Versal™ AI Edge Series Announcement

Rehan Tahir, Senior Product Line Manager
What’s Happening at the Edge

The Edge

Low Latency

AI Compute

Low Power

Safety and Security
Hypergrowth at the Edge

“Edge computing … solves for weaknesses of the cloud”¹

Edge AI chipset opportunity is 3X that of data center - $65B in 2025²

¹: Gartner, “2021 Strategic Roadmap for Edge Computing”, November 2020

Deep learning chipset revenue, enterprise vs. edge, world markets: 2019–25

© Copyright 2021 Xilinx
Now Bringing Versal ACAPs to the Edge

- Versal™ ACAPs first introduced breakthrough compute for the cloud and network
- Now ‘miniaturizing’ this technology for performance/watt at the edge
New Versal™ Platform for Intelligence at the Edge
Versal™ AI Edge: Intelligence Unleashed

- 4X AI Performance/Watt vs. GPUs¹ with Innovations in AI Engines and Memory Hierarchy

- 10X Compute Density² with Highest Levels of Safety and Security

- World’s Most Scalable and Adaptable Platform for Edge and Endpoint

---

² Compared to Zynq® UltraScale+™ MPSoCs
4X AI Performance/Watt
Proven AI Engine Architecture

Array of Compute Core
- Flexible compute: fixed- & floating-point vector processors
- HW adaptable to evolving algorithms

Tightly Coupled Memory
- Cache-less memory hierarchy
- Maximizes bandwidth, ensures determinism & low latency

Flexible Interconnect
- Connect any tile to any tile for custom microarchitecture
- High bandwidth

Architected for Adaptability, Low Power, and Low Latency
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 AI engine array
- Higher bandwidth memory access

Delivering 4X ML Compute at ½ the Latency\(^1\)

---

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
AIE-ML Complements AI Engines for Diverse Workloads

AI Engine Architecture
Balanced for ML & DSP

4X ML Compute at ½ the Latency

AI Engine-ML Architecture
Optimized for ML

AI Engine

Advanced Signal Processing
BEAMFORMING, RADAR PROCESSING, HIGH PERF. COMPUTING

Machine Learning
AI Engine-ML
CNN, RNN, MLP

AI Engine-ML delivers 2X INT8 compute, 4X INT4 compute, and 16X BFLOAT16 compute vs. AI Engine (per core)

Native 32-bit support in AI Engines only
Innovations in Memory Hierarchy: Accelerator RAM

4MB of On-Chip RAM for Massive Bandwidth
Avoid DDR to store AI compute data or safety-critical code

Part of the Adaptable Memory Hierarchy
Select the right memory for bandwidth requirement

35GB/s Memory Bandwidth

12.8GB/s Memory Bandwidth

3200
LPDDR4-4266

PCIe® w/DMA & CCIX

DDR4 / LPDDR4

32Gbps SerDes

40G Multirate Ethernet

MIPI

LVDS

GPIOD

© Copyright 2021 Xilinx
Up to 4X Performance/Watt vs. GPUs

Intelligent Edge Sensor

Jetson Xavier NX 1
Versal AI Edge (VE2102)

1.9X Performance/Watt

Jetson AGX Xavier
(V15 Mode) 2
Versal AI Edge (VE2302)

3.3X Performance/Watt

Jetson AGX Xavier
(MAX N-Mode) 3
Versal AI Edge (VE2802)

4.2X Performance/Watt

Autonomous System or Edge Aggregation

ResNet50 224x224, batch=1

CPU Accelerator

ResNet50 224x224, batch=1

1: Jetson NX Xavier: https://mlcommons.org/en/inference-edge-10, batch size not provided
3: Jetson AGX Xavier MAX N-Mode and Versal™ VE2802 ACAP represent the highest performing device configuration in their respective portfolios
Jetson Xavier device power estimated by subtracting published memory & I/O power from total module power
All charts are normalized
10X Compute Density with Highest Levels of Safety and Security
**10X Compute Density: Level 3 Semi-Automated Driving**

### Previous Gen Adaptive SoC

- **Compute**: 6x cameras (2MP, 4MP) + AI (4TOPs)
- **Area**: 3 devices = 1,259mm²
- **Power**: ZU3(6W) + ZU5(10W) + Z-7020(5W)

### Versal™ ACAP

- **Compute**: 6x cameras (2MP, 8MP) + AI (17.4TOPs)
- **Area**: 1 device = 529mm²
- **Power**: ~20W

---

*Power levels are typical, approximate, and estimated at room temp*
AI as Part of the Whole Application

**Sense**
Adaptable Engines

- Cameras
- Radar
- LiDAR
- V2X
- Other Sensors

**Think**
Intelligent Engines

- Localization
- Sensor Fusion
- Mapping
- Perception
- Object Classification
- Path/Motion Planning

**Act**
Scalar Engines

- Vehicle Control
- Driver Interaction

- Braking
- Steering
- Other Actuators
- Displays
- Driver Monitor
- Other HMI

Human Machine Interface
Whole Application Acceleration for Real-Time Systems
From Sensor to AI to Real-Time Control

EXECUTION TIME

Sense | Think (AI) | Act

Sense | Think | Act | Actuator Response
Versal™ AI Edge ACAP 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
- Scalable compute from edge sensor to domain controller

1: Diagram demonstrates capabilities of architecture; does not represent a single chip AD system
Fully Automotive-Qualified and Safety Certified

Architected to Meet Stringent ISO 26262 Requirements
Supporting Multiple Safety Standards

Versal™ AI Edge
ACAP

ISO 26262
Automotive

IEC 61508
Safety across All Industries

DO-254/178
Avionics HW/SW

IEC 61511
Process Industry

IEC 61800
Electrical Drives

ISO 13849
Machinery Control

IEC 62061
Machinery

EN 60601
Medical
Collaborative Robotics: AI-Based Systems Need to be Safe and Secure

- **Real-Time Precision and Control to Augment AI**
  Deterministic response, AI to navigate unpredictable movement of workers

- **Environmental Awareness and Perception**
  Sensor fusion for perception, self-learning to improve capabilities over time

- **Predictive Maintenance**
  Analyze sensor data for actionable insights to reduce downtime

- **Safety and Security are Connected Matters**
  Cyber-attack creates safety and data privacy risks, robotic systems require IEC 62443 compliance
Whole Application Acceleration for Collaborative Robotics

Robotic Perception Systems for Real-Time Control, Safety Critical, and Predictive Maintenance

- Adaptable Engines for perception, control/networking, navigation
- AI to augment control for dynamic execution, predictive maintenance
- Scalar Engines for cybersecurity (IEC 62443), safety control, UI
AI-Enabled Multi-Mission Payloads for UAVs

*AI with Software Defined Radio (SDR), Signal Intelligence (SIGINT), Image/Video Processing*

---

- **Vision AI for Real-Time Analysis and Response**
  - Autonomous flight control, optimize navigation paths

- **Cognitive RF**
  - Optimizing radio communication and protecting against malicious intrusion

- **Diverse and Emerging Forms of AI**
  - AI is rapidly evolving in tactical applications and vendors will need to adapt over time

---

Need AI Compute in Limited Size, Weight, and Power (SWaP) and Thermal Envelope
Versal AI Edge for Unmanned Aerial Vehicles

- Adaptable Engines for sensor fusion and pre-processing
- Intelligent Engines for low power, low latency AI and signal conditioning
- Scalar Engines for command and control
- Ruggedized packaging and military-temp grade (XQ)
Versal ACAP Development Experience for All Developers

- **HW Developer**
- **SW Developer**
- **Data Scientist**

**Versal™ AI Edge ACAP**

- OS & Embedded Run-Time
- Custom HW
- HW IP & Accelerated Libraries
- HW Accelerated Libraries

- Scalar Engines
- Adaptable Engines
- Intelligent Engines

C, C++, Python

License: © Copyright 2021 Xilinx
Market-Specific Application Stacks
Examples for Automotive, Robotics, and Multi-Mission Payload Applications

- One platform with market-specific libraries, frameworks, and ecosystem to enable all developers
- Following industry standards for developing safety critical software on silicon

### Automotive Applications
- Caffe
- TensorFlow
- PyTorch

### Robotics Applications
- Gazebo
- ROS

### Multi-Mission Payload: SDR, SIGINT, Image/Video
- TensorFlow
- PyTorch

**Xilinx Runtime (XRT)**

**Scalar Engines**

**Adaptable Engines**

**Intelligent Engines**

**Versal™ AI Edge ACAP**
World’s Most Adaptable and Scalable Edge Platform
**Adaptability:** From Domain Specific Architectures (DSAs) to Dynamic Function Exchange

**DSAs for Diverse Platform Requirements**
- Implement custom AI, vision, sensor strategies
- Design for different safety and security targets
- One platform for diverse end-customers' requirements

**Hardware/Software Over-the-Air Updates**
- Update your AI accelerator or fusion algorithms
- Future proof for emerging security threats
- Avoid recalls or costly re-deployment

**Dynamic Function Exchange (DFx)**
- Swap functionality in milliseconds
- Available in Adaptable Engines, DSP, AI Engines
- Fewer system components → reduce power and cost
Dynamic Function Exchange (DFx) in Automotive

Swap Functionality in Milliseconds

Dynamic Regions (Engines, Integrated Cores, I/O)

Drive Mode (Lane Departure Warning)

Low Speed Mode (Parking Assist)

Post-Drive Mode (Dog Left Behind)

Fewer Devices to Reduce System-Wide Power and Cost
# Scale from Edge Sensor to CPU Accelerator

## Lead Device

<table>
<thead>
<tr>
<th>Lead Device</th>
<th>Intelligent Edge Sensor &amp; End Point</th>
<th>Accelerator</th>
</tr>
</thead>
<tbody>
<tr>
<td>VE2002</td>
<td>IntEgo V11, VE2102</td>
<td>VE2202</td>
</tr>
<tr>
<td>VE2102</td>
<td>IntEgo V12, VE2202</td>
<td>VE2302</td>
</tr>
<tr>
<td>VE2202</td>
<td>IntEgo V13, VE2302</td>
<td>VE2602</td>
</tr>
<tr>
<td>VE2302</td>
<td>IntEgo V14, VE2602</td>
<td>VE2602</td>
</tr>
<tr>
<td>VE2602</td>
<td>IntEgo V15, VE1752</td>
<td>VE2802</td>
</tr>
<tr>
<td>VE1752</td>
<td>IntEgo V16, VE2802</td>
<td>VE2802</td>
</tr>
<tr>
<td>VE2802</td>
<td>IntEgo V17, VE2802</td>
<td>VE2802</td>
</tr>
</tbody>
</table>

## Engines

<table>
<thead>
<tr>
<th>Engines</th>
<th>VE2002</th>
<th>VE2102</th>
<th>VE2202</th>
<th>VE2302</th>
<th>VE2602</th>
<th>VE1752</th>
<th>VE2802</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total AI Compute (INT4)</td>
<td>14 TOPS</td>
<td>22 TOPS</td>
<td>47 TOPS</td>
<td>67 TOPS</td>
<td>256 TOPS</td>
<td>166 TOPS</td>
<td>479 TOPS</td>
</tr>
<tr>
<td>Total AI Compute (INT8)</td>
<td>7 TOPS</td>
<td>10 TOPS</td>
<td>21 TOPS</td>
<td>31 TOPS</td>
<td>120 TOPS</td>
<td>124 TOPS</td>
<td>228 TOPS</td>
</tr>
<tr>
<td>AIE / AIE-ML&lt;sup&gt;1&lt;/sup&gt;</td>
<td>8</td>
<td>12</td>
<td>24</td>
<td>34</td>
<td>152</td>
<td>304</td>
<td>304</td>
</tr>
<tr>
<td>Adaptable Engines</td>
<td>20K LUTs</td>
<td>37K LUTs</td>
<td>105K LUTs</td>
<td>150K LUTs</td>
<td>375K LUTs</td>
<td>449K LUTs</td>
<td>521K LUTs</td>
</tr>
</tbody>
</table>

## Processing Subsystem

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

## RAM

<table>
<thead>
<tr>
<th>RAM</th>
<th>VE2002</th>
<th>VE2102</th>
<th>VE2202</th>
<th>VE2302</th>
<th>VE2602</th>
<th>VE1752</th>
<th>VE2802</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accelerator RAM (4MB)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Total Memory</td>
<td>95Mb</td>
<td>103Mb</td>
<td>156Mb</td>
<td>172Mb</td>
<td>554Mb</td>
<td>253Mb</td>
<td>575Mb</td>
</tr>
<tr>
<td>32G Transceivers</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>8</td>
<td>8</td>
<td>32</td>
</tr>
<tr>
<td>PCIe®</td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PCIe + CCIX</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Estimated Power</td>
<td>6–9W</td>
<td>7–10W</td>
<td>15–20W</td>
<td>~20W</td>
<td>50–60W</td>
<td>50–60W</td>
<td>75W</td>
</tr>
</tbody>
</table>

<sup>1</sup>: VE2xx based on AIE-ML, VE1752 device base based on AIE

© Copyright 2021 Xilinx
# The Only Edge AI Platform that Scales from Sensor to Accelerator on a Single Architecture\(^1,2\)

<table>
<thead>
<tr>
<th>Power Range</th>
<th>0–10TOPS</th>
<th>10–25TOPS</th>
<th>25–75TOPS</th>
<th>100+ TOPS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Versal CV-800</strong></td>
<td>![Symbol]</td>
<td>![Symbol]</td>
<td>![Symbol]</td>
<td>![Symbol]</td>
</tr>
<tr>
<td><strong>NVIDIA (Jetson)</strong></td>
<td>![Symbol]</td>
<td>![Symbol]</td>
<td>![Symbol]</td>
<td>![Symbol]</td>
</tr>
<tr>
<td><strong>NVIDIA (T4)</strong></td>
<td>![Symbol]</td>
<td>![Symbol]</td>
<td>![Symbol]</td>
<td>![Symbol]</td>
</tr>
<tr>
<td><strong>NVIDIA (ORIN)</strong></td>
<td>![Symbol]</td>
<td>![Symbol]</td>
<td>![Symbol]</td>
<td>![Symbol]</td>
</tr>
<tr>
<td><strong>Mobileye</strong></td>
<td>![Symbol]</td>
<td>![Symbol]</td>
<td>![Symbol]</td>
<td>![Symbol]</td>
</tr>
<tr>
<td><strong>Texas Instruments</strong></td>
<td>![Symbol]</td>
<td>![Symbol]</td>
<td>![Symbol]</td>
<td>![Symbol]</td>
</tr>
<tr>
<td><strong>Qualcomm Snapdragon®</strong></td>
<td>![Symbol]</td>
<td>![Symbol]</td>
<td>![Symbol]</td>
<td>![Symbol]</td>
</tr>
<tr>
<td>Qualcomm® Cloud AI 100</td>
<td>![Symbol]</td>
<td>![Symbol]</td>
<td>![Symbol]</td>
<td>![Symbol]</td>
</tr>
<tr>
<td><strong>NXP (iMX8)</strong></td>
<td>![Symbol]</td>
<td>![Symbol]</td>
<td>![Symbol]</td>
<td>![Symbol]</td>
</tr>
<tr>
<td><strong>Renesas</strong></td>
<td>![Symbol]</td>
<td>![Symbol]</td>
<td>![Symbol]</td>
<td>![Symbol]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Power Range</th>
<th>1–100 Watts</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Versal CV-800</strong></td>
<td>![Symbol]</td>
</tr>
<tr>
<td><strong>NVIDIA (Jetson)</strong></td>
<td>![Symbol]</td>
</tr>
<tr>
<td><strong>NVIDIA (T4)</strong></td>
<td>![Symbol]</td>
</tr>
<tr>
<td><strong>NVIDIA (ORIN)</strong></td>
<td>![Symbol]</td>
</tr>
<tr>
<td><strong>Mobileye</strong></td>
<td>![Symbol]</td>
</tr>
<tr>
<td><strong>Texas Instruments</strong></td>
<td>![Symbol]</td>
</tr>
<tr>
<td><strong>Qualcomm Snapdragon®</strong></td>
<td>![Symbol]</td>
</tr>
<tr>
<td>Qualcomm® Cloud AI 100</td>
<td>![Symbol]</td>
</tr>
<tr>
<td><strong>NXP (iMX8)</strong></td>
<td>![Symbol]</td>
</tr>
<tr>
<td><strong>Renesas</strong></td>
<td>![Symbol]</td>
</tr>
</tbody>
</table>

1: Shown in INT8 TOPS
2: Based on published sources

© Copyright 2021 Xilinx
Scalable for Different Requirements and Product Features

Scale for Varying Levels of Compute and Safety
- Scale number of sensors, AI compute, vision and video processing
- e.g., Scale from Level-3 ADAS to Level-5 automated drive on a single platform

Scale a Low-End to High-End End-Product Portfolio
- Design once, scale with same tools, SW, ecosystem, safety certification
- Scale for different price points and capabilities

Explore Distributed vs. Centralized Architectures
- “Load Balance” across the system
- Shift compute from edge sensor to central compute across a single system
How Customers Can Get Started
Availability

- Documentation Available Now
- Tools Available in 2\textsuperscript{nd} Half of ‘21
- ES & Production Silicon in 1\textsuperscript{st} Half ‘22
- Versal\textsuperscript{TM} AI Edge ACAP Eval Kit in 2\textsuperscript{nd} Half ‘22
Start Prototyping Now

Start Now with Versal AI Core ACAP VCK190 Evaluation Kit
Migrate Later to Versal AI Edge Device

Evaluate Key Blocks in Versal™ AI Edge
Leverage Vitis™ Accelerated Libraries
Breadth of Interfaces for System Testing
System-Design Methodology Guides
Guided Flows in Vitis and Vivado® Tools

www.xilinx.com/vck190
Versal AI Edge ACAP: Intelligence Unleashed

*From Sensor to AI to Real-Time Control*

- 4X AI Performance/Watt vs. GPUs\(^1\) with Innovations in AI Engines and Memory Hierarchy

- 10X Compute Density\(^2\) with Highest Levels of Safety and Security

- World’s Most Scalable and Adaptable Platform for Edge Systems

---

2: Compared to Zynq® UltraScale™ MPSoCs

© Copyright 2021 Xilinx
Thank You