Versal:
AI Engine & Programming Environment

Presented By
Ambrose Finnerty
Xilinx DSP Technical Marketing Manager

October 16, 2018
Motivation for AI Engine
Motivation for AI Engine

Dynamic Markets Require Adaptable Compute Acceleration

- 5G
- ADAS / AD
- Smart City
- Smart Factory
- Data Center Workloads
- Machine Learning
- AI Everywhere

Applications
- Compute Intensity
- Real Time Capability
- Power Efficiency
- Moore’s Law
- Performance & Power Scaling
- Traditional Single / Multi-core

Technology Scaling
Delivering Adaptable Compute Acceleration

<table>
<thead>
<tr>
<th></th>
<th>CPU (Sequential)</th>
<th>GPU (Parallel)</th>
<th>ACAP</th>
<th>Custom ASIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>SW Programmable</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>HW Adaptable</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Workload Flexibility</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Throughput vs. Latency</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Device / Power Efficiency</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

ACAP w/ AI Engine

© Copyright 2018 Xilinx
Introducing the AI Engine

- SW Programmable
- Deterministic
- Efficient

Artificial Intelligence

- 1GHz+ Multi-precision Vector Processor
- High bandwidth extensible memory
- Up to 400 AI Engines per device
- 8X Compute Density
- 40% Lower Power

Adaptable. Intelligent.

- CNN
- LSTM / MLP

Signal Processing

Computer Vision
Software Programmable: Any Developer

1. Design
   - C/C++
   - Frameworks: mxnet, TensorFlow, Caffe
   - 4G/5G/Radar Library
   - AI Library
   - Vision Library

2. Compile
   - AI Engine Compiler

3. Run
   - Programming Abstraction Levels
     - Architecture Overlay
     - Data Flow w/ Xilinx libraries
     - Kernel Program
       Data Flow w/ user defined libraries

Page 6 © Copyright 2018 Xilinx
Hardware Adaptable: Accelerating the Whole Application

- **Hardware Features**
  - Scalar, Sequential & Complex Compute
  - Flexible Parallel Compute, Data manipulation
  - ML & Signal Processing, Vector, Compute Intensive

- **Performance**
  - 160 GB/s of Memory B/W per Core
  - 1 TB/s of Bandwidth PL-to-AI Engine

- **Connectivity**
  - Any-to-Any Connectivity
  - Custom Memory Hierarchy
  - NETWORK-ON-CHIP
  - I/O

- **Applications**
  - Video + AI
  - Genomics + AI
  - Risk Modeling + AI
  - Database + AI
  - Network IPS + AI
  - Storage + AI

**Heterogeneous Acceleration from Data Center to the Edge**

**Delivering Deterministic Performance & Low Latency**
AI Engine Application Performance & Power Efficiency

- **Image Classification (GoogleNet, <1ms)**: 20x
- **Massive MIMO Radio (DUC, DDC, CFR, DPD)**: 5x

- **AI Inference Compute**: Xilinx UltraScale+
- **5G Wireless Bandwidth**: Xilinx UltraScale+
- **Power Consumption**: 40% Less Power

© Copyright 2018 Xilinx
AI Engine

Architecture, Programming & Applications
AI Engine: Tile-Based Architecture

Non-Blocking Interconnect
Up to 200+ GB/s bandwidth per tile

Local Memory
Multi-bank implementation
Shared across neighbor cores

Cascade Interface
Partial results to next core

Data Mover
Non-neighbor data communication
Integrated synchronization primitives

ISA-based Vector Processor
Software Programmable (e.g., C/C++)

AI Vector Extensions
5G Vector Extensions

Data Mover

Interconnect

© Copyright 2018 Xilinx
AI Engine: Array Architecture

Modular and scalable architecture
• More tiles = more compute
• Up to 400 per device
  • Versal AI Core VC1902 device

Distributed memory hierarchy
Maximize memory bandwidth

Array of AI Engines
• Increase in compute, memory and communication bandwidth

Deterministic Performance & Low Latency
AI Engine: Processor Core

32-bit Scalar RISC Processor
- Scalar Unit
  - Scalar Register File
  - Scalar ALU
  - Non-linear Functions

Local, Shareable Memory
- 32KB Local, 128KB Addressable

Vector Processor
- 512-bit SIMD Datapath
- Vector Unit
  - Vector Register File
  - Fixed-Point Vector Unit
  - Floating-Point Vector Unit

Memory Interface
- AGU
  - AGU
  - AGU

Stream Interface
- Instruction Fetch & Decode Unit
- Load Unit A
- Load Unit B
- Store Unit

Instruction Parallelism: VLIW
- 7+ operations / clock cycle
  - 2 Vector Loads / 1 Mult / 1 Store
  - 2 Scalar Ops / Stream Access

Data Parallelism: SIMD
- Multiple vector lanes
  - Vector Datapath
  - 8 / 16 / 32-bit & SPFP operands

Highly Parallel

Up to 128 MACs / Clock Cycle per Core (INT 8)
Multi-Precision Support

### AI Data Types

<table>
<thead>
<tr>
<th>Data Type</th>
<th>MACs / Cycle (per core)</th>
</tr>
</thead>
<tbody>
<tr>
<td>32x32 SPFP Real</td>
<td>8</td>
</tr>
<tr>
<td>32x32 Real</td>
<td>8</td>
</tr>
<tr>
<td>32x16 Real</td>
<td>16</td>
</tr>
<tr>
<td>16x16 Real</td>
<td>32</td>
</tr>
<tr>
<td>16x8 Real</td>
<td>64</td>
</tr>
<tr>
<td>8x8 Real</td>
<td>128</td>
</tr>
</tbody>
</table>

### Signal Processing Data Types

<table>
<thead>
<tr>
<th>Data Type</th>
<th>MACs / Cycle (per core)</th>
</tr>
</thead>
<tbody>
<tr>
<td>32x32 Complex</td>
<td>2</td>
</tr>
<tr>
<td>32x16 Complex</td>
<td>4</td>
</tr>
<tr>
<td>16x16 Complex</td>
<td>8</td>
</tr>
<tr>
<td>16 Complex x 16 Real</td>
<td>16</td>
</tr>
</tbody>
</table>
Data Movement Architecture

Memory Communication

Dataflow Pipeline

Dataflow Graph

Streaming Communication

Non-Neighbor

Streaming Multicast

Cascade Streaming

Memory Interface

Stream Interface

Cascade Interface
AI Engine Integration with Versal ACAP

TB/s of Interface Bandwidth
- AI Engine to Programmable Logic
- AI Engine to NOC

Leveraging NOC connectivity
- PS manages Config / Debug / Trace
- AI Engine to DRAM (no PL req’d)
### AI Engine: Xilinx Reinvents Multi-Core Compute

**Traditional Multi-core (cache-based architecture)**

- **Block 0**
  - Core: red
  - Level 0 (L0)
  - Level 1 (L1)
  - Level 2 (L2)
  - DRAM: D0

- **Block 1**
  - Core: red
  - Level 0 (L0)
  - Level 1 (L1)

**Fixed, shared Interconnect**
- Blocking limits compute
- Timing not deterministic
- Data Replicated
  - Robs bandwidth
  - Reduces capacity

**AI Engine Array (intelligent engine)**

- **Dedicated Interconnect**
  - Non-blocking
  - Deterministic

- **Local, Distributed Memory**
  - No cache misses
  - Higher bandwidth
  - Less capacity required

© Copyright 2018 Xilinx
AI Engine Delivers High Compute Efficiency

> Adaptable, non-blocking interconnect
  > Flexible data movement architecture
  > Avoids interconnect “bottlenecks”

> Adaptable memory hierarchy
  > Local, distributed, shareable = extreme bandwidth
  > No cache misses or data replication
  > Extend to PL memory (BRAM, URAM)

> Transfer data while AI Engine Computes

Vector Processor Efficiency

<table>
<thead>
<tr>
<th>Block-based Matrix Multiplication (32x64) x (64x32)</th>
<th>1024-pt FFT/IFFT</th>
<th>Volterra-based forward-path DPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML Convolutions</td>
<td>95%</td>
<td>80%</td>
</tr>
</tbody>
</table>

Overlap Compute and Communication
## Versal ACAP Development Tools:
**Any Application, Any Developer**

<table>
<thead>
<tr>
<th>TOOLS</th>
<th>USER</th>
<th>SUPPORTED FRAMEWORKS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frameworks</td>
<td>AI and Data Scientists</td>
<td>Caffe, TensorFlow, mxnet, FFmpeg</td>
</tr>
<tr>
<td>New Unified Software Development Environment</td>
<td>Software Application Developers</td>
<td></td>
</tr>
<tr>
<td>Vivado Design Suite</td>
<td>Hardware Developers</td>
<td></td>
</tr>
</tbody>
</table>
Software Development Environment

- Unified development environment
  - Full chip programming

- SW programmable for whole application
  - Heterogeneous SW acceleration

- Full system simulation, debug & profiling
  - Software development experience

New Unified SW Development Environment

- Application e.g. C/C++
- Performance Constraints

- Processing Sub-system
- Programmable Logic
- AI Engines

- System Simulation
- Hardware
- System Debug & Profiling
AI Engine Programming Environment

Application (e.g. C/C++)

New Unified SW Development Environment

PS  PL  AI Engines

Full SW Programming Tool Chain
(Single-engine and Multi-engine)

IDE
Compiler
Debugger
Performance Analysis

Performance-Optimized Software Libraries
(Examples)

4G/5G/Radar Library
AI Library
Vision Library

Run-Time Software
(Examples)

Error Management
Memory Management
Boot + Configuration
Power/Thermal Management

IDE
Compiler
Debugger
Performance Analysis

4G/5G/Radar Library
AI Library
Vision Library

Error Management
Memory Management
Boot + Configuration
Power/Thermal Management

© Copyright 2018 Xilinx
AI Engine Programming Experience: Dataflow Model

1. User defines dataflow logic

   a → b → c → e

   a → d

2. User describes dataflow graph using C/C++ APIs

3. Compiler transparently manages placement & interconnect

   Physical Mapping to AI Engines

   Memory → Vector Core → Memory

   Memory → Vector Core → Memory

   Memory → Vector Core → Memory

   Memory → Vector Core → Memory
Frameworks for Any Developer

<table>
<thead>
<tr>
<th>TOOLS</th>
<th>USER</th>
<th>SUPPORTED FRAMEWORKS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frameworks</td>
<td>AI and Data Scientists</td>
<td>Caffe, TensorFlow, mxnet, FFmpeg</td>
</tr>
<tr>
<td>New Unified Software Development Environment</td>
<td>Software Application Developers</td>
<td></td>
</tr>
<tr>
<td>Vivado Design Suite</td>
<td>Hardware Developers</td>
<td></td>
</tr>
</tbody>
</table>

Domain Specific Architecture (e.g. AI Inference)

- Architecture Overlay
- Data Flow w/ Xilinx libraries
- Kernel Program Data Flow w/ user defined libraries

Target Domain Specific Architectures – No HW Design Experience Required
Accelerating AI Inference in the Data Center

1. User works in Framework of choice
   - Develop & train custom network
   - User provides trained model

2. Xilinx DNN Compiler implements network
   - Targets AI Inference Domain Specific Architecture
   - Quantize, merge layers, prune
   - Compile to AI Engines

3. Scalable across hardware targets
   - Start with Alveo today

Deep Learning Frameworks
- mxnet
- TensorFlow
- Caffe

Xilinx DNN Compiler

Xilinx AI Inference Domain Specific Architecture

Future Alveo Accelerator Cards
- Powered by Versal with AI Engine

Alveo U200 / U250
AI Inference on Versal ACAP

Convolutions

Fully Connected Layers

Pooling

Activations

- Video
- Genomics
- Storage
- Database
- Network IPS
- Risk modeling

*Figure credit: https://en.wikipedia.org/wiki/Convolutional_neural_network*
AI Inference Mapping on Versal ACAP

- Custom memory hierarchy
  - Buffer on-chip vs off-chip: Reduce latency and power
- Stream Multi-cast on AI interconnect
  - Weights and Activations
  - Read once: reduce memory bandwidth
- AI-optimized vector instructions (128 INT8 mults/cycle)
AI Engine Delivers Real-time Inference Leadership (75W Power Envelope)

Note: Versal device achieves 8X performance increase in 150W power envelope

(1) 12-nanometer T4 GPU device, Projected Batch=1 performance based on currently available vendor benchmarks
(2) 7-nanometer Versal AI Core Series VC1902 Device, 75W card power figures based on 2018.3 XPE power estimates, Latency <500us
Market Requirements and Trends: Wireless 5G

5G Complexity is 100X that of 4G
Still Evolving Standard

- Peak Data Rate: 20 Gbps
- User Experienced Data Rate: 100 Mbps
- Area Traffic Capacity: 10 Mbps/m²
- 3 Times Spectrum Efficiency
- Network Energy Efficiency: 100 Times
- Connection Density: 10⁶ per km² (1 per m²)
- 8 KPIs
  - Enhanced Mobile Broadband
  - Massive IoT
  - Ultra-reliable & Low latency
  - Mobility: 500 km/h
- Latency: 1 msec (Radio Interface)

New Technologies in 5G

- Massive MIMO
- Multiple antenna, frequency bands
- Changing functional partitioning

ETRI RWS-150029,
5G Vision and Enabling Technologies: ETRI Perspective 3GPP RAN Workshop
Phoenix, Dec. 2015
http://www.3gpp.org/ftp/tsg_ran/TSG_RAN/TSGR_70/Docs
5G Wireless on Versal ACAP

5G Wireless Infrastructure (i.e., base-station)

- Packet Processing
- Higher Layer Processing
- Baseband Processing
- Switching
- Beamform & MIMO
- Transforms
- Digital Radio
- ADC/DAC
- Analogue Radio
- Antenna Array

Mapping Example

Digital Radio with ADC/DAC

- DPD Update
- CPRI
- DUC
- DPD
- ADC/DAC

I/O Maps to PL

AI Engines

- DUC
- DPD

Programmable Logic

- CPRI

I/O

- ADC/DAC

Control Maps to PS

- DPD Update

Compute Maps to AI Engine

- Processing System

1: DUC: Digital Up Converter
2: DPD: Digital Pre-Distortion
3: Direc RF: ADC/DAC
4: CPRI: Common Public Radio Interface

© Copyright 2018 Xilinx
AI Engine Delivers 5X more 5G Wireless Compute

Xilinx Zynq UltraScale+ RFSoC

- 16x TXRX Antennae
- 5G
- CPRI Optics
- ~1M sLC
- ~4K DSP48

Partial Spectrum

Frequency Spectrum

Xilinx Versal with AI Engine

- 16x TXRX Antennae
- 5G
- CPRI/eCPRI Optics w. TSN

AI Engine:

- 100’s of 5G Wireless Accelerators

Full Spectrum

Frequency Spectrum

Enabling Single Chip Massive MIMO 16x16 800 MHz Radio
Wrapping Up
AI Engine: Accelerating AI Inference & Signal Processing

Software Programmable
- Frameworks & C/C++
- SW Compile, Debug & Deploy

Deterministic
- Max throughput w/ low latency
- Real-time inference leadership

Efficient
- Up to 8X compute density
- At ~40% lower power
See What AI Engine Can Do For You

Read the AI Engine White Paper
• Visit: www.xilinx.com/versal

Find out more about AI Engine
• Early Access Program Open Now
• Contact Your Xilinx Sales Representative