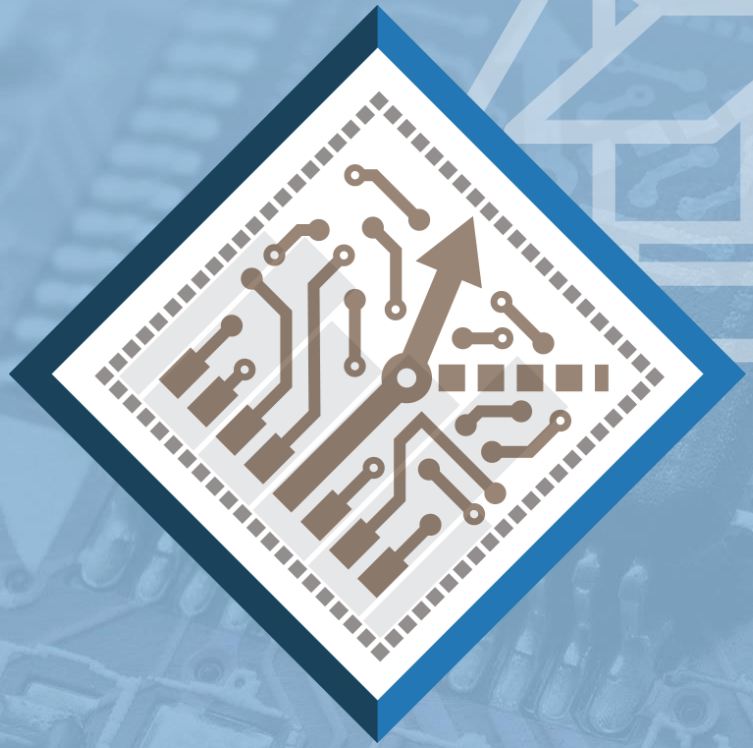


FOUNDATIONS REQUIRED FOR NOVEL COMPUTE (FRANC)

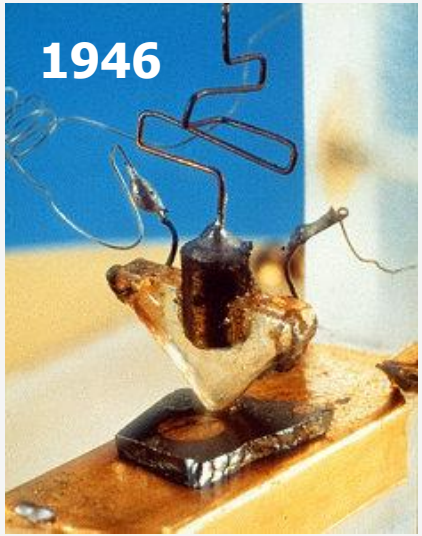
YOUNG-KAI CHEN

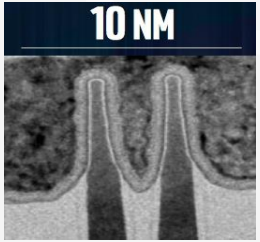
PROGRAM MANAGER
DARPA/MTO

CURRENT COMPUTING

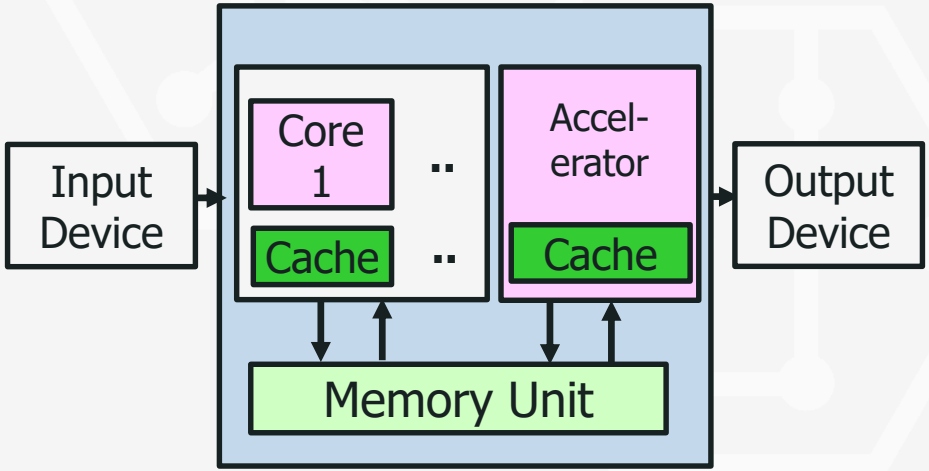
A robust von Neumann architecture powered by the CMOS scaling with Moore's Law

Highly Scaled CMOS



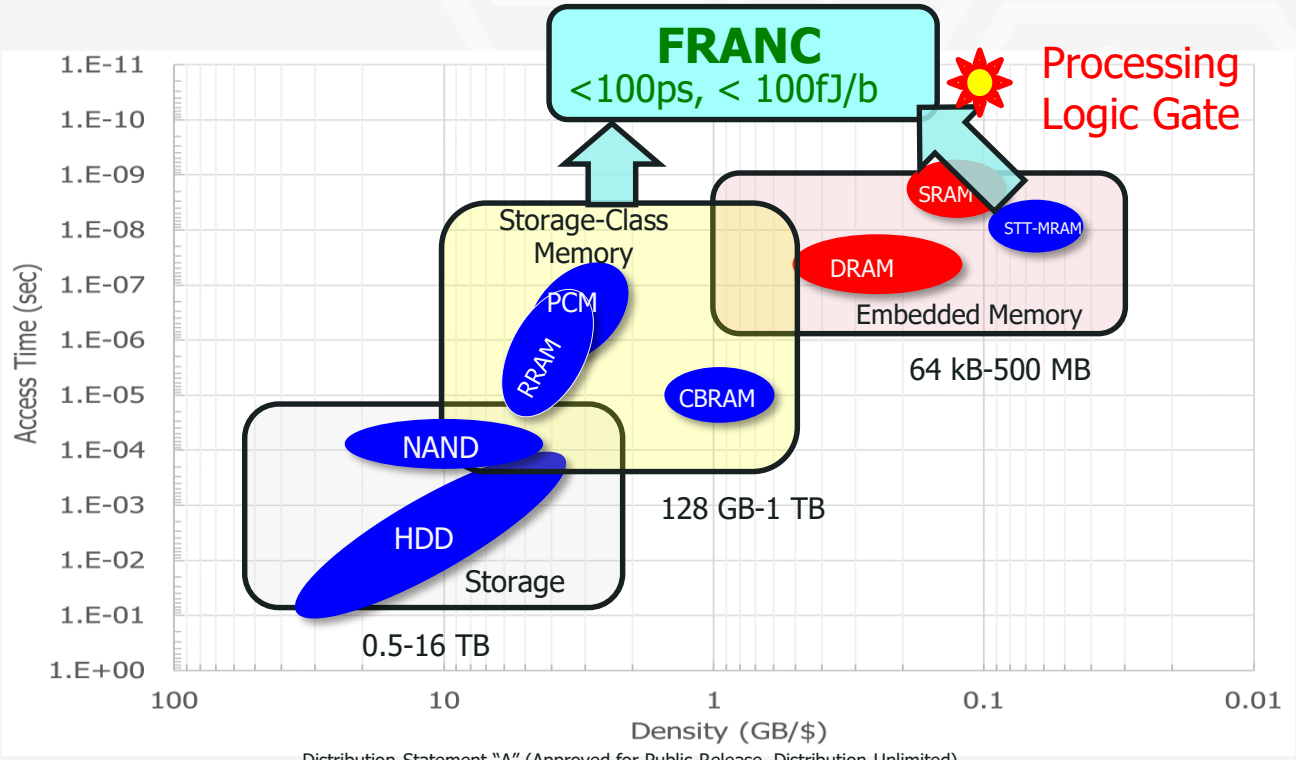


Parallelization to reduce latency



TODAY'S PROCESSOR SPEED IS 100X FASTER THAN MEMORY

FRANC utilizes new materials and devices to make 10x advances in embedded non-volatile memories with speed as SRAM and density as storage-class memory



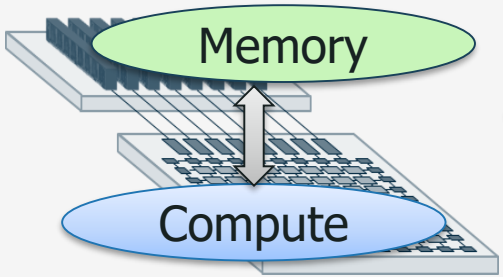
Non-Volatile

Volatile

Reference: Siva Sivaram, Western digital Corporation, August 9, 2016, Flash Memory Summit

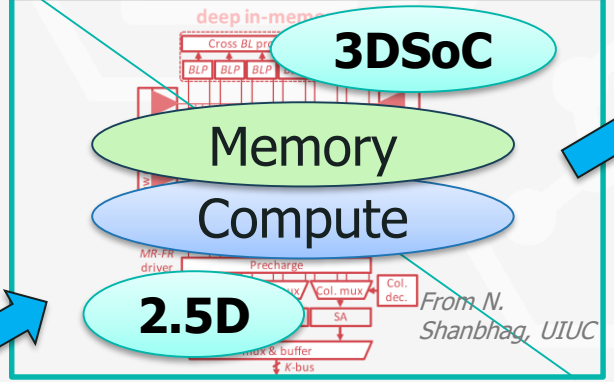
FRANC WILL ENHANCE MEMORY-CENTRIC COMPUTING ARCHITECTURE

Present von Neumann Computing



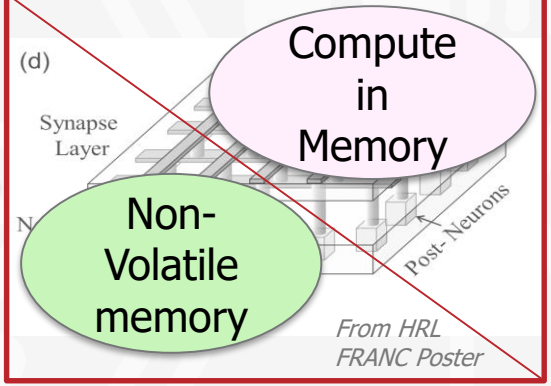
From S. Mitra at Stanford

Enhanced von Neumann Computing



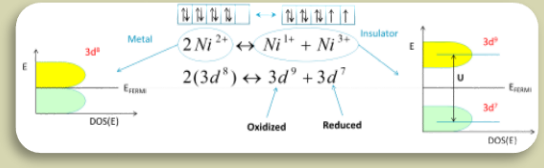
From N. Shanbhag, UIUC

Emerging Memory-Centric Computing

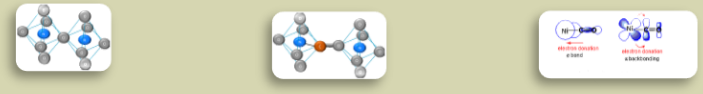


From HRL FRANC Poster

DEVELOPMENT OF NEW MATERIALS



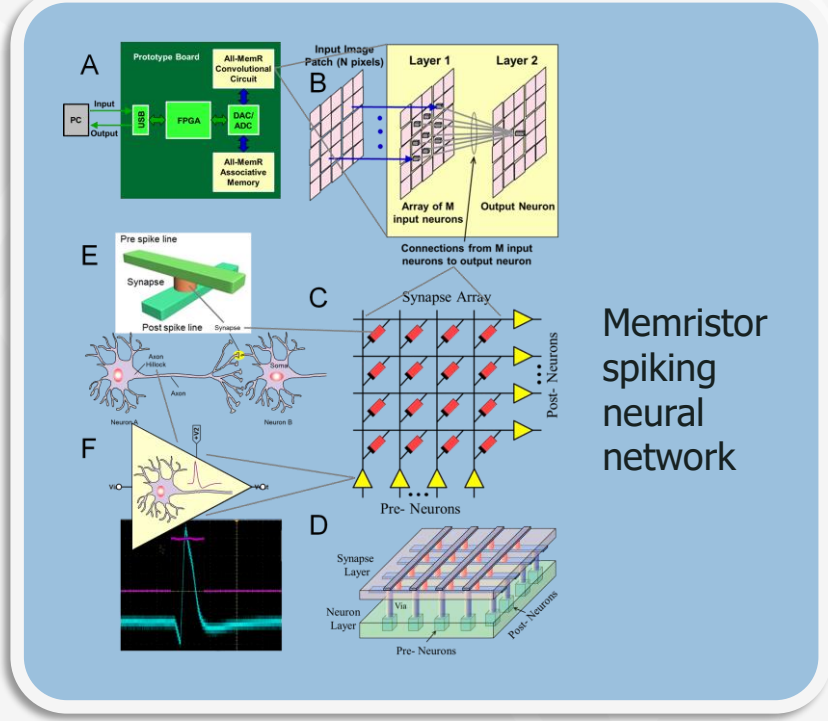
Correlated electron RAM (ceRAM)



ceRAM contruction

ceRAM preliminary data

Applied Materials



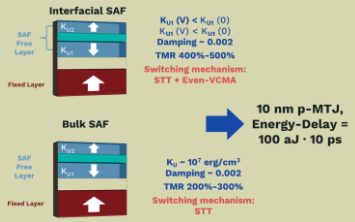
Memristor spiking neural network

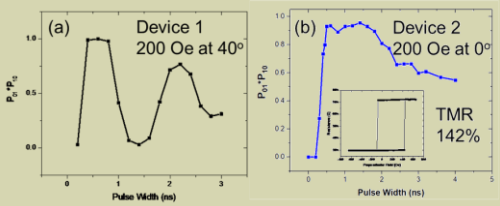
HRL Laboratories

NEW DEVICE DEVELOPMENT

Voltage-Controlled Magnetic tunnel junction memory

- Ultra fast switch
- Low switch energy
- Small feature size

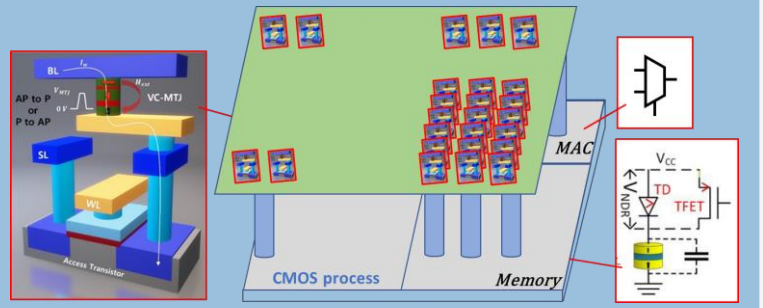


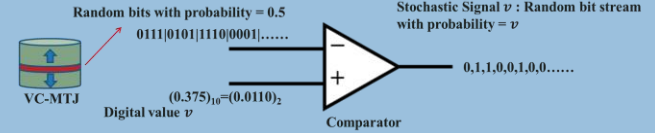


Preliminary data

University of Minnesota

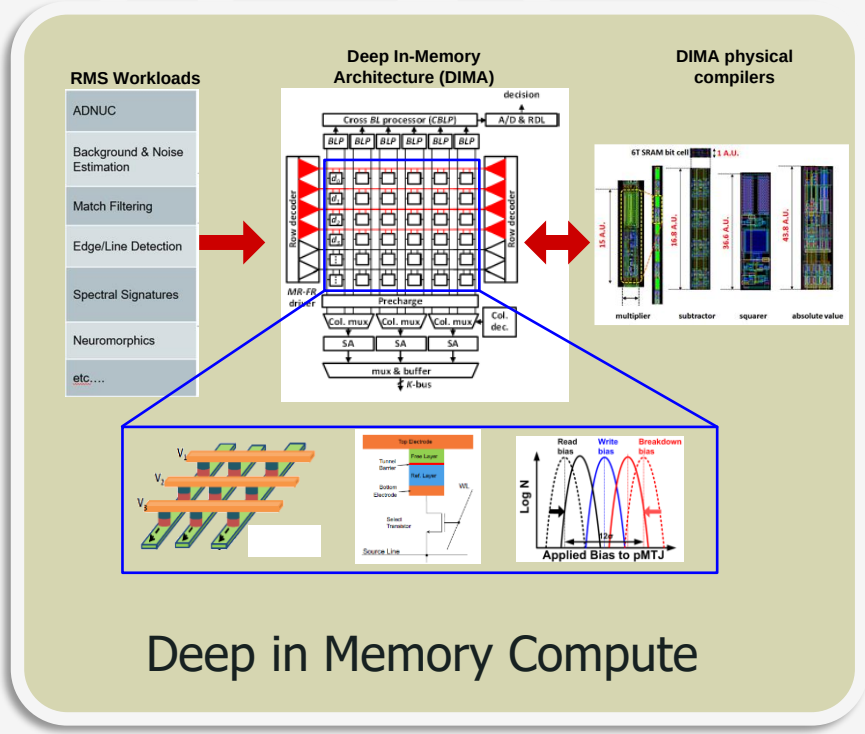
Magnetic tunnel junction memory for stochastic computing





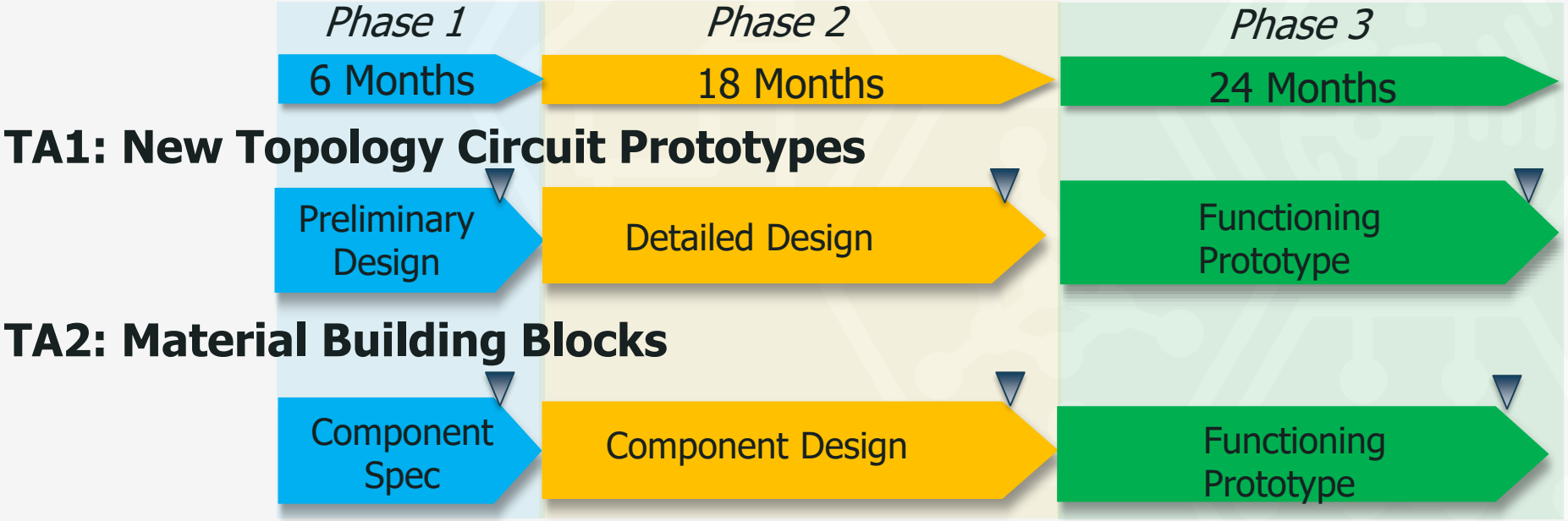
University of CA, Los Angeles

NOVEL COMPUTING ARCHITECTURES



University of Illinois at UC

FRANC PROGRAM STRUCTURE AND METRICS



Preliminary Design:
- Simulated > 10x performance enhancement over state of art
- Define detailed metrics

Detailed Design:
- Implement test samples
- Emulation of performance on benchmarks
- Down selections

Functioning Prototype:
- Execution on benchmark performance
- Transition for commercialization

ERI ELECTRONICS RESURGENCE INITIATIVE

SUMMIT

2018 | SAN FRANCISCO, CA | **JULY 23-25**

STEVE PAWLOWSKI

VICE PRESIDENT
MICRON, ADVANCED COMPUTING SOLUTIONS

NEW MEMORY TECHNOLOGIES ARE RARE

- Today's available memory technologies emerged in the early 70s
- The way memory is used in systems, the memory hierarchy, has been defined by the evolution of these memories over four decades.

The hierarchy and the hardware and software wrapped around it is as much defined by each memory technology's "limitations" as its "features"

Year of 1st Shipment	Memory Technology
1969	SRAM
1970	DRAM
1971	EPROM
1986	NOR Flash
1995	NAND Flash
1997	MLC Flash
2008	PCM

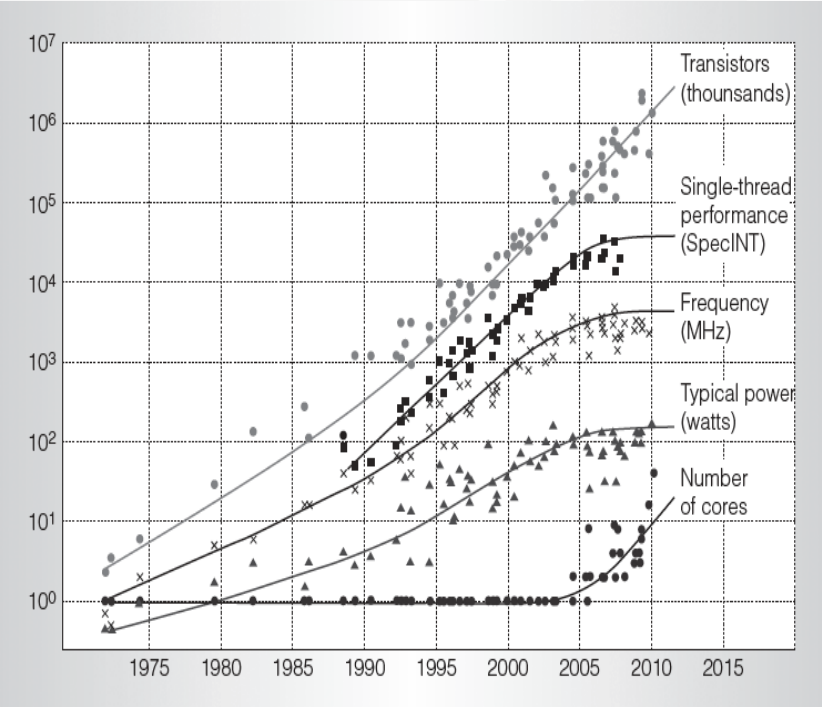
Electron Based

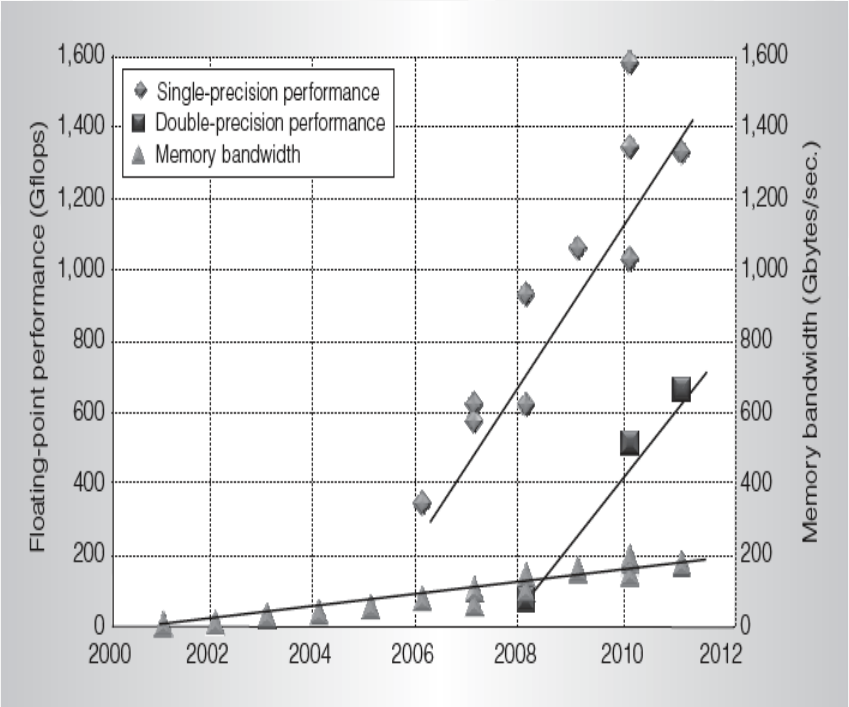
EPROM Derivatives

Memories that have shipped ≥ 1Gb densities

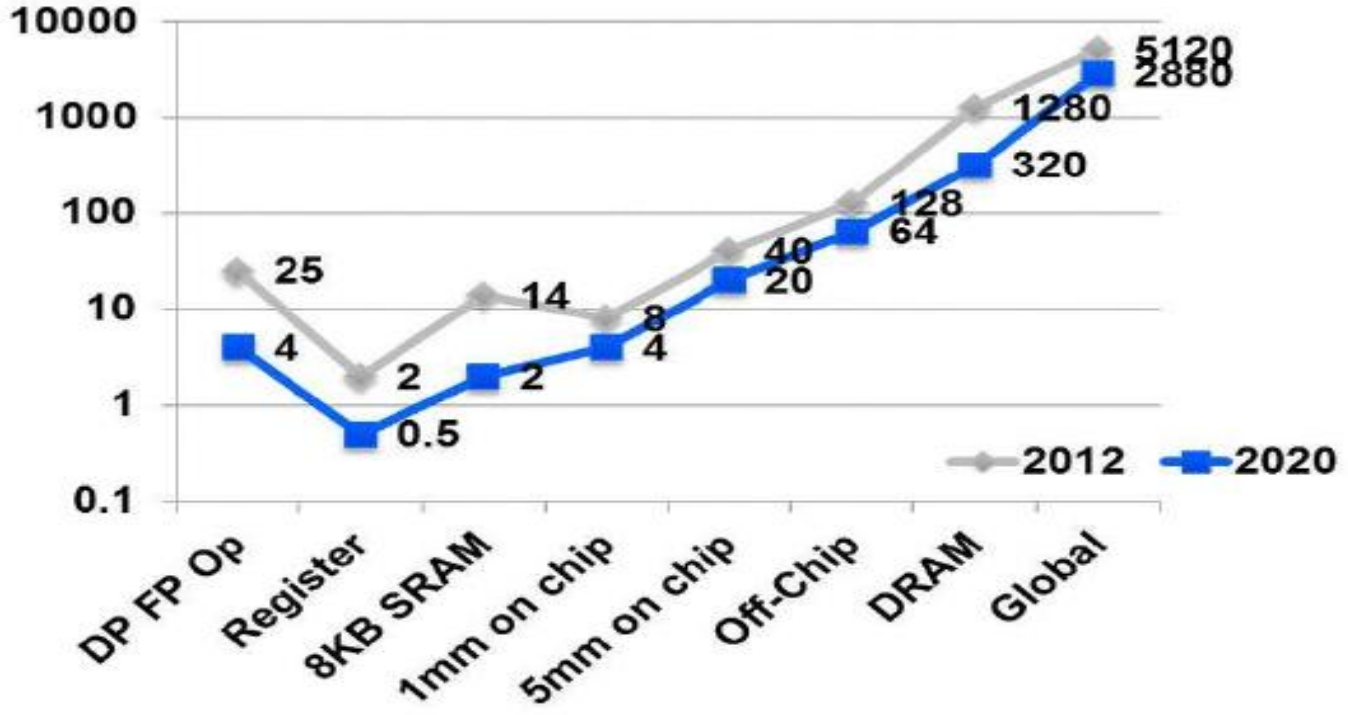
CHALLENGES IN INCREASING PERF/VALUE

Energy and Power Dissipation





DATA MOVEMENT DOMINATES ENERGY USAGE



From Dan McMorrow, Technical Challenges of Exascale Computing, JSR-12-310, JASON, MITRE Corporation, April 2013.

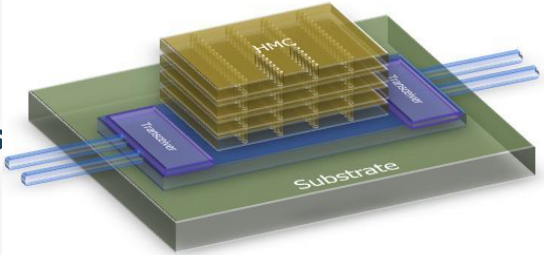
MEMORY FUTURES

- **Scaling continues**
- **Latency will not improve significantly**
- **BW continues to increase with evolving interfaces**
- **Refresh mitigated using ECC and potentially other management**

Ideally, the memory controller function splits the memory physics and defines control closer to the memory bits

- **How do we:**
  - Avoid moving un-used data over the bus?
  - More effectively buffer for lower latency?
- **Can some memory intensive compute be done on the memory?**
  - Application specific memory based accelerators
  - General purpose local memory low level computation

Hybrid Memory Cube

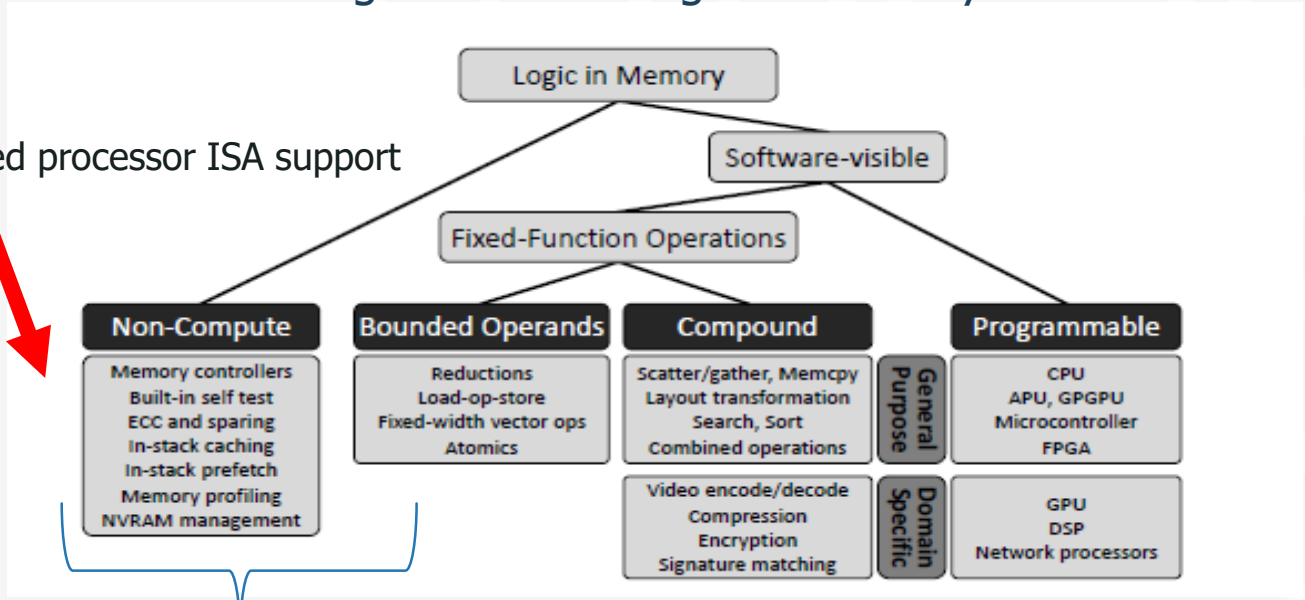


Self Test, Self Repair, Scrubbing, Refresh, Autonomous Functions

PERFORMING MORE PROCESSING IN MEMORY (PIM)
(MORE APPROPRIATELY – LOGIC IN MEMORY) TO LOWER POWER?

Observation: 3D-integration of logic and memory (e.g., HMC) is driving a resurgence in the Logic-in-memory conversation

Doesn't need processor ISA support



Grabiel H Loh, et.al. A Processing –in-Memory Taxonomy and a Case for Studying Fixed-function PIM. *IEEE/ACM International Symposium on Microarchitecture (MICRO-46) 2013.*

CELL PHONES ARE AN INSPIRATION…

- **What can you get for ~ 1 Watt of Power?**
  - **If designed correctly… a lot. For example**

Benchmark	Average System Power (mW)		
	Freerunner	G1	N1
Suspend	103.2	26.6	24.9
Idle	333.7	161.2	333.9
Phone call	1135.4	822.4	746.8
Email (cell)	690.7	599.4	-
Email (WiFi)	505.6	349.2	-
Web (cell)	500.0	430.4	538.0
Web (WiFi)	430.4	270.6	412.2
Network (cell)	929.7	1016.4	825.9
Network (WiFi)	1053.7	1355.8	884.1
Video	558.8	568.3	526.3
Audio	419.0	459.7	322.4

Table 9: Freerunner, G1 and N1 system power (excluding backlight) for a number of micro- and macro-benchmarks.

	G1	N1
SoC	Qualcomm MSM7201	Qualcomm QSD 8250
CPU	ARM 11 @ 528 MHz	ARMv7 @ 1 GHz
RAM	192 MiB	512 MiB
Display	3.2" TFT, 320x480	3.7" OLED, 480x800
Radio	UMTS+HSPA	UMTS+HSPA
OS	Android 1.6	Android 2.1
Kernel	Linux 2.6.29	Linux 2.6.29

Table 6: G1 and Nexus One specifications.

The "compute module" could be an SoC or… a bunch of die on a package…

HYPOTHETICALLY, WHAT IF WE HAD 1 MILLION OF THESE DEVICES?

- **Assumptions per Compute Molecule**
  - 2-3 Watts of Power (Proc Core, Specialized Compute elements (e.g. AI/ML inference engines), compute assist in memory, etc.)

  Goal:
  - 256G DP-Flops/1T – 16-bit Flops ((8-64bit Mul + 8-64 bit Add)*16 cores*1GHz) OR
  - ~65 ResNet-50 Inferences/sec (assuming 50% of Peak Efficiency)
  - 32 GBytes of DRAM
- **Deployed Configuration**
  - 256 Peta DP-FLOPS/Peak
  - ~65M ResNet-50 inferences/s.
  - 32 PBytes RAM
  - 2-3MW of Power
- **Greater resiliency to faults.**
- **The software lift and programming models need to be comprehended.**

We have the technology – we need to re-design/arrange the pieces of the puzzle

MICRON - COMPUTE NEAR MEMORY DEVICE

INNOVATIVE CLAIMS

- Improving application performance by 10-100x without relying on Moore's law is a challenging problem that requires new architectural approaches and integration solutions.
- Micron proposes a memory-centric compute architecture to provide a high-performance, energy efficient solution for a broad range of applications.

TECHNICAL APPROACH

- Define a scalable compute-near-memory architecture that efficiently uses novel compute elements and memory bandwidth to solve a broad range of problems.
- This program implements the architecture using a chiplet approach
  with 2.5D integration, validating performance and energy metrics.

ERI ELECTRONICS RESURGENCE INITIATIVE

SUMMIT

2018 | SAN FRANCISCO, CA | **JULY 23-25**

NARESH SHANBHAG

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

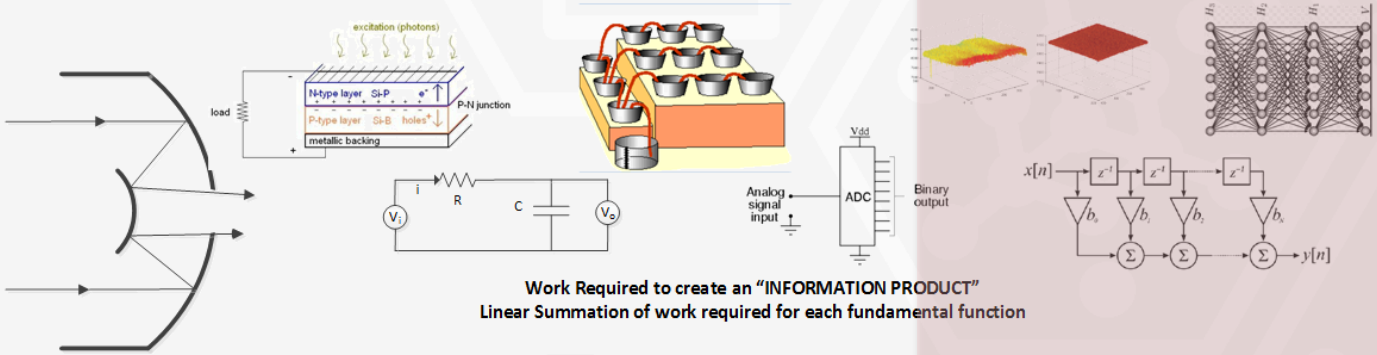
THERMAL LIMITS IN AUTONOMOUS PLATFORMS

High Power Signal Processing Stages

SENSING FUNCTIONS (EO Example)

Radiation Collection	Focusing	Conversion: Photoelectrons	Photoelectron Collection (integration)	Analog Voltage Signal Generated	Analog Voltage to Digital Signal Conversion	Pre-possessing (NUC)	Filtering (FIR)	Signal Processing

→ **INFORMATION PRODUCT**



Work Required to create an "INFORMATION PRODUCT"
Linear Summation of work required for each fundamental function

$$W_{sys}^{t} = \sum W_n \geq W_o$$

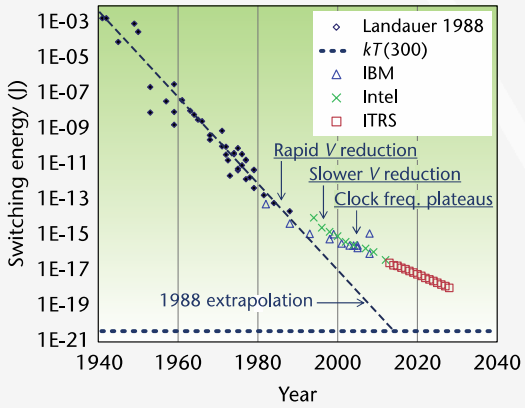
W_o – theoretical minimum limit of work necessary to produce "INFORMATION PRODUCT"

Current Technology Prevents Achievability of Lower Bounds on System-level Work W_{sys}^{t} along with Rad Hard Requirements

Thermal Compensation limits

TRADITIONAL SOLUTIONS RUNNING OUT OF STEAM

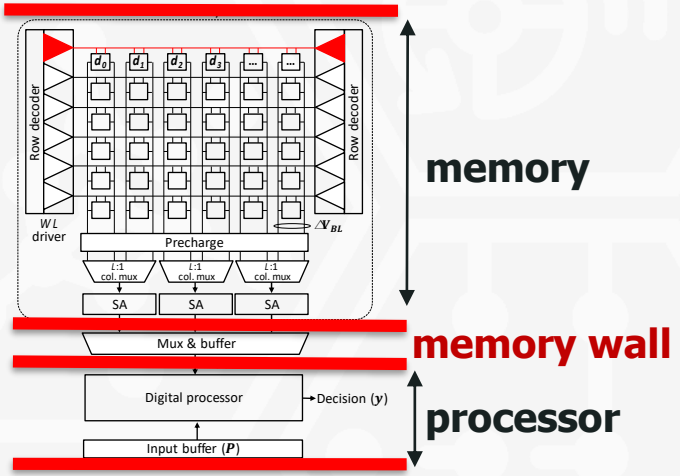
Moore's Law (slowing down)



[Wong, et al., "CMOS Technology Scaling Trend," https://nano.stanford.edu/cmos-technology-scaling-trend, 2017]

- **energy efficiency** gains stagnating
- increased **variability**;
- increased **susceptibility to SEU**

von Neumann Architecture (the Memory Wall)

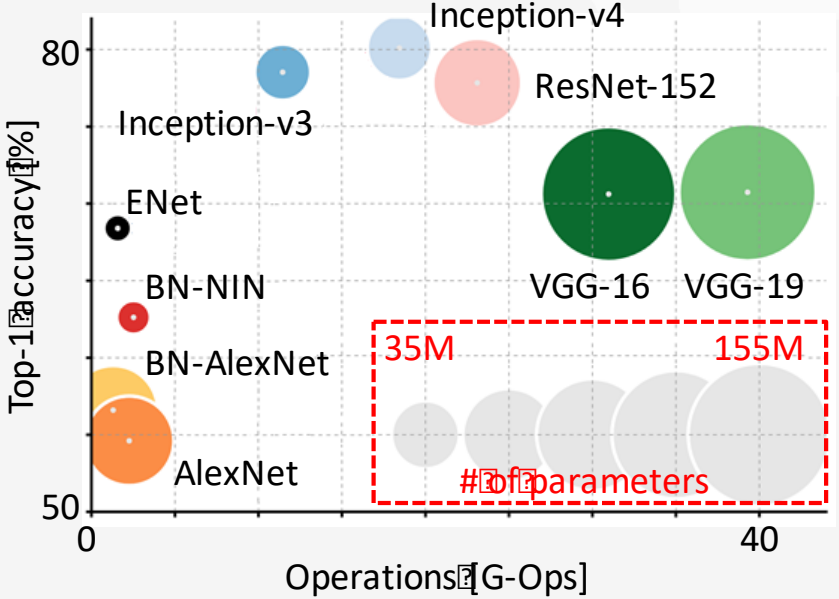


- **memory access costs** dominate in data-centric DoD workloads
- **deterministic roots** limits device options

THE VON NEUMANN ARCHITECTURE'S MEMORY WALL PROBLEM

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

[*Canziani, Arxiv16*]



[*Horowitz, ISSCC'14*]

Computation energy (45nm)

Integer	ADD	Mult
8 bits	0.03 pJ	0.2 pJ
32 bits	0.1 pJ	3 pJ

Memory access energy (45nm)

Memory	64 bits
Cache 8 KB	10 pJ
Cache 32 KB	20 pJ
Cache 1 MB	100 pJ
DRAM	1.2 – 2.6 nJ

OUR FRANC PROJECT
MRAM-BASED DEEP IN-MEMORY ARCHITECTURE (DIMA)

Mission: to realize **> 200X in EDP gains** in DoD workloads

Method: by leveraging MRAM-based DIMA within Shannon-inspired Statistical Computing

Raytheon Missile Systems

DoD System Benchmarks & Specifications

UIUC & Princeton

Shannon-inspired Statistical Compute Models ③

Deep In-memory Architecture (DIMA) and Circuits ②

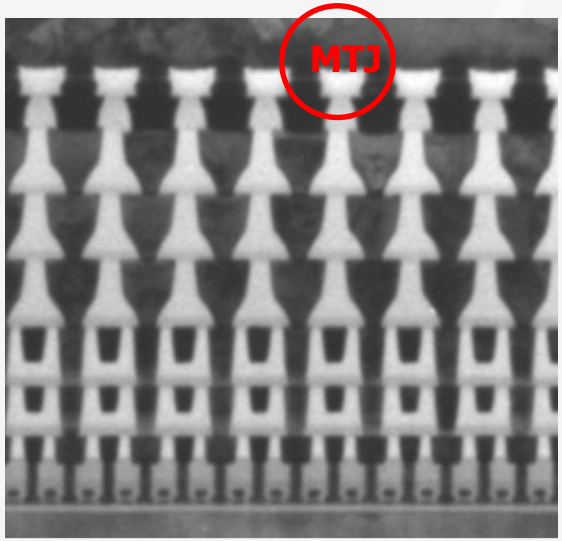
MRAM device + MPW ①

Raytheon Missile Systems

System Test & Validation

GLOBALFOUNDRIES

SOLUTION PART 1: GLOBALFOUNDRIES'S 22-NM FD-SOI E-MRAM

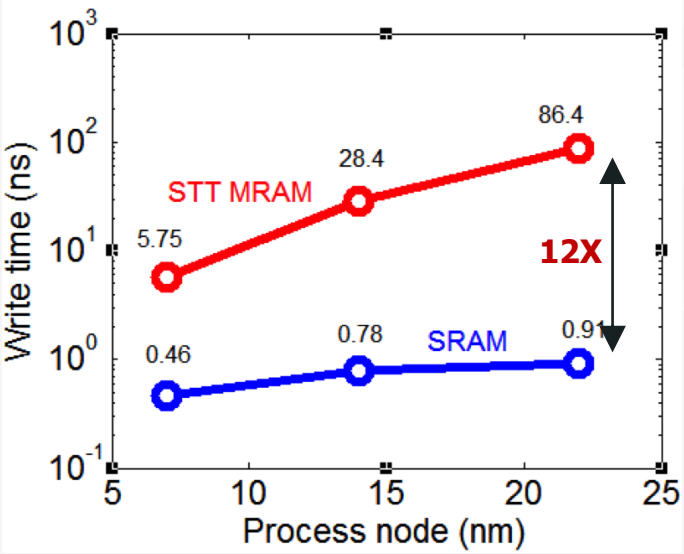


Modified from source: W. Hao, GLOBALFOUNDRIES, Shanghai MRAM Workshop 2018

- BEOL integrated with 22nm FD-SOI
- perpendicular STT MRAM
- bit endurance 100,000 cycles
- > 10 year data retention at 105 ^{0}C

- **The MRAM Advantage** (over 22nm SRAM):
  - $\sim 46 \times$ higher storage density
  - $\sim 13 \times$ lower power consumption
  - $\sim 6 \times$ lower standby current
  - intrinsically robust to SEU

BUT……MRAM ACCESSES LIMITED BY ENERGY-DELAY-ERROR RATE TRADE-OFF



Power-Performance-Area Benchmarking of STT MRAM for Server Cache Applications, N. Sharma, **A. P. Jacob**, G. Gomba, IEEE NE Technology forum (Aug'2016)

- ~12 × slower write speed
- fundamentally stochastic R/W behavior

$$\epsilon(E, T_g) \approx e^{-\sqrt{ET_g}}$$

$$E(i, T_g) = i^2 I_{crit}^2 R T_g$$

- mismatched to the von Neumann architecture:
  - aggravates the **memory wall problem**
  - high **cost of determinism**

SOLUTION PART 2: THE DEEP IN-MEMORY ARCHITECTURE (DIMA)

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**

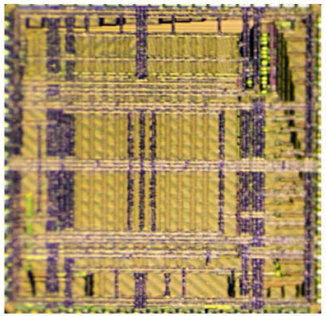


Image: Sujan Gonugondla

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)

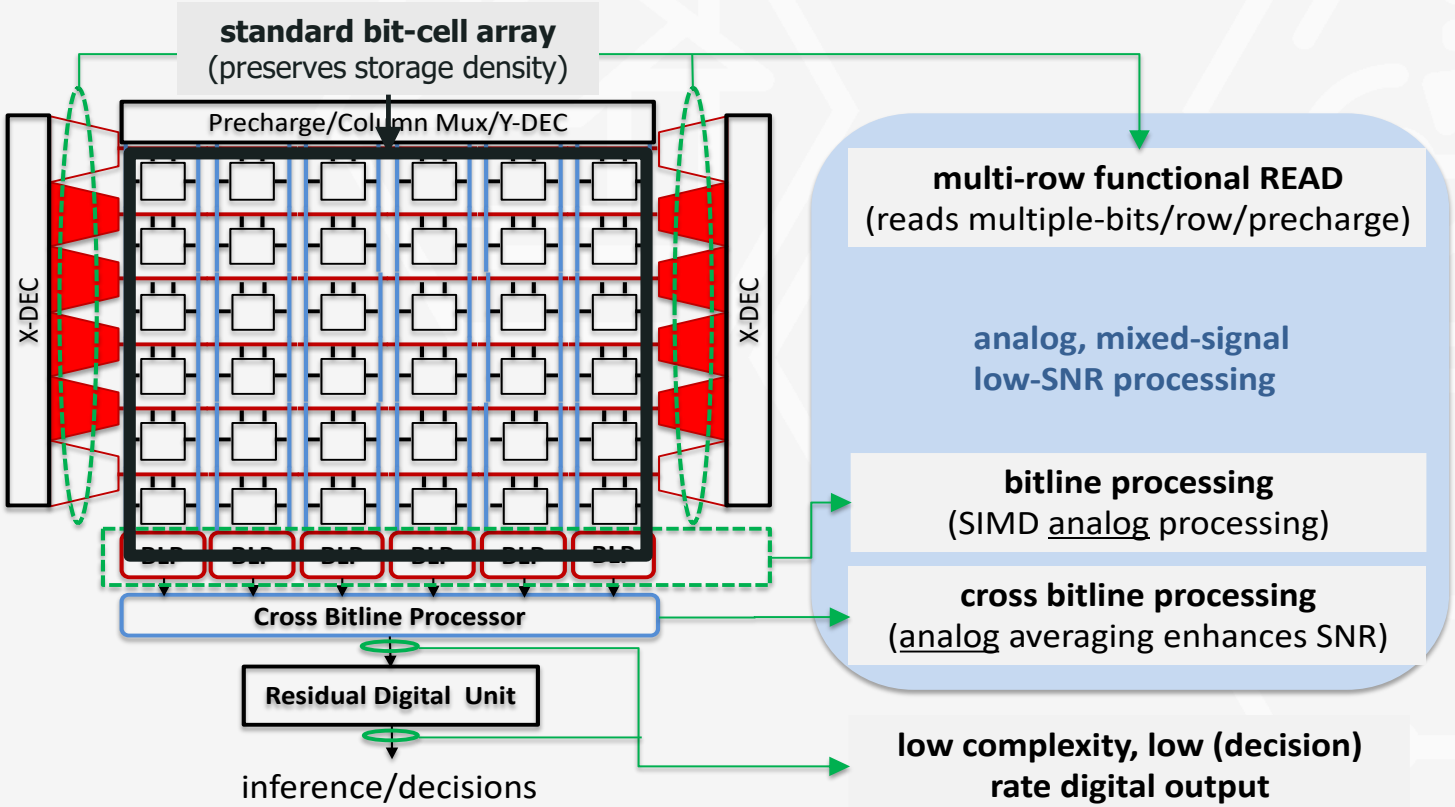
"tearing down the memory wall"

[Verma, Shanbhag, STARnet SONIC]

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

THE DEEP IN-MEMORY ARCHITECTURE (DIMA)

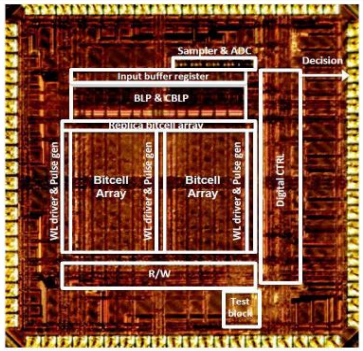
read **functions of data** (never the **raw data**)



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

SRAM DIMA PROTOTYPES

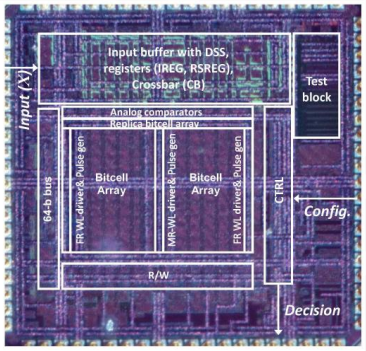
53× EDP ↓



Multi-functional inference processor (65nm CMOS)

[UIUC: *Kang, JSSC'18*]

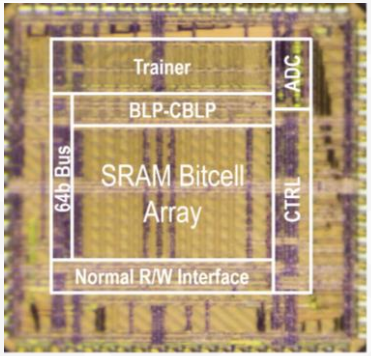
7× EDP ↓



Random forest processor (65nm CMOS)

[UIUC: *Kang, JSSC'18, ESSCIRC'17*]

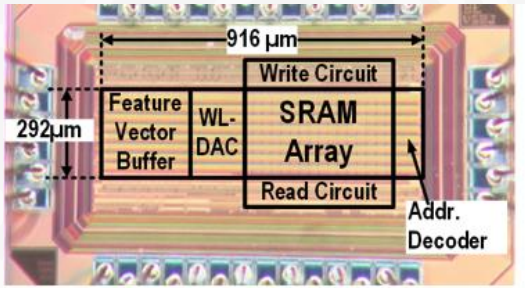
100× EDP ↓



On-chip training processor (65nm CMOS)

[UIUC: *Gonugondla, ISSCC'18*]

175× EDP ↓



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

[Princeton: *Zhang, JSSC'17, VLSI'16*]

←——256 column-parallel 8 bit-parallel compute——→

SOLUTION PART 3: SHANNON-INSPIRED STATISTICAL COMPUTING

29 Oct 2010 | 19:32 GMT

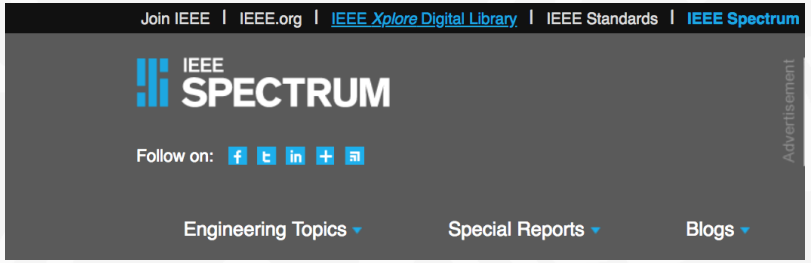
The Era of Error-Tolerant Computing

Errors will abound in future processors...and that's okay

By **David Lammers**

The computer's perfectionist streak is coming to an end. Speaking at the International Symposium on Low Power electronics and Design, experts said power consumption concerns are driving computing toward a design philosophy in which errors are either allowed to happen and ignored, or corrected only where necessary. Probabilistic outcomes will replace the deterministic form of data processing that has prevailed for the last half century.

Naresh Shanbhag, a professor in the department of electrical and computer engineering at the University of Illinois at Urbana-Champaign, refers to error-resilient computing (also called probabilistic computing) by the more formal name of stochastic processing. Whatever the name, the approach, Shanbhag says, is not to automatically circle back and correct errors once they are identified, because that consumes power. "If the application is such that small errors can be tolerated, we let them happen," he says. "Depending on the application, we keep error rates under a threshold, using algorithmic or circuit techniques." For many applications such as graphics processing or



"reliable computing with unreliable components"

[Shanbhag, Verma, Varshney, Grover, STARnet SONIC faculty]

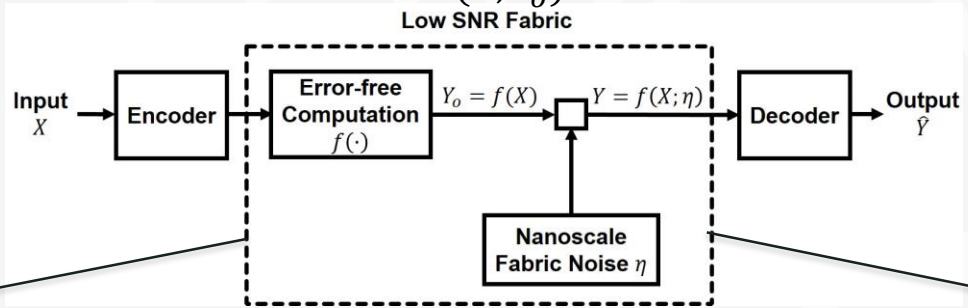
https://spectrum.ieee.org/semiconductors/processors/the-era-of-errortolerant-computing

SHANNON-INSPIRED STATISTICAL MODEL OF COMPUTING

computation as *information flow* across *low-SNR* nanoscale fabrics

information-based design metrics - **mutual information** (MI)

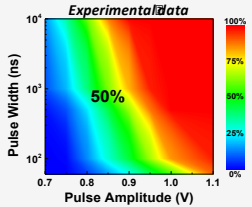
$$I(\hat{Y}; Y_o)$$



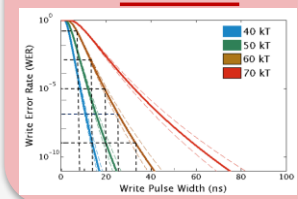
low-SNR nanoscale fabrics

stochastic nanofunctions ***deep in-situ architectures***

RRAM

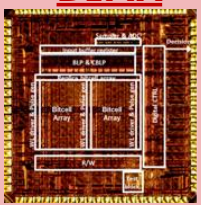


MRAM

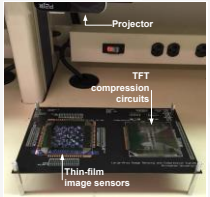


+

DIMA

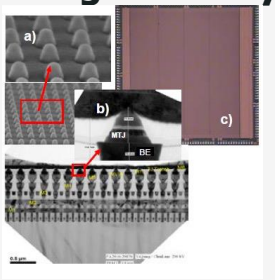


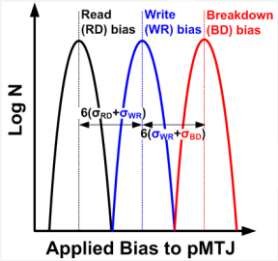
DISA



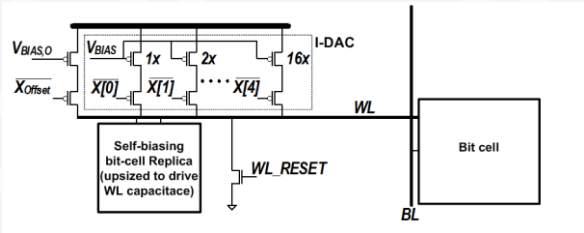
EXPECTED PROJECT OUTCOMES

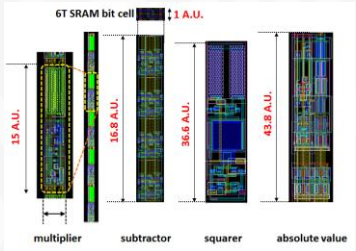
DIMA-driven MRAM Device Optimization for High-density & Low-energy



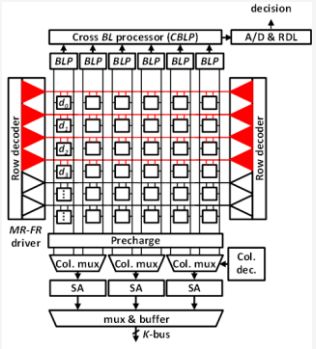


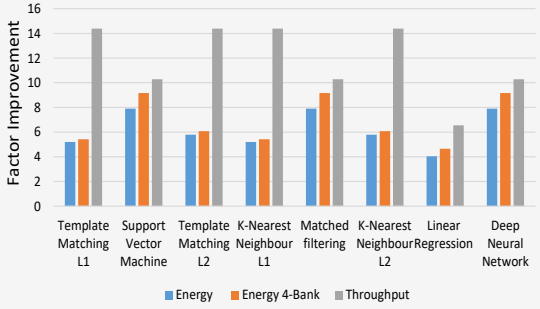
Parameterized DIMAs for Platform-design Tools



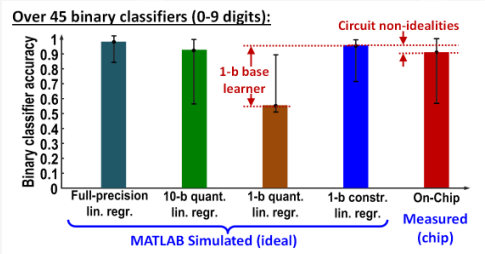


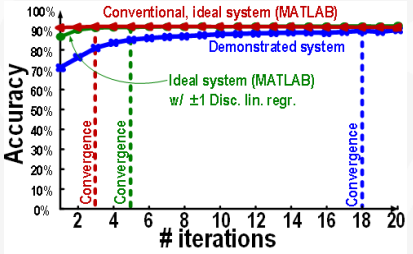
Multi-functioned DIMAs for Broad Application-level Usage





Statistical Compute Models for Aggressive Scaling





SUMMARY

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

Raytheon Missile Systems

DoD System Benchmarks & Specifications

UIUC & Princeton

Shannon-inspired Statistical Compute Models ③

Deep In-memory Architecture (DIMA) and Circuits ②

MRAM device + MPW ①

GLOBALFOUNDRIES

Raytheon Missile Systems

System Test & Validation

ERI ELECTRONICS RESURGENCE INITIATIVE

SUMMIT

2018 | SAN FRANCISCO, CA | **JULY 23-25**