All Programmable SoCs?
Platforms to enable the future of Embedded Machine Learning

Tomas Evensen
CTO Embedded Software, Xilinx
CPU Architectures not Scaling with Workloads

- Processor frequency scaling ended in 2007
- Multicore architecture scaling has flattened

Workloads require higher performance, lower latency
- **Cloud**: video, big data, AI...
- **Edge**: auto, surveillance, AI...
Zynq UltraScale+ MPSoC

**Application Processor**
- 64-bit Dual/Quad-Core

ARM Cortex® R5
- Real-Time Processors
- 32-bit Dual-Core

ARM Cortex® A53
- Application Processor
- 64-bit Dual/Quad-Core

Fabric Acceleration
- Customizable Engines
- High Speed Connectivity

Platform & Power Management
- Granular Power Control
- Functional Safety

Memory Subsystem
- High Bandwidth
- Low Latency

Graphics Processor
- ARM Mali-400MP2

High Speed Peripherals
- Key Interfaces

Video Codec
- 8K4K (15fps)
- 4K2K (60fps)

Configuration & Security Unit
- Anti-Tamper & Trust
- Industry Standards

Memory Subsystem
- ARM® Mali™-400MP2

Processing System
- Quad-Core ARM® Cortex™-A53
- Memory Subsystem
- ARM® Cortex™-R5
- Platform Management
- Config and Security
- System Functions

Programmable Logic
- UltraRAM
- Video Codec

Peripherals
- PCIe® Gen4
- 10G Ethernet
- 100G Interconnect

High Bandwidth
- Low Latency

Video Codec
- HEVC
- MPEG-4

© Copyright 2018 Xilinx
FPGA: The “Chameleon” Chip

What is FPGA/Fabric/Programmable Logic:
- Is it glue logic?
- Is it a powerful parallel DSP engine?
- Is it an RTL simulator?

Yes!!! And more...

FPGA includes:
- Programmable logic (LUTs)
- Hardened DSP blocks
- Hardened memory (BRAM, URAM)

FPGAs are great to implement:
- Parallel compute (e.g. MAC)
  - With variable precision
- Parallel, flexible dataflows
  - Build your own buses
- Flexible, multiport memory hierarchies
Breakout in Programming Model

Development Productivity

SW Programmability

15x productivity with HLS, IPI

Traditional HW design

© Copyright 2018 Xilinx
SDSoC Example: Matrix Multiply + Add

```c
main()
{
    malloc(A,B,C);
    mmult(A,B,D);
    madd(C,D,E);
    printf(E);
}
```

```c
madd(inA,inB,out){
    HLS C/C++
}
```

```c
mmult(inA,inB,out){
    HLS C/C++
}
```

**SDSoC Environment**

Generated

PS

PL

Application

Driver

AXI Bus

datamovers

© Copyright 2018 Xilinx
Supporting the Whole Stack

Accelerated Open Frameworks
- Caffe
- FFmpeg
- STORM
- gatk

Accelerated Libraries
- Machine learning
- H.265 HEVC (High efficiency video coding)
- Database Analytics
- OpenCV

Development Environment
- SDAccel Environment
- SDx Environments
- SDSSoC Environment

Boards w/ HLx-based platform
- VCU1525 Acceleration card

Development Stack

Platform
Example of Embedded Vision Application at the Edge

Deep learning-based multi-object recognition for smart-city

Live video object detection
SSD @ 480x360 on Zynq MPSoC

End-customer obtained:
- 5x NVIDIA TX2 performance
- Better accuracy
Xilinx and Avnet is partnering and is announcing the Ultra96 board
- Equipped with Zync Ultrascale+ MPSoC (ZU3EG)
- https://www.96boards.org/product/ultra96/

Ultra96 board makes ARM® based Xilinx SoCs available to developers at a low price point
- Built to 96Boards standard, suitable for software prototyping with standardized expansion kits
- Targeting a range of applications including Machine Learning, IoT, and compute

Leverages an open-source software development platform
- **96boards community**: 12K actively contributing software engineers
- Supports both self-hosted and cross development
  - Self-hosted: Compile on the board itself
  - Cross: Develop on your workstation/laptop
  - C to fabric/FPGA: SDSoC tools available later this year
  - Unboxing to coding in less than 2 minutes

$249 Ultra96 Board Targeted for Software Designers
Available from Avnet in April
The Future is *Ultra96 Xilinx Contest*

» Submit your most creative, most out-of-the-box AI or ML application at the Xilinx or Avnet table during Demo Friday (12:00 – 14:00).

» The best **30** get a **FREE Ultra96 board** plus software to help you realize your vision.

» The 1st twenty to submit a working design by MAY 25th, 2018 get a $25 Amazon Gift Card.

» **ONE Winner** announced through Xilinx social media channels. If it’s you, you’re invited to present your design to your peers in industry at Xilinx Developer Forum 2018.
German Road Sign Database
- 50,000+ 32x32 bit images for training
- 44 classes (43 road signs, 1 background)
- Training via Amazon Web Services
  - AWS: p2.xlarge Instance – 8 hours → $7.78 6.5e

Binary Neural Network Characteristics
- 6 convolutional layers
- 2 max pool layers
- 3 fully connected layers
Neural Network Performance Results

<table>
<thead>
<tr>
<th>Performance Metric</th>
<th>Software Only</th>
<th>Programmable Logic Accelerated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tiles per second</td>
<td>2.2</td>
<td>19,000</td>
</tr>
<tr>
<td>Scene rate (fps)</td>
<td>0.011 (92 sec per frame)</td>
<td>94</td>
</tr>
<tr>
<td>Overall Acceleration</td>
<td>-</td>
<td>8,600X</td>
</tr>
</tbody>
</table>

Up to 8,600 times faster when accelerated with programmable logic
Learn More About FPGA’s and Software Acceleration

- Dramatically Accelerate 96Board Software via an FPGA with Integrated Processors
  - Wednesday 16:00-16:55

- Accelerating Neural Networks for Vision Systems via FPGAs
  - Thursday 11:00-11:25
Questions?/ Thank You