Open Source On-Chip Communication from Edge to Cloud: the PULP experience

Davide Rossi <davide.rossi@unibo.it>
Thomas Benz <tbenz@iis.ee.ethz.ch>
Luca Bertaccini <lbertaccini@iis.ee.ethz.ch>
Florian Zaruba <zarubaf@iis.ee.ethz.ch>
Agenda

- Davide Rossi (UNIBO): “PULP: An Open-Source RISC-V Based Multi-Core Platform for In-Sensor Analytics”
- Luca Bertaccini (ETHZ): “HERO: A Heterogenous Research Platform to Explore HW/SW Codesign and RISC-V manycore accelerators”
- Florian Zaruba (ETHZ): “Manticore as an NoC Case Study: A 4096 Chiplet-based Architecture for Ultra-Efficient Floating-Point Computing”
Near-Sensor Computing challenge

AI capabilities in the power envelope of an MCU: 100mW peak (10mW avg)

LOGISTICS
$28BN

MANUFACTURING / INDUSTRIAL AUTOMATION
$22BN

SMART CITIES/BUILDING
$12BN

RETAIL
$8BN

70B$ in 5Y

AVERAGE CAGR 27.3%

#1 Customer Question on Amazon.com (out of 1,000+):

1. I don’t want any of my (private, personal) videos on any servers not in my control. Is this possible?

Source: www.amazon.com/dp/7768671737

#2 Customer Question on Amazon.com (out of 1,000+):

2. How long does the battery charge last?

Source: www.amazon.com/dp/7768671737

E. Gousev, Qcomm research
Energy efficiency @ GOPS is **THE** Challenge

![Graph showing energy efficiency vs. performance for high and low-power MCUs]

- **High performance MCUs**
- **Low-Power MCUs**

J Pineda, NXP + Updates
RI5CY Processor

3-cycle ALU-OP, 4-cycle MEM-OP → IPC loss: LD-use, Branch

V1  Baseline RISC-V RV32IMC (not good for ML)
V2  HW loops, Post modified Load/Store, Mac
V3  SIMD 2/4 + DotProduct + Shuffling
    Bit manipulation, Lightweight fixed point

XPULP 25 kGE → 40 kGE (1.6x) but 9+ times DSP!

Nice – But what about the GOPS? Faster+Superscalar is not efficient!

M7: 5.01 CoreMark/MHz-58.5 µW/MHz
M4: 3.42 CoreMark/MHz-12.26 µW/MHz

ML & Parallel, Near-threshold: a Marriage Made in Heaven

- As VDD decreases, operating speed decreases
- However efficiency increases → more work done per Joule
- Until leakage effects start to dominate
- Put more units in parallel to get performance up and keep them busy with a parallel workload

ML is massively parallel and scales well (P/S ↑ with NN size)

Better to have N× PEs running at lower voltage than one PE at nominal voltage!
Multiple RI5CY Cores (1-16)
Low-Latency Shared TCDM

Tightly Coupled Data Memory

Mem

Mem

Mem

Mem

Mem

Mem

Logarithmic Interconnect

RISC-V core

RISC-V core

RISC-V core

RISC-V core

Shared instruction cache with private “loop buffer”

Fast synchronization and Atomics

Tightly Coupled Data Memory

Mem
Mem
Mem
Mem
Mem

Logarithmic Interconnect

RISC-V core
RISC-V core
RISC-V core
RISC-V core

Results: RV32IMCXpulp vs RV32IMC (DNN)

- 8-bit convolution
  - Open source DNN library
- 10x through xPULP
  - Extensions bring real speedup
- Near-linear speedup
  - Scales well for regular workloads.
- 75x overall gain

Garofalo, Angelo et al. “PULP-NN: Accelerating Quantized Neural Networks on Parallel Ultra-Low-Power RISC-V Processors.” Philosophical Transactions of the Royal Society A
Xpulp Extensions Performance (non-DNN)

- up to 1.8x
- up to 4x
- up to 11x
Parallel Speed-Up (non-DNN)

DMA for data transfers from/to L2

An additional I/O controller is used for IO

How do we work: Initiate a DMA transfer
Data copied from L2 into TCDM

PULPissimo

Tightly Coupled Data Memory

Mem
Mem
Mem
Mem
Mem

I$
Once data is transferred, event unit notifies cores
Cores can work on the data transferred
Once our work is done, DMA copies data back
During normal operation all of these occur concurrently.

PULP: An Open-Source RISC-V Based Multi-Core Platform for In-Sensor Analytics
Explicit Memory Management: MobileNet Example

- ~4 Mparameters → need to store weights in off-chip memory (L3)

- L1 Bandwidth: 256 Gbit/s @ 250 MHz
- L2 Bandwidth: 32 Gbit/s @ 250 MHz
- L3 Bandwidth: 1.6 Gbit/s @ 100 MHz
Tensor tiling

L3 / L2 tiling
64 MB / 512 kB

small memory

big memory
Tensor tiling

**L3 / L2 tiling**

64 MB / 512 kB

**L2 / L1 tiling**

512 kB / 64 kB

small memory

big memory
Tile Data Movement

L2 memory

Input feature map I

Filters weights W

Output feature map O

L1 memory

L1 buffer 1

x TILE 1

y TILE 1

W TILE 1

L1 buffer 2

x TILE 2

y TILE 2

W TILE 2

CONVOLUTIONAL PIPELINE

\[ t_0 \quad t_1 \quad t_2 \quad t_3 \quad \ldots \quad t_n \]

In.
copy

DMA ch. 0-1

DMA ch. 2

Cluster computation
Tile Data Movement

CONVOLUTIONAL PIPELINE

$\mathbf{L1} \text{ memory}$

- $x \text{ TILE 1}$
- $y \text{ TILE 1}$
- $W \text{ TILE 1}$

$\mathbf{L2} \text{ memory}$

- Input feature map $I$
- Filters weights $W$
- Output feature map $O$

$t_0 \quad t_1 \quad t_2 \quad t_3 \quad \ldots \quad t_n$

- In. copy
- Convol. kernel

DMA ch. 0-1

Cluster computation

DMA ch. 2
Tile Data Movement

CONVOLUTIONAL PIPELINE

\[ t_0 \rightarrow t_1 \rightarrow t_2 \rightarrow t_3 \rightarrow \ldots \rightarrow t_n \]

Input feature map \( I \)

Output feature map \( O \)

Filters weights \( W \)

L1 memory

L1 buffer 1

x TILE 1

y TILE 1

W TILE 1

L1 buffer 2

x TILE 2

y TILE 2

W TILE 2

L2 memory

DMA ch. 0-1

DMA ch. 2

Cluster computation

In. copy

Convol. kernel

In. copy

PULP: An Open-Source RISC-V Based Multi-Core Platform for In-Sensor Analytics
### Tile Data Movement

#### Input feature map $I$
- $h_M$ x $i_M$ x $w_M$

#### Filters weights $W$
- $h_M$ x $i_M$ x $w_M$

#### Output feature map $O$
- $h_M$ x $i_M$ x $w_M$

#### L1 memory
- $L1$ buffer 1
  - $x$ TILE 1
  - $y$ TILE 1
  - W TILE 1

- $L1$ buffer 2
  - $x$ TILE 2
  - $y$ TILE 2
  - W TILE 2

#### L2 memory

#### CONVOLUTIONAL PIPELINE
- $t_0$, $t_1$, $t_2$, $t_3$, ..., $t_n$

<table>
<thead>
<tr>
<th></th>
<th>In. copy</th>
<th>Convol. kernel</th>
<th>Out. copy</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_0$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$t_1$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$t_2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$t_3$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### DMA channels
- DMA ch. 0-1
- DMA ch. 2

#### Cluster computation
Tile Data Movement

CONVOLUTIONAL PIPELINE

<table>
<thead>
<tr>
<th>t₀</th>
<th>t₁</th>
<th>t₂</th>
<th>t₃</th>
<th>...</th>
<th>tₙ</th>
</tr>
</thead>
<tbody>
<tr>
<td>In. copy</td>
<td>Convol. kernel</td>
<td>Out. copy</td>
<td>In. copy</td>
<td>• • •</td>
<td>Convol. kernel</td>
</tr>
<tr>
<td>In. copy</td>
<td>Convol. kernel</td>
<td>Out. copy</td>
<td>• • •</td>
<td>In. copy</td>
<td></td>
</tr>
<tr>
<td>In. copy</td>
<td>Convol. kernel</td>
<td>• • •</td>
<td>Out. copy</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

DMA ch. 0-1

DMA ch. 2

Cluster computation
PULP includes Cores+Interco+IO+HWCE → Open Platform

**RISC-V Cores**
- RI5CY: 32b
- Ibex: 32b
- Snitch: 32b
- Ariane + Ara: 64b

**Platforms**
- Single Core:
  - PULPino
  - PULPissimo
- Multi-core:
  - Fulmine
  - Mr. Wolf
- Multi-cluster:
  - Hero
  - Manticore

**Accelerators**
- HWCE: convolution
- Neurostream: (ML)
- HWCrypt: (crypto)
- PULPO: (1st ord. opt.)

**Interconnect**
- Logarithmic interconnect
- APB – Peripheral Bus
- AXI4 – Interconnect

**Peripherals**
- JTAG
- SPI
- UART
- I2S
- DMA
- GPIO
- DMA
- I2S
- UART
- SPI

**Integrations**
- PULP: An Open-Source RISC-V Based Multi-Core Platform for In-Sensor Analytics
From Edge to Cloud: The Communication Wall

40 Years of Microprocessor Trend Data

- Big Gap
- 1.2 times/year
- (2.5 times/5 years)
- Performance Progress of CPU


Ref.: Ahmet Ceyhan, INTERCONNECTS FOR FUTURE TECHNOLOGY GENERATIONS—CONVENTIONAL CMOS WITH COPPER/LOW-κ AND BEYOND, Fig 2, 9