Extreme Energy Efficiency for Extreme Edge AI Acceleration
a RISC-V platform design perspective
Luca Benini
<lbenini@iis.ee.ethz.ch,luca.Benini@unibo.it>

Prof. of Digital Circuit and Systems
@ ETHZ and UNIBO. h-index=112,
57’000+ citations, 1’000+
publications, fellow IEEE, ACM,
Chief Architect in STMicroelectronics
(2009-2012) Group of 100+ people

http://pulp-platform.org  @pulp_platform
TinyML opportunity

TinyML challenge
AI capabilities in the power envelope of an MCU: 10-mW peak (1mW avg)
AI Workloads - DNNs

High OP/B ratio
Massive Parallelism
MAC-dominated
Low precision OK

90.2%, 480M-param, many GOPS

70%, “Tiny” DNNs

5M-param
Energy efficiency @ GOPS is the Challenge

ARM Cortex-M MCUs: M0+, M4, M7 (40LP, typ, 1.1V)*

1pJ/MAC=1GMAC/mW

*data from ARM’s web
RI5CY – An Open MCU-class RISC-V Core for EE-AI

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

ISA is extensible by construction (great!)

V1  Baseline RISC-V RV32IMC (not good for ML)
    HW loops

V2  Post modified Load/Store
    Mac

V3  SIMD 2/4 + DotProduct + Shuffling
    Bit manipulation unit
    Lightweight fixed point

XPULP extensions: 25 kGE → 40 kGE (1.6x)

[Gautschi et al. TVLSI 2017]
PULP-NN: Xpulp ISA exploitation

8-bit Convolution

RV32IMC

addi a0, a0, 1
addi t1, t1, 1
addi t3, t3, 1
addi t4, t4, 1
lbu a7, -1(a0)
lbu a6, -1(t4)
lbu a5, -1(t3)
lbu t5, -1(t1)
mul s1, a7, a6
mul a7, a7, a5
add s0, s0, s1
mul a6, a6, t5
add t0, t0, a7
mul a5, a5, t5
add t2, t2, a6
add t6, t6, a5
bne s5, a0, 1c000bc

RV32IMCXpulp

lp.setup
p.lw w1, 4(a0!)
p.lw w2, 4(a1!)
p.lw x1, 4(a2!)
p.lw x2, 4(a3!)

lp.setup
p.lw w1, 4(a0!)
p.lw w2, 4(a1!)
p.lw x1, 4(a2!)
p.lw x2, 4(a3!)

lp.setup
p.lw w1, 4(a0!)
p.lw w2, 4(a1!)
p.lw x1, 4(a2!)
p.lw x2, 4(a3!)

lp.setup
p.lw w1, 4(a0!)
p.lw w2, 4(a1!)
p.lw x1, 4(a2!)
p.lw x2, 4(a3!)

9x less instructions than RV32IMC

14.5x less instructions at an extra 3% area cost (~600GEs)
Supporting dotp + load

NN Register File: 6 32-bit registers (weights and input activations)

\[ P \uparrow T \downarrow \downarrow \text{so, } E = P \times T \downarrow \downarrow \]

Nice…But, what about the GOPS?

Faster clock + Superscalar is not efficient! (M4 \rightarrow M7)
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 optimum Energy than 1 PE running fast at low Energy efficiency
Multiple RISCY Cores (1-16)

CLUSTER

RISC-V core  RISC-V core  RISC-V core  RISC-V core
Low-Latency Shared TCDM

CLUSTER

Tightly Coupled Data Memory BF=2

Mem0 Mem1 Mem2 Mem3 Mem4 Mem5 Mem6 Mem7

Logarithmic Interconnect

RISC-V core RISC-V core RISC-V core RISC-V core
DMA for data transfers from/to L2 + Fast synchro

Tightly Coupled Data Memory BF=2

Mem4  Mem5  Mem6  Mem7

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

~15x latency and energy reduction for a barrier [Glaser TPDS20]
Shared instruction cache with private “loop buffer”

- Most IFs from I$-$P
  - Low IF energy
- I$-$S for capacity
  - Reduces miss latency

Tightly Coupled Data Memory BF=2

L2 Mem

MEM

DMA

HW SYNC

Logarithmic Interconnect

RISC-V core

RISC-V core

RISC-V core

RISC-V core

CLUSTER
Vector Lockstep Execution Mode (VLEM)

~45% energy saving

- MIMD: 1.0
- VLEM: 1.7
- VLEM + BRD + MIS. DATA: 1.02

Tightly Coupled Data Memory BF=2

Mem0, Mem1, Mem2, Mem3

Mem4, Mem5, Mem6, Mem7

[Garofalo et al. ESSCIR21]
Results: RV32IMCXpulp vs RV32IMC

- **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**
  - 2 orders of magnitude with DOTP+LW (122x)
  - Sub-byte (nibble, crumb) supported (537x, 939x)

Overall Speedup of 75x

Near-Linear Speedup

10x Speedup w.r.t. RV32IMC (ISA does matter 😊)

[Garofalo et al. Philos. Trans. R. Soc 20]
An additional I/O controller is used for IO & PM

Open sourced since 2017: github.com/pulp-platform/pulp
Deploying DNNs on PULP

QuantLab
Quantization Laboratory

NEMO
NEural Minimization for pytOrch

DORY
Deployment Oriented to memoRY

PULP-NN
PULP Neural Network backend

github.com/pulp-platform/nemo, /dory, /pulp-nn

[Burrello et al. TCOMP21]
What’s next? Tightly-coupled HW Compute Engine

Acceleration with flexibility: zero-copy HW-SW cooperation
Hardware Processing Engines (HWPEs)

HWPE efficiency vs. optimized RISC-V core

1. Dedicated control (no I-fetch) with shadow registers (overlapped config-exec)
2. Specialized high-BW interco into L1 (on data-plane)
3. Specialized datapath: supporting configurable & aggressive quantization
Binary-Based Quantization (BBQ)

QNN layer:

\[ y(k_{out}) = \text{quant} \left( \sum_{k_{in}} \left( W(k_{out}, k_{in}) \otimes x(k_{in}) \right) \right) \]

- Q-bit output fmaps
- M-bit weights
- N-bit input fmaps
- INT32 accumulator

Many \( M \times N \) bits products...

... but one \( M \times N \) product is the superposition of \( M \times N \) 1-bit products!

\[ y(k_{out}) = \text{quant} \left( \sum_{i=0}^{M} \sum_{j=0}^{N} \sum_{k_{in}} 2^i 2^j \left( w_{bin}(k_{out}, k_{in}) \otimes x_{bin}(k_{in}) \right) \right) \]

- Q-bit output fmaps
- 1-bit weights
- 1-bit input fmaps
- power-of-2 scaling factors

One quantized NN can be emulated by superposition of power-of-2 weighted \( M \times N \) binary NN
Reconfigurable Binary Engine

\[ y(k_{out}) = \text{quant} \left( \sum_{i=0..M} \sum_{j=0..N} 2^i 2^j (W_{\text{bin}}(k_{out}, k_{in}) \otimes x_{\text{bin}}(k_{in})) \right) \]

github.com/pulp-platform/rbe

RBE Block
Peak throughput
10368 = 9×9×4×32

Energy efficiency 10-20x (0.1pJ/OP) wrt to SW on cluster @same accuracy

Vector LD/ST Unit (9×32 bit)

Fmap Buffer

Block
Block
Block
Block
Block
Block

Block
Block
Block
Block
Block
Block

Block
Block
Block
Block
Block
Block

Acc
Acc
Acc
Acc
Acc
Acc

Quant
Quant
Quant
Quant
Quant
Quant

Ctrl & Regfile

Scale & Add

32x BinConv
32x BinConv
32x BinConv
32x BinConv
32x BinConv
32x BinConv

 reduction tree
 reduction tree
 reduction tree
 reduction tree

 inputs stationary (32b)
 weights streamed (32b)

BinConv

psum streamed

31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 0
31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 0
Mixed-Precision Quantized Networks – CMIX-NN

Apply tensor-wise quantization to fit memory constraints with low accuracy drop

- Best Top1: 70.1%
- Best Mixed: 68%
- Best Top4 Fit 60.5%
- Best Top1 Fit 48%

Only -2% wrt most accurate INT8 mobilenetV1 (224_1.0) which does not fit on-chip
+8% wrt most accurate INT8 mobilenetV1 fitting on-chip (192_0.5)
+7.5% wrt most accurate INT4 mobilenetV1 (224_1.0) fitting on chip

Rule-based bit precision selection based on memory constraints
Avgbit 6.7 Act Avgbit 5.9 Wgt

[Capotondi et al. TCAS II, 2020]
HW acceleration in perspective

Using 22FDX tech, NT@0.6V, High utilization, minimal IO & overhead

Energy-Efficient RV Core $\rightarrow$ 20pJ (8bit)

ISA-based 10-20x $\rightarrow$ 1-2pJ (8bit) $\rightarrow$ XPULPV2 & V3

Configurable DP 10-20x $\rightarrow$ 50-100fJ (4bit) $\rightarrow$ HWCE, RBE, NE

Fully specialized DP 10-20x $\rightarrow$ 5-10fJ (ternary) $\rightarrow$ XNE, CUTIE*

*sub 1fJ in 7nm [Scherer et al. TCAD21]
All together in VEGA: Extreme Edge IoT Processor

- RISC-V cluster (8 cores + 1)
  614GOPS/W @ 7.6GOPS (8 bit DNNs), 79GFLOPS/W @ 1GFLOP (32 bit FP appl)
- Multi-precision HWCE (4b/8b/16b)
  3×3×3 MACs with normalization/activation: 32.2GOPS and 1.3TOPS/W (8 bit)

In cooperation with Rossi et al. ISSCC21
All together in VEGA: Extreme Edge IoT Processor

- RISC-V cluster (8 cores + 1)
  614 GOPS/W @ 7.6 GOPS (8 bit DNNs), 79 GFLOPS/W @ 1 GFLOP (32 bit FP appl)
- Multi-precision HWCE (4b/8b/16b)
  3×3×3 MACs with normalization / activation: 32.2 GOPS and 1.3 TOPS/W (8 bit)
- Fully-on chip DNN inference with 4 MB MRAM (high-density NVM with good scaling)

<table>
<thead>
<tr>
<th>Technology</th>
<th>22nm FDSOI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chip Area</td>
<td>12 mm²</td>
</tr>
<tr>
<td>SRAM</td>
<td>1.7 MB</td>
</tr>
<tr>
<td>MRAM</td>
<td>4 MB</td>
</tr>
<tr>
<td>VDD range</td>
<td>0.5V - 0.8V</td>
</tr>
<tr>
<td>VBB range</td>
<td>0V - 1.1V</td>
</tr>
<tr>
<td>Fr. Range</td>
<td>32 kHz - 450 MHz</td>
</tr>
<tr>
<td>Pow. Range</td>
<td>1.7 µW - 49.4 mW</td>
</tr>
</tbody>
</table>

In cooperation with GREENWAVES Technologies
Full DNN Energy (MobileNetV2)

Bandwidth [MB/s]

Energy per byte [pJ/B]

weights on MRAM
weights on HyperRAM

end-to-end on-chip computation

1.19 mJ

3.5x less energy

1.4 <1

4.16 mJ
When you count mWatts, everything matters!

What about IO power? (Mem, Sensor)

- **SPIs**
  - I/O VDD=1.8V
  - f_{spi-max}=50MHz,
  - Assuming duty-cycled operation @ various bandwidths

- **ULP serial link** (duty-cycled)
  - 10.2x less energy and 15.7x higher maximum BW compared to single SPI
  - 2.56x higher efficiency than the DDR
  - Octal SPI @787Mbps
  - 5 → 3pJ/bit
  - However it’s still 2mW@ 500Mbps
  - 3D integration: 0.15pJ/bit and below

From near-sensor to in-sensor (3D IC)
Closing thoughts – Open Platform for TinyML

PULP is an Open Platform

- For science … fundamental “research infrastructure”
  Reduce “getting up to speed” overhead for partners
  Enables fair and well controlled benchmarking
- For Business … it is truly disruptive
  Reduces NRE + faster innovation path for startups (e.g. Greenwaves tech.), new business models (eg. OpenHWGroup), helps collaboration with foundries (e.g. GF)

Heterogeneous & Flexible

- 1-3 orders of magnitude improvement (wrt to efficient RV) by acceleration
  ISA → Configurable → Fully customized + heterogeneous architectural combinations
- Focus on IO energy (memory, sensor) to achieve sub pJ/OP @ full platform
  3D-IC technology is a key enabler
Luca Benini, Alessandro Capotondi, Alessandro Ottaviano, Alessio Burrello, Alfio Di Mauro, Andrea Borghesi, Andrea Cossettini, Andreas Kurth, Angelo Garofalo, Antonio Pullini, Arpan Prasad, Bjoern Forsberg, Corrado Bonfanti, Cristian Cioflan, Daniele Palossi, Davide Rossi, Fabio Montagna, Florian Glaser, Florian Zaruba, Francesco Conti, Georg Rutishauser, Germain Haugou, Gianna Paulin, Giuseppe Tagliavini, Hanna Müller, Luca Bertaccini, Luca Valente, Manuel Eggimann, Manuele Rusci, Marco Guermandi, Matheus Cavalcante, Matteo Perotti, Matteo Spallanzani, Michael Rogenmoser, Moritz Scherer, Moritz Schneider, Nazareno Bruschi, Nils Wistoff, Pasquale Davide Schiavone, Paul Scheffler, Philipp Mayer, Robert Balas, Samuel Riedel, Segio Mazzola, Sergei Vostrikov, Simone Benatti, Stefan Mach, Thomas Benz, Thorir Ingolfsson, Tim Fischer, Victor Javier Kartsch Morinigo, Vlad Niculescu, Xiaying Wang, Yichao Zhang, Frank K. Gürkaynak, all our past collaborators and many more that we forgot to mention.