Prepared by
Malcolm Chen

AI Mustang Accelerators

Mustang-V100-MX4

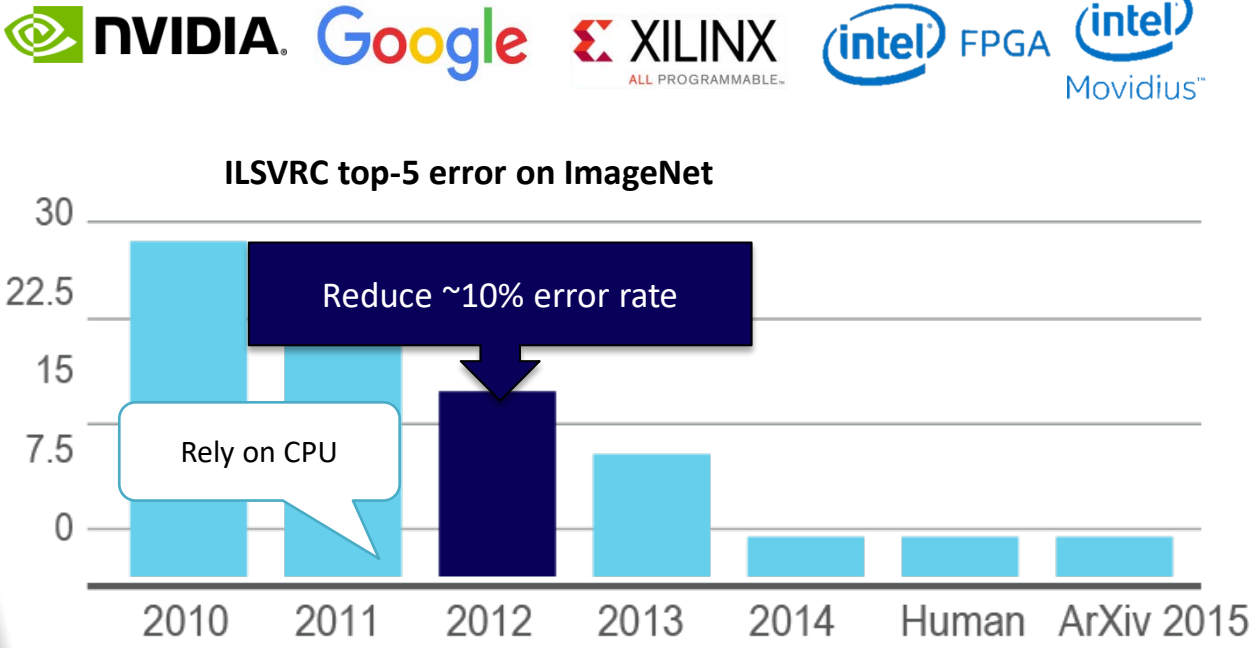
Mustang-M2AE-MX1

Mustang-M2BM-MX2

Mustang-MPCIE-MX2

Why AI Accelerator

- In 2012, AlexNet used two GPU as accelerators for training model and won the ImageNet contest. Since that, developers started to use AI Accelerators to reduce CPU workloads.

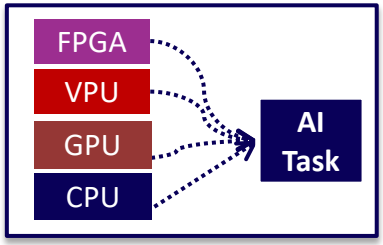


Why Intel® Solution

Advantages of Intel solution

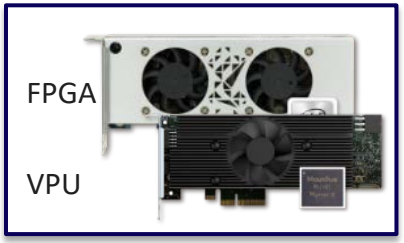
- Cross platform: easy to be implemented in the existing projects
- Heterogeneous: integrate all Intel's accelerators
- Fast integration: easy to convert a trained model to Intel platform.
- Compact size: small in size and scalable.
- Power efficiency: FPGA with 40W TDP; VPUx8 with 25W TDP.
- Better performance/$/W: Intel® solutions have better FPS/$/W, compared to other solutions

Cross platform



Heterogeneous

Fast integration



Compact size & Power Efficiency

IEI Mustang Accelerators

FPGA

Mustang-F100-A10

- Intel® Vision Accelerator Design with Intel® Arria® 10 FPGA
- Intel® Movidius™ VPU

VPU

Mustang-V100-MX8

Mustang-V100-MX4
2019/08

Mustang-M2AE-MX1
2019/09

Mustang-MPCIE-MX2
2019/07

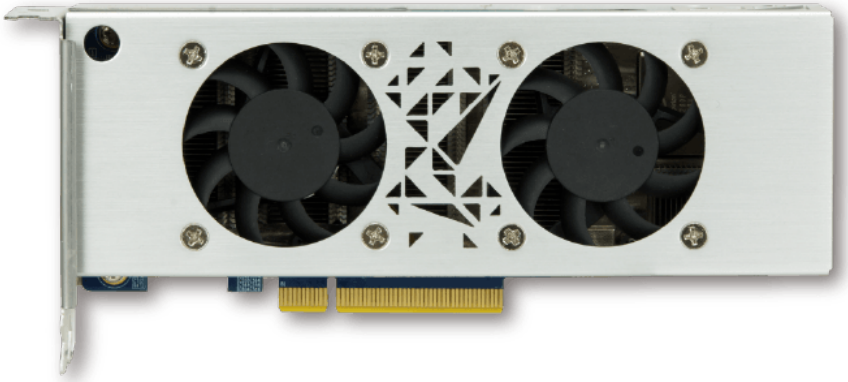
Mustang-M2BM-MX2
2019/09

Launched — **2019 Q3**

Systems for Mustang Accelerators

Accelerator / Platform	FPGA PCIe Gen3x8	VPUx8 PCIe Gen2x4	VPUx4 PCIe Gen2x2	VPUx2 minipcie	VPUx2 M.2 B+M
TANK AIoT Dev. Kit Intel® SkyLake AI Dev. Kit					
FLEX-BX-200-Q370 Intel® Coffee Lake AI Modular Box PC					
ITG-100AI Intel® Atom™ x5-E3930					

Mustang-F100-A10



Compact Size

Multiple Cards

Low Power consumption

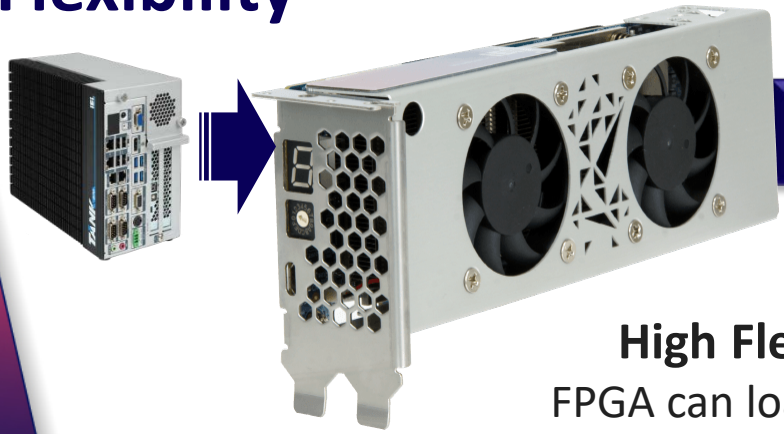
Low Latency

Accelerate To The Future

- Intel® Vision Accelerator Design with Intel® Arria® 10 FPGA
- A Perfect Choice for AI Deep Learning Inference Workloads

Mustang-F100-A10-Flexibility

Flexibility

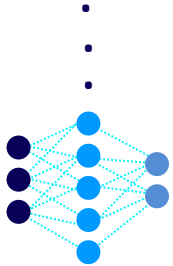


Optimized Bitstreams

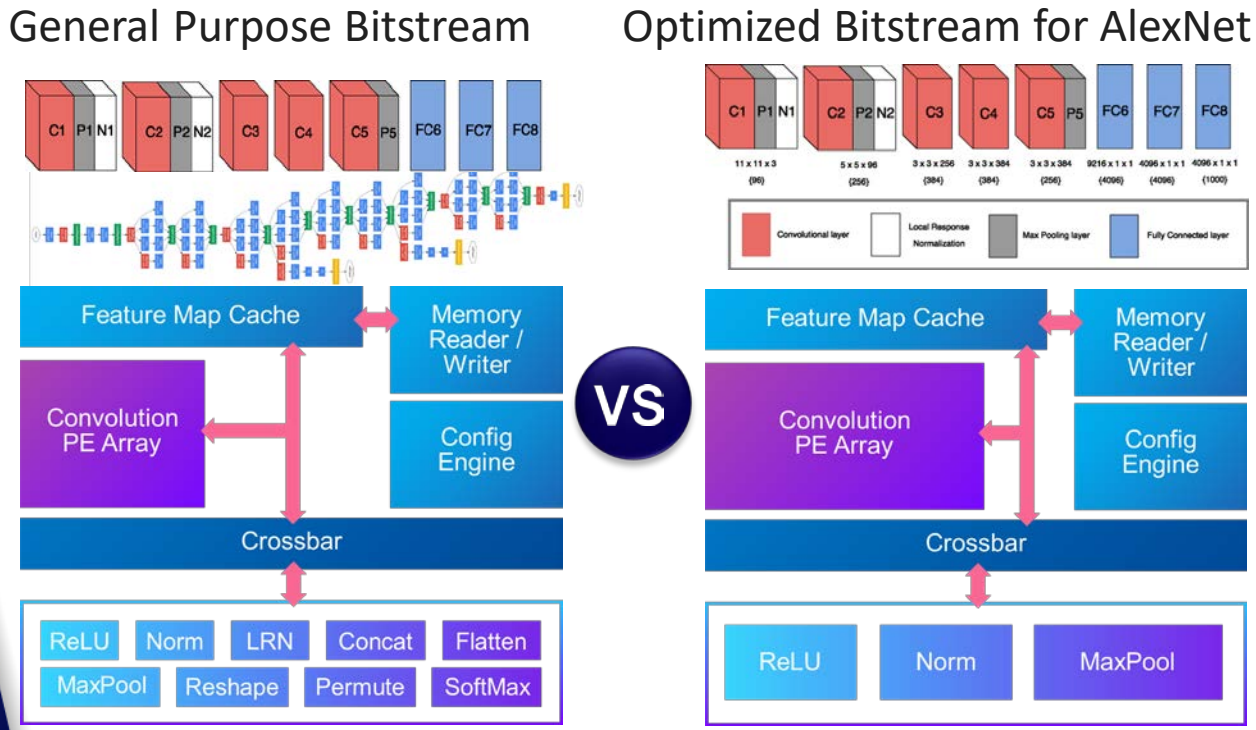
- ResNet
- GoogleNet
- AlexNet

High Flexibility

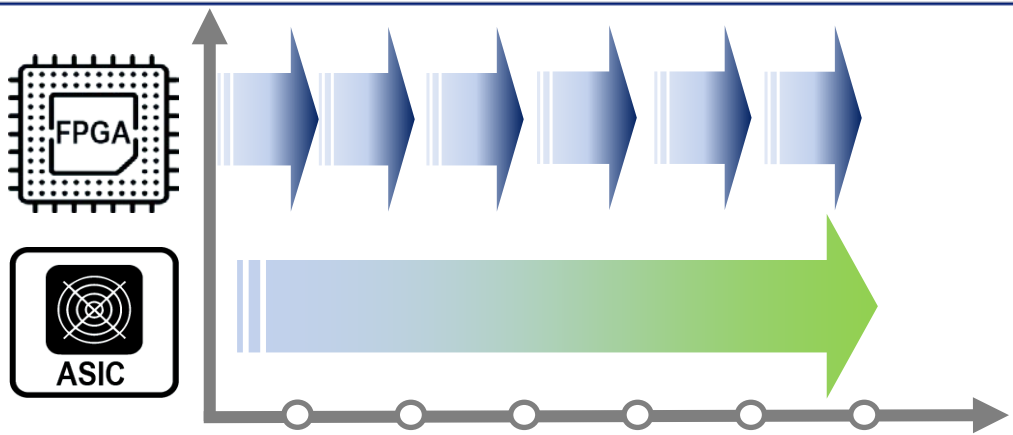
FPGA can load different bitstreams optimized for targeted topologies.



Mustang-F100-A10-Flexibility

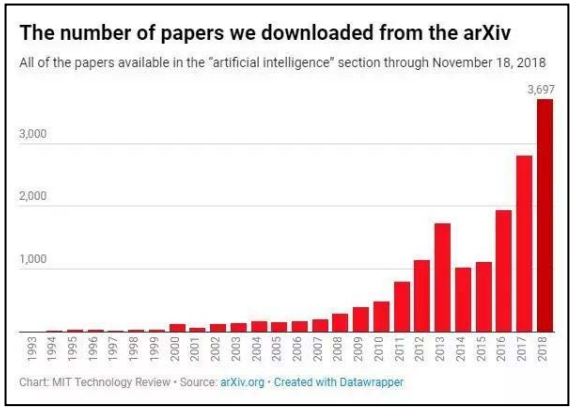


Mustang-F100-A10 Latest algorithm



Long term project supported

=> FPGA bitstreams are updated and optimized in quarterly cadence.



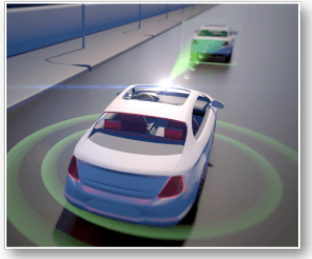
New neural network layers & topologies announced rapidly.

Application Field for FPGA

Advantages of using FPGA-based acceleration card

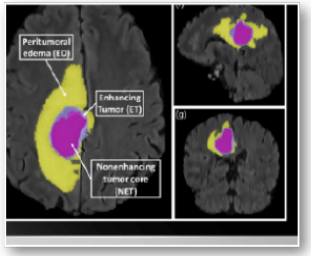
- Low latency
- Continued improved performance
- Industrial grade
- 10-year longevity
- High resolution image (>1080P)

High speed Inspection



Driving safety

Real-time monitoring



Medical image AI

Mustang-V100-MX8

Compact Size

Multiple Cards

Low Power consumption

x8
Multi-Tasks

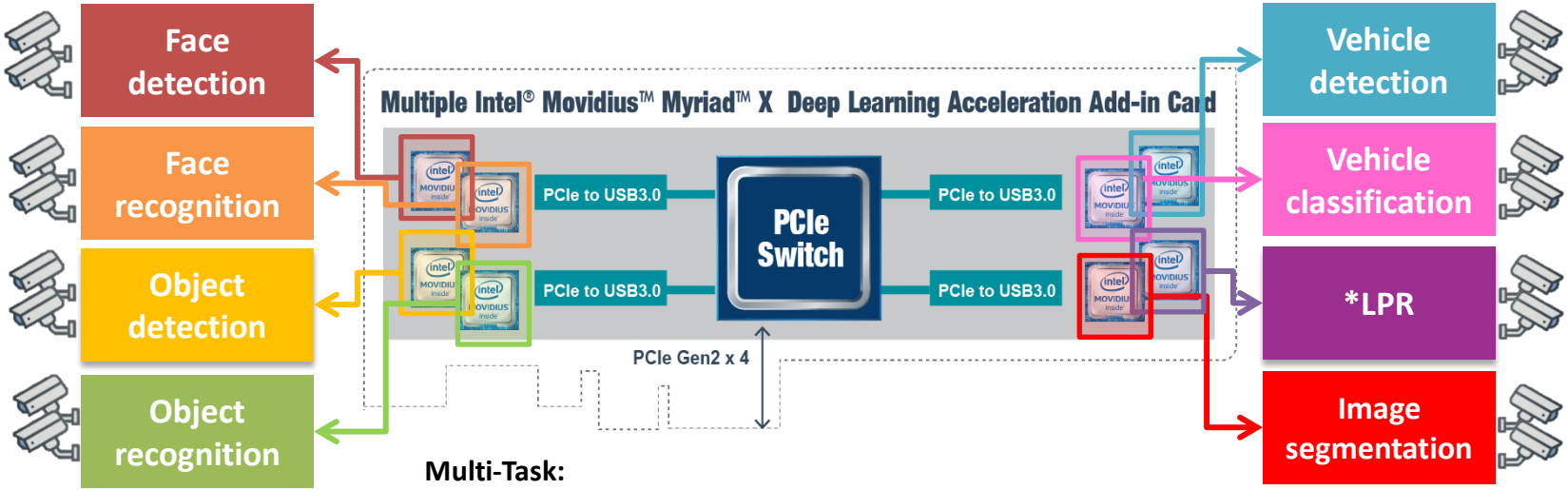
intel MOVIDIUS inside

Better Performance /$/W

Accelerate To The Future

- Intel® Vision Accelerator Design with Intel® Movidius™ VPU
- A Perfect Choice for AI Deep Learning Inference Workloads

Mustang-V100-MX8- Multi-Tasks



Multi-Task:
Can execute multiple neural networks on the same card simultaneously.

Distributed computing:
Can assign VPU to specific video stream and network to achieve guaranteed throughput.

*LPR: License plate recognition

Mustang-V100-MX4

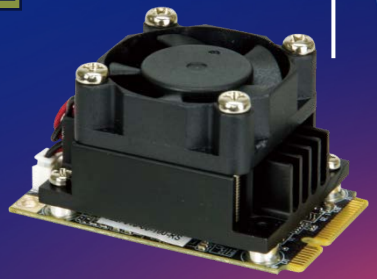
Compact Size

Multiple Cards

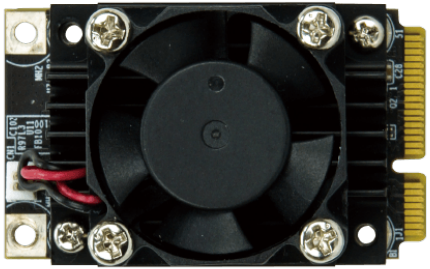
Low Power consumption

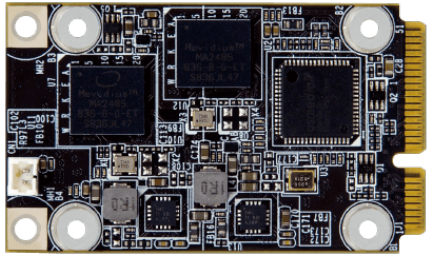
Multi-Tasks 4

Features	
Main Chip	4 x Intel® Movidius™ Myriad™ X MA2485 VPU
Operating Systems	Ubuntu 16.04.3 LTS 64bit, CentOS 7.4 64bit, Windows® 10 64bit
Dataplane Interface	PCIe Gen 2 x 2 Single slot
Power Consumption	15W
Cooling	Active Fan
Dimensions	113 x 56 x 23 mm



Mustang-MPCIE-MX2





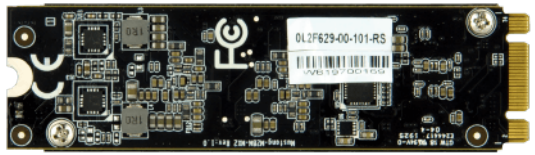
Compact Size

Low Power consumption

x 2
Multi-Tasks

Features	
Main Chip	2x Intel® Movidius™ Myriad™ X MA2485 VPU
Operating Systems	Ubuntu 16.04.3 LTS 64bit, CentOS 7.4 64bit, Windows® 10 64bit
Dataplane Interface	minipcie
Power Consumption	7.5W
Cooling	Active Fan
Dimensions	50.5 x 30 x 29.25 mm

Mustang-M2BM-MX2

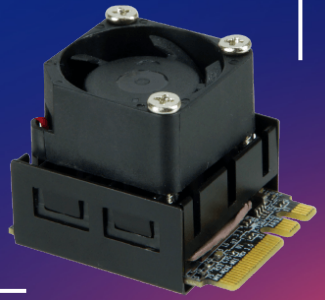


Compact Size

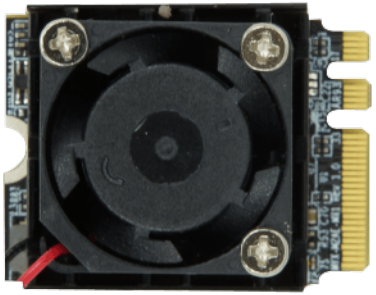
Low Power consumption

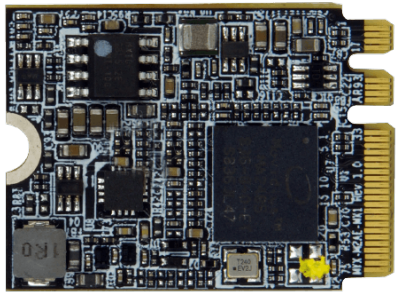
2 Multi-Tasks

Features	
Main Chip	4 x Intel® Movidius™ Myriad™ X MA2485 VPU
Operating Systems	Ubuntu 16.04.3 LTS 64bit, CentOS 7.4 64bit, Windows® 10 64bit
Dataplane Interface	M.2 B/M key 2280
Power Consumption	15W
Cooling	Active Fan
Dimensions	113 x 56 x 23 mm



Mustang-M2AE-MX1





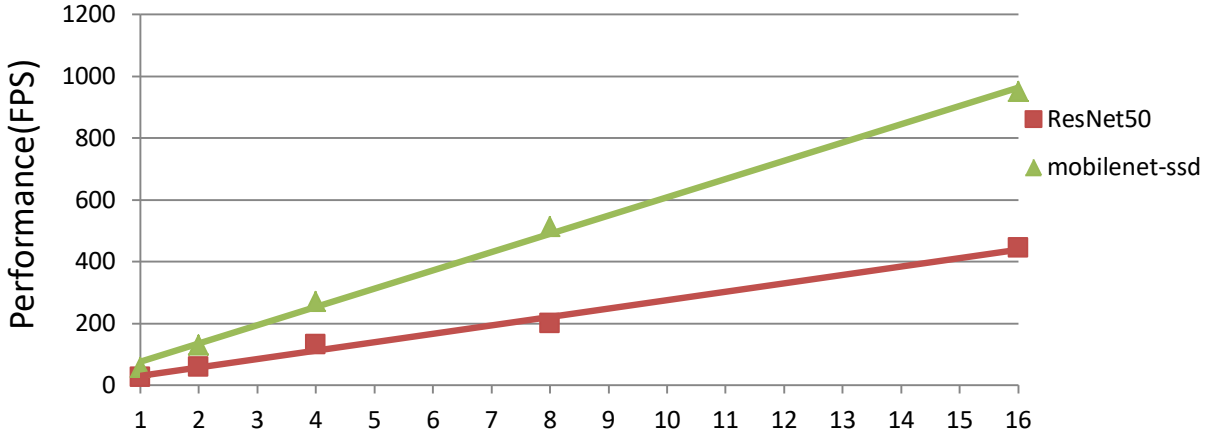
Compact Size

Low Power consumption

Features	
Main Chip	1x Intel® Movidius™ Myriad™ X MA2485 VPU
Operating Systems	Ubuntu 16.04.3 LTS 64bit, CentOS 7.4 64bit, Windows® 10 64bit
Dataplane Interface	M.2 A/E Key 2230
Power Consumption	5W
Cooling	Active Fan
Dimensions	113 x 56 x 23 mm

Mustang-Myriad-Scalability

Performance was increased by the factor of VPU quantity.



Batch Size=1 FP16

Application **Field** for VPU

Advantages of using VPU-based Acceleration card

- Low power consumption
- High scalability
- Multi-Tasks
- Compact size
- Longevity 5 years
- Medium & low resolution image(< 1080P)

Smart City

Self-check out

Digital Signage

Face recognition

Topology Support List

OpenVINO 2019 R1

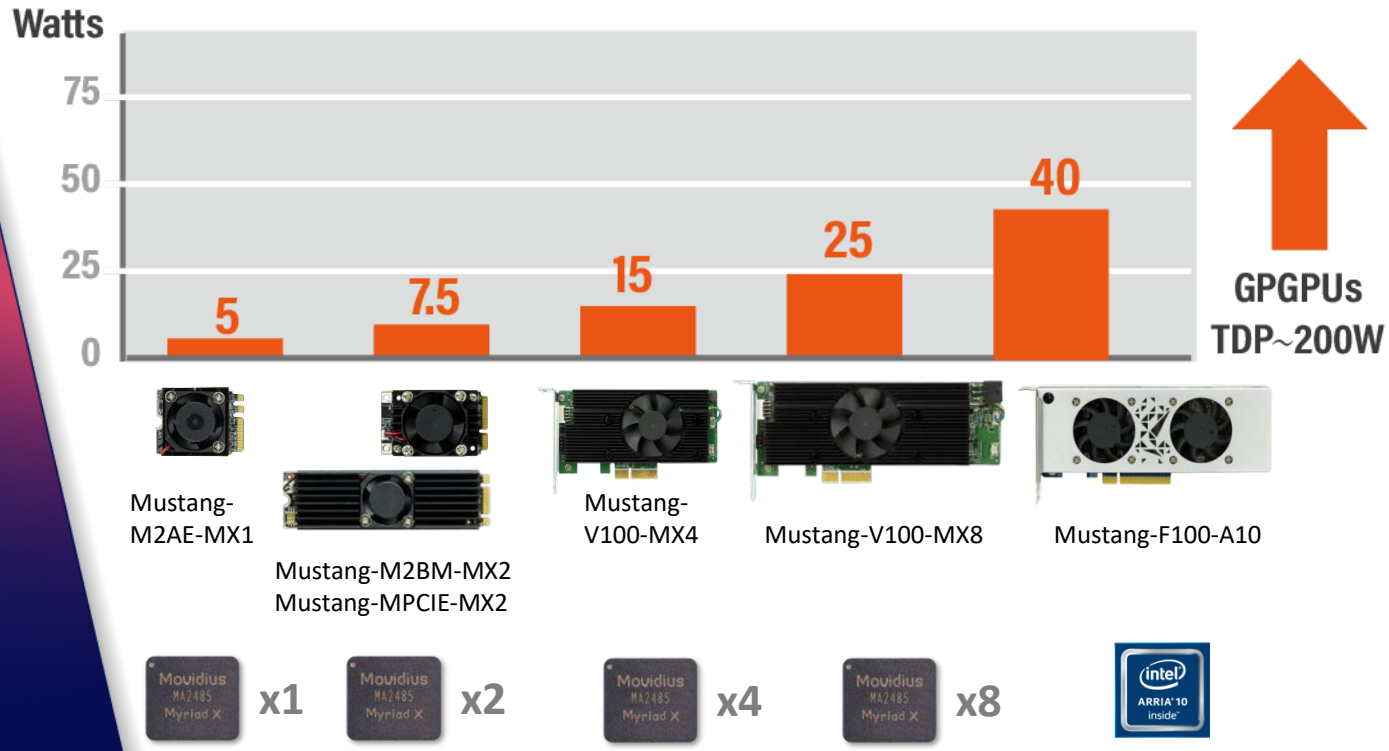
	Mustang-F100-A10	Mustang-V100-MX8
AlexNet	V	V
CaffeNet		V
DenseNet-121, -161, -169, -201	V	V
GoogLeNet v1, v2, v3, v4	V	V
Inception v1, v2, v3, v4	V	V
LSTM: CTPN	V	V
MobileNet v1, v2 ; MobileNet SSD	V	V
MTCNN-o, -p, -r	V	V
ResNet-18, -50, -101, -152; ResNet v2-50, -101, -152	V	V
ResNext-101		V
Sphereface	V	
SqueezeNet v1.0, v1.1	V	V
SSD MobileNet v1, v2	V	V
SSD GoogLeNet		
SSD Inception v2, v3	V	V
SSD ResNet	V	
SSD300, SSD512	V	V
U-Net	V	V
VGG16, VGG19	V	V
YoloTiny v1, v2, v3	V	V
Yolo v2, v3	V	V

Performance Benchmark (Perfcheck)

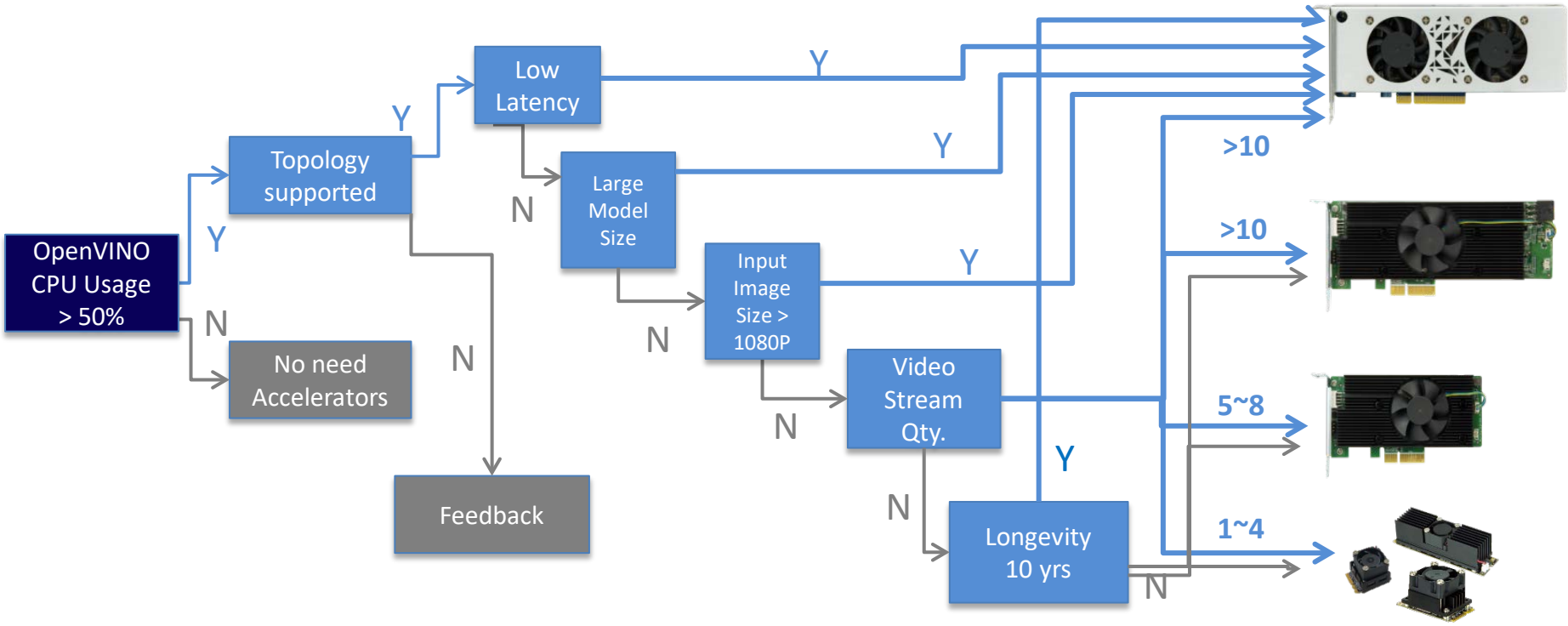
OpenVINO 2019 R1 perfcheck can help to evaluate performance.

Unit: FPS	CPU (Atom E3930)	Mustang-MPCIE-MX2	CPU(i7)	GPU(i7)	Mustang-V100-MX8	Mustang-F100-A10	Mustang-F100-A10
Floating Point	FP32	FP16	FP32	FP16	FP16	FP11	FP16
Alexnet	9.57	119.3	47.13	155.56	477.38	166.50	87.15
Googlenet/v1	5.11	182.4	79.98	85.27	729.64	575.01	173.46
Densenet/201	1.73	37.1	28.67	18.78	148.38	143.44	47.18
Inception-resnet/v2	0.61	14.2	10.66	10.91	56.64	53.43	10.24
Resnet/v1/50	2.35	60.3	40.29	49.49	241.01	271.11	64.14
Resnet/v1/101	1.16	30.0	20.01	29.25	120.04	168.55	35.87
Resnet/v1/152	0.77	19.9	13.35	20.81	79.66	119.62	24.19
Squeezenet/1.1	20.78	562.0	327.86	252.21	2247.99	2112.33	730.74
VGG19	0.46	17.2	6.51	18.95	68.67	48.58	14.06
Mobilenet-SSD	6.59	114.5	109.34	65.45	458.12	329.52	138.88
SSD512	-	2.5	1.63	2.82	9.81	12.89	3.70

Power Budget

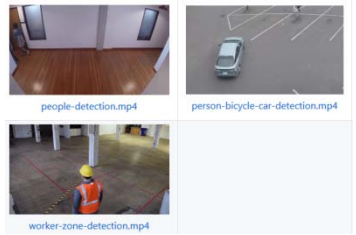


Mustang AI Accelerators Decision Tree



What OpenVINO offers

- **Inference engine samples**
  https://iei.pse.is/KW2D3



- **Pre-trained models**
  https://iei.pse.is/JJZ7C

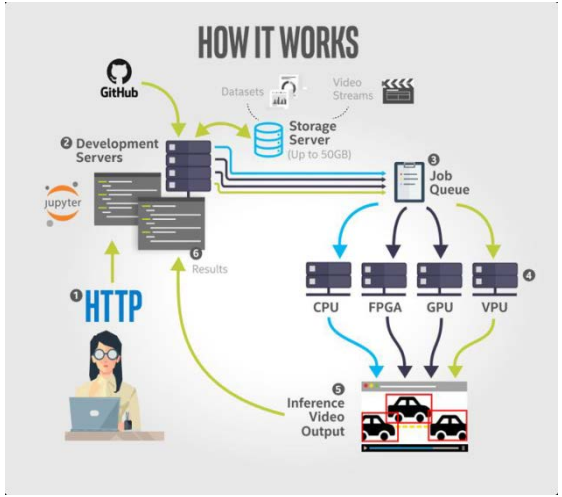
- **Sample videos**
  https://iei.pse.is/KLD43

What Intel® offers

IoT DevCloud

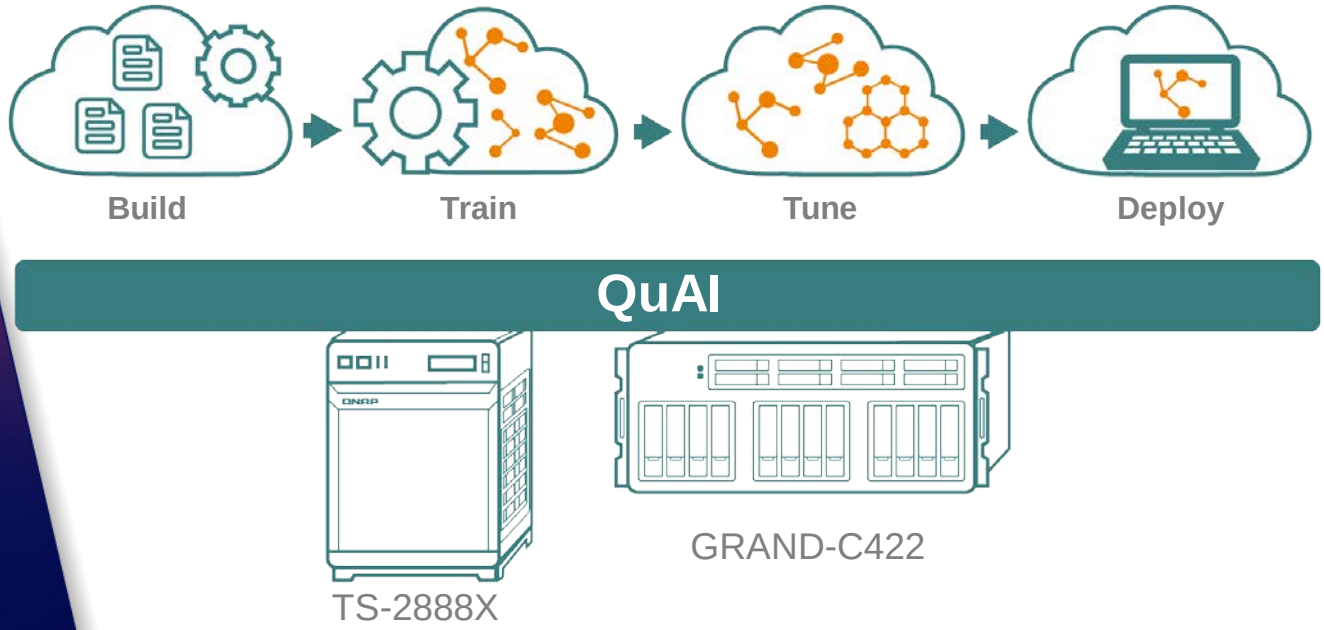
Develop your computer vision applications using the Intel® DevCloud, which includes a preinstalled and preconfigured version of the Intel® Distribution of OpenVINO™ toolkit.

Access reference implementations and pretrained models to help explore real-world workloads and hardware acceleration solutions.



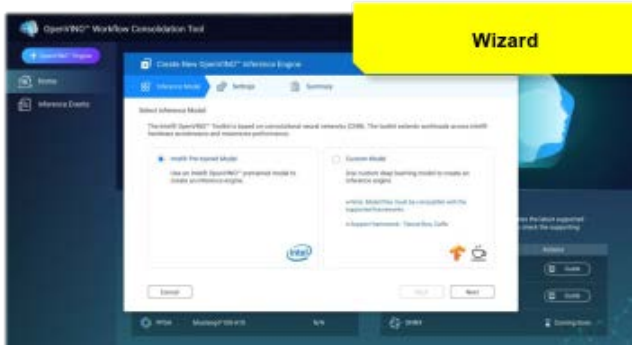
https://software.intel.com/en-us/devcloud/edge

What IEI x QNAP offers



What IEI x QNAP offers

OWCT: OpenVINO Workflow Consolidation Tool



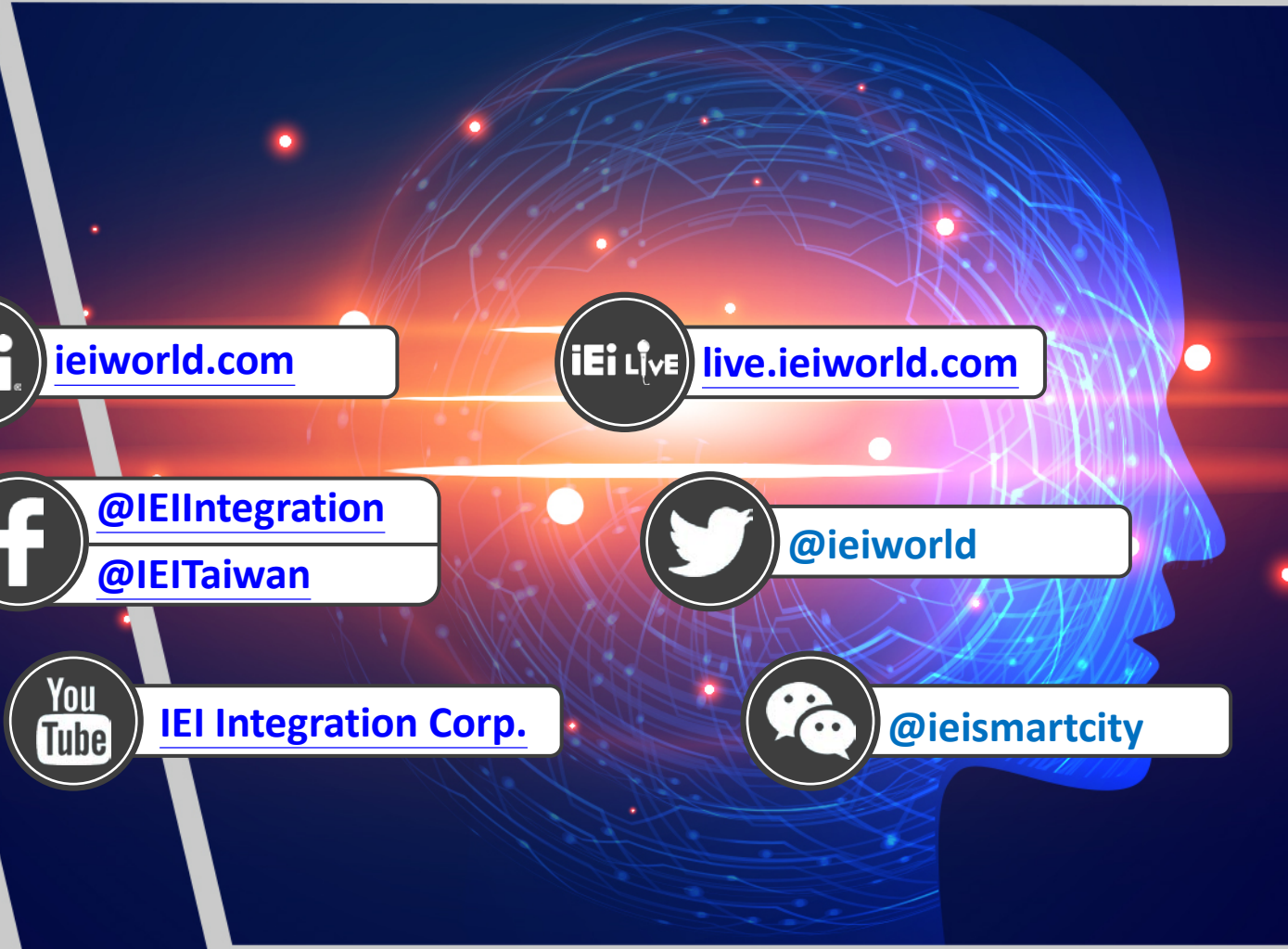
AI Model Upload

Model Optimize

Convert Format

Inference Engine

Follow us


ieiworld.com
live.ieiworld.com
@IEIIntegration
@IEITaiwan
@ieiworld
IEI Integration Corp.
@ieismartcity