RedMulE: A Compact FP16 Matrix-Multiplication Accelerator for Adaptive Deep Learning on RISC-V-Based Ultra-Low-Power SoCs

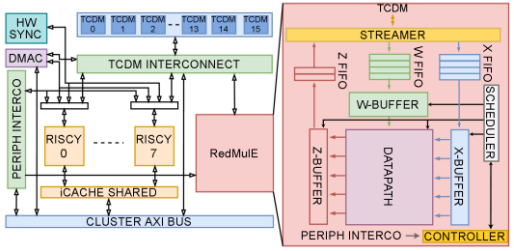
Yvan Tortorella, Luca Bertaccini, Davide Rossi, Luca Benini, Francesco Conti

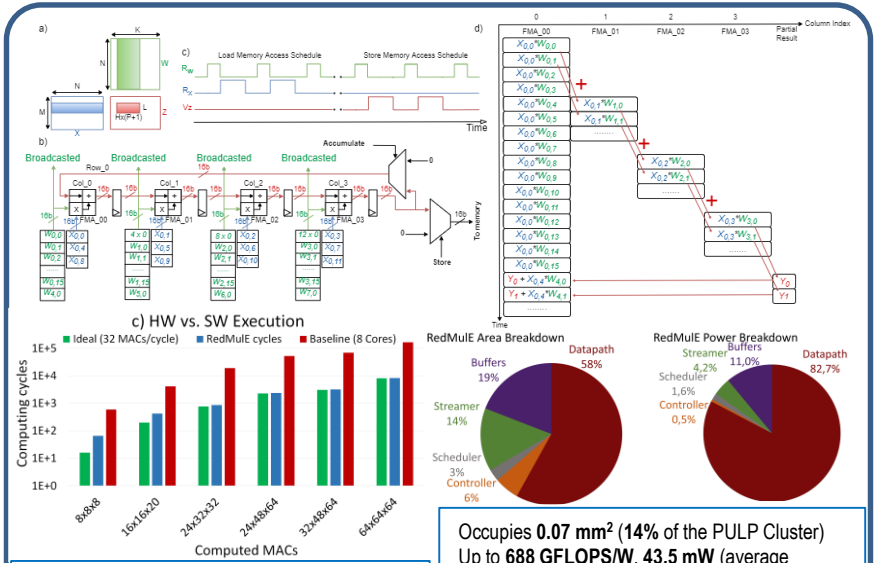
The request for **Energy-efficient** hardware enabling **extreme-edge** Deep Learning (DL) is increasing

Narrow integer operations suffice for extreme-edge **inference**, helping reducing avarege power consumption and increase energy efficiency

Training on-the-edge is **challanging** and hard to achieve in **sub-100 mW** domain due to **FP operations** requirement for accuracy and precision

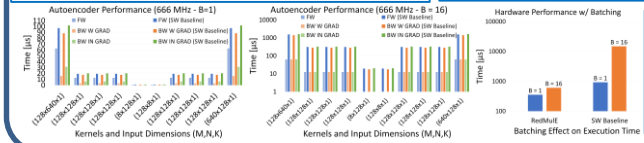
RedMulE accelerates **FP16** matrix multiplications to fasten **online training** of **genealized DL models**
First PULP-Based reduced precision matrix multiplication accelerator enabling **extreme-edge training** at very **low HW cost**



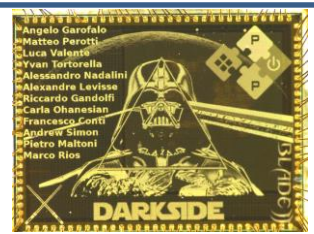


Up to **22x** HW speedup and **31.6 MAC/cycle (98.8%** utilization), **42 GFLOPS** at **666 MHZ**

Occupies **0.07 mm²** (**14%** of the PULP Cluster)
Up to **688 GFLOPS/W**, **43.5 mW** (average Cluster power consumption) at **475 MHz**, with **30 GFLOP/sec** (from **post-layout** power simulations)



2.6x Speedup over SW (**B = 1**)
24.4x speedup over SW (**B = 16**)



Cathegory	Design	Tech nm	Area mm2	Freq MHz	Volt V	Power mW	Perf GOPS	Energy Eff GOPS/W	Mac Units	Precision
GPU	*NVIDIA A100*	7	-	1410	-	300000	-	-	256	FP16
Inference Chips	*Eyeriss*	65	12.25	250	1.0	278	46	166	168	INT16
	EIE	45	40.8	800	-	590	102	173	64	INT8
	Zeng et al.	65	2.14	250	-	478	1152	2410	256	INT8
	Simba	16	6	161	0.42	-	-	9100	1024	INT8
				2000	1.2	-	4000	-		
Training Chips	*IBM*	7	19.6	1000	0.55	4400	8000	1800	4096	FP16
				1600	0.75	13000	12800	980		
	Cambricon-Q	45	888	1000	0.6	133	2000	2240	1024	INT8
HPC	*Manticore*	22	888	500	0.6	200	25	188	24	FP64
				1000	0.9	900	54	50		
Mat-Mul Acc.	*Anders et al.*	14	0.024	2.1	0.26	0.023	0.068	2970	16	FP16
				1090	0.8	82.7	34	420		
Our Work	*PULP (w/ RedMulE)*	22	0.5	476	0.65	43.5	30	688	32	FP16
				666	0.8	90.7	42	462		
	Darkside	65	3.85	200	1.2	89.1	12.6	152	32	FP16

We presented **RedMulE**, an example of accelerator to enable **on-chip leaning** efficiently at **low HW cost**

Future work:

- explore HW solution for efficient **inference and training** on lower precision (e.g. **Hybrid-FP8**)
- Introduction of **reconfiguration** features (e. g. **zero skipping**)