



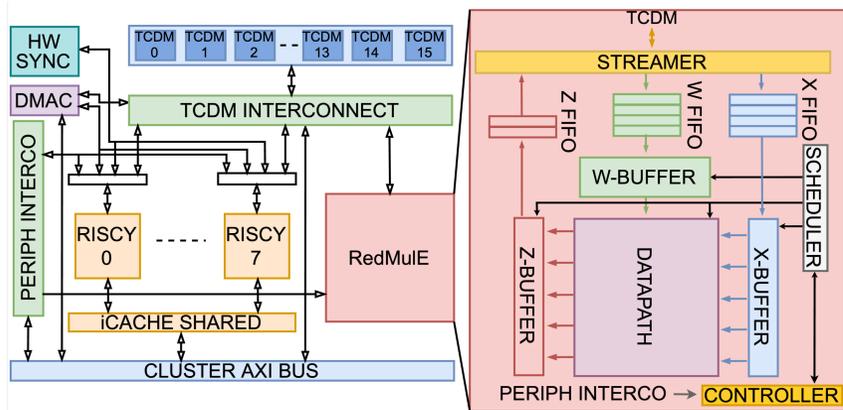
RedMule – Reduced-Precision Matrix Multiplication Engine

Yvan Tortorella, Luca Bertaccini, Davide Rossi, Luca Benini, Francesco Conti
University of Bologna
ETH Zurich



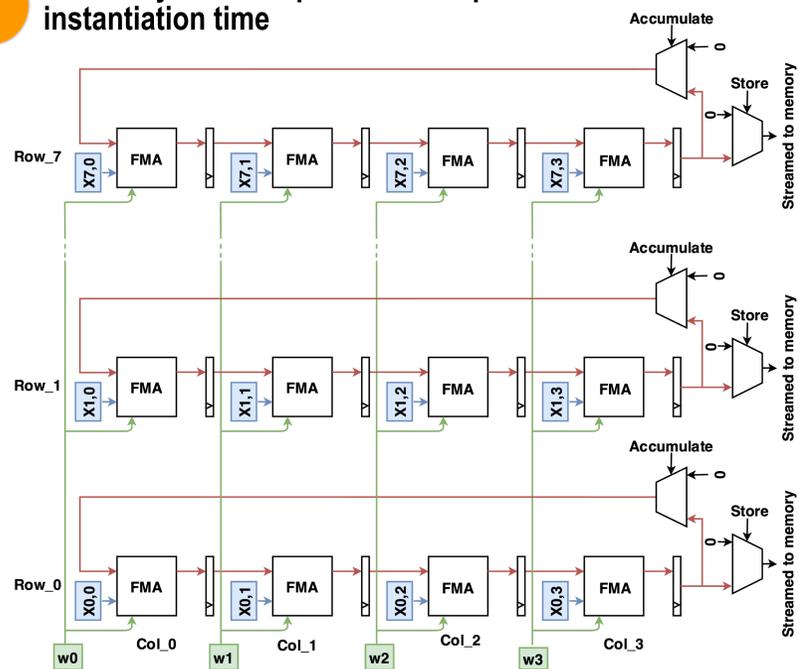
Introduction

- The request for specialized **energy-efficient** hardware capable of enabling **extreme-edge** Deep Learning (DL) is increasing
- Training on-the-edge** is **challenging** and hard to achieve in **sub-100 mW** domain due to **FP operations** requirement
- RedMule** accelerates **FP16** matrix multiplications to fasten **online adaptation** of **generalized DL models**



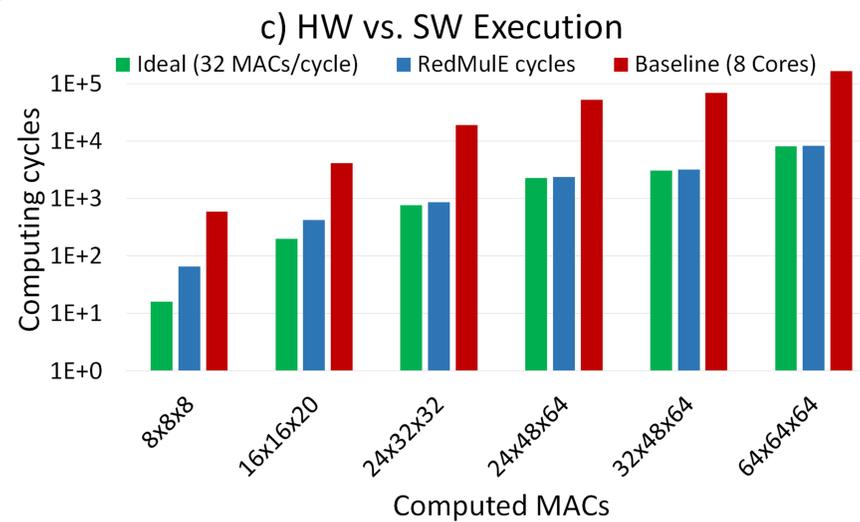
Design Overview

- Designed for integration in **PULP Clusters**, **RedMule** is based on a **semi-systolic array** made to maximize **data reuse**
- Geometry** and **data precision** are **parametric** and defined at **instantiation time**



Performance Results

- RedMule** introduces **22x** speedup over SW (parallelized on 8xRISCV-V cores), reaching **98,8%** of utilization (**31,6 MAC/cycle**)

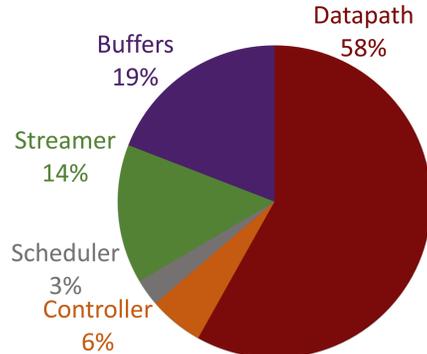


Area & Power Results

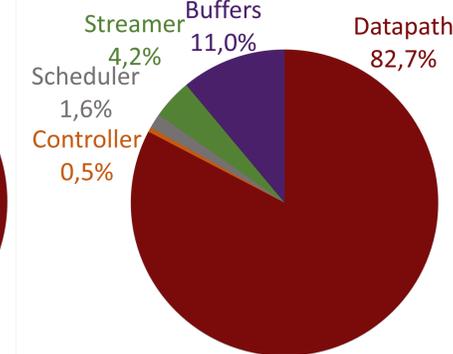
In 22 nm technology:

- RedMule** occupies **0.07 mm²**, meaning **14%** of the PULP Cluster (targeting 208 MHz at 125°C)
- Best energy efficiency point: 475 MHz, 688 GFLOPS/W, 43.5 mW** (average Cluster power consumption), **30 GFLOP/sec** (from post-layout power simulations)
- Best performance point: 666 MHz, 42 GFLOP/sec, 90.7 mW** (average Cluster power consumption), **432 GFLOPS/W** (from post-layout power simulations)

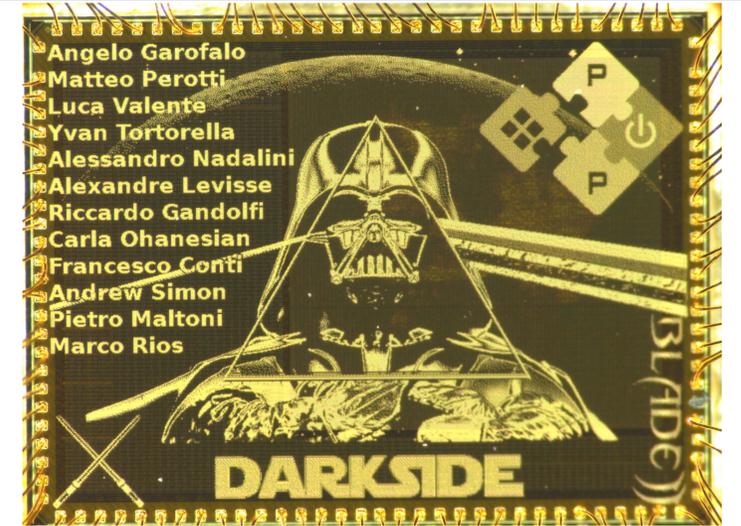
RedMule Area Breakdown



RedMule Power Breakdown



Darkside: PULP + RedMule in TSMC 65



State of the Art

The **State of the Art** (SoA) Comparison Table includes also **Darkside**, our PULP Cluster with RedMule prototyped in **65 nm** technology.

| Category | Design | Tech nm | Area mm2 | Freq MHz | Volt V | Power mW | Perf GOPS | Energy Eff GOPS/W | Mac Units | Precision |
|-----------------|-------------------|---------|----------|----------|--------|----------|-----------|-------------------|-----------|-----------|
| GPU | NVIDIA A100 [1] | 7 | - | 1410 | - | 300000 | - | - | 256 | FP16 |
| | Eyeriss [2] | 65 | 12.25 | 250 | 1.0 | 278 | 46 | 166 | 168 | INT16 |
| | EIE [3] | 45 | 40.8 | 800 | - | 590 | 102 | 173 | 64 | INT8 |
| Inference Chips | Zeng et al. [4] | 65 | 2.14 | 250 | - | 478 | 1152 | 2410 | 256 | INT8 |
| | Simba [5] | 16 | 6 | 161 | 0.42 | - | - | 9100 | - | 1024 |
| Training Chips | IBM [6] | 7 | 19.6 | 1000 | 0.55 | 4400 | 8000 | 1800 | 4096 | FP16 |
| | Cambricon-Q [7] | 45 | 888 | 1000 | 0.6 | 133 | 2000 | 2240 | 1024 | INT8 |
| HPC | Manticore [8] | 22 | 888 | 500 | 0.6 | 200 | 25 | 188 | 24 | FP64 |
| | | 1000 | 0.9 | 900 | 54 | 50 | - | - | - | - |
| Mat-Mul Acc. | Anders et al. [9] | 14 | 0.024 | 2.1 | 0.26 | 0.023 | 0.068 | 2970 | 16 | FP16 |
| | | 1090 | 0.9 | 82.7 | 34 | 420 | - | - | - | - |
| Our Work | PULP (w/ RedMule) | 22 | 0.5 | 476 | 0.65 | 43.5 | 30 | 688 | 32 | FP16 |
| | Darkside | 65 | 3.85 | 200 | 1.2 | 89.1 | 12.6 | 152 | 32 | FP16 |

Acknowledgement

This work was supported in part by the EU H2020 "WiPLASH" (g.a. 863337), by the ECSEL H2020 "AI4DI" (g.a. 826060), and by Thales Alenia Space.

References

- [1] J. Choquette et al., 2021 IEEE International Solid-State Circuits Conference (ISSCC), 2021
- [2] Y. Chen et al., IEEE Journal of Solid-State Circuits (Volume: 52, Issue: 1, Jan. 2017), 2017.
- [3] S. Han et al., ISCA, 2016.
- [4] Y. Zeng et al., IEEE International Symposium on Circuits and Systems (ISCAS), 2021.
- [5] Y. S. Shao et al., MICRO '21: Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture, 2019.
- [6] A. Agrawal et al., 2021 IEEE International Solid-State Circuits Conference (ISSCC), 2021.
- [7] Y. Zhao et al., 2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA), 2021.
- [8] F. Zaruba et al., IEEE Micro (Volume: 41, Issue: 2, March-April 1), 2021.
- [9] M. Anders et al., IEEE Symposium on VLSI Circuits, 2018.