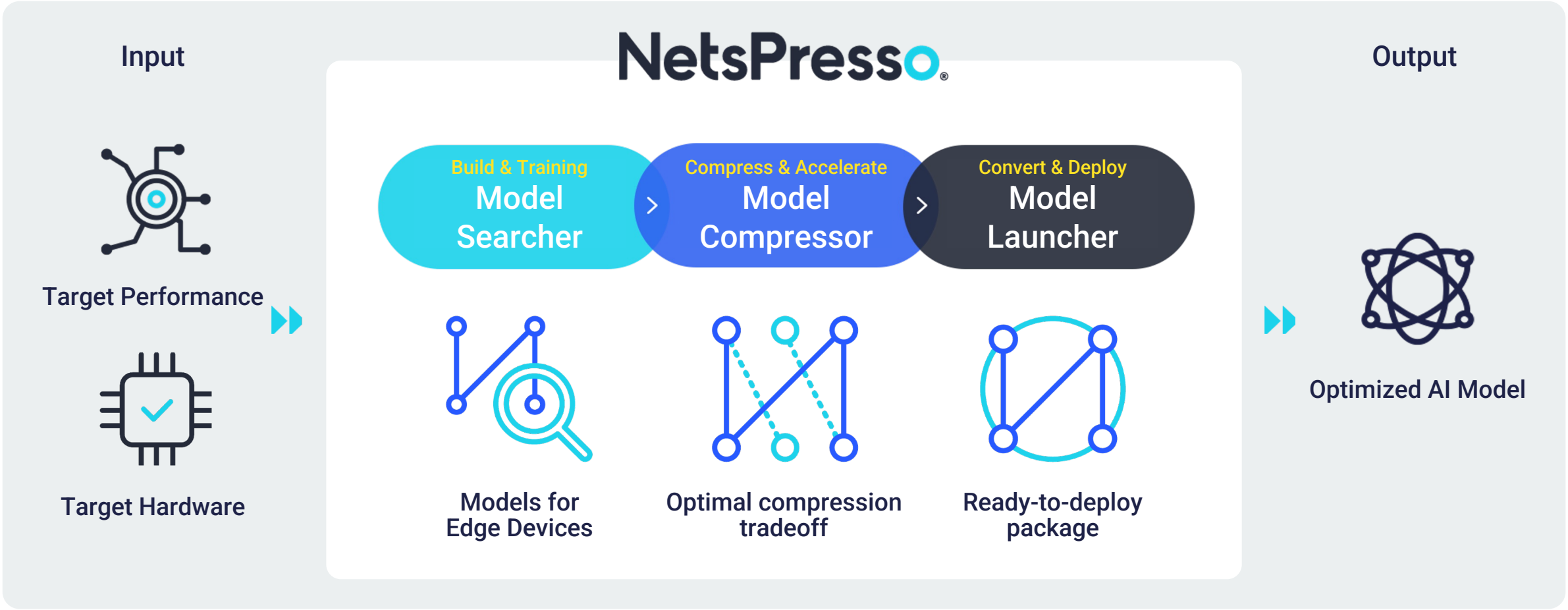
Unified Interface

Operator Conversion

AI Model Compression

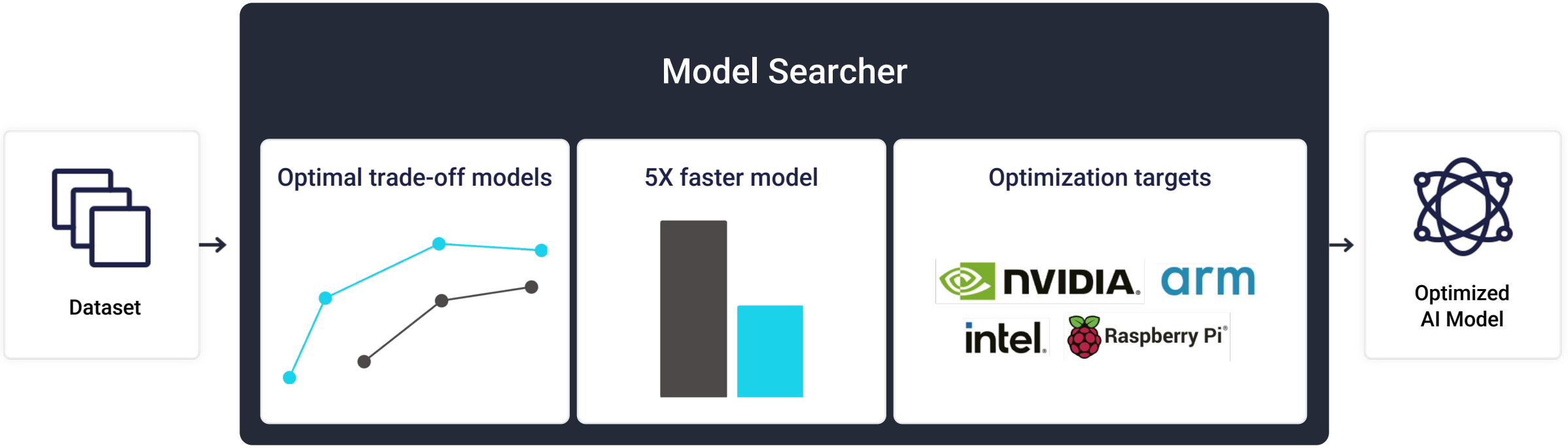
Platform Overview: AI Model Optimization

NetsPresso® simplifies AI model optimization for target devices with automated processes.



Model Searcher Module: Automated Model Search for Your Device

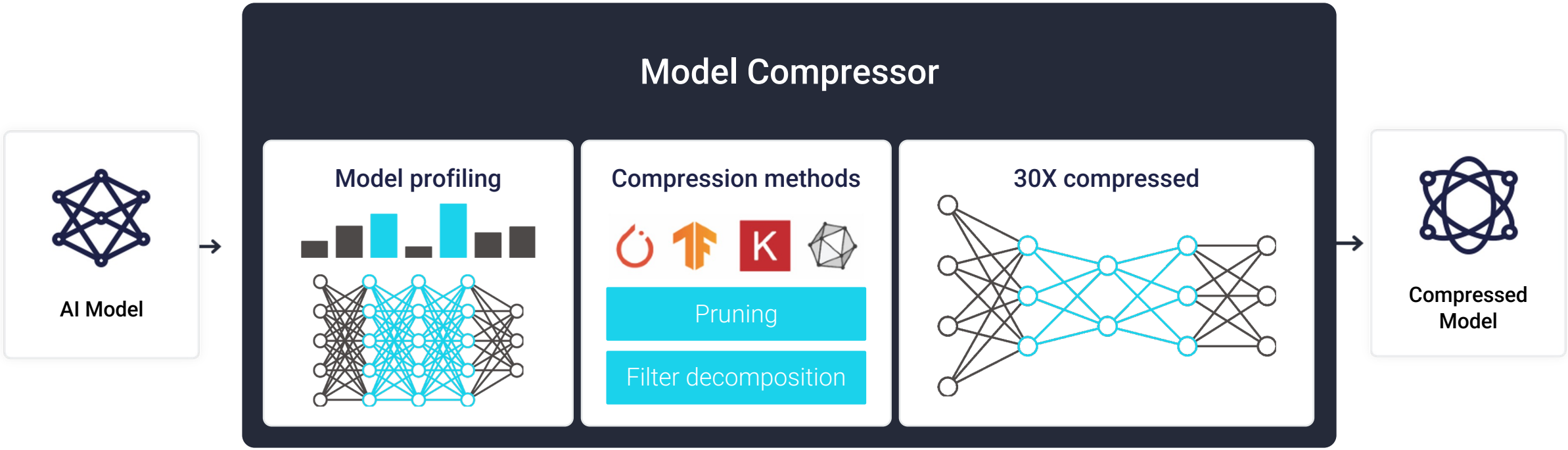
The Model Searcher module provides models optimized for your target device.



- Optimized model design for target hardware
- Multiple models with various options
- Produce models close to production level based on actual hardware testing
- Create models with lower latency

Model Compressor Module: Balancing Performance and Efficiency

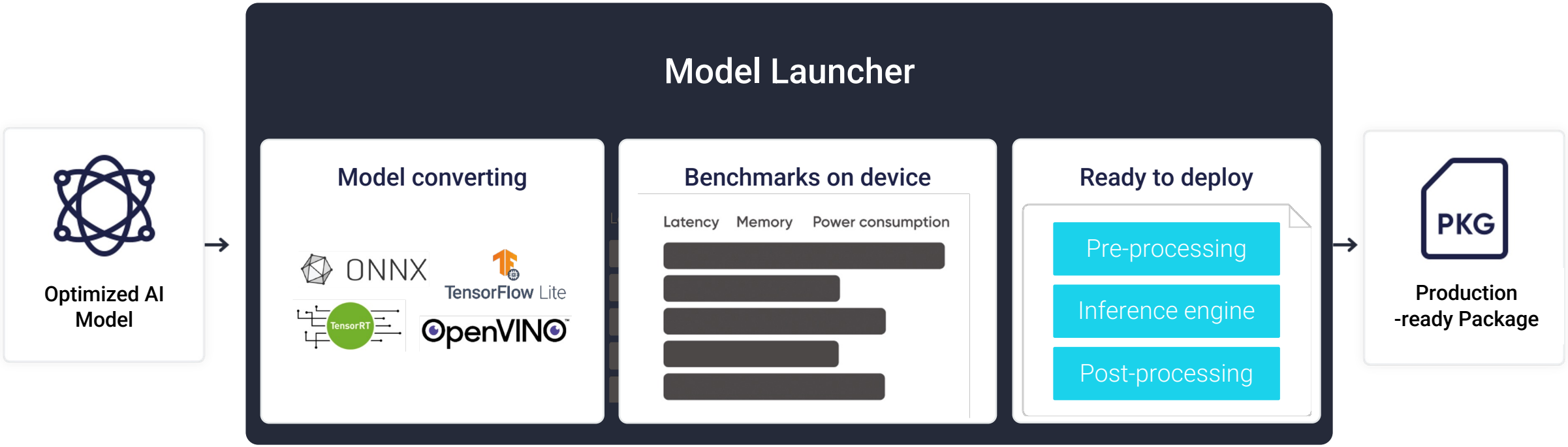
Achieve optimal performance-efficiency tradeoffs with the Model Compressor module.



- Supports all CNN architectures (limitedly transformer models)
- Recommends optimal compression ratios
- Eliminates months of paper implementation period
- Minimal loss of information

Model Launcher Module: Swift Launch for Accelerated Models

Accelerate model deployment with the Model Launcher module.



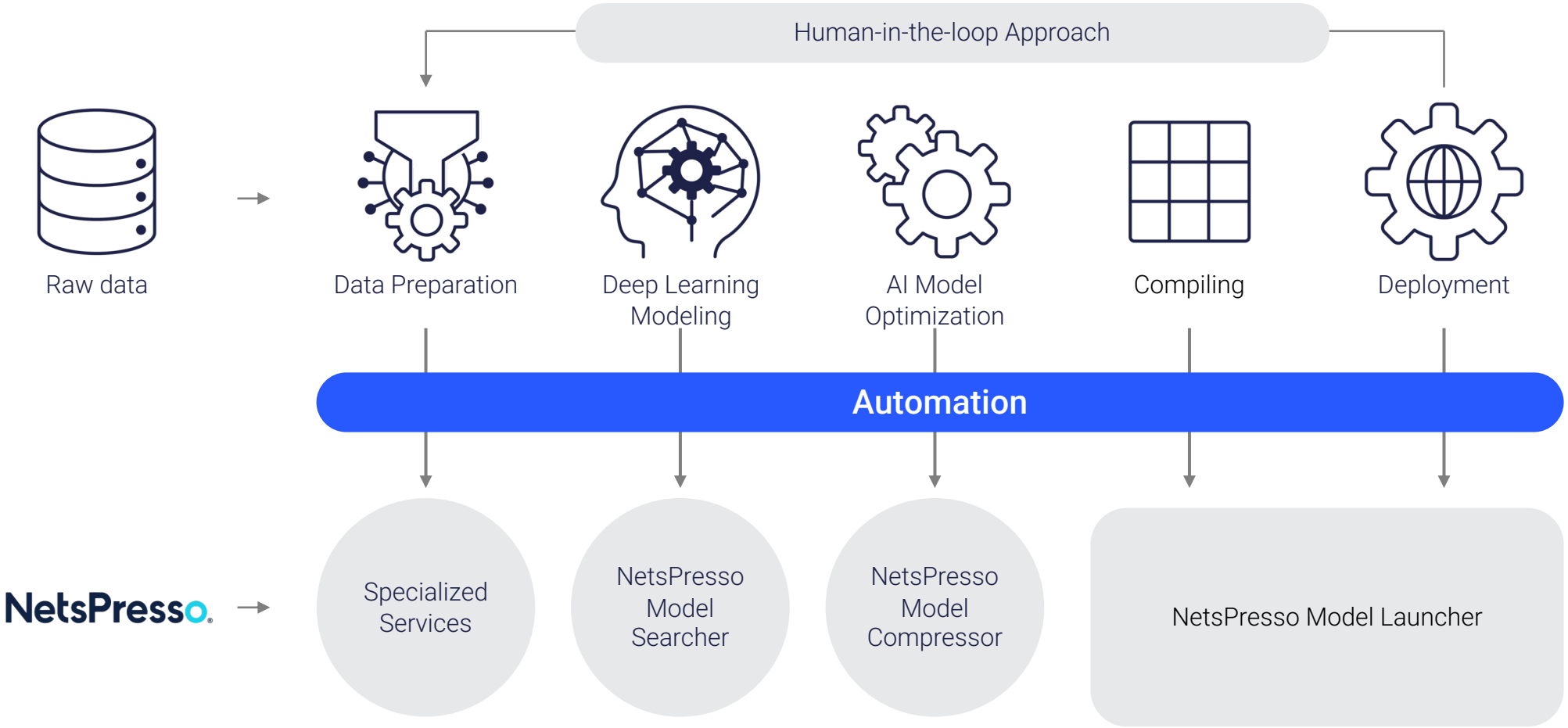
- Unified interface for various SDKs
- Performance benchmarks and recommendations on actual devices
- Conversion to operations supported by target hardware
- Production-ready package

Challenges vs NetsPresso

Challenges	Build & Search Model Searcher	Compress & Accelerate Model Compressor	Convert & Deploy Model Launcher
Various types of SDKs			✓
Diverse model conversion options			✓
Out of Memory	✓	✓	✓
Latency is too high	✓	✓	✓
Operation not supported	✓		✓

Platform Pipeline: Streamlined AI Model Development

NetsPresso facilitates seamless transitions from optimization to deployment.



Success Cases

With Renesas

- Target: RA6M3 (ARM Cortex-M4)
- Model: YOLOv2
- Replace unsupported operation to be supported and HW-aware compression.
- Offering an extensive array of options to cater to diverse customer preferences.

Without Nota

Cannot operate DET model

Enabling →

With Nota

Can operate with higher accuracy

With Sony

- Target: IMX500
- Model: YOLO based model
- IMX500 is commonly used for smart city use cases.
- While preserving latency, Nota maximizes the capacity of the model so we could train higher performing model.

Without Nota

DET Model with 31.4% mAP

10% More Accurate →

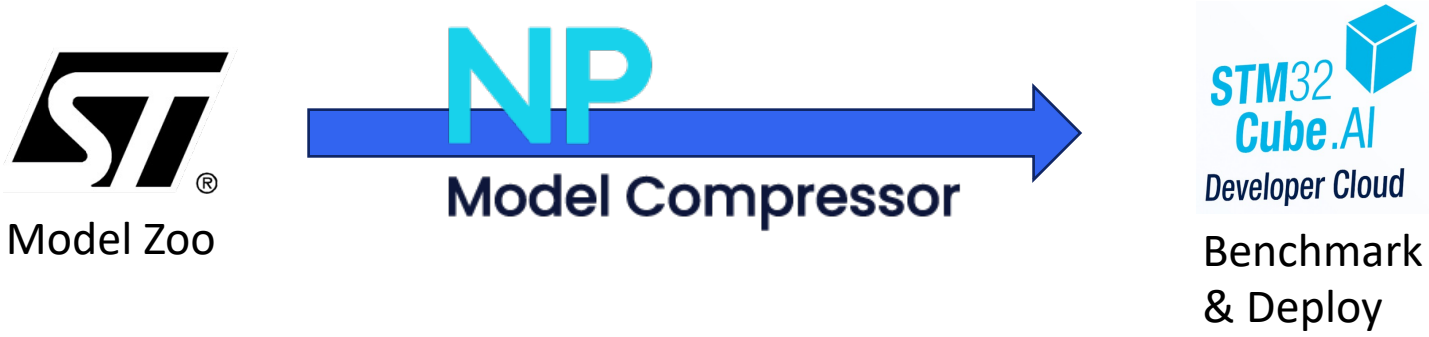
With Nota

DET Model with 41.1% mAP

Success Cases

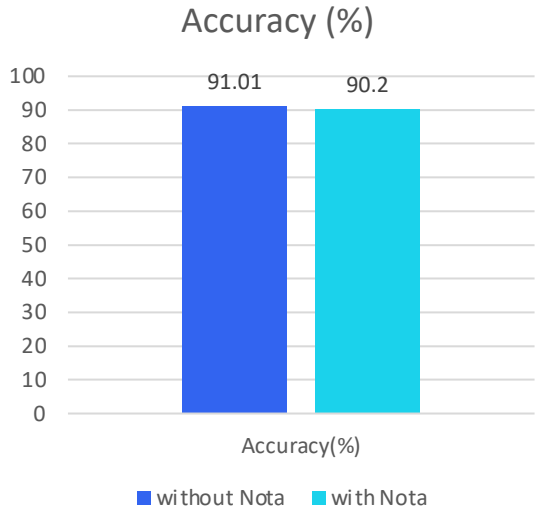
With STMicroelectronics

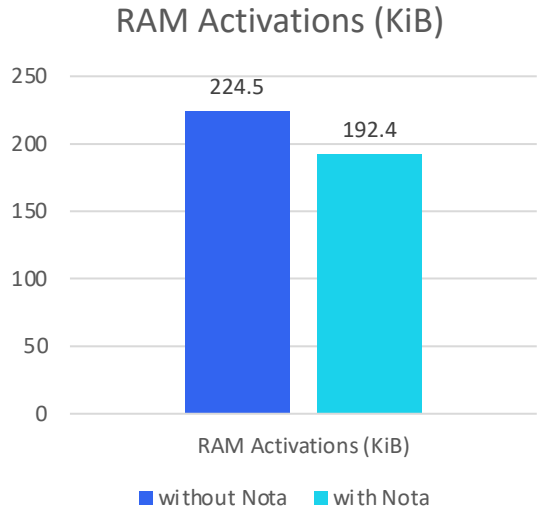
- Target: STM32H747I-DISCO
- Model: Mobilenet v2 (Tensorflow Flower Dataset)
- Compressed a model from the STMicroelectronics model zoo using NetsPresso Model Compressor and deployed it to a device using STM32 Cube AI
- Significant benefits in terms of memory and latency optimization with minimal performance loss
- This experiment was conducted by engineers of STMicroelectronics (sample code will be released)

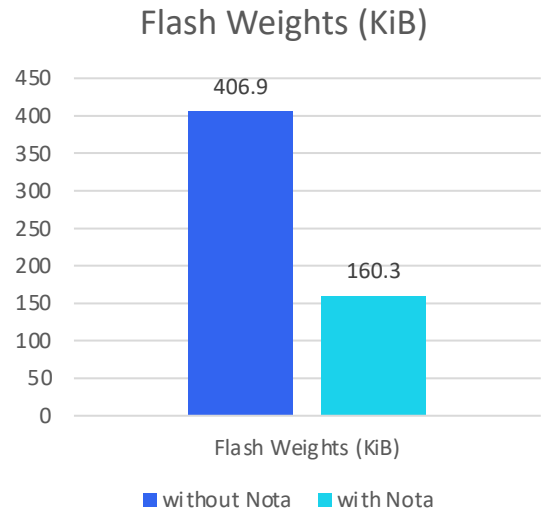


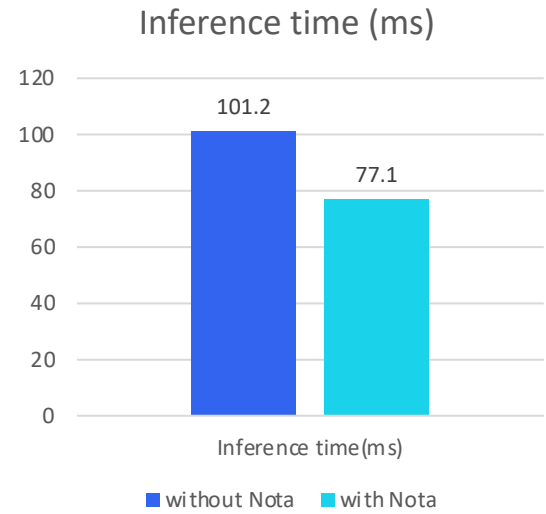
Success Cases

With STMicroelectronics









0.8%p drop

14% lighter

61% lighter

24% faster

Basic pruning options were used
(Global pruning ratio 0.5, using recommended layer-wise pruning ratio)

Ecosystem for Edge AI Development

With integration to model/device ecosystem, NetsPresso can bring benefit to broader area.

STM32 Cube.AI Developer Cloud

arm ARM Virtual Hardware

NVIDIA TAO

Nota's Device Farm

Nvidia, ARM, Renesas, Intel, Arduino, …

…

Demo: PyNetsPresso and LaunchX

PyNetsPresso

PyNetsPresso

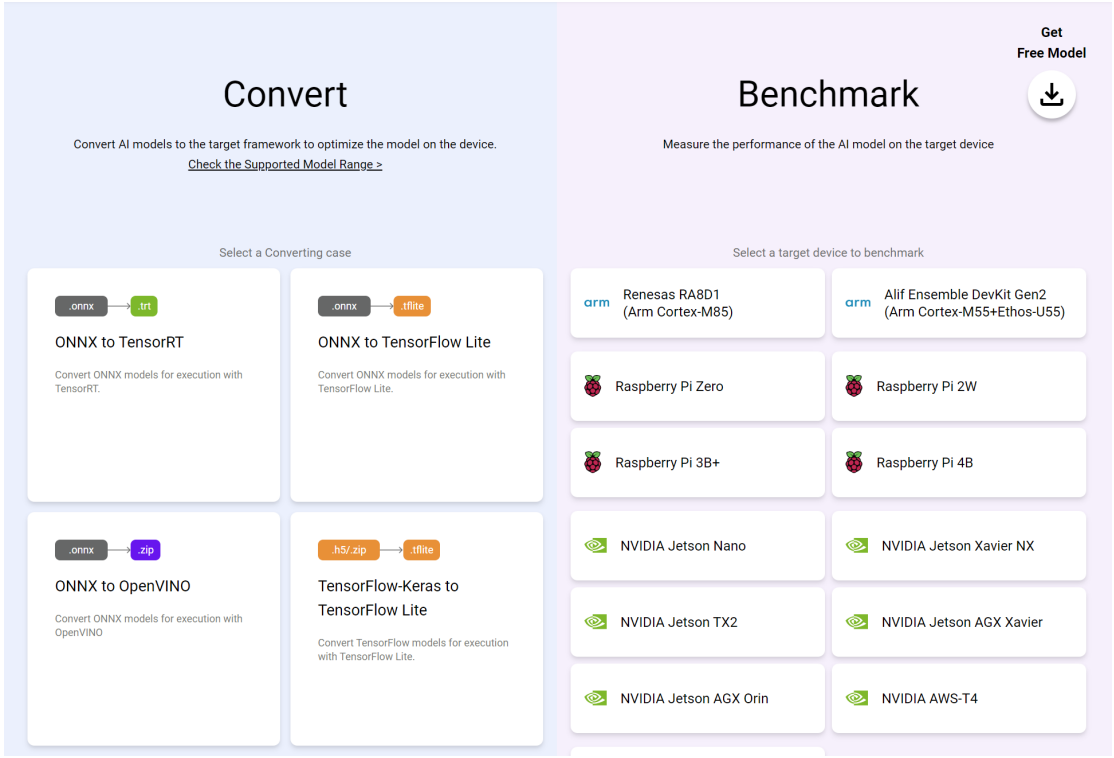
Streamline your AI model **optimization** workflow

YOLOX | YOLOv8 | YOLOv7 | YOLOv5 | PIDNet | PyTorch-CIFAR-Models

python 3.8 | 3.9 | 3.10 TensorFlow 2.3.x ~ 2.8.x. PyTorch 1.11.x ~ 1.13.x.

NetsPresso Open in Website ModelZoo Open in Github Best Practice Open in Colab

LaunchX



* You can see the real demo at Nota AI booth

PyNetsPresso

Login

```
import getpass
from netspresso.client import SessionClient
from netspresso.compressor import ModelCompressor

email ='xxx.yyy@st.com'
print('Enter you password')
password = getpass.getpass()

session = SessionClient(email=email, password=password)
compressor = ModelCompressor(user_session=session)
```

```
Enter you password
2023-10-27 09:42:35.715 | INFO     | netspresso.client:__login:50 - Login successfully
2023-10-27 09:42:37.787 | INFO     | netspresso.client:__get_user_info:67 - successfully got user information
```

PyNetsPresso

Uploading model

```
from netspresso.compressor import Task, Framework

model = compressor.upload_model(
    model_name='model',
    task=Task.IMAGE_CLASSIFICATION,
    framework=Framework.TENSORFLOW_KERAS,
    file_path= 'experiments_outputs/training/saved_models/best_model.h5',
    input_shapes=[{'batch': 1, 'channel': 3, 'dimension': [128, 128]}])
```

```
2023-10-27 09:52:50.120 | INFO     | netspresso.compressor:upload_model:72 - Uploading Model...
2023-10-27 09:53:06.984 | INFO     | netspresso.compressor:upload_model:83 - Upload model successfully. Model ID: 94d350dc-0297-4964-aa4b-d033b93ffc9f
```

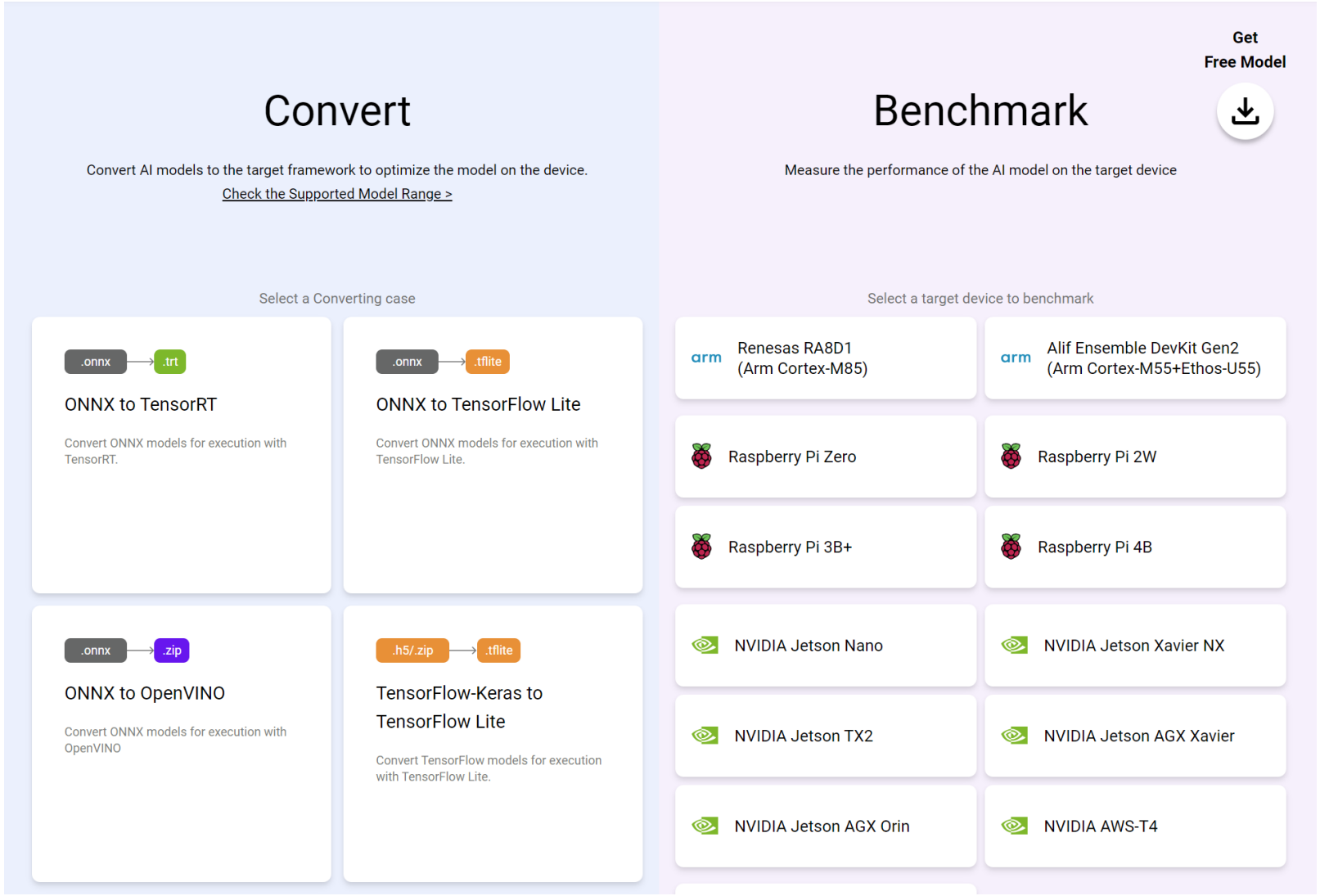
PyNetsPresso

Compression

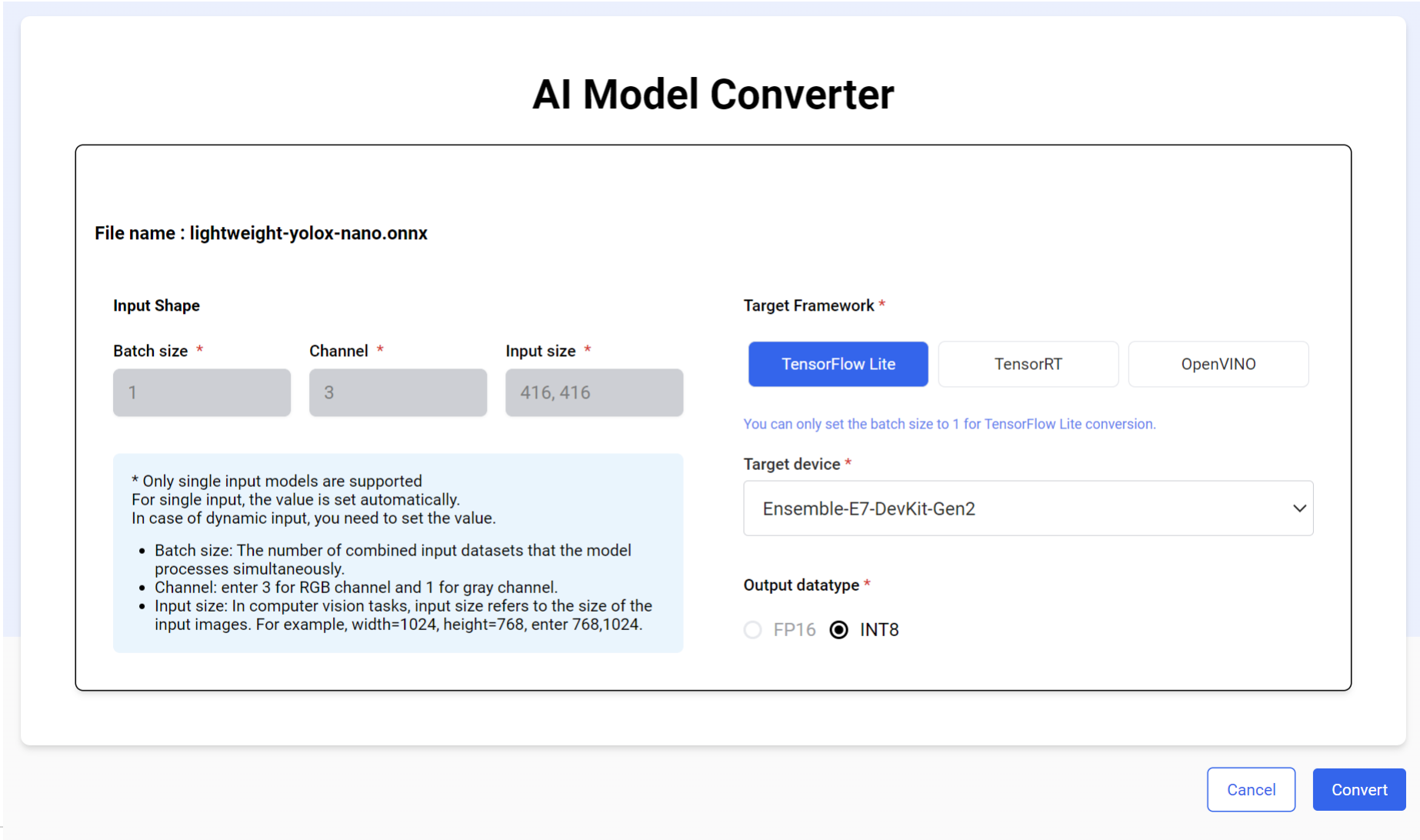
```python
import os

from netspresso.compressor import CompressionMethod
from netspresso.compressor import RecommendationMethod

if not os.path.exists('experiments_outputs/compressed_models'):
    os.makedirs('experiments_outputs/compressed_models')

compressed_model = compressor.recommendation_compression(
    model_id='94d350dc-0297-4964-aa4b-d033b93ffc9f',
    model_name='compressed_model.h5',
    compression_method=CompressionMethod.PR_L2,
    recommendation_method=RecommendationMethod.SLAMP,
    recommendation_ratio=0.5,
    output_path='experiments_outputs/compressed_models/compressed_model.h5',)
```

```
2023-10-27 09:53:59.084 | INFO     | netspresso.compressor:recommendation_compression:448 - Compressing recommendation-based model...
2023-10-27 09:54:19.376 | INFO     | netspresso.compressor:download_model:223 - Downloading model...
2023-10-27 09:54:31.058 | INFO     | netspresso.compressor:download_model:226 - Download model successfully. Local Path: experiments_outputs/compressed_m
odels/compressed_model.h5
2023-10-27 09:54:31.063 | INFO     | netspresso.compressor:get_model:196 - Getting model...
2023-10-27 09:54:33.535 | INFO     | netspresso.compressor:get_model:202 - Get model successfully.
2023-10-27 09:54:33.538 | INFO     | netspresso.compressor:recommendation_compression:505 - Recommendation compression successfully. Compressed Model ID:
49ac11c4-2c56-4575-b215-b7f184ca4693
2023-10-27 09:54:33.542 | INFO     | netspresso.compressor:recommendation_compression:506 - 50 credits have been consumed.
```

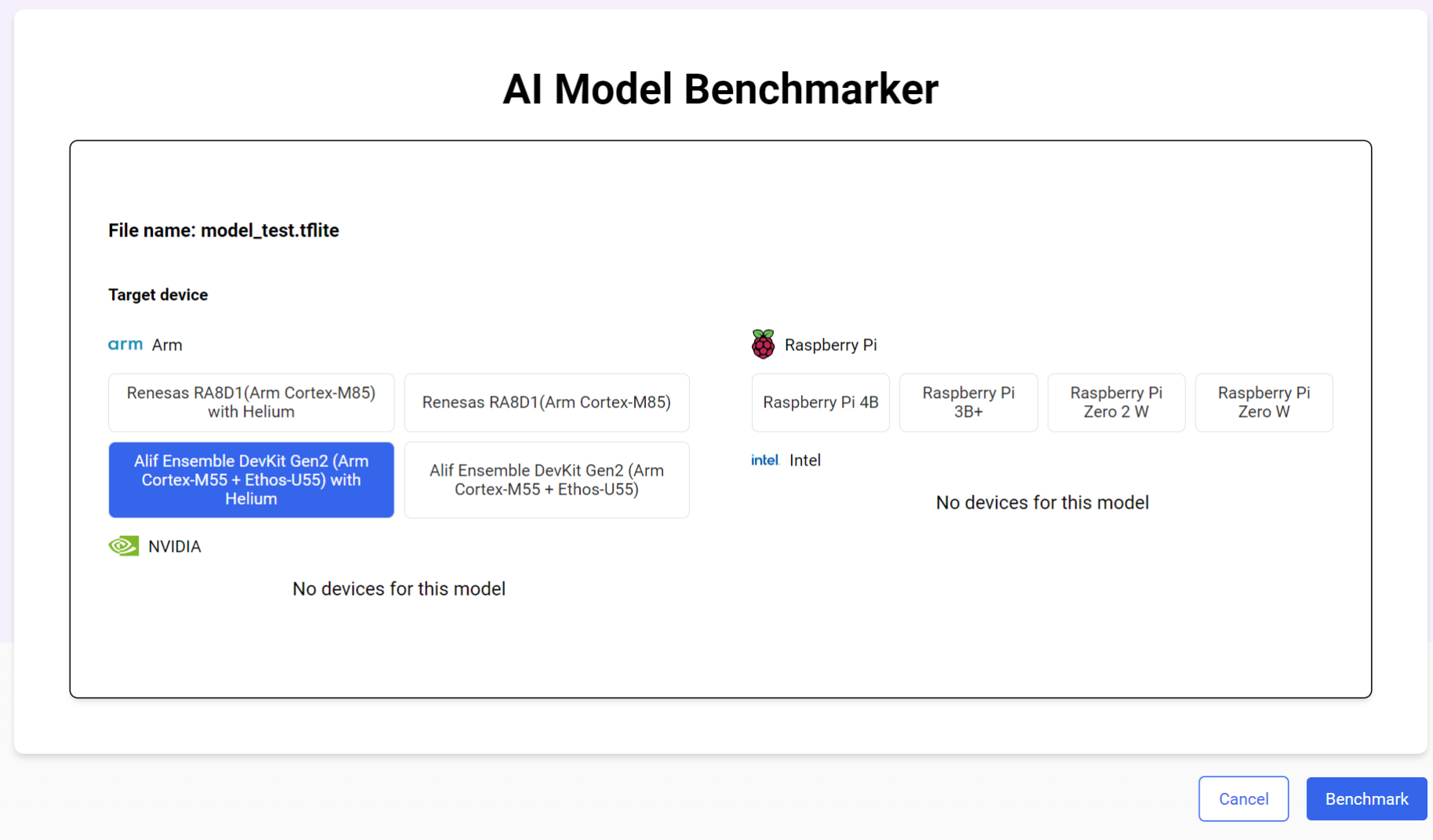
LaunchX



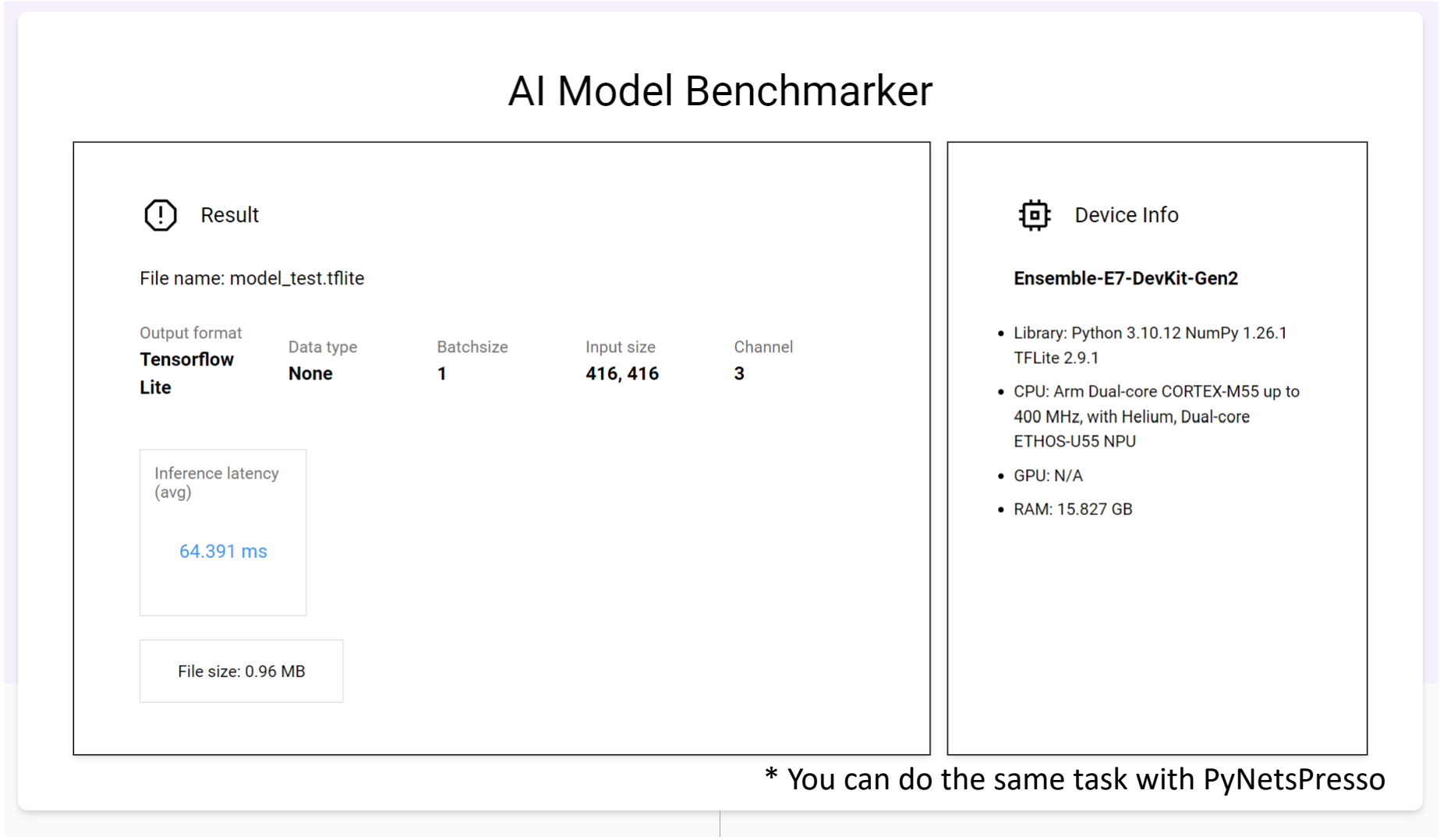
LaunchX



LaunchX

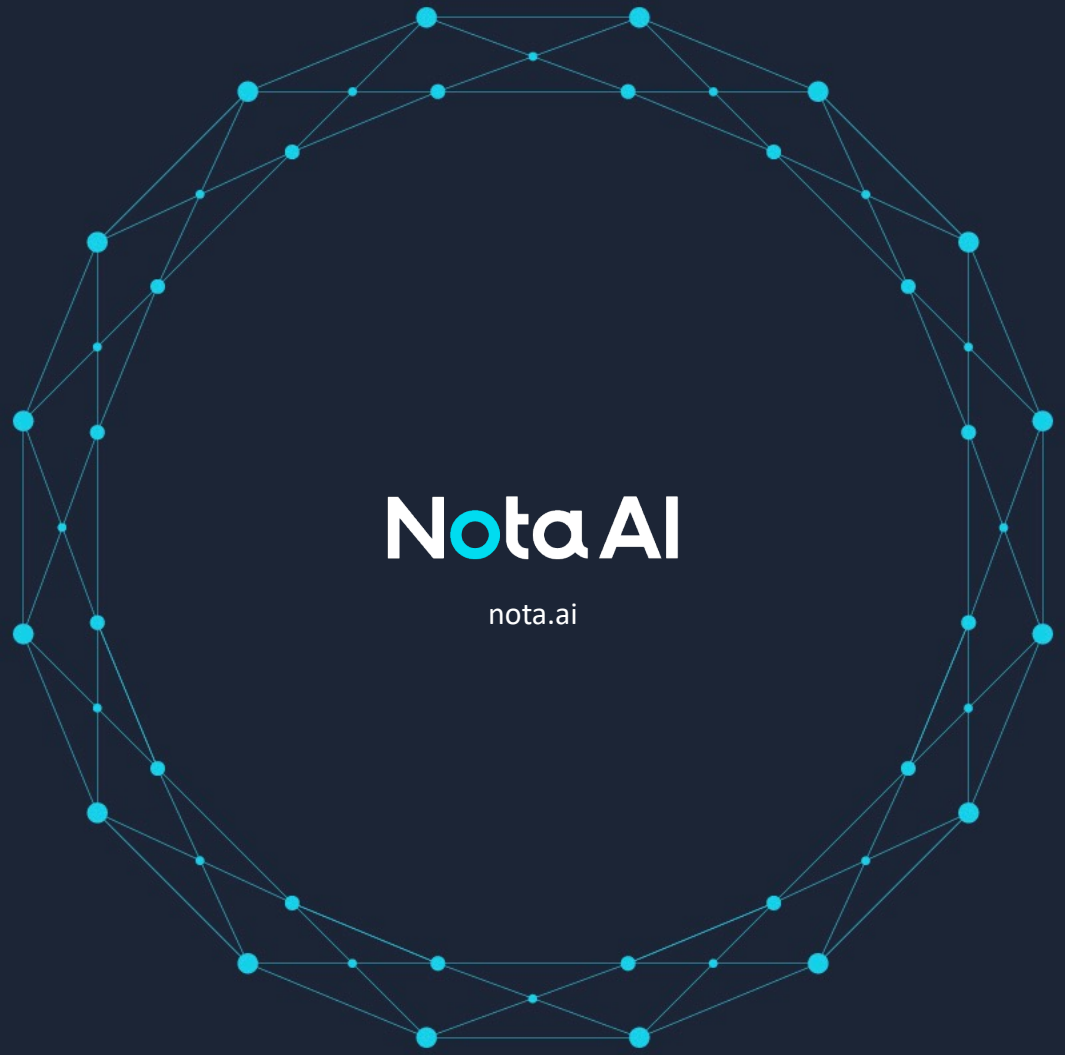


LaunchX



* You can do the same task with PyNetsPresso

Let’s do happy edge AI engineering with NetsPresso!



nota.ai

TINY ML ASIA
Seoul, South Korea
November 16, 2023

Copyright Notice

www.tinyml.org