Building the Adaptable, Intelligent World

Ivo Bolsens
Senior Vice President & Chief Technology Officer
Mountains of Unstructured Data

One Architecture Can’t Do It Alone

This is the Era of Heterogeneous Compute
Today’s Developer Needs

- Performance for a Diverse Range of Applications
- Software Programmability
- Adaptability to Keep Pace with Rapid Innovation
Enter the ACAP
A New Class of Devices for Today’s Challenges
Adaptive Compute Acceleration Platform
Compute Acceleration

Scalar Engines

Adaptable Engines

AI Engines
The Industry’s First ACAP

- Heterogeneous
- Parallel
- Scalable
- SW Programmable
- HW Adaptable
Network-on-Chip (NoC)

Ease of Use
Inherently Software Programmable
Available at Boot, No Place-and-Route Required

High Bandwidth and Low Latency
Multi-Terabit/Sec Throughput
Guaranteed QoS

Power Efficiency
8x Power Efficiency vs. Soft Implementations
Arbitration Across Heterogeneous Engines

© Copyright 2018 Xilinx
AI Engines

- High Throughput, Low Latency, and Power Efficient
- Ideal for AI Inference and Advanced Signal Processing

© Copyright 2018 Xilinx
Platform for Any Developer

Adaptable for Any Application
Software Programmable
Heterogeneous Platform

User Application
C, C++, Python

Application-Specific Frameworks
Machine Learning | Video | Genomics | Search | Financial Modeling | Database

New Unified Software Development Environment
OS & Embedded Run-Time
Custom HW
Xilinx & Ecosystem HW Libraries
C, Xilinx Libraries
Scalar Engines
Adaptable Engines
Intelligent Engines
VERSAL

© Copyright 2018 Xilinx
Versal Multi-Market Platform

CLOUD

Data Center

NETWORK

Wired

Wireless

EDGE

Endpoints
VERSAL AI Core Series
For 5G Beamforming & CloudRAN
AI Engines Provides >5X Compute Density for Advanced Wireless Compute
Projected Growth in AI Inference

Barclays Research, Company Reports May 2018

© Copyright 2018 Xilinx
Challenges

- The Rate of AI Innovation
- Performance at Low Latency
- Low Power Consumption
- Whole App Acceleration

Inference

Unlabeled Data → Model → “Dog” → Estimate
The Rate of AI Model Innovation

DIVERSE MODELS OVER A BROAD RANGE OF APPLICATIONS

Classification

Object Detection

Segmentation

Speech Recognition

Recommendation Engine

Anomaly Detection

CNN

RNN, LSTM

MLP

© Copyright 2018 Xilinx
The Rate of AI Model Innovation

Classification


© Copyright 2018 Xilinx
Rate of Innovation Outpaces Silicon Cycles

AlexNet

GoogLeNet

DenseNet

Start Design

Silicon Design Cycle (time)

Production Design

© Copyright 2018 Xilinx
Low Latency is Critical for Inference

High Throughput OR Low Latency

CPU/GPU
- Input 1
- Input 2
- Input 3
- Input 4
- Result 1
- Result 2
- Result 3
- Result 4

FPGA
- Input 1
- Input 2
- Input 3
- Input 4
- Result 1
- Result 2
- Result 3
- Result 4

50ms Latency Response

3ms Latency Response

© Copyright 2018 Xilinx
Inference Moving to Lower Precision

<table>
<thead>
<tr>
<th>Operation</th>
<th>Energy (pJ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>8b Add</td>
<td>0.03</td>
</tr>
<tr>
<td>16b Add</td>
<td>0.05</td>
</tr>
<tr>
<td>32b Add</td>
<td>0.1</td>
</tr>
<tr>
<td>16b FP Add</td>
<td>0.4</td>
</tr>
<tr>
<td>32b FP Add</td>
<td>0.9</td>
</tr>
</tbody>
</table>

Source: Bill Dally (Stanford), Cadence Embedded Neural Network Summit, February 1, 2017
Reduced Precision Arithmetic

ResNet-50L ImageNet Top5 Error vs Hardware Cost

- Notation: 3b/5b: 3 bit weights/5 bit activation
- Reduced Precision Arithmetic
- Retrained
- Float
- Direct Quantization
- Retrained

1.00
10.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00

1.00
10.00
100.00
1000.00

Error (%)

Hardware Cost (LUT + 100*DSP)

1b/2b
2b/8b
4b/6b
3b/5b
5b/5b
8b/8b
Floating Point Baseline

Notation: 3b/5b: 3 bit weights/5 bit activation
Need for Adaptable Hardware

- Custom Data Flow
- Custom Memory Hierarchy
- Custom Precision

Domain Specific Architectures (DSAs)
on Adaptable Platforms
Low Latency
Xilinx’s Unique Advantage

Latency Tolerant Inference

GoogLeNet v1

2X vs. GPU

43X vs. CPU

High-End CPU
High-End GPU
Xilinx

AI Inference Acceleration
Leveraging AI Engines
Majority of Adaptable & Scalar Engines
Available for Whole App Acceleration

(1) Measured on EC2 Xeon Platinum 8124 Skylake, c5.18xlarge AWS instance, Intel Caffe: https://github.com/intel/caffe
(2) V100 results taken from Oct 9th updates on www.Nvidia.com
(3) Versal Core Series
(4) GoogLeNet V1 throughput (img/sec)

© Copyright 2018 Xilinx
Low Latency

Xilinx’s Unique Advantage

Sub – 7ms Latency

GoogLeNet v1

2X vs. GPU

72X vs. CPU

High-End CPU

High-End GPU

Xilinx

AI Inference Acceleration

Leveraging AI Engines

Majority of Adaptable & Scalar Engines Available for Whole App Acceleration

---

(1) Measured on EC2 Xeon Platinum 8124 Skylake, c5.18xlarge AWS instance, Intel Caffe: [https://github.com/intel/caffe](https://github.com/intel/caffe)
(2) V100 results taken from Oct 9th updates on [www.Nvidia.com](http://www.Nvidia.com)
(3) Versal Core Series
(4) GoogLeNet V1 throughput (img/sec)

© Copyright 2018 Xilinx
Low Latency
Xilinx’s Unique Advantage

Sub – 2ms Latency

GoogLeNet v1

4X vs. GPU

High-End GPU

Xilinx

AI Inference Acceleration

Leveraging AI Engines

Majority of Adaptable & Scalar Engines Available for Whole App Acceleration

---

(1) Measured on EC2 Xeon Platinum 8124 Skylake, c5.18xlarge AWS instance, Intel Caffe: [https://github.com/intel/caffe](https://github.com/intel/caffe)
(2) V100 results taken from Oct 9th updates on [www.Nvidia.com](http://www.Nvidia.com)
(3) Versal Core Series
(4) GoogLeNet V1 throughput (img/sec)

© Copyright 2018 Xilinx
Whole Application Acceleration

Intelligent Video Analytics

CPU-GPU

- H.264 Decode
- Motion Analysis
- CNN
- PCIe

GPU

- Decode 16ms
- OpenCV 16ms
- CNN 50ms

- Power: 75W
- Latency: 82 ms
- Throughput: 4x12 fps

CPU-Xilinx FPGA

- H.264 Decode
- Motion Analysis
- CNN
- PCIe

FPGA

- 16ms
- 0.9ms
- CNN 9.2ms

- Power: 50W
- Latency: 26.1 ms
- Throughput: 4x38 fps

© Copyright 2018 Xilinx
Enabling the Development Community

- Accelerated Libraries
- Pruning / Compression
- Compiler & Quantization Tools
- Runtime
- Processor Overlays (DNN, LSTM, RNN, MLP)

- FPGA-as-a-Service
- Alveo
- Custom Board

© Copyright 2018 Xilinx
<table>
<thead>
<tr>
<th>Role</th>
<th>Platforms/Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Scientists</td>
<td>Frameworks: Python, APIs</td>
</tr>
<tr>
<td></td>
<td><img src="#" alt="DeepHi" />, Caffe, mxnet, FFmpeg, TensorFlow</td>
</tr>
<tr>
<td>SaaS Developers</td>
<td>FaaS Platform</td>
</tr>
<tr>
<td></td>
<td><img src="#" alt="AWS" />, Huawei, Aliyun, Nimbix</td>
</tr>
<tr>
<td>Application Developers</td>
<td>SDX: C++, OpenCL, Libraries, XRT open source runtime</td>
</tr>
<tr>
<td></td>
<td><img src="#" alt="Linux" />, RTOS, Xen</td>
</tr>
<tr>
<td>Embedded Developers</td>
<td>MPSoC Software Environment</td>
</tr>
<tr>
<td>Hardware-Aware Software</td>
<td>HLS: C++ IP Functions</td>
</tr>
<tr>
<td>Developers</td>
<td></td>
</tr>
<tr>
<td>System Integrators</td>
<td>IP Integrator</td>
</tr>
<tr>
<td>Hardware Developers</td>
<td>Vivado Design Suite: RTL Full Design</td>
</tr>
</tbody>
</table>
Data Center First
XILINX
ALVEO™

Fast
Faster than CPUs & GPUs
Latency Advantage Over GPUs

Adaptable
Optimized for Any Workload
Adapt to Changing Algorithms

Accessible
Deploy in the Cloud or On-Premises
Applications Available Now

© Copyright 2018 Xilinx
<table>
<thead>
<tr>
<th>TOOLS</th>
<th>USER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frameworks</td>
<td>Data Scientists and AI Developers</td>
</tr>
<tr>
<td>Libraries, Compilers, Middleware</td>
<td>Application Developers</td>
</tr>
<tr>
<td>Firmware and Runtime</td>
<td>Software Developers</td>
</tr>
<tr>
<td>Integrated Development Environment</td>
<td>Hardware and Software Developers</td>
</tr>
</tbody>
</table>

© Copyright 2018 Xilinx