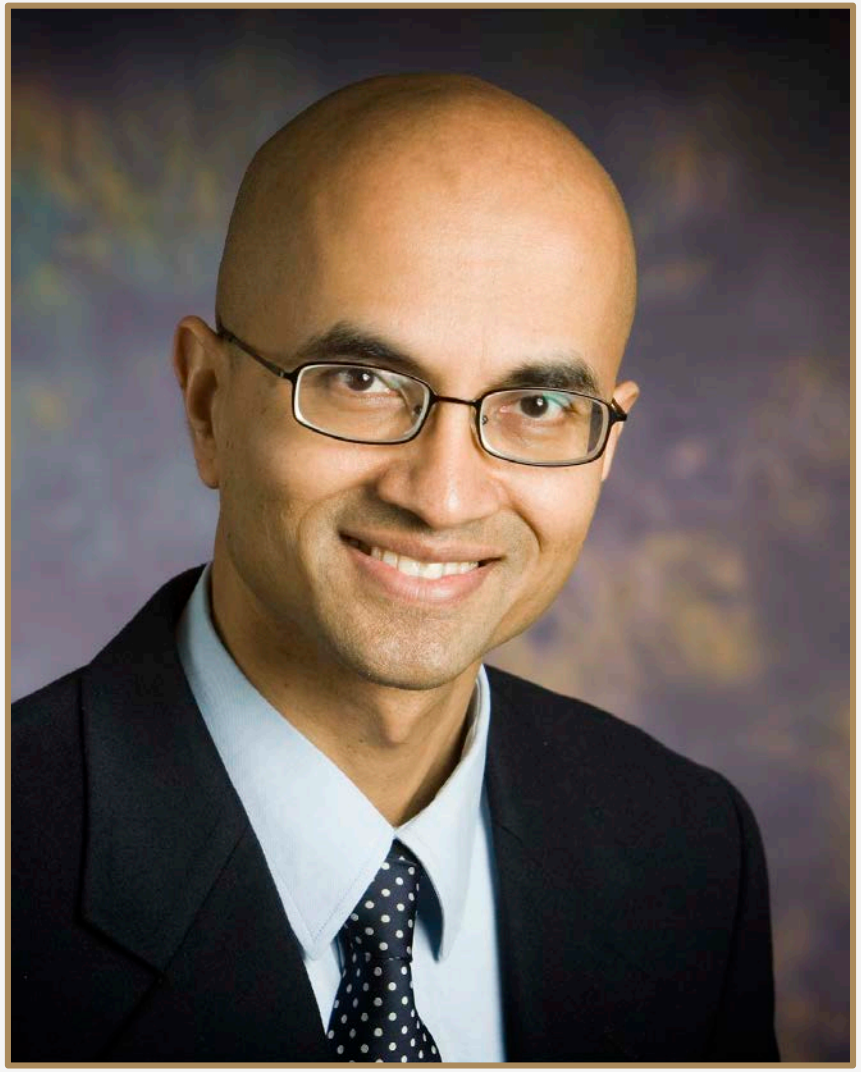


NARESH SHANBHAG

UNIVERSITY OF ILLINOIS AT
URBANA-CHAMPAIGN

MRAM-BASED DEEP IN-MEMORY ARCHITECTURES

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA). The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

THE TEAM

Raytheon MS

Mike Burkland

[DoD systems]

- J. Broussard
- L. Suantak
- J. Goulding
- S. Cole
- T. Gangwer
- R. Kelley
- C. Contreras

GLOBALFOUNDRIES

Ajey Jacob

[devices, foundry]

- S. Soss, D. Brown
- N. Gaul, B. Paul, B. Lanza

Princeton

Naveen Verma (Princeton)

[systems & ckts]

- L.-Y. Chen
- P. Deaville
- B. Zhang

University of Illinois at Urbana-Champaign

Pavan Hanumolu (UIUC)

[mixed-signal ICs]

Naresh Shanbhag (UIUC)

[systems & ckts]

- A. Patil, H. Hua, S. Gonugondla, K.-H. Kim

REALIZING ARTIFICIAL INTELLIGENCE @ THE EDGE

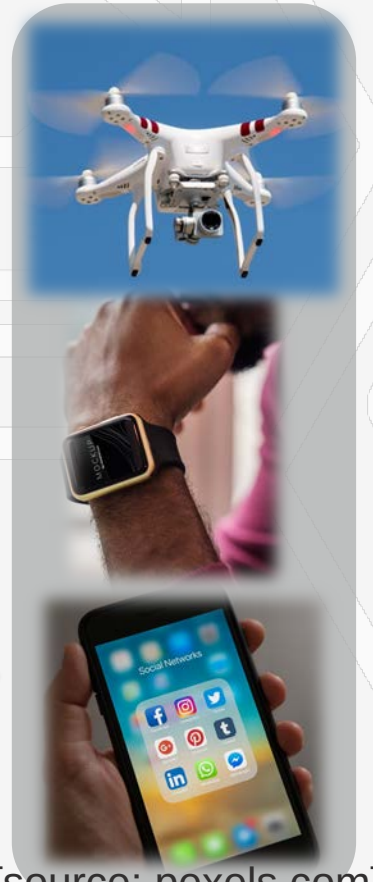
AI in the Cloud

[source: pexels.com]

- client-server model
- efficiency challenged
- security + privacy issues

data

model

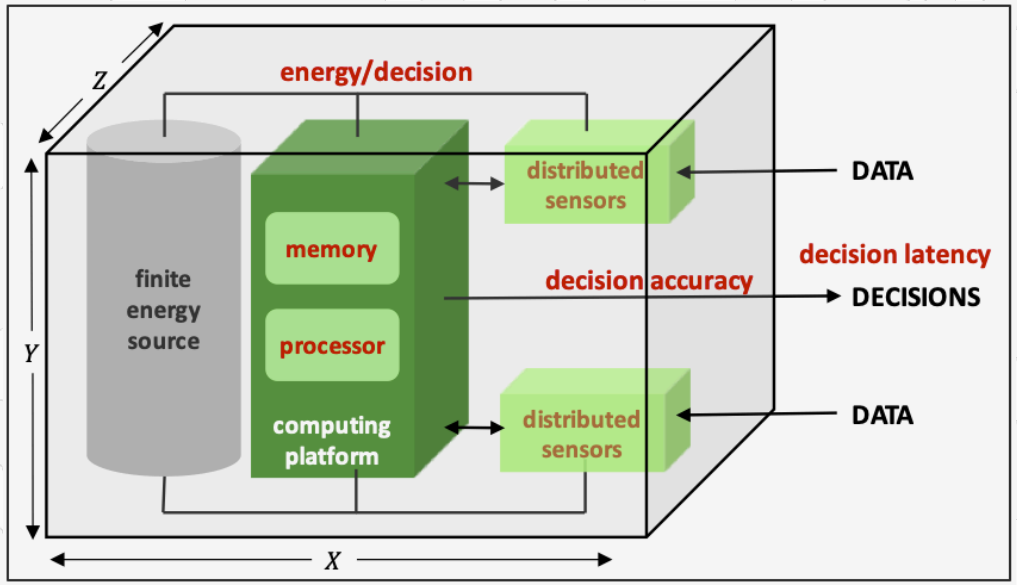


[source: pexels.com]

decisions

AI at the Edge

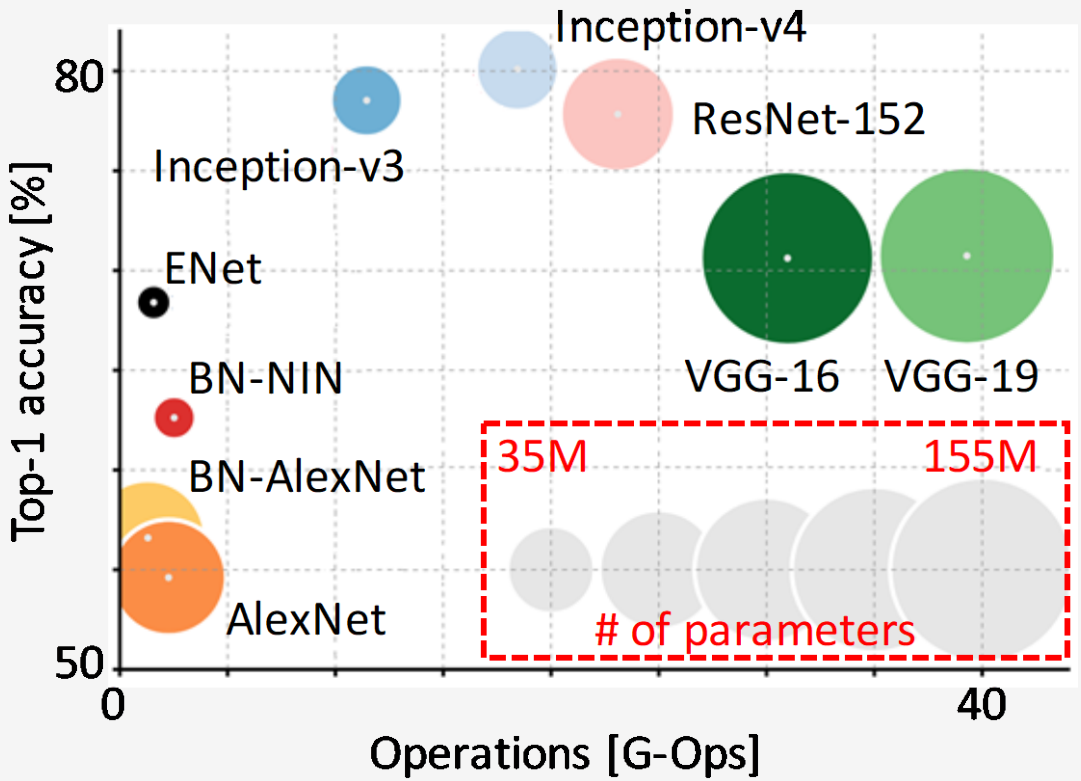
(autonomous cognitive agent)



- energy-latency efficient, secure but……
- volume + resource constrained

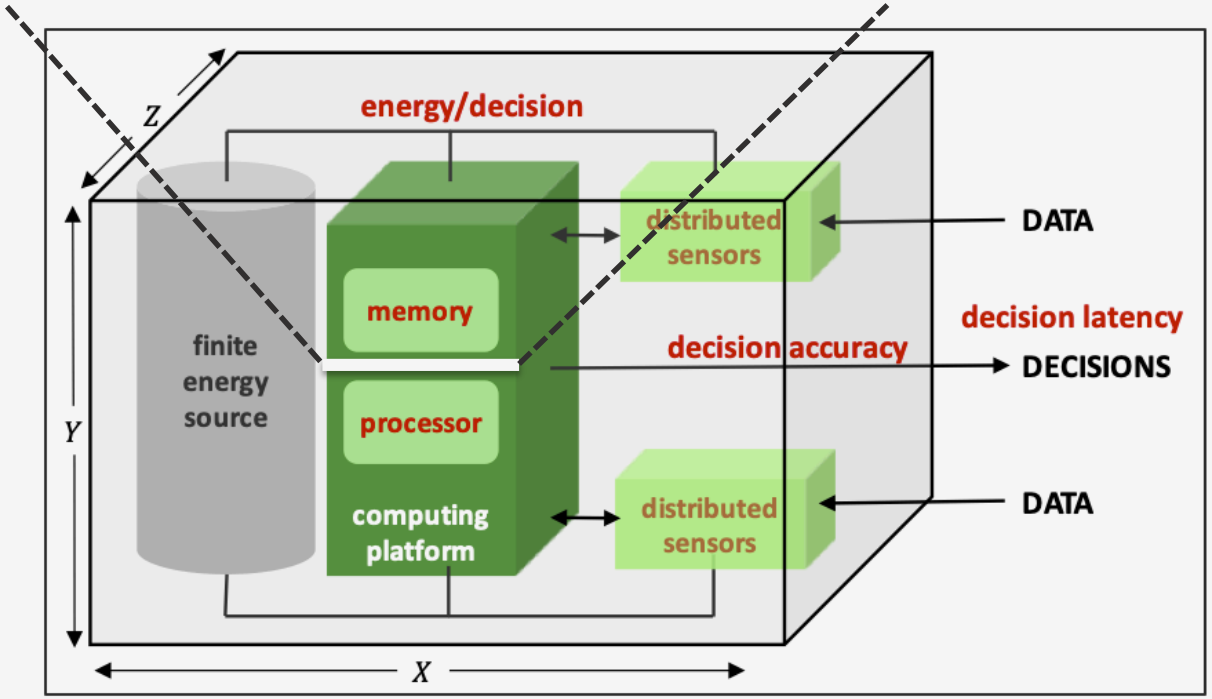
THE DATA MOVEMENT PROBLEM

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



[Canziani, Arxiv'16]

The Memory Wall in von Neumann Architectures



fundamental question

how do we design **intelligent autonomous machines** that **operate at the limits** of **energy-delay-accuracy**?

Systems on Nanoscale Information fabriCs

www.sonic-center.org

[2013-'17]

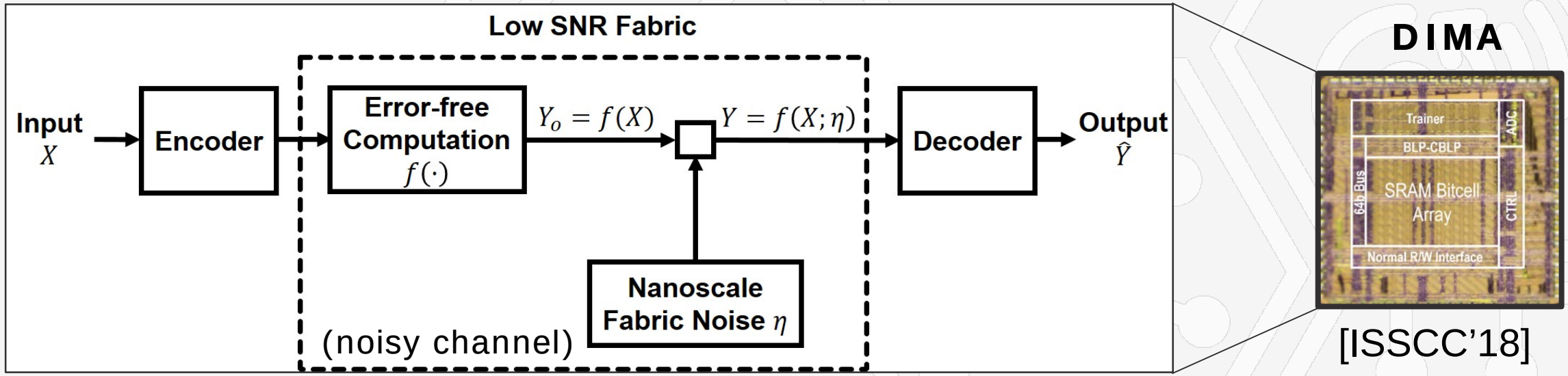
(STARnet Program by Semiconductor Research Corporation & DARPA)

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*, January 2019.

SHANNON-INSPIRED MODEL OF COMPUTING



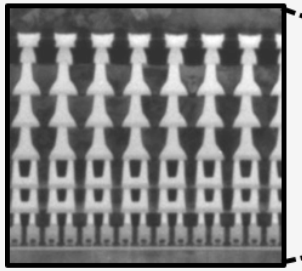
1. use **information-based metrics** e.g., mutual information $I(Y_o; \hat{Y})$
2. design **low SNR fabrics**, e.g., deep in-memory architecture (DIMA)
3. develop **statistical error-compensation (SEC)** techniques

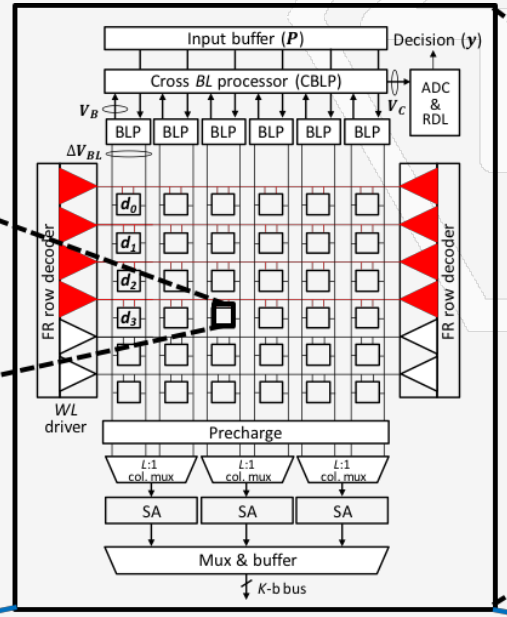
FRANC OBJECTIVE

GLOBALFOUNDRIES **PRINCETON & UIUC** **RAYTHEON MISSILE SYSTEMS**

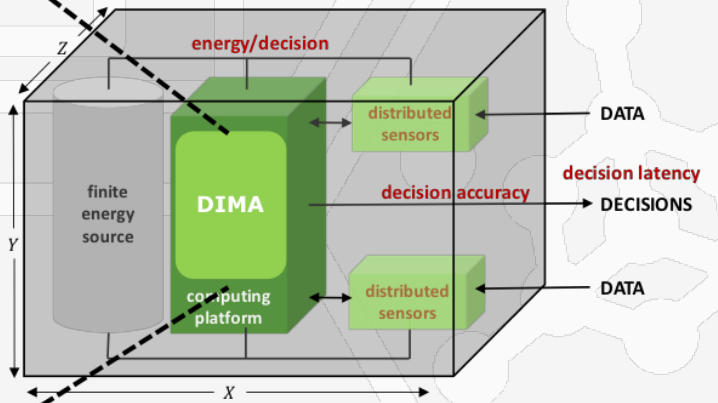
deep in-memory arch.

MRAM device

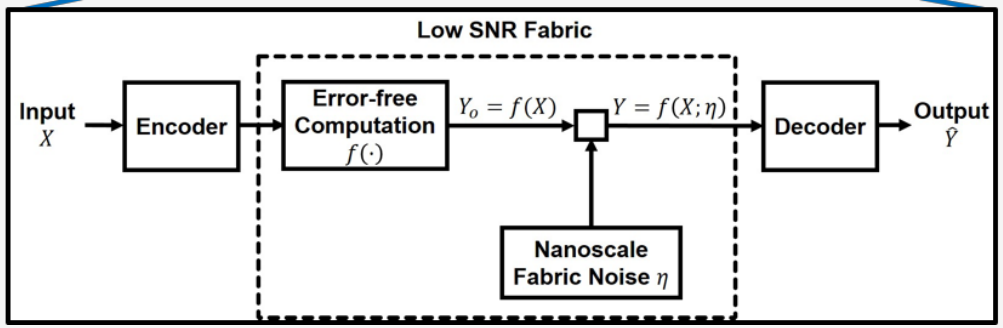




autonomous platforms



thermally limited



Shannon-inspired compute models

integrate

the MRAM device

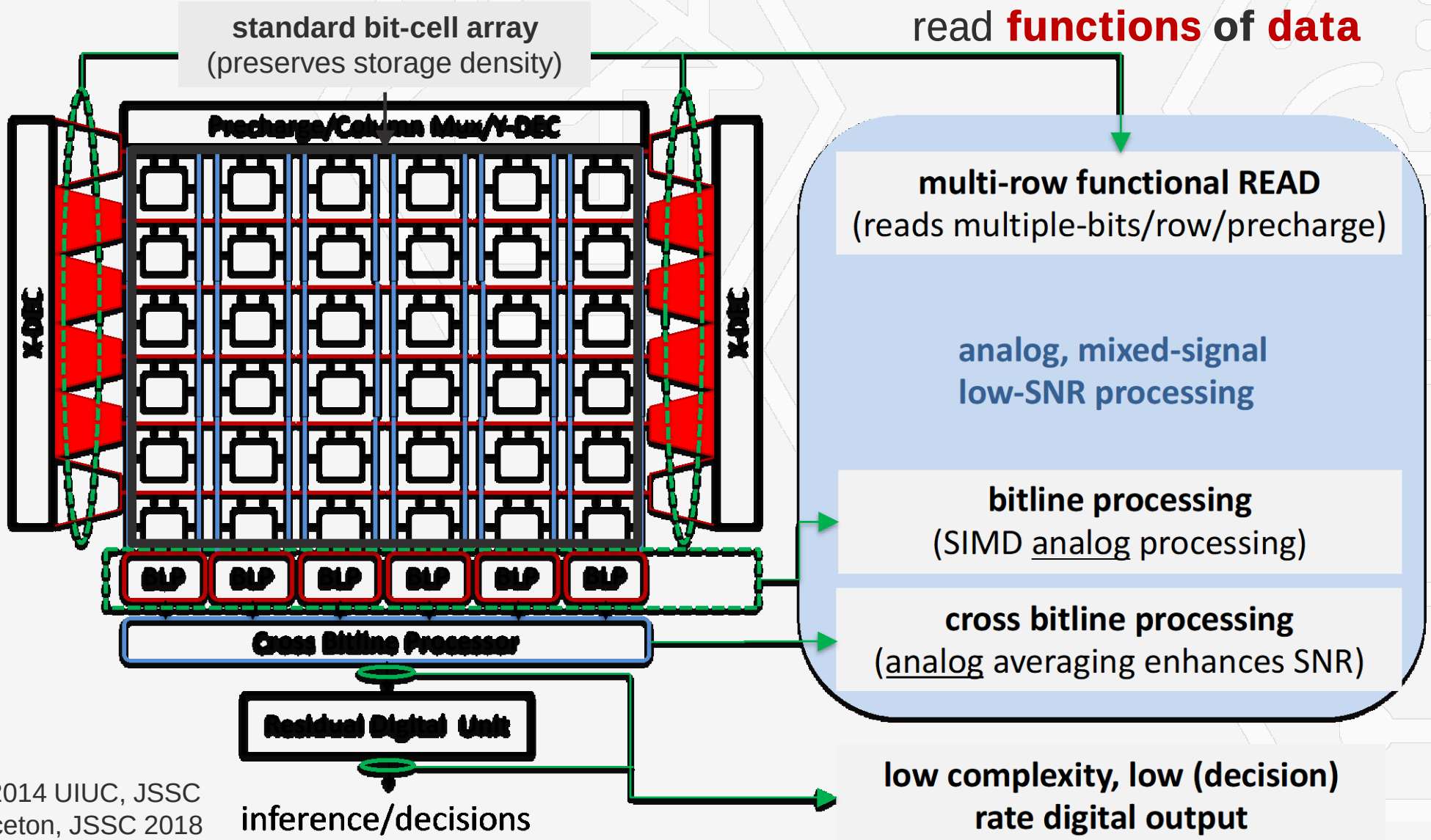
within a

deep in-memory architecture (DIMA)

using

Shannon-inspired compute models

THE DEEP IN-MEMORY ARCHITECTURE (DIMA)

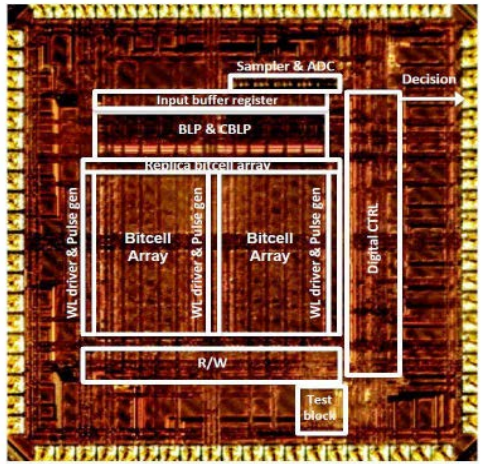


[ICASSP 2014 UIUC, JSSC 2017 Princeton, JSSC 2018 UIUC]

SRAM DIMA PROTOTYPES

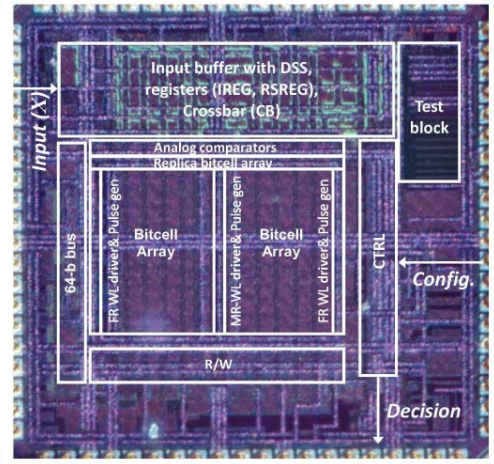
100X EDP reduction over von Neumann equivalent* @ iso-accuracy

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



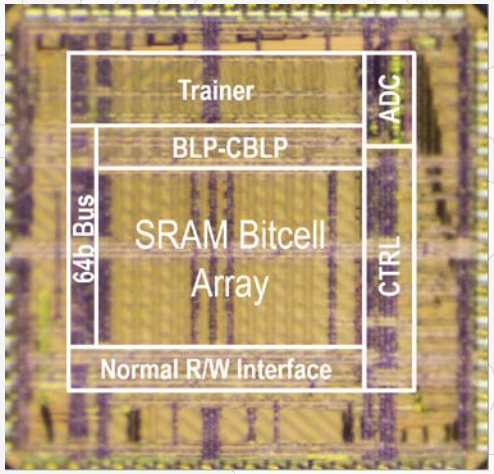
Multi-functional inference processor (65nm CMOS)

[Feb. JSSC'18]



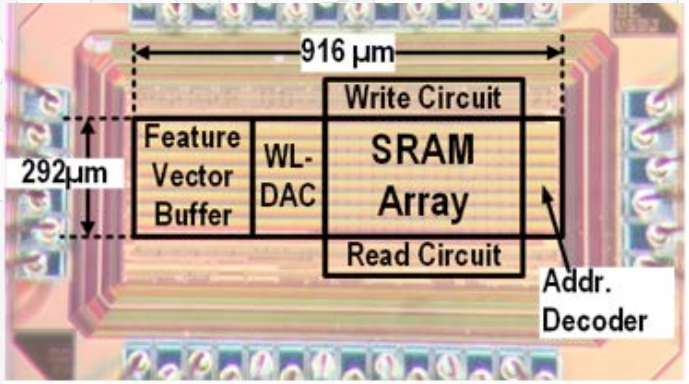
Random forest processor (65nm CMOS)

[ESSCIRC'17, JSSC July'18]



On-chip training processor (65nm CMOS)

[ISSCC'18, JSSC Nov.'18]

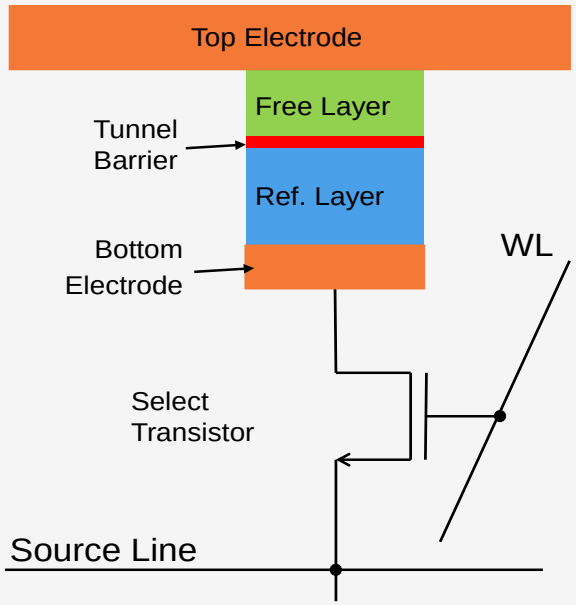


Fully (128) row-parallel compute (130nm CMOS)

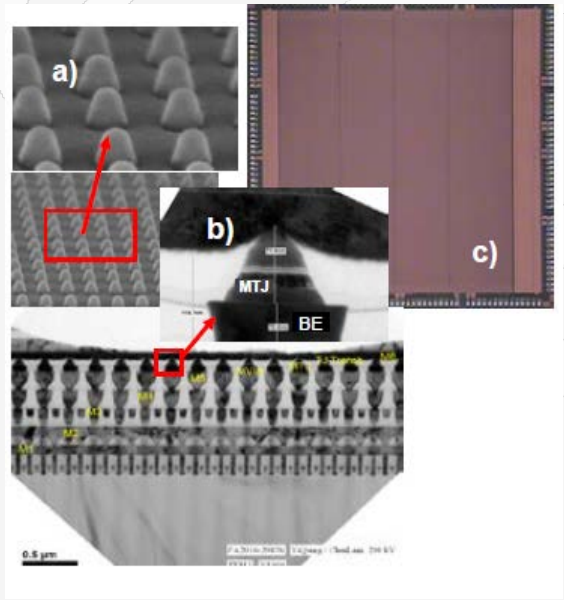
[VLSI'16, JSSC'17]

MRAM OPPORTUNITIES & CHALLENGES

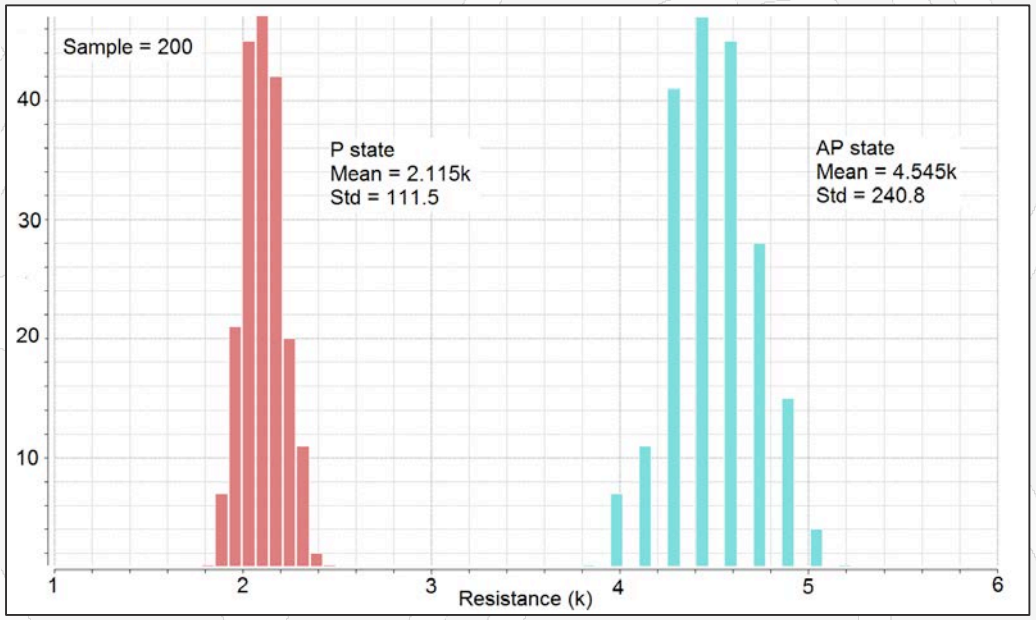
pMTJ stack



40Mb MRAM demo



MTJ resistance distribution



Opportunities:

high density → high on-chip compute density

non-volatility → ultra low-power duty-cycled operation

in production → enables system prototyping &
reduces time to DoD availability

Challenges:

high conductance → large array currents

small TMR → high sensitivity readout

huge $R_{MTJ} - R_{MOS}$ limits cell compute models

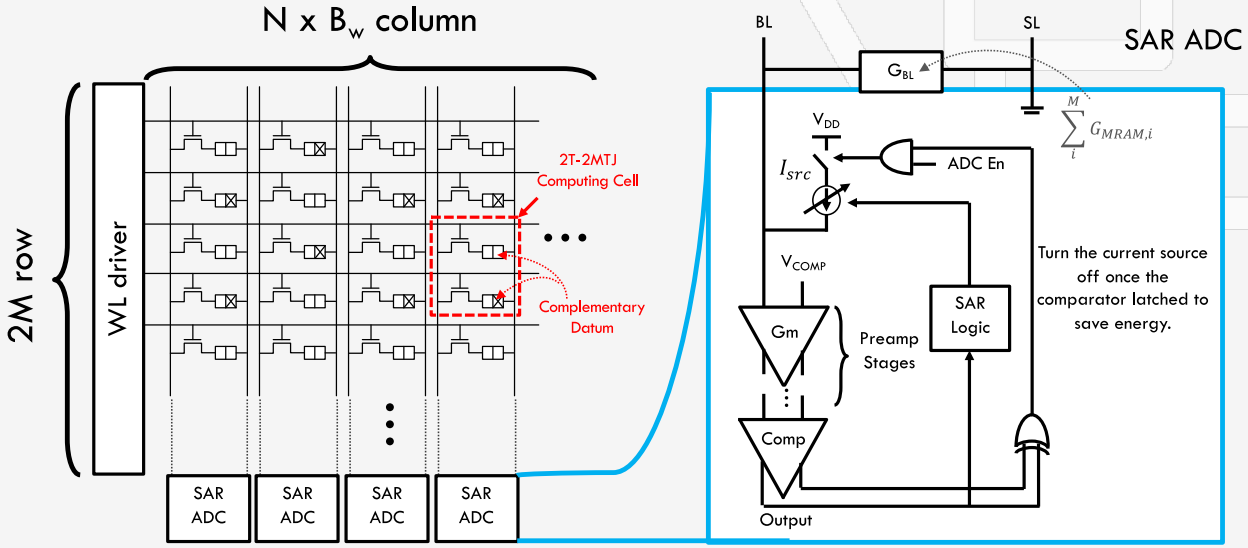
high cell density → severe pitch-matching const.

high MTJ process variation restricts bit-line SNR

MRAM-BASED DEEP IN-MEMORY ARCHITECTURES

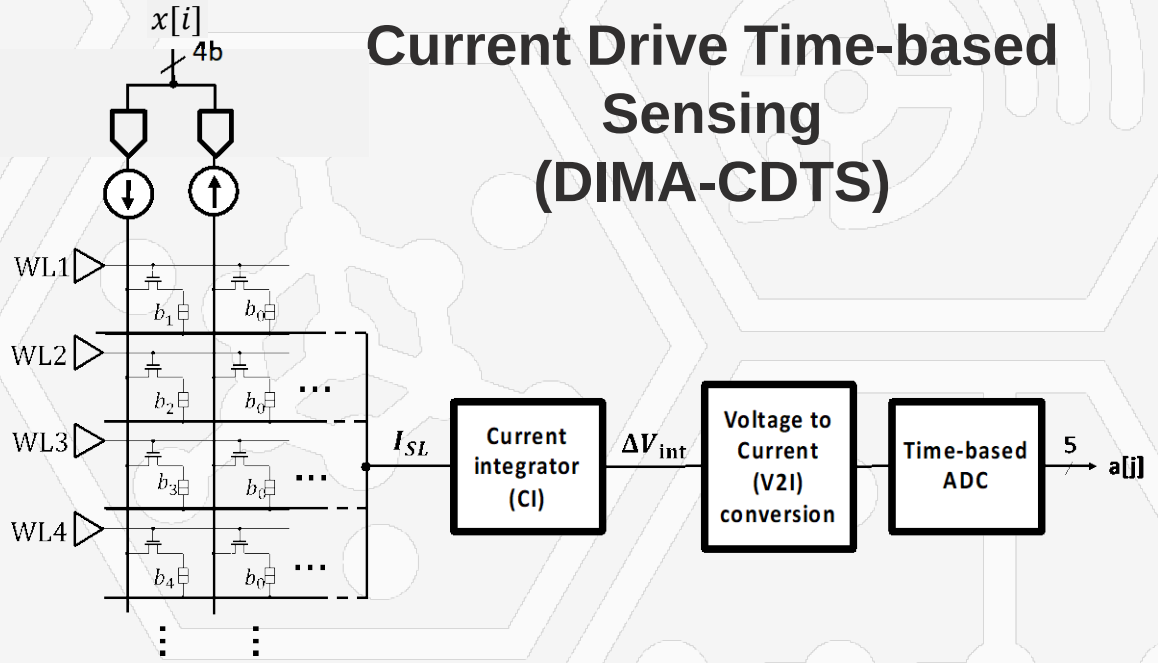
four DIMAs proposed – two selected for prototyping

Bit-line Sensing (DIMA-BLS)



- $M \times N$ single-shot matrix vector multiply
- Functional read: binary dot products on BLs
- Compute cell: 2T-2MTJ (1b XNOR/AND)
- BL voltage sensing via 4-bit SAR ADCs
- SNR vs. energy trade-off due to sensing circuits

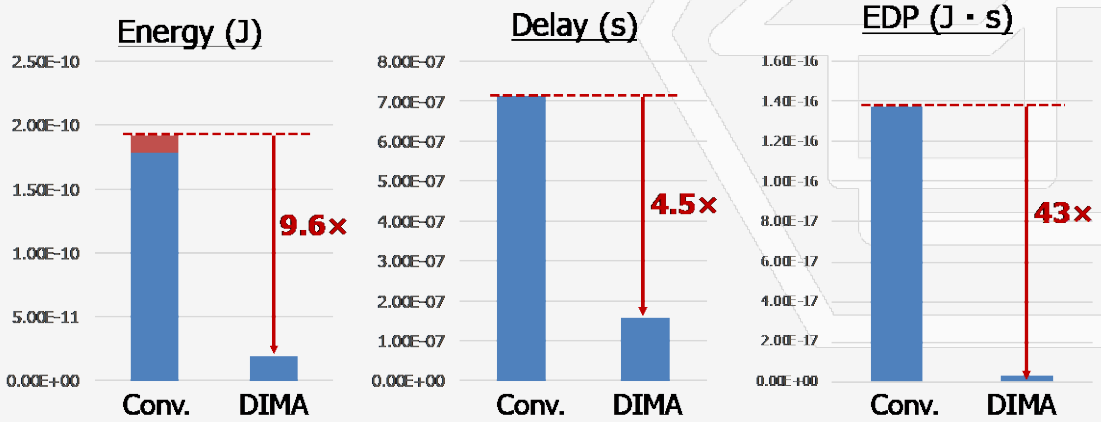
Current Drive Time-based Sensing (DIMA-CDTS)



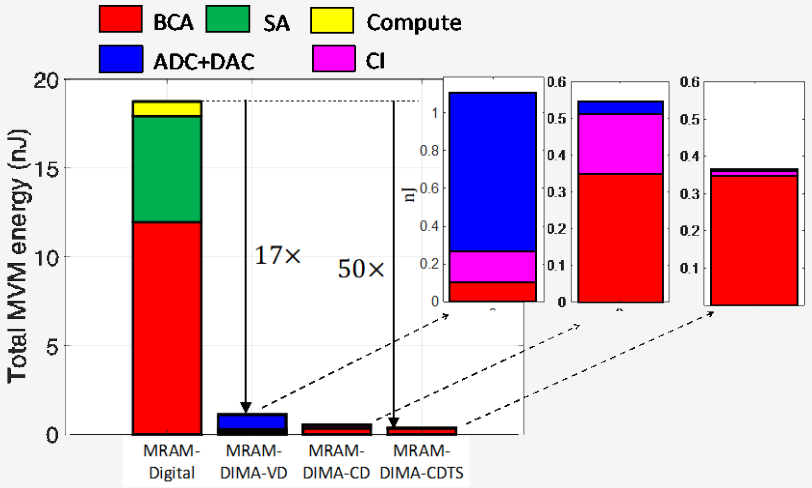
- $M \times N$ single-shot matrix vector multiply
- Functional read: 5-bit $\times$ 4-bit dot products on SLs
- Compute cell: 1T-1MTJ (4×1 multiplication)
- SL current sensing via time-based 5-bit ADCs
- SNR vs. energy trade-off due to sensing circuits

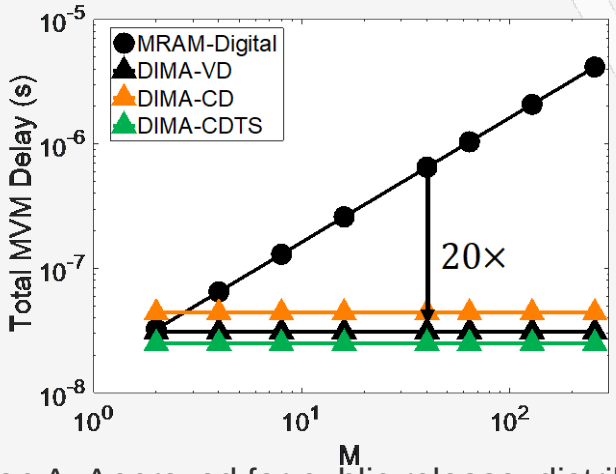
EDP GAINS OVER VON NEUMANN EQUIVALENT

DIMA-BLS



DIMA-CDTS





MVM: matrix-vector multiply

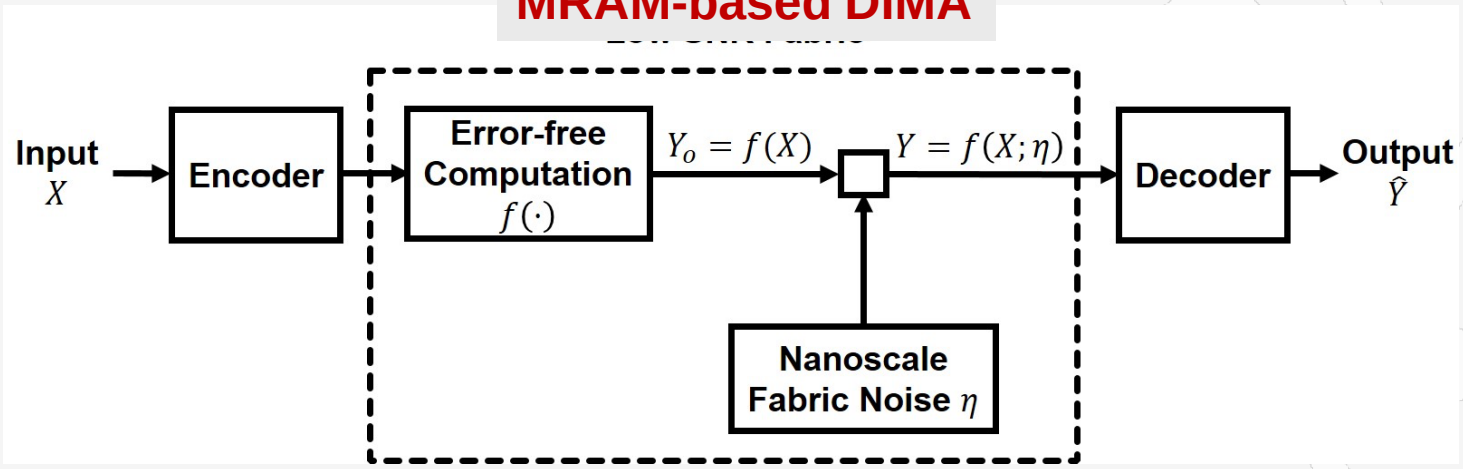
43× EDP reduction
(256-dimensional dot products)

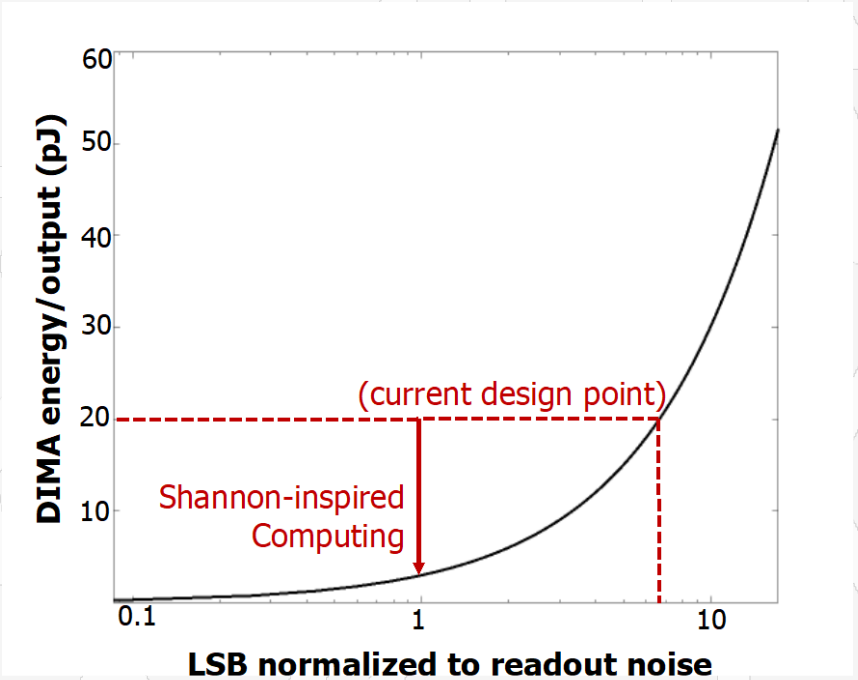
(includes all peripherals)

250×-to-1000× EDP reduction
(128-dimensional dot products)

SHANNON-INSPIRED COMPUTE MODELS

MRAM-based DIMA





- use **information-based metrics**
- develop **error-compensation techniques**

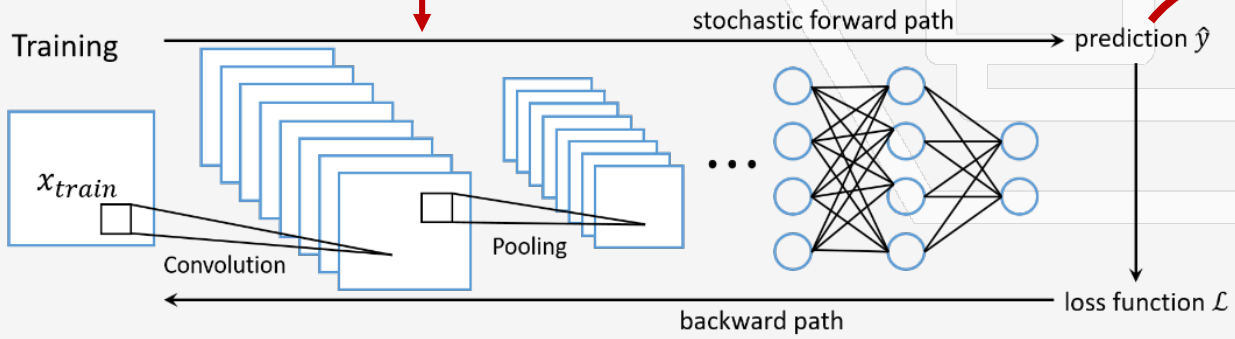
- relaxes MRAM-based DIMA's compute SNR requirements → enables substantial energy savings
- two methods: stochastic data-driven hardware resilience & coded DIMA

STOCHASTIC DATA-DRIVEN HARDWARE RESILIENCE

DIMA-aware Stochastic Training

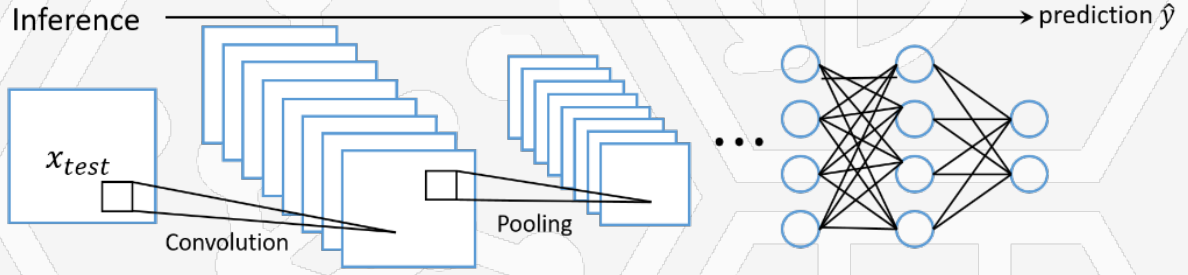
G (MRAM variation, ADC noise)

Model Parameters $\theta_{STOCH}(x, G)$

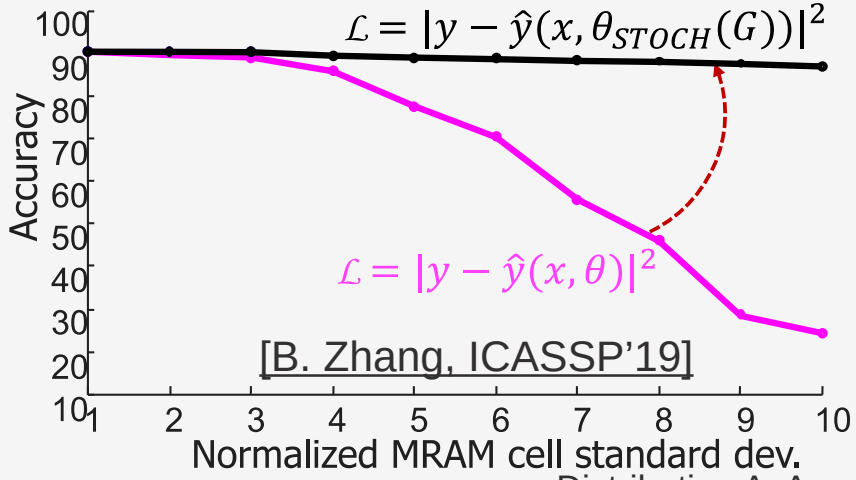


MRAM-based DIMA Inference

g_i



Simulated MRAM-based BNN (CIFAR-10):

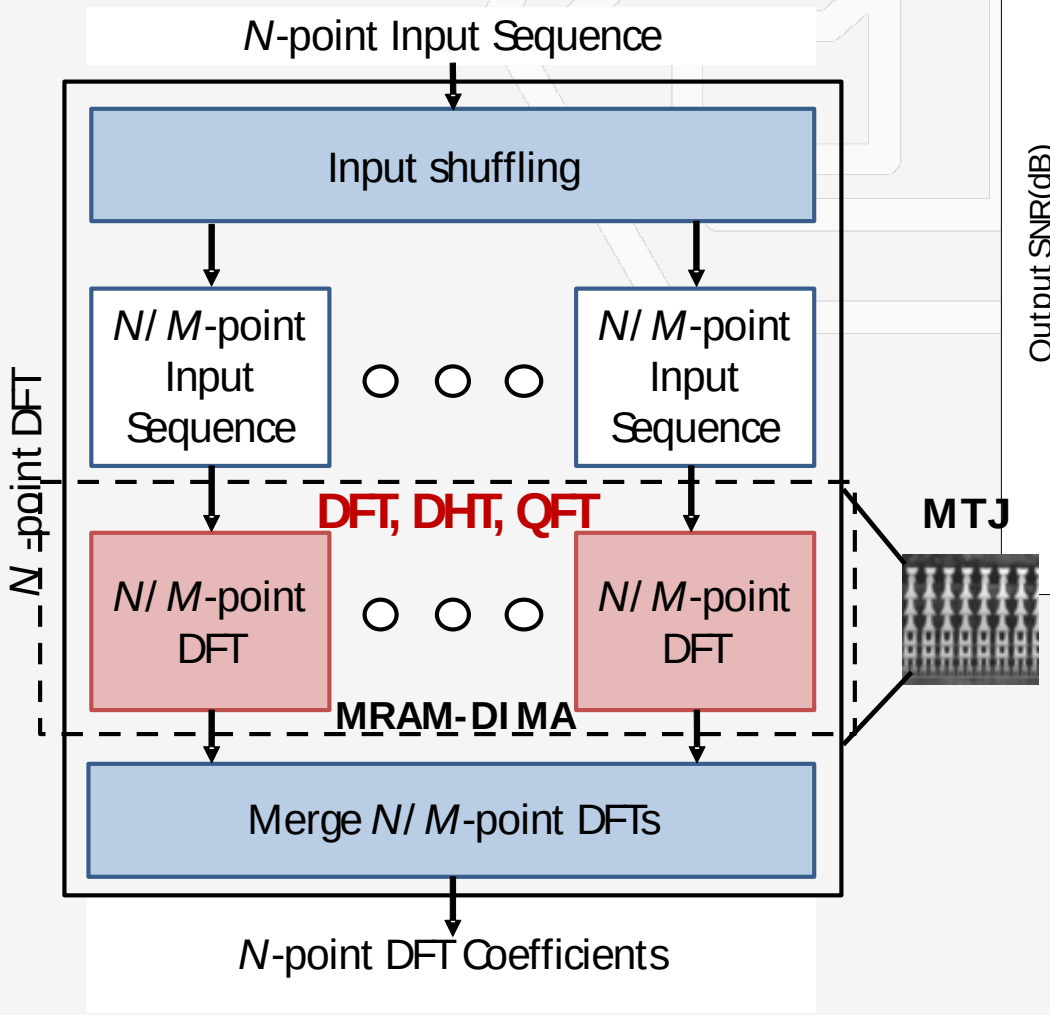


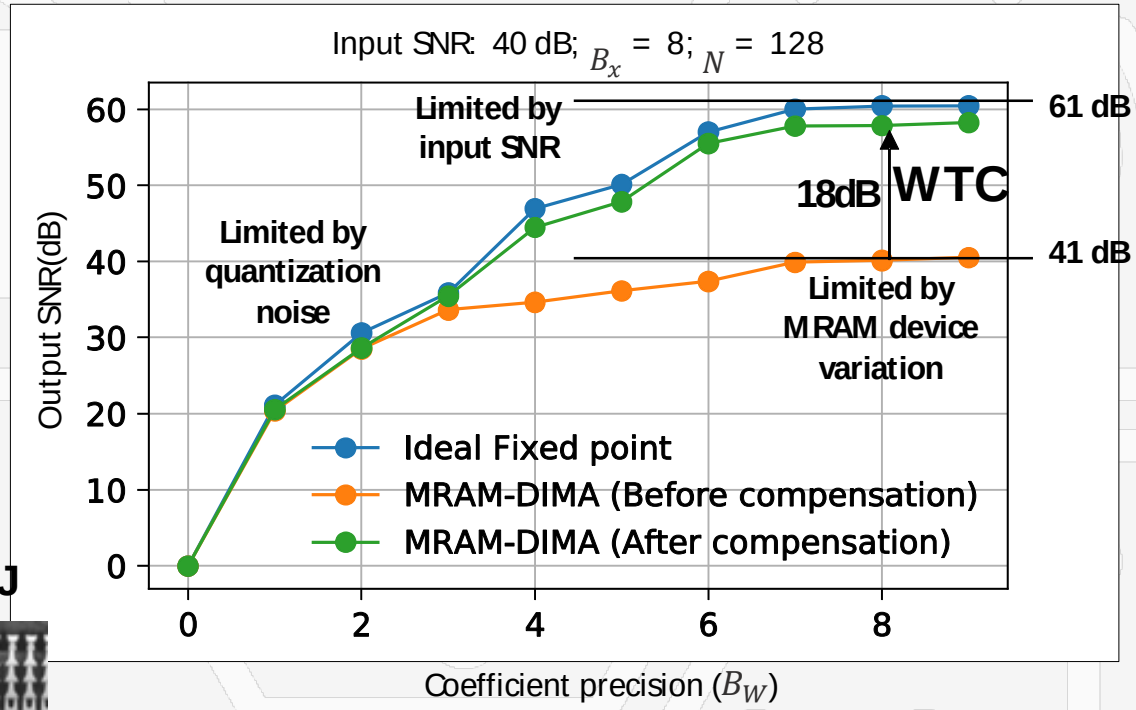
- noise-tolerance enhanced well beyond MTJ variations
- can be exploited for energy-SNR tradeoff

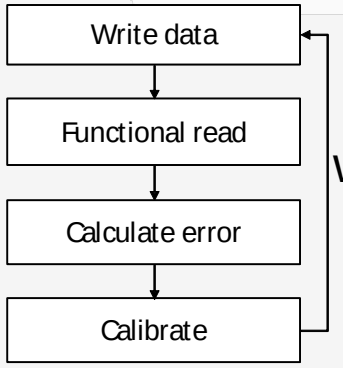
CODED DIMA - ENHANCING DIMA'S COMPUTE SNR

MTJ variation aware coding enables HIGH system SNR in spite of LOW circuit SNR

partial decimation-in-time Architecture for mapping FFT on DIMA





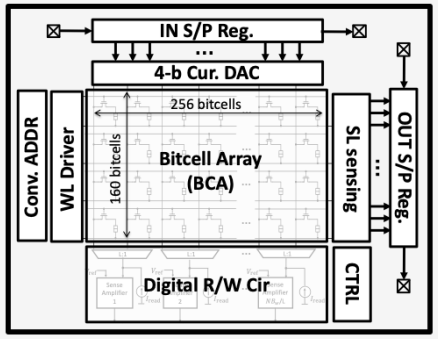


write-time compensation (WTC) of MRAM variations

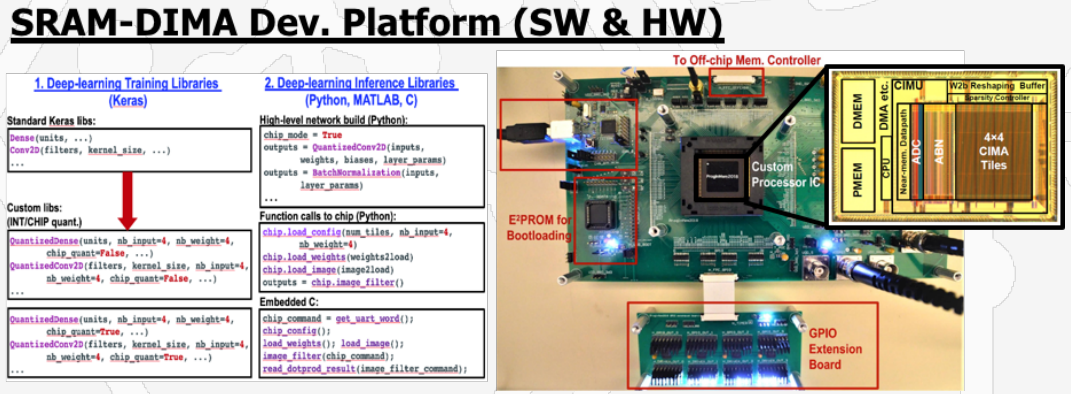
CURRENT STATUS

- **verified & quantified RMS's mission** capability enabled by MRAM-based DIMA
- MRAM-based DIMA demonstrates > **2 orders-of-magnitude EDP reduction**

DIMA prototype designs taped out



Technology transition via SRAM-based DIMA



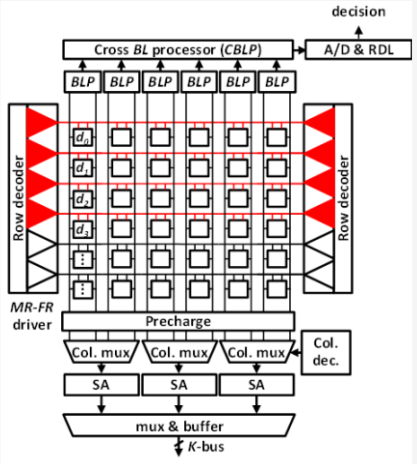
[H. Jia, *arXiv*:1811.04047, Princeton]

- (2019) [single-bank] validate EDP gains & calibrate models
- (2020) [multi-bank] enable scaling and system integration

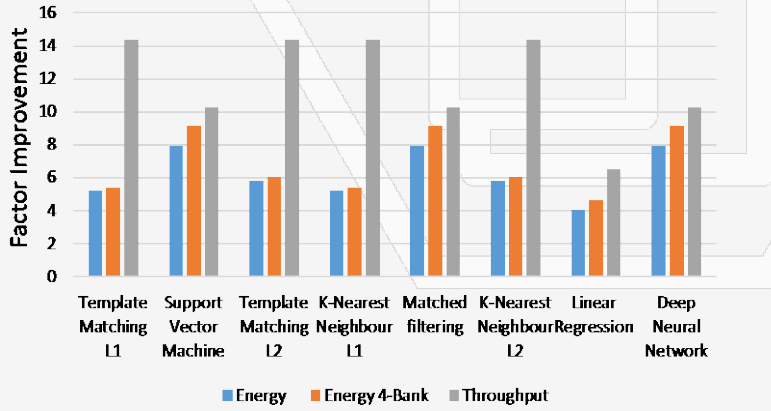
- DIMA technology transition to RMS for algorithm prototyping & evaluation
- facilitates future MRAM-based DIMA integration by RMS

NEXT STEPS & SUMMARY

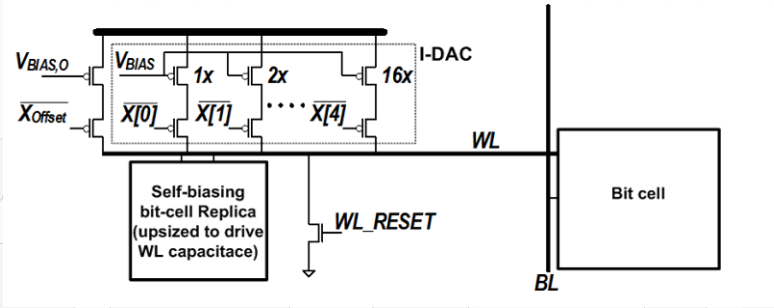
Multi-functioned DIMAs for Diverse Applications

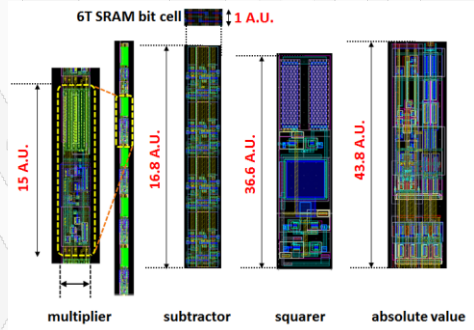


[Srivastava, et al., ISCA'18]



Parameterized DIMAs for Platform-design Tools

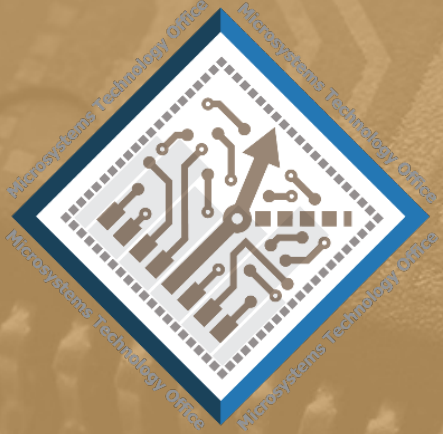




Summary

Mission: To realize > 200X in EDP gains in DoD workloads by integrating **MRAM device** within **Deep In-memory Architectures** using **Shannon-inspired compute models**

FRANC Program: "*to provide the foundation for new materials technology and new integration approaches to be exploited in pursuit of novel compute architectures*"



ERI ELECTRONICS RESURGENCE INITIATIVE

SUMMIT

2019 | Detroit, MI | **July 15 - 17**