ISCAS 2020
Virtual, October 10-21
IEEE INTERNATIONAL SYMPOSIUM ON CIRCUITS AND SYSTEMS
CAS

A Variation Robust Inference Engine Based on STT-MRAM with Parallel Read-Out

Yandong Luo[1], Xiaochen Peng[1], Ryan Hatcher[2], Titash Rakshit[2], Jorge Kittl[2], Mark S Rodder[2] , Jae-sun Seo[3] and Shimeng Yu[1]

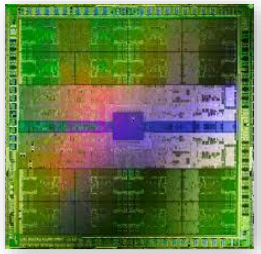
[1]Georgia Institute of Technology, Atlanta, GA 30332, USA,
[2]Samsung Semiconductor Inc., Austin, TX 78754, USA,
[3]Arizona State University, Tempe, AZ 85281, USA

Presentation Outline

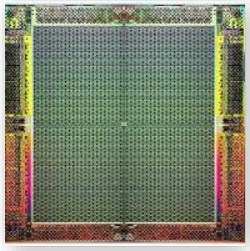
1. Background introduction
2. Challenges using STT-MRAM for computing
3. Variation-robust design strategies
4. Evaluation results and discussions
5. Conclusion

1. Background introduction

- **Compute-in-memory (CIM) achieves good energy efficiency for deep neural network (DNN) inference**

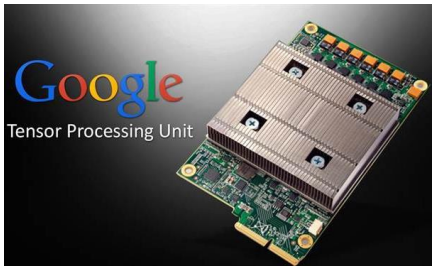


GPU



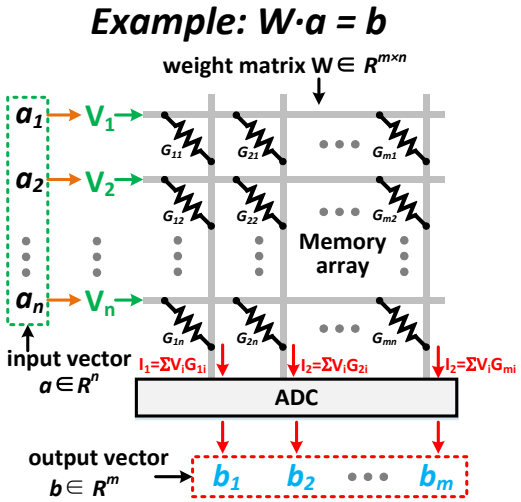
FPGA

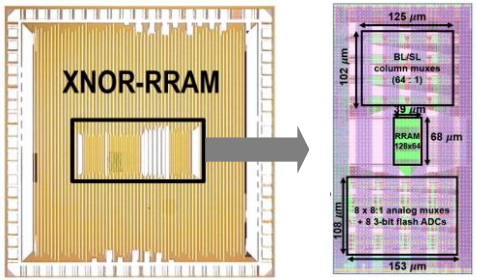
Conventional computing platforms
~ 0.1 TOPS/W



Google TPU

CMOS ASICs
~ 0.1-1TOPS/W





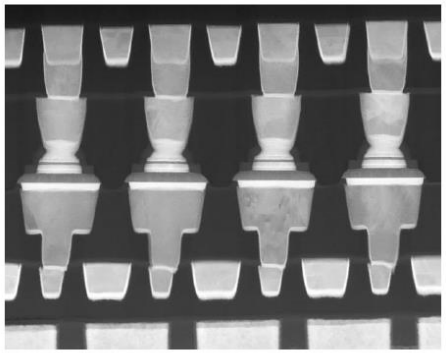
Compute-in-memory

Beyond CMOS (eNVMs)
~ 1-100 TOPS/W

Fig. 1 The inference energy efficiency for different computing platforms

1. Background introduction

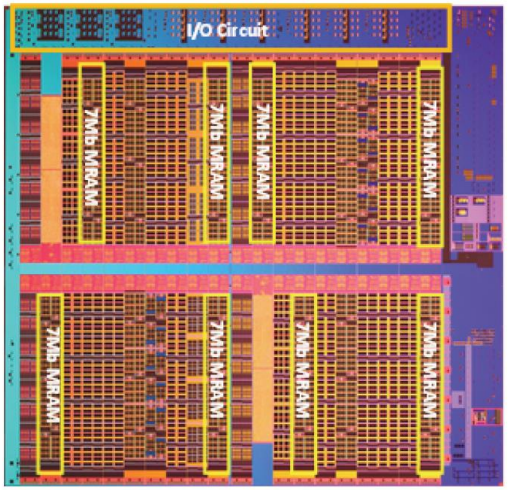
- **STT-MRAM Technology**
  - Magnetic tunneling junction (MTJ) with binary state (parallel and anti-parallel)
  - Low write voltage (~1V), good data retention and write endurance
  - Foundries provide 22nm tech. node process (e.g. Intel, TSMC, Samsung, GlobalFoundries et.al)



O. Golonzka, IEDM, 2018

(a)

Technology	22FFL FinFET Technology
Memory	Perpendicular STT-MRAM
TMR	>180%@25C
Cell type	1T1MTJ
Cell size	$0.0486\mu m^2$
Capacity	7Mb
Subarray Density (Incl. ECC bits)	10.6 $Mbits/mm^2$
Read Sense Time	4ns@0.9V, 8ns@0.6V
Bit Yield	> 99.997%
Retention	200C 10 years
Write Endurance	>1E06
READ Disturb	>1E12
Temp Range	-40C to 105C



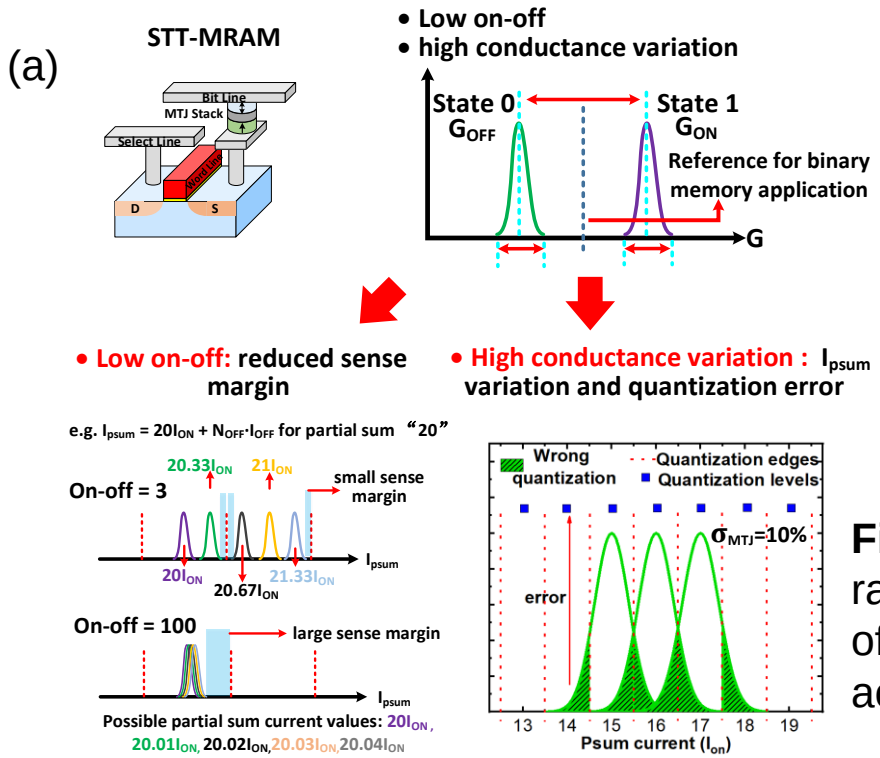
(b)

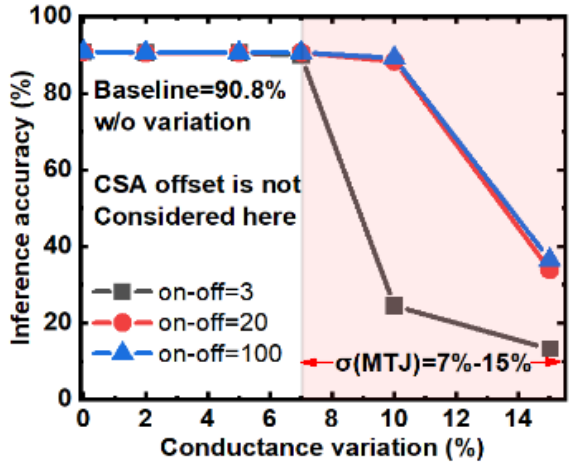
L. Wei et.al, ISSCC, 2019

Fig. 2 (a) The TEM image of the MTJ array (Intel's 22FFL process) (b) The die photo and performance metrics of a STT-MRAM chip macro

2. Challenges using STT-MRAM for computing

- **MTJ conductance variation (σ=7%~15%)**
  - ➢ partial sum current variation
- **Low on-off ratio (<3)**
  - ➢ I_{off} is not negligible and may leads to wrong partial sum



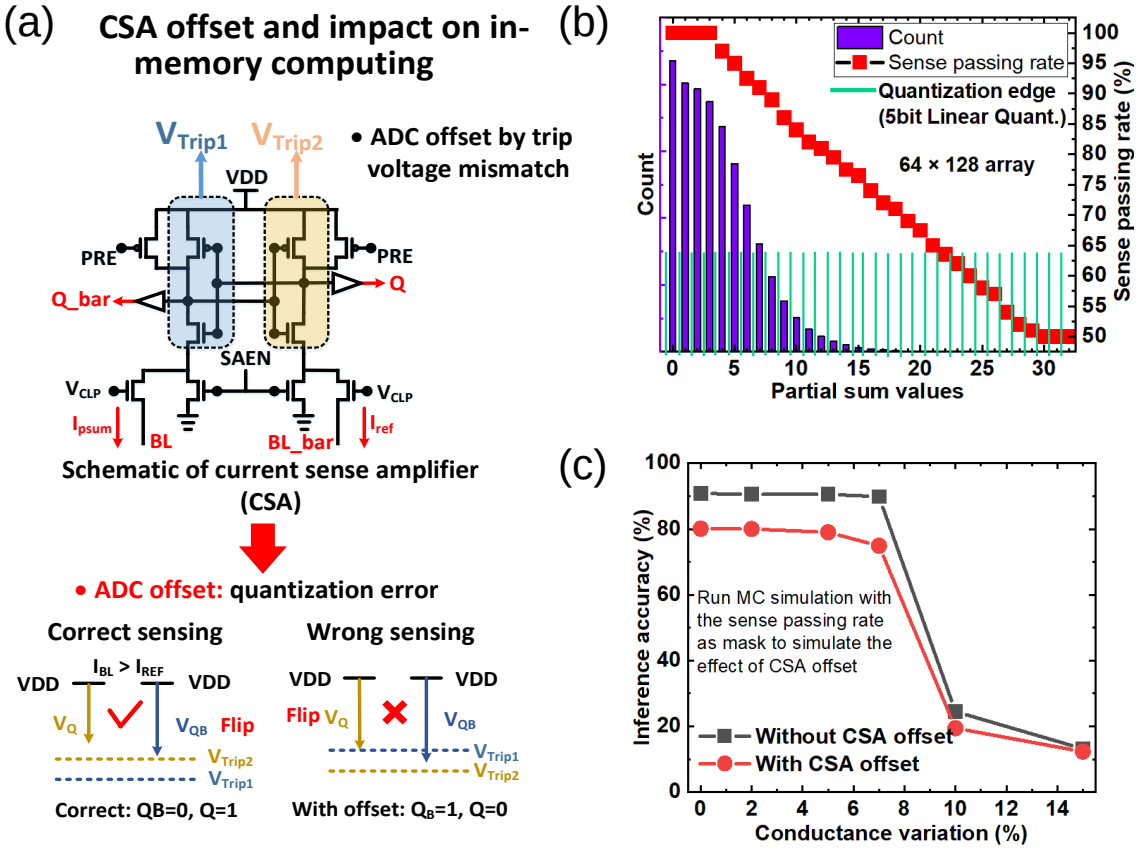


7-layer CNN CIFAR-10 dataset

Fig. 3 (a) An illustration of the impact of low on-off ratio and device conductance variation (b) The impact of device conductance variation on inference accuracy with different on-off ratio

2. Challenges using STT-MRAM for computing

- **Low R_{ON}** (<20KΩ, can be as low as a few kΩ)
- **Sense amplifier (or ADC) offset** : due to the large current to sense
  - ➢ Quantize the partial sum current to a wrong digital level

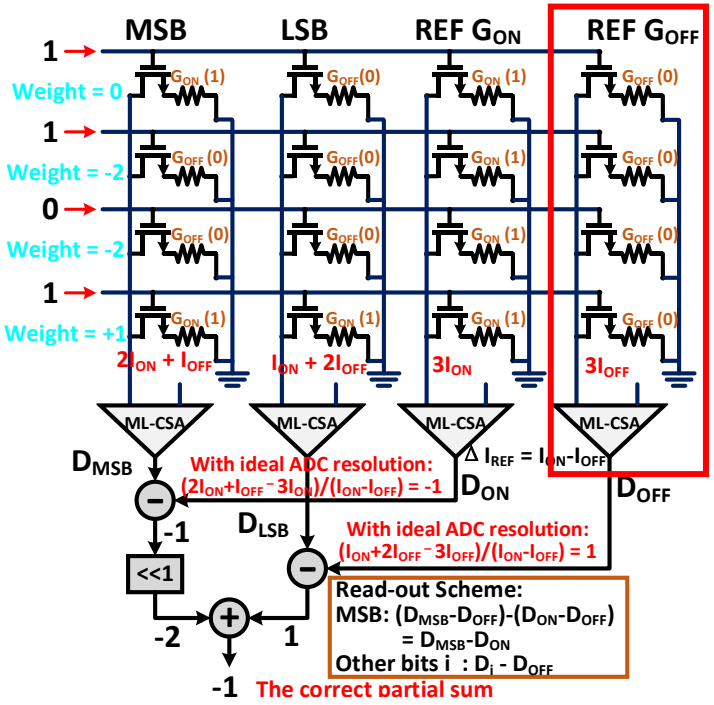


Sense passing rate: the probability that a partial sum current is correctly quantized.

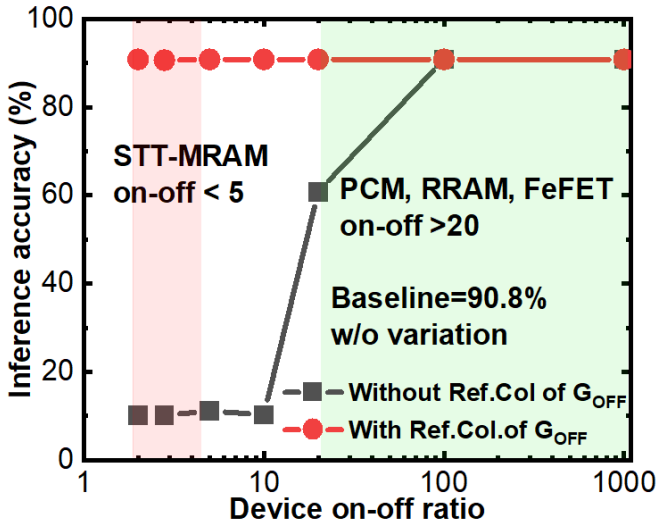
Fig. 3 (a) An illustration of the current sense amplifier (CSA) offset (b) The sensing passing rate of CSA for a 64 x 128 array for in-memory compute. It was simulated using 28nm PDK. (c) The impact of CSA offset on the inference accuracy

3. Variation-robust design strategies

- **A differential sensing scheme to compensate low on-off ratio**
  - Subtract the partial sum value by that from a dummy column with all G_{OFF}
  - Achieves good accuracy without considering the device variation



(a)

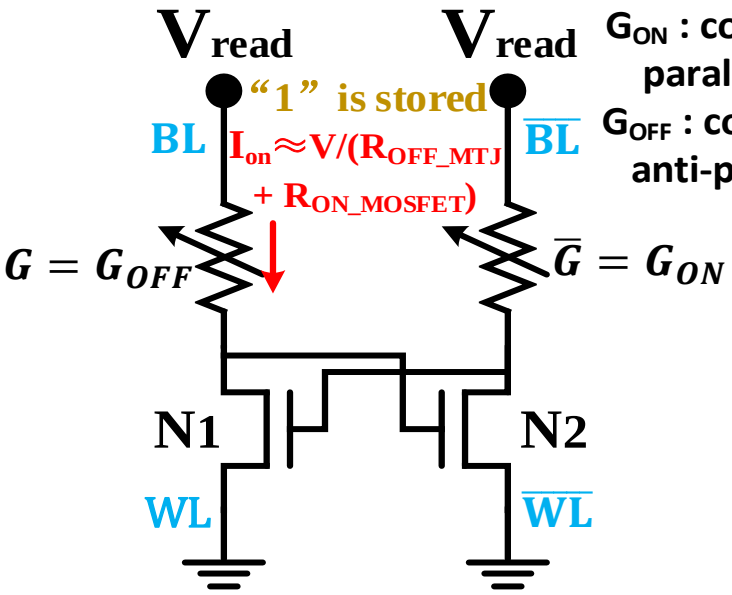


(b)

Fig. 4 (a) Parallel read-out scheme to represent negative weight and eliminate the effect of low on/off ratio. (b) The inference accuracy with the reference column

3. Variation-robust design strategies

- **2T-2MTJ bit cell design to increase the on-off ratio**
  - Cross-coupled MTJ
  - $R_{ON} = R_{OFF\_MTJ} + R_{ON\_MOSFET}$



G_{ON} : conductance at parallel (P) state

G_{OFF} : conductance at anti-parallel state

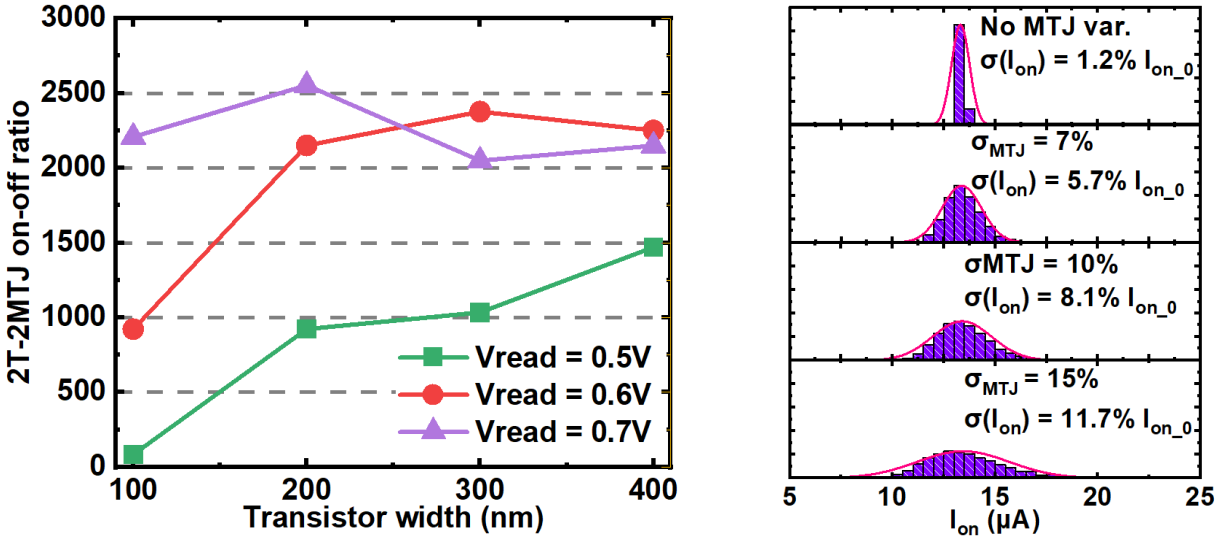
TABLE I. The binary weight representation of the 2T-2MTJ cell

Weight	G	$\bar{G}$
0	G_{ON}	G_{OFF}
1	G_{OFF}	G_{ON}

Fig. 5 Schematic of proposed 2T-2MTJ bit-cell with cross-coupled MTJs

3. Variation-robust design strategies

- **2T-2MTJ bit cell simulation**
  - 28nm foundry PDK
  - > 1000 On-off if V_{read} > 0.6V
  - Simulated the variation of the cell considering the MTJ variation and transistor variation



(a) (b)

Fig. 6 (a) Simulated on/off ratio for the 2T-2MTJ bit cell design (b) The Monte Carlo simulation results of the I_{ON} distribution for the proposed 2T-2MTJ bit cell.

3. Variation-robust design strategies

- **2T-2MTJ cell: array level operation**
  - Parallel partial sum read: similar as regular 1T1R cell
  - G_{OFF} to G_{ON} state: column-by-column programming
  - G_{ON} to G_{OFF} state: row-by-row programming
  - One-time program for inference

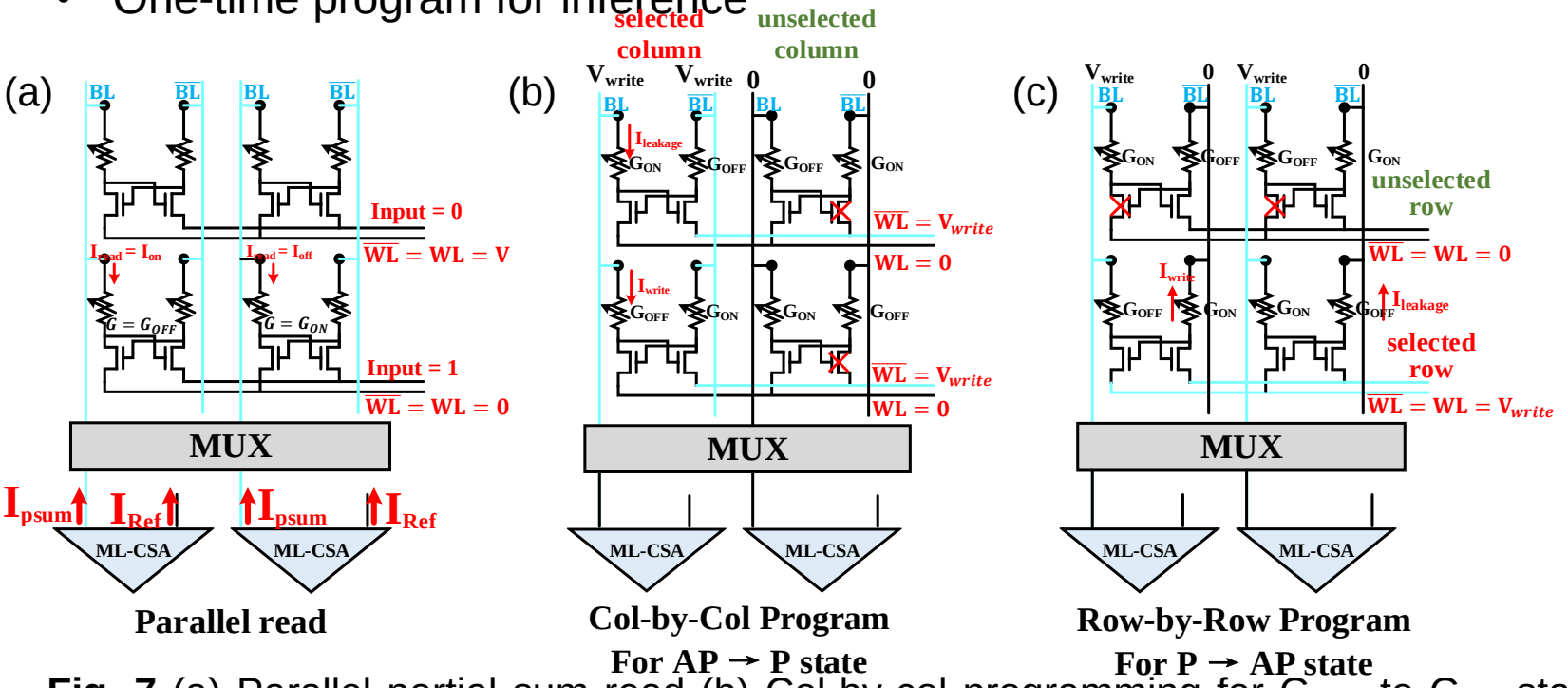


Fig. 7 (a) Parallel partial sum read (b) Col-by-col programming for G_{OFF} to G_{ON} state. (c) Row-by-row programming for G_{ON} to G_{OFF} state

3. Variation-robust design strategies

- **MSB redundancy:** mitigate the impact of conductance variation
  - MSB is more vulnerable due to higher numerical significance
  - Use a redundancy column for MSB
  - Take the average of partial sum of the MSB columns

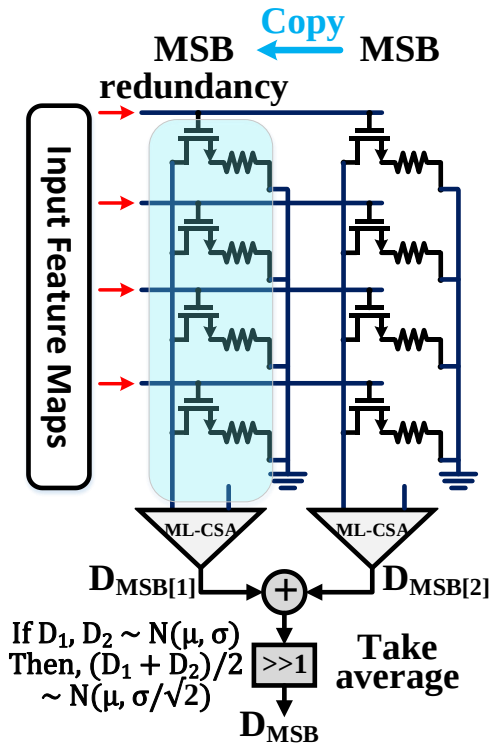


Fig. 8 The MSB redundancy scheme

3. Variation-robust design strategies

- **Partial parallel read: compensate the ADC offset**
  - Conventional mapping: weights of the same layer maps into one array
    - Increased latency
  - Layer-hybrid mapping: weights of two layers share one array
    - Less rows are activated
    - Add up the outputs from two arrays
    - No latency increase

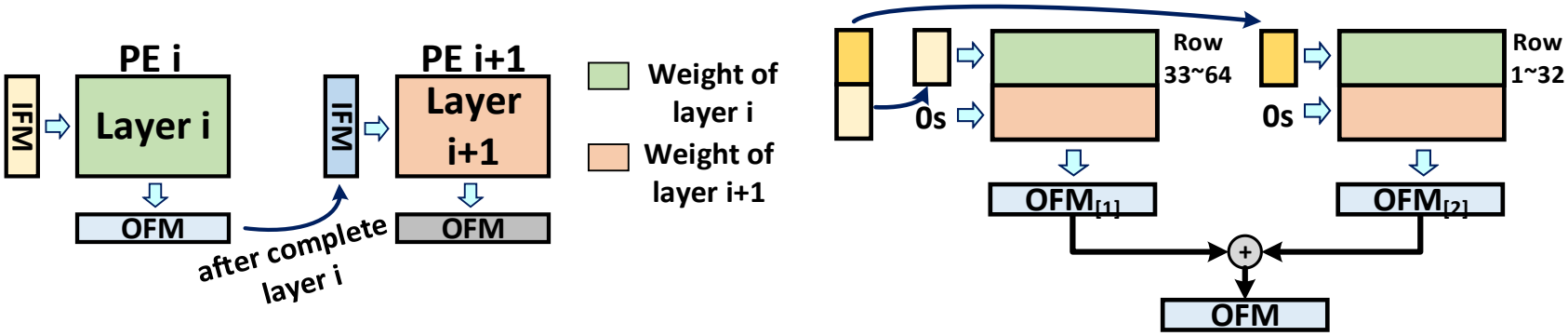


Fig. 9 (a) The conventional mapping scheme that maps the weights of a layer into one PE (b) The proposed layer hybrid mapping scheme to reduce the number of rows read in parallel

4. Evaluation Results and Discussions

- **Methodology for software simulation**
  - Use a 7-layer CNN to evaluate inference accuracy
  - 4-bit weight (4 memory cells)
  - 90.8% software accuracy for CIFAR-10 dataset
  - Incorporate the non-ideal effects in the tensorflow-based simulation platform

- **Methodology for hardware performance estimation**
  - Modify DNN+NeuroSim
  - Assume array size 64 x 128, ADC precision 5-bit
  - 28nm technology node

4. Evaluation Results and Discussions

- **Conductance variation only**
  - 2T-2MTJ design is more robust to 1T-1MTJ design
  - MSB redundancy improves the robustness to conductance variation
- **ADC offset only**
  - Partial parallel read scheme improves the robustness to ADC offset

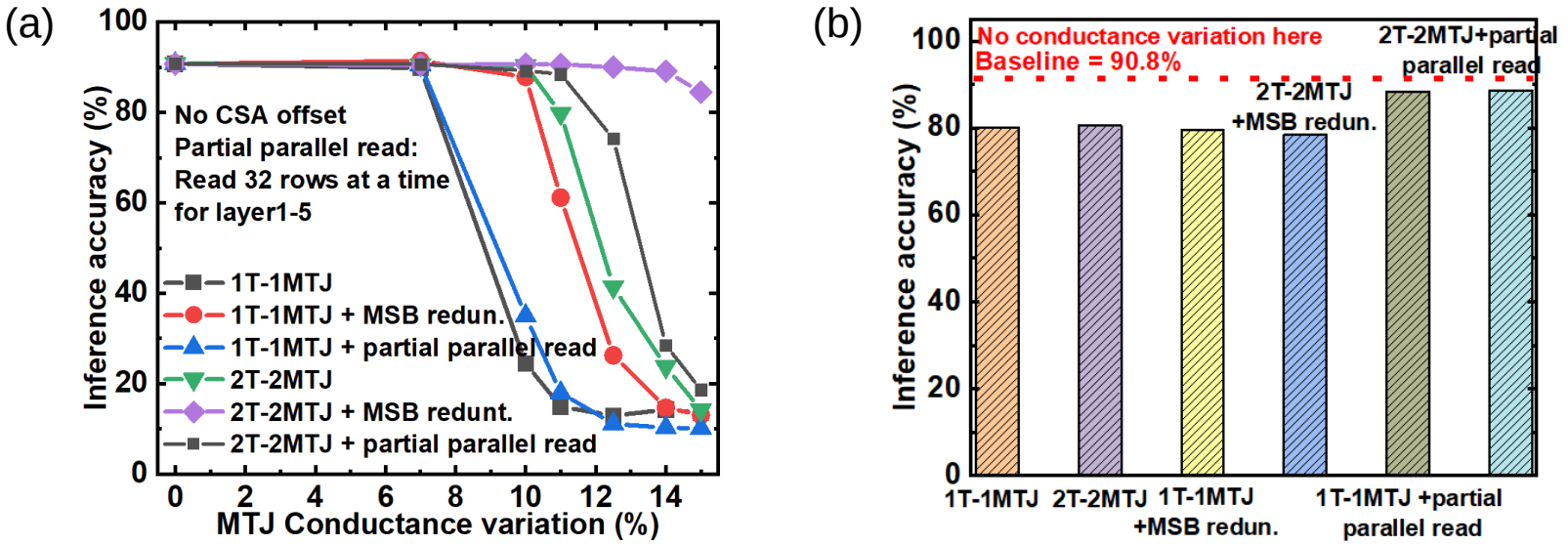


Fig. 10 (a) Inference accuracy vs. MTJ conductance variations for different design schemes. (b) Inference accuracy considering CSA offset.

4. Evaluation Results and Discussions

- **Combine different strategies**
  - Scheme1: 1T-1MTJ cell + MSB redundancy + 32row partial parallel read
    Scheme2: 2T-2MTJ cell + MSB redundancy + 32row partial parallel read
- **Considering both conductance variation and ADC offset**
  - 2T-2MTJ alone is only robust to MTJ conductance variation
  - MSB redundancy and partial parallel read makes it robust to both conductance variation and ADC offset

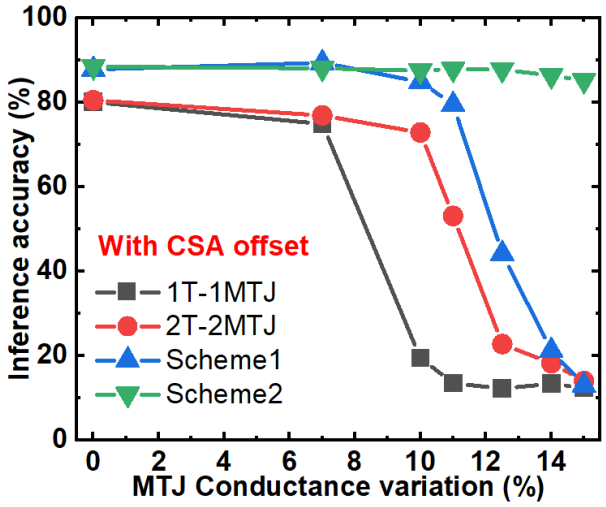


Fig.11 Inference accuracy vs. MTJ conductance variations with CSA offset. Scheme2 shows robustness against conductance variations.

4. Evaluation Results and Discussions

- **System level performance**
  - Low accuracy degradation using Scheme 2 (~87.5%)
  - 8% higher energy and 4% chip area overhead compared with 1T-1MTJ baseline

TABLE II. Estimated chip performance for different design schemes (28nm node)

	1T-1MTJ	2T-2MTJ	Scheme 1	Scheme 2
CIFAR10 Inference accuracy (σ_{MTJ}=10%,w/CSA offset)	**~19.45%**	**~73%**	~85%	**~87.5%**
Chip area (mm^2)	11.65	10.33	**13.68**	12.09
Read Dynamic Energy (layer-by-layer, μJ)	52.09	29.67	**89.09**	56.26
Leakage Energy (μJ)	0.11	0.044	0.130	0.053
Latency (ms)	2.875	1.167	2.881	1.173
Energy efficient (TOPS/W)	2.93	**5.14**	**1.712**	2.71
Throughput (FPS)	347.86	856.625	347.16	852.74

4. Evaluation Results and Discussions

- **Chip area and energy breakdown**
  - 2T-2MTJ design shows smaller chip area and energy consumption
    - Due to the increased R_{ON}
    - Smaller mux area due to the reduced TG size
    - Smaller ADC energy due to the smaller current

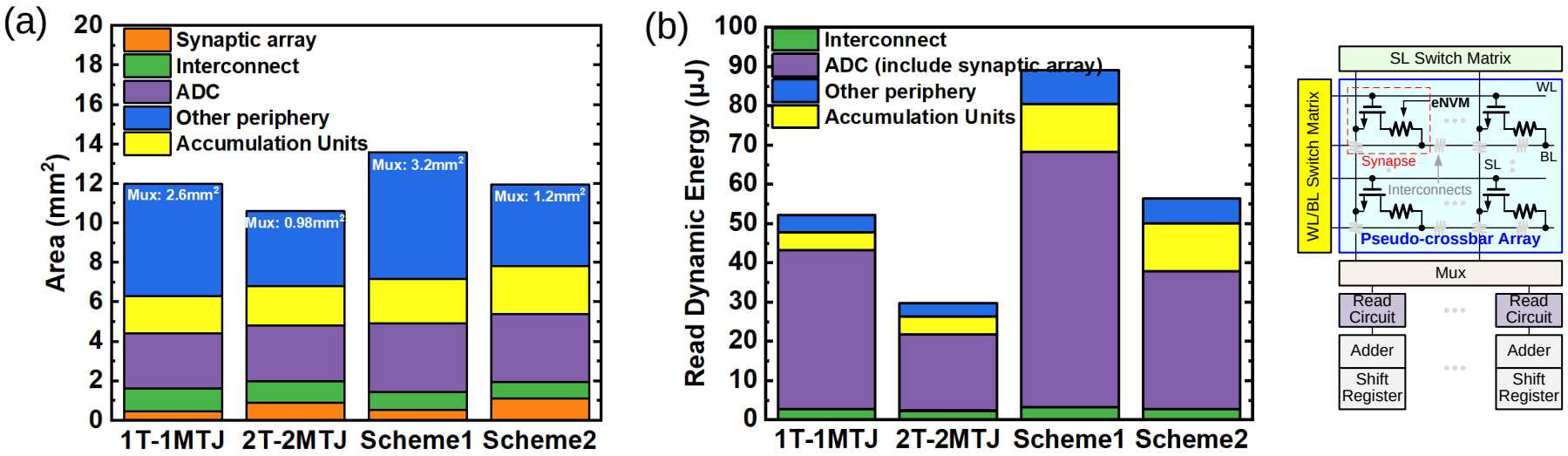


Fig. 12 (a) Chip area and (b) read dynamic energy breakdown for the CIFAR-10 benchmark

5. Conclusions and Acknowledgement

Conclusions

- **Investigate the impact of non-ideal device property on DNN inference for STT-MRAM technology**
- **Proposed design strategies to mitigate the impact**
  - 2T-2MTJ cell design
  - MSB redundancy
  - Hybrid-layer mapping
- **Benchmarked the system level performance**
  - ➢ Maintains high inference accuracy with device variation and ADC offset
  - ➢ 4% area and 8% energy consumption overhead

Acknowledgement

- This work is supported by ASCENT, one of the SRC/DARPA JUMP Centers, and Samsung Electronics.

Thank you for your attention

Questions are welcomed