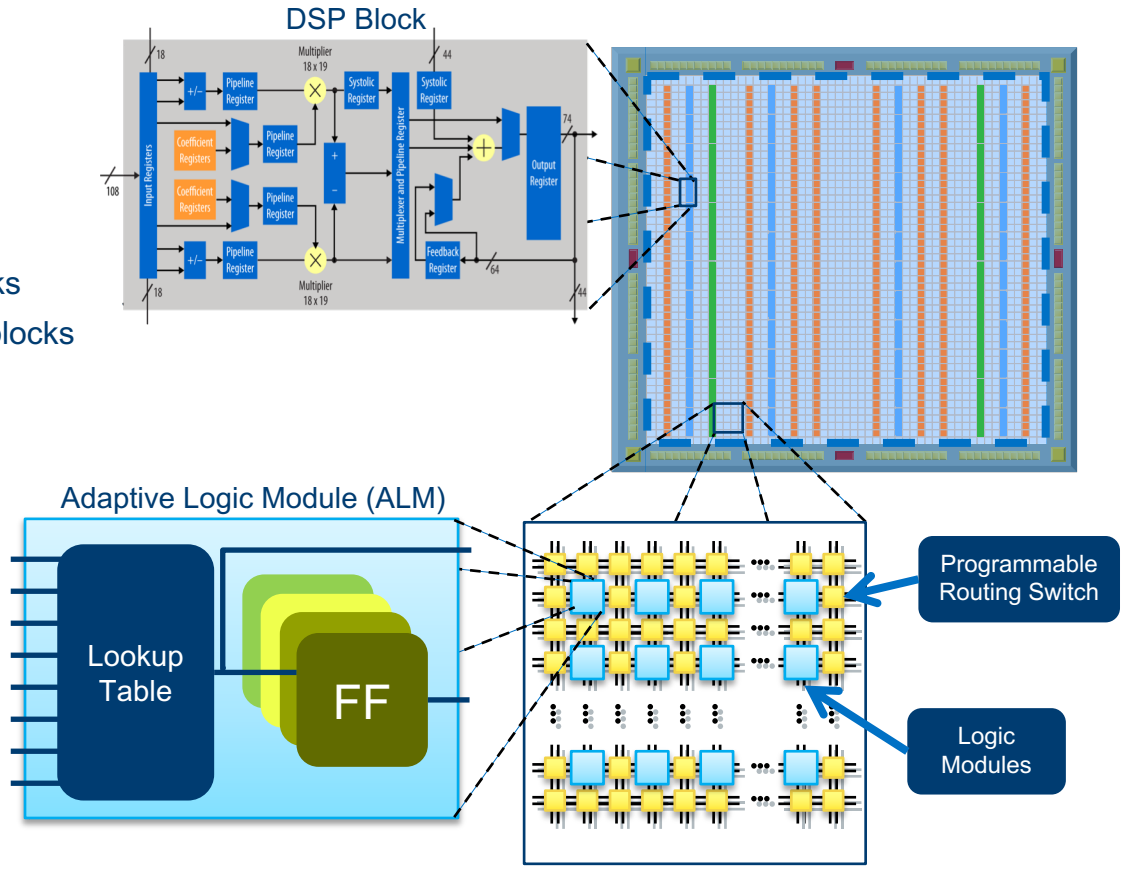
Deep Learning Inference with FPGAs

November 2018

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.

What's Inside Intel® Distribution of OpenVINO™ toolkit

Intel® Deep Learning Deployment Toolkit

Model Optimizer
Convert & Optimize

IR

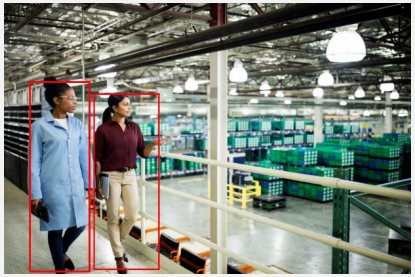
Inference Engine
Optimized Inference

20+ Pre-trained Models

Computer Vision Algorithms

Samples

IR = Intermediate Representation file



Traditional Computer Vision

Optimized Libraries & Code Samples

OpenCV*

OpenVX*

Code Samples

For Intel® CPU & GPU/Intel® Processor Graphics

Tools & Libraries

Increase Media/Video/Graphics Performance

Intel® Media SDK
Open Source version

OpenCL™ Drivers & Runtimes

For GPU/Intel® Processor Graphics

Optimize Intel® FPGA (Linux* only)

FPGA RunTime Environment
(from Intel® FPGA SDK for OpenCL™)

Bitstreams

OS Support CentOS* 7.4 (64 bit) Ubuntu* 16.04.3 LTS (64 bit) Microsoft Windows* 10 (64 bit) Yocto Project* version Poky Jethro v2.0.3 (64 bit)

Intel® Architecture-Based Platforms Support

CELERON inside | ATOM inside | CORE inside | XEON inside | ARRIA 10 inside | MOVIDIUS inside | IRIS Pro GRAPHICS

Intel® Vision Accelerator Design Products & Intel/Partner Developer Kits

An open source version is available at 01.org/openvinotoolkit (some deep learning functions support Intel CPU/GPU only).

Intel® Distribution of OpenVINO™ with DLA User Flows

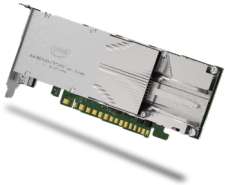
Application Developer

Design

Intel® Distribution of OpenVINO™ toolkit

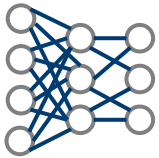
Run

Program



Turnkey Software Deployment Flow

Neural Net



Bitstream Library

Choose from many **precompiled** FPGA Images

Or

Custom create FPGA bitstream

FPGA Architecture Development Flow

IP Architect

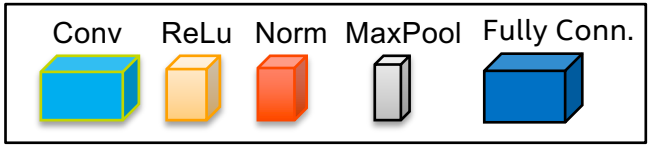
Design

User Customization of DLA Suite Source Code

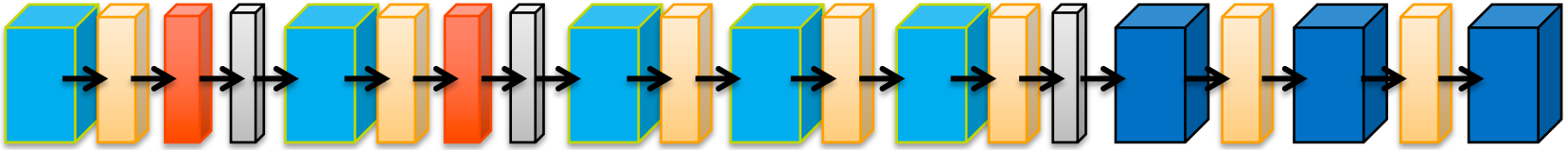
Compile

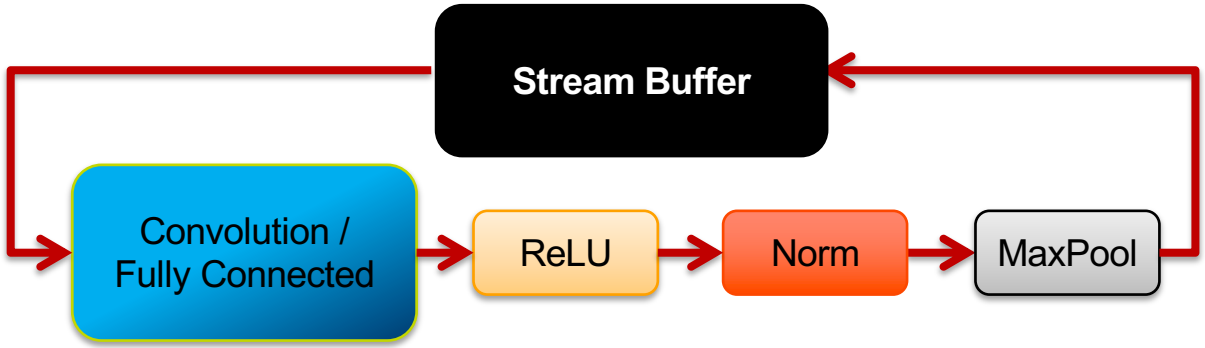
Intel® FPGA SDK for OpenCL™

Mapping Graphs in DLA



AlexNet Graph





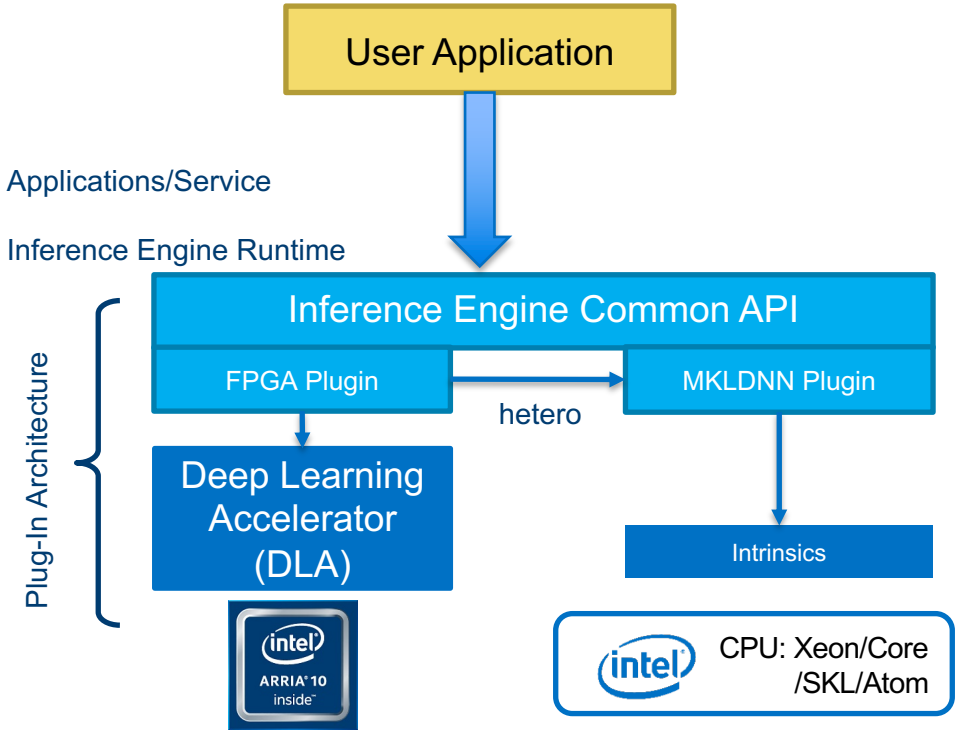
Blocks are run-time reconfigurable and bypassable

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 needs 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/squeezenet1.1/squeezenet1.1.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 Intel® Distribution of OpenVINO™ toolkit
- Available for Intel® Programmable Acceleration Card


intel®
experience
what's inside™