Adaptable Intelligence
The Next Computing Era

Hot Chips, August 21, 2018
Victor Peng, CEO, Xilinx
Pervasive Intelligence from Cloud to Edge to Endpoints
Exponential Growth and Opportunities

Data Explosion

Source: Patrick Cheesman

Hyperscale Data Centers

Source: Cisco Global Cloud Index, 2016-2021

Hyperscale Market Value

Source: Bloomberg; company reports; The Economist estimates
Challenges: The End of Moore’s Law and Scaling

40 Years of Processor Performance

- 2um VAX
- .75um VAX
- .35um Alpha 21264
- 180nm MIPS IP Core
- 55nm Radeon HD4870
- 16nm Zynq RFSoC
- End of Dennard Scaling Multicore 2x / 3.5yrs (23%/yr)
- End of the Line? 2x / 20yrs (3%/yr)
- Amdahl’s Law 2x / 6yrs (12%/yr)

Source: John Hennessy and David Patterson, Computer Architecture: A Quantitative Approach, 6/e 2018
Challenge: Exponential Power Density Growth

Source: John Hennessy and David Patterson: A New Golden Age for Computer Architecture
Domain-Specific Hardware/Software Co-Design, Enhanced Security, Open Instruction Sets, and Agile Chip Development

Power consumption based on models in “Dark Silicon and the End of Multicore Scaling”
Hadi Esmaeilizadeh, ISCA, 2011
The Third Wave: Domain Specific Architectures on Adaptable HW

CPU
CISC → RISC → Multi-Core

Fixed HW Accelerators
GPU, ASSP, ASIC

DSA’s on Adaptable Platforms
FPGA, Zynq, ACAP
Massive Scale Out Requires DSA’s and Adaptable Platforms

Integration of AI

<table>
<thead>
<tr>
<th>Era</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mainframe Era</td>
<td>Units (M)</td>
</tr>
<tr>
<td>PC Era</td>
<td>Units (100M’s)</td>
</tr>
<tr>
<td>Mobile Era</td>
<td>Units (B’s)</td>
</tr>
<tr>
<td>Pervasive Intelligence Era</td>
<td>Units (50B’s)</td>
</tr>
</tbody>
</table>
The Innovation to Deployment Acceleration Imperative

AI Papers Published

Over-the-air Updates

Source: Scopus
Application Acceleration with DSA’s on FPGA Platforms

FPGA’s Accelerate
Entire Application

Hyperscale Data Centers
with FPGA Accelerators

Execution Time

CPU only
Front End | ML | Back End

CPU + ML ASIC
Front End | ML | Back End

CPU + FPGA
Front End | ML | Back End

App 1
Front End | ML | Back End

App 2
Front End | ML | Back End

App N
Front End | ML | Back End

Server CPU

FPGA Accelerator

Database Workloads
Search (Database + ML)
Speech (ML) Workloads
Security Workloads
FPGA Plane
CPU Plane

Virtualized & Scaled Out Adaptable Acceleration
Dynamic Optimization for Changing Workloads & Mix
Development for DC Compute, Storage, Network Apps
Stack for Application Acceleration including ML

<table>
<thead>
<tr>
<th>User Applications</th>
<th>C++ with Framework API</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Application-Specific Framework SW</strong></td>
<td></td>
</tr>
<tr>
<td>TensorFlow</td>
<td>SQL</td>
</tr>
</tbody>
</table>

**Application-Specific Acceleration**
- OpenCV
- Genomics
- Video
- Search
- Financial
- Database
- SmartNIC

**ML Acceleration**
- xDNN
- xDNN

**DSA Shell**

**Silicon**
- FPGA or SoC or ACAP

---

Xilinx & Partner Provided Accelerators

Frameworks Provide API to User Applications

Xilinx Provided ML Engines

Standard Platform Interface (DDR, PCIe, Network)
Cloud: Latency-sensitive High Resolution Imaging

**CPU/GPU**

- **Motion Analysis**
- **Custom Operation**
- **PCle**
- **CPU**
- **CNN**
- **GPU**

**OpenCV**

- Pre-processing a Bottleneck
- Data Sharing Between CPU and GPU

**CNN**

- 100ms
- 2.3s

**CPU/GPU Results**

**Xilinx FPGA**

- **PCle**
- **CPU**
- **FPGA**
- **CNN**
- **Image Scaling, Tiling**
- **Custom Operation**

**OpenCV**

- 20ms
- 0.85s

**FPGA Results**

**2.7X Faster**

**OpenCV**

- 5x Faster
- Up to 3x Faster
- Based on Image Size
- Model Parallelism for High Res Processing
- Lower Power

**2.7X Faster and Lower Power**

© Copyright 2018 Xilinx
Cloud: Security / Anomaly Detection

CPU/GPU

- CPU
- Custom Operation
- PCIe
- NIC
- MLP/Random Forrest
- GPU
- PCIe

CPU/GPU Results

- OpenCV
- MLP
- 50 µs

Xilinx FPGA

- CPU
- PCIe
- FPGA
- Custom Operation
- MLPRandom Forrest
- Smart NIC
- PCIe

FPGA Results

- OpenCV
- MLP
- 10 µs

5X Lower Latency with High Security

- High Performance NIC: 100+ G
- MLP: MLP/RF Acceleration
- Security Vulnerability: Data Travels thru CPU to GPU
- Integrated Single Card/Chip Solution
- High Speed Smart NIC
- Real-time MLP/RF at 5x Lower Latency
- TCO and Power Advantage: 1 Card vs 2 Cards
- In-line Security Detection is Much Higher Security
Cloud: Smart City / Security

CPU/GPU

- Motion Analysis
- Custom Operation
- H.264 Decode
- OpenCV: Pre-Processing/Motion Analysis on CPU
- CNN: Object Detection on GPU
- Data Sharing Between CPU and GPU

CPU/GPU Results

<table>
<thead>
<tr>
<th></th>
<th>Decode</th>
<th>OpenCV</th>
<th>CNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>H.264</td>
<td>16 ms</td>
<td>10 ms</td>
<td></td>
</tr>
</tbody>
</table>

CPU/Xilinx FPGA

- Motion Analysis
- Custom Operation
- H.264 Decode
- CNN: 5ms Object Detection (xDNN Acceleration)
- Integrated Single Chip Solution: Cloud (VU9/13P)
- Lower Power

FPGA Results

<table>
<thead>
<tr>
<th></th>
<th>Decode</th>
<th>OpenCV</th>
<th>CNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>H.264</td>
<td>0.9 ms</td>
<td>1.7 ms</td>
<td></td>
</tr>
</tbody>
</table>

10x Lower Latency and Lower Power
Xilinx Programmable Architecture Milestones

- **First FPGA Introduced**: 1980
- **First Virtex FPGA**: 1990
- **Virtex-2 Pro**: 2000
- **First 3D FPGA & HW/SW Programmable SoC**: 2010
- **First MPSoC & RFSoC**: 2016
- **ACAP**: 2019

© Copyright 2018 Xilinx
New Device Category for Adaptive Workload-specific Acceleration

- HW/SW Programmable Engines
- IP Subsystems and a Network-on-Chip
- Platform Offerings for Compute / Storage / Networking
Adaptive Compute Acceleration Platform

- Dynamically Adaptable to Workloads
- Exponential Increase in Acceleration
- Software Programmable
Adaptive Compute Acceleration Platform

> **20x** ML Inference Performance

> **4x** 5G Communications Bandwidth

> **112G** Transceivers
New HW/SW Programmable Architecture

Application-level Performance Enabled by SW Programmable Engine

- Everest 7nm
- 16nm

<table>
<thead>
<tr>
<th>ML Inference</th>
<th>5G Wireless</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>20x*</td>
<td>4x</td>
<td>40% Less Power</td>
</tr>
</tbody>
</table>

Compute Efficiency
- Domain Specific Engine
- Greater Compute Density
- Xilinx 7nm Everest

Multiple Applications
- ML Inference for Cloud DC
- Wireless 5G: Radio, Baseband
- ADAS/AD Embedded Vision
- Wired: DOCSIS Cable Access

Heterogenous Architecture
- PE Throughout and Efficiency
- PL Flexibility
- Customized Memory Hierarchy

SW Programmable
- SW Programmable (e.g., C/C++)
- Compile, Execute, Debug
- Increased Productivity

Array Architecture

>> 18

© Copyright 2018 Xilinx
xDNN: Adaptable Overlay DNN Processor for Xilinx FPGA

No FPGA Expertise Needed

Trained Model → Compiler + Runtime → Overlay DNN Processor

60-80% Computation Efficiency

Low Latency & High Throughput (Batch = 1)
Soft Overlay Architectures for ML

DeePhi Aristotle Architecture

DeePhi Descartes Architecture

[Diagram showing the architecture of the Aristotle and Descartes platforms, including components such as High Performance Scheduler, Instruction Fetch Unit, Global Memory Pool, High Speed Data Tube, RAM, Host CPU, AXI4 Interface, Interconnect, Memory, DMA, AXI4 Lite Interface, Sparse_LSTM IP, Acceleration Kernel, and Host.]
Building the Adaptable, Intelligent World