RISC-V Summit
December 3 - 6, 2018
Santa Clara Convention Center
CA, USA

REVOLUTIONIZING THE COMPUTING LANDSCAPE AND BEYOND.

https://tmt.knect365.com/risc-v-summit
AI AT THE EDGE USING PULP + EFPGA

Luca Benini
Professor
ETH Zurich

Timothy Saxe
CTO & SVP Engineering
QuickLogic

https://tmt.knect365.com/risc-v-summit
From Cloud …

- Optimal when sensors are simple (thermostat or switch)
- Applications have higher latency & power consumption
- Data security can be a factor
- Local insights are trivial and non-actionable
From Cloud … to Endpoint

- Optimal when sensors are simple (thermostat or switch)
- Applications have higher latency & power consumption
- Data security can a factor
- Local insights are trivial and non-actionable

- Smart Sensors ➔ rich data ➔ actionable if real-time
- Determine real-time local response
- Network sends insightful data (less bandwidth needed)
- Cloud focuses on aggregate data insights and actions
From Cloud … to Endpoint

- Optimal when sensors are simple (thermostat or switch)
- Applications have higher latency & power consumption
- Data security can a factor
- Local insights are trivial and non-actionable

Cloud and Endpoint AI should be **cooperative**, not competitive
“Within the wide-area IoT segment, two distinct sub-segments with different requirements have emerged: massive and critical applications.

Massive IoT connections are characterized by high connection volumes and small data traffic volumes, low cost devices and low energy consumption.”

Dealing with severe resource constraints

1. Reduce vast amounts of raw sensor data into meaningful events
   ➔ AI appears to be the most practical way to map raw data into meaningful events

2. Use hardware processing engines to deliver energy efficiency
   ➔ Typical hardware accelerators deliver 3x to 10x more energy efficiency

3. Use hardware processing engines to augment CPU performance
   ➔ Typical hardware processing engines deliver 3x to 8x more performance
Excellent starting point for resource constrained devices

- 32b RISC core with ISA extensions
  - Increases energy efficiency of signal processing applications
- Autonomous I/O system
  - Increases energy efficiency by handling sensor I/O in hardware, not software
- Support for custom hardware processing elements
  - Enables further increases in either energy efficiency or performance
Future proofing hardware processing elements

- Your software crystal ball is hazy?
  → No problem, just send an over-the-air update to the software

- Your hardware crystal ball is hazy?
  - Don’t have eFPGA?
    → Workaround with software and pay the power penalty
  - Got eFPGA?
    → No problem, just send an over-the-air update to the eFPGA
Arnold = PULPissmo + eFPGA Testbed

- Cooperative effort between ETH Zurich and QuickLogic
- ETH supplied the PULPissimo
- QuickLogic supplied the eFPGA
- Uses GLOBALFOUNDRIES 22FDX
- Goal is to demonstrate tightly coupled hardware programmable processing elements deployed in the eFPGA
Three use cases for eFPGA

- **Co-processor use case**
  - Hardware processing element implemented in eFPGA to off-load the RISCY CPU

- **Pre-processor use case**
  - Hardware processing element inserted between the sensors and the RISCY CPU

- **Sensor/Actuator/Accelerator interface use case**
  - eFPGA directly interfaces with sensor, actuator or accelerator device with non-standard interface requirements
Co-processor use case

- RISCY sets up data in memory
- Data retrieved via Memory Interconnect
- State machines and data paths in eFPGA process the data using local DP memories as scratch memory
- Data sent back to memory via Memory Interconnect

- Lower power than pure software, higher than dedicated hardware
  - FFT, MFCC, DWT, BNN, etc.
Pre-processor use case

- RISCY sets up sensor and uDMA
- uDMA manages sensor and supplies sensor data to eFPGA
- State machines and data paths in eFPGA process the data using local DP memories as scratch memory
- Data sent to memory via Memory Interconnect to be processed by RISCY

- Lower power than pure software, higher than dedicated hardware
  - FFT, MFCC, DWT, ROI, Subsampling, Histogramming, reshaping, etc.
Sensor/Actuator/Accelerator use case

- eFPGA directly connected to I/O
- State machines and data paths handle sensor/actuator/accelerator interface
- Data sent to/from memory

- Provides precise I/O timing and data formatting required to interface with non-standard sensors, actuators or accelerator devices
  - Laser scanners, image sensors, PDM microphones, multi-color LEDs, CNN accelerators
Future combined use cases

- **HWPE implements AI engine**
- **eFPGA manages sensor and formats data for the AI engine**
- **Covers the full energy efficiency flexibility space**
Future combined use cases

- HWPE implements AI engine
- eFPGA manages sensor and formats data for the AI engine
- Covers the full energy efficiency flexibility space
THANK YOU

https://tmt.knect365.com/risc-v-summit