Xilinx ML Suite Overview

Jim Heaton
Sr. FAE
Deep Learning explores the study of algorithms that can learn from and make predictions on data.

Deep Learning is re-defining many applications:

- Cloud Acceleration
- Security
- E-commerce
- Financial
- Surveillance
- Industrial IOT
- Medical Bioinformatics
- Autonomous Vehicles
Accelerating AI Inference into Your Cloud Applications

Featuring the Most Powerful FPGA in the Cloud

Cloud

Virtex® Ultrascale+™ VU9P

On Premises

Zynq® Ultrascale+™ MPSoC

Edge
Want to try the out Xilinx ML Suite?

https://github.com/Xilinx/ml-suite
Xilinx ML Suite - AWS Marketplace

**ML Suite**

- **Supported Frameworks:**
  - Caffe
  - MxNet
  - Tensorflow
  - Keras
  - Python Support
  - Darknet

- **Jupyter Notebooks available:**
  - Image Classification with Caffe
  - Using the xfDNN Compiler w/ a Caffe Model
  - Using the xfDNN Quantizer w/ a Caffe Model

- **Pre-trained Models**
  - Caffe 8/16-bit
    - GoogLeNet v1
    - ResNet50
    - Flowers102
    - Places365
  - Python 8/16-bit
    - Yolov2
  - MxNet 8/16-bit
    - GoogLeNet v1

- **xDNN Tools**
  - Compiler
  - Quantizer

Unified Simple User Experience from Cloud to Alveo

Develop

- User Guides
- Tutorials
- Examples

Publish

- User Guides
- Tutorials
- Examples

Deploy

- AWS
  - Launch Instance
  - Download
  - User Choice

© Copyright 2018 Xilinx
Deep Learning Models

- **Multi-Layer Perceptron**
  - Classification
  - Universal Function Approximator
  - Autoencoder

- **Convolutional Neural Network**
  - Feature Extraction
  - Object Detection
  - Image Segmentation

- **Recurrent Neural Network**
  - Sequence and Temporal Data
  - Speech to Text
  - Language Translation

**Classification**
- “Dog”

**Object Detection**
- Image

**Segmentation**
- Image
Overlay Architecture
Custom Processors Exploiting Xilinx FPGA Flexibility

- Customized overlays with ISA architecture for optimized implementation
- Easy plug and play with Software Stack

**MLP Engine**
Scalable sparse and dense implementation

**xDNN** – CNN Engine for Large 16 nm Xilinx Devices
**Deephi DPU** – Flexible CNN Engine with Embedded Focus

**Deephi ESE**
LSTM Speech to Text engine
Available on AWS

**Random Forest**
Configurable RF classification

© Copyright 2018 Xilinx
### xDNN-v1
- 500 MHz
- URAM for feature maps without caching
- Array of accumulator with
- 16 bit(batch 1), 8 bit(batch 2)
- Instructions: Convolution, Relu, MaxPool, AveragePool, Elementwise
- Flexible kernel size(square) and strides
- Programmable Scaling
- Q4CY17

### xDNN-v2
- 500 MHz
- All xDNN-v1 features
- DDR Caching: Larger Image, CNN Networks
- Instructions: Depth wise Convolution, Deconvolution, Convolution, Transpose Upsampling
- Rectangular Kernels
- Q2CY18

### xDNN-v3
- 700 MHz
- Feature compatible with xDNN-v2
- New Systolic Array Implementation: 50% Higher FMAX and 2.2x time lower latency
- Batch of 1 for 8 bit implementation
- Non-blocking Caching and Pooling
- Q4CY18
Break Through on Peak Performance

GPU: Introduce new architectures and silicon

- Native INT8 operation
- Mixed FP16 for TensorCore

Xilinx: Adapt the break through of emerging domain knowledge

- INT8 Optimization
- Adaptable Precision

GPU Deep Learning Peak Power Efficiency

FPGA Deep Learning Peak Power Efficiency

© Copyright 2018 Xilinx
Seamless Deployment with Open Source Software

RESTful API

From Community
- Caffe
- Scala
- R
- JS
- Python
- Julia
- Go
- MXNet

From Xilinx
- xDNN Middleware, Tools and Runtime
- xDNN Processing Engine

© Copyright 2018 Xilinx
xfDNN flow

- Tensorflow
- MxNet
- Caffe
- CNTK
- Caffe2
- PyTorch
- ONNX

- Framework Tensor Graph to Xilinx Tensor Graph
- xfDNN Tensor Graph Optimization
- xfDNN Compiler
- xfDNN Compression
- xfDNN Runtime (python API)

- Model Weights
- Calibration Set

- Image
- CPU Layers
- FPGA Layers

https://github.com/Xilinx/ml-suite

© Copyright 2018 Xilinx
xfDNN Inference Toolbox

Graph Compiler

- Python tools to quickly compile networks from common Frameworks – Caffe, MxNet and Tensorflow

Network Optimization

- Automatic network optimizations for lower latency by fusing layers and buffering on-chip memory

xfDNN Quantizer

- Quickly reduce precision of trained models for deployment
- Maintains 32bit accuracy at 8 bit within 2%
xfDNN Graph Compiler

Pass in a Network

Microcode for xDNN is Produced

xfDNN Graph Compiler

© Copyright 2018 Xilinx
xfDNN Network Optimization
Layer to Layer

Unoptimized Model

xfDNN Intelligently Fused layers
Streaming optimized for URAM

DDR Buffered

On-Chip URAM Buffered
xFDNN Network Deployment

Fused Layer Optimizations
- Compiler can merge nodes
  - (Conv or EltWise)+Relu
  - Conv + Batch Norm
- Compiler can split nodes
  - Conv 1x1 stride 2 -> Maxpool+Conv 1x1 Stride 1

On-Chip buffering reduces latency and increases throughput
- xFDNN analyzes network memory needs and optimizes scheduler
  - For Fused and “One Shot” Deployment

“One Shot” deploys entire network to FPGA
- Optimized for fast, low latency inference
- Entire network, schedule and weights loaded only once to FPGA
**Problem:**

- Nearly all trained models are in 32-bit floating-point
- Available Caffe and TensorFlow quantization tools take hours and produce inefficient models

**Introducing: xfDNN Quantizer**

- A customer friendly toolkit that automatically analyses floating-point ranges layer-by-layer and produces the fixed-point encoding that loses the least amount of information
  - Quantizes GoogleNet in under a minute
  - Quantizes 8-bit fixed-point networks within 1-3% accuracy of 32-bit floating-point networks
  - Extensible toolkit to maximize performance by searching for minimal viable bitwidths and prune sparse networks
1) Provide FP32 network and model
   • E.g., prototxt and caffemodel

2) Provide a small sample set, no labels required
   • 16 to 512 images

3) Specify desired precision
   • Quantizes to <8 bits to match Xilinx’s DSP
# Xilinx ML Processing Engine – xDNN

<table>
<thead>
<tr>
<th>Features</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Supported Operations</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Strides</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Padding</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Dilation</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Activation</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Bias</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Scaling</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Strides</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Padding</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Strides</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Padding</strong></td>
</tr>
<tr>
<td>Element-wise Add</td>
<td><strong>Width &amp; Height</strong></td>
</tr>
<tr>
<td>Memory Support</td>
<td><strong>On-Chip Buffering, DDR Caching</strong></td>
</tr>
<tr>
<td>Expanded set of image sizes</td>
<td><strong>Square, Rectangular</strong></td>
</tr>
<tr>
<td>Upsampling</td>
<td><strong>Strides</strong></td>
</tr>
<tr>
<td>Miscellaneous</td>
<td><strong>Data width</strong></td>
</tr>
</tbody>
</table>

- Programmable Feature-set
- Tensor Level Instructions
- 700+MHz DSP Freq (VU9P)
- Custom Network Acceleration
Seamless Deployment with Open Source Software

From Xilinx
- xDNN CNN Processing Engine

From Community
- Caffe
- Scala
- R
- JS
- Python
- Julia
- Go

{RESTful API}

Deploy
- TensorFlow Q4 2017

*TensorFlow Q4 2017

© Copyright 2018 Xilinx
Adaptable
- AI algorithms are changing rapidly
- Adjacent acceleration opportunities

Realtime
- 10x Low latency than CPU and GPU
- Data flow processing

Efficient
- Performance/watt
- Low Power
### xDNN PEs Optimized for Your Cloud Applications

#### Throughput, Multi-Network Optimized

<table>
<thead>
<tr>
<th>Overlay Name</th>
<th>DSP Array</th>
<th>#PEs</th>
<th>Cache</th>
<th>Precision</th>
<th>GOP/s</th>
<th>Optimized For</th>
<th>Examples Networks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overlay_0</td>
<td>28x32</td>
<td>4</td>
<td>4 MB</td>
<td>Int16</td>
<td>896</td>
<td>Multi-Network, Maximum Throughput</td>
<td>ResNet50 (224x224)</td>
</tr>
<tr>
<td>Overlay_1</td>
<td>28x32</td>
<td>4</td>
<td>4 MB</td>
<td>Int8</td>
<td>1,792</td>
<td>Multi-Network, Maximum Throughput</td>
<td>ResNet50 (224x224)</td>
</tr>
<tr>
<td>Overlay_2</td>
<td>56x32</td>
<td>1</td>
<td>5 MB</td>
<td>Int16</td>
<td>1,702</td>
<td>Lowest Latency</td>
<td>Yolov2 (224x224)</td>
</tr>
<tr>
<td>Overlay_3</td>
<td>56x32</td>
<td>1</td>
<td>5 MB</td>
<td>Int8</td>
<td>3,405</td>
<td>Lowest Latency</td>
<td>Yolov2 (224x224)</td>
</tr>
</tbody>
</table>
Inference with batches

- Requires batch of input data to improve data reuse and instruction synchronization
- High throughput depends on high number of batch size
- High and unstable latency
- Low compute efficiency while batch is not fully filled or at lower batch size

Real Time Inference

- No requirement for batch input data
- Throughput less related to batch size
- Low and deterministic latency
- Consistent compute efficiency
Xilinx - High Throughput at Real-Time

GoogLeNet V1 Performance

Consistent High Throughput at Low Latency

Degradation of Throughput with Lower Latency

Throughput (images/sec)

Latency (ms)

Alveo U200 v3 XDNN

P4 with Tensor RT4.0

P4 with Tensor RT3.0

© Copyright 2018 Xilinx
Fast
Advantages in Machine Learning Inference

INCREASE REAL-TIME MACHINE LEARNING* THROUGHPUT BY 20X

20x advantage

* Source: Accelerating DNNs with Xilinx Alveo Accelerator Cards White Paper
Alveo Overview

Jim Heaton
Sr. FAE
Alveo – Breathe New Life into Your Data Center

16nm UltraScale™ Architecture

Off-Chip Memory Support
• Max Capacity: 64GB
• Max Bandwidth: 77GB/s

Internal SRAM
• Max Capacity: 54MB
• Max Bandwidth: 38TB/s

PCIe Gen3x16

Accelerate Any Application
• IDE for compiling, debugging, profiling
• Supports C/C++, RTL, and OpenCL

Cloud Deployed

Cloud ↔ On-Premise Mobility

Ecosystem of Applications
• Many available today
• More on the way

Server OEM Support
• Major OEMs in Qualification

© Copyright 2018 Xilinx
Alveo Accelerator Card Value Proposition

**Fast**
*Highest Performance*
- Faster than CPUs & GPUs
- Latency advantage over GPUs

**Adaptable**
*Accelerate Any Workload*
- Optimize for any workload
- Adapt to changing algorithms

**Accessible**
*Cloud ↔ On-Premises Mobility*
- Deploy in the cloud or on-premises
- Applications available now
Accelerator Cards That Fit Your Performance Needs

**Alveo U200**
- 18.6 Peak INT8 TOPs
- 77GB/s DDR Memory Bandwidth
- 31TB/s Internal SRAM Bandwidth
- 892,000 LUTs

**Alveo U250**
- 33.3 Peak INT8 TOPs
- 77GB/s DDR Memory Bandwidth
- 38TB/s Internal SRAM Bandwidth
- 1,341,000 LUTs

**Alveo U280**
- 24.5 Peak INT8 TOPs
- 460GB/s HBM2 Memory Bandwidth
- 30TB/s Internal SRAM Bandwidth
- 1,079,000 LUTs

**Buy Now**
- **Product Brief >**

**Buy Now**
- **Product Brief >**

**Learn More**
- **Coming Soon**
Expanding Accelerator Card Portfolio

- U250: Available Now
- U200: Available Now
- U280
- SmartNIC

Performance & Capability
- 16nm UltraScale+™
- 7nm Versal

2018

2019

Planned

Broad Portfolio of Acceleration Cards
Accessible
Unified Simple User Experience from Cloud to Alveo

Develop

Publish

Deploy

User Guides  Tutorials  Examples

User Guides  Tutorials  Examples

AWS
Launch Instance

User Choice

Download
Adaptable - On-chip memory
The critical asset on-chip to assist and feed the compute
Store parameters, buffer intermediate activations, and move data around

> Alveo: Highest on-chip memory capacity
> Alveo: Most adaptable memory architecture
Adaptable - Neural Network Inference Efficiency

GoogLeNet v1 Peak Efficiency

- Xilinx Alveo U200: 50%
- Xilinx Alveo U250: 40%
- Nvidia Tesla T4: 20%

Fast Real-time Performance

GoogLeNet v1 Real-time (batch = 1) Performance/Efficiency

Xilinx Performance Roadmap

Model Compression, Low and Custom precision arithmetic

Silicon Improvement (Versal)

© Copyright 2018 Xilinx
Beyond Machine Learning
Fast Advantages Across Many Workload Types

Database: 90x
Financial: 89x
Machine Learning: 20x
Video: 12x
HPC & Life Sciences: 10x
# It’s All About the Applications

## Database
- Elasticsearch
- Low Latency KVS
- ETL, Streaming, SQL Analytics
- Deepgreen DB
- Spark-MLlib
- Hyperion 10G RegEx

## Machine Learning
- ML Suite for Inference
- Image Classification
- Text Analysis

## Video
- ABR Transcoding
- Experience Engine
- Image Processing

## Financial
- Real-Time Risk Dashboard
- SIMM Calculation

## HPE & Life Sciences
- Accelerated Genomics Pipelines

Application Ecosystem Continues to Grow
Infuse Machine Learning with other accelerations

Database

Video

Machine Learning

HPC & Life Sciences

Financial
Solution Stack

**Accelerated Solutions**

100% Growth of Published Applications

Hundred of Developers Trained

RTL, C, C++, OpenCL

**Developer Package**

SDAccel Environment

**Platforms**

FPGA as a Service (FaaS)

Cloud

On-premise

**End user**

Xilinx ML Suite

Framework, API, Python/Java/C++ Programmability

HPC & LIFE SCIENCES

FINANCIAL

VIDEO

MACHINE LEARNING

DATABASE

Solutions Xilinx ISVs

Platform
Xilinx is Qualifying with Major Server OEMs

Server Support & Qualification Strategy

<table>
<thead>
<tr>
<th>Interop Level Qual</th>
<th>Stringent Qual</th>
<th>Very Stringent Qual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xilinx Validated</td>
<td>3rd Party Option (3PO) Qualified</td>
<td>OEM Factory Qualified</td>
</tr>
<tr>
<td>• Dell R730</td>
<td>• SuperMicro SYS-4028GR-TR</td>
<td>• HPE ProLiant DL380 G10</td>
</tr>
<tr>
<td>• Dell R740</td>
<td>• SuperMicro SYS-4029GP-TRT</td>
<td></td>
</tr>
<tr>
<td>• SuperMicro SYS-7049GP-TRT</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

© Copyright 2018 Xilinx
## Ordering Information

<table>
<thead>
<tr>
<th>Alveo Accelerator Card</th>
<th>Part Number</th>
<th>SRP 1pc</th>
</tr>
</thead>
<tbody>
<tr>
<td>U200</td>
<td>A-U200-A64G-PQ-G</td>
<td>$8,995</td>
</tr>
<tr>
<td></td>
<td>A-U200-P64G-PQ-G</td>
<td></td>
</tr>
<tr>
<td>U250</td>
<td>A-U250-A64G-PQ-G</td>
<td>$12,995</td>
</tr>
<tr>
<td></td>
<td>A-U250-P64G-PQ-G</td>
<td></td>
</tr>
</tbody>
</table>

Two Cooling Options

- Passive
  - No Fan
- Active
  - Built in Fan

 Blo via standard quote/PO process from Xilinx or Avnet

> Buy via eCOM (at SRP) from Xilinx, Avnet, DigiKey, EBV, or Premier Farnell

> Developer discount available
  >> Requires qualification through [Accelerator Program](#)
More Information Available on Xilinx.com

Xilinx.com

- Product Brief
- Product Selection Guide
- Getting Started Guide
- Data Sheet
- ML Solution Brief
- SDAccel Solution Brief
- ABR Transcoding Solution Brief
- Accelerating DNNs with Alveo White Paper
- Applications Directory

Accelerating DNNs with Alveo White Paper

Applications Directory
Adaptable.
Intelligent.