deci

How to Successfully Deploy Deep Learning on Edge Devices

December 13th, 2022

Yonatan Geifman, PhD
Co-founder & CEO, Deci

About the speaker

Yonatan Geifman
Co-Founder & CEO, Deci

- CEO and Co-Founder of Deci
- PhD in Computer Science from the Technion-Israel Institute of Technology
- Former member of Google AI's MorphNet team
- Research focuses on making Deep Neural Networks (DNNs) more applicable for mission-critical tasks

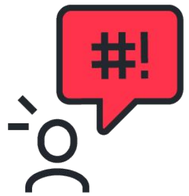
deci

Agenda

- The AI efficiency gap - implications for edge deployments
- The importance of production aware development
- Key considerations when developing DL for edge deployments
- How to quickly deploy on edge devices with Deci

deci

Common barriers to deployment on edge devices

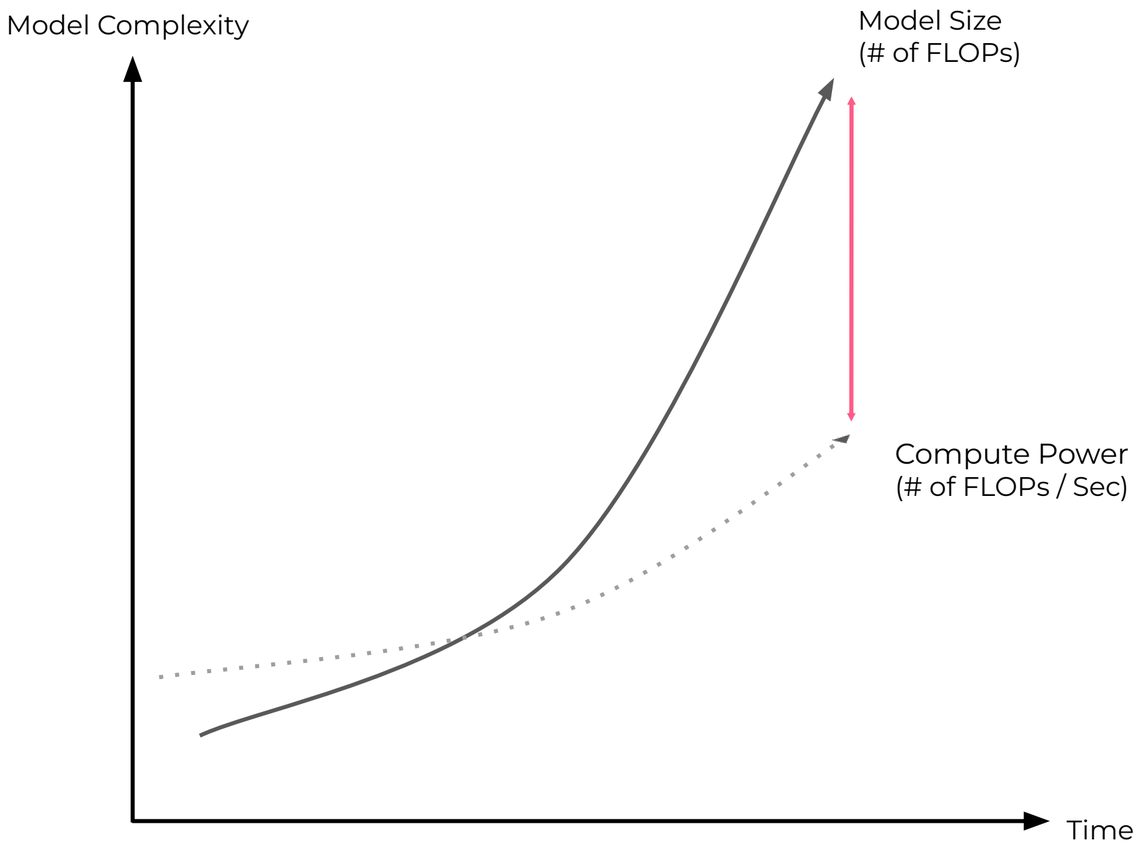


Inability to deploy on edge devices

Unsatisfactory Accuracy or performance

Long development cycle and high dev cost

Models’ power hunger is increasingly rapidly



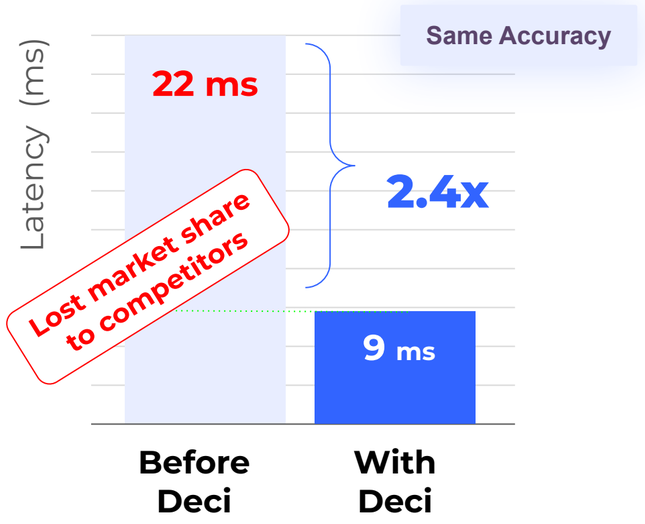
DL Efficiency Gap leads to:

- ✖ *Insufficient accuracy*
- ✖ *High latency*
- ✖ *Low throughput*
- ✖ *Large model size*
- ✖ *Large memory footprint*

deci

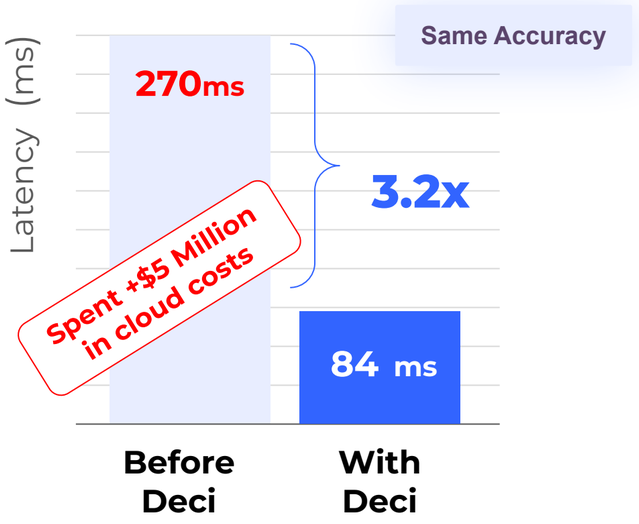
Leading AI teams are facing these development barriers

Insufficient latency resulting in <u>poor UX</u> on client laptop

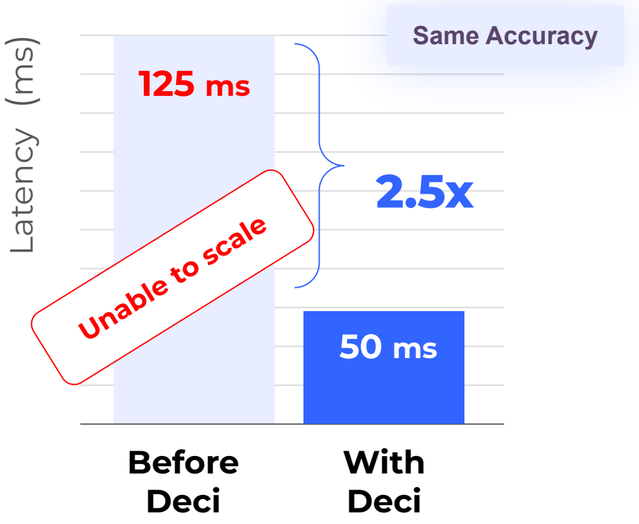


Measured on Macbook pro with Intel i5-8257U

<u>Inability to deploy on edge devices</u> resulting in high cloud cost and poor UX



<u>Inability to run in real-time</u> OD on a Jetson connected to multiple streams

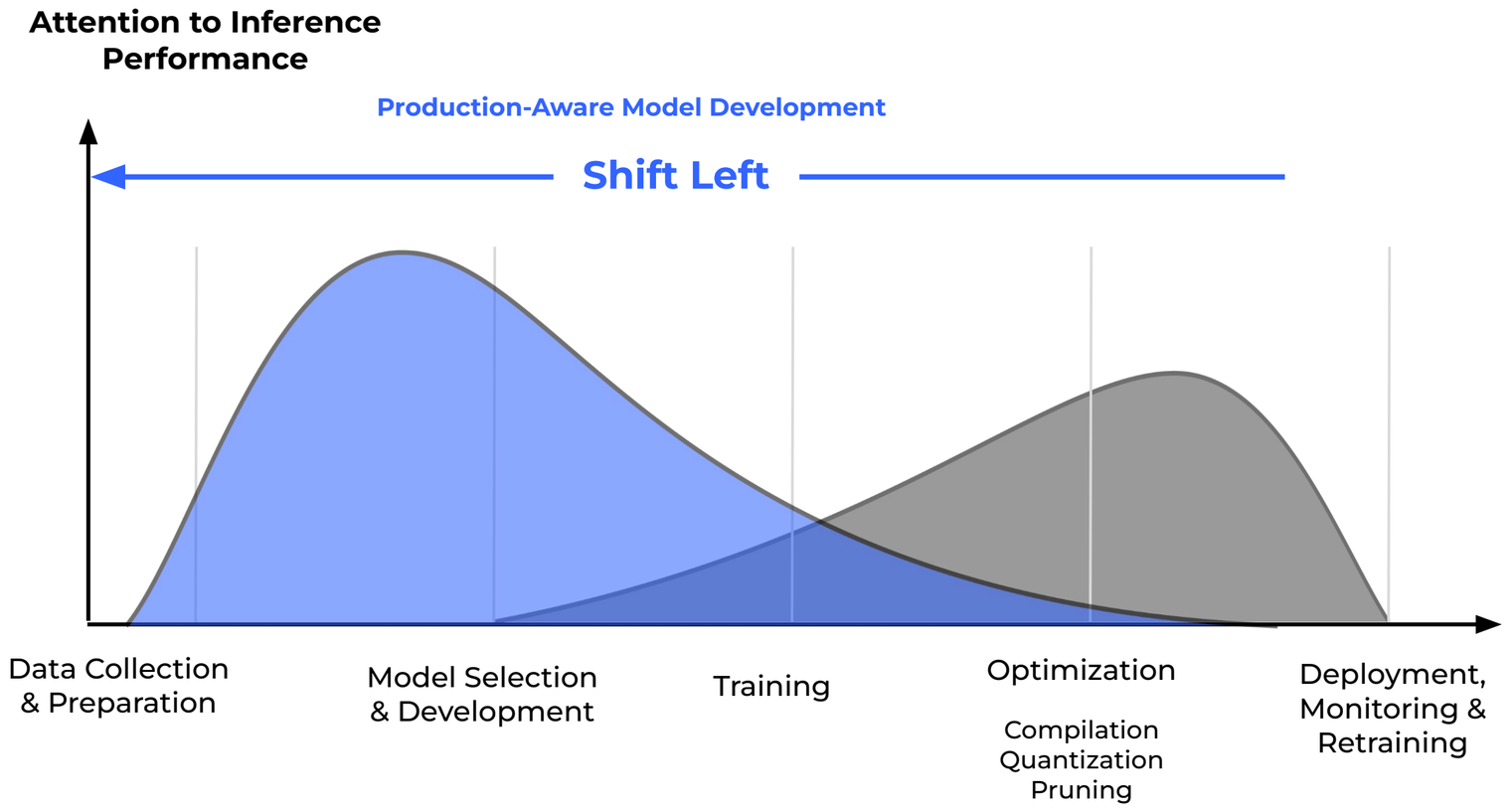


x8

NVIDIA Jetson Xavier NX GPU

deci

Teams must adopt efficient processes for developing production-grade models



Results using Deci early

5X better performance

80% reduction in dev time

-30% reduction in dev cost

Guarantee of success

deci

Key considerations for developing DL for edge deployments

Model Architecture Design

Model Runtime Optimization

Efficient Deployment

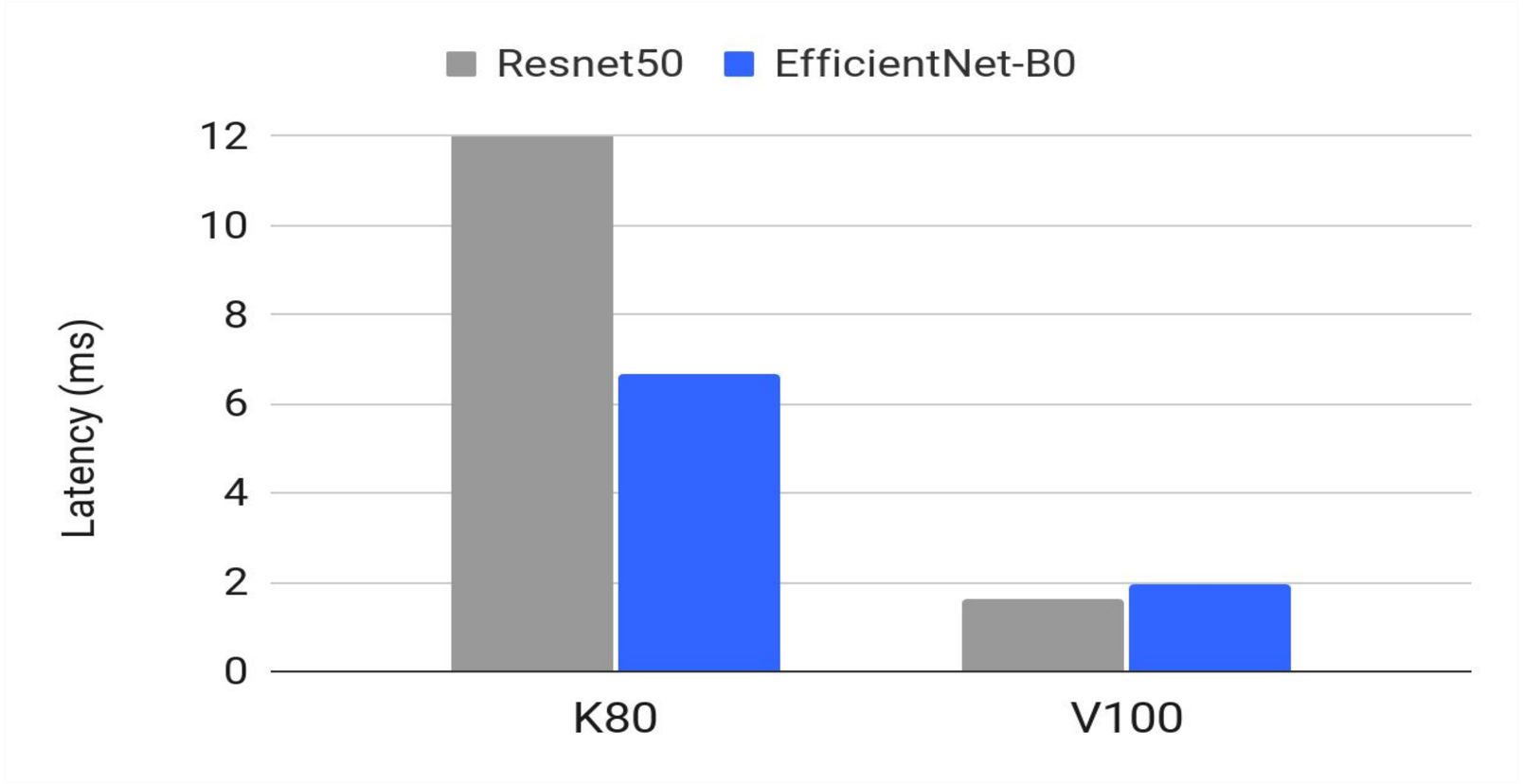
deci

Model Architecture Design

The importance of hardware awareness in model design

deci

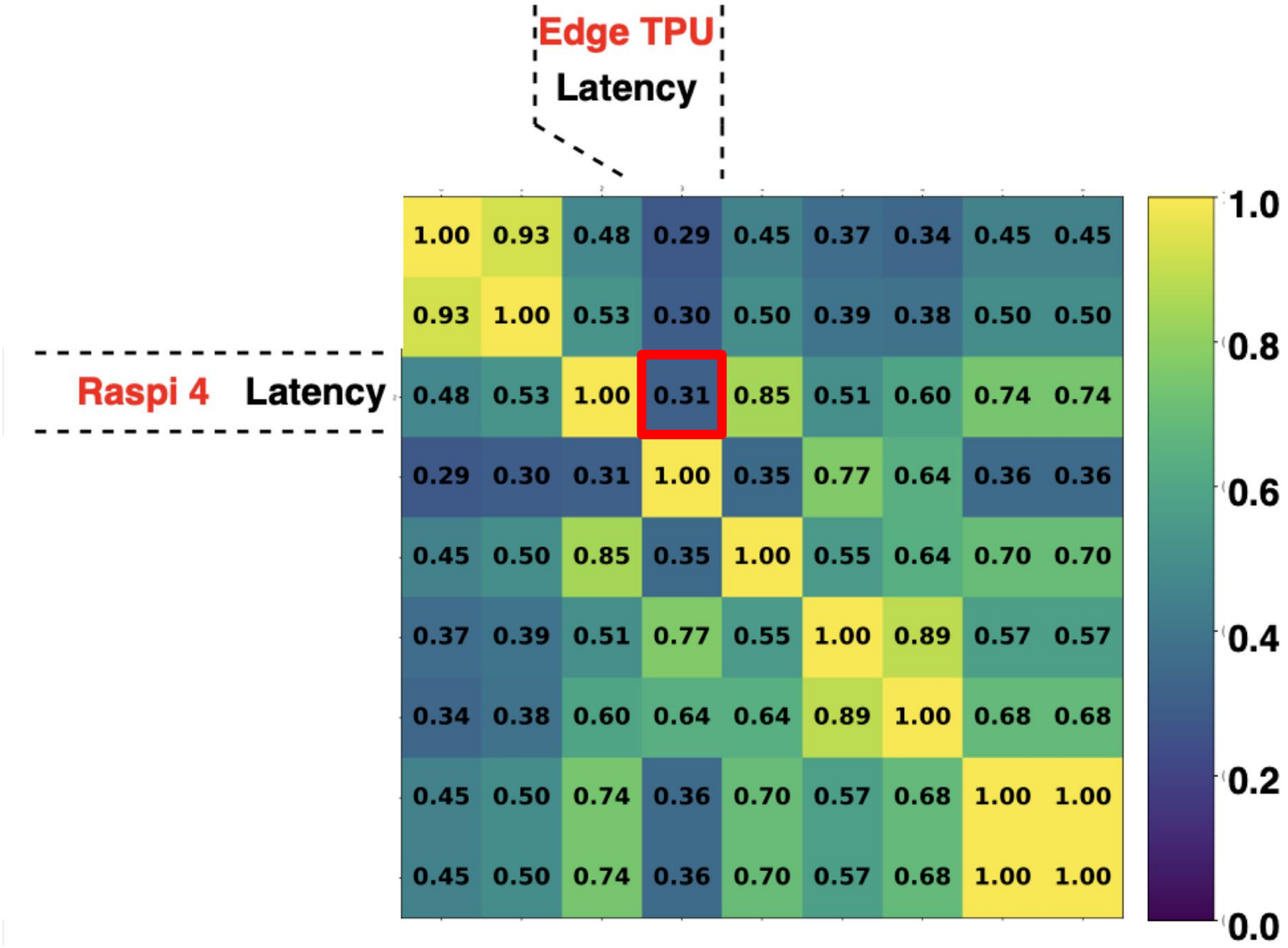
Models will Perform Differently on Different Hardware



[TRT FP16, CUDA-10]

deci

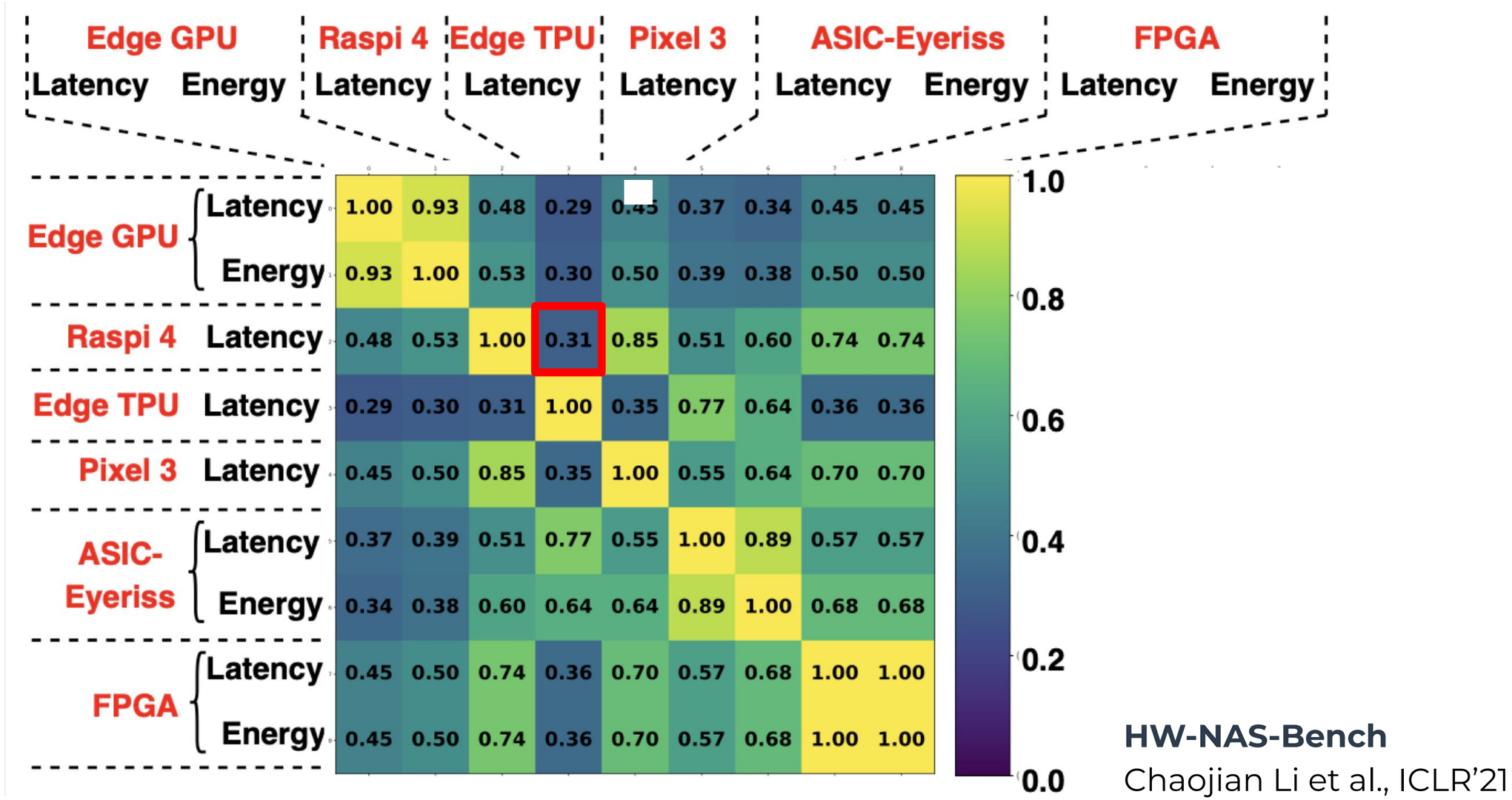
Inconsistent Performance Across HW!



HW-NAS-Bench
Chaojian Li et al., ICLR'21

deci

Inconsistent Performance Across HW!



HW-NAS-Bench
Chaojian Li et al., ICLR'21

deci

Open source SOTA models VS HW-Aware NAS Generated Models

Efficient Frontier - Semantic Segmentation Measured on NVIDIA Jetson

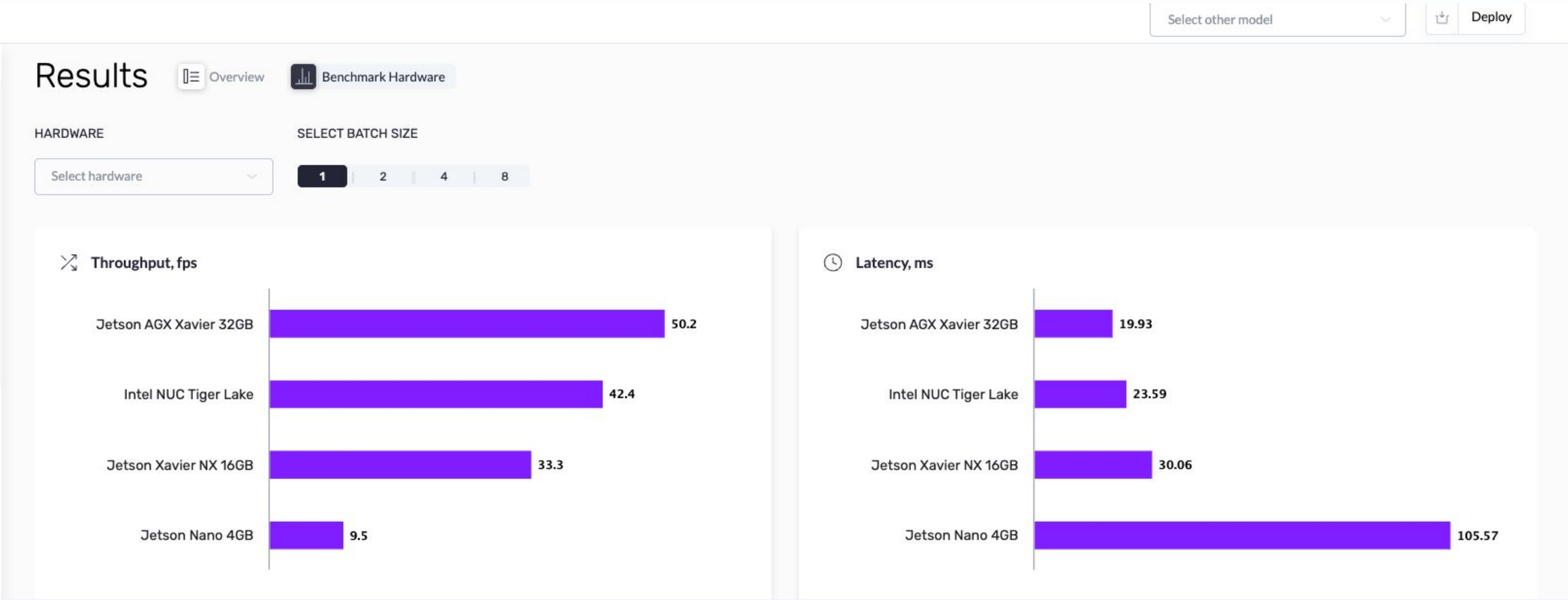
All models were complied and quantized to FP16 with NVIDIA TensorRT

deci

Model Runtime Optimization

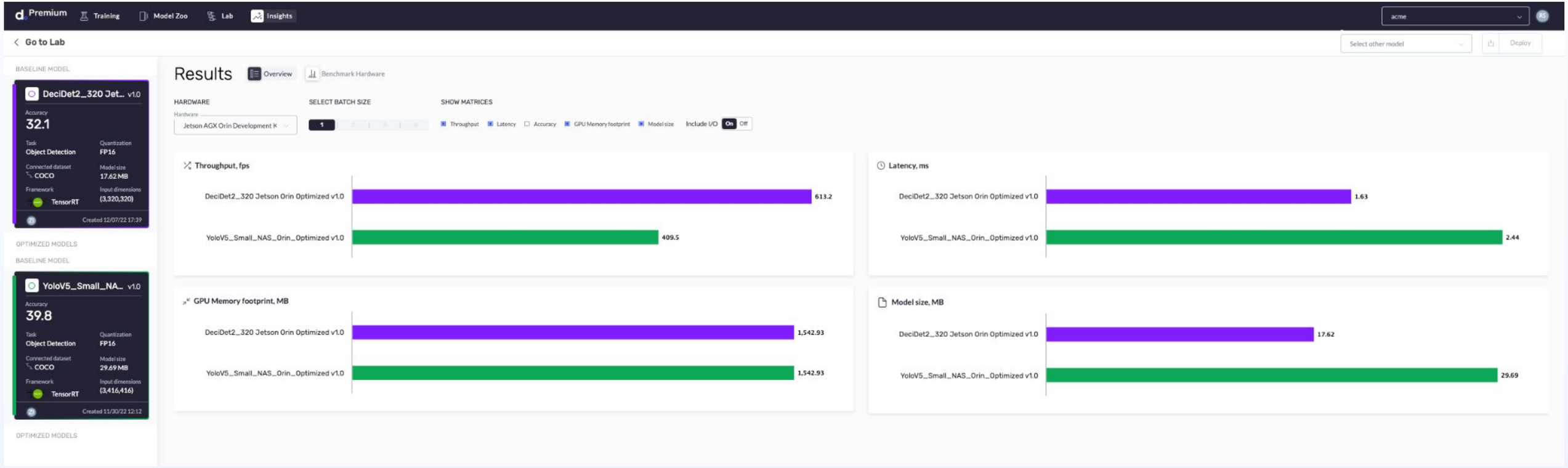
deci

Benchmark Your Models on various edge devices



deci

Accelerate inference speed and reduce model size and memory footprint with model Compilation and Quantization

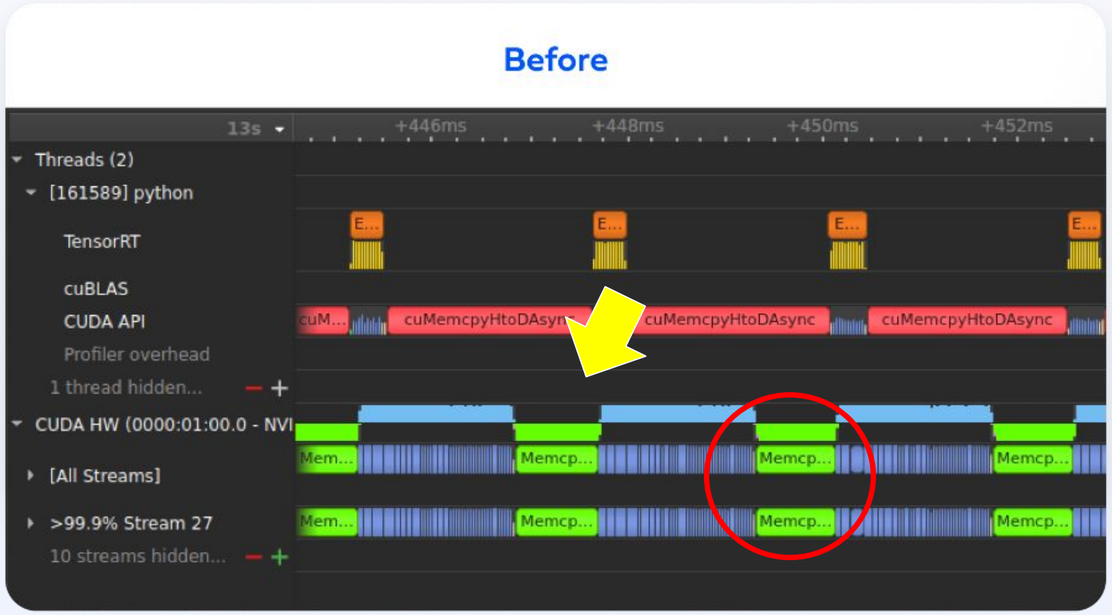


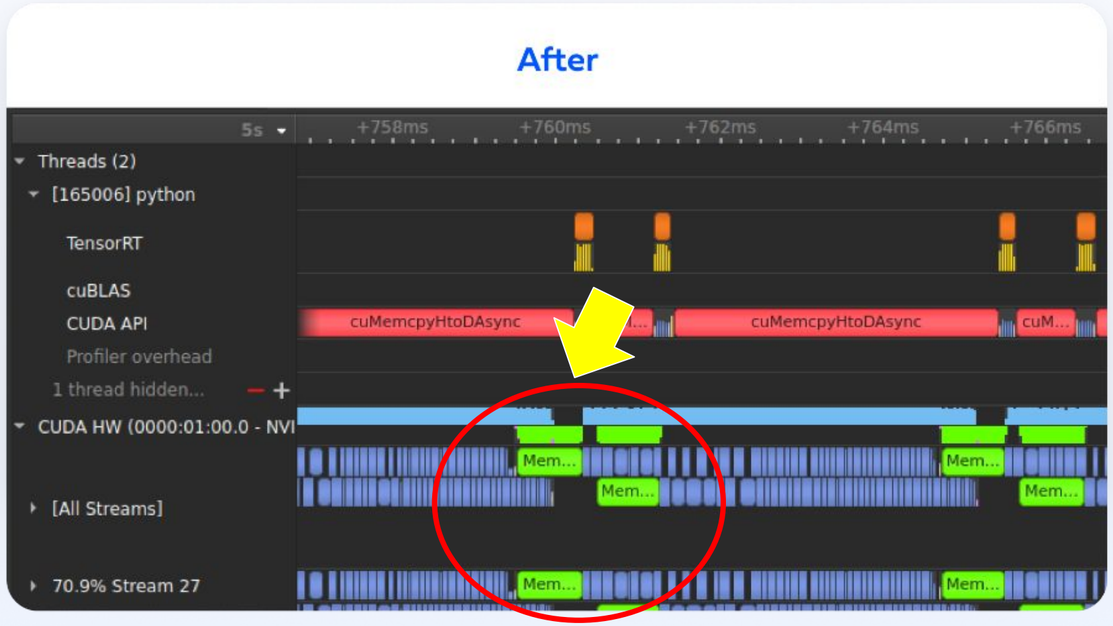
deci

Efficient Deployment

Accelerate Inference with Better HW utilization

- Is your GPU being fully utilized?
- Leverage advanced capabilities as asynchronous inference and concurrent inferencing.



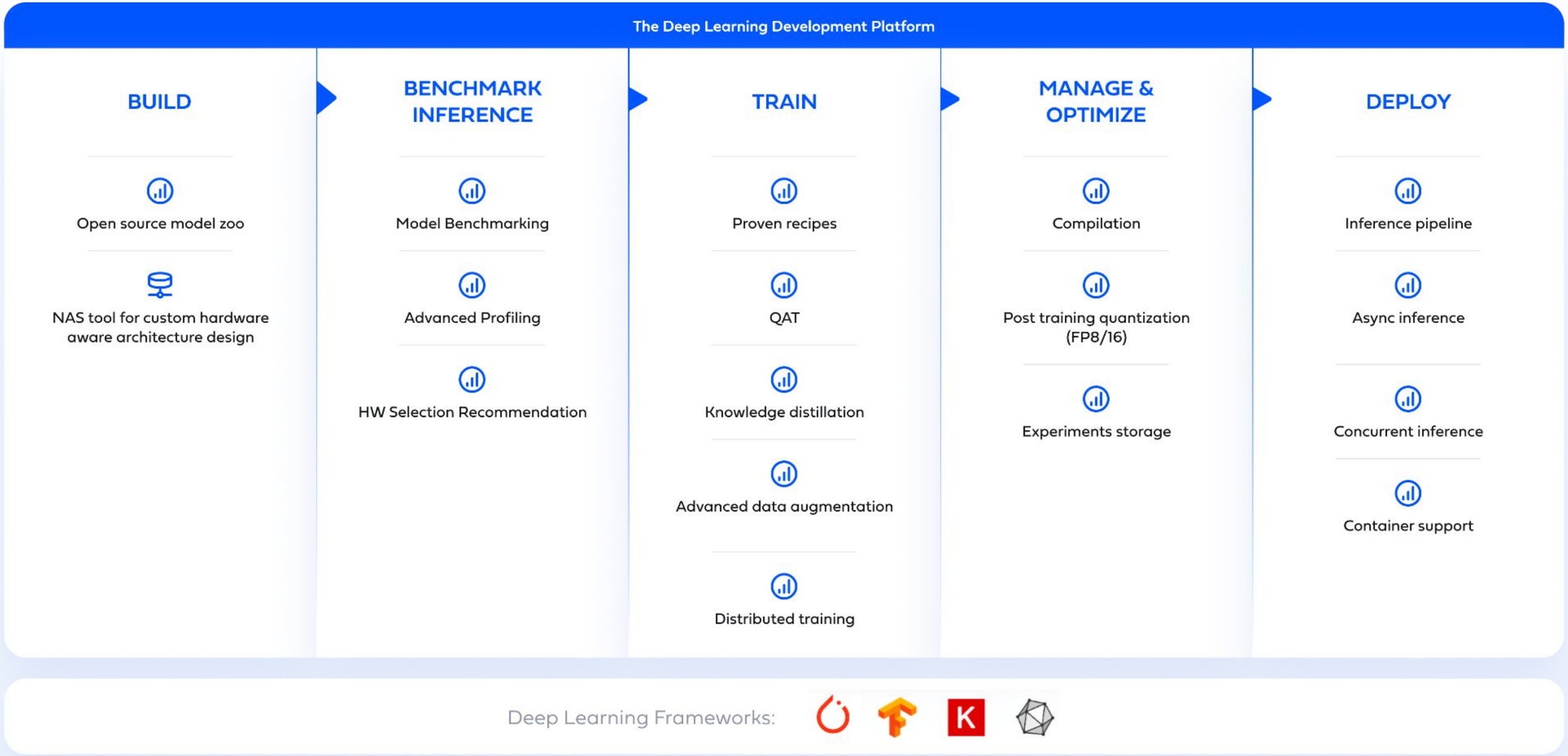


deci

How to Quickly Deploy on Edge Devices with Deci

Deci Deep Learning Development Platform

Powered by Neural Architecture Search



Deci Deep Learning Development Platform

Powered by Neural Architecture Search

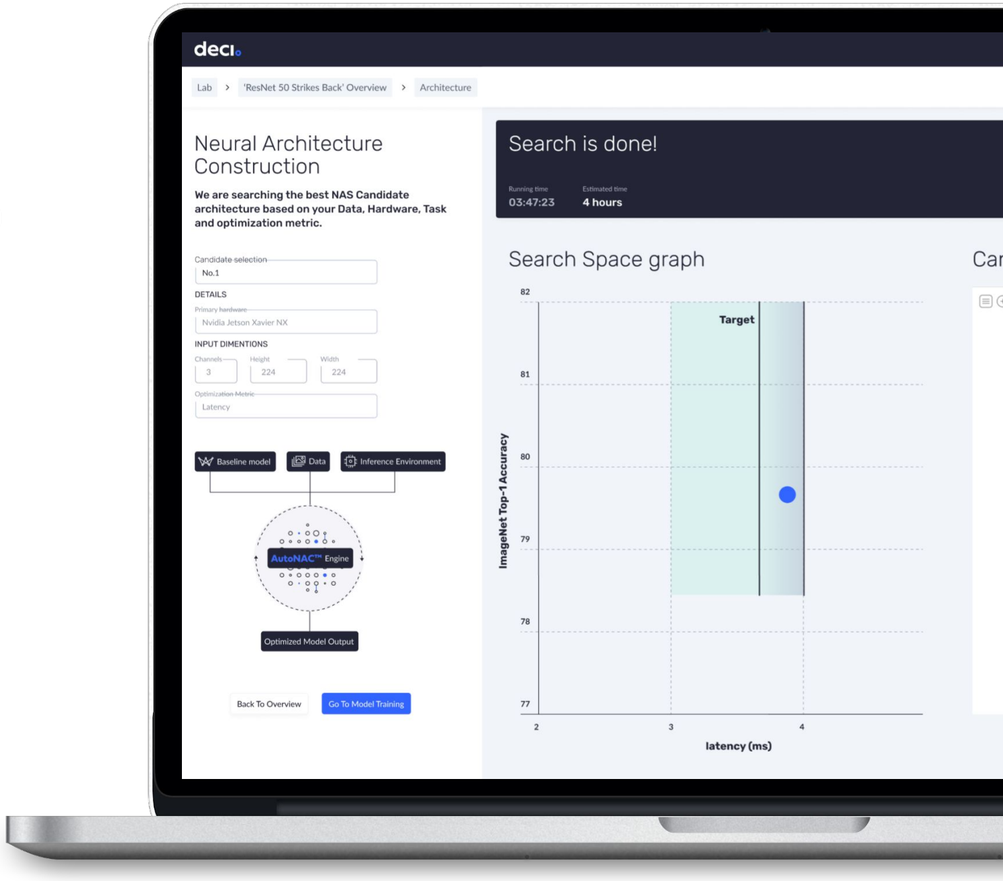
- **Outperform SoTA with Custom NN Architectures**
  Save time and guarantee success by building accurate & fast architectures tailored for your performance targets & hardware

- **Fast and Efficient Training Library**
  - Easily leverage advanced training techniques (Quantization Aware Training, Knowledge distillation)
  - Get SOTA hyperparameter recipes

- **Automated Compilation & Quantization**
  Optimize your trained models for your HW with a click of a button
  (leveraging TensorRT, OpenVino etc.)

- **Inference Engine**
  Deploy with 3 lines of code using Deci's Python Inference Runtime Engine

- **Expert Support**
  Dedicated deep learning expert support



deci

With Deci, You can Build Better Models, Faster.

Gain Unparalleled Inference Performance

Up to **5X**

Acceleration

Shorten Time to Market

80%

Reduction in dev time

Lower Development Cost

30%

Reduction in dev cost

deci

Core Technology - Deci's AutoNAC Engine

Hardware-Aware Neural Architecture Search for DL Inference Efficiency

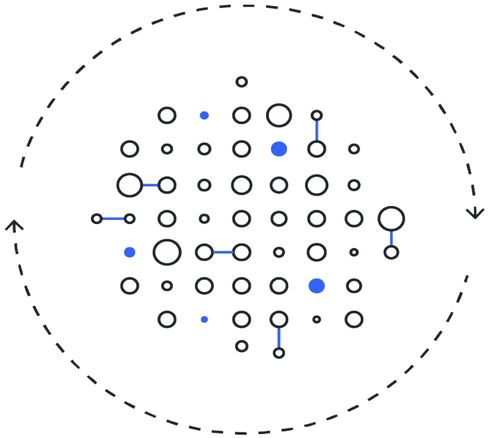
Input

Task

Data characteristics

Inference Environment & HW

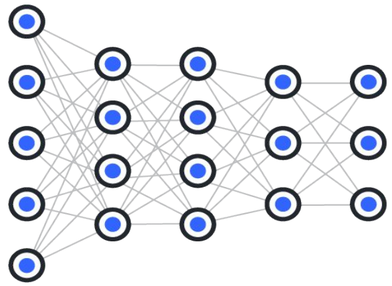
AutoNAC™ Engine



- ✓ Hardware aware
- ✓ Compilation & quantization aware
- ✓ Search time ±2.5 of training time
- ✓ Multi-objective search:

Accuracy | Latency | Throughput | Model size | Memory footprint

New Architecture



- ✓ Output format - NN Module (PyTorch)
- ✓ Can accommodate other formats per need
- ✓ Trainable on user's premises with its own dataset

deci

Build custom models with Deci

Set Targets & HW

Set your Inference HW, accuracy & performance targets

AutoNAC Engine

The engine searches for the most suitable DNN & delivers an architecture

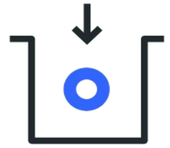
Train

Get your custom architecture & train it with your data on prem

Auto Compile & Quantize

Automatically compile and quantize your model

Deploy



Use Deci's tools to gain additional boost

Performance Targets

- **Accuracy Preserving**
- **Latency**
- **Throughput**
- **Model size**
- **Memory footprint**

deci

Scaling up AI-based Security Application

The Challenge

A security-centric Fortune 500 enterprise struggled to run real-time object detection on a Jetson device connected to multiple cameras streams in order to support its pedestrian detection application.

- Security
- Object Detection
- NVIDIA Jetson Xavier NX

Results

Using Deci's AutoNAC engine, the company was able to increase the number of video streams connected to its existing Jetson Xavier NX devices while also improving the accuracy from 40.5 to 41.3 mAP. Throughput was increased by 1.5x from 60 FPS to 94 FPS.

1.5x	Boost in Throughput
+0.8%	Increase in Accuracy
2x	More Video Streams per Device

deci

Enabling Real-Time Performance at the Edge

The Challenge

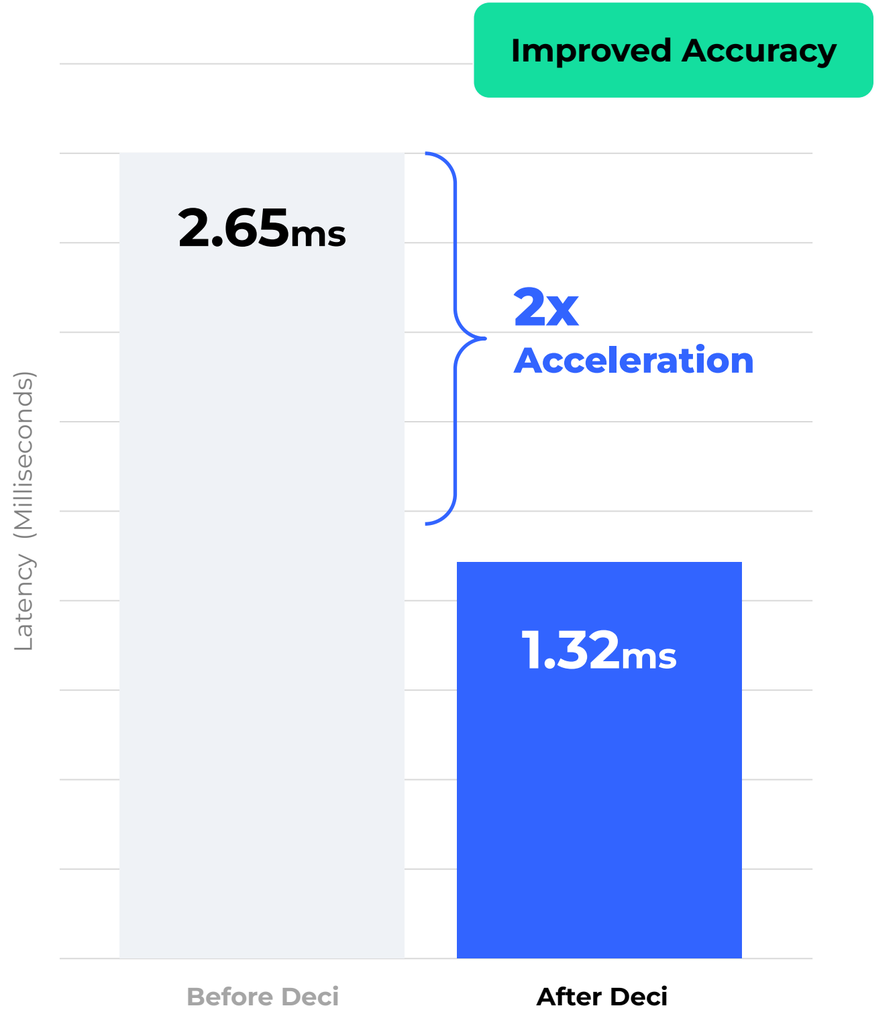
An automotive company wanted to improve the latency of a model powering their road condition estimation system. The poor latency of their baseline ResNet50 model was impacting the real-time performance on their target hardware, a NVIDIA Xavier AGX.

- Automotive
- Image Classification
- NVIDIA Jetson Xavier AGX

Results

The team used Deci's AutoNAC engine to build a customized model and was able to gain real-time performance on edge and shorten time to market.

2x	Latency Acceleration
+4%	Increase in accuracy
2x	Boost in Throughput
3.2x	Model Size Reduction



deci

Use Cases - How AI teams are using Deci?

Enables Inference on Edge Devices

Enable inference on resource constrained devices (e.g. CPU, Edge devices, mobile etc.)

Ship Better Products with Improved Inference Performance

Outperform SOTA models with better accuracy, latency, throughput, smaller memory footprint & model size.

Reduce Training & Inference Costs

Maximize Hardware utilization. Make the of most of your current hardware or more to a more affordable one. Cut up to 80% of your cloud costs.

Simplify Development, Shorten Time to Market

Automate model development & optimization steps. Eliminate uncertainty, guarantee success in production and reach production faster.

deci

Empowering Product Creators to Harness Edge AI and Vision

The Edge AI and Vision Alliance (www.edge-ai-vision.com) is a partnership of 100+ leading edge AI and vision technology and services suppliers, and solutions providers

Mission: To inspire and empower engineers to design products that perceive and understand.

The Alliance provides low-cost, high-quality technical educational resources for product developers

Register for updates at www.edge-ai-vision.com

The Alliance enables edge AI and vision technology providers to grow their businesses through leads, partnerships, and insights

For membership, email us: membership@edge-ai-vision.com

edge ai + vision ALLIANCE™

Join us at the Embedded Vision Summit
May 22-25, 2023—Santa Clara, California

The only industry event focused on practical techniques and technologies
for system and application creators

- *"Awesome! I was very inspired!"*
- *"Fantastic. Learned a lot and met great people."*
- *"Wonderful speakers and informative exhibits!"*

Embedded Vision Summit 2023 highlights:

- **Inspiring keynotes** by leading innovators
- High-quality, practical **technical, business and product talks**
- Exciting **demos**, **tutorials** and **expert bars** of the latest applications and technologies

Visit www.EmbeddedVisionSummit.com to learn more and register

Q&A

Thank You.

deci