Key Take-Aways for This Talk

1. Xilinx Optimized Devices for AI/ML Edge Inference Acceleration

2. Whole Application Acceleration

3. AIE-ML: Optimized AIE Architecture for Machine Learning
AI/ML Edge Inference: Use-cases and Requirements

Automotive
ADAS
Automated Driving

Industrial Robotics
Vision guidance
Environmental awareness

Medical
Ultrasound
Endoscopy, Surgical Robotics

Industrial IoT
Industrial PCs
Smart Grid Controllers

Vision & Smart City
Machine Vision Camera
Edge AI Box

Aerospace & Defense
Unmanned Aerial Vehicles
Multi-Mission Payload Systems

Challenge: Delivering massive AI compute at low-latency and low-power
Versal™ AI Edge: Architecture Overview

ADAPTABLE TO MULTIPLE WORKLOADS
- Versal AI Core
- Versal Premium
- Versal AI Edge

COMPUTE ACCELERATION
- Scalar Engines
- Adaptable Engines
- Intelligent Engines

PLATFORM
- Software programmable NoC
- Platform Management Controller
- Dedicated Interfaces (e.g., PCIe, DDR)

Technology Node: TSMC 7nm FinFET
# Optimized Devices for AI/ML Edge Inference

## Key Features

- **AI Compute (INT8x4)**
  - Total AI compute includes AI Engines, DSP Engines, and Adaptable Engines
  - VE2302: 67 TOPS
  - VE2802: 479 TOPS

- **AI Compute (INT8)**
  - VE2302: 31 TOPS
  - VE2802: 228 TOPS

- **AIE-ML Tiles**
  - VE2302: 34
  - VE2802: 304

- **Adaptable Engines**
  - VE2302: 150K LUTs
  - VE2802: 521K LUTs

- **Processing Subsystem**
  - Dual-Core Arm® Cortex®-A72 Application Processing Unit / Dual-Core Arm Cortex-R5F Real-Time Processing Unit

- **Accelerator RAM (4MB)**
  - VE2302: ✓
  - VE2802: -

- **Total Memory**
  - VE2302: 172Mb
  - VE2802: 575Mb

- **32G Transceivers**
  - VE2302: 8
  - VE2802: 32

- **PCIe®**
  - VE2302: ✓
  - VE2802: ✓

- **Video Decode Unit (VDU)**
  - VE2302: ✓
  - VE2802: ✓

- **Power**
  - VE2302: 15-20W
  - VE2802: 75W

---

1: Total AI compute includes AI Engines, DSP Engines, and Adaptable Engines
2: Power Projections

© Copyright 2021 Xilinx
Inference Form Factor

Small Module Form Factor
High Degrees of Scalability, Thermal Challenges

HHHL PCIe Form Factor
Lowest Common Denominator
Deployments in Core, AI Edge Cloud

200W+ FH3/4L Form Factor
Some restrictions due power and form factor
Optimized **AIE-ML** for Machine Learning

**Optimized** the compute core for ML
- Doubled the multipliers, doubled INT8 performance
- Native support for INT4 and BFLOAT16

**Doubled** the data memory
- From 32kB to 64 kB
- Improved localization of data

**New** Memory Tile
- Up to 38 Megabytes across the AIE array
- Higher bandwidth memory access

---

1: AIE-ML delivers 2X INT8 compute, 4X INT4 compute, and 16X BFLOAT16 compute vs. AI Engine (per core)
2: Native 32-bit support in AI Engines only

---

1: AI Engine-ML delivers 2X INT8 compute, 4X INT4 compute, and 16X BFLOAT16 compute vs. AI Engine (per core)
2: Native 32-bit support in AI Engines only

© Copyright 2021 Xilinx
### Xilinx AIE Tile: Architecture Features

<table>
<thead>
<tr>
<th></th>
<th>AIE Tile</th>
<th>AIE-ML Tile</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Target Markets</strong></td>
<td>Wireless 5G, AI/ML inference, A&amp;D</td>
<td>AI/ML inference</td>
</tr>
<tr>
<td><strong>Compute ((\text{Mults} / \text{tile}))</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BFLOAT16</td>
<td>—</td>
<td>▶️ New ▶️</td>
</tr>
<tr>
<td>INT8</td>
<td>128</td>
<td>▶️ 2X ▶️</td>
</tr>
<tr>
<td>INT4</td>
<td>—</td>
<td>▶️ New ▶️</td>
</tr>
<tr>
<td><strong>Tile Local Data Memory</strong></td>
<td>32 KB</td>
<td>▶️ 2X ▶️</td>
</tr>
<tr>
<td><strong>AIE Array Interconnect B/W</strong></td>
<td>1X</td>
<td>▶️ 1X ▶️</td>
</tr>
<tr>
<td><strong>Compression and Sparsity</strong></td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Scratchpad On-Chip Memory</strong></td>
<td>PL uRAM</td>
<td>AIE Memory (512KB/tile)</td>
</tr>
</tbody>
</table>
AIE-ML Tile Architecture

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

Local Memory
Multi-bank implementation
Shared across neighbor cores

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

Cascade Interface
Partial results to next core

Data Mover
Non-neighbor data communication
Integrated synchronization primitives

Data Mover
AI Vector Extensions

Interconnect
Flexible Dataflow: AIE-ML Array Examples

**DATA REDUCTION**

\[ C = (A_0 \times B_0) + (A_1 \times B_1) + (A_2 \times B_2) + (A_3 \times B_3) \]

**AIE-ML IMPLEMENTATION**

- **Compute Single Output Matrix**
  - \( C_0 = (A_0 \times B_0) \)
  - \( C_1 = C_0 + (A_1 \times B_1) \)
  - \( C_2 = C_1 + (A_2 \times B_2) \)
  - \( C_3 = C_2 + (A_3 \times B_3) \)

**DATA MULTICASTING**

**Compute Multiple Output Matrices**

- \( C_0 \)
- \( C_1 \)
- \( \cdots \)
- \( C_N \)

**AIE-ML IMPLEMENTATION**

- Weights
- \( N \)
- \( N \)

© Copyright 2021 Xilinx
Flexible Dataflow: Device-level Examples

- DDR ↔ Scratchpad using Programmable Logic
- DDR ↔ Scratchpad using NoC
- DDR ↔ AIE-ML Tiles using NoC
Compression to DDR: Acceleration of activation Spill/Restore

- Reduction in external memory bandwidth using compression
  - Quantized int8 weights, int8 activations
  - HD images require tiling, with spill/restore to DDR
    - Without compression ~56GB/s
    - With compression ~29GB/s

HD ResNet-50 Intermediate Feature Map Compression Factor

- 58.9%
- 52.3%
AIE-ML Delivers High Compute Efficiency

Vector Processor Efficiency

*Peak Kernel Theoretical Performance*

- 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 AIE-ML computes in parallel
AIE-ML: Multi-Core Architecture Comparison

Traditional Multi-core (cache-based architecture)

- **Fixed, shared Interconnect**
  - Timing not deterministic

- **Data Replicated**
  - Additional bandwidth
  - Reduces capacity

AIE-ML Array (intelligent engine)

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

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

---

© Copyright 2021 Xilinx
AIE Architecture: Source Code Portability

- Backward compatible
- Evolving
  New devices can add new functionality
- Efficient
  Applications can be optimized to maximize compute efficiency

Portable common high-level API across the AIE architecture
Vitis AI: Full AI Software Stack Support

Frameworks
- PyTorch
- TensorFlow
- Caffe
- tvm

Vitis AI models

Vitis AI development kit
- AI Optimizer
- AI Quantizer
- AI Compiler
- AI Profiler
- AI Library

Platforms
- CNN DPU
- LSTM DPU
- MLP DPU

Deep Learning Processing Unit (DPU)

Xilinx runtime library (XRT)

© Copyright 2021 Xilinx
Smart World Application: Use-Case Mapping

Smart Retail

Smart Hospital

Smart City

Accelerating the Whole Application: From Video Decode to Decision Making
Versal AI Edge in ADAS and Automated Driving

Accelerating the Whole Application From Sensor to AI to Real-Time Control

- Adaptable Engines for sensor fusion and pre-processing
- Intelligent Engines for signal conditioning and low latency AI
- Scalar engine for decision making and vehicle control

1: Demonstrates capabilities of architecture, not representing a single chip AD system

© Copyright 2021 Xilinx
Performance/Watt Results

**Autonomous System or Edge Aggregation**

3.3X Performance/Watt

- Jetson AGX Xavier (15W Mode)\(^1\)
- Versal AI Edge (VE2302)

**CPU Accelerator**

4.2X Performance/Watt

- Jetson AGX Xavier (MAX N-Mode)\(^2\)
- Versal AI Edge (VE2802)

---

2. Jetson AGX Xavier MAX N-Mode and VE2802 represent the highest performing device configuration in their respective portfolios. Jetson AGX Xavier device power estimated by subtracting published memory and IO power from total module power.
Summary

Xilinx Optimized Devices for AI/ML Edge Inference Acceleration
- Built using Xilinx 7nm Versal™ Architecture
- Versal AI Edge VE2302: 15-20W, small form factor
- Versal AI Edge VE2802: 75W, HHHL PCIe form factor

Whole Application Acceleration
- Edge Aggregation & Autonomous Driving
- Smart City for Edge Server Accelerator

AIE-ML: Optimized AIE Architecture for Machine Learning
- Major improvements in TOPs/W
- ML-optimized features (e.g., optimized datatypes and memory hierarchy)
Thank You