mi2

Machine Intelligence

Modern Infrastructure

http://mi2.live

Accelerating Neural Networks with Intel Movidius

What is MI2?

MI2 Webinars focus on the convergence of **machine intelligence** and **modern infrastructure**. Every alternate week, I deliver informative and insightful sessions covering cutting-edge technologies. Each webinar is complemented by a tutorial, code snippets, and a video.

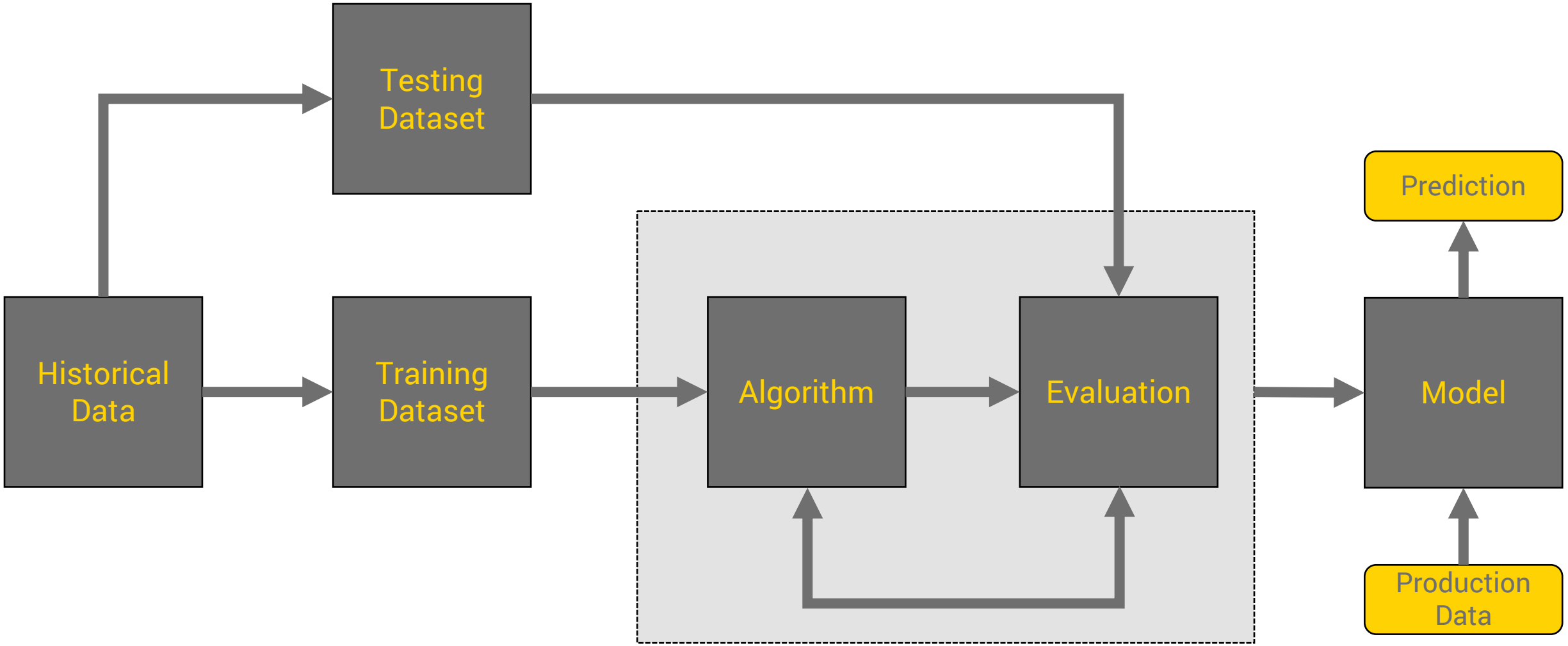
MI2 strives to be an independent and neutral platform for exploring emerging technologies.

Register at http://mi2.live

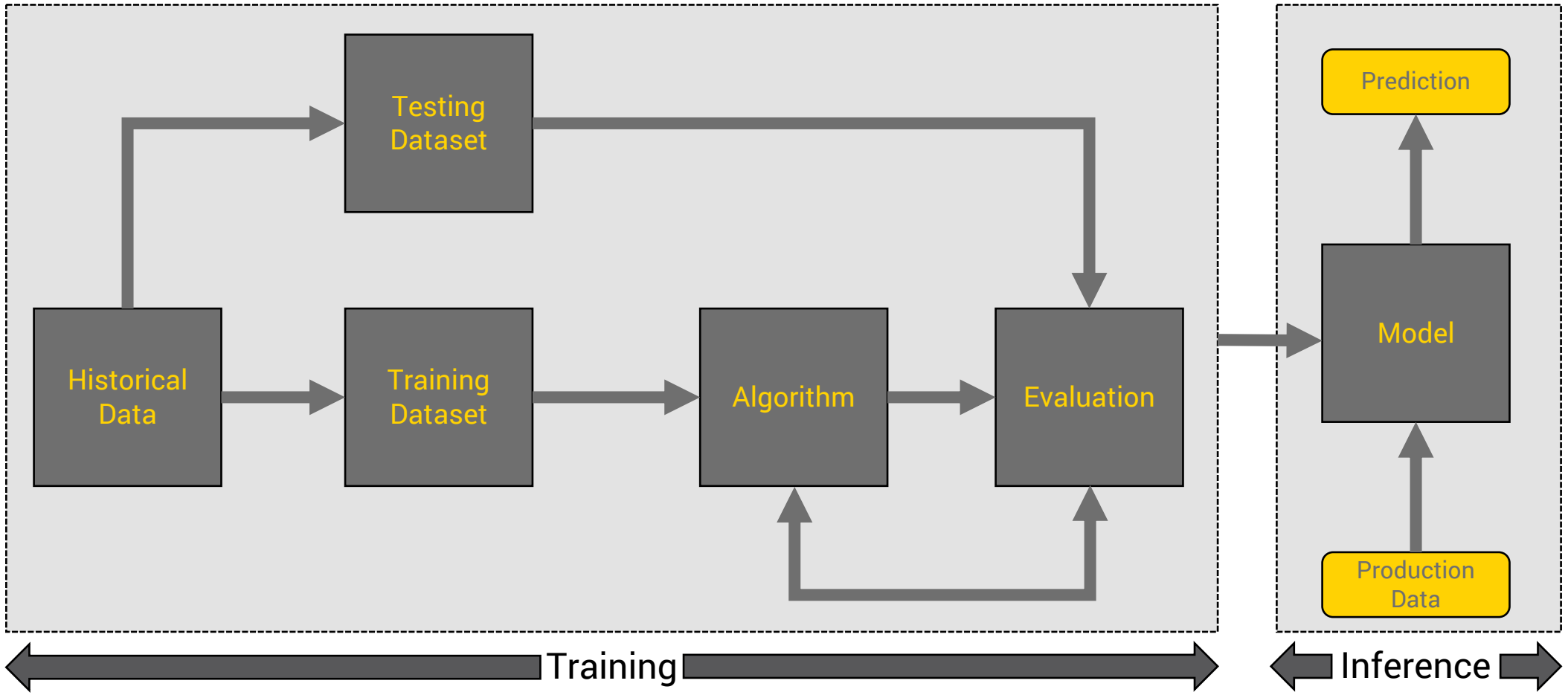
Objectives

- Lifecycle of an ML model
- The need to accelerate ML model inferencing
- Overview of Intel Movidius
- Optimizing models for Intel Movidius
- Demo
- Summary

Lifecycle of an ML model



Lifecycle of an ML model

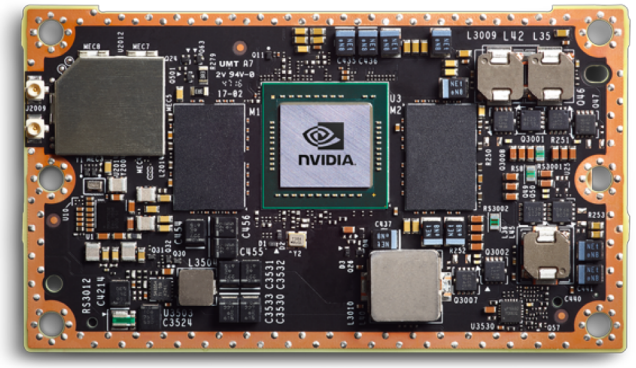


Accelerating ML Model Training and Inferencing

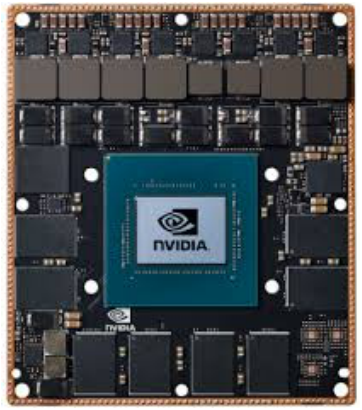
- Training & inferencing are compute intensive
- Each epoch goes through complex computation
- GPUs help accelerate training through massive parallelization
- **Training goal – High throughput**
- Fully trained models run in constrained environments
- AI @ Edge is the most common deployment scenario
- Edge computing environments may not support GPUs
- **Inference goal – Reduced latency**

Accelerating ML Model Inferencing at Edge

- Complement CPUs with purpose-built chips and co-processors (accelerators)
- Compute-intensive tasks are offloaded from CPUs

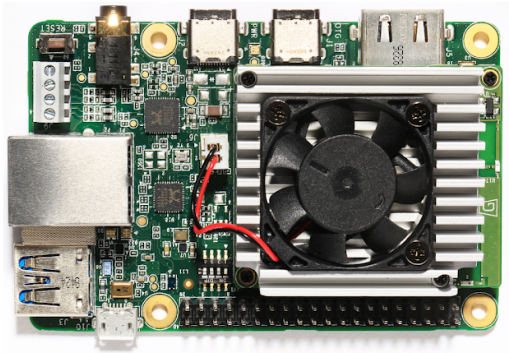


NVIDIA Jetson TX2



NVIDIA Xavier

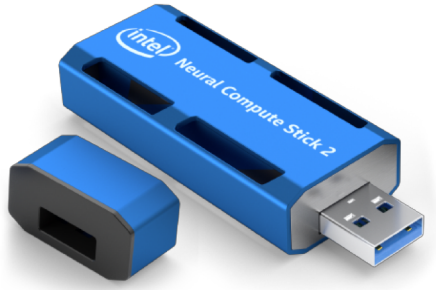
Intel Myriad



Google Edge TPU

Overview of Intel Movidius Neural Compute Stick

- Small, fanless, USB-based deep learning accelerator device based on Intel Myriad X Vision Processing Unit (VPU)
- Available in two versions
  - NCS 1
  - NCS 2

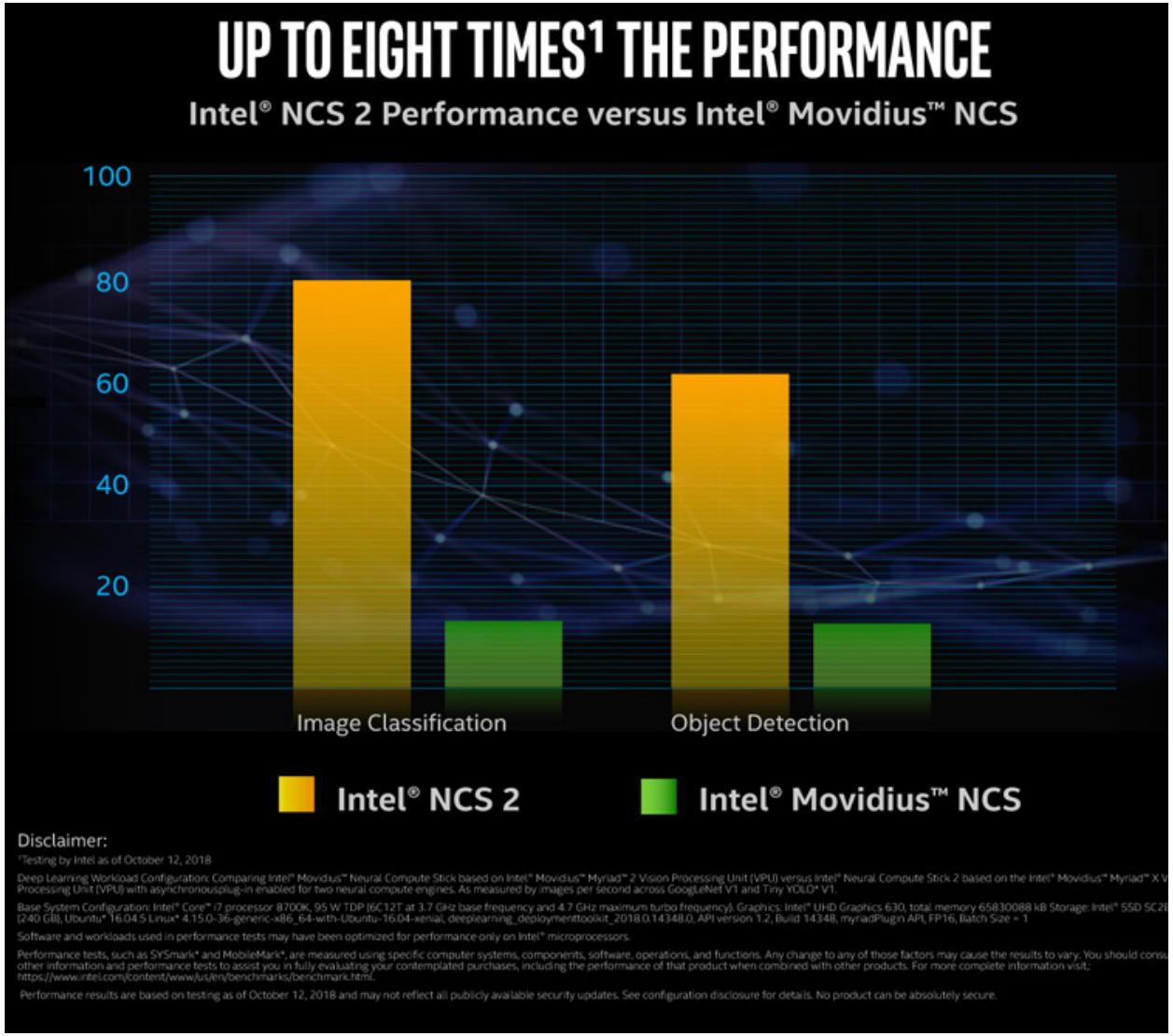


Intel NCS V1

- Processor: Intel® Movidius™ Myriad™ 2 Vision Processing Unit (VPU)
- 8 processing cores
- Supported frameworks: TensorFlow and Caffe
- Connectivity: USB 3.0 Type-A
- USB stick dimensions: 2.85 in. x 1.06 in. x 0.55 in. (72.5 mm x 27 mm x 14 mm)
- Operating temperature: 0° C to 40° C
- Minimum system requirements:
  - Ubuntu 16.04
  - Raspberry Pi 3 Model B running Stretch desktop
  - USB 2.0 Type-A port (USB 3.0 recommended)
- 1 GB RAM
- 4 GB free storage space
- Software: NCS SDK

Intel NCS V2

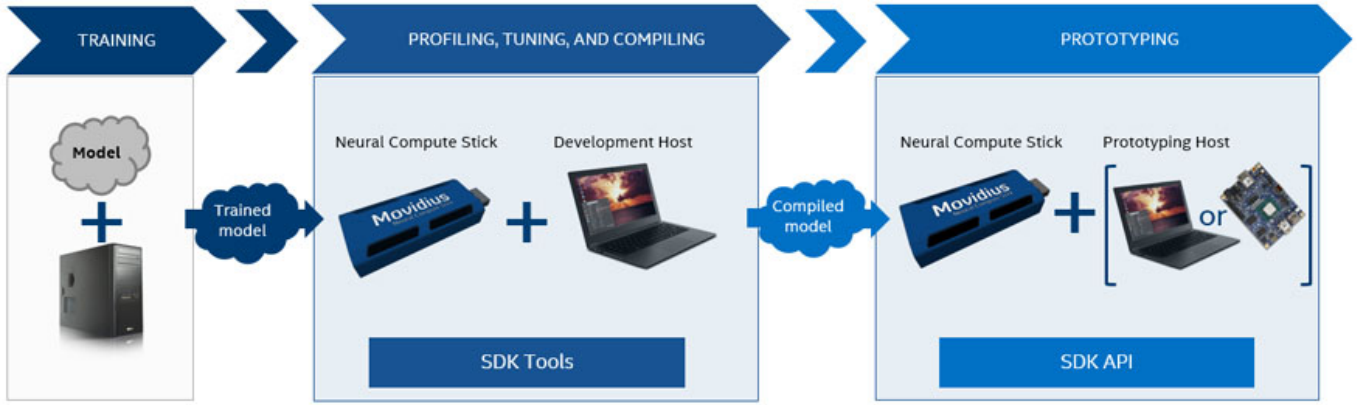
- Processor: Intel® Movidius™ Myriad™ 2 Vision Processing Unit (VPU)
- 16 processing cores
- Supported frameworks: TensorFlow and Caffe
- Connectivity: USB 3.0 Type-A
- USB stick dimensions: 2.85 in. x 1.06 in. x 0.55 in. (72.5 mm x 27 mm x 14 mm)
- Operating temperature: 0° C to 40° C
- Minimum system requirements:
  - Ubuntu 16.04.3 LTS (64 bit)
  - CentOS 7.4 (64 bit)
  - Windows 10 (64 bit)
  - USB 3.0
- 1 GB RAM
- 4 GB free storage space
- 8 times faster than NCS 1
- Software: Intel OpenVINO Toolkit


UP TO EIGHT TIMES[1] THE PERFORMANCE
Intel® NCS 2 Performance versus Intel® Movidius™ NCS
100
80
60
40
20
Image Classification
Object Detection
Intel® NCS 2
Intel® Movidius™ NCS
Disclaimer:
[1]Testing by Intel as of October 12, 2018
Deep Learning Workload Configuration: Comparing Intel® Movidius™ Neural Compute Stick based on Intel® Movidius™ Myriad™ 2 Vision Processing Unit (VPU) versus Intel® Neural Compute Stick 2 based on the Intel® Movidius™ Myriad™ X V
Processing Unit (VPU) with asynchronous plug-in enabled for two neural compute engines. As measured by images per second across GoogLeNet V1 and Tiny YOLO* V1.
Base System Configuration: Intel® Core™ i7 processor 8700K, 95 W TDP (6C12T at 3.7 GHz base frequency and 4.7 GHz maximum turbo frequency). Graphics: Intel® UHD Graphics 630, total memory 65830088 kB Storage: Intel® SSD SC2
(240 GB), Ubuntu* 16.04.5 Linux* 4.15.0-36-generic-x86_64-with-Ubuntu-16.04-xenial, deeplearning_deploymenttoolkit_2018.0.14348.0, API version 1.2, Build 14348, myriadPlugin API, FP16, Batch Size = 1
Software and workloads used in performance tests may have been optimized for performance only on Intel® microprocessors.
Performance tests, such as SYSmark* and MobileMark*, are measured using specific computer systems, components, software, operations, and functions. Any change to any of those factors may cause the results to vary. You should consi
other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit:
https://www.intel.com/content/www/us/en/benchmarks/benchmark.html.
Performance results are based on testing as of October 12, 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure.

NCS 1
vs.
NCS 2

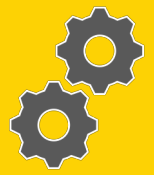
Getting Started with Intel NCS 1

- **Development Host**
  - Ubuntu 16.04 Desktop
  - NC SDK and tools
- **Inference Host**
  - Ubuntu 16.04 Desktop or Raspberry Pi 3 Model B+
  - NC SDK in API Mode



NC SDK Workflow

Train the model in the cloud



Profile & generate graph from the model

Load the graph for inference

Installing NC SDK and Tools on Dev Host

```
mkdir -p ~/workspace
cd ~/workspace
git clone -b ncsdk2 http://github.com/Movidius/ncsdk
cd ncsdk
make install
```

```
cd ~/workspace
git clone -b ncsdk2 https://github.com/movidius/ncappzoo.git
cd ncappzoo
make install
```

Optimizing Caffe Models for NCS

1. Train the Caffe model through NVIDIA DIGITS
2. Download the model
3. Generate graph
4. Profile Graph
5. Check Graph
6. Run Inference

DEMO

Optimizing Caffe Models for Inference

MI2 Sponsors

Next Webinar

Configuring Blue/Green Deployments with Istio

Istio is a service mesh designed to make communication among microservices reliable, transparent, and secure. In this webinar, I will help you understand how to configure blue/green deployment of microservices running in Kubernetes with Istio. You don't need to have any prerequisites to explore this scenario except a basic idea of deploying pods and services in Kubernetes.

Thursday, March 28th, 2019
9:00 AM PST / 9:30 PM IST

Register at http://mi2.live