Zynq Ultrascale+ Architecture

Stephanie Soldavini and Andrew Ramsey

CMPE-550

Dec 2017
Agenda

- Heterogeneous Computing
- Zynq Ultrascale+
  - History
  - Architecture
  - Applications
Problem: Flexibility/Performance Trade Off

Selection Factors:
- Type and complexity of computational algorithms (general purpose vs. Specialized)
- Desired level of flexibility - Performance
- Development cost - System cost
- Power requirements - Real-time constraints
- Specialization, Development cost/time
- Performance/Chip Area/Watt (Computational Efficiency)
- Software/Hardware

Solution: Use some of each in a single system
Problem: Flexibility/Performance Trade Off

Selection Factors:
- Type and complexity of computational algorithms (general purpose vs. Specialized)
- Desired level of flexibility
- Development cost
- Power requirements

Solution: Use some of each in a single system
Heterogeneous Computing

- Combine the use of different devices, for example:
  - Hardware accelerator used to speed up one function in a program
  - Offload matrix calculations to a GPU
  - Cloud system with GPP, GPU, and/or FPGA resources
- Allows for each part of a task to run on the device it is best suited for
Zynq Ultrascale+ History

- Made by Xilinx
- “Microheterogenous”
  - Integrates GPP, GPU, FPGA, Co-Proc, & ASIC in one SoC
  - Increases speed by reducing off-chip data transfer
- Predecessors
  - Kintex-UltraScale and Virtex-UltraScale (20/16nm FPGA fabric)
  - Zynq-7000 (Dual-core ARM Cortex A9 & 28nm FPGA fabric)
General Architecture

Processing System (PS)

- Application Processing Unit (APU)
  - 64-bit quad-core or dual-core ARM Cortex-A53
- Real-time Processing Unit (RPU)
  - 32-bit dual-core ARM Cortex-R5
- Graphics Processing Unit (GPU)
  - ARM Mali-400
- On-Chip Memory (OCM)
  - 256 kB RAM with Error-Correcting Codes (ECC)

Programmable Logic (PL)

- 16nm FinFET+ programmable logic
- Configurable Logic Blocks (CLB)
- 36 kb Block RAMs
- UltraRAM
- DSP Blocks
- Processing System (PS)
- Programmable Logic (PL)
- Interconnects & I/O
Application Processing Unit (APU)

- 64-bit quad-core or dual-core ARM Cortex-A53
- Up to 1.5 GHz
- ARMv8-A Architecture
  - 64-bit mode: A64 instruction set
  - 32-bit mode: A32/T32 instruction set
- Single/double precision floating point unit (FPU)
- Cache
  - IL1: 32 kB 2-way set-assoc with parity (independent for each CPU)
  - DL1: 32 kB 4-way set-assoc with ECC (independent for each CPU)
  - L2: 1 MB 16-way set-assoc with ECC (shared between CPUs)
Real-time Processing Unit (RPU)

- 32-bit dual-core ARM Cortex-R5
- Up to 600 MHz
- ARMv7-R Architecture: A32/T32 instruction set
- Single/double precision FPU
- Caches/Tightly Coupled Memory (TCM)
  - L1: 32 kB 4-way set-assoc with ECC (independent for each CPU)
  - TCM: 128 kB (independent, but can be combined into one 256 kB)
Graphics Processing Unit (GPU)

- ARM Mali-400
- Up to 667 MHz
- One geometry processor
- Two pixel processors
- Supports OpenGL 1.1 & 2.0, OpenVG 1.1
- Advanced anti-aliasing support
- Cache: L2: 64 kB
Programmable Logic (PL)

- 16nm FinFET+ programmable logic
- Configurable Logic Blocks (CLB)
  - Look Up Tables (LUT)
  - Flip flops (FF)
  - Cascadable adders
- 36 kb Block RAMs
  - True dual-port
  - Up to 72 bits wide
  - Configurable as dual 18 kb
- UltraRAM
  - 288 kb
  - 72 bits wide
  - ECC
- DSP Blocks
  - $27 \times 18$ signed multiply
  - 48-bit adder/accumulator
  - 27-bit pre-adder
Vivado Design Suite

Bright green shows configurable components
Customize components, for instance the DDR controller
Applications

- Data Center: Networked Storage/Service Platform[2]
- Multimedia, video encoding/decoding[1]
- Particle physics[4]
- Automotive driver assistance, driver information, and infotainment.
- LTE radio and baseband.
- Medical diagnostics and imaging.
- Video and night vision equipment.
- Wireless radio.
- Single-chip computer.
Application: Data Center

- A sample configuration used for a networked storage platform
- 4.5X performance speed-up & 20X power reduction over x86 implementations
Questions?


