



# Deep Learning Inference with FPGAs

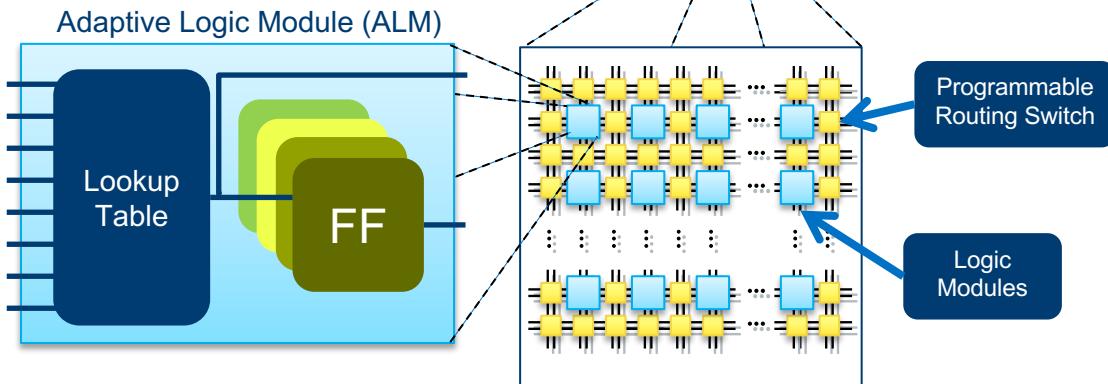
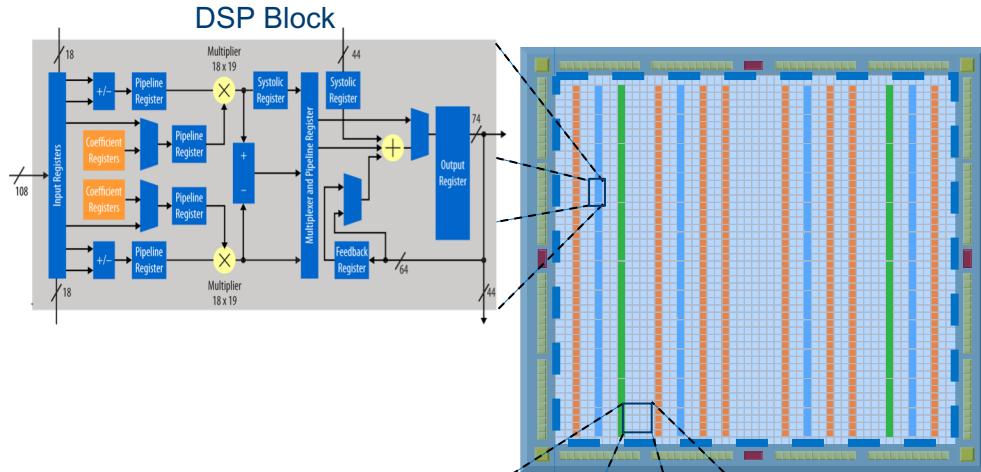
# FPGA Architecture

- **Massive Parallelism**

- Millions of logic elements
- Thousands of embedded memory blocks
- Thousands of Variable Precision DSP blocks
- Programmable routing
- Dozens of High-speed transceivers
- Various built-in hardened IP

- **FPGA Advantages**

- **Custom hardware!**
- Efficient processing
- Low power
- Ability to reconfigure
- Fast time-to-market



# Solving Machine Learning Challenges with FPGA



## EASE-OF-USE

SOFTWARE ABSTRACTION,  
PLATFORMS & LIBRARIES

*Intel FPGA solutions enable software-defined programming of customized machine learning accelerator libraries.*

## REAL-TIME

DETERMINISTIC  
LOW LATENCY

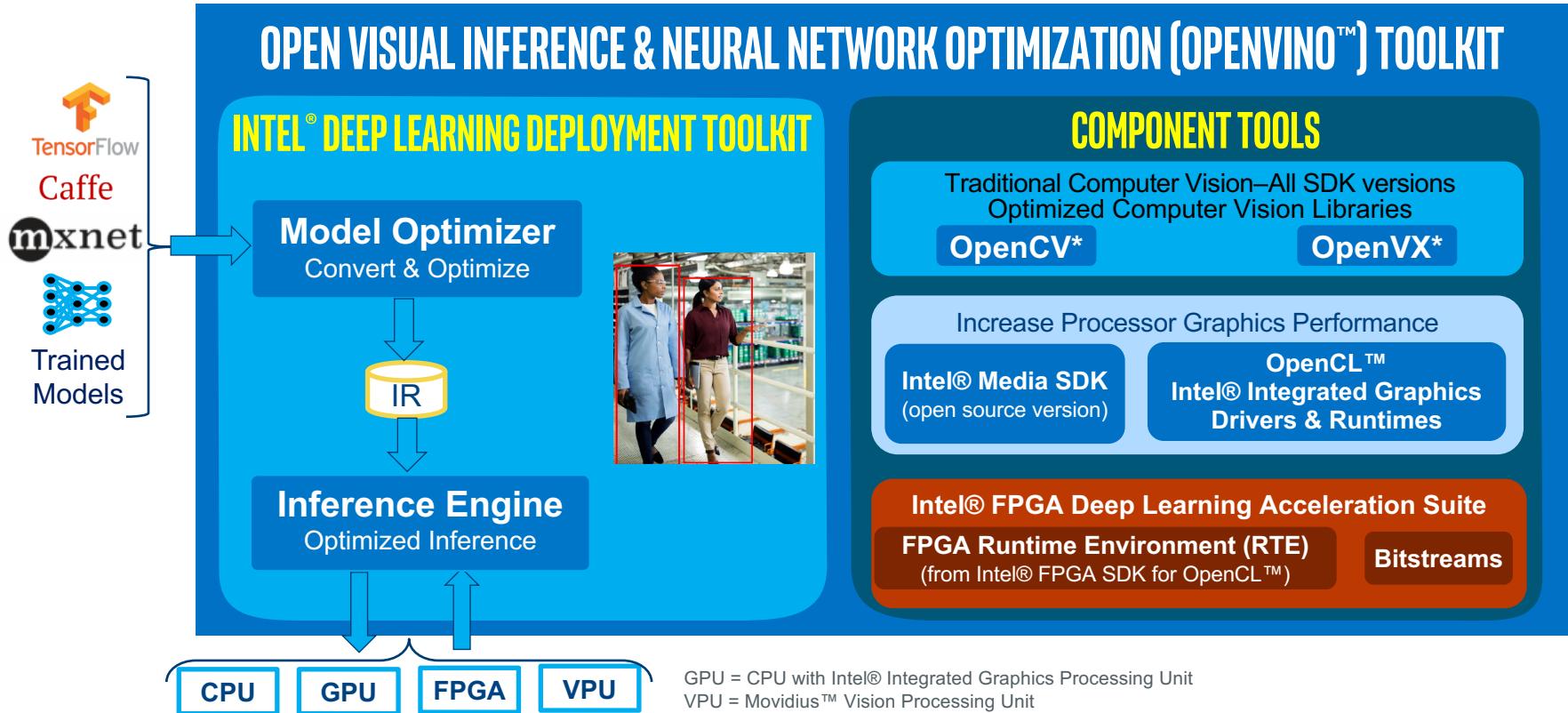
*Intel FPGA hardware implements a deterministic low-latency data path unlike any other competing compute device.*

## FLEXIBILITY

CUSTOMIZABLE HARDWARE  
FOR NEXT GEN DNN ARCHITECTURES

*Intel FPGAs can be customized to enable advances in machine learning algorithms.*

# OpenVINO™ Toolkit



GPU = CPU with Intel® Integrated Graphics Processing Unit

VPU = Movidius™ Vision Processing Unit

# OpenVINO™ with DLA User Flows

Application Developer



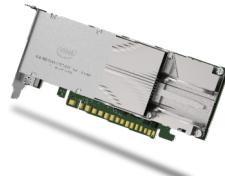
Design

OpenVINO™ toolkit

Run



Program



Turnkey Software Deployment Flow

Neural Net



Choose from many  
**precompiled** FPGA Images

Or

Custom create  
FPGA bitstream

Bitstream Library

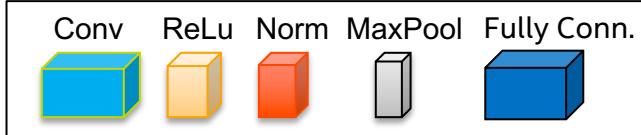
Design

User Customization  
of DLA Suite  
Source Code

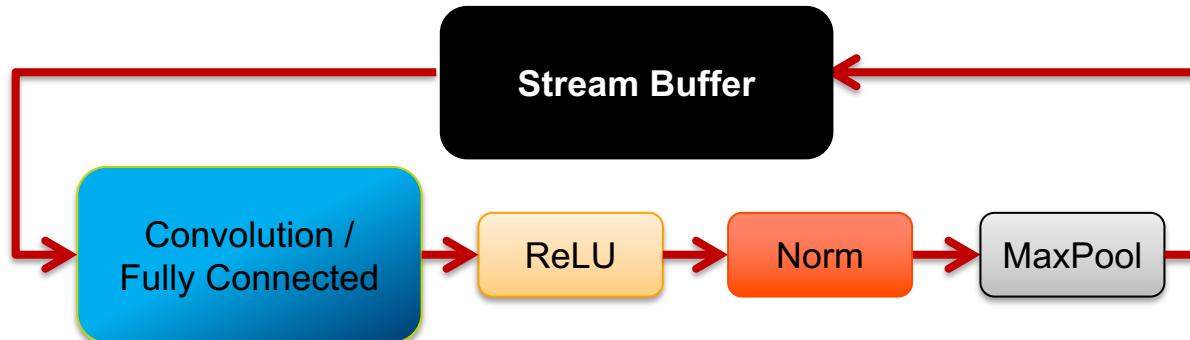
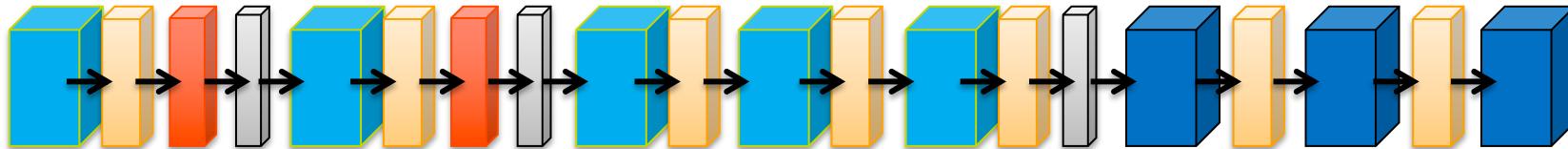
Compile

**Intel® FPGA SDK  
for OpenCL™**

# Mapping Graphs in DLA



AlexNet Graph



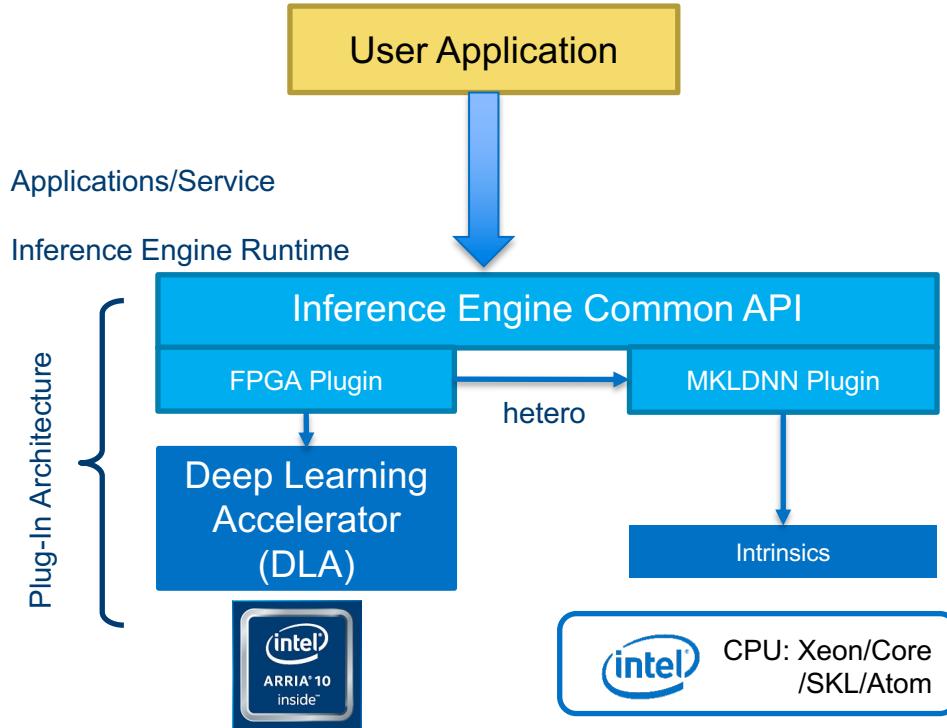
Blocks are run-time reconfigurable and bypassable

# DLA Architecture Selection

- Find ideal FPGA image (arch) that meets your needs
  - Right combination of data type support, primitives support, and parallelism
- Optional: Create custom FPGA image based on need

Network	Bitstreams (Programmable Acceleration Card with Intel® Arria® 10 GX FPGA)
AlexNet	arch2, arch3, arch10, arch11, arch12, arch16, arch17, arch23, arch25, arch26
GoogleNet v1	arch2, arch3, arch11, arch12, arch16, arch17, arch23, arch25, arch26
VGG-16, VGG-19	arch2, arch3, arch16
SqueezeNet v1.0, v1.1	arch2, arch3, arch9, arch11, arch12, arch16, arch17, arch18, arch20, arch23, arch25, arch26
ResNet-18	arch2, arch3, arch9, arch16, arch20
ResNet-50	arch3, arch9, arch20
ResNet-101	arch3, arch9, arch20

# Software stack



FPGA implementation is “Deep Learning Accelerator” (DLA)

Additional step: FPGA RTE (aocl) loads bitstream with desired version of DLA to FPGA before running

# FPGA Supported Primitives

- The following layers are supported by the DLA plugin:

✓ Batch norm	✓ Concat	✓ Convolution
✓ Fully Connected	✓ ReLU, leaky ReLU	✓ LRN normalization
✓ Pooling	✓ ScaleShift	✓ Power
✓ Eltwise	✓ prelu	

- Heterogeneous execution

- In case when topology contains layers not supported on FPGA, you need to use Heterogeneous plugin with dedicated fallback device.

# Automatic Fallback with Hetero Plugin

```
$ classification_sample -d HETERO:FPGA,CPU ...
```

```
InferenceEngine::InferenceEnginePluginPtr enginePtr;  
enginePtr = dispatcher.getPluginByDevice("HETERO:FPGA,CPU");
```

- The “priorities” define search order
- Keeps all layers that can be executed on the device (FPGA)
- Carefully respecting the topological and other limitations
- Then follows priorities when searching ( e.g. CPU)

# Prepare FPGA Environment for Inference

- Intel® FPGA Runtime Environment for OpenCL™
- Prepare FPGA Board for OpenCL
- Set environment
  - Use script to ensure DLA and OpenCL libraries part of LD\_LIBRARY\_PATH

```
[xkqi@centos-z620 ~]$ aocl diagnose
aocl diagnose: Running diagnose from /opt/intelFPGA_pro/17.0/hld/board/a10_ref/linux64/libexec

----- acl0 -----
Vendor: Intel(R) Corporation

Phys Dev Name  Status   Information

acla10_ref0  Passed   Arria 10 Reference Platform (acla10_ref0)
               PCIe dev_id = 2494, bus:slot.func = 04:00.00, Gen3 x8
               FPGA temperature = 68.5547 degrees C.

DIAGNOSTIC_PASSED
-----
```

# Load FPGA Image and Execute IE Application

- FPGAs need to be preconfigured with primitives prior to application execution
- Choose FPGA bitstream from the DLA suite
  - Based on topology needs and data type requirements
  - Option to create custom FPGA bitstream based on requirements

```
-bash-4.2$ aocl program acl0 $DLA_AOCX
aocl program: Running program from /opt/altera/aocl-pro-rte/aclrte-linux64/board/a10_ref/linux64/libexec
Programming device: a10gx : Arria 10 Reference Platform (acla10_ref0)
Reprogramming device [0] with handle 1
Program succeed.
```

- Execute User or Example Application

```
-bash-4.2$ classification_sample -d HETERO:FPGA,CPU -i car.png -m ir/squeezeenet1.1/squeezeenetc1.1.xml
```

# Summary

- FPGAs provide a flexible, deterministic low-latency, high-throughput, and energy-efficient solution for accelerating AI applications
- Intel® FPGA DLA Suite supports CNN inference on FPGAs
- Accessed through OpenVINO™ toolkit
- Available for Intel® Programmable Acceleration Card

