SCALING AT THE EDGE

Accelerating Deep Learning Performance on Edge Devices

deci

Yonatan Geifman, PhD

CEO and Co-Founder, Deci

Common barriers to deployment on edge devices

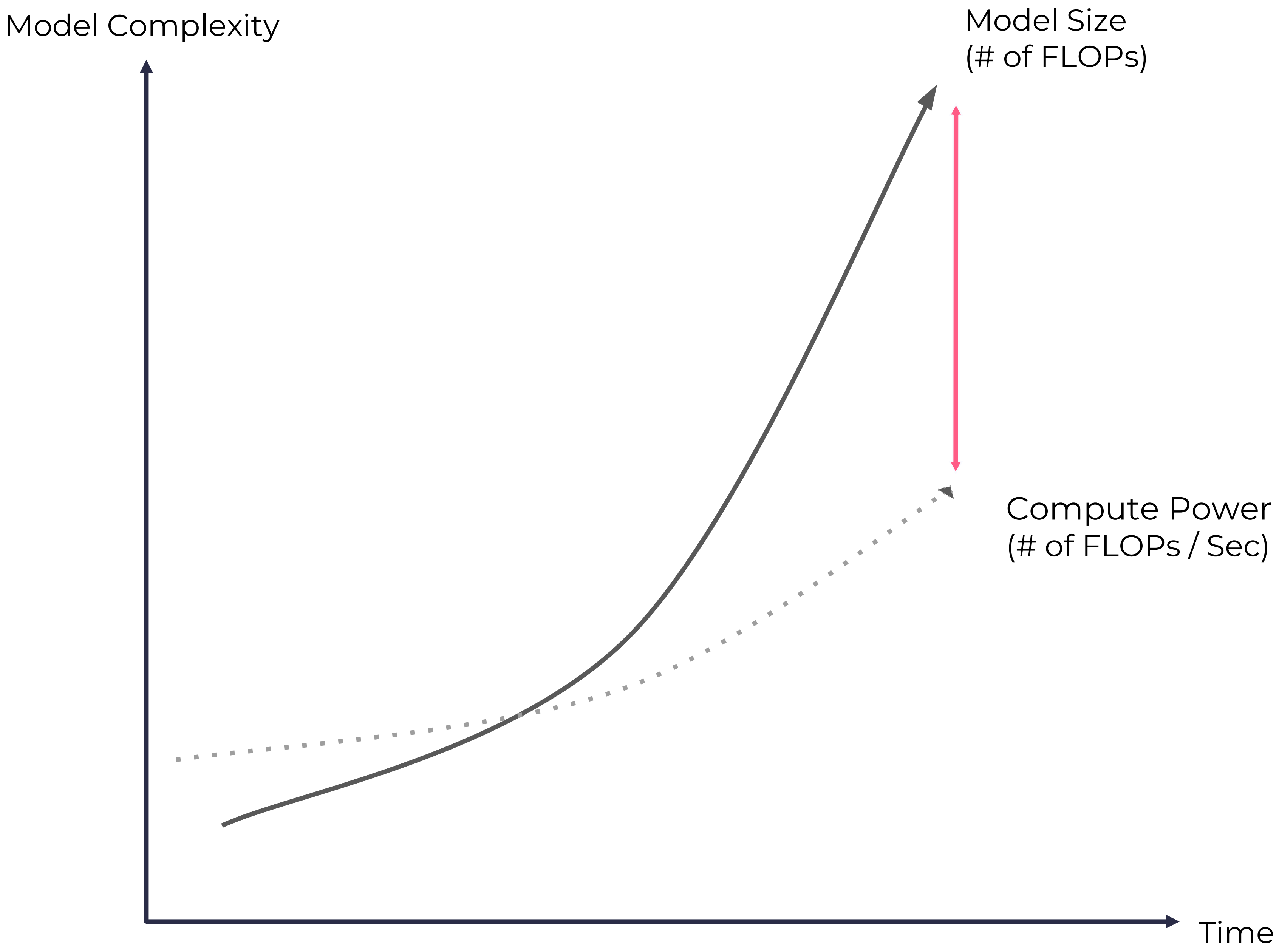
Inability to deploy on edge devices

Unsatisfactory Accuracy or performance

Long development cycle

© 2022 Deci. All Rights Reserved.

Models’ power hunger is increasing rapidly

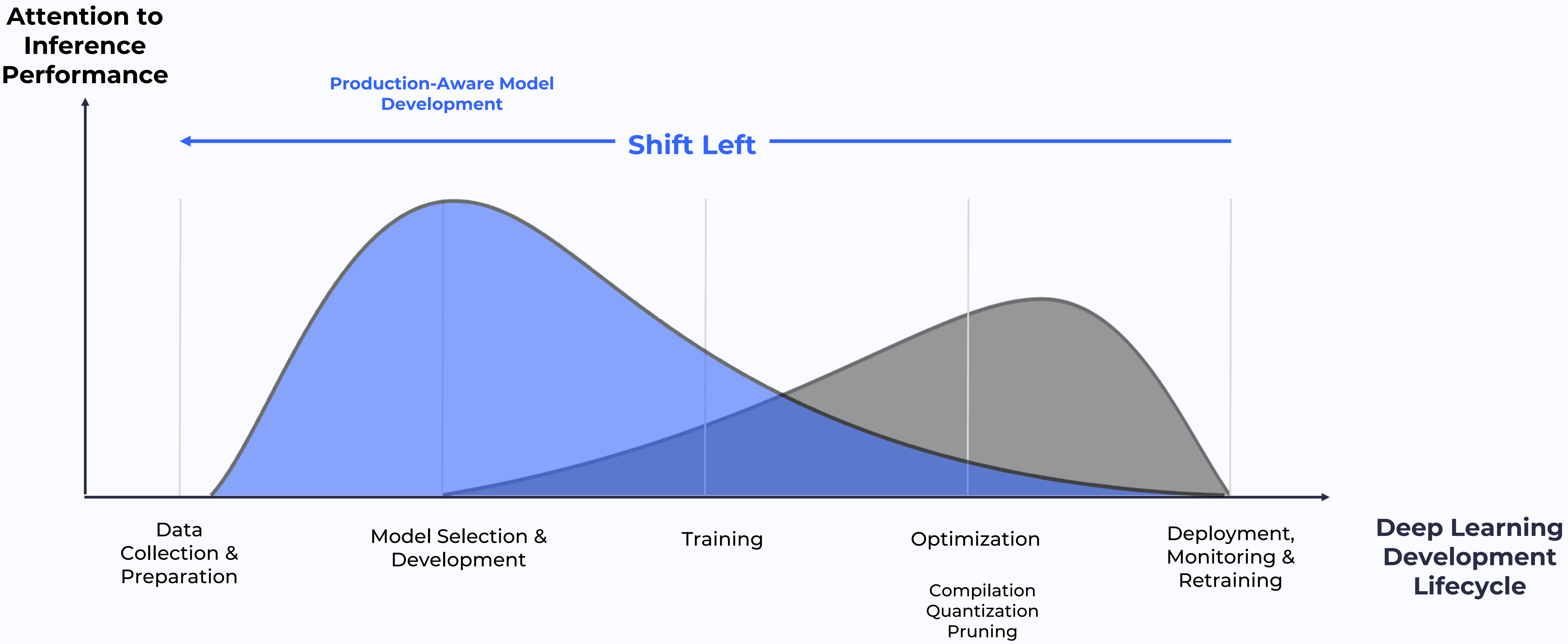


The AI Efficiency Gap leads to:

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

© 2022 Deci. All Rights Reserved.

AI Efficiency calls for a new development paradigm



© 2022 Deci. All Rights Reserved.

Deci Deep Learning Development Platform

Powered by Neural Architecture Search

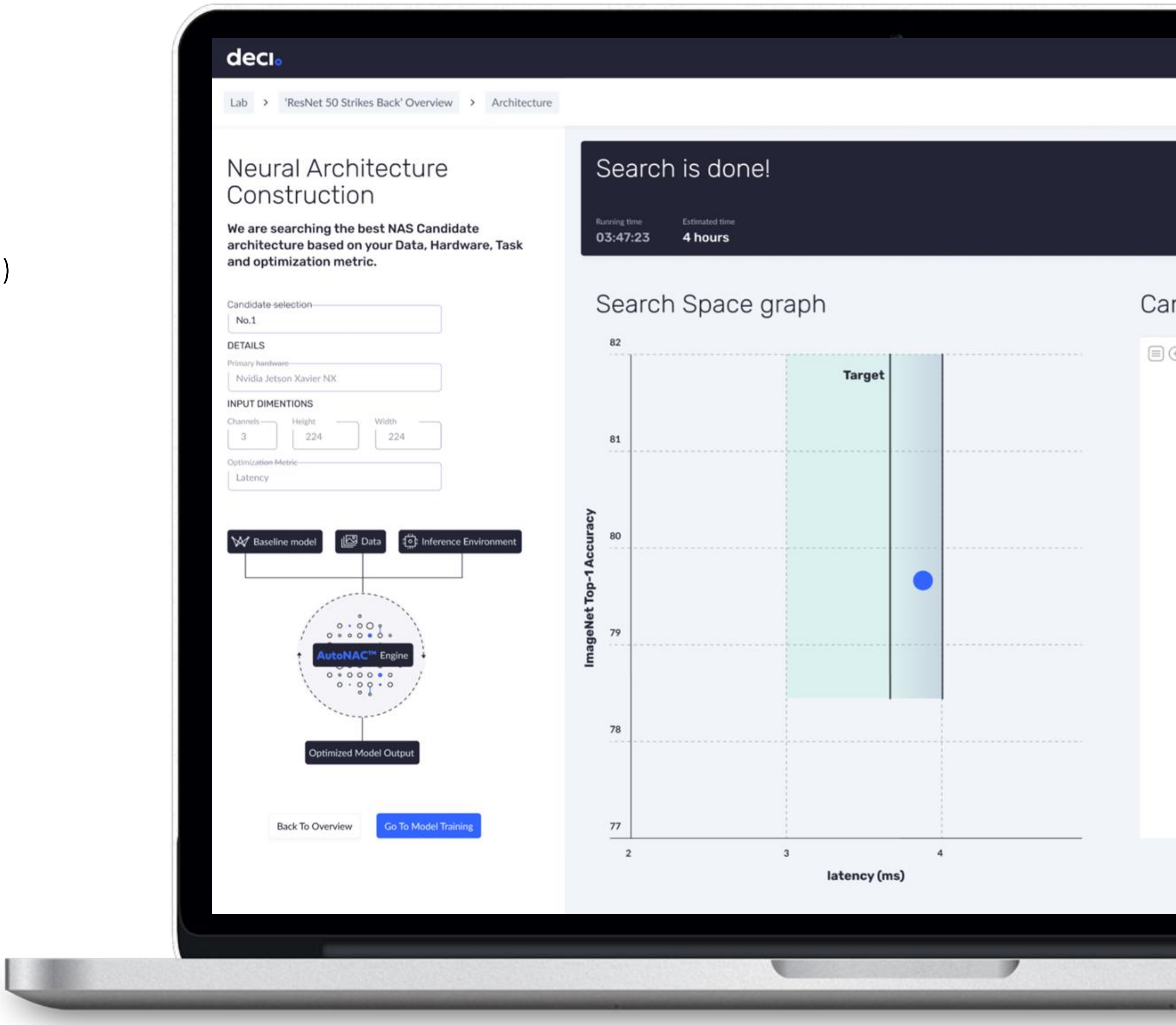
- **Outperform SoTA with Custom NN Architectures**
  Save time and guarantee success by building accurate & fast architectures tailored for your performance targets & hardware

- **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

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

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



Neural Architecture Construction

We are searching the best NAS Candidate architecture based on your Data, Hardware, Task and optimization metric.

Candidate selection: No.1

DETAILS

Primary hardware: Nvidia Jetson Xavier NX

INPUT DIMENTIONS

Channels: 3 Height: 224 Width: 224

Optimization Metric: Latency

Search is done!

Running time 03:47:23 Estimated time 4 hours

© 2022 Deci. All Rights Reserved.

Deci's AutoNAC Engine:
Hardware-Aware Neural Architecture Search for DL Inference Efficiency

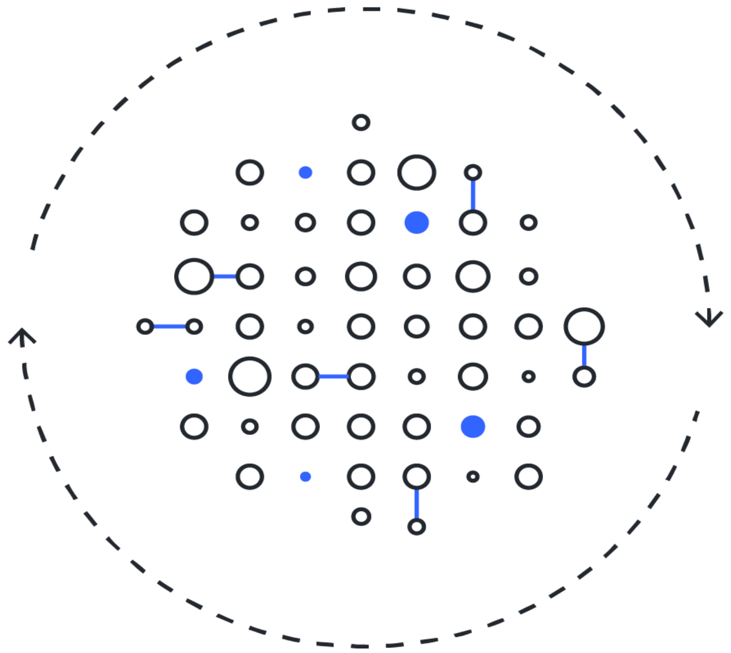
Input

Task

Data characteristics

Inference Environment & HW

AutoNAC™ Engine



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

Accuracy

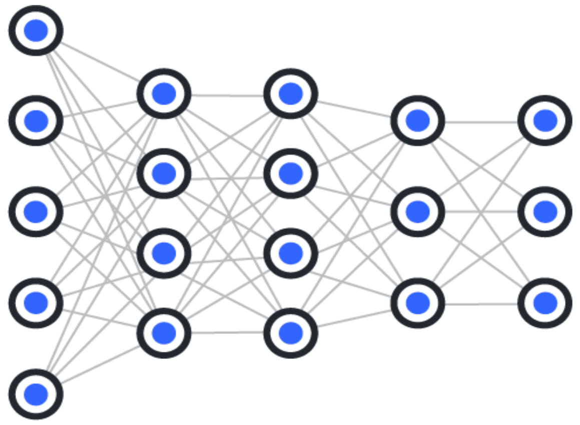
Latency

Throughput

Model size

Memory footprint

New Architecture

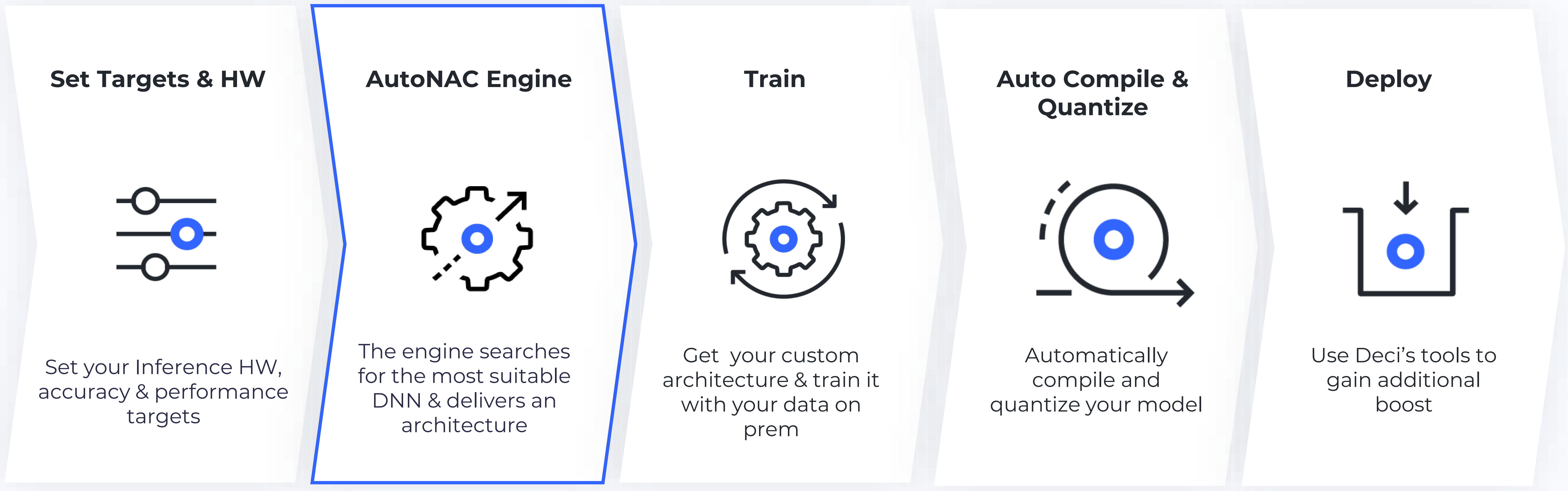


- ✓ Output format - NN Module (PyTorch)
- ✓ Can accommodate other formats per need
- ✓ Trainable on user's premises with its own dataset

© 2022 Deci. All Rights Reserved.

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

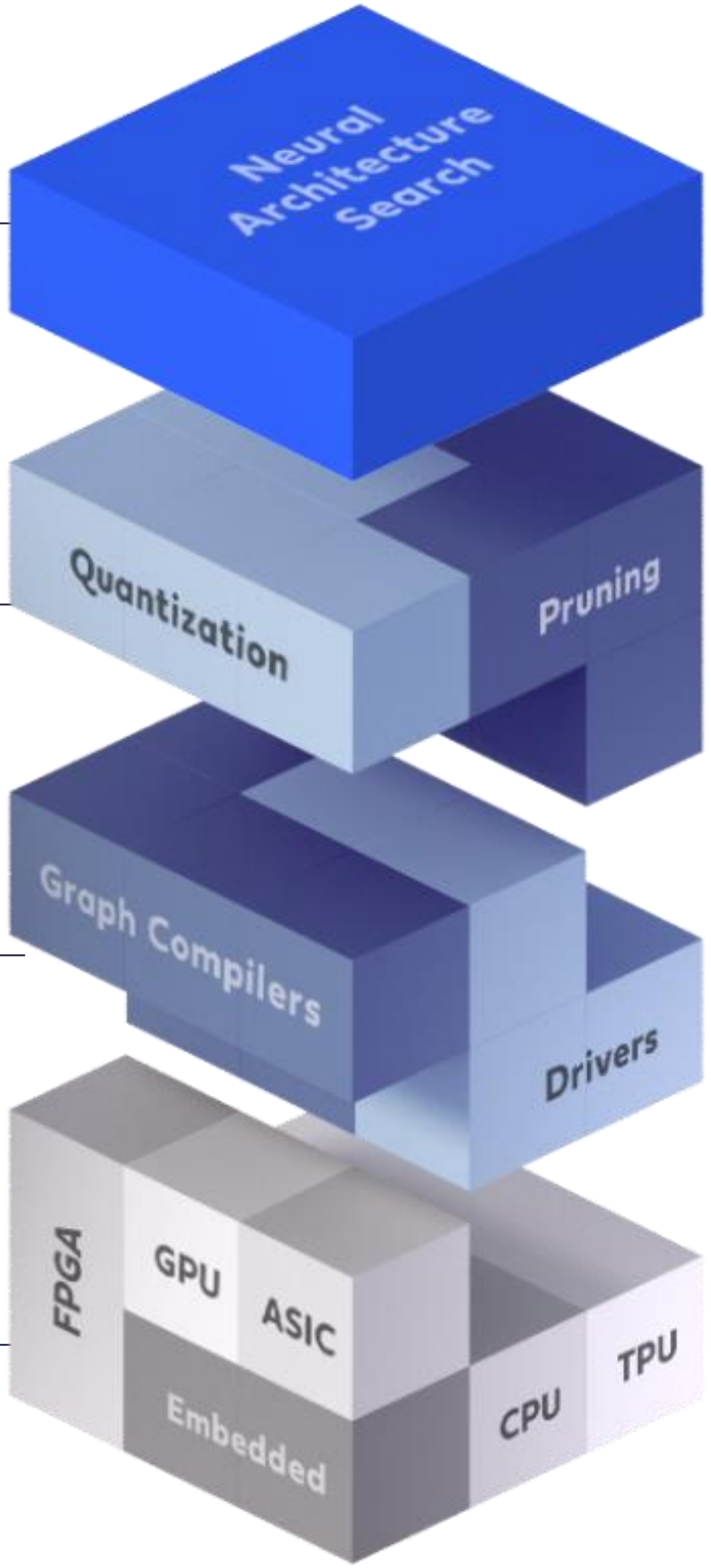
Use Deci's tools to gain additional boost

Performance Targets

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

© 2022 Deci. All Rights Reserved.

An **inherent advantage** driven by Deci's algorithmic level optimization



Hardware Aware Model Architecture construction

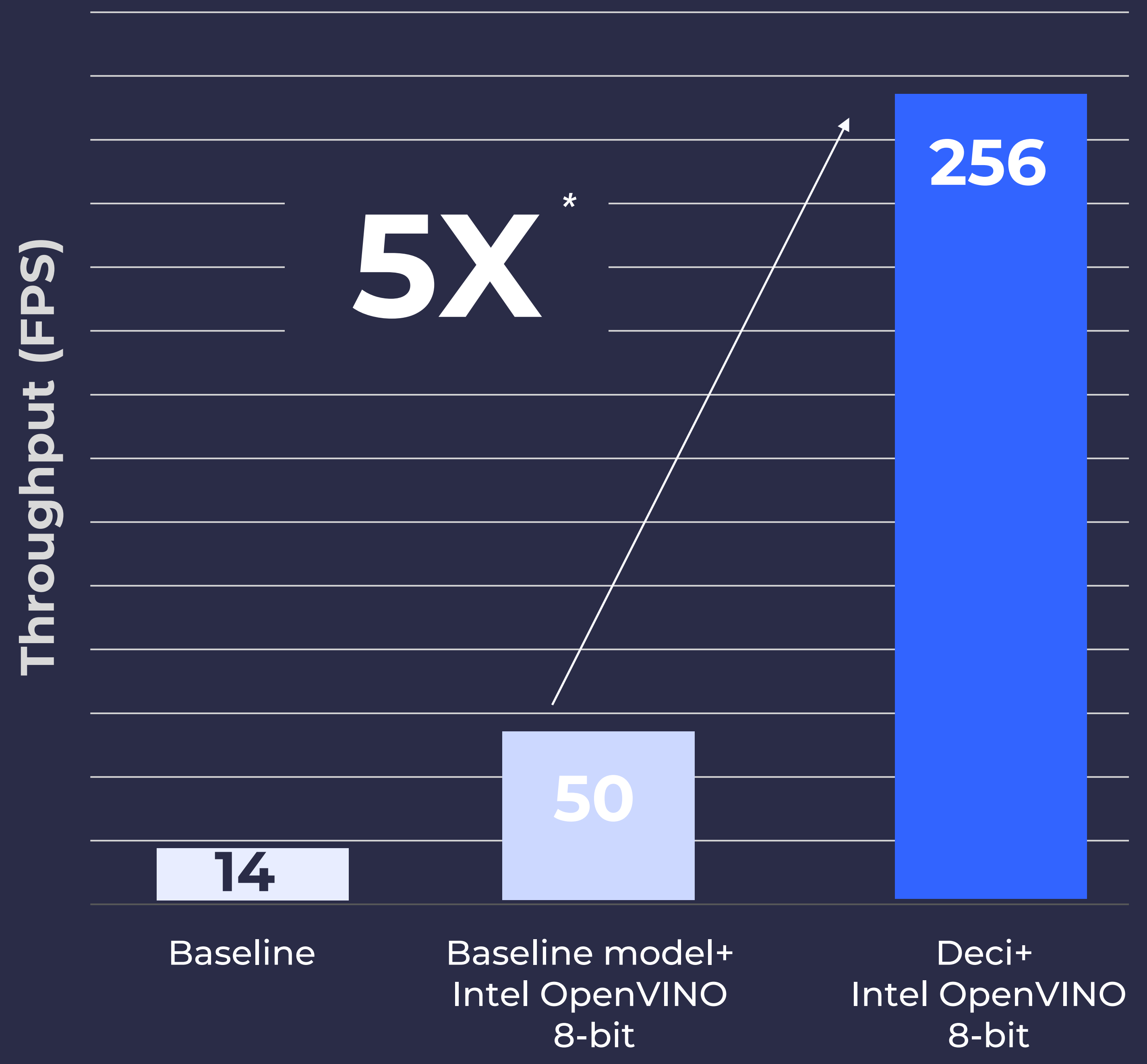
Model Compression Techniques

Runtime Optimization

Inference Hardware

MLPerf

A comparison of ResNet 50 & DeciNet on ImageNet



Measured Intel Intel® Xeon® Platinum 8375C CPU @ 2.90GHz

© 2022 Deci. All Rights Reserved.

*Performance claim data: https://community.intel.com/t5/Blogs/Tech-Innovation/Artificial-Intelligence-AI/Intel-Collaboration-With-Deci-Boosts-AI-Performance-on-Intel/post/1404029

With Deci, You can Build Better Models, Faster.

Gain Unparalleled Inference Performance

Up to **5X***

acceleration

Shorten Time to Market

3 weeks

on average to reach a production-ready model

Guarantee Success In Production

Built for purpose & Expert Support

© 2022 Deci. All Rights Reserved.

*Performance claim data: https://community.intel.com/t5/Blogs/Tech-Innovation/Artificial-Intelligence-AI/Intel-Collaboration-With-Deci-Boosts-AI-Performance-on-Intel/post/1404029

Use Cases - How AI teams are using Deci?

Enables Inference on Edge Devices

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

Boost 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.

© 2022 Deci. All Rights Reserved.

Thank You.

deci

Notices and Disclaimers

Performance varies by use, configuration and other factors. Learn more on the Performance Index site.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

Intel is committed to respecting human rights and avoiding complicity in human rights abuses. See Intel Global Human Rights Principles. Intel's products and software are intended only to be used in applications that do not cause or contribute to a violation of an internationally recognized human right.

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.