Xilinx First 7nm Device: Versal AI Core (VC1902)

Hot Chips 31, Aug 20, 2019

Sagheer Ahmad, Sridhar Subramanian
Vamsi Boppana, Shankar Lakka, Fu-Hing Ho, Tomai Knopp, Juanjo Noguera, Gaurav Singh, Ralph Wittig
@Xilinx Inc
Agenda

> Versal Overview
  > What is Versal
  > Versal series overview
  > First Versal device

> Key Blocks & Features
  > NOC, Memory, Interfaces, IOs, and SerDes
  > PS/PMC, Security, Config and Debug
  > Programmable Logic

> AI Engine
  > Array, Core
  > Compute, memory, and throughput
  > Benchmarks and use-cases performance
# Xilinx Device Categories

<table>
<thead>
<tr>
<th>DEVICE CATEGORY</th>
<th>FPGA</th>
<th>SoC</th>
<th>ACAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>FEATURED PRODUCTS</td>
<td>Spartan</td>
<td>Zynq-7000</td>
<td>Versal</td>
</tr>
<tr>
<td></td>
<td>Artix</td>
<td>Zynq UltraScale+ MPSoC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kintex</td>
<td>Zynq UltraScale+ RFSoC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Virtex</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**ACAP = Adaptive Compute Acceleration Platform**

**Versal = First ACAP device series**
Versal Series Overview

Compute Engines
- Scalar Processors in every device
- Enhanced Programmable Logic
- New AI and enhanced DSP Engines

NoC and Memory
- High BW Network-on-Chip
- Hardened [LP]DDR4/5, and HBM

High-Speed Interfaces
- PCIe & CCIX up to Gen5
- Ethernet MAC up to 600Gbps

SerDes and RF
- SerDes Up to 112G PAM4
- Integrated ADC/DAC
First Versal AI Core (VC1902)

- Process Technology: TSMC 7nm FF
- # Transistors: 37B
- On-die memory: 855Mb
- # AI engine cores: 400
- # IOs: 785
- # SerDes: 44

Shipping to Early Customers
Versal NoC (Network-on-Chip)

**Packetized High-speed NoC**
- All of SoC building blocks & PL connected via NoC
- Packetized w/ VCs & end-to-end ECC protection

**Highly Configurable & Scalable**
- Configurable topology, ports, routing, and QoS
- Compiler to generate use-case specific routing, QoS, ...
- NoC extends for die-to-die connectivity

**Clocking & Power Management**
- Clock forwarding to minimize clock jitter & power
- Aggressive clock-gating & data bypass

Data movement efficiency critical for compute acceleration
Memory Subsystem and IO

Unified Memory Subsystem
- Unified memory subsystem, but can be customized
- Transaction reordering & QoS for multiple traffic types

256b DDR w/ 4x 64b or 8x 32b channels
- Optimized for 64b or 32b memory channels
- 32b granularity more efficient for some use-cases
- DDR4 up to 3200 and LPDDR4 up to 4266 Mbps

Parallel IOs
- 644x high performance XPIOs for DDR, MIPI, …
- 137x high density multiprotocol IOs for up to 3.3v
PCle, Ethernet, and SerDes

6x PCle Gen4
- Up to Gen4 x16 with End-point & Root-port
- Smart storage or IO-Hub accelerator

4x 100G Multi-rate Ethernet
- Multi-rate (100/50/40/25/10Gbps) Ethernet
- MACs with RS-FEC. 1588 support.

SerDes
- 44x 32Gbps multi protocol transceivers
- Supports 100+ protocol/rate combinations
Versal CCIX

**CCIX and PCIe (CPM)**
- 2nd generation of CCIX coherent accelerator link

**Coherent Home-node and L2 cache**
- Home-node for coherent peer processing
- L2 for caching capability for PL accelerator kernels

**CCIX ESM (Extended Speed Mode)**
- Supports PCIe Gen4 x16 for CCIX & PCIe
- Supports up to CCIX 25Gbps 2x8

**Coherent Load/Store Memory Semantics**
Versal Processor System (PS)

PS in all Versal devices
- 3rd generation of PS integration
- First generation with PS in all devices
- Host for embedded, control for acceleration

Dual-core A72 APU
- 2x Cores with 1MB L2 Cache with ECC
- Coherency and virtualization support

Dual core R5 RPU w/ lockstep
- 2x Cores with 256KB TCM & 256KB OCM
- ASIL-C(D) capable functional safety
Versal PMC and Security

Platform Management Ctrl (PMC)

- Gateway for Boot/Config, Security, Power mgmt, ...
- Dual-core triple redundant MicroBlaze subsystem
- Crypto accelerator engines (RSA, ECDSA, AES, SHA)

Security & Monitors

- Hardware RoT with authentication and encryption
- Key storage & management including PUF support
- Distributed Voltage & Temp monitors

50Gbps Configuration Interface

- Typical PL kernel configuration in sub 10msec
- 8x faster PL configuration time per config-bit

10Gbps Debug & Trace Interface

- New HSDP (High-Speed Debug Port) serial interface.
- 100x faster than JTAG for debug & trace
Versal Programmable Logic (PL)

900K LUTs (2M LC) and 1.8M Flops
- 4x Larger CLB (8 LUTs → 32 LUTs)
- 16 Flip-Flops → 64 Flip-Flops
- Increased local routing (→ lower global routing)
- Imux registers for pipeline & time borrowing

158Mb of URAM & BRAM
- Distributed URAM and BRAM columns
- Customizable memory hierarchy
- 50% lower power than previous gen
Versal Programmable Logic DSP

Versal DSP58
- FP32/16 floating point
- INT8/16/24 and CMPLX18 fixed point

1968x DSP58
- Distributed DSP columns
- 2.8TFLOP/s FP32 Peak
- 11.8TOP/s (INT8) Peak
Versal Silicon

**Engines**

**Scalar Engines**
- Boots 64-bit Linux
- A72, R5, PMC all running

**AI Engines**
- All 400 AI Engine Tiles functional

**Programmable Logic**
- PL state machines
- Data across NoC & AI Engines

**Network-on-Chip (NoC)**
- Running error-free @ 3200 Mb/s
- Arbitration across engines

**Interfaces & IP**

**PCle & CCIX**
- Passed Gen3 compliance
- Clean link at Gen4 x4

**DDR Memory**
- DDR4 running at 3200Mb/s
- LPDDR4 running at 4266Mb/s

**32Gb/s Transceivers**
- Passed backplane PRBS31
- 7.16ps total &180fs random jitter

**100G Multi-Rate MAC**
- Passed internal loopback without GTY at full speed
AI Engine: Array

400 AI Engine Tiles
- 133 TOPs (INT8) Peak

Non-blocking Interconnect Mesh
- 20Tbps row x-sectional bandwidth
- 10 32-bit channels per column and 8 per row

Distributed Memory Hierarchy
- 12.5MB distributed L1 memory
- Multi-bank local memory shared w/ neighboring tiles
- Distributed DMA
AI Engine: Core

32b Scalar RISC Processor
- 2 Scalar Ops / Stream Access

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

Vector Processor
- 512-bit SIMD Datapath
- 2 Vector Loads / 1 Mult / 1 Store
- vec128int8
- vec8fp32

7+ Ops per cycle VLIW
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</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>
AI Engine Memory Hierarchy

- DRAM
- L2 SRAM
- L1 SRAM

Flexible XBAR
Adaptable L1 NOC with DMA
AI Engine Memory Hierarchy

- DRAM
- L2 SRAM
- L1 SRAM
- Multicast / Broadcast
AI Engine Memory Hierarchy

- 1.6 TB/s (128 kByte 4 Core Cluster) L1 SRAM
- 38 TB/s
- 12.5 MByte (128 kByte 4 Core Cluster) L1 SRAM
- 16.3 MByte L2 SRAM
- 102 GB/s
- > 64 GByte DRAM

© Copyright 2019 Xilinx
AI Engine: 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)

- Distributed DMA for overlapping Compute and Comm.

Vector Processor Efficiency

<table>
<thead>
<tr>
<th></th>
<th>95%</th>
<th>80%</th>
<th>98%</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML Convolutions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FFT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DPD</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- ML Convolutions: Block-based Matrix Multiplication (32×64) x (64×32)
- FFT: 1024-pt FFT/iFFT
- DPD: Volterra-based forward-path DPD
AI Engine: Performance Benchmark

GoogleNet Inference Performance
(sub 2ms latency)

<table>
<thead>
<tr>
<th>UltraScale+ series</th>
<th>Versal AI Core Series*</th>
</tr>
</thead>
<tbody>
<tr>
<td>4087</td>
<td>29250</td>
</tr>
</tbody>
</table>

ResNet-50 Inference Performance

<table>
<thead>
<tr>
<th>UltraScale+ series</th>
<th>Versal AI Core Series*</th>
</tr>
</thead>
<tbody>
<tr>
<td>2043</td>
<td>11812</td>
</tr>
</tbody>
</table>

*Versal AI Core (VC1902) projected performance
Accelerating AI Applications on Versal

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

- Arm Dual-Core Cortex-A72
- Arm Dual-Core Cortex-R5

- 463 x 32KB + 967 x 4KB of RAM
- Any-to-Any Connectivity
- Custom Memory Hierarchy
- TB/s of Bandwidth PL-to-AI Engine

Heterogeneous Acceleration from Data Center to the Edge

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

Deterministic Performance & Low Latency
Accelerating 5G Wireless on Versal

5G Wireless Infrastructure

Packet Processing
Higher Layer Processing
Baseband Processing
Switching
Beam Forming & MMIO

Digital Radio w/ ADC/DAC

CPRI → DUC → CFR → DPD → ADC/DAC

Versal

Programmable Logic (PL)

Processor System (PS) : APU

DUC: Digital Up Converter
DPD: Digital Pre-Distortion
CPRI: Common Public Radio Interface

30.72 MHz 30.72 MHz 61.44 MHz
491.52 MHz
122.88 MHz
614.4 MHz

N_{HB3} = 43  N_{HB4} = 21

N_{HB1} = 23  N_{HB2} = 11

N_{HB5} = 41

N_{HB} = 43

AI Engine Array

Peak Detect and Scale Find

Peak Detect and Scale Find

DPD Update

CABS

Coefficient to LUT Conversion

Memory Active/Shadow

DPD LUTs

Gain

Frequency Domain Measurements

Power Spectrum Estimate

Gain

Coefficients

Coefficients

DPD LUTs

9x9 DPD kernel

DPD Filter 1/4
DPD Filter 2/4
DPD Filter 3/4
DPD Filter 4/4

Digital Pre-distortion

Crest Factor Reduction

Shaping Up-sample Heterodyne

© Copyright 2019 Xilinx

>> 24
Software Programmable

1 Design

C/C++

Frameworks

mxnet

TensorFlow

Caffe

4G/5G/Radar Library

AI Library

Vision Library

2 Compile

AI Engine Compiler

Run

Programming Abstraction Levels

Architecture Overlay

Data Flow
w/ Xilinx libraries

Kernel Program
Data Flow w/ user defined libraries

1

2

3

>> 25
Unified Tool Chain for Programming

- **Application**
- **Performance & Partitioning Constraints**

**SDK**
- **System-Level Performance Analysis (using core profiling)**
- **System-Level Debugger (using core debugger)**

**Xilinx SDK: Eclipse GUI**
- **User-Directed System Partitioner**

**Base Platform**

**PS**
- **ARM C Compiler**

**AI Engine Array**
- **AI Engine Compiler**

**PL**
- **Vivado**
  - **HLS**
  - **RTL IP**

**Binaries & Bitstream**

**System-Level Performance Analysis**
- Using core profiling

**System-Level Debugger**
- Using core debugger

**System-C Virtual Simulation Platform**
- QEMU
- Core ISS

**Targets**

**Versal Device**

© Copyright 2019 Xilinx
Summary

Versal is the first generation of ACAP device
- ACAP is a new class of device from Xilinx

Versal employs adaptable heterogeneous system architecture
- New SW programmable AI Engine for diverse compute acceleration workloads
- New High-bandwidth Network-on-Chip integrated w/ hardened DDR Subsystem
- Processor System in all Versal devices
- Re-architected Programmable Logic

Xilinx first 7nm device: Versal AI Core VC1902
- 133TOPs AI Engines, 12TOPs DSP Engines, and 900K LUTs
- 256b DDR4/LPDDR4, PCIe Gen4 & CCIX up to 25Gpbs
- For more details, refer to: www.xilinx.com/versal
Adaptable.
Intelligent.