AIDC

INTEL AI DEVCON 2018

AIDC
INTEL AI DEVCON 2018

DEEP LEARNING IN RESOURCE-CONSTRAINED ENVIRONMENTS

Dave Ojika, PhD
Bhavesh Patel, Dell EMC AI Architect

Intel Collaborators:
Shawn Slockers, Chris Branch

intel AI


ASPLOS2018 Thanks our Sponsors

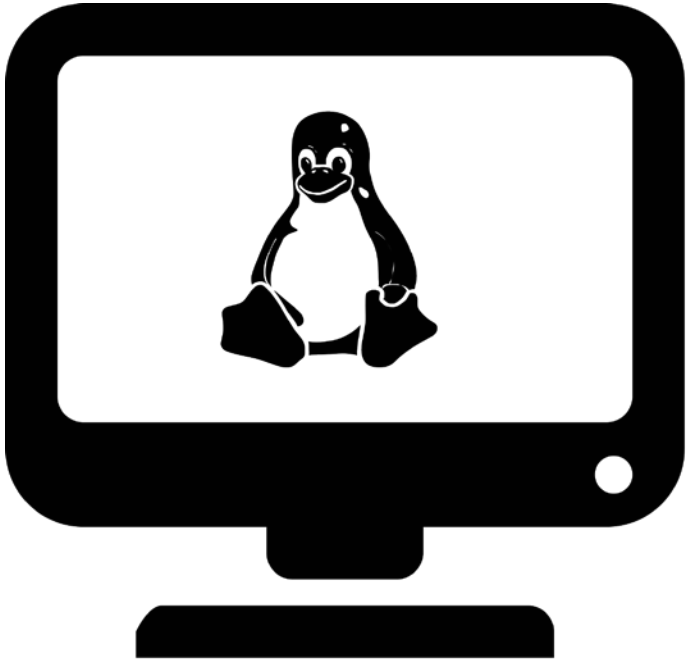
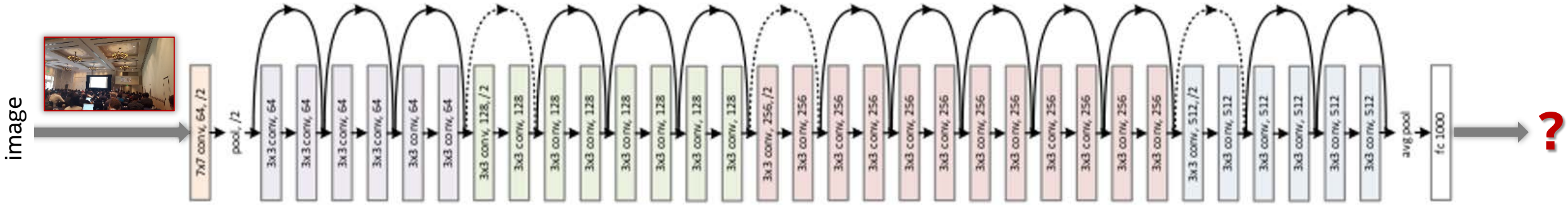


IMAGE CLASSIFICATION WITH CNN



ResNet architecture

		Input	Output	Layers	Operations
CNN	**AlexNet**	150,528	1,000	4	61 millions
	ResNet-50	150,528	1,000	50	3.8 billions

CNN PREDICTIONS


```
Image ../../../asplos.jpeg

819 0.2742987 label #819
762 0.0785879 label #762
498 0.0540126 label #498
406 0.0476660 label #406
887 0.0420651 label #887
892 0.0371223 label #892
632 0.0289109 label #632
851 0.0225158 label #851
624 0.0175353 label #624
818 0.0154749 label #818
```

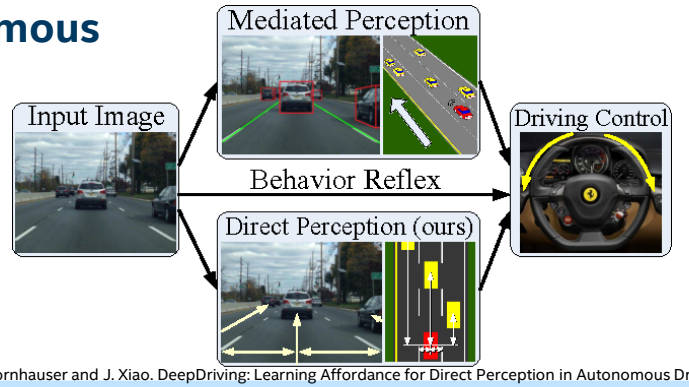
CNN result

```
816   n04275548 spider web, spider's web
817   n04277352 spindle
818   n04285008 sports car, sport car
819   n04286575 spotlight, spot
820   n04296562 stage
821   n04310018 steam locomotive
822   n04311004 steel arch bridge
```

synset_words.txt

CNN IN MISSION-CRITICAL APPLICATIONS

Autonomous vehicle



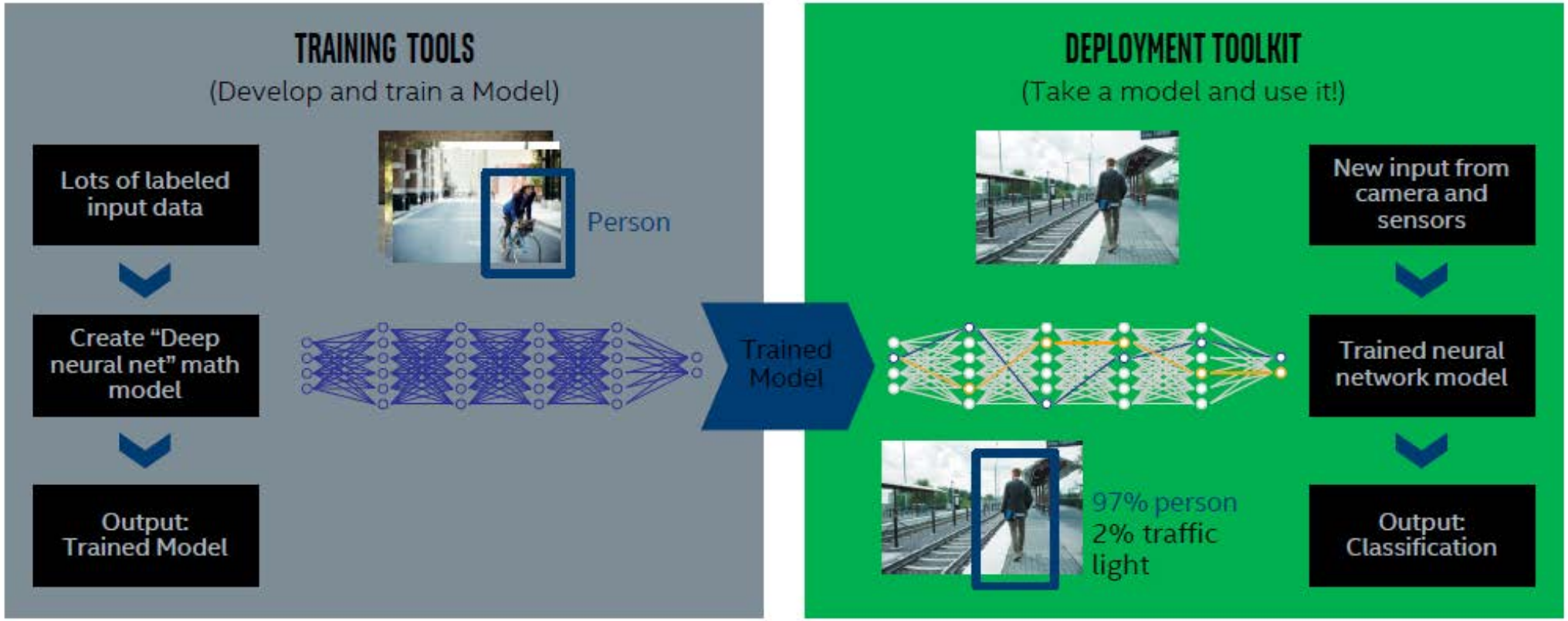
C. Chen, A. Seff, A. Kornhauser and J. Xiao. DeepDriving: Learning Affordance for Direct Perception in Autonomous Driving, ICCV2015.

Manufacturing

IEEE Computer

1. Latency: streaming at batch size of 1
2. Power efficiency: particularly in embedded solutions
3. Developer productivity: emergence of new hardware

CNN OVERVIEW



POPULAR CNN MODELS

method	top-1 err.	top-5 err.
VGG [41] (ILSVRC'14)	-	8.43†
GoogLeNet [44] (ILSVRC'14)	-	7.89
VGG [41] (v5)	24.4	7.1
PReLU-net [13]	21.59	5.71
BN-inception [16]	21.99	5.81
ResNet-34 B	21.84	5.71
ResNet-34 C	21.53	5.60
ResNet-50	20.74	5.25
ResNet-101	19.87	4.60
ResNet-152	**19.38**	**4.49**

ACCELERATING INFERENCE

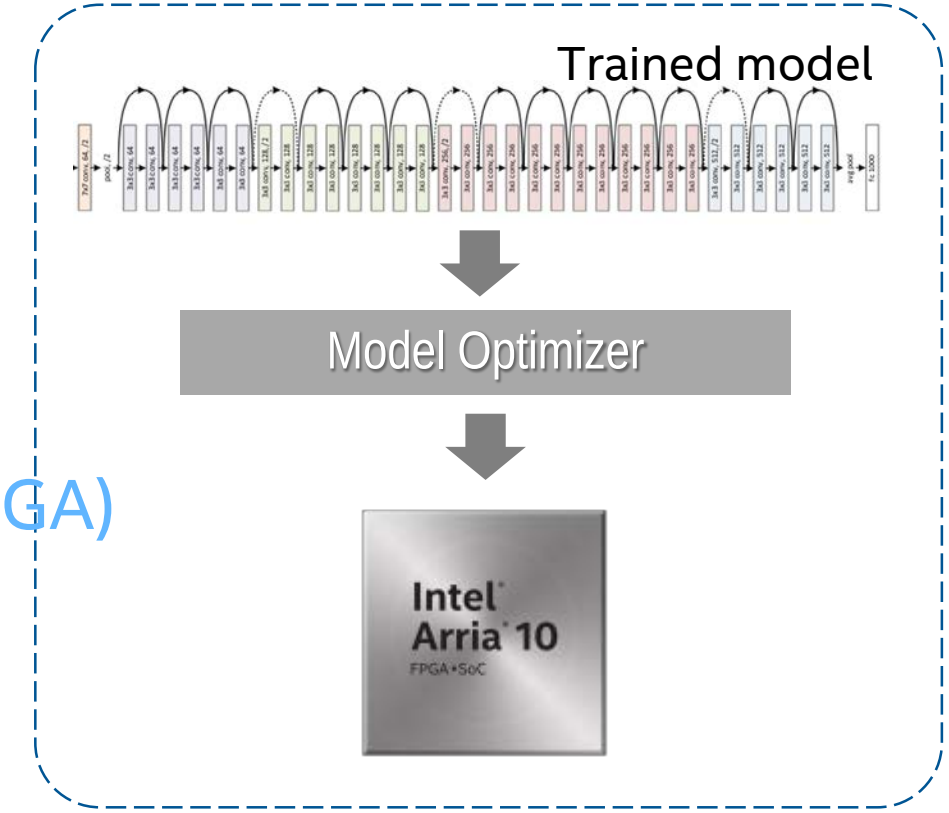
Goal

A. Flexibly import model

B. Optimize FP computation

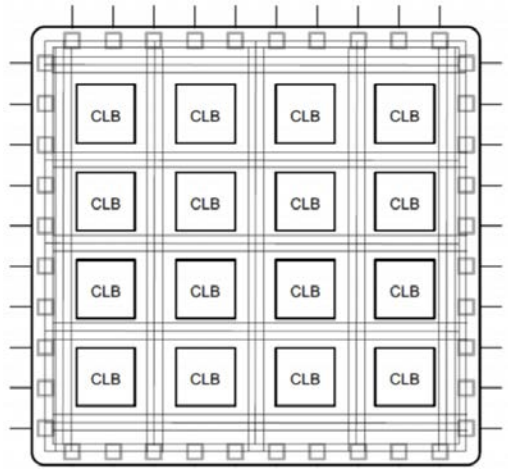
C. Deterministic low-latency (with Intel FPGA)

D. Increased developer-productivity
(with Intel OpenVINO)

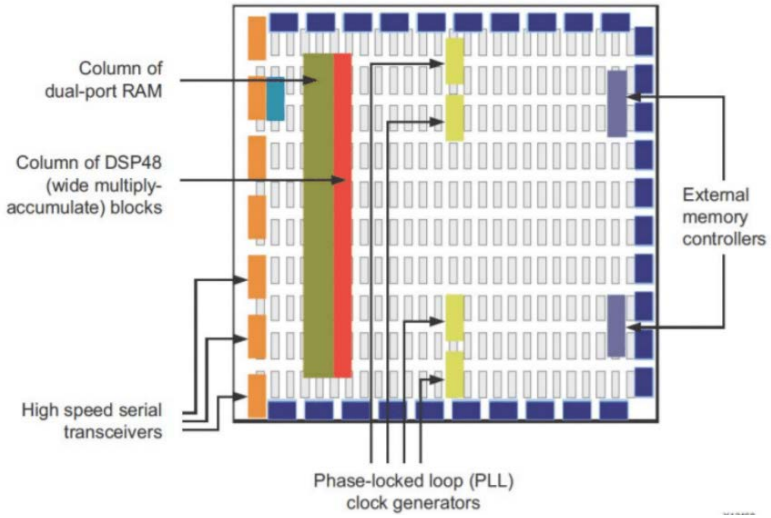


FIELD-PROGRAMMABLE GATE ARRAY (FPGA)

Basic



CLB: configurable logic block



Contemporary

Benefits:

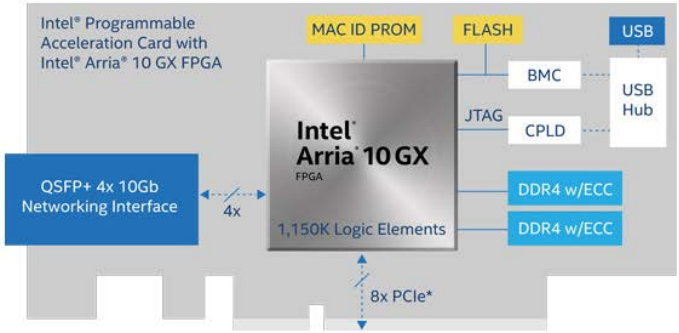
- Flexibility
- High-throughput
- Low-latency
- Power-efficient

Why FPGA for CNN

- 1.5 TFLOPs floating-point performance (on A10)
- 8 TB/s on-chip memory bandwidth

1. INTEL PROGRAMMABLE ACCELERATOR CARD

Programmable Accelerator Card (PAC) with Arria-10



- Low-power (50 W) add-in card
- Accelerates image recognition CNNs
- Configurable with reduced-precision arithmetic (FP8 … FP11)

2. SOFTWARE STACK

Dynamically Allocate Intel® FPGAs for Workload Optimization

Rack-Level Solutions

Simplified Application Development

User Applications

Leverage Common Frameworks

Industry Standard Software Frameworks

Fast-Track Your Performance

Acceleration Libraries

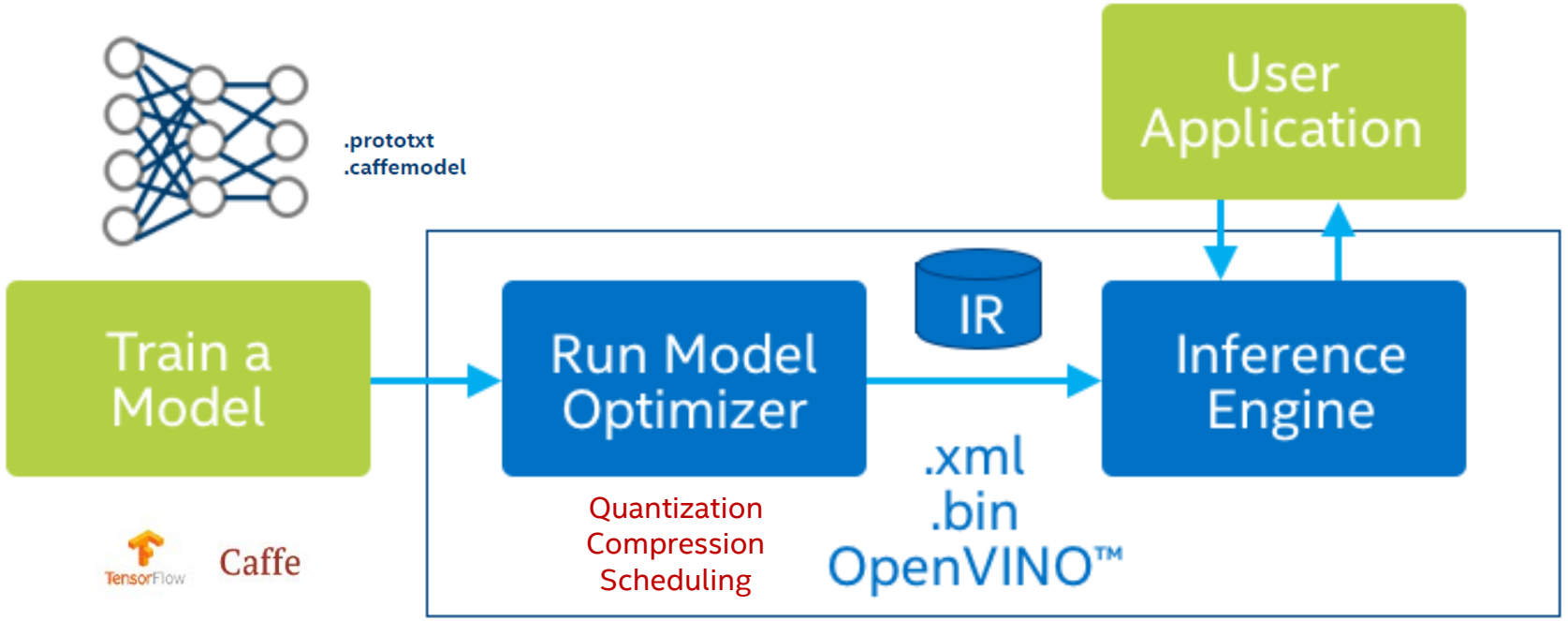
Workload Optimization with Less Effort

Intel® Developer Tools
(Intel® Parallel Studio XE, Intel® FPGA SDK for OpenCL™, Intel Quartus® Prime software)

Common Developer Interface for Intel® FPGA Data Center Products

Acceleration Environment
(Intel® Acceleration Engine with OPAE Technology, FPGA Interface Manager (FIM))

3. DEVELOPMENT TOOLKIT - OPENVINO



CNN INFERENCE: COMPLETE SOLUTION

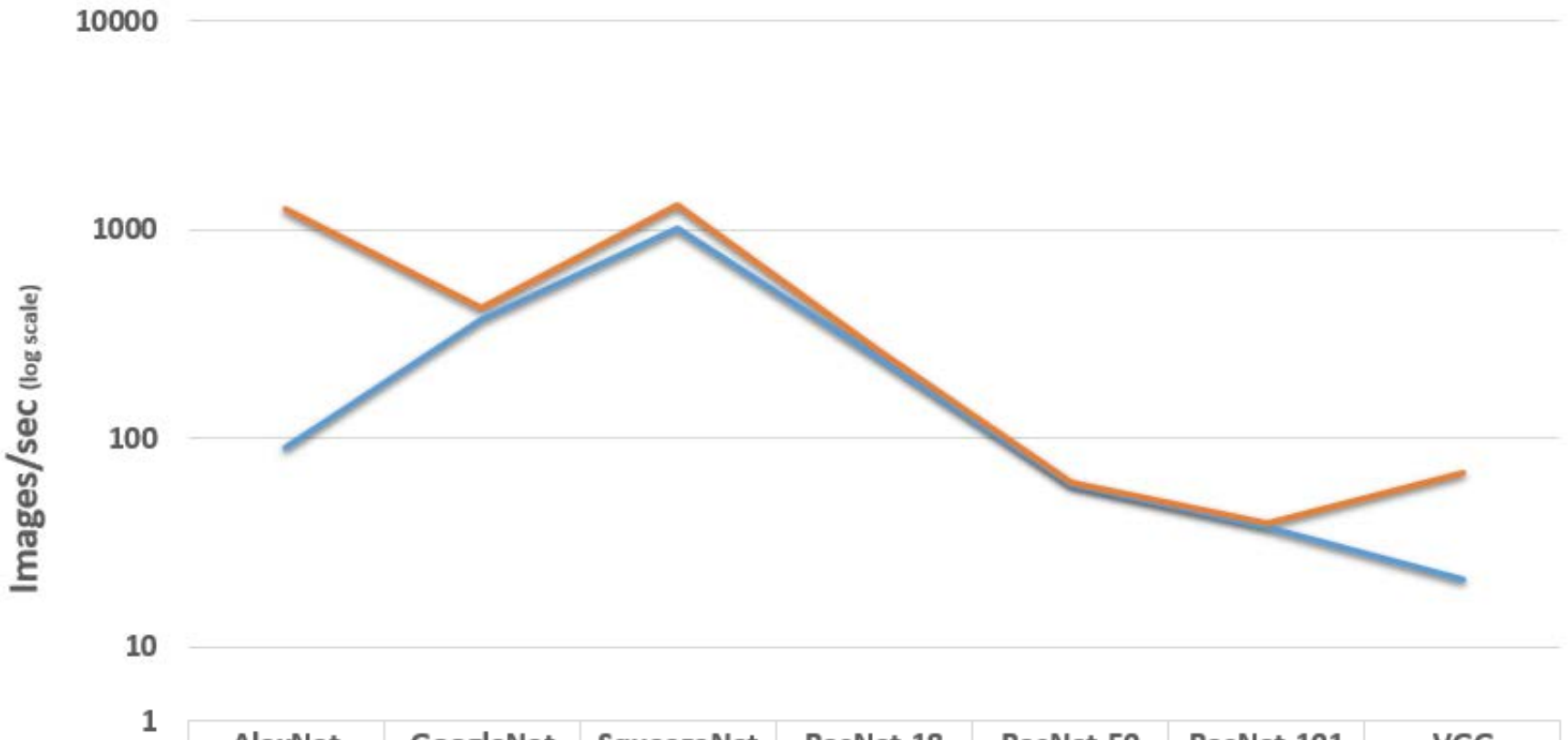
1. PAC hardware
2. Acceleration Stack software
3. OpenVINO developer toolkit

System Configuration	
Server	PowerEdge R740xd with Intel PAC
CPU	2-socket Xeon Gold CPU @ 2.1GHz
RAM	196 GB
SSD	600 GB

OPENVINO AT WORK BY UF GRAD STUDENTS

Shreyas and Nirali. Thanks to Intel and Dell

INFERENCE PERFORMANCE WITH FP 11 PRECISION

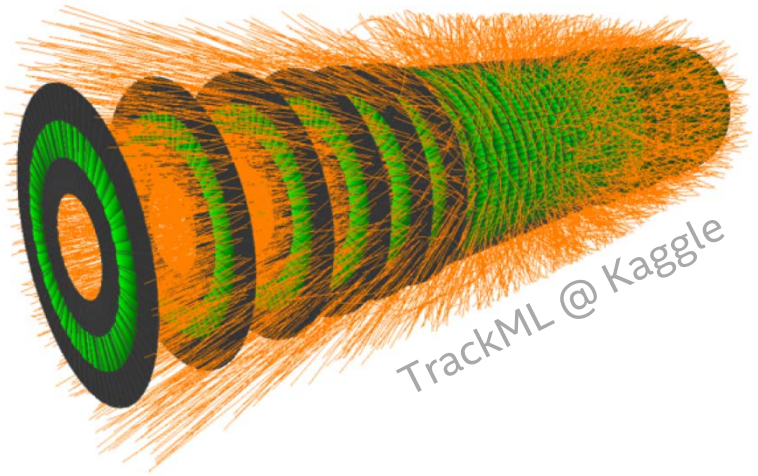


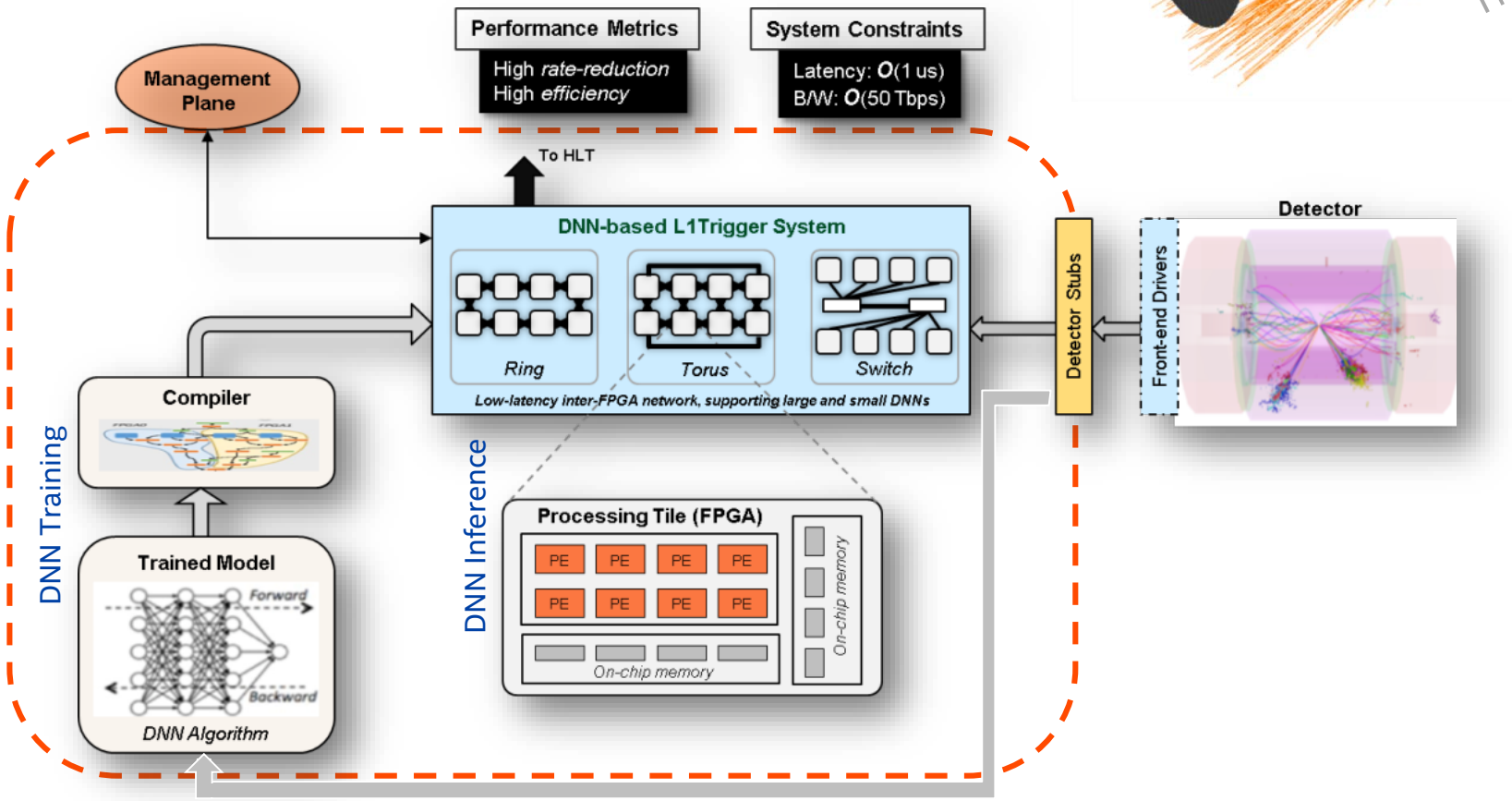
	AlexNet	GoogleNet	SqueezeNet	ResNet-18	ResNet-50	ResNet-101	VGG
Batch 1	90	369	1024	255	60	38	21
Batch 64	1254	426	1330	277	61	39	68

LATENCY (TURN-AROUND TIME) WITH FP11 PRECISION

CNN IN HEP

Fast particle-prediction with deep learnir





WIN AWS DEEPLENS

AIDEVCONAPP.INTEL.COM

Completing a session evaluation in the mobile app by 10:00 a.m. tomorrow automatically enters you in a drawing to win.

Copies of the complete sweepstakes rules are available at the Concierge Desks.

AIDC

INTEL AI DEVCON 2018