MACHINE LEARNING WITH INTEL® FPGAS

Adrian Macias
Sr. Manager, Software Planning
5/23/2018
AGENDA

• FPGAs Success in Machine Learning
• Introduction to FPGAs and Software Evolution
• Introducing the Intel® FPGA Deep Learning Acceleration Suite
WHY FPGAS WIN IN DEEP LEARNING TODAY

System-Level Optimization
Customizable datapath and precision creating energy-efficient dataflow.
Fine-grained parallelism enabling high throughput on low-batch workloads

Memory-Bound Problems
Extremely high, fine-grained, on-chip memory bandwidth (S10: 58 TBps) that can be more efficiently used to solve break-the-memory wall

Leadership for optimized low-latency systems
(Performance, Power, Cost)

Leadership performance on memory-bound workloads

Common: Quickly support new features; flexible system integration; lower system latency
Scaling in the Data Center

Datacenter

Per-chip performance **increases** when scaled

![Graph showing performance increase with number of FPGAs](image)

- 2x improvement w/ ResNet-101

Higher on-chip memory bandwidth and more usable structure than GPUs

![Graph showing bandwidth and density comparison](image)

- Higher is better

FPGAs scale better when models and data can be resident on-chip
WINNING IN CLOUD: INTEL® FPGAS FOR WORKLOAD ACCELERATION

Microsoft

95% GAIN IN THROUGHPUT†

29% DECREASE IN LATENCY†

8X INCREASE IN SPEED WITH 15% LESS POWER†
FPGA wins on balance of metrics over competitive system solutions
FPGA wins on form factor flexibility (cards vs. chips)
SUCCESS STORY: VIDEO SURVEILLANCE

Competing Vendor

GPU  GPU  GPU  multiple devices

Face-Attribute Classification

Performance
Power
Latency
INTRODUCTION TO FPGAS
Compute Architecture
Compute Evolution
Software Development for FPGA
WHAT IS AN FPGA?

- FPGA architecture: Fine-grained massively parallel
  - Millions of reconfigurable logic elements
  - Thousands of 20Kb memory blocks
  - Thousands of variable precision digital signal processing (DSP) blocks
  - Dozens of high-speed transceivers
  - Multiple high-speed configurable memory controllers
FPGA ARCHITECTURE: CONFIGURABLE ROUTING

Blocks are connected into a custom datapath that matches your application.

Your custom 64-bit bit-shuffle and encode

16-bit add

32-bit sqrt
INTEL® FPGA COMPUTE EVOLUTION

- **Introduction of Variable Precision DSP**
- **First Floating-Point FPGA**
  - 1.5 TFLOPS
  - 50 GFLOPs per Watt
  - 400-450 MHz
- **First 1 GHz FPGA**
  - 9.2 TFLOPS
  - 23 TMACS
  - 80 GFLOPs per Watt
  - 750 MHz-800 MHz
ADVANTAGE OF DEDICATED FLOATING-POINT MATH PRIMITIVES
Removing the Barriers of Adoption

Hardware Developers

- Intel Quartus® Prime Design Software
- High-Level Design Backend Compiler
- LLVM Compiler
- OpenCL™
- DSP Builder for Intel FPGAs
- Intel® HLS Compiler

Software Stacks
- Libraries Overlay
- Primitives
- Deep Learning Acceleration
- Parallel Compilers
- Software Stacks
UNLOCKING THE BENEFIT OF FPGAS WITH HIGH-LEVEL DESIGN

Programming methodology for acceleration
- Pipeline parallelism and single-threaded task
- Software-defined data movement

Custom compute unit synthesis
- C-based programming
- Customized data precision and data flow
SYSTOLIC ARRAY-BASED SGEMM IN OPENCL™

Proof-of-Concept Design using state-of-the-art architecture

- Written in OpenCL
- Highly scalable
- Leverage hardened floating point
- Matrices in external DDR4 SDRAM

Results: > 1TFLOP (FP32)

| ALUTs: 253,755 |
| Registers: 500,735 |
| ALMs: 203,532 / 427,200 (47%) |
| DSP blocks: 1,280 / 1,518 (84%) |
| RAM blocks: 1,195 / 2,713 (44%) |
| Kernel $f_{\text{MAX}}$: 367 MHz |

Evaluated on the following Intel® Arria® 10 FPGA 10AX115
FPGAS FOR AI

Why FPGA for Artificial Intelligence (AI)?
**Design Flow with Machine Learning**

- **Data Collection**
  - Data Store
  - Choose Network
  - Architecture
  - Train Network
  - Parameters
  - Inference Engine
  - Improvement Strategies
    - Collect more data
    - Improve network

---

**Choose Network**
- Use framework (e.g. Caffe, TensorFlow)

**Train Network**
- A high-performance computing (HPC) workload from large dataset
- Weeks-to-months process

**Inference Engine (FPGA focus)**
- Implementation of the neural network performing real-time inferencing
Evolving AI Requirements Benefit from Flexibility (FPGA)

2017

Convolutional neural network (CNN)

Floating point

FP32

2018

ResNet-50

RNN

Floating point

FP16, FP11, FP9, BFLOAT
WHY INTEL® FPGAS FOR MACHINE LEARNING?
**CONSTANT INNOVATIONS IN AI IMPLEMENTATION**

Many efforts to improve efficiency

- Batching
- Reduce bit width
- Sparse weights
- Sparse activations
- Weight sharing
- Compact network

<table>
<thead>
<tr>
<th><strong>Implementation Techniques</strong></th>
<th><strong>Networks</strong></th>
<th><strong>Conferences</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Batching</td>
<td>LeNet</td>
<td>IEEE</td>
</tr>
<tr>
<td>Reduce bit width</td>
<td>AlexNet</td>
<td>ILSVRC’12</td>
</tr>
<tr>
<td>Sparse weights</td>
<td>VGG</td>
<td>ILSVRC’14</td>
</tr>
<tr>
<td>Sparse activations</td>
<td>GoogleNet</td>
<td>ILSVRC’14</td>
</tr>
<tr>
<td>Weight sharing</td>
<td>ResNet</td>
<td>ILSVRC’15</td>
</tr>
<tr>
<td>Compact network</td>
<td>XNORNet</td>
<td></td>
</tr>
</tbody>
</table>

**SparseCNN**
- Spatially SparseCNN [CIFAR-10 winner ‘14]
- Pruning [NIPS’15]
- SparseCNN [CVPR’15]

**DeepComp** [ICLR’16]
- TernaryConnect [ICLR’16]

**SqueezeNet**
- BinaryConnect [NIPS’15]
- HashedNets [ICML’15]

**TernaryConnect** [NIPS’15]

**HashedNets** [ICML’15]

**XNORNet** [ICML’15]

**Bar Chart**

**Diagram**
- Shared Weights
- I
- O
- 3
- 2

**Figure**
- M20K
- M20K
- M20K
- M20K
## Why Intel® Arria® 10 FPGAs for Deep Learning?

<table>
<thead>
<tr>
<th>Feature</th>
<th>Benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highly parallel architecture</td>
<td>Facilitates efficient low-batch video stream processing and reduces latency</td>
</tr>
<tr>
<td>Configurable distributed floating-point DSP blocks</td>
<td>FP32 9 TFLOPS, FP16, FP11, INTx Accelerates computation by tuning compute performance</td>
</tr>
<tr>
<td>Tightly-coupled high-bandwidth memory</td>
<td>&gt;50 TBps on chip SRAM bandwidth, random access, reduces latency, minimizes external memory access</td>
</tr>
<tr>
<td>Programmable Datapath</td>
<td>Reduces unnecessary data movement, improving latency, and efficiency</td>
</tr>
<tr>
<td>Configurability</td>
<td>Support for variable precision (trade-off throughput and accuracy). Future-proof designs and system connectivity</td>
</tr>
</tbody>
</table>

### Feature Benefit

- **Highly parallel architecture**: Facilitates efficient low-batch video stream processing and reduces latency.
- **Configurable distributed floating-point DSP blocks**: FP32 9 TFLOPS, FP16, FP11, INTx Accelerates computation by tuning compute performance.
- **Tightly-coupled high-bandwidth memory**: >50 TBps on chip SRAM bandwidth, random access, reduces latency, minimizes external memory access.
- **Programmable Datapath**: Reduces unnecessary data movement, improving latency, and efficiency.
- **Configurability**: Support for variable precision (trade-off throughput and accuracy). Future-proof designs and system connectivity.
Deterministic System Latency

FPGAs can perform in-line, real-time acceleration on the data ingest and avoid costly data movement within the system.
INTEL® FPGA DEEP LEARNING ACCELERATION SUITE

Turnkey AI Solutions for FPGA
## What’s Inside the OpenVINO™ Toolkit

### Intel Deep Learning Deployment Toolkit
- Cross-platform approach to deep learning inference
- **Model Optimizer**
  - Convert optimized trained models
- **Inference Engine**
  - Run optimized inferences

### OpenCV*
- Optimized functions for Intel processors
- Create own customer kernels or use a library of functions

### Optimized Libraries and OpenVX*
- Runtimes, emulator, kernels, workload samples
- Enhanced, graphical development using Vision Algorithm Designer

---

OpenVX is a trademark of Khronos Group Inc.
Getting Started with FPGAs for Deep Learning Inferencing

Targeting

Edge

Data Center

OpenVino™ Toolkit

Intel® Deep Learning Deployment Toolkit

Intel FPGA Deep Learning Acceleration Suite

Intel Programmable Acceleration Card
with Intel Arria® 10 GX FPGA

Intel FPGA Evaluation:
Intel Programmable Acceleration Card for Data Centers
(Production is available via original equipment manufacturer (OEM))
IEI for Edge* (coming soon)

Intel FPGA Deep Learning Acceleration Suite enables Intel FPGA for deep learning inferencing via the OpenVino™ toolkit
Intel® FPGA Deep Learning Acceleration Suite

Supported Deep Learning Frameworks
- Caffe
- TensorFlow

OpenVino™ Toolkit
- Model Optimizer
- Inference Engine

Intel® FPGA Deep Learning Acceleration Software API
A collection of software graph, compiler, libraries, and runtime

Intel Xeon® Processor
Heterogeneous CPU/FPGA Deployment

Intel FPGA

Current Supported Topologies
(more variants are coming soon)
- AlexNet
- GoogleNet
- Tiny Yolo
- LeNet
- SqueezeNet
- VGG16
- ResNet 18
- ...
- ResNet 50
- ResNet 101

Pre-Compiled Graph Architectures
- GoogleNet optimized template
- ResNet Optimized Template
- SqueezeNet optimized template
- VGG optimized template
- Additional, generic convolutional neural network (CNN) templates

Feature Map Cache
Conv PE Array
Crossbar
Configuration Engine
Memory Reader/Writer
DDR
DDR
DDR
DDR
DDR
DDR
DDR
FPGAs provide the system flexibility and unique compute-memory architecture to differentiate.

FPGA core architecture is well suited for machine learning.

FPGA deploying turnkey with set primitives and customizable solutions for deep learning.
Intel’s compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness or any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice. Notice Revision #20110804
LEGAL NOTICES AND DISCLAIMERS

- Intel technologies’ features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at www.intel.com.
- Performance estimates were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown." Implementation of these updates may make these results inapplicable to your device or system.
- Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.
- This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.
- Any forecasts of goods and services needed for Intel's operations are provided for discussion purposes only. Intel will have no liability to make any purchase in connection with forecasts published in this document.
- ARDUINO 101 and the ARDUINO infinity logo are trademarks or registered trademarks of Arduino, LLC.
- Altera, Arria, the Arria logo, Intel, the Intel logo, Intel Atom, Intel Core, Intel Nervana, Intel Xeon Phi, Movidius, Saffron and Xeon are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries.
- *Other names and brands may be claimed as the property of others.
- OpenCL and the OpenCL logo are trademarks of Apple Inc. used by permission by Khronos.
- Copyright 2018 Intel Corporation.
**LEGAL NOTICES AND DISCLAIMERS**

- This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.
- Intel technologies’ features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at intel.com, or from the OEM or retailer. No computer system can be absolutely secure.
- Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit [http://www.intel.com/performance](http://www.intel.com/performance).
- Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.
- Statements in this document that refer to Intel’s plans and expectations for the quarter, the year, and the future, are forward-looking statements that involve a number of risks and uncertainties. A detailed discussion of the factors that could affect Intel’s results and plans is included in Intel’s SEC filings, including the annual report on Form 10-K.
- The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.
- Performance estimates were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as “Spectre” and “Meltdown.” Implementation of these updates may make these results inapplicable to your device or system.
- No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.
- Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.
- Intel, the Intel logo, Pentium, Celeron, Atom, Core, Xeon, Movidius, Saffron and others are trademarks of Intel Corporation in the U.S. and/or other countries.
*Other names and brands may be claimed as the property of others.
- © 2018 Intel Corporation.