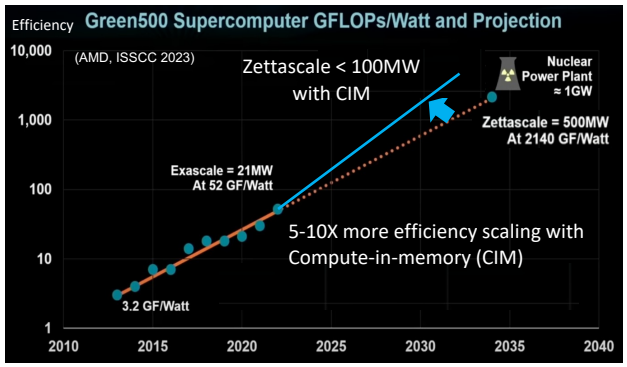
(*Presentation: Wednesday 2/23 5:20pm*) **Paper No. : 30.6**

Vecim: A 289.13GOPS/W RISC-V Vector Co-processor with Compute-in-memory Vector Register File for Efficient High-performance Computing

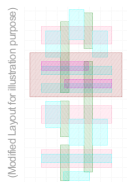
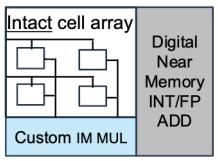
Yipeng Wang(yipeng.wang@utexas.edu), Mengtian Yang, Chieh-pu Lo, Jaydeep Kulkarni(jaydeep@austin.utexas.edu)
Circuit Research Lab, University of Texas, Austin, TX

Motivation

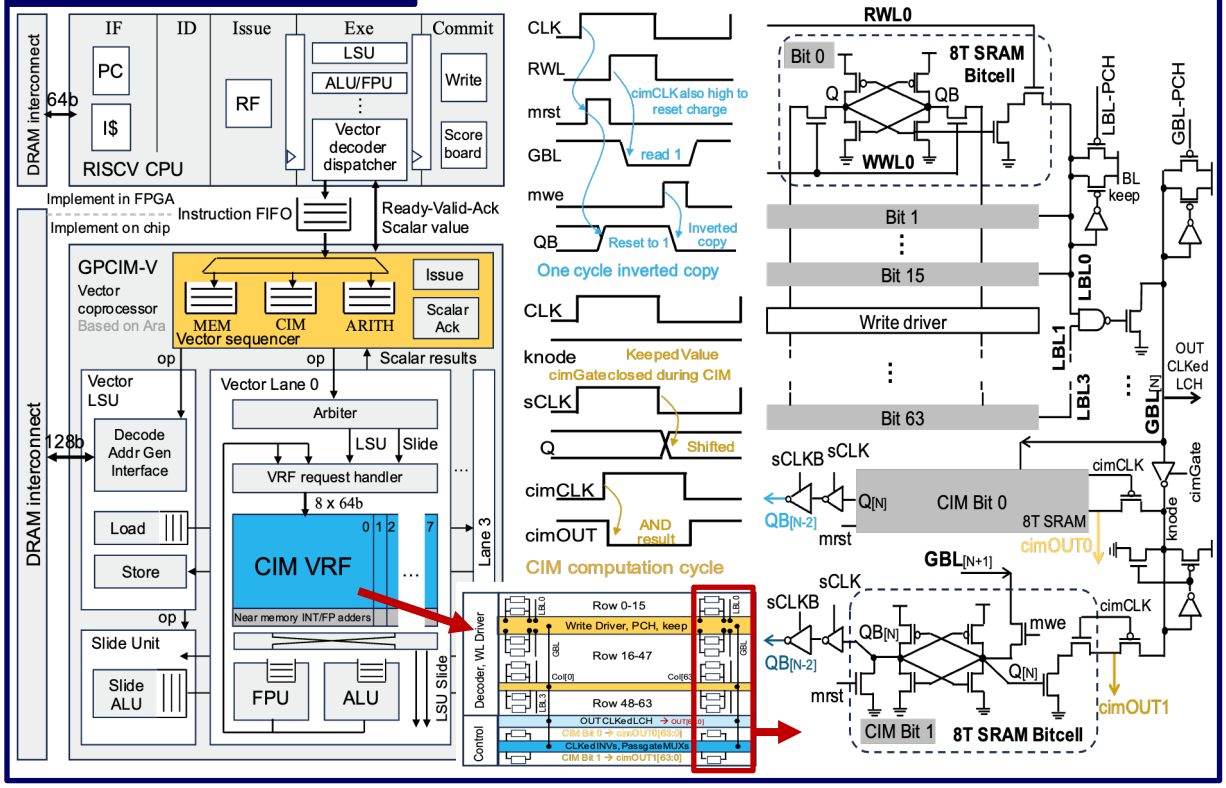
Mitigate the efficiency gap of HPC:



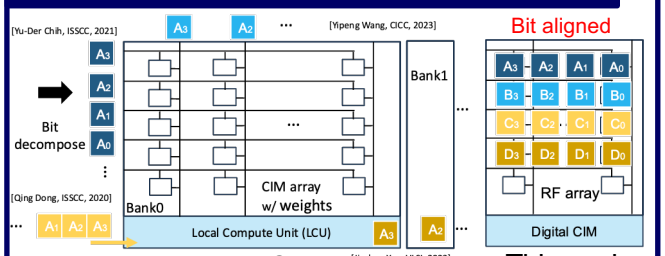
Compute-in-memory application challenges:

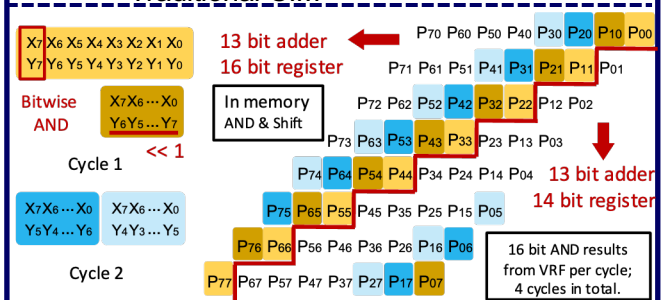
Large area footprint	Low Frequency Low Robustness	Low Programmability
		RISC-V® RVV0.9 ISA supported (Fine, it's partial supported Mask operation is not full)
Use foundry 8T cell	Pipelined MAC Near memory FPU	CIM: RISCV Vector Processor RF

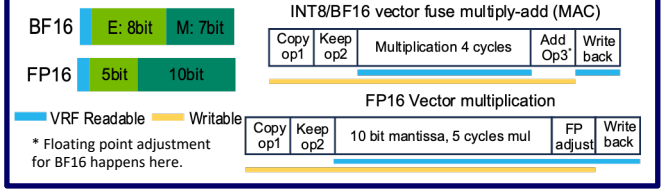
Architecture and Circuit



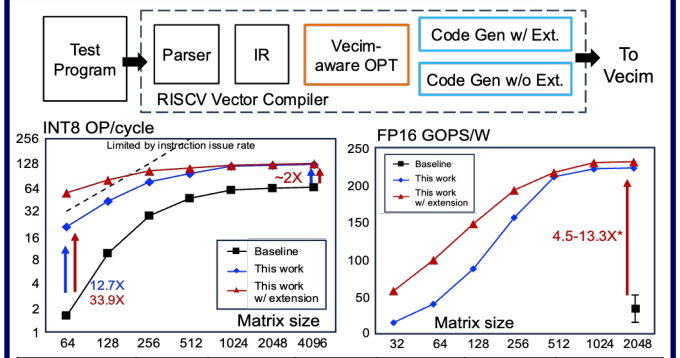
New Dataflow, Floating point support







Compiler, tests and comparison



	Intel Xeon Max 9480 CPU	NVIDIA H100 GPU	Ara [2]	ISSCC 2019 [5]	VLSI 2023 [6]	This work
Technology	Intel 7 (10nm)	TSMC 4nm	GF 22nm	TSMC 28nm	TSMC 65nm	TSMC 65nm
ISA	X86	CUDA	RISCV	Customed	Customed	RISCV
Category	General purpose processors			Customed CIM design with general purpose operations		General purpose processor
CIM type	-	-	-	Custom 8T	Custom 8T/9T	Foundry 8T
Bit precision	All	All	All	All (potentially)	INT8/32	(INT8/BF16/FP16 enhanced ALL)
Design level	System	System	Processor (simulation)	Macro + ctrl	Macro + ctrl	Co-processor
Nominal frequency	1.9/3.5GHz	1.83GHz	1.25GHz	475MHz	200MHz	250MHz
Nominal voltage			0.8V	1.1V	1V	1V
Die area	400mm^2	819mm^2	1.125mm^2 (No pads)	2.7mm^2	1.512mm^2	4mm^2
Peak performance INT8/FP16	79.8 / 39.9 T(FL)OPS	3000 / 1979 T(FL)OPS	78.4 / 39.2 G(FL)OPS	40.5 / ~0.5 G(FL)OPS	2.3*4 / X GOPS	31.8 / 25.3 G(FL)OPS
Energy efficiency INT8/FP16 (power normalized to 65nm^2)	5.44 / 2.60 G(FL)OPS/W	32.45 / 21.39 G(FL)OPS/W	34.64 / 17.32 G(FL)OPS/W	439.78 / ~5 G(FL)OPS/W	473.35 / X GOPS/W	289.13/230.10 G(FL)OPS/W
Area efficiency INT8/FP16 (area normalized to 65nm)	4.7 / 2.4 M(FL)OPS/mm^2	13.86 / 8.1 M(FL)OPS/mm^2	7.98 / 4.00 *3 G(FL)OPS/mm^2	2.78 / 0.034 G(FL)OPS/mm^2	1.52 / X GOPS/mm^2	7.95 / 6.33 G(FL)OPS/mm^2

Demo Details