Working with RISC-V

Part 4 of 4: PULP based chips

Frank K. Gürkaynak <kgf@ee.ethz.ch>
Luca Benini <lbenini@iis.ee.ethz.ch>
Summary

- Part 1 – Introduction to RISC-V ISA
- Part 2 – Advanced RISC-V Architectures
- Part 3 – PULP concepts
- Part 4 – PULP based chips
  - From concept to reality
  - Single core microcontrollers: PULPino to PULPissimo
  - Many core systems: OpenPULP
  - Advanced systems with accelerators
  - Lessons learned, the good, the bad and the ugly.
We will discuss chips we have made with PULP

- Why make chips at all?
  - MPW: Only limited samples
  - Use cases

- Single core PULP chips
  - PULPino (Imperio)
  - PULPissimo (Arnold)

- Many core PULP chips
  - Cluster only (Honey Bunny)
  - PULPopen (Mr. Wolf)

- Advanced PULP chips
  - Kosmodrom: 2x 64b Ariane cores + ML accelerators
  - Making use of technology: Body biasing

- Lessons learned
  - There are many pitfalls
  - We had great success, but..
    .. sometimes you have embarrassing failures. Part of the process
Multi Project Wafer, chips for prototyping

- **Cost sharing method for ICs**
  - Multiple ICs are manufactured together. They share the mask costs
    - 1.5M cost / 10 projects = 150k per project
    - But you only get 1 / 10 of the area
  - Dedicated MPW services available
    - Europractice-IC for SMEs and academia

- **You only get a few chips**
  - Usually 50 to 200
  - Per chip costs very high (few kUSD)

- **All our chips through MPWs**

Image taken from https://europractice-ic.com/mpw-prototyping/general/mpw-minisic/
Our ASICs have different use cases

- Chips characterized on an IC tester *(Poseidon 22nm)*
- Research demonstrators *(Nano drone with Mr. Wolf/GAP8)*
- Industrial uses of our cores/peripherals *(open-isa.org Vega) board*
Most of what we show is openly available

- All our development is on GitHub
  - HDL source code, testbenches, software development kit, virtual platform
    [https://github.com/pulp-platform](https://github.com/pulp-platform)

- PULP is released under the permissive Solderpad license
  - Allows anyone to use, change, and make products without restrictions.
PULP has released a large number of IPs

### 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, Open Piton

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

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

### Accelerators
- HWCE (convolution)
- Neurostream (ML)
- HWCrypt (crypto)
- PULPO (1st ord. opt)
PULPino: Our first open source release

- Simple design
  - Meant as a quick release

- Separate data and inst. mem
  - Makes it easy in HW
  - Not meant as a Harvard arch.

- Can use all our 32bit cores
  - RI5CY (CV32E40P), Zero/Micro-Riscy (Ibex)

- Peripherals from other projects
  - Any AXI and APB peripherals could be used
Imperio – 65nm RISC-V core

- Chip implemented in 65nm
  - Using RI5CY (RV32IMC) core
  - 64 kBytes of memory
  - Basic peripherals (GPIO, SPI, I2C)
  - Working debug interface

- Functional up to 500 MHz
  - Main challenge was to find fast memory cuts to work at that speed.
  - Memory made of multiple smaller cuts to maximize the operating speed.
Working chip on an Arduino compatible board
#5 - Arnold (2018) – Fastest collaboration

- **GF22nm**
  - RISC-V microcontroller with eFPGA
  - Based around PULPissimo

- **Collaboration with Quicklogic**
  - Met at GTC 2017 by coincidence
  - In one year chip was taped out
  - Only possible because of open source nature

- **Quicklogic is going open source**
  - They announced June 2020 the Quicklogic Open Reconfigurable Computing
  - [https://www.quicklogic.com/QORC/](https://www.quicklogic.com/QORC/)
PULPissimo: very good platform for extensions

- eFPGA added as accel.
  - Easy plug and play
  - Configuration over APB
  - Additional ALU and memory
  - Uses the same memory

- Multiple operation modes
  - Configurable peripheral
  - Accelerator for core
  - Accelerator for independent I/O
Experimental platform with many configurations

- I/O subsystem accel
  - 6.0mW, 2.5x

- Custom I/O interface
  - BNN interface 12.5mW 2.2x

- CPU accelerator
  - CRC 7.5mW 42x

- Many more ideas
  - Dynamic reconfiguration
Arnold test board with D. Schiavone
Many cores running at low VDD is more efficient.

Better to have 4 cores running at lower voltage than one core at nominal voltage!
Instead of using a single fast core
Let us have a cluster of cores
Many cores connected to many memory banks
DMA copies data from an external memory
Add a SoC part that includes memory and I/O

Tightly Coupled Data Memory

SoC

L2 Mem

interconnect

DMA

Mem

Mem

Mem

Mem

Event Unit

interconnect

RISC-V core

RISC-V core

RISC-V core

RISC-V core

I$

I$

I$

I$
Honey Bunny – GF28 SLP

- Our first RISC-V many-core chip
  - Four RI5CY cores (RC32IMC) in one cluster
  - 64 kBytes of TCDM memory inside cluster
  - 256 kBytes of L2 memory
  - Runs at 400MHz+

- New technology for us
  - Needed to port the clock generator (FLL)
  - Design has analog parts
    - Can not be made open source directly
  - Major effort needed for every new technology

Size and number of blocks in the drawing are indicative and not to scale.
Visiting card with 4x RISC-V cores in 28nm

See a video of how the board is assembled under: https://www.youtube.com/watch?v=OEgPXQMRRyc
Mr. Wolf (TSMC 40): 8+1 core IoT Processor

- One cluster with
  - 8 RISC-V cores
  - 2x shared FPU units
  - 64 kByte of TCDM

- One controller with
  - 512 kByte L2 RAM
  - Peripherals

- On chip voltage regulators
  - By Dolphin Integration
On-chip regulators allow different power modes

<table>
<thead>
<tr>
<th>Power Mode</th>
<th>VDD</th>
<th>Frequency Range</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep Sleep</td>
<td>0.8 V</td>
<td>n.A.</td>
<td>72 µW</td>
</tr>
</tbody>
</table>

**Power Control**

```
Controller: M M M M
Cluster: R5 M M M M M M M M
Interconnect: M R5 R5 R5 R5 R5 R5 R5
```
It is possible to keep memory state intact

<table>
<thead>
<tr>
<th>Power Mode</th>
<th>VDD</th>
<th>Frequency Range</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep Sleep</td>
<td>0.8 V</td>
<td>n.A.</td>
<td>72 µW</td>
</tr>
<tr>
<td>State Retentive Deep Sleep</td>
<td>0.8 V</td>
<td>n.A.</td>
<td>77 – 108 µW</td>
</tr>
</tbody>
</table>
SoC is awake but is clock gated

<table>
<thead>
<tr>
<th>Power Mode</th>
<th>VDD</th>
<th>Frequency Range</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep Sleep</td>
<td>0.8 V</td>
<td>n.A.</td>
<td>72 µW</td>
</tr>
<tr>
<td>State Retentive Deep Sleep</td>
<td>0.8 V</td>
<td>n.A.</td>
<td>77 – 108 µW</td>
</tr>
<tr>
<td>SoC Idle</td>
<td>0.8 – 1.1V</td>
<td>SoC clock gated</td>
<td>0.55 – 1.96 mW</td>
</tr>
</tbody>
</table>
Only SoC with a single RISC-V core running

<table>
<thead>
<tr>
<th>Power Mode</th>
<th>VDD</th>
<th>Frequency Range</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep Sleep</td>
<td>0.8 V</td>
<td>n.A.</td>
<td>72 μW</td>
</tr>
<tr>
<td>State Retentive Deep Sleep</td>
<td>0.8 V</td>
<td>n.A.</td>
<td>77 – 108 μW</td>
</tr>
<tr>
<td>SoC Idle</td>
<td>0.8 – 1.1V</td>
<td>SoC clock gated</td>
<td>0.55 – 1.96 mW</td>
</tr>
<tr>
<td>SoC active</td>
<td>0.8 – 1.1V</td>
<td>32 kHz – 450 MHz</td>
<td>0.97 – 38 mW</td>
</tr>
</tbody>
</table>

Controller

<table>
<thead>
<tr>
<th>Power Control</th>
<th>R5</th>
</tr>
</thead>
<tbody>
<tr>
<td>M M M M M</td>
<td></td>
</tr>
</tbody>
</table>

Cluster

<table>
<thead>
<tr>
<th>Interconnect</th>
<th>R5</th>
<th>R5</th>
<th>R5</th>
<th>R5</th>
<th>R5</th>
<th>R5</th>
<th>R5</th>
<th>R5</th>
<th>R5</th>
</tr>
</thead>
<tbody>
<tr>
<td>M M M M M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Cluster is active, but clock gated

<table>
<thead>
<tr>
<th>Power Mode</th>
<th>VDD</th>
<th>Frequency Range</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep Sleep</td>
<td>0.8 V</td>
<td>n.A.</td>
<td>72 μW</td>
</tr>
<tr>
<td>State Retentive Deep Sleep</td>
<td>0.8 V</td>
<td>n.A.</td>
<td>77 – 108 μW</td>
</tr>
<tr>
<td>SoC Idle</td>
<td>0.8 – 1.1V</td>
<td>SoC clock gated</td>
<td>0.55 – 1.96 mW</td>
</tr>
<tr>
<td>SoC active</td>
<td>0.8 – 1.1V</td>
<td>32 kHz – 450 MHz</td>
<td>0.97 – 38 mW</td>
</tr>
<tr>
<td>Cluster Idle</td>
<td>0.8 – 1.1V</td>
<td>Cluster clock gated</td>
<td>1.2 – 4.6 mW</td>
</tr>
</tbody>
</table>
Cluster with 8 RISC-V cores is active

<table>
<thead>
<tr>
<th>Power Mode</th>
<th>VDD</th>
<th>Frequency Range</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep Sleep</td>
<td>0.8 V</td>
<td>n.A.</td>
<td>72 µW</td>
</tr>
<tr>
<td>State Retentive Deep Sleep</td>
<td>0.8 V</td>
<td>n.A.</td>
<td>77 – 108 µW</td>
</tr>
<tr>
<td>SoC Idle</td>
<td>0.8 – 1.1V</td>
<td>SoC clock gated</td>
<td>0.55 – 1.96 mW</td>
</tr>
<tr>
<td>SoC active</td>
<td>0.8 – 1.1V</td>
<td>32 kHz – 450 MHz</td>
<td>0.97 – 38 mW</td>
</tr>
<tr>
<td>Cluster Idle</td>
<td>0.8 – 1.1V</td>
<td>Cluster clock gated</td>
<td>1.2 – 4.6 mW</td>
</tr>
<tr>
<td>Cluster Active</td>
<td>0.8 – 1.1V</td>
<td>32 kHz – 350 MHz</td>
<td>1.6 – 153 mW</td>
</tr>
</tbody>
</table>

Controller: Power Control R5

Cluster: M M M M M

Interconnect: R5 R5 R5 R5 R5 R5 R5 R5
Our OpenPULP release is essentially Mr. Wolf

- OpenPULP contains most of what we have as open source
  - This is a complex IoT processor, not like the much simpler PULPino
  - 8 + 1 cores, FPUs, shared accelerators, multiple power down modes.

- Still many parts still can not be open source
  - Technology specific information, P&R scripts
  - Memory macros, selected cuts, their performance
  - I/O cells
  - FLL, analog macros, I/O cells, memory cuts (affects performance), P&R scripts

- OpenPULP facilitated interesting industry collaboration
  - Greenwaves, BitCraze, Dolphin
Mr. Wolf has been used in multiple systems

- Designed as an application processor
  - We still build boards with it
  - Despite only 200 manufactured

- Widespread industrial use:
  - Dolphin IP was validated on this chip
  - Greenwaves GAP8 is based on the open source release OpenPULP
  - BitCraze AI Deck is related
  - GAP9 (Vega) is a follow up project
Complete Application: DroNET on NanoDrone

Pluggable PCB: PULP-Shield
- ~5g, 30×28mm
- GAP8 SoC
- 8 MB HDRAM
- 16 MB HFlash
- QVGA ULP HiMax camera
- Crazyflie 2.0 nano-drone (27g)

Only onboard computation for autonomous flight + obstacle avoidance
no human operator, no ad-hoc external signals, and no remote base-station!
Moving to more advanced nodes: Kosmodrom

- **Globalfoundries 22FDX**
  - In 2018, most advanced node for us
  - Minimum size 3mm x 3mm
    - That fits about **100 million transistors**
  - Allows body biasing

- **With great power comes...**
  - Designs in 22FDX are more involved
  - More blocks, more functionality
    - More things that can go wrong
  - Challenging design
  - Collaboration with Globalfoundries
Kosmodrom: Main components

- **2x Ariane 64b RISC-V cores**
  - AHP optimized for high speed
  - ALP optimized for low power

- **Automatic Body Bias Gen.**
  - IP by INVECAS
  - Allows body bias to be tuned

- **NTX: Neural Training Accelerator**
  - 260 Gflops/Watt efficiency

- **Common infrastructure**
  - SRAM, Debug, I/Os
Fine-Grained Shared-Memory Accelerators

Similar concept as OpenPULP, but fewer RISC-V cores and more accelerators
NTX uses 1 RISC-V core to control 8 units

- NTX runs at up to 1.25 GHz
- Compute of 20 Gflop/s
- Bandwidth of 5 GB/s
- At 9.3 pJ/flop and using only 0.51 mm²
- Scale up by replicating cluster
Kosmodrom ABB Demonstration Board

- STM microcontroller for control
- Test socket for Kosmodrom chip
- USB connection to computer
- Analog to Digital Converter module
- Body bias voltage generation
- Supply voltage generation
- Measurement points for all supplies
Boosting performance with Body Biasing

- We set the performance target (730 MHz, @0.65 V, ~40 mW)
- Actual chip performance is measured
- Forward VBB is applied (positive VBP and negative VBN)

Until we reach the performance goals

By individually applying VBB to chips we can improve yield

50% Performance gain with Body Biasing
Gaining Energy Efficiency with Body Biasing

- We set the desired operating frequency (800MHz).
- We decrease the voltage to the minimum level chip will work (0.8V).
- At this point we start reducing voltage further (0.65V).
- Maximum operating frequency will also drop (~500MHz).
- We compensate for the lost performance with forward VBB (positive VBP and negative VBN).
- Until we reach the desired operating frequency.

At least 20% more efficiency with VBB

20% less power
The good the bad and the ugly

- We designed and tested 37 chips as part of PULP project (as of now)
  - Three more planned until end of year
- Most worked great
- But there were also mistakes made
- Here is a look at some **highs** and some **lows**
Good: Fulmine the award winning one

- UMC65
- Earlier chip (2015)
  - 4x OpenRISC cores (not yet RISC-V)
  - 192 kBytes L2 + 64 kBytes TCDM
  - 2x HW accelerators
    - HW – Crypt (together with TU-Graz)
    - HW – Convolution Engine
- Publication from this chip
Bad: Bonding issues on Poseidon

- First GF22nm chip
  - Used Europractice IC service
  - Cost 150k CHF for 50 samples
- Has three parts (trident..)
  - PULPissimo system
  - Ariane core
  - Independent ML accelerator
- 30 of 50 chips were packaged
  - We provide a bonding diagram
  - Mostly simple manual work

Work with RISC-V

ACACES 2020 - July 2020
Bad: Bonding issues on Poseidon

- Look closer on the right side
  - There is a pad that is not bonded
- We skipped one pad
  - All connections are shifted by one
- VDD and GND are one after other
  - Bonding causes shorts between VDD and GND
  - Pretty much catastrophic
- Fortunately: unpackaged dies
  - There were 20 unpackaged dies
  - We could bond those correctly
Downright Ugly, reset problem of Urania

- 2 PULP clusters, each with
  - 4x RV32 RI5CY cores
  - 4x transprecision FPUs
  - 1x PULPO accelerator
  - 64 kB TCDM in 8 banks

- Ariane RV64 host processor
  - 128 KiB Shared LLC
  - software-managed IOMMU

- DDR3 DRAM Controller + PHY by TU-Kaiserslautern
The reset can not be released for clusters

- Chip has many modules
  - 1x Ariane core
  - 1x DDR interface
  - 2x Clusters

- Reset to clusters is stuck 0
  - Design flow mistake
  - Some other control signals are stuck as well affecting Ariane performance

- DDR interface is functional
  - Not everything is lost
IC Design is tricky and demands attention

- Even the simplest things can derail a complex chip
  - A copy paste error in a bonding diagram, a mistake in reset
- Academic research chips are not industrial products
  - Designed to test and verify ideas, not mass production
  - Much more effort needed in DfT and verification to make a successful product
- Experience is key in IC Design
  - All the mistakes we make, add to our future success
  - Some lessons you learn the hard way
  - But these stay with you and help you for your future designs
We hope this was helpful/fun for you

- Covered the basics of RISC-V
  - Explained the ISA
  - Examples of Implementations
  - Advanced cores and Concepts

- Talked about building open source systems around RISC-V
  - Showed the main concepts and talked about our ICs

- You can find PULP related information
  - GitHub: [http://github.com/pulp_platform](http://github.com/pulp_platform)
  - PULP Webpage: [http://pulp-platform.org](http://pulp-platform.org)
  - Follow us on Twitter: [@pulp_platform](http://twitter.com/pulp_platform)
Luca Benini, Davide Rossi, Andrea Borghesi, Michele Magno, Simone Benatti, Francesco Conti, Francesco Beneventi, Daniele Palossi, Giuseppe Tagliavini, Antonio Pullini, Germain Haugou, Manuele Rusci, Florian Glaser, Fabio Montagna, Bjoern Forsberg, Pasquale Davide Schiavone, Alfio Di Mauro, Victor Javier Kartsch Morinigo, Tommaso Polonelli, Fabian Schuiki, Stefan Mach, Andreas Kurth, Florian Zaruba, Manuel Eggimann, Philipp Mayer, Marco Guermandi, Xiaying Wang, Michael Hersche, Robert Balas, Antonio Mastrandrea, Matheus Cavalcante, Angelo Garofalo, Alessio Burrello, Gianna Paulin, Georg Rutishauser, Andrea Cossettini, Luca Bertaccini, Maxim Mattheeuws, Samuel Riedel, Sergei Vostrikov, Vlad Niculescu, Hanna Mueller, Matteo Perotti, Nils Wistoff, Luca Bertaccini, Thorir Ingulfsson, Thomas Benz, Paul Scheffler, Alessio Burello, Moritz Scherer, Matteo Spallanzani, Andrea Bartolini, Frank K. Gurkaynak, and many more that we forgot to mention

http://pulp-platform.org  @pulp_platform