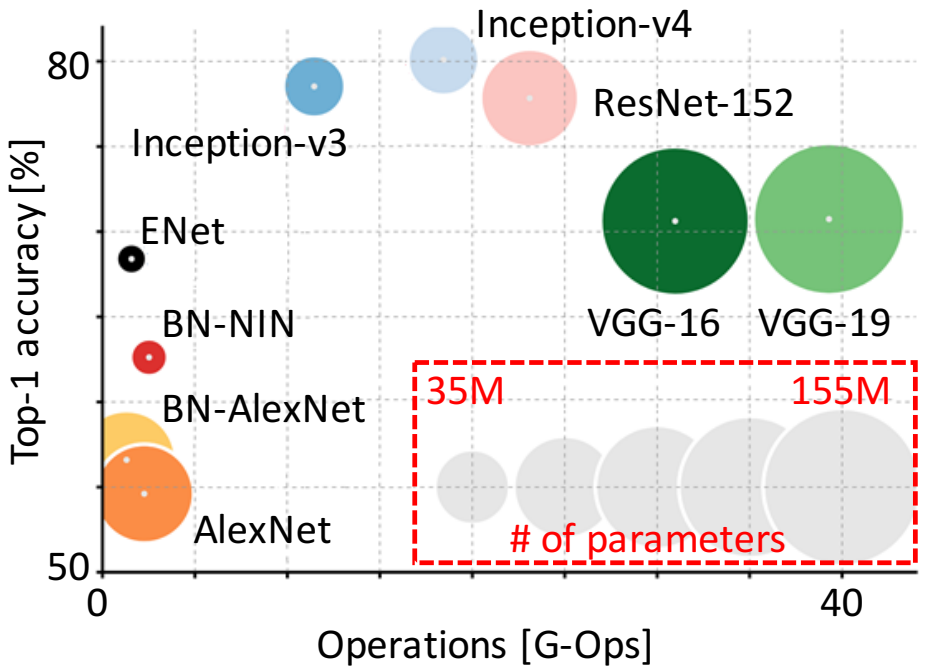
The Deep In-memory Architecture for Energy Efficient Machine Learning

Naresh Shanbhag

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign

The Energy Cost of Data Movement

$$\frac{E_{mem}}{E_{mac}} \approx \sim 100\times \text{ (SRAM)} \rightarrow \sim 500\times \text{ (DRAM)} \rightarrow \sim 1000\times \text{ (Flash)}$$



[Canziani, Arxiv'16]

Memory Energy (45nm)

Memory	
Cache	(64bit)
8KB	10pJ
32KB	20pJ
1MB	100pJ
DRAM	1.3-2.6nJ

Computation Energy (45nm)

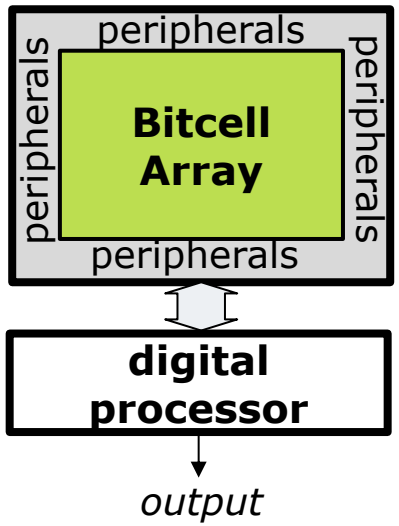
Integer	
Add	
8 bit	0.03pJ
32 bit	0.1pJ
Mult	
8 bit	0.2pJ
32 bit	3 pJ

FP	
FAdd	
16 bit	0.4pJ
32 bit	0.9pJ
FMult	
16 bit	1pJ
32 bit	4pJ

[*Horowitz, ISSCC'14*]

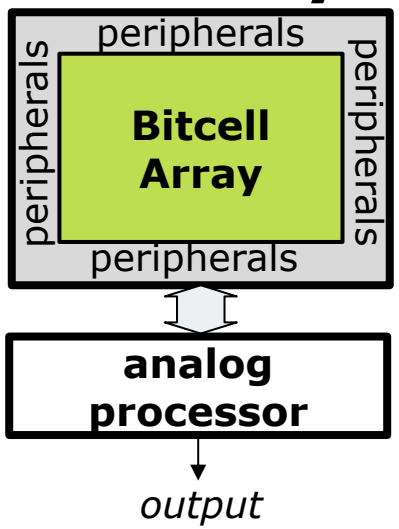
Inference Architectures

digital



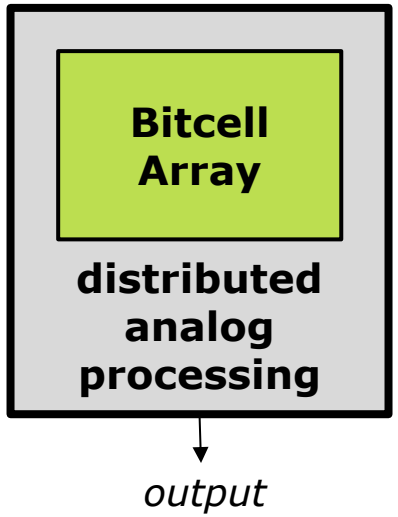
- memory accesses dominates energy & latency costs
- data reuse & comp. alleviate costs

analog near memory



- compute energy (↓)
- memory accesses still dominate energy & latency costs

in-memory



- merges memory access and compute
- energy & latency (↓)
- compute density (↑)

SONIC

Systems on Nanoscale Information fabriCs

www.sonic-center.org

[2013-'17]

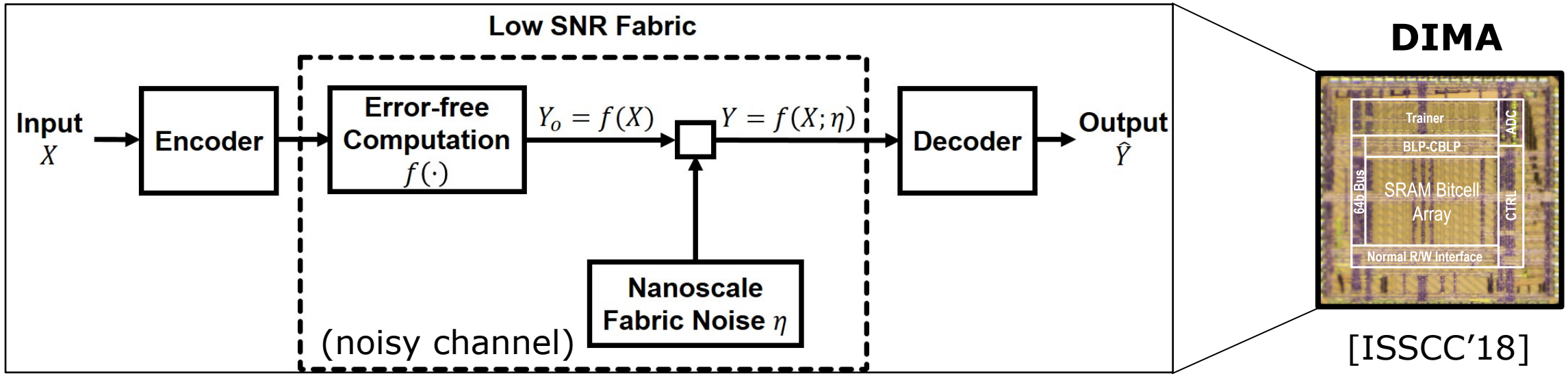
(STARnet Program by Semiconductor Research Corporation & DARPA)

Shannon-Inspired Statistical Computing for the Nanoscale Era

By Naresh R. Shanbhag, *Fellow IEEE*, Naveen Verma, *Member IEEE*, Yongjune Kim, *Member IEEE*, Ameya D. Patil, *Student Member IEEE*, and Lav R. Varshney, *Senior Member IEEE*

- Proceedings of IEEE, Special Issue on *Non von Neumann Computing*, 2018.

Shannon-inspired Model of Computing



1. use **information-based metrics** e.g., mutual information $I(Y_o;\hat{Y})$
2. design **low SNR fabrics**, e.g., deep in-memory architecture (DIMA)
3. develop **statistical error-compensation (SEC)** techniques

23 Mar 2018 | 15:00 GMT

To Speed Up AI, Mix Memory and Processing

New computing architectures aim to extend artificial intelligence from the cloud to smartphones

By **Katherine Bourzac**

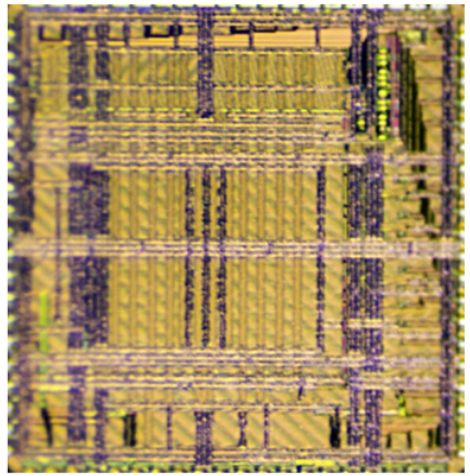


Image: Sujan Gonugondla

Tearing Down Walls: This prototype features a new chip design called deep in-memory architecture.

If John von Neumann were designing a computer today, there's no way he would build a thick wall between processing and memory. At least, that's what computer engineer Naresh Shanbhag of the University of Illinois at Urbana-Champaign believes. The eponymous von Neumann architecture was published in 1945. It enabled the first stored-memory, reprogrammable computers—and it's been the backbone of the industry ever since.

Now, Shanbhag thinks it's time to switch to a design that's better suited for today's data-intensive tasks. In February, at the International Solid-State Circuits Conference (ISSCC), in San Francisco, he and others made their case for a new architecture that brings computing and memory closer together. The idea is not to replace the processor altogether but to add new functions to the memory that will make devices smarter without requiring more power.

The Deep In-memory Architecture (DIMA)

"breaching the memory wall"

[Verma, Shanbhag]

https://spectrum.ieee.org/computing/hardware/to-speed-up-ai-mix-memory-and-processing

Timeline

2014 — DIMA concept paper (ICASSP)[UIUC]

2016 — AdaBoost DIMA IC (VLSI)[Princeton]

2017 — Multifunction & Random Forest DIMA ICs (ESSCIRC)[UIUC]

DIMA US Patent [UIUC]

2018 — In-memory Special Session/Forums (ISSCC,VLSI,ICCAD)

SGD-SVM DIMA IC (ISSCC)| Scalable DIMA IC (VLSI)[Princeton]

In-memory ICs (ISSCC) [MIT,NTHU,NTHU-ASU]

In-memory ICs (VLSI)[Columbia-ASU]

2019 — In-memory ICs (ISSCC)[UMich, NTHU-ASU-GIT, SU(China)], ISSCC Forum…….

DIMA @ UIUC – Concept to ICs

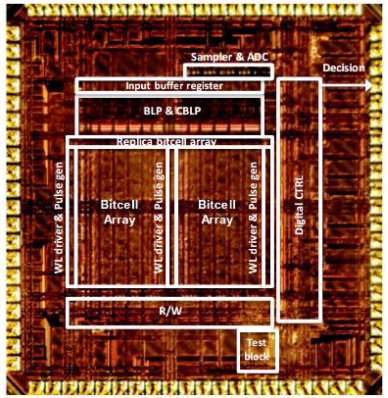
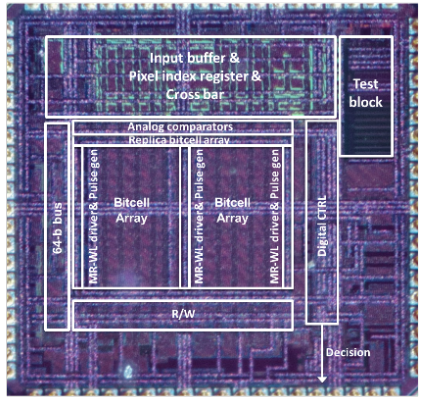
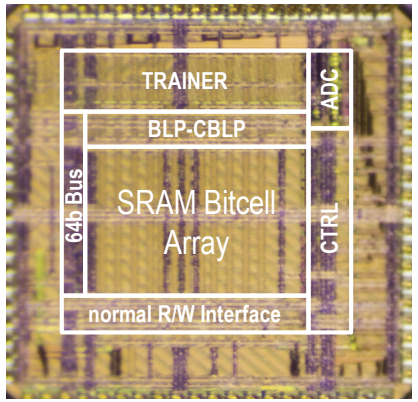
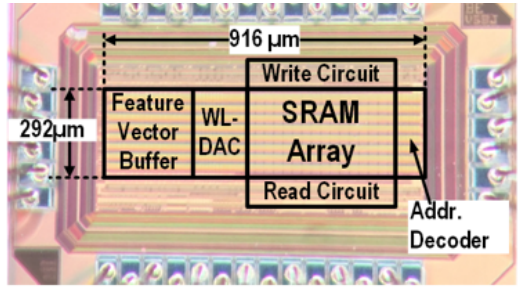
- M. Kang, M.-S. Keel, N. Shanbhag, S. Eilert, and K. Curewitz, "An energy-efficient VLSI architecture for pattern recognition via deep embedding of computation in SRAM," *ICASSP,* Florence, Italy, May 2014.
- M. Kang, S. Gonugondla, M.-S. Keel, and N. Shanbhag, "An energy-efficient memory-based high-throughput, VLSI architecture for convolutional networks,"*ICASSP*, Australia, April 2015.
- M. Kang, E. P. Kim, M.-S. Keel, and N. Shanbhag, "Energy-efficient and high throughput sparse distributed memory architecture," *ISCAS*, Lisbon, Portugal, May 24-27, 2015. [**Best Paper Award**]
- M. Kang and N. Shanbhag, "In-memory computing architectures for sparse distributed memories," *IEEE Trans. on BIOCAS*, 2016.
- M. Kang, S. Gonugondla, A. Patil, and N. R. Shanbhag, "A multi-functional in-memory inference processor using a standard 6T SRAM array," *IEEE Journal of Solid-State Circuits*, February 2018.
- M. Kang, S. Gonugondla, N. Shanbhag, "A 19.4 nJ/decision 364K decisions/s in-memory random forest classifier in 6T SRAM array", *ESSCIRC*, 2017 [*JSSC* version, May 2018]
- S. Gonugondla, M. Kang, and N. Shanbhag, "A 42pJ/decision 3.12 TOPS/W robust in-memory machine learning classifier with on-chip training," *ISSCC 2018*. [*JSSC* version Nov. 2018]
- M. Kang, S. Lim, S. Gonugondla, and N. Shanbhag, "An in-memory VLSI architecture for convolutional neural networks" *IEEE Journal on Emerging and Selected Topics in Circuits and Systems,* Sept. 2018.
- P. Srivastava, M. Kang, S. Gonugondla, S. Lim, J. Choi, V. Adve, N. S. Kim, N. Shanbhag,"PROMISE: An end-to-end design of a programmable mixed-signal accelerator for machine learning algorithms," *ISCA* 2018.
- S.K. Gonugondla, M. Kang, Y. Kim, M. Helm, S. Eilert, and N. Shanbhag, "Energy-efficient deep in-memory architectures for NAND flash memories," *ISCAS*, Italy, May 2018. [**Best Paper Award**]
- Y. Kim, M. Kang, L. Varshney, and N. Shanbhag, "Generalized water-filling for source-aware energy-efficient SRAMs," *IEEE Trans. On Communications*, Oct. 2018.

6T SRAM DIMA Prototypes

100X EDP reduction over von Neumann equivalent* @ iso-accuracy

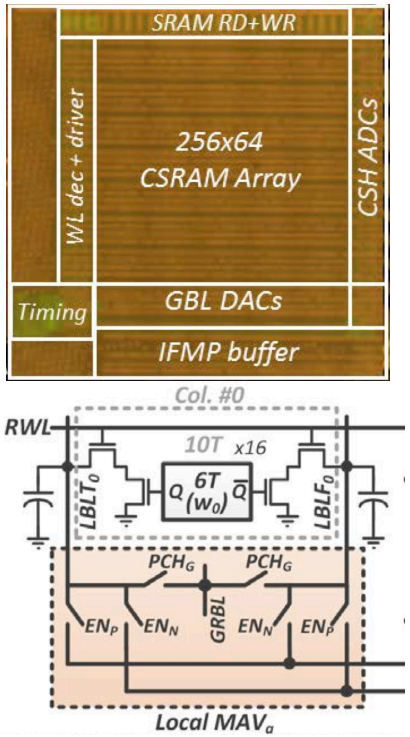
* 8b fixed-function digital architecture with identical SRAM size

8b compute; 16kB SRAM in 65nm CMOS (UIUC)

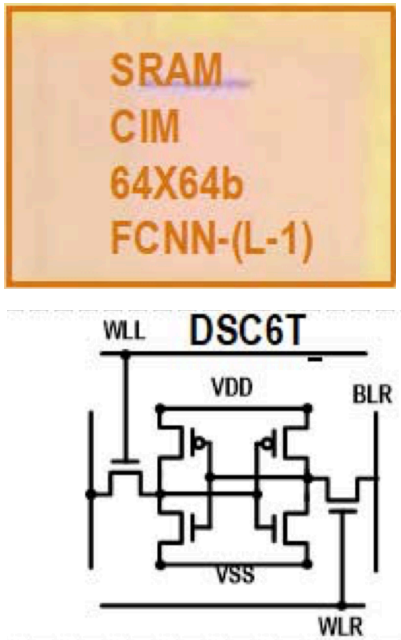
multi-functional	random forest	on-chip learning	RNN attention ntwk	AdaBoost
				
[JSSC'18]	[ESSCIRC'17 JSSC'18]	[ISSCC'18, JSSC'18]	[On-going(UIUC)]	(130nm CMOS) [VLSI'16, JSSC'17 (Princeton)]

Other In-memory Works - 6T+ Bitcells

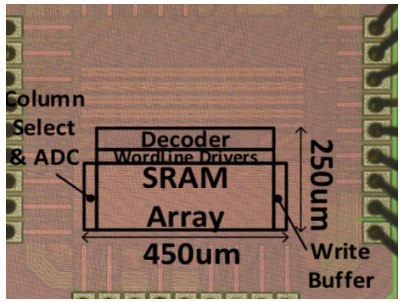
[ISSCC'18(MIT)]

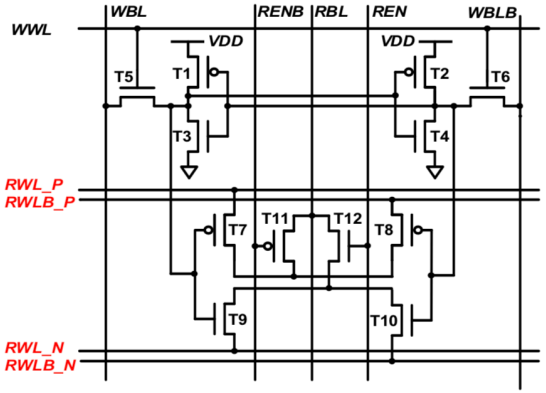


[ISSCC'18(NTHU, UESTC,ASU)]

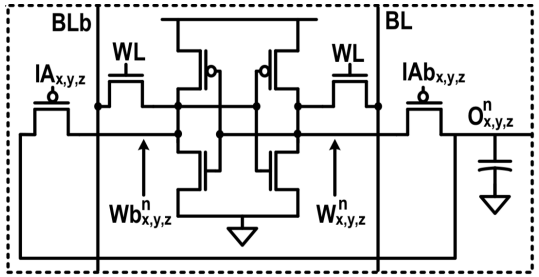


[VLSI'18(Columbia,ASU)]





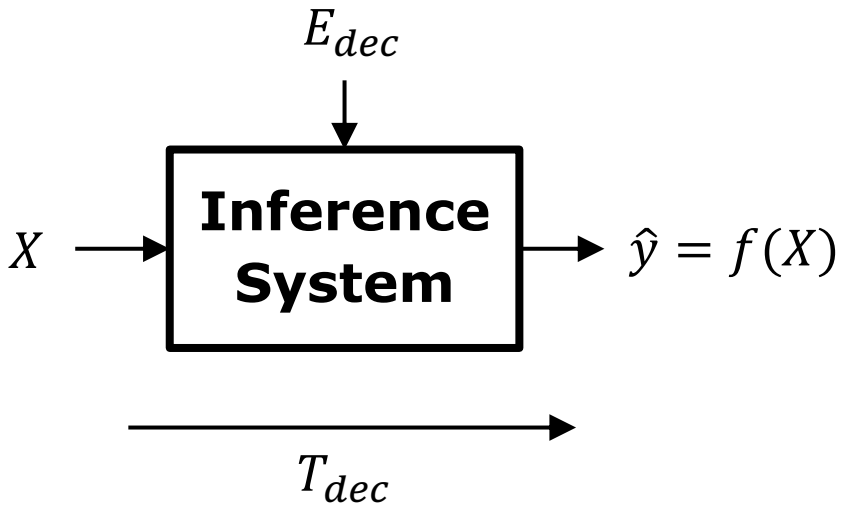
[VLSI'18(Princeton)]

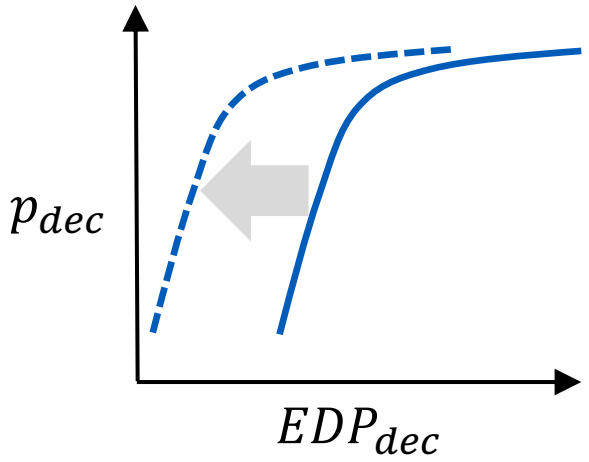


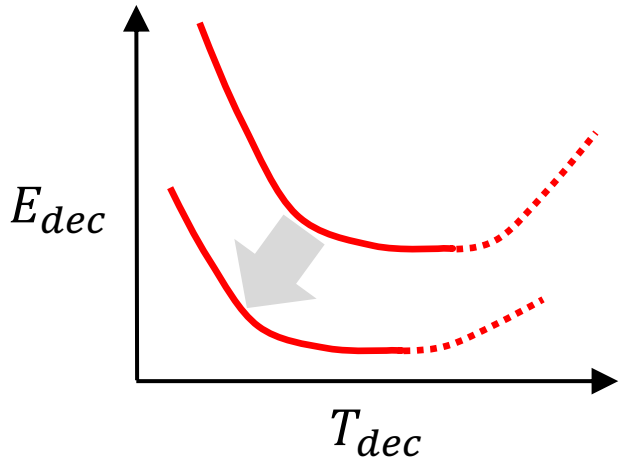
- 6T+ bitcell/bitcell array – 8T/10T/12T
- restricted to binary weights and/or activations

System Metrics for Inference

system-level energy vs. latency vs. accuracy trade-off





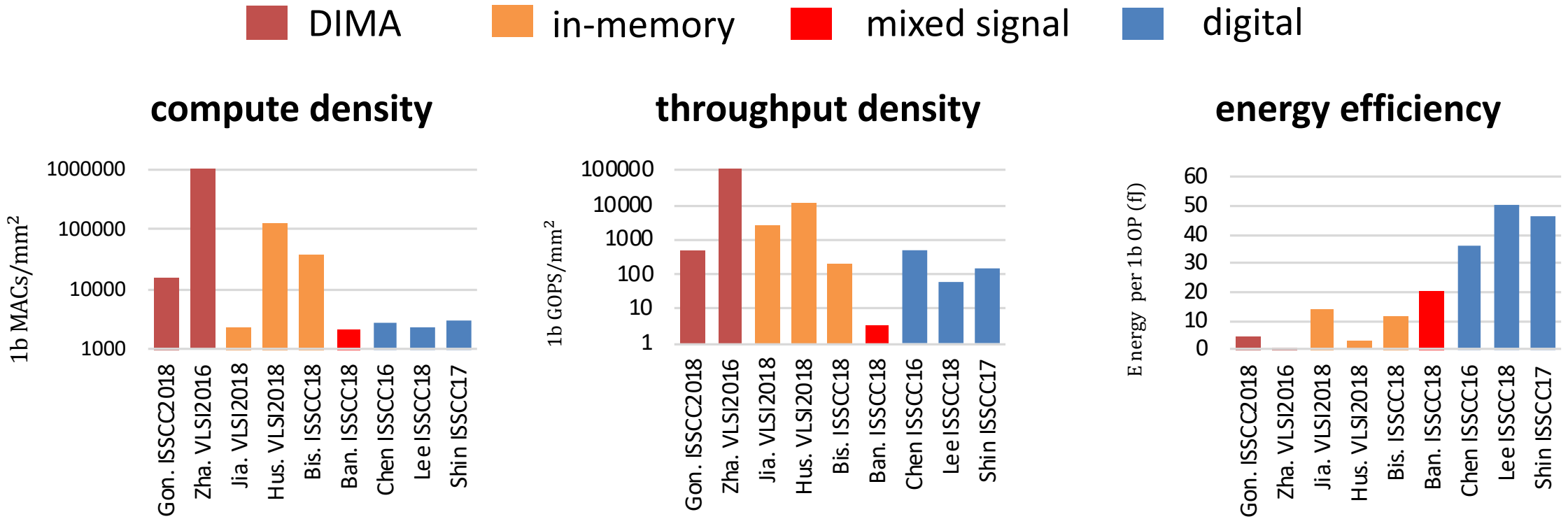


- E_{dec}: energy per decision
- f_{dec}: decision throughput
- T_{dec}: decision latency
- p_{dec}: decision accuracy
- $EDP_{dec} = E_{dec} \times T_{dec}$: decision EDP

- $E_{dec} \propto 1/T_{dec} \propto$ data & network size
- $p_{dec} \propto EDP_{dec}$

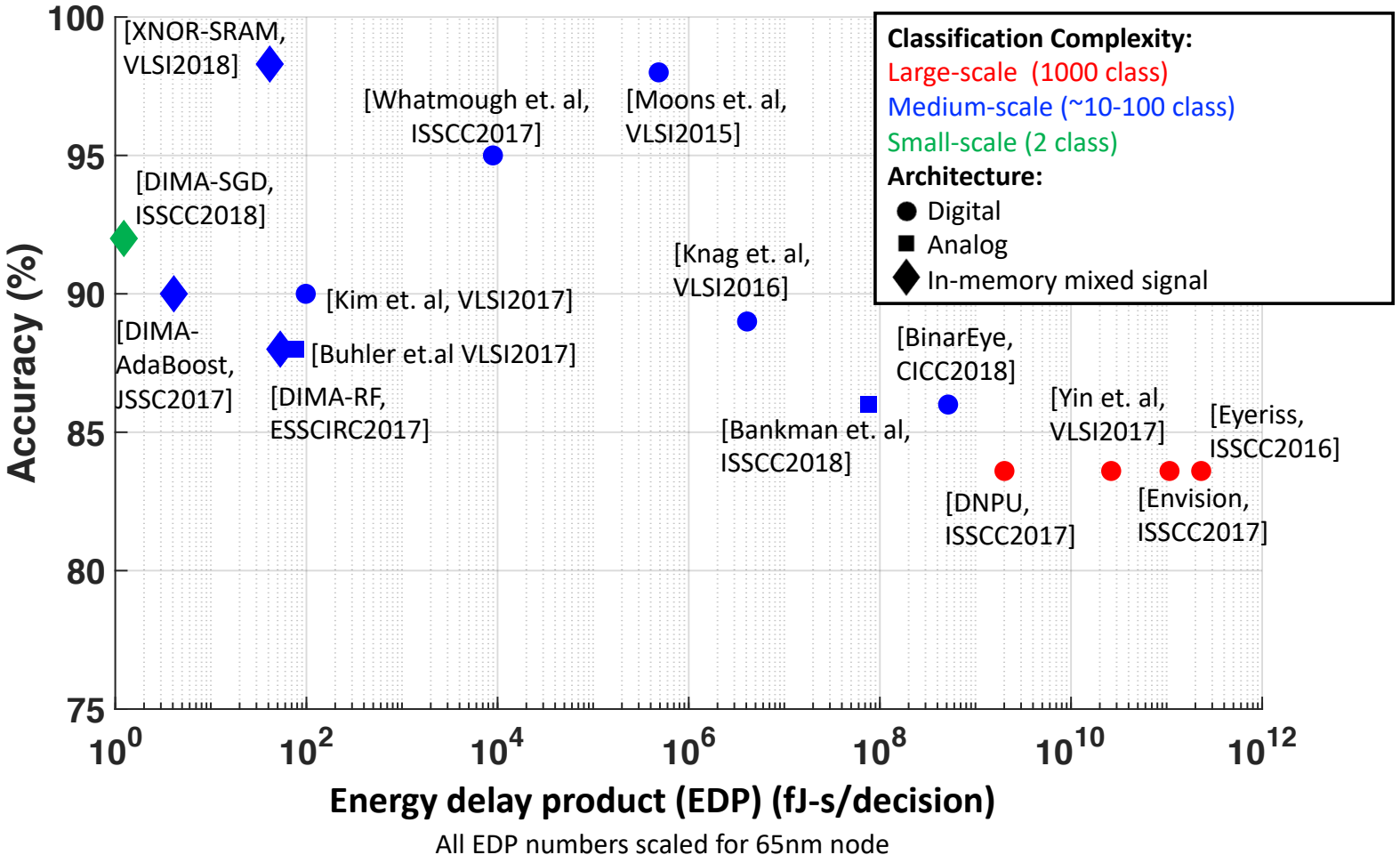
Arithmetic Efficiency = System Efficiency/System Complexity

orders-of-magnitude improvement



(scaled to 65nm CMOS)

System Efficiency

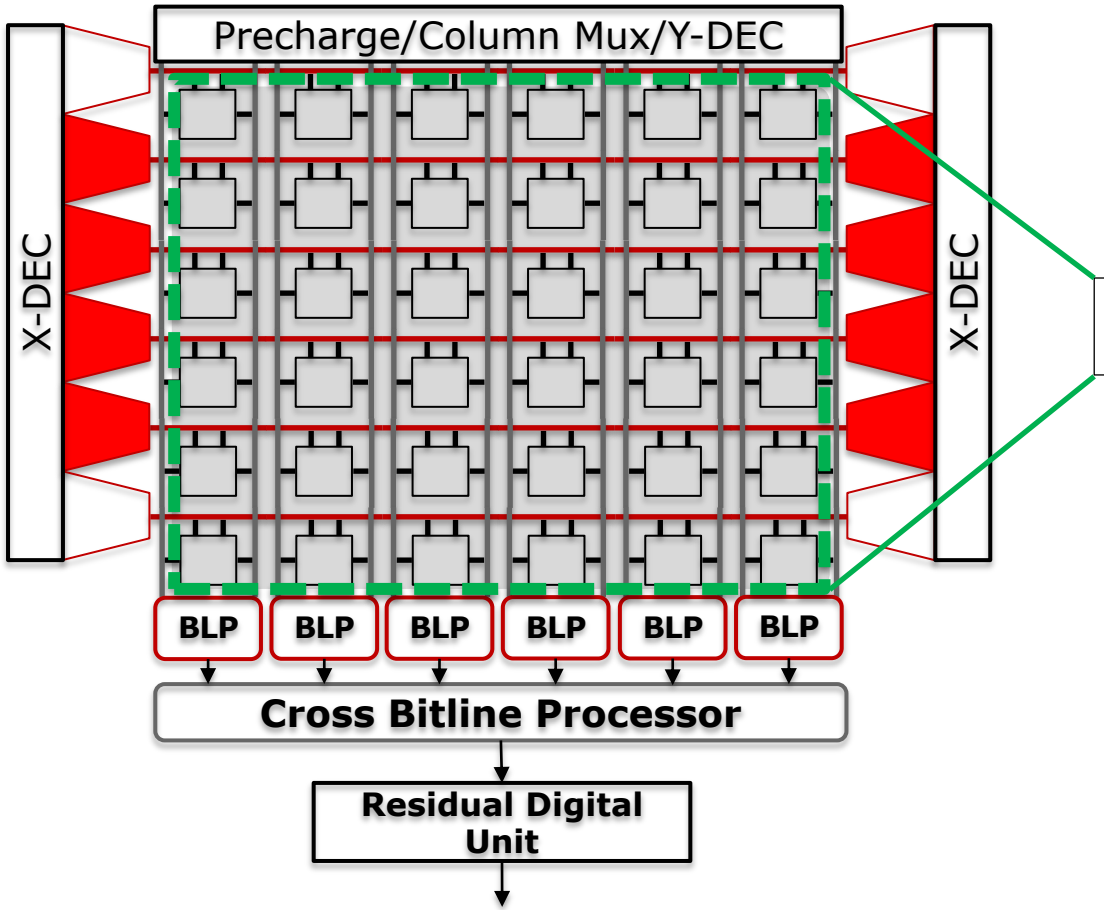


All EDP numbers scaled for 65nm node

- clear system EDP benefits of in-memory over digital for medium-scale tasks with no loss in accuracy
- more work needed to evaluate benefits for large-scale tasks

The Deep In-memory Architecture

Deep In-memory Architecture (DIMA)

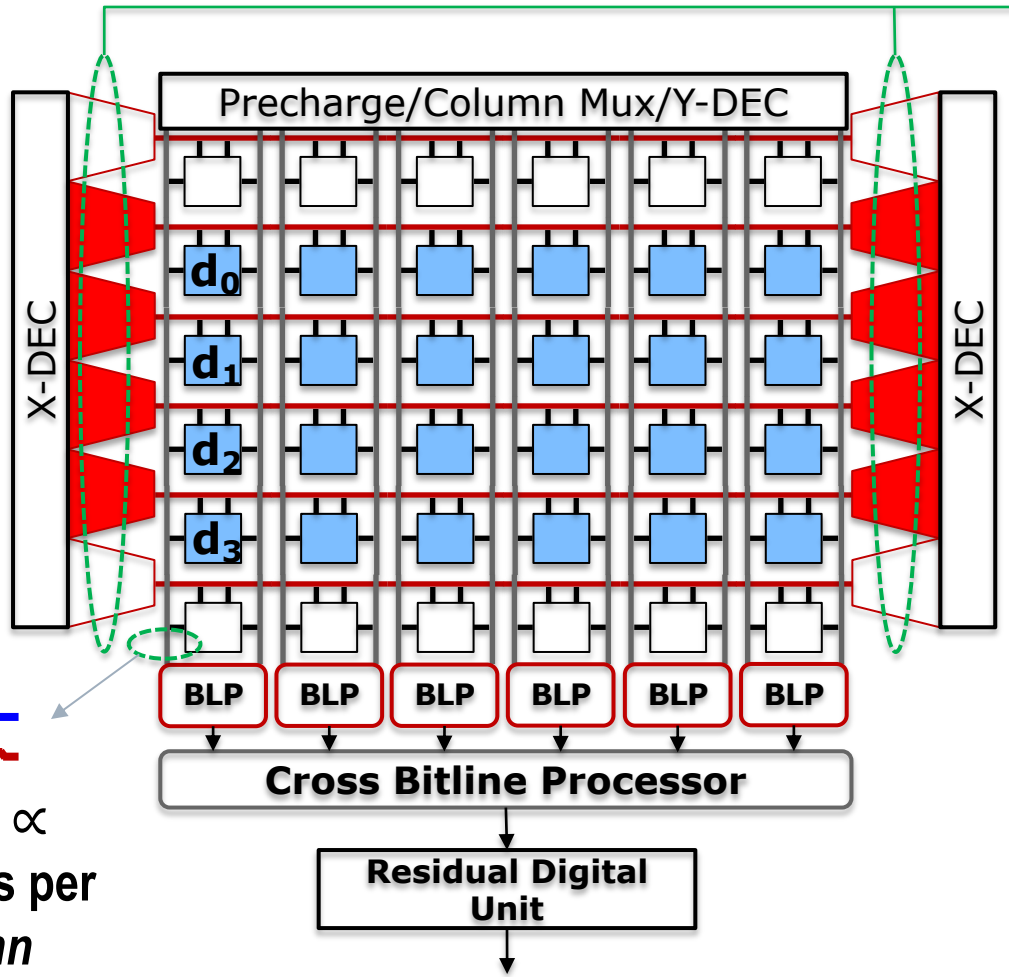


standard 6T SRAM BCA

- preserves storage density (bits/mm^2)
- enhances compute density (TOPS/mm^2)
- reduces both E & D in EDP
- distributed analog compute amortizes the cost of read accesses

[UIUC (ICASSP'14,3xJSSC'18); Princeton (VLSI'16,JSSC'17)]

Deep In-memory Architecture (DIMA)



multi-row functional READ
(reads multiple-bits/row/precharge)

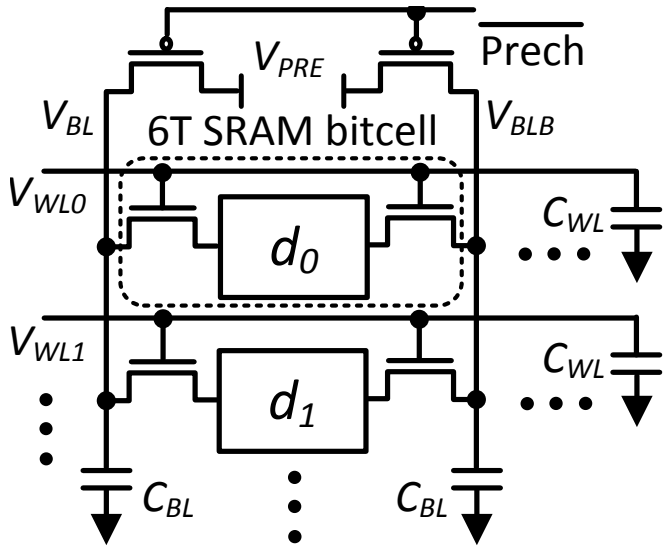
①

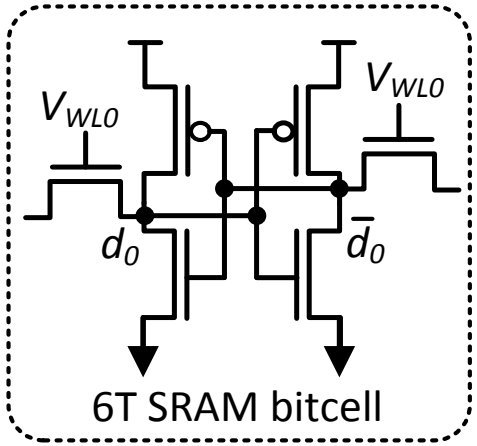
analog, mixed-signal
low-SNR processing

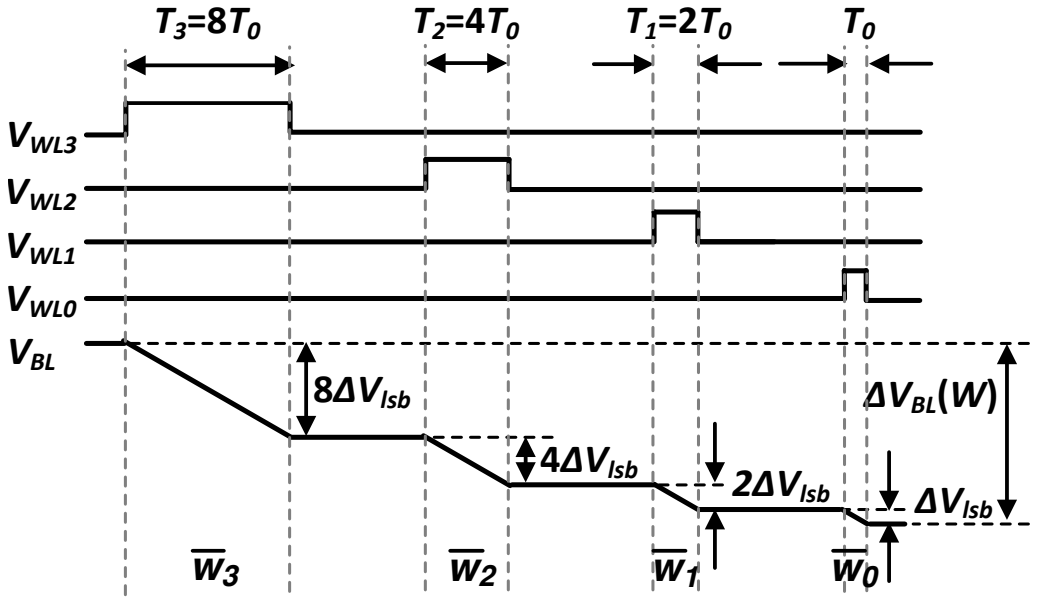
improved throughput
reduced memory access
energy minimized

[UIUC (ICASSP'14,3xJSSC'18); Princeton (VLSI'16,JSSC'17)]

Functional READ - Overview







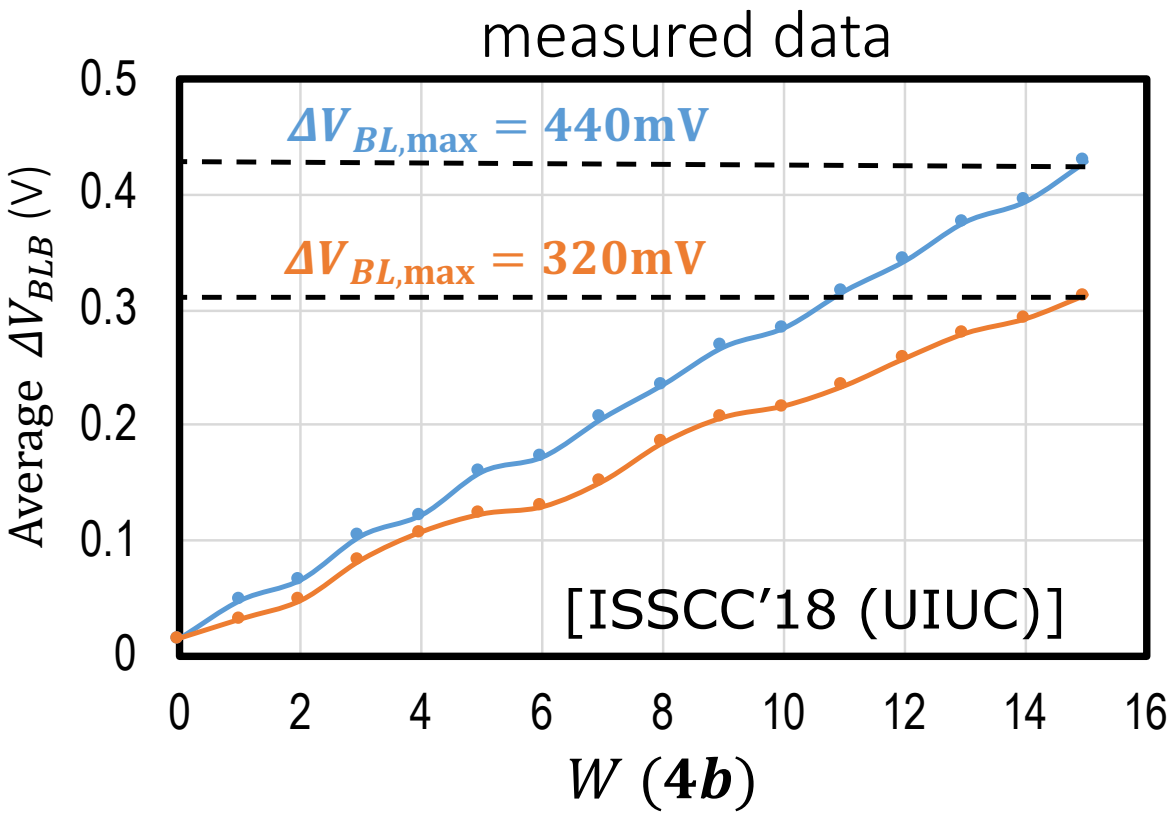
- matrix-vector product/read cycle
- column-major format
- multi-row access/read cycle
- modulated access pulses

Per-column dot-product

$$\Delta V_{BL} = \frac{I_{cell}}{C_{BL}} \sum_{i=0}^{N-1} T_i w_i$$

4b Functional READ

$$T_i = 2^i$$



charge-summing

if $T_i \ll R_{BL}C_{BL}$

$$\Delta Q_i = I_{cell}T_i w_i \rightarrow \Delta Q = I_{cell}\sum_i T_i w_i$$

Per-column dot-product

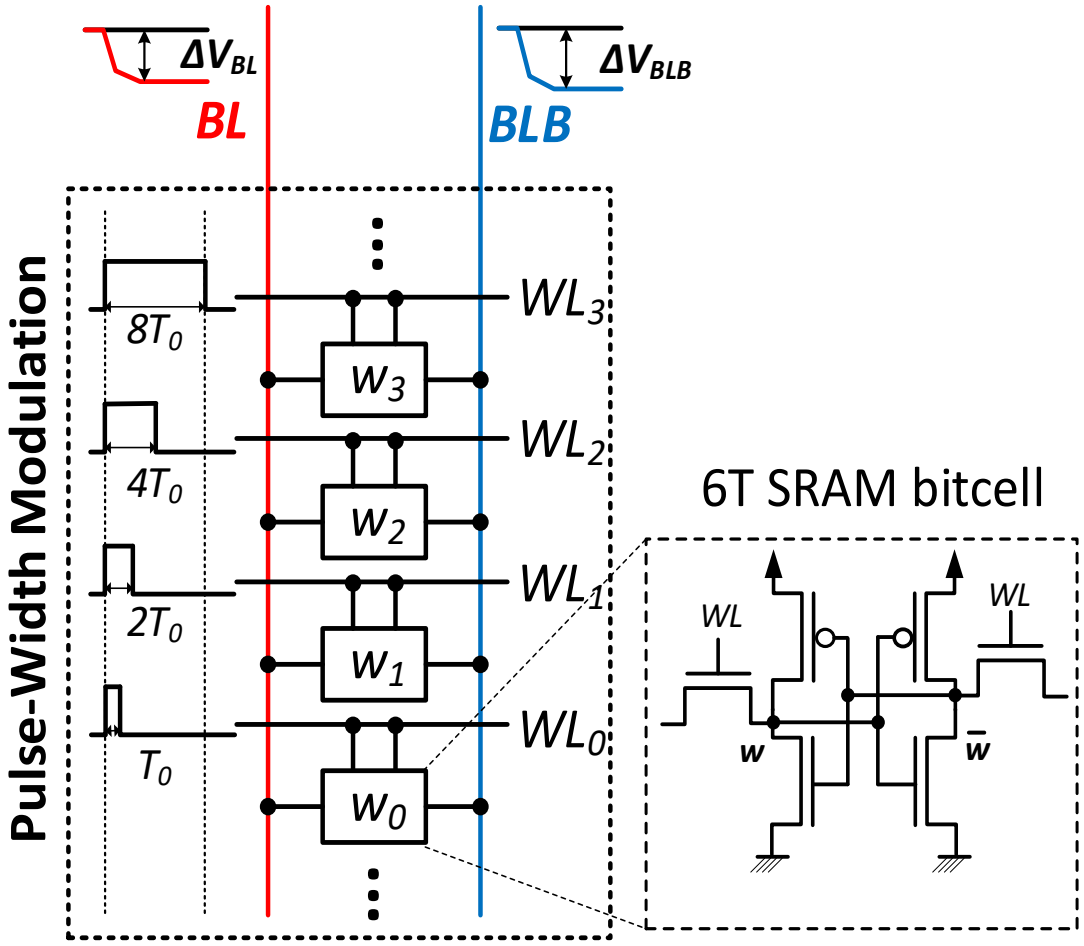
$$\Delta V_{BL} = \frac{I_{cell}}{C_{BL}}\sum_{i=0}^{N-1} T_i w_i$$

Assumptions for Per-Column Dot Product

$$\Delta V_{BL} = \frac{I_{cell}}{C_{BL}} \sum_{i=0}^{N-1} T_i w_i = \frac{V_{PRE}}{R_{BL} C_{BL}} \sum_{i=0}^{N-1} T_i w_i$$

1) $T_i \ll R_i C_{BL}$ (good linear approximation to an exponential)
2) $T_i \propto X_i$ or $T_i = 2^i T_0$ (pulse widths proportional to inputs)
3) $R_i = R_{BL}$ (row-invariance of discharge path resistance)
4) R_{BL} independent of V_{BL} (no channel length modulation in access TX)

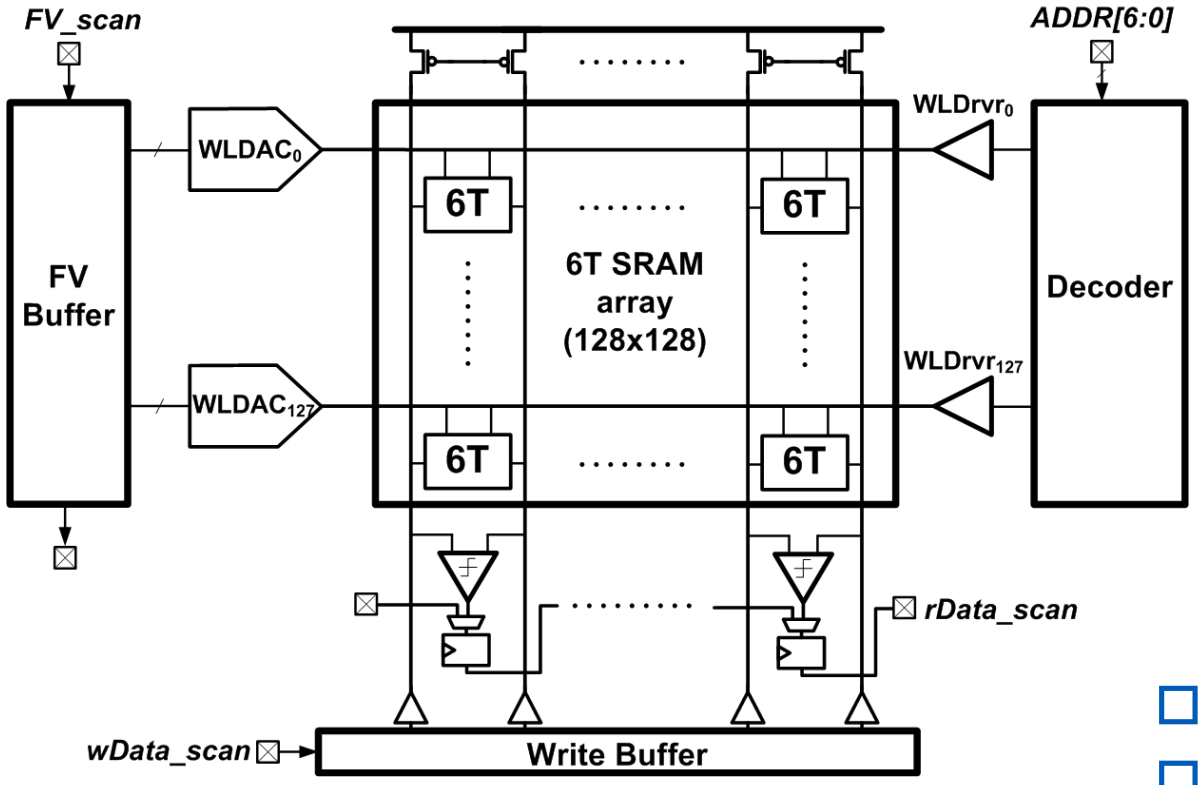
Design Challenges



[Feb. JSSC'18 (UIUC)]

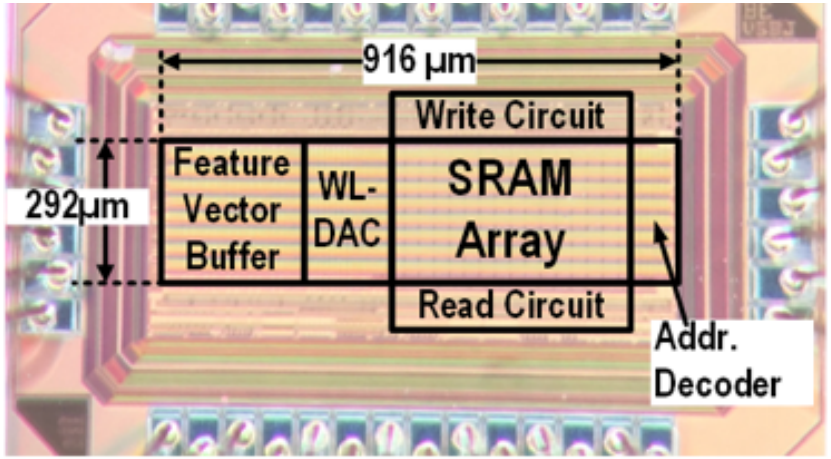
- $T_r + T_f < 0.2T_0 \rightarrow$ to drive large WL caps with PWM pulses (2)
- $2^{B-1}T_0 \ll R_{BL}C_{BL} \rightarrow$ to reduce time non-linearity (1)
- $V_{WL} < 0.8V_{PRE} \rightarrow$ to avoid read upsets & improve current linearity (4)
- $\Delta V_{BL,max} \approx 500mV \rightarrow$ to reduce spatial V_t variations (3)

Functional READ via PAM Access Pulses



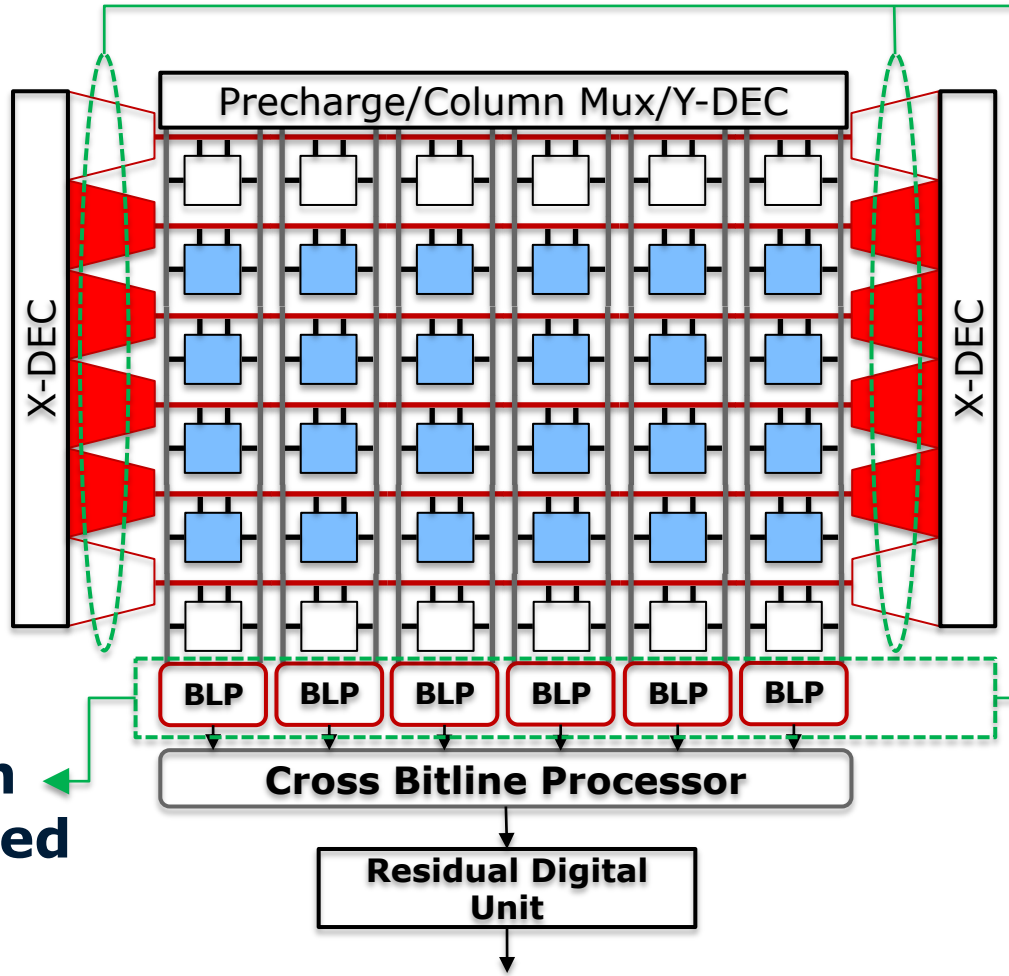
$$\Delta V_{BL} = \frac{V_{pre} T}{R_{BL} C_{BL}} \sum_{i=0}^{N-1} f(V_i)\, d_i$$

[JSSC'17, Princeton]



- 128b x 128b – BCA; 5b input; 1b weights
- $R_P = 81; C_P = 128$ (massively parallel)
- AdaBoost compensates for low SNR
- EDP gain = 172× (13× energy reduction)

Deep In-memory Architecture (DIMA)



multi-row functional READ
(reads multiple-bits/row/precharge)

①

analog, mixed-signal low-SNR processing

bitline processing (BLP)
(SIMD analog processing)

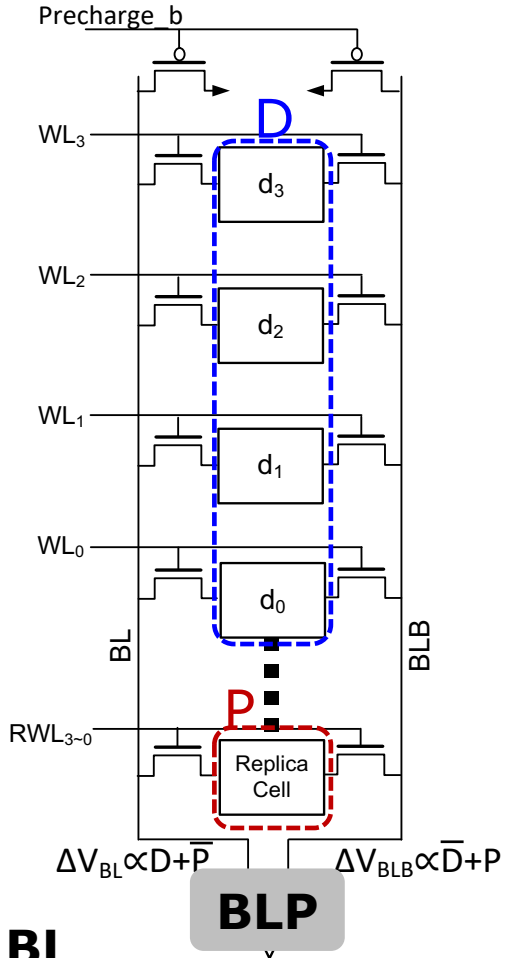
②

massively parallel processing

[UIUC (ICASSP'14,3xJSSC'18); Princeton (VLSI'16,JSSC'17)]

Bit-line Analog Processors

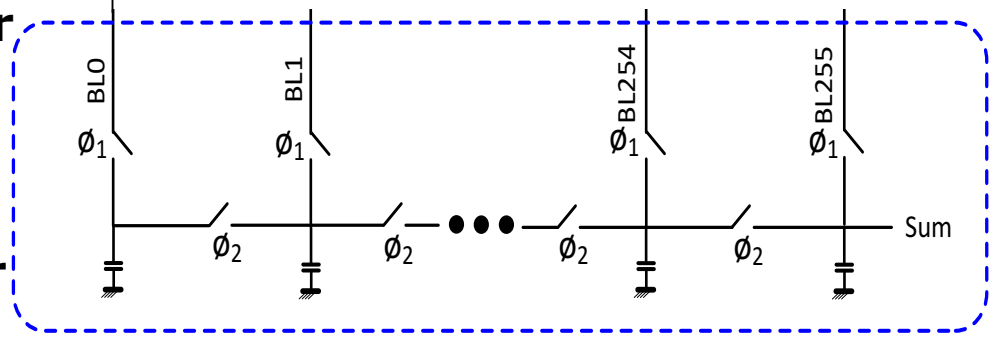
charge-redistribution based processing

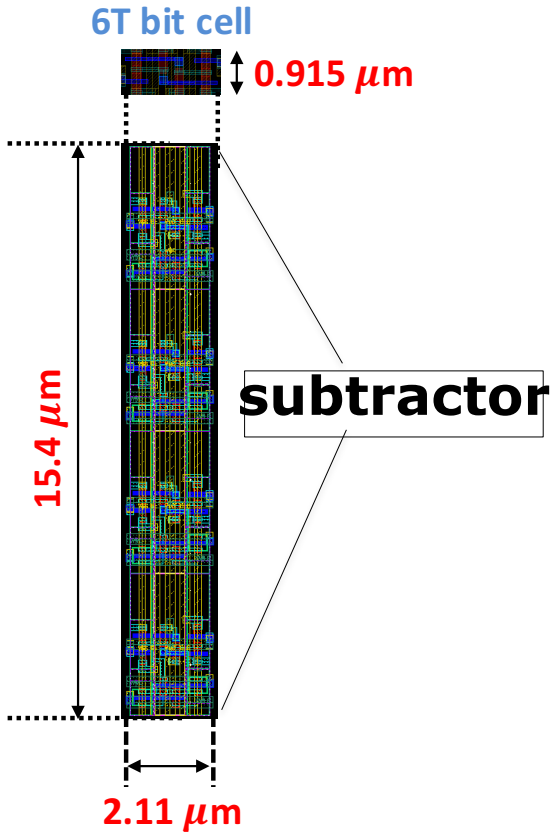


Kernel	BLP	CBLP
Manhattan distance	subtract-compare	aggregation
Euclidean distance	subtract-square	aggregation
Dot product	multiply	weighted aggregation
Hamming distance	XOR	aggregation

BL processor (BLP)

cross BL processor (CBLP)

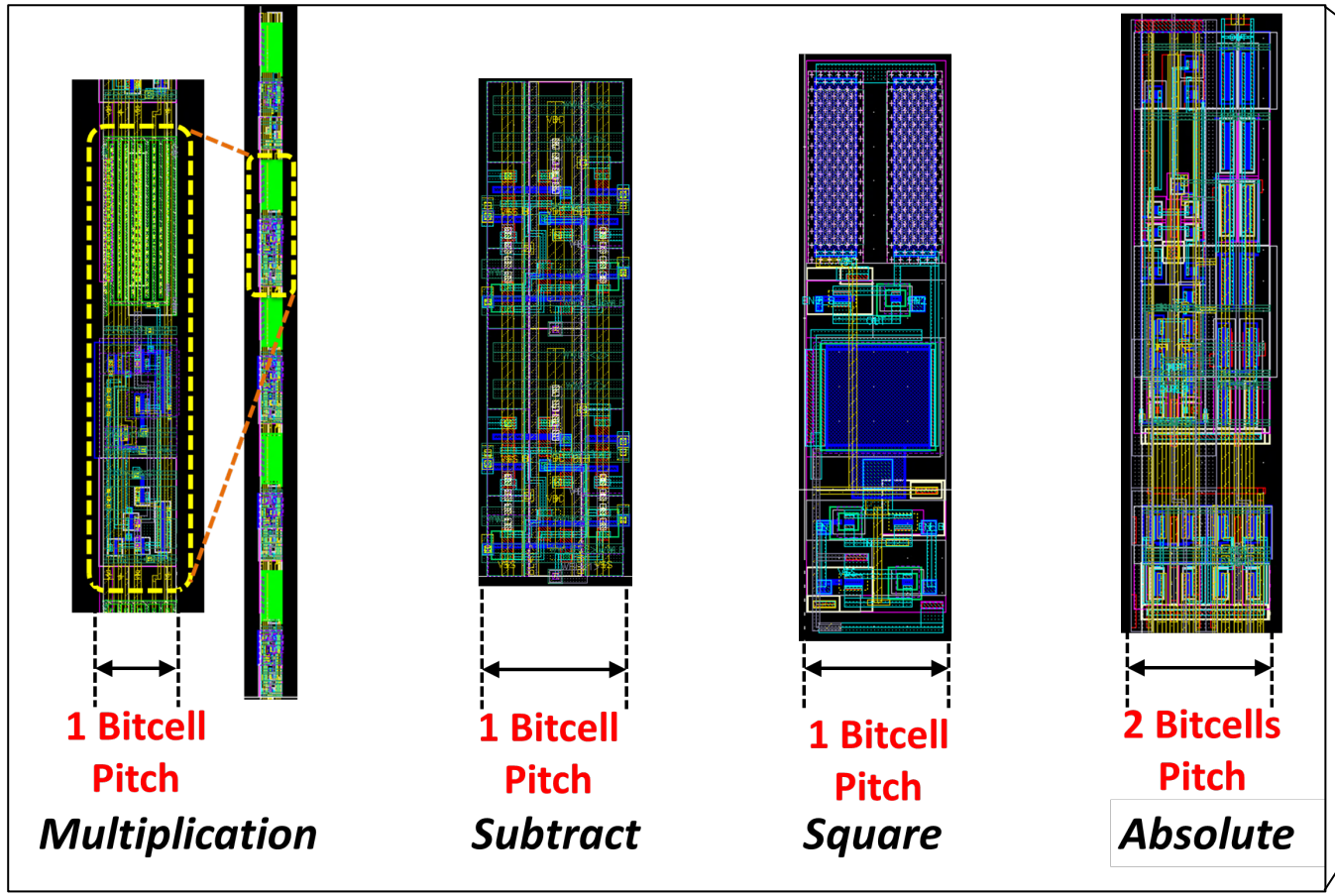


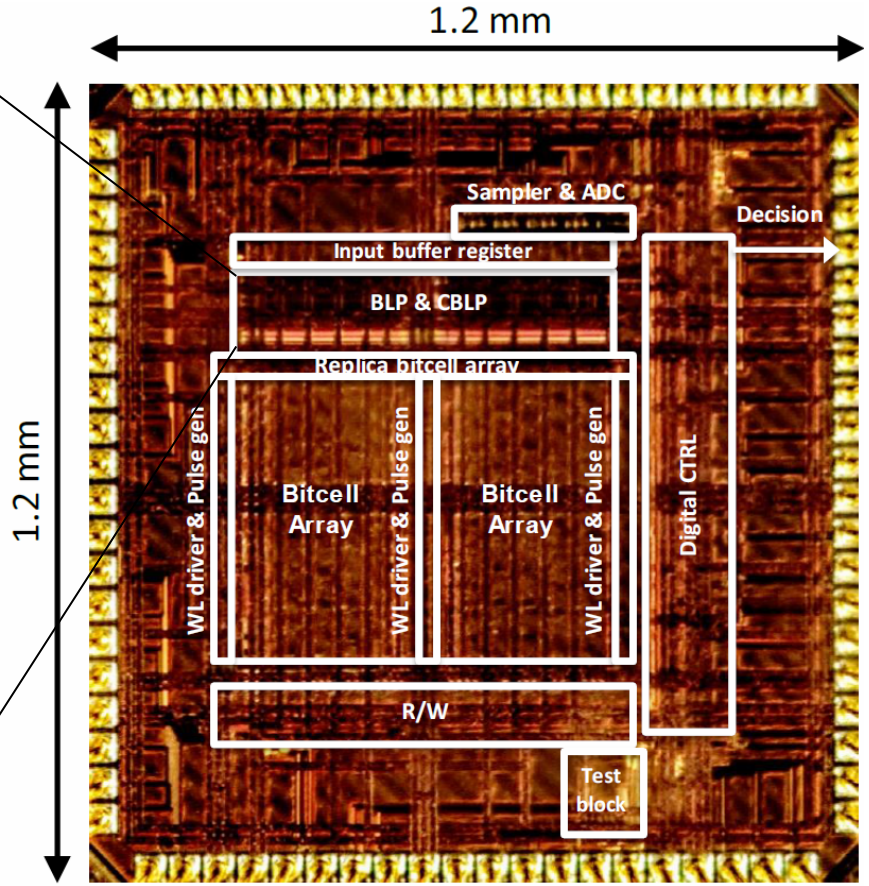


Pitch-matching in BL Processors

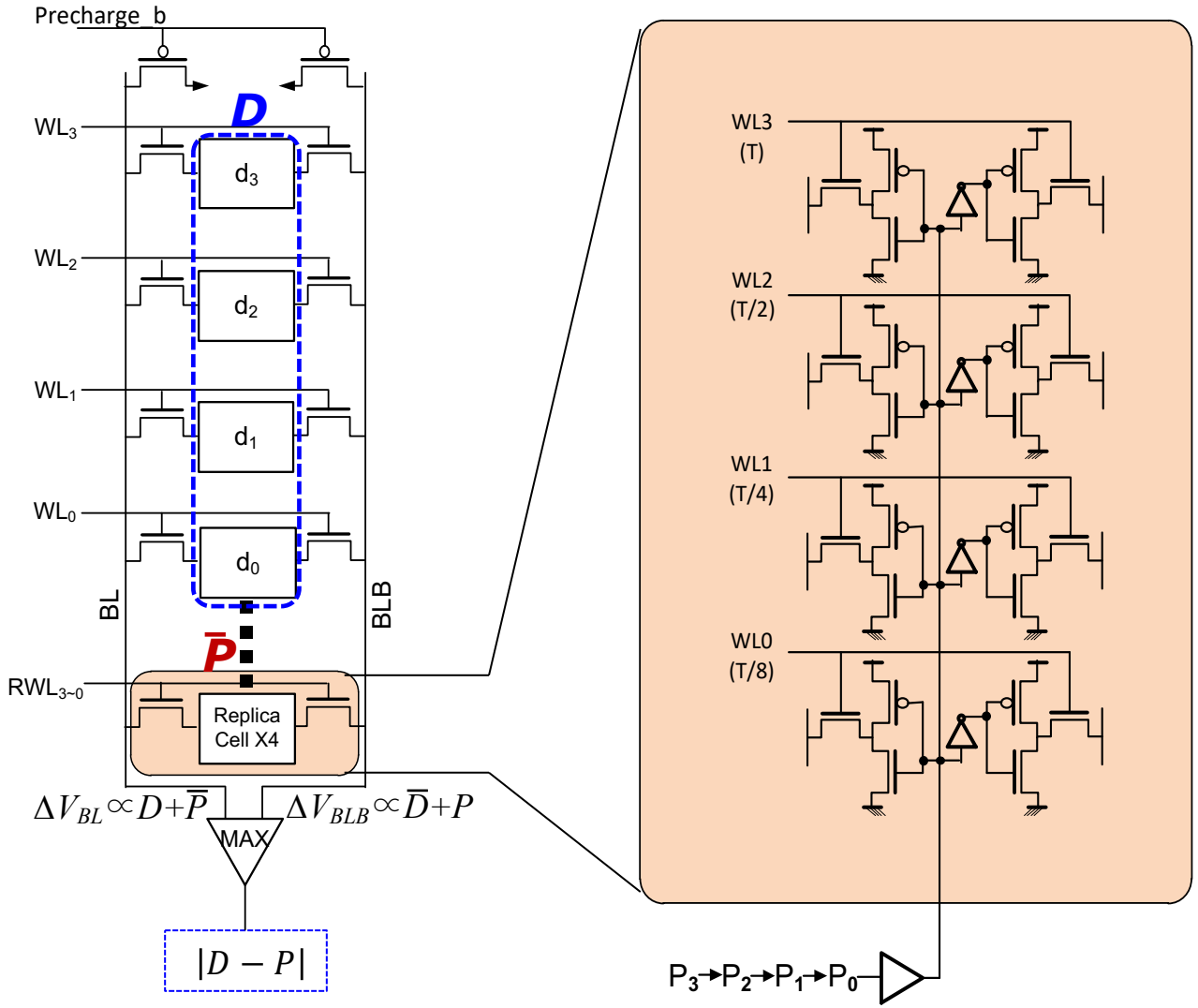
BLP area overhead reduces with BCA size

area overhead < 10%
(512x256 BCA)





Bit-Line Processor – Absolute Difference



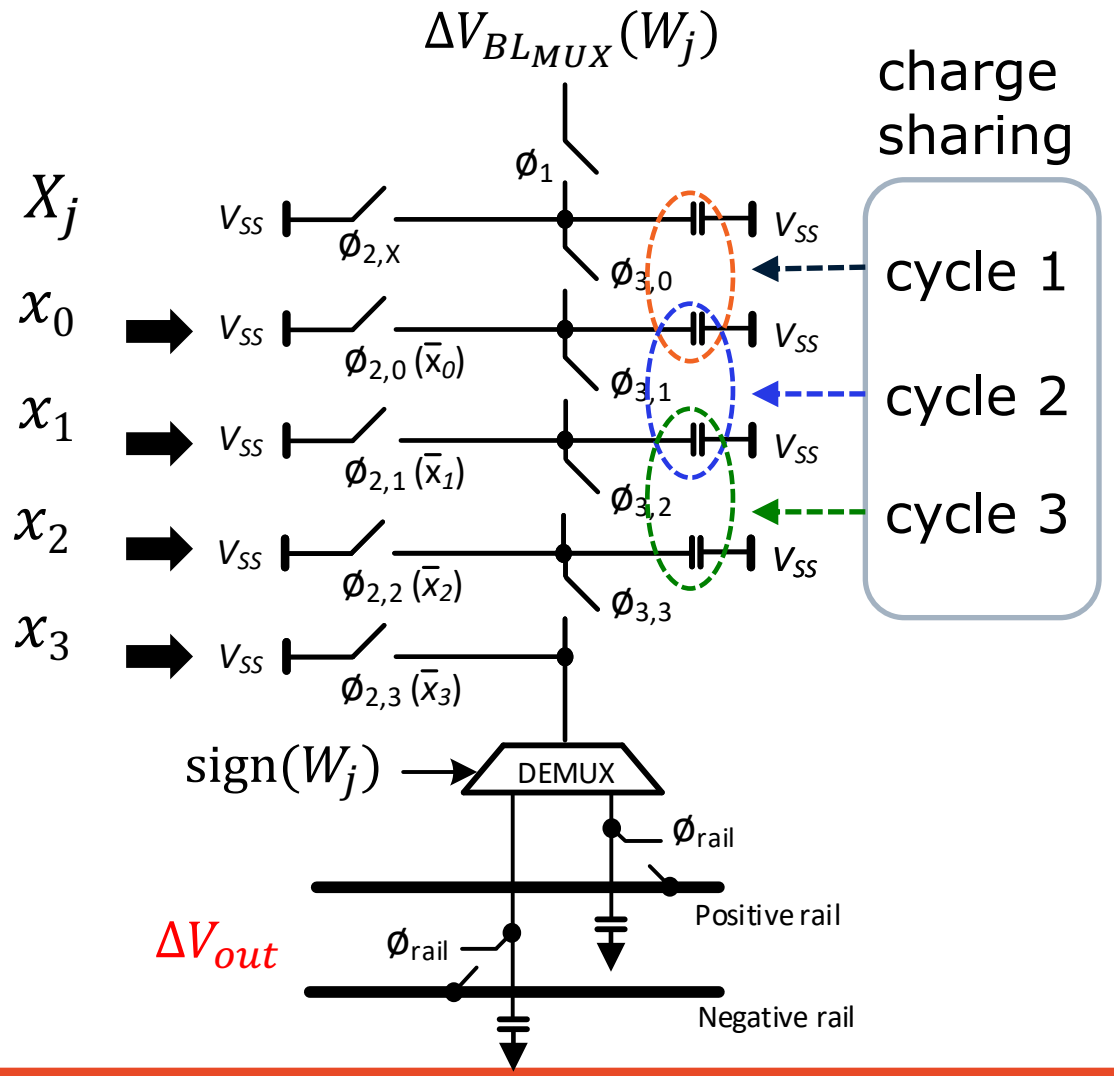
- Efficient computation of word-level difference due to differential nature of SRAM bitcell + functional READ

$$|D - P| = \max(D - P, P - D)$$
$$\max(D + \bar{P}, \bar{D} + P)$$

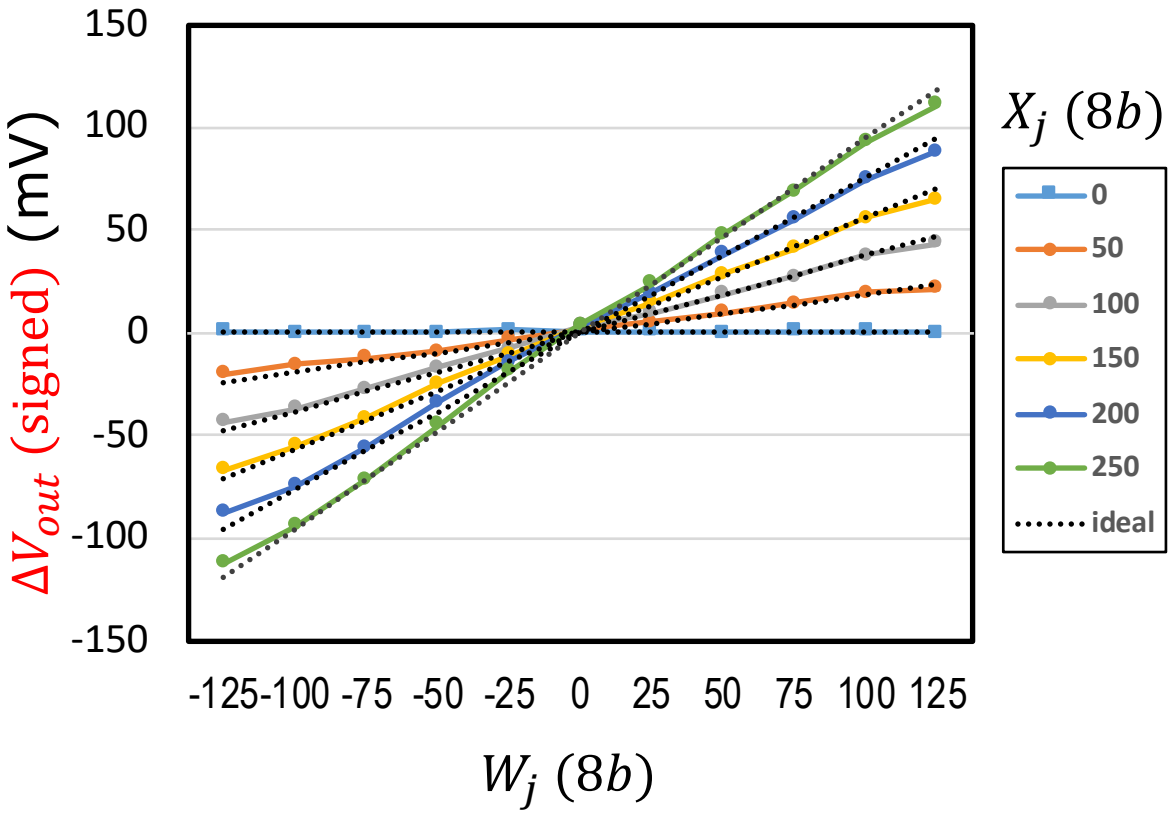
$$\Delta V_{BL} = \frac{V_{pre}}{RC} \sum_{i=1}^{N} T_i \, (d_i + \bar{p}_i)$$

$$\Delta V_{BLB} = \frac{V_{pre}}{RC} \sum_{i=1}^{N} T_i \, (\bar{d}_i + p_i)$$

Bit-Line Processor – Signed 4x4b Multiplier

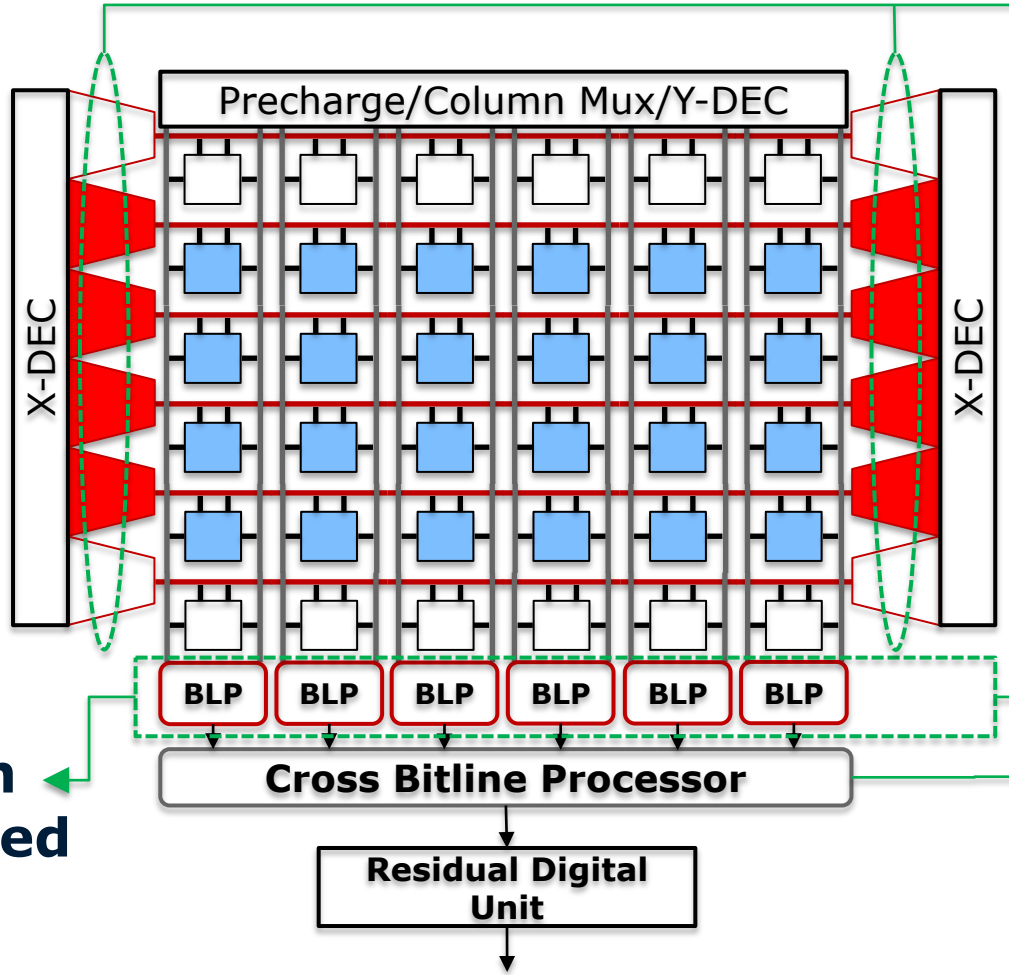


(measured across 128 locations)



<4% error magnitude

Deep In-memory Architecture (DIMA)



multi-row functional READ
(reads multiple-bits/row/precharge) ①

analog, mixed-signal low-SNR processing

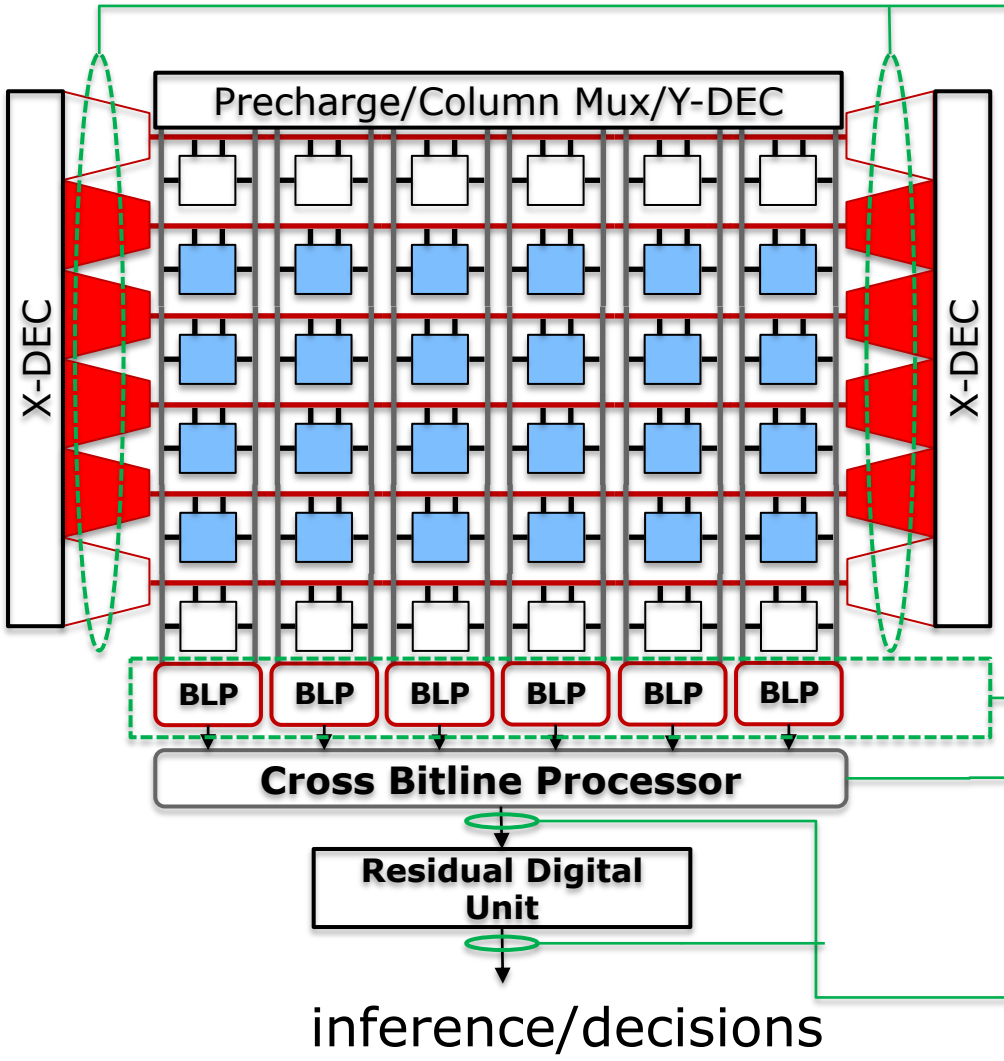
bitline processing (BLP)
(SIMD analog processing) ②

cross bitline processing (CBLP)
(analog averaging enhances SNR) ③

efficient implementation of vector operations

[ICASSP 2014 UIUC, JSSC 2017 Princeton, JSSC 2018 UIUC]

Deep In-memory Architecture (DIMA)



multi-row functional READ
(reads multiple-bits/row/precharge)

①

analog, mixed-signal low-SNR processing

bitline processing (BLP)
(SIMD analog processing)

②

cross bitline processing (CBLP)
(analog averaging enhances SNR)

③

reduced complexity, bandwidth efficient digital output

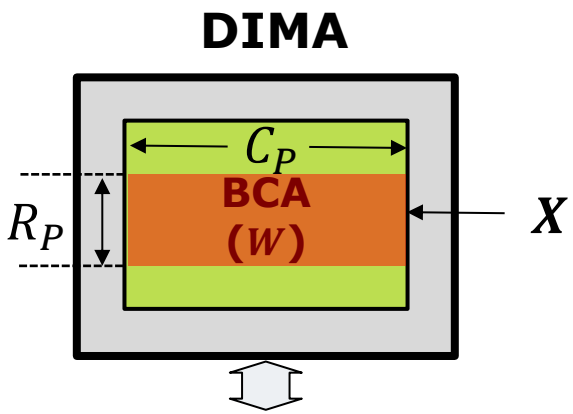
④

[UIUC (ICASSP'14,3xJSSC'18); Princeton (VLSI'16,JSSC'17)]

Fundamental Energy-Latency-Accuracy Trade-offs

DIMA's Efficiency vs. SNR Trade-off

DIMA computations exhibit a fundamental SNR vs. EDP trade-off



$$E_{DIMA} = C_{BL}\Delta V_{BL,max} V_{PRE} + \frac{E_{lk}}{S}$$

$$SNR_{DIMA} = \frac{\Delta V_{BL,max}}{M \times \sigma_\eta}$$

$S = LR_P$: speed-up over digital architecture
M: # of BL levels

- row-parallelism (R_P) trades-off with SNR, dynamic range, read stability
- column (C_P) vs. read sensing noise trade-off

Example – SRAM in 65 nm CMOS

Parameter	Value	Parameter	Value
k'	$220\,\mu\text{A/V}^2$	α	1.8
σ_{To}	$0.43\,\text{ps}^{0.5}$	σ_{V_t}	23.8 mV
WLC_{ox}	0.21 fF	κ	$0.008\,\text{fF}^{0.5}$
V_t	0.4 V	T_o	100 ps
T	270 K	p	0.5
V_{dd}	1 V	V_{WL}	$0.4 - 0.8$ V
$V_{BL,max}$	0.8 V	C_{BL}	270 fF

$$E_{DIMA} = C_{BL}\Delta V_{BL,max}V_{PRE} + \frac{E_{lk}}{S} = 0.237\text{pJ}$$

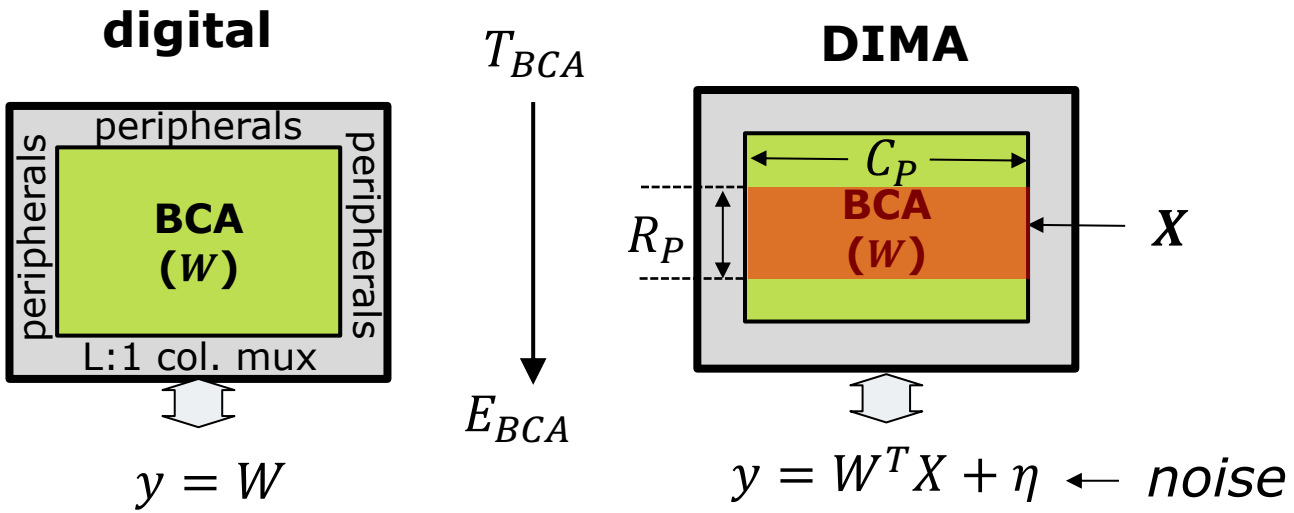
$$SNR_{DIMA} = \frac{\Delta V_{BL,max}}{M \times \sigma_\eta} = 24.4\,\text{dB}$$

- $N_{col} = 256; N_{row} = 512;$

- $\sigma_\eta = \frac{3mV}{LSB}$ @ $V_{WL} = 0.65V$
- $E_{lk} = 337\frac{fJ}{BL}; S = 16$

Comparing DIMA vs. Digital

DIMA computed functions are noisy but its inferences remain accurate



- DIMA reads are functions of data
- fewer read cycles to realize a function of B bits → reduces both E & D in EDP

- fundamental EDP vs. SNR trade-off (no free lunch)
- system accuracy derived from its Shannon-inspired roots (affordable lunch)

Energy vs. Latency

- assume identical BCA and free compute
- E_{BCA}: array energy per read cycle
- T_{BCA}: read delay
- DIMA's row-(R_P) & col. (C_P) parallelism

metric	digital	DIMA	digital/DIMA
# of bits/cycle	$\frac{N_{col}}{L}$	$N_{col}R_P$	$\left(\frac{1}{LR_P}\right)$
cycles/B bits	$\frac{LB}{N_{col}}$	$\frac{B}{N_{col}R_P}$	LR_P
energy per B bits	$\left(\frac{LB}{N_{col}}\right)E_{BCA}$	$\left(\frac{B}{N_{col}R_P}\right)E_{BCA}$	LR_P
delay per B bits	$\left(\frac{LB}{N_{col}}\right)T_{BCA}$	$\left(\frac{B}{N_{col}R_P}\right)T_{BCA}$	LR_P
EDP	$\left(\frac{LB}{N_{col}}\right)^2 EDP_{BCA}$	$\left(\frac{B}{N_{col}R_P}\right)^2 EDP_{BCA}$	$(LR_P)^2$

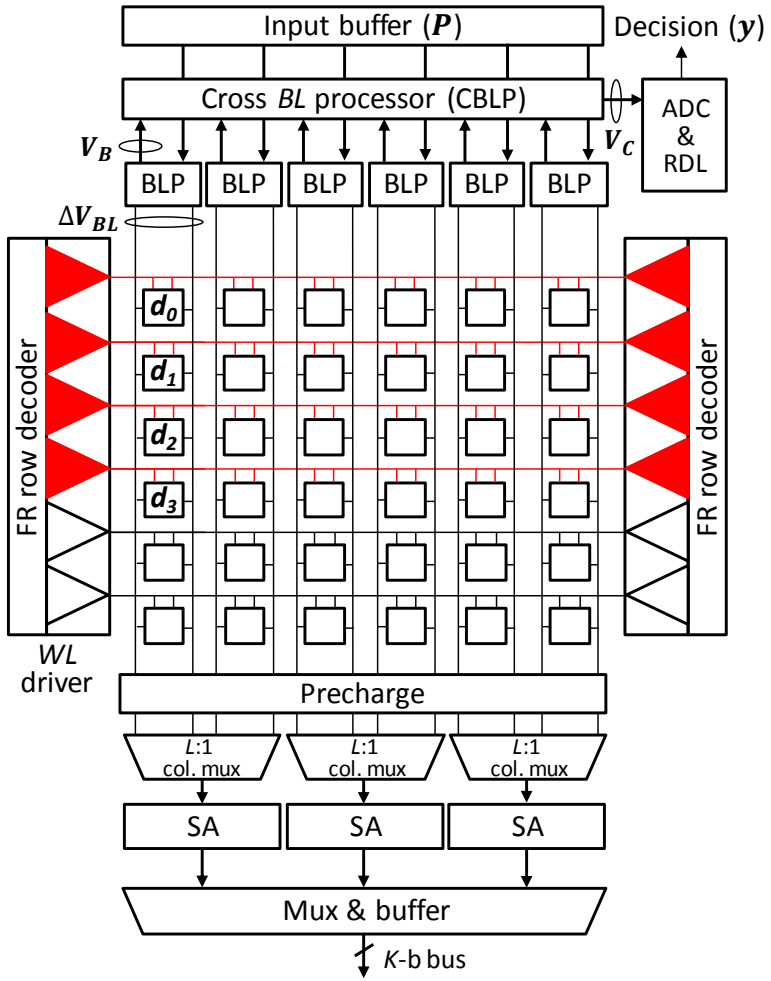
Typical values: $C_P = N_{col}$

- $L = 4,8,16; R_P = 4,5;$
- **idealized** EDP gains: 256×-to-6400×
- **realized** EDP gains [ISSCC'18]:

  100× with $L = 4; R_P = 4$

- much room to improve!

DIMA's Compute SNR



spatial noise in functional READ dominates

Error type	FR (η_f)	BLP (η_b) DP	BLP (η_b) SAD	CBLP (η_c)
% distortion (μ)	$2.6^{(1)}$	$2.1^{(1)}$	$2.5^{(1)}$	$0.8^{(3)}$
% noise (σ/μ)	$(5.3$ -to- $22)^{(2)}$	$2.8^{(2)}$	$3.2^{(2)}$	$0.2^{(3)}$

(1) silicon measured values
(2) Monte Carlo simulations with $0.4\,\mathrm{V} \leq V_{WL} \leq 0.8\,\mathrm{V}$.
(3) estimated from the capacitor sizes in [9].

digital

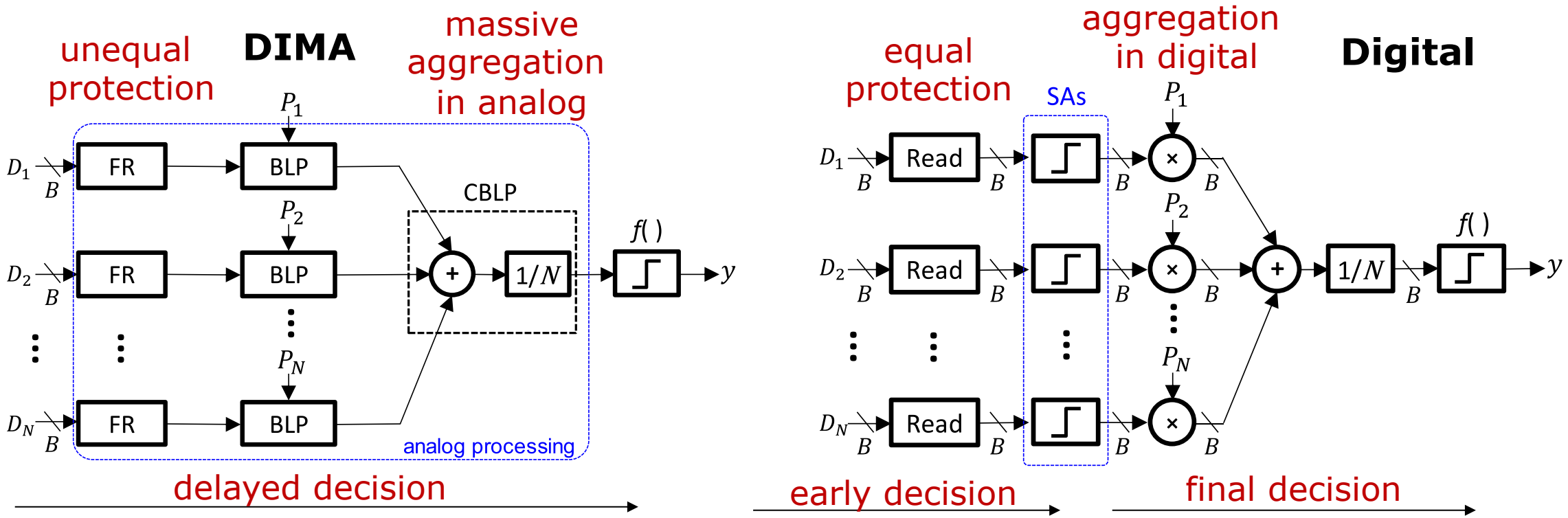
$$y = \mathbf{W}^T\mathbf{X} = \sum_{i=1}^{N} w_i x_i$$

DIMA

$$y = \sum_{i=1}^{N}[(w_i + \eta_{f,i})x_i + \eta_{b,i}] + \eta_c$$

FR BLP CBLP

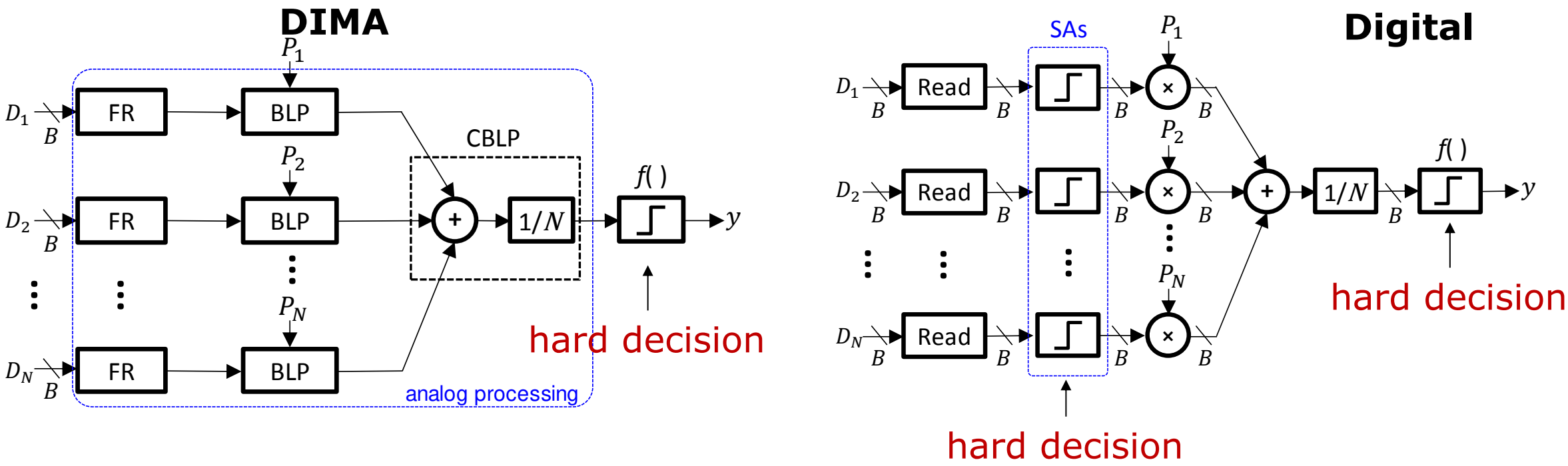
Sources of DIMA's Accuracy



data → noisy channel → decoder

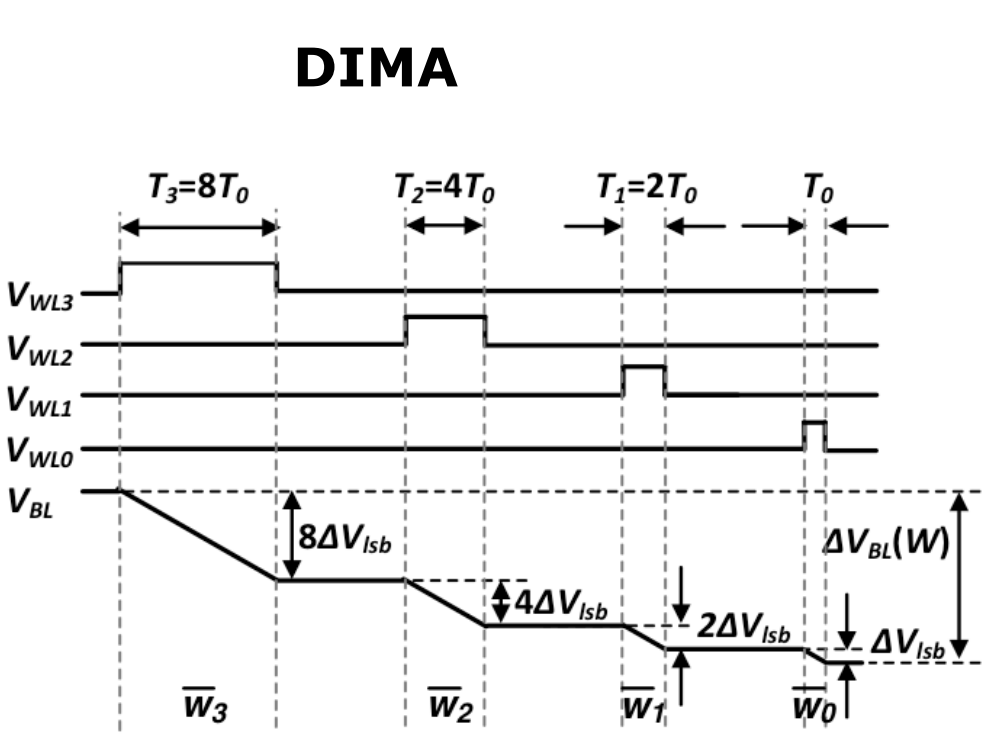
- **accuracy** via **delayed decision**, **unequal protection**, **massive aggregation**

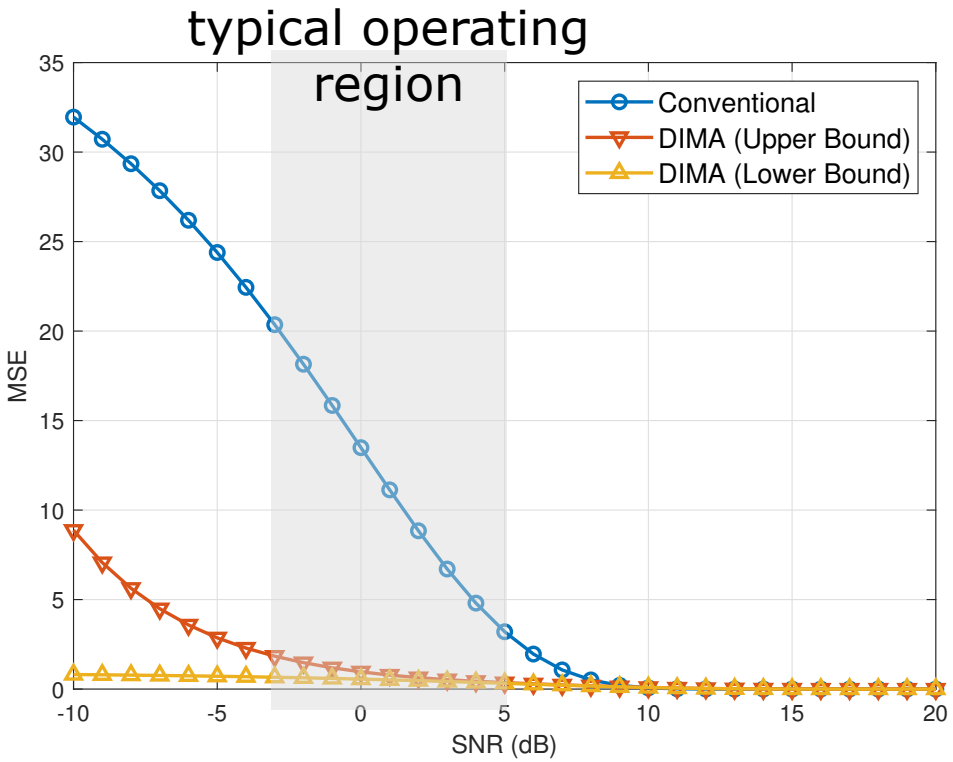
Delayed Decision



- a key rule in decision making – delay making hard decisions
- DIMA follows this rule; digital arch. violates this rule & suffers from MSB errors
- DIMA's accuracy better than digital *in low-SNR regime* – [JSSC'18 (UIUC)]
- accuracy gap increases with dimensionality N of data $\boldsymbol{X}$

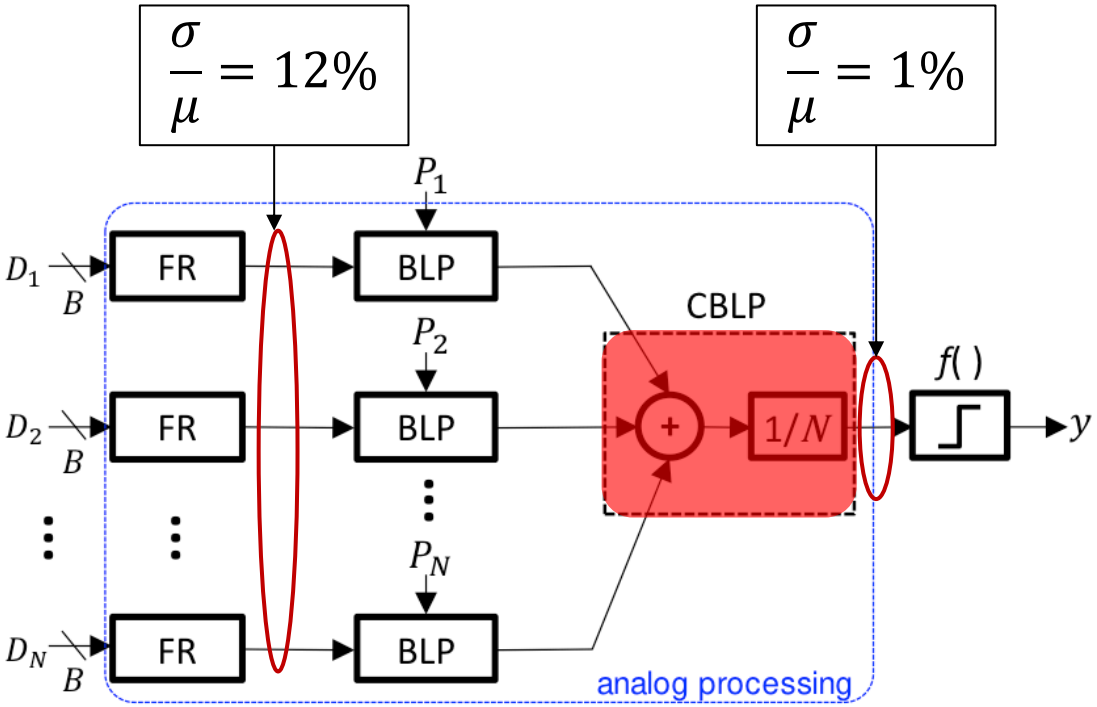
Non-uniform Bit Protection



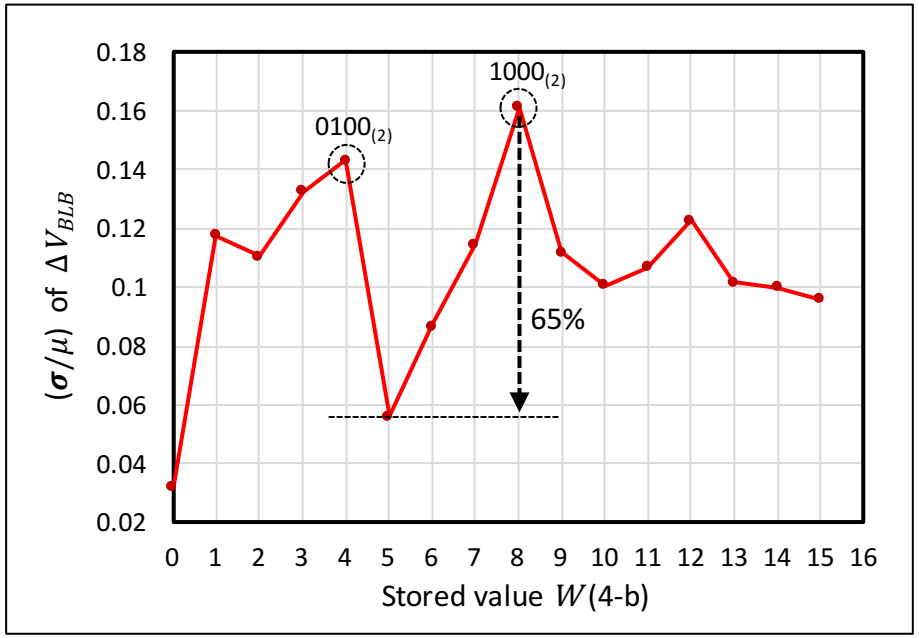


- DIMA allocates more swing to MSBs vs. LSBs – better use of limited swing
- digital architecture allocates uniform swing to all bits
- $\mathrm{MSE}(w) = \mathbb{E}[(\hat{w} - w)^2]$ – mean-squared error

Massive Aggregation



[ISSCC'18 (UIUC)]



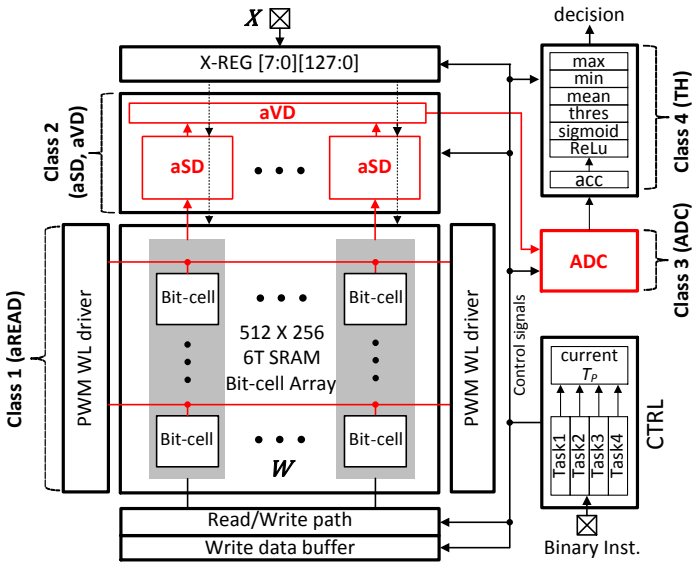
- DIMA's analog domain aggregation boosts SNR
  - SNR boost factor = $\sqrt{N_{col}}$
- Digital architectures miss out on this opportunity

Next Class - DIMA IC Case Studies

Future Prospects

Scaling-up DIMA

DIMA accelerator (ISCA'18, UIUC)



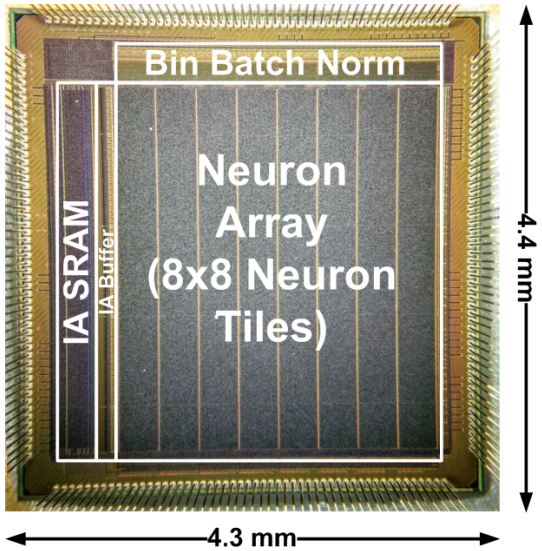
instruction set

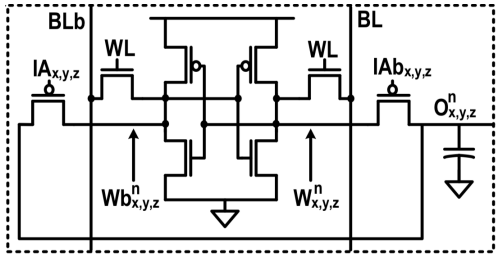
Class	Operation [operand]	Bit length	OP CODE	Option
1	*none*	3 bits (OPCODE)	000	*X_PRD: X_ADDR1* circulates from 0 to "*X_PRD*-1"
	write[W_ADDR]		001	
	read[W_ADDR]		010	
	aREAD[W_ADDR]		011	
	aSUBT[W_ADDR, X_ADDR1]		100	
	aADD[W_ADDR, X_ADDR1]		101	
2	*none*	4 bits (OPCODE + aVD)	000	aVD bit: 0: no aggregation 1: aggregation *X_PRD: X_ADDR2* circulates from 0 to "*X_PRD*-1"
	compare		001	
	absolute		010	
	square		011	
	sign_mult[X_ADDR2]		100	
	unsign_mult[X_ADDR2]		101	
3	*none*	1 bit (OPCODE)	0	
	ADC		1	
4	*accumulation*	3 bits (OPCODE)	000	*DES, ACC_NUM*
	mean		001	*DES*
	threshold		010	*DES, THRES_VAL*
	max		011	*DES*
	min		100	*DES*
	sigmoid		101	*DES*
	ReLu		111	*DES*

[2019 IEEE MICRO Top Picks Honorable Mention]

- analog-aware ISA
- Julia (compiler) software stack
- energy-accuracy optimization pass
- 22X EDP reduction for a 8b, 5-layer DNN

multi-bank DIMA (VLSI'18, Princeton)

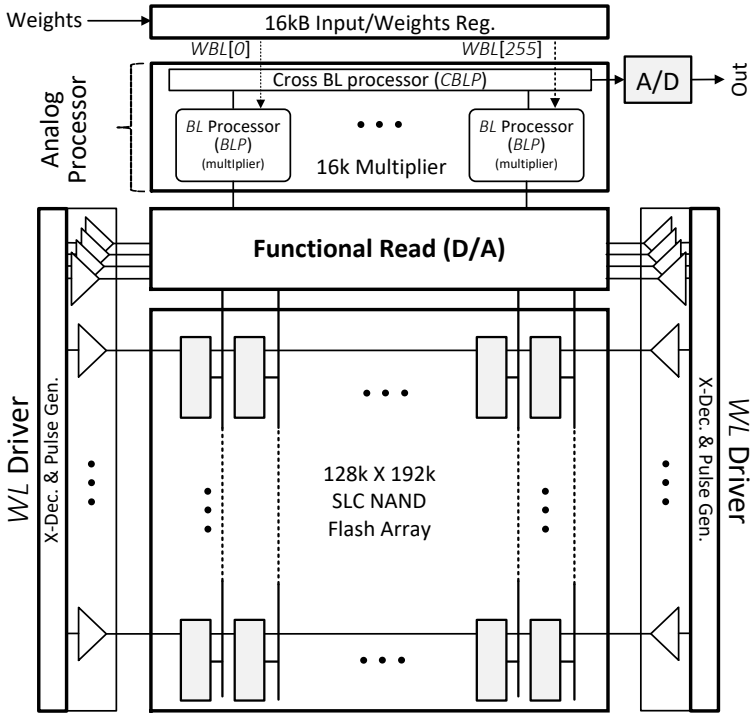




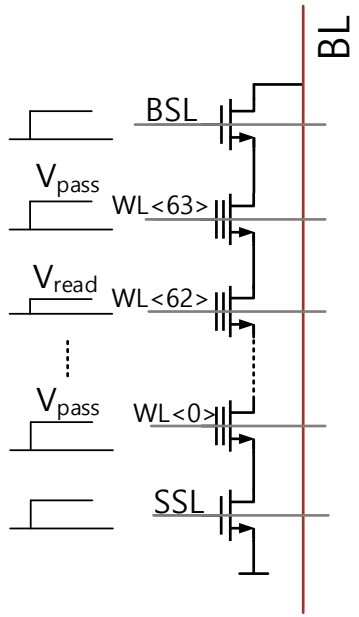
- 8T BC; 2.4Mb BCA; 64 bank; 1b compute; charge-based summing;
- accuracy - SVHN (94%); CIFAR-10 (83%) with 9-layer ConvNets
- 658 TOPS/W

DIMA in Flash and Emerging Memories

Flash DIMA (ISCAS'18)

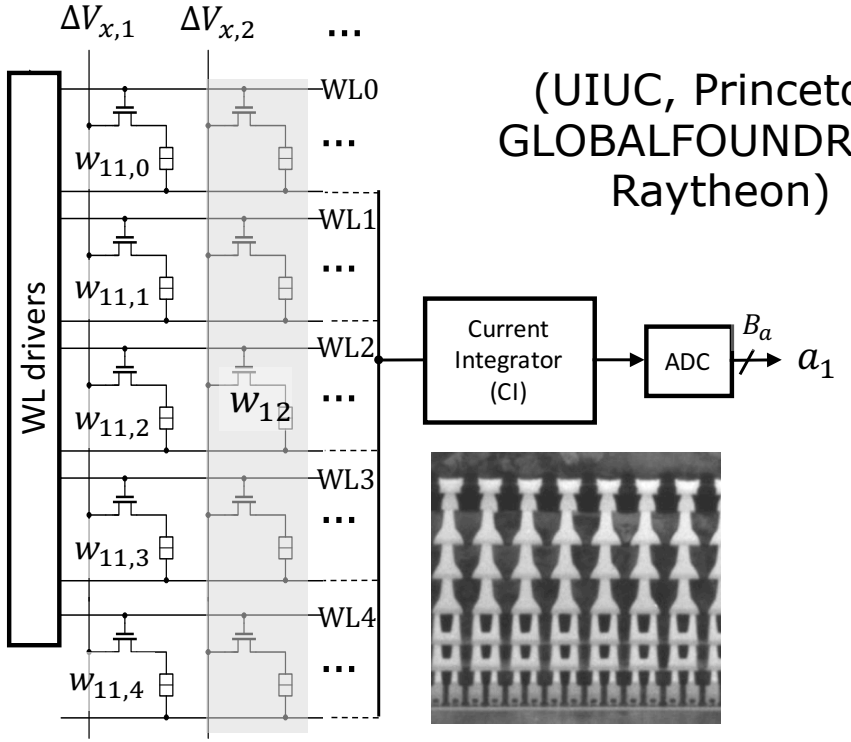


(UIUC, Micron)



- Challenges – larger BL caps, smaller BCs, slower devices, higher V_t variations

MRAM DIMA (ISCAS'19)

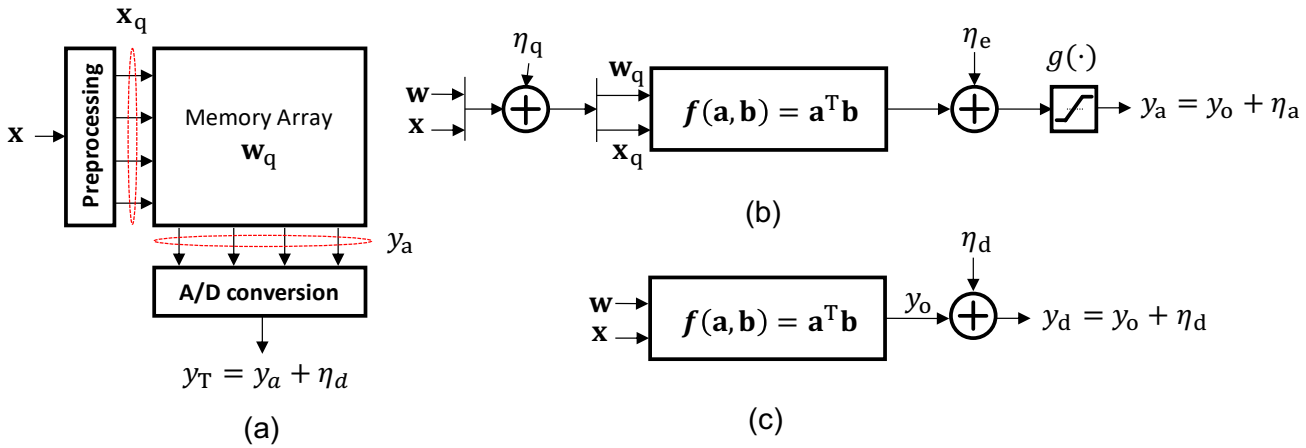


(UIUC, Princeton, GLOBALFOUNDRIES, Raytheon)

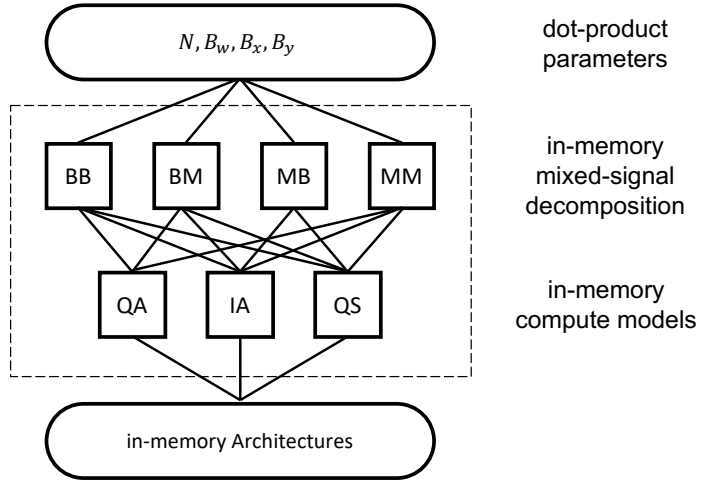
- Challenges – high conductance, small TMR, high cell density, high MTJ variations

DIMA Synthesis and Optimization

SNR & energy metrics and models

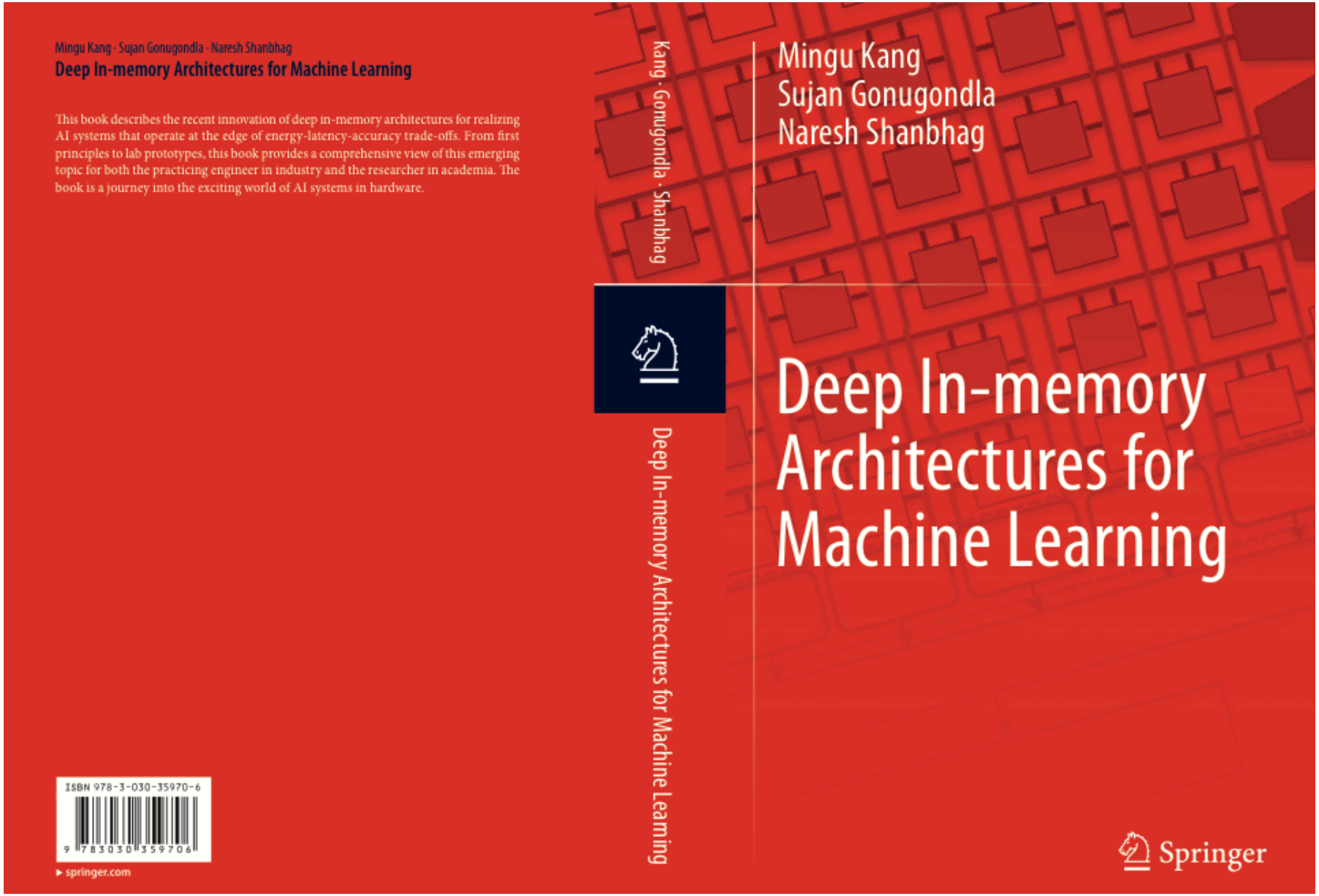


Compositional Framework



- analog vs. digitization SNR
- weight, activation, output precision
- iso-accuracy minimum energy operating point

- synthesize/explain in-memory architectures
- in-memory architectural space = mixed-signal decomposition x compute models



2020 (to appear)

Course Web Page

https://courses.grainger.illinois.edu/ece598nsg/fa2019/
https://courses.grainger.illinois.edu/ece498nsu/fa2019/

http://shanbhag.ece.uiuc.edu

Thank You!

http://shanbhag.ece.illinois.edu