

The BRAM is the Limit: Shattering Myths, Shaping Standards, and Building Scalable PIM Accelerators

MD Arafat Kabir, Tendayi Kamucheka, Nathaniel Fredricks, Joel Mandebi, Jason Bakos, Miaoqing Huang, and David Andrews

Introduction

- Existing PIM Architectures promise impressive parallelism but lack desired performance and scalability.
- The community has accepted these limitations without empirical study.
- This study establishes a true theoretical upper limit of FPGA-PIM accelerators, we call a Gold Standard.
- IMAGine, a GEMV PIM accelerator, is developed that achieves the proposed standard.

Gold Standard

- PIM accelerators must run at the maximum BRAM frequency.
- PIM array peak-performance must scale linearly with the on-chip BRAM.
- Reduction latency must fit $aN \log(P) + bP + c$ with parameters a, b, and c within the ideal range to balance logic utilization and latency.

Results

- IMAGine achieves the Gold Standard.
- No other existing FPGA PIMs achieve Gold Standard.
- IMAGine clocks faster than TPU v1-v2 and Alibaba Hanguang 800.
- IMAGine is fast and scalable due to targeting the Gold Standard.

A Gold Standard for FPGA-PIM Designs The Fastest GEMV PIM Accelerator Beating ASIC with FPGA Overlay

- How Fast : BRAM F_{max} , 737 MHz on US+
- How Big : BRAM 100%, 64K MAC on U55
- Why Care: Clocks Faster than TPU v1-v2

Existing PIM Designs

PIM Design	Type	Device	F_{BRAM}	F_{PIM}	Rel.	F_{sys}	Rel.
CCB	Custom	Stratix 10	1000	624	62%	455	46%
CoMeFa-A	Custom	Arria 10	730	294	40%	288	39%
CoMeFa-D	Custom	Arria 10	730	588	81%	292	40%
BRAMAC-2SA	Custom	Arria 10	730	586	80%	-	-
BRAMAC-1DA	Custom	Arria 10	730	500	68%	-	-
M4BRAM	Custom	Arria 10	730	553	76%	-	-
SPAR-2	Overlay	UltraScale+	737	445	60%	200	27%
PiMulator	Overlay	UltraScale+	737	-	-	333	45%
PiCaSO	Overlay	UltraScale+	737	737	100%	-	-

Path Delay Breakdown

	FF-C2Q ¹	LUT ²	FF-Setup	Total ³	BRAM ⁴	Net Budget	Min ⁵
V7	0.290	0.34	0.255	0.885	1.839	0.954	0.272
US+	0.087	0.15	0.098	0.335	1.356	1.021	0.102

¹ Clock-to-Q delay of flip-flops
² Average delay through LUTs
³ Total cell delay
⁴ BRAM pulse-width requirement, clock period for Fmax
⁵ Minimum net delay through a switchbox

Reduction Parameters

Parameter	Ideal Range	Related to
a	$1/N \leq a \leq 2$	Latency of reduction steps (addition)
b	$0 \leq b \leq 1$	Latency of data movement
c	$0 \leq c$	Cycles spent outside reduction network

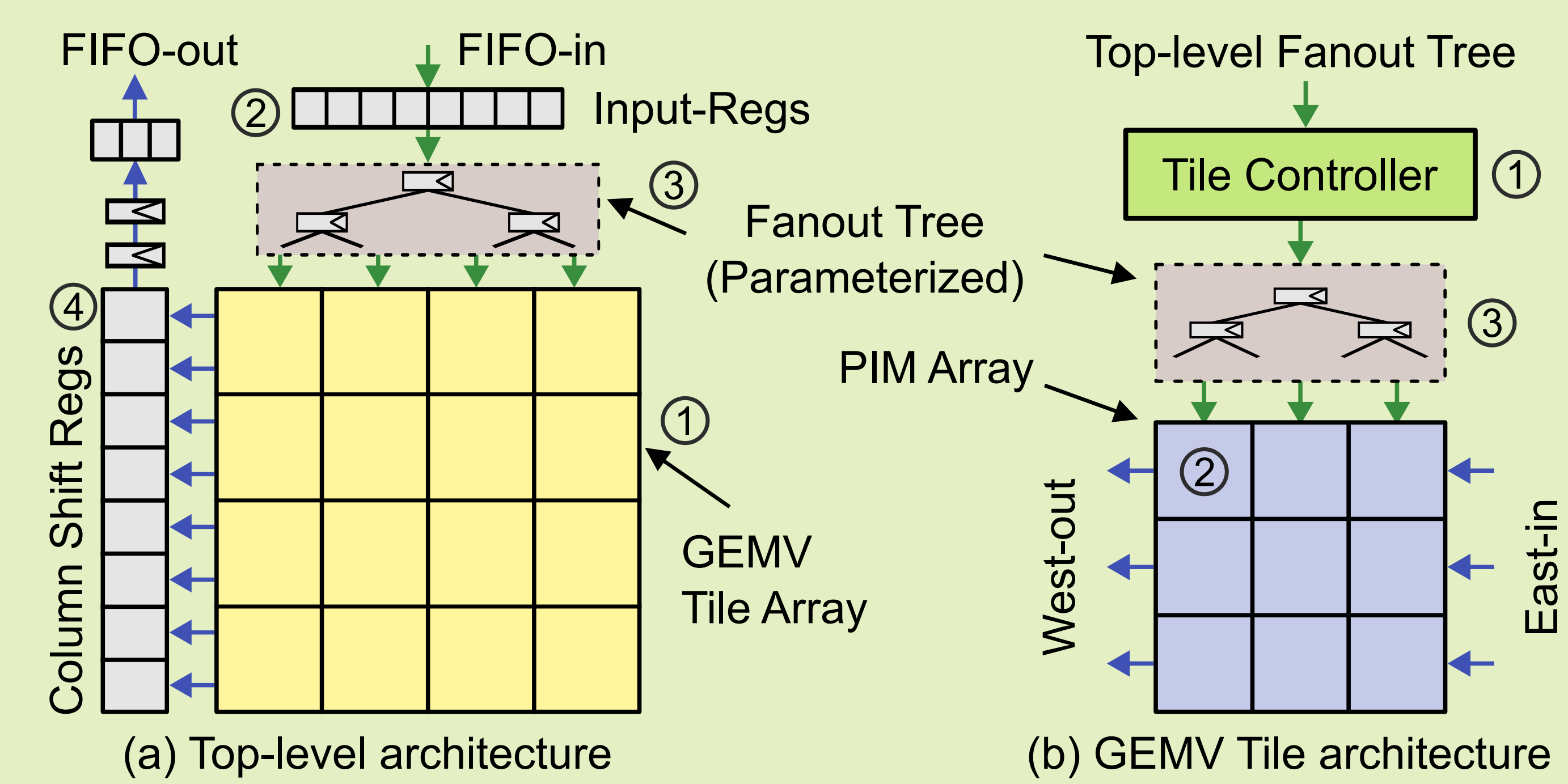
12x2 GEMV Tile

	Controller	Rel.	Fanout	Rel.	PIM Array	Rel.	Tile
LUT	167	5.8%	0	0.0%	2736	94.2%	2903
FF	155	4.0%	615	15.9%	3096	80.1%	3866
DSP	0	-	0	-	0	-	0
BRAM	0	0.0%	0	0.0%	12.0	100.0%	12
Freq. (MHz)	890	1.2x	890	1.2x	737	1x	737

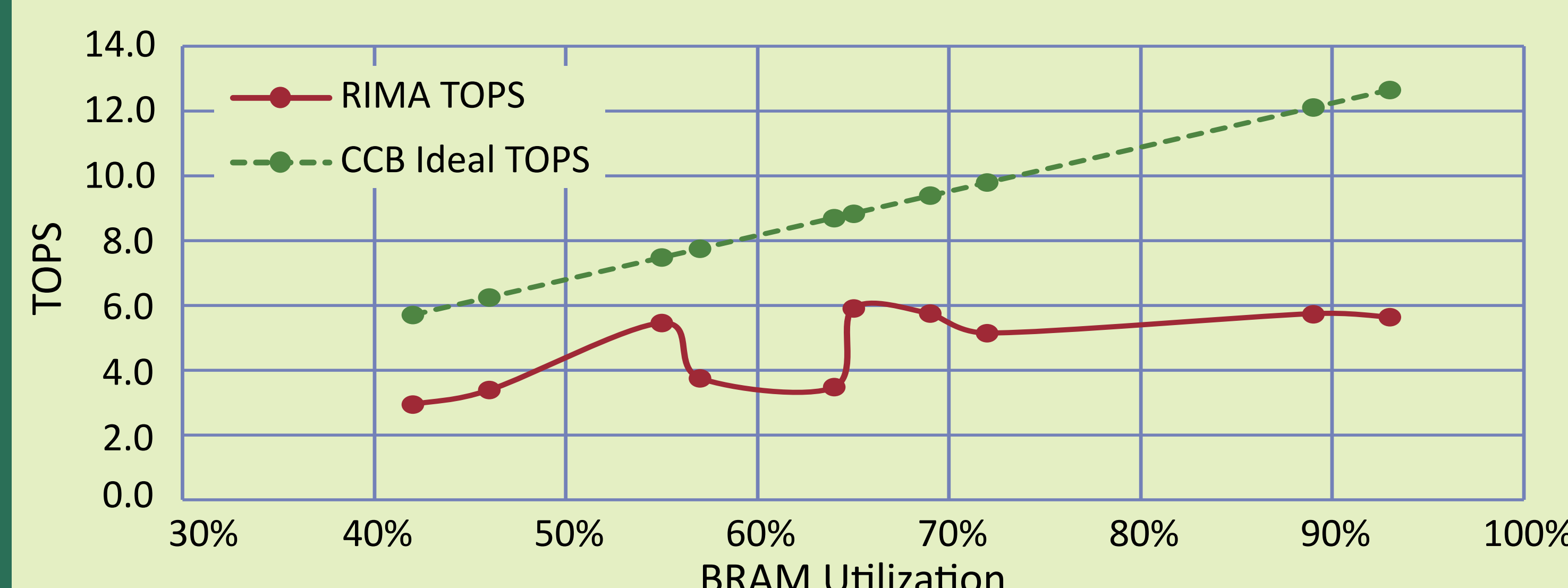
GEMV Engine Compare

	LUT	FF	DSP	BRAM	F_{sys}^1	Rel. Freq
RIMA-Fast	60%		50%	55%	455	45.5%
RIMA-Large	89%		50%	93%	278	27.8%
CCB GEMV	27.9%		90.1%	91.8%	231	31.6%
CoMeFa-A GEMV	27.9%		90.1%	91.8%	242	33.2%
CoMeFa-D GEMV	25.5%		92.4%	86.7%	267	36.6%
SPAR-2 (US+)	11.3%	2.4%	0.0%	14.5%	200	27.1%
SPAR-2 (V7)	28.5%	7.0%	0.0%	30.4%	130	23.9%
IMAGine	35.6%	24.8%	0.0%	100.0%	737	100.0%
IMAGine-CB ²	10.1%	7.2%	0.0%	100.0%	737	100.0%

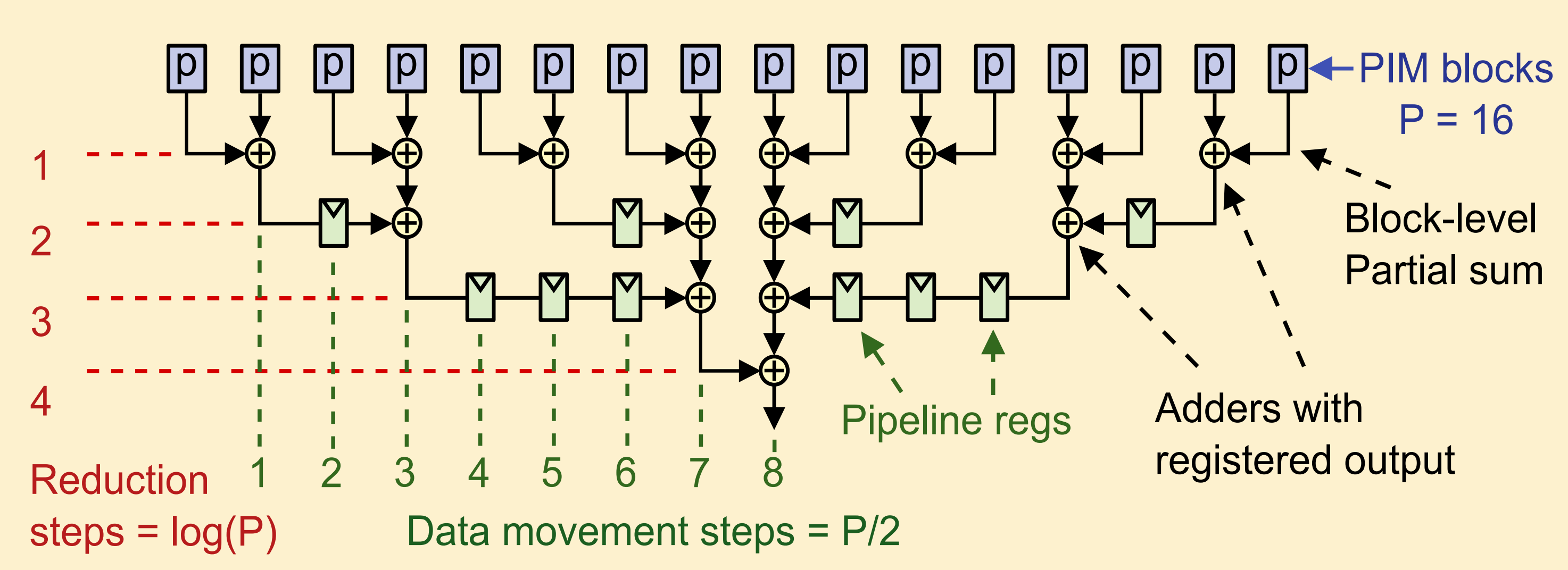
¹ System frequency in MHz
² IMAGine with custom-BRAM PiCaSO-F (PiCaSO-CB)



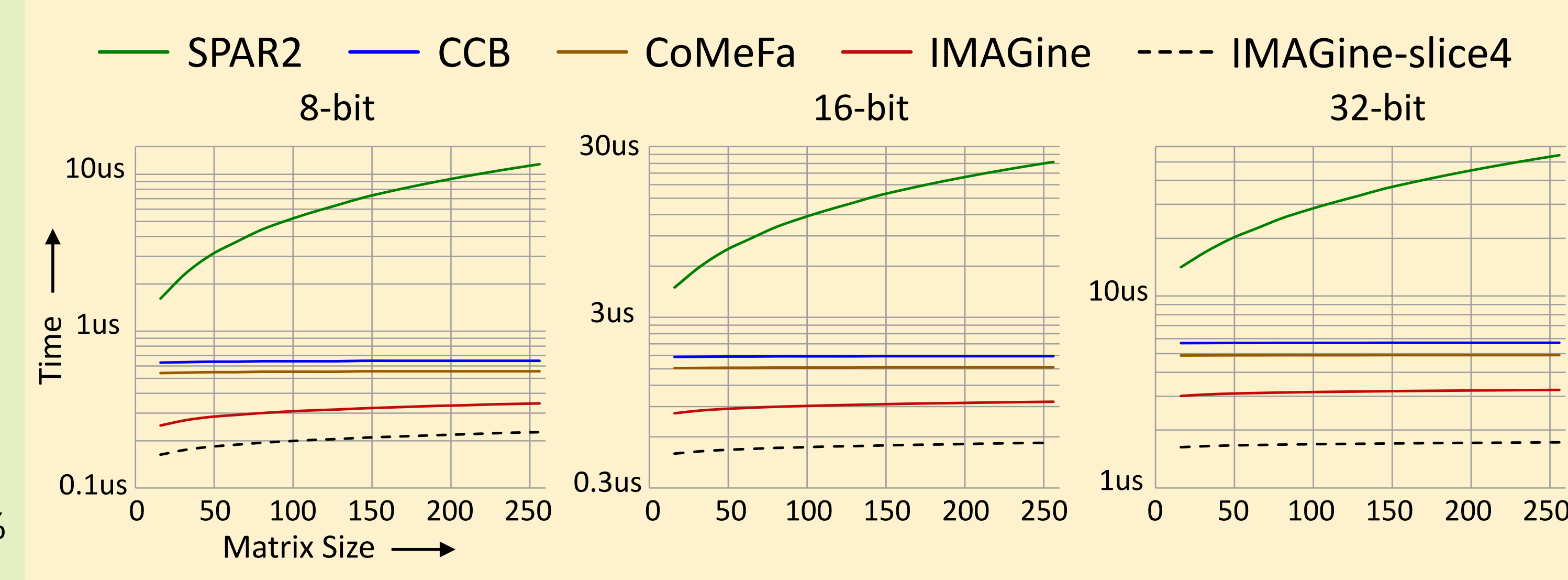
System Architecture of IMAGine




Linear Scaling of Peak-Performance (RIMA)



Pipelined Reduction Tree to Achieve BRAM Fmax



GEMV Execution latency on PIM Array Accelerators



Download This Poster

