PULP: Looking Back and Looking Forward

Luca Benini  lbenini@ethz.ch, luca.Benini@unibo.it

PULP Platform
Open Source Hardware, the way it should be!
Looking Back: April 2012, Job Talk @ ETHZ

Digital Platform Design in the Twilight of Moore's Law

Luca Benini
Università di Bologna & STMicroelectronics

The Twilight of Moore’s Law: Power

- Transistor Scaling (Moore’s Law)
- Power Scaling (ITRS)
- Chip Power (ITRS)
- Dark Silicon!!!

Thermal wall: transistor count still increases exponentially but we can no longer power the entire chip (voltages, cooling do not scale)

The Twilight of Moore’s Law: IO Bandwidth

Memory wall: larger datasets and limited bandwidth at high power cost for accessing external memory

The Twilight of Moore’s Law: Economics

Market volume wall: only the largest volume products will be manufactured with the most advanced technology
Looking Back: April 2012, Job Talk @ ETHZ

STMicroelectronics’ Platform 2012

**GOPS/mm² – GOPS/W**

<table>
<thead>
<tr>
<th>1</th>
<th>3</th>
<th>6</th>
<th>&gt; 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>General-purpose Computing</td>
<td>Throughput Computing</td>
<td>SW</td>
<td>Mixed</td>
</tr>
<tr>
<td>CPU</td>
<td>GPGPU</td>
<td>1GOPS/mW</td>
<td></td>
</tr>
</tbody>
</table>

Closing The Accelerator Efficiency Gap

Heterogeneous, Accelerated Computing, 3D integration…

P2012 in a nutshell…

P2012 Fabric

Fabric Controller

Cluster

Cluster

Cluster

Cluster

3D-stackable SW accelerator

Customization design flow

ETHZH zurich
Looking Back: April 2012, Job Talk @ ETHZ

A Killer Application (domain) for P2012

P2012 was too early + Crashed against ARM dominance

A Killer Application (domain) for P2012

P2012 SoC in 28nm

- 4 Clusters, 69 processors
- 80 GFlops
- 1MB L2 mem
- 2D flip chip or 3D stacked
- 600 MHz typ
- < 2 W
- 3.7 mm² per cluster

Energy efficiency 40GOPS/W → 0.04GOPS/mW

The next killer app: Machines that see (J. Bier)
Looking Back: A few good ideas

Parallel, Ultra Low Power Processors Target pJ/OP @ GOPS and beyond

But for what?

And how to escape the proprietary ISA cage?
Looking Back: Serendipity!

**AI training:** 10x every year!!!


10x every 2 years

GPT-4 (OpenAI'23)

Training Compute: 2.1E+25 (FLOP)
Looking Back: Serendipity!
Looking Back: More Serendipity!

2023 RISC-V International more than 26% membership growth year-over-year, with over 3,180 members across 70 countries. More than 10 billion RISC-V cores in the market, 10K+ engineers working on RISC-V.
Open Source Hardware! $\rightarrow$ RTL source code (permissive*, e.g. Apache is key for industrial adoption)
Later stages contain closed IP of various actors $\rightarrow$ not open source by default (working on that…)

Open Source Platform

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

Platforms
- Single Core
  - PULPino
  - PULPissimo
- Multi-core
  - Open-PULP
  - PULP-PM
- Multi-cluster
  - Hero
  - MANTICORE

Peripherals
- JTAG
- SPI
- UART
- I2S
- DMA
- GPIO
- AXI4 – Interconnect
- APB – Peripheral Bus
- Logarithmic interconnect

IOT
- HWCE (convolution)
- Neurostream (ML)
- HWCrypt (crypto)
- PULPO (1st ord. opt)

HPC

Tens of active users, many use-cases
HW, SW specialization, verification, documentation, training

Cannot be sustained by one University, or two...
OpenHW Group is a not-for-profit, global organization (EU, NA, Asia) where HW and SW designers collaborate in the development of open-source cores, related IP, tools and SW such as the Core-V family.

OpenHW Group provides an infrastructure for hosting high quality open-source HW developments in line with industry best practices.

A Fast Growing Industrial Open Source Ecosystem

Rick O’Connor (OpenHW CEO, former RISC-V foundation director)

80+ members today
Creating Product Value with OSHW

**First iteration**: test-chip for IP qualification, early customer engagement (MPW)

**Second iteration**: first low volume production (most effort on c and d) (MLR or full mask set)

---

**NOTE** – **aggressive** (e.g. Greenwaves: IoT processor) vs. **cost-sensitive** fabless (e.g. Eggtronics: cellular charger IC) users

- **Aggressive**: customizing OSHW to provide differentiation wrt to ARM (differentiation). Targets advanced nodes
- **Cost-sensitive**: using OSHW “as is” to reduce cost wrt to ARM, and TtM, effort wrt to in-house, Targets older nodes
With a Little Help from my Friends...

Co-sponsored resources

Custom IP

Mr. Wolf
TSMC 40nm

Vega
GF 22nm

clone
issues

Now to Eric+Loic!
Forward to 2022: Job Done?
Forward to 2022: Job Done?
Forward to 2022: Job Done?

**RISC-V Cluster**
- Comparable 32bits-8bits SOA Energy efficiency to other PULPs [7]
- The highest energy efficiency on sub-byte SIMD operations (4b-2b)

**SNE**
- 1.7X higher than SOA [5] energy/efficiency

**CUTIE**
- 2X higher energy efficiency improvement over SOA [6]

CUTIE, SNE can work concurrently for SNN + TNN “fused” inference (never done so far)

Fast Forward: Perceptive $\rightarrow$ Generative $\rightarrow$ Embodied AI

Precise

Interactive, creative

Efficient, RT-safe, secure
Disruptive Embodied AI: Automotive

- **GF12, target 1GHz (typ)**
- 2 AXI NoCs (multi-hierarchy)
  - 64-bit
  - 512-bit with “interleaved” mode
- Peripherals
- Linux-capable manager core CVA6
- 6 Quadrants: 216 cores/chiplet
  - 4 cluster / quadrant:
    - 8 compute +1 DMA core / cluster
    - 1 multi-format FPU / core (FP64,x2 32, x4 16/alt, x8 8/alt)
- 8-channel HBM2e (8GB) 512GB/s
- D2D link (Wide, Narrow) 70+2GB/s
- System-level DMA
- SPM (2MB wide, 512KB narrow)

Peak 384 GDPflop/s per chiplet
Conclusion

• Efficient, RT, Safe Secure: PE, Cluster, SoC, System

• Key ideas
  • Deep PE optimization $\rightarrow$ extensible ISAs (RISC-V!)
  • Low-overhead work distribution. Latency hiding $\rightarrow$ large “mempools”
  • Heterogeneous architecture $\rightarrow$ host+accelerator(s)

• Game-changing technologies
  • “Commoditized” chiplets: 2.5D, 3D
  • Computing “at” memory (DRAM mempool)
  • Coming: optical IO and smart NICs, switches

• Challenges:
  • High performance RV Host
  • RV HPC software ecosystem?
  • Access to technology!
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.