AmPEC: Approximate MRAM with Partial Error Correction for Fine-grained Energy-quality Trade-off

Lan-yang Sun[1], Yaoru Hou[2], Hao Cai[1]

[1]Southeast University, Nanjing, China

[2]The Hong Kong University of Science and Technology, Hong Kong, China

hao.cai@seu.edu.cn

Outline

1 Background

2 Proposed Approximate Scheme

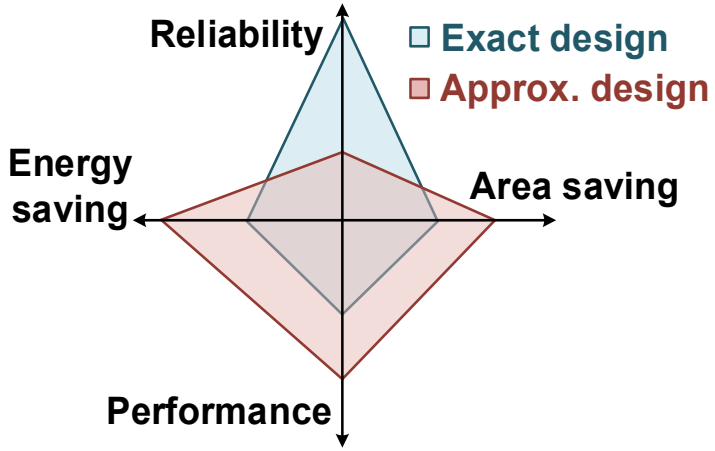
3 Evaluation

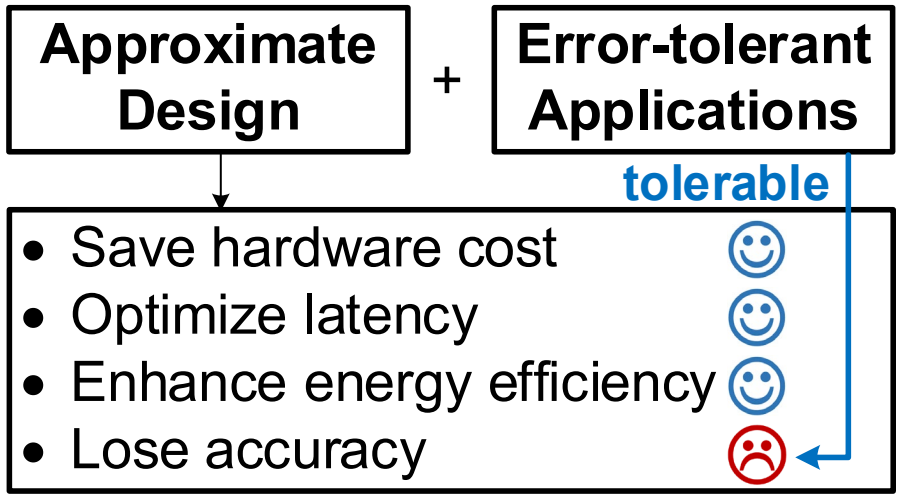
4 Conclusion

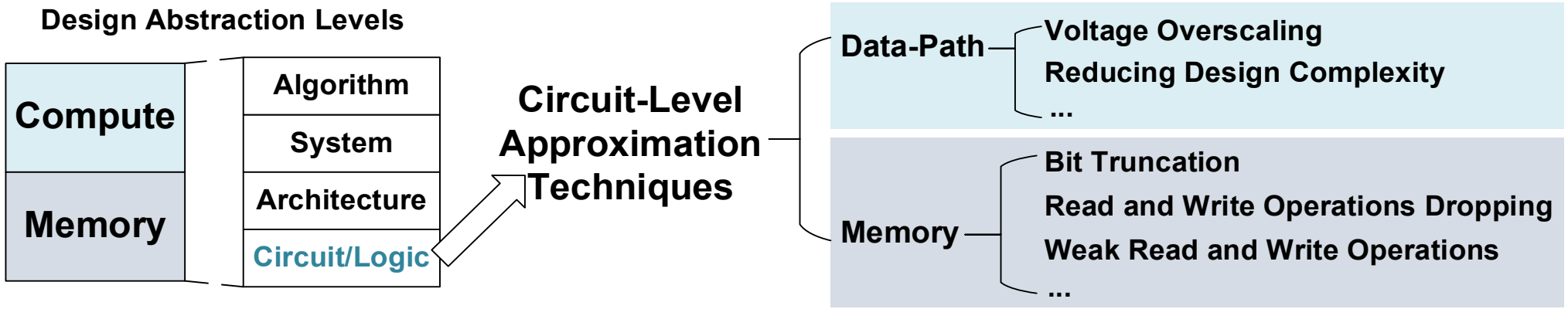
Outline

1 Background

☐ Approximate Design Paradigms







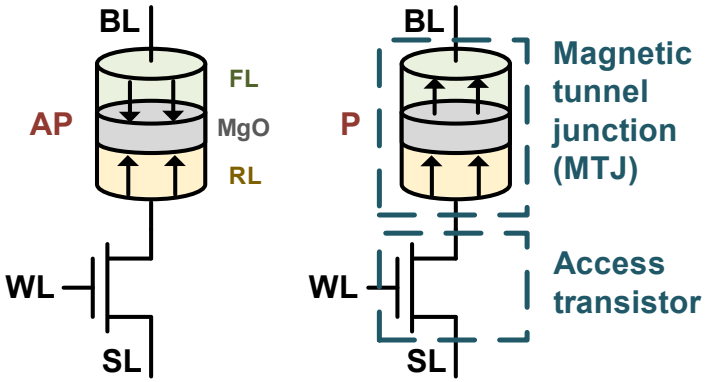
1 Background

☐ Spin Transfer Torque-Magnetic RAM

High current, long access time

→ High energy consumption

→ Approximate design for read/write access of MRAM

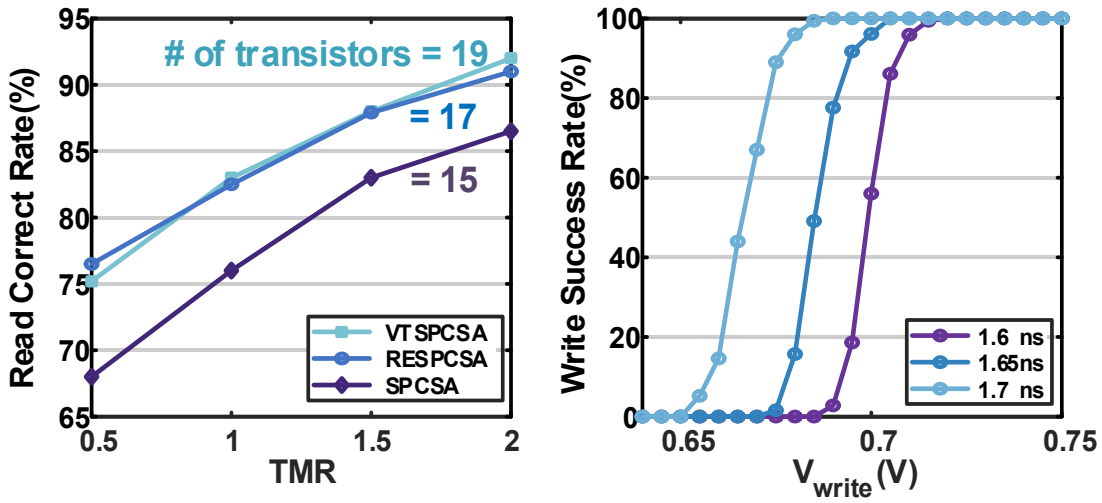


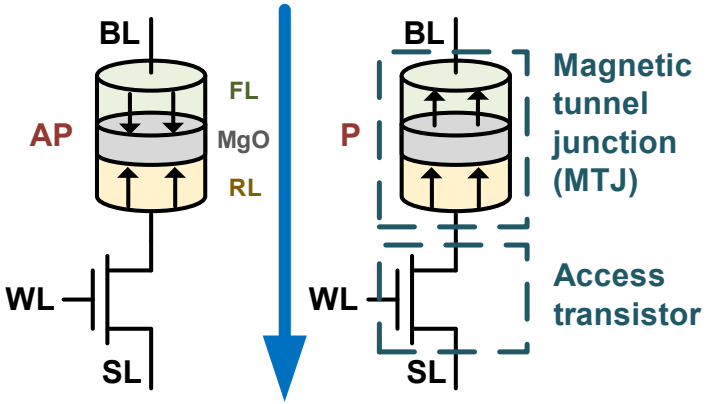
➢ **Characteristics of various types of storage**

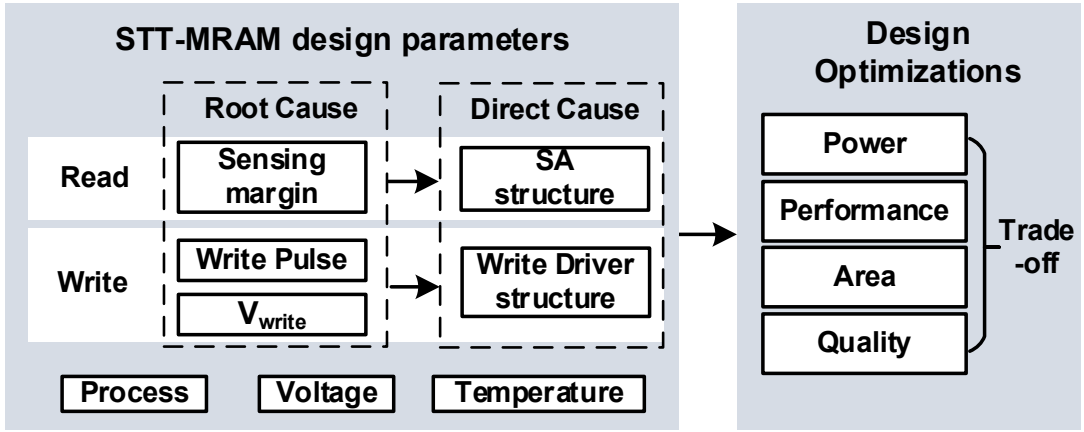
	SRAM	DRAM	Flash	STT-MRAM	PCM	RRAM
Cell size	120-150F^2	10-30F^2	10-30F^2	**10-30F^2**	10-30F^2	10-30F^2
Non-volatility	YES	NO	YES	**YES** ☺	YES	YES
Write voltage	<1V	<1V	~10V	**<1.5V**	<3V	<3V
Write energy	~fJ	~10 fJ	~100 pJ	**~1 pJ** ☹	~10 pJ	~1 pJ
Standby power	HIGH	MEDIUM	LOW	**LOW**	LOW	LOW
Write speed	~1 ns	~10 ns	0.1-1 ms	**~5 ns**	~10 ns	~10 ns
Read speed	~1 ns	~3 ns	~10 ns	**~5 ns**	~10 ns	~10 ns
Endurance	10^16	10^16	10^4-10^6	**10^15**	10^12	10^7

1 Background

STT-MRAM







- Read: Simpler SA structure
- Write: Modify write driver structure

→sacrifice **precision** for lower **power and area** consumption

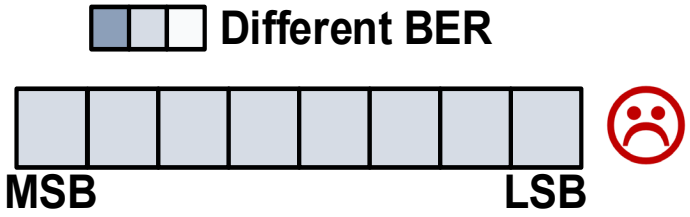
1 Background

Data Mapping

- Different bit positions within a word have varying weights
- Different applications need varying accuracy

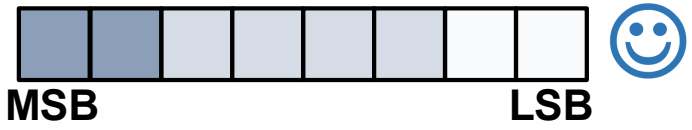
✗ **Uniform approximate approach:**

- Coarse-grained, word-level precision adjustment
- MSBs have greater influence: exponential increase in errors



✓ **Bit-level approximate approach:**

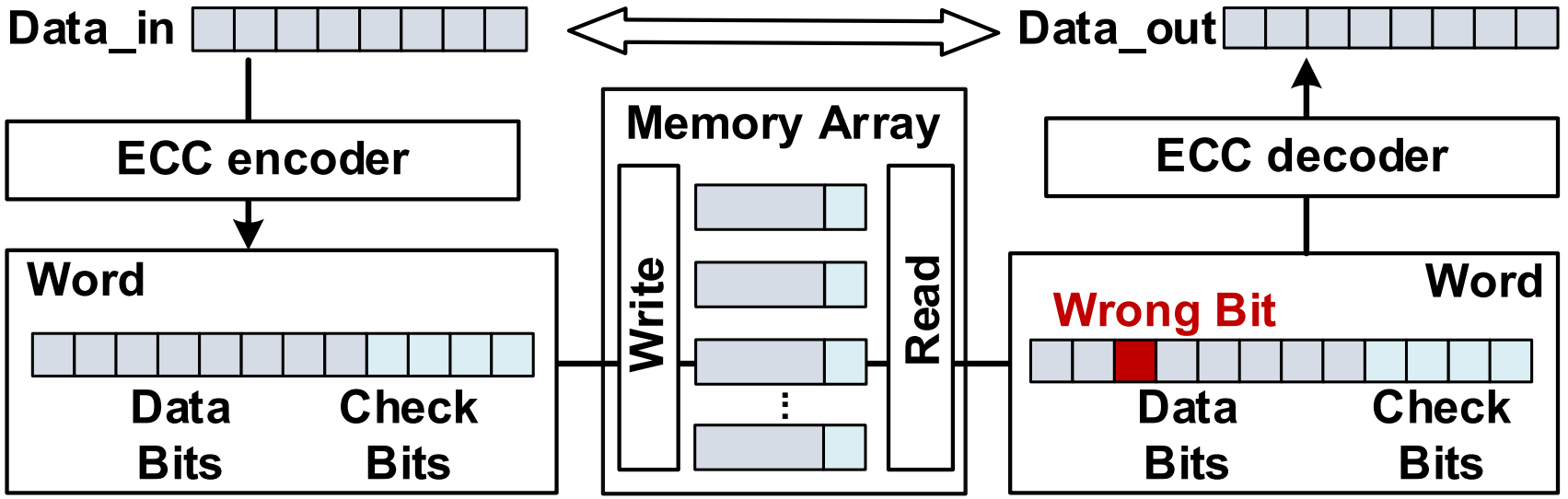
- Fine-grained precision adjustment
- Flexible adjustment based on specific needs of different applications



1 Background

Error Correction Code (ECC)

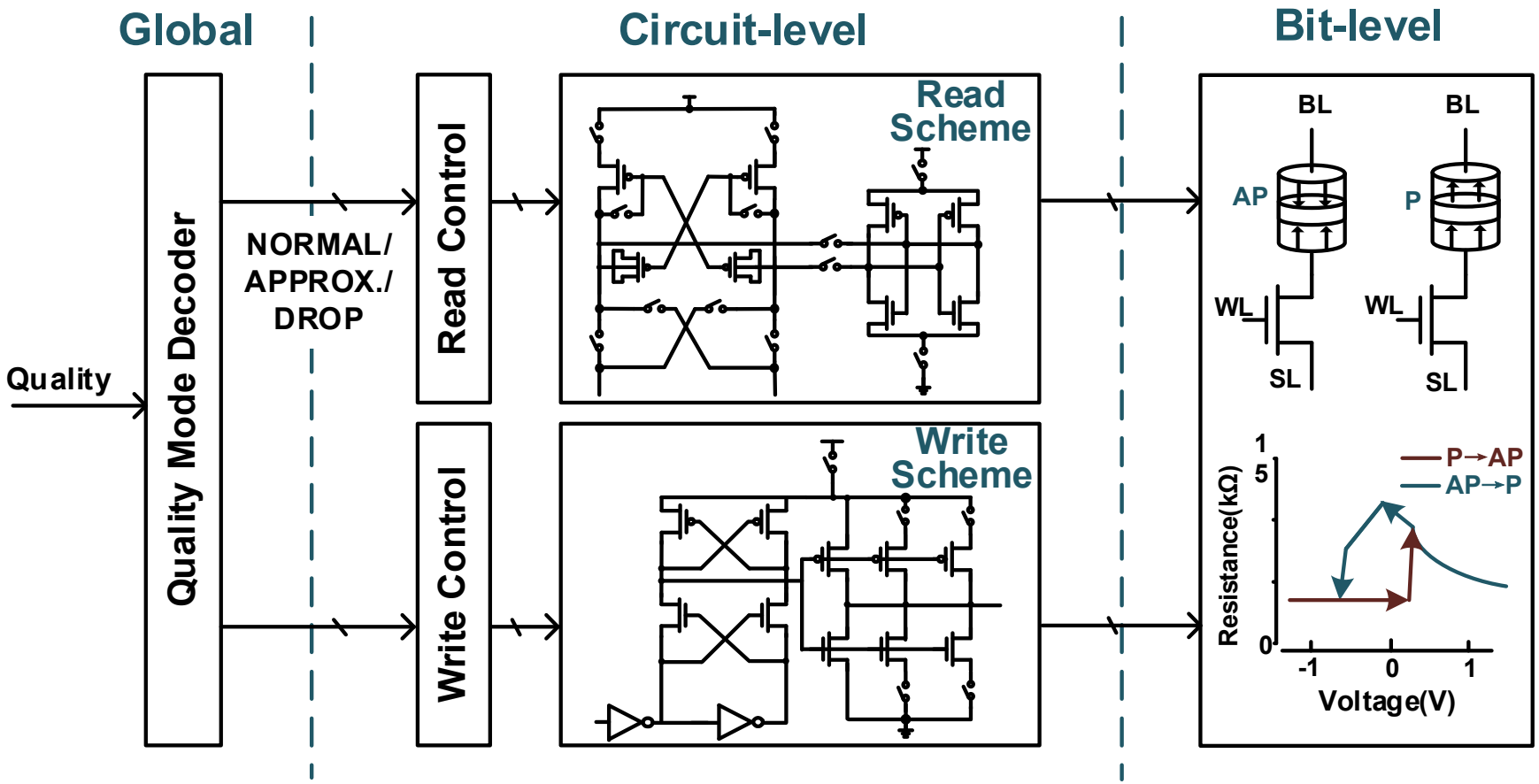
- Cases like worse Process/Voltage/Temperature(PVT), accurate read and write **cannot** guarantee MSBs' correctness
- ECC: **protect** MSBs
- Traditional ECC need extra check bits: **additional area**



Outline

2 Proposed Approximate Scheme

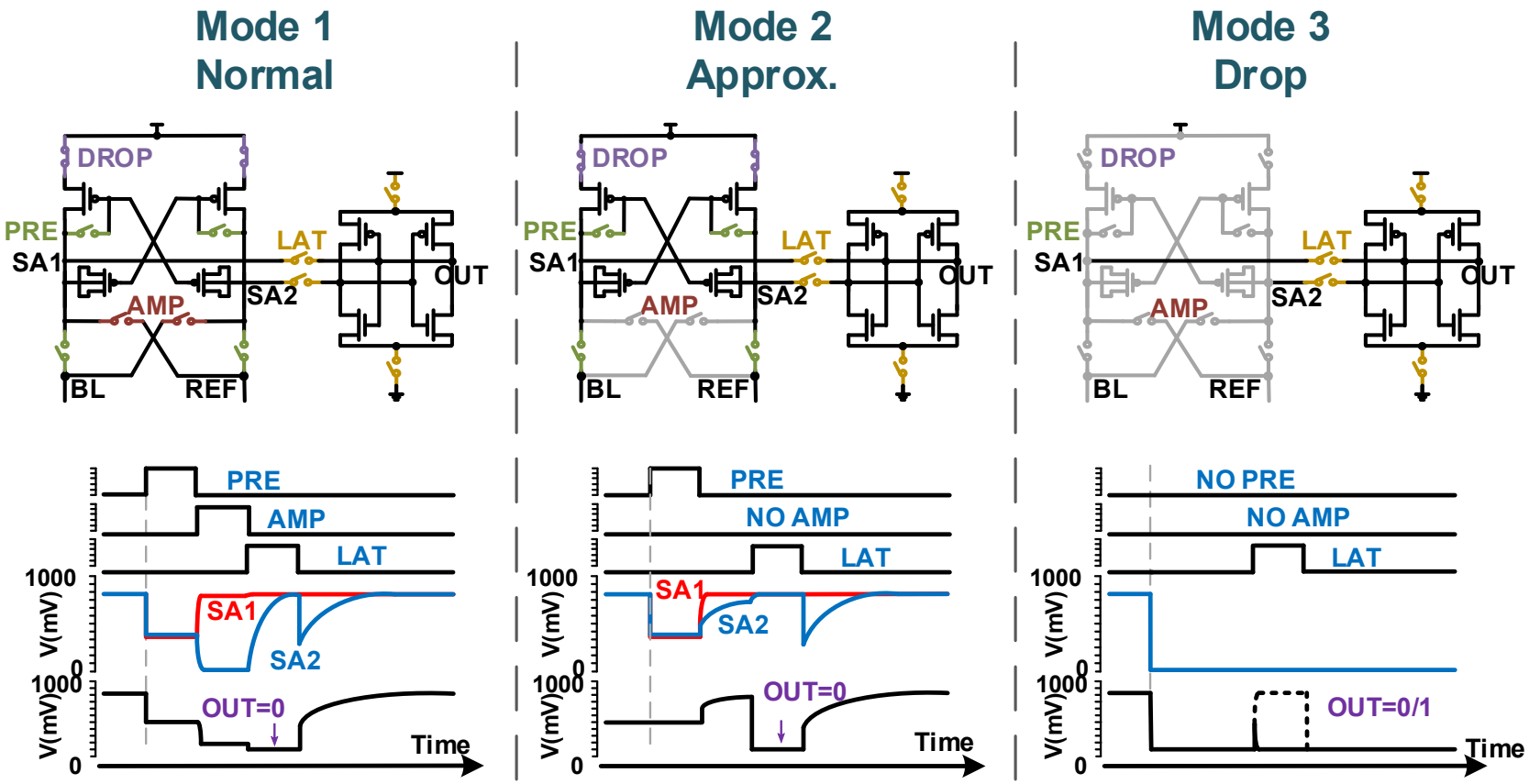
☐ Approximate Scheme Structure



2 Proposed Approximate Scheme

☐ Approximate Read Scheme

- ✘ Traditional scheme: need multiple sets of read circuits
- ✓ Proposed read scheme: Leverage one shared circuit



2 Proposed Approximate Scheme

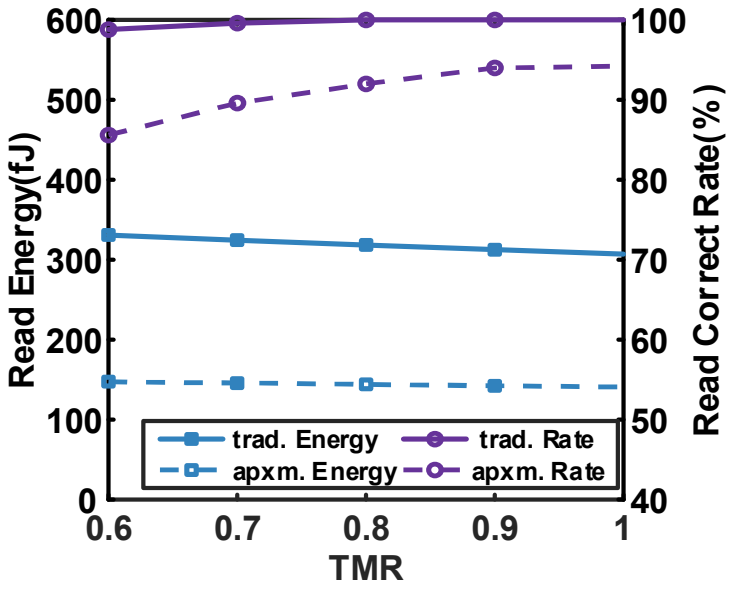
☐ Approximate Read Scheme

➢ Mode selection

- Normal:
  - ✓ Offset cancellation
  - ✓ Large sensing margin
  - ✓ Strong positive feedback
- Approximate:
  - ✓ Low energy consumption
  - ✓ Low correct rate
- Drop:
  - ✓ No power supply

➢ Single read access

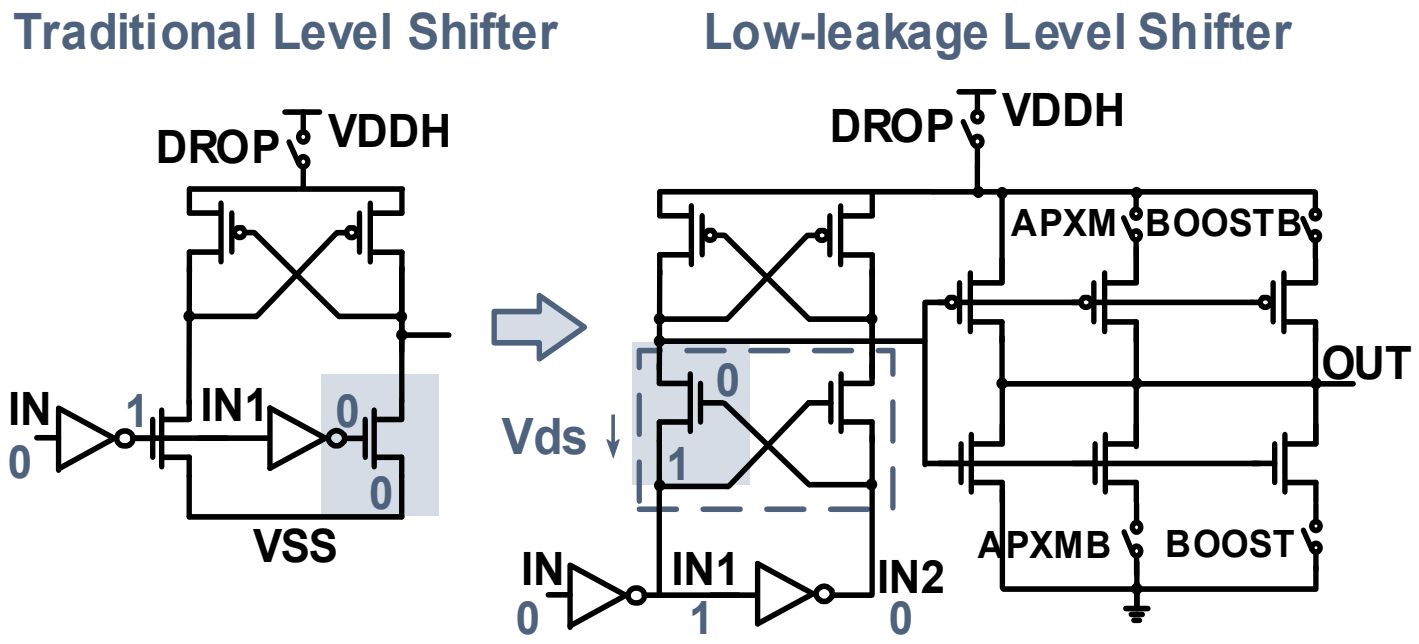
- 6.13uW ↓ 54.1% ↓ power
- 94.1% read correct rate



2 Proposed Approximate Scheme

☐ Approximate Write Scheme

- Level Shifter: common write driver structure that converts voltage levels
- Cross-coupled NMOSs' source: VSS → VDD
- Lower the Vds and minimize the static leakage current



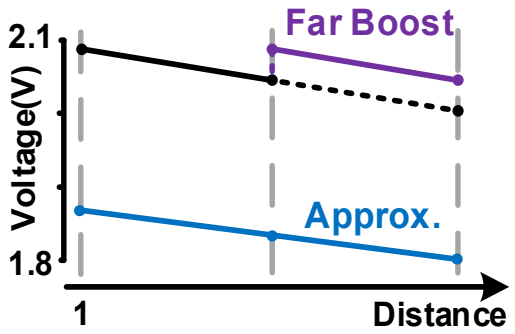
2 Proposed Approximate Scheme

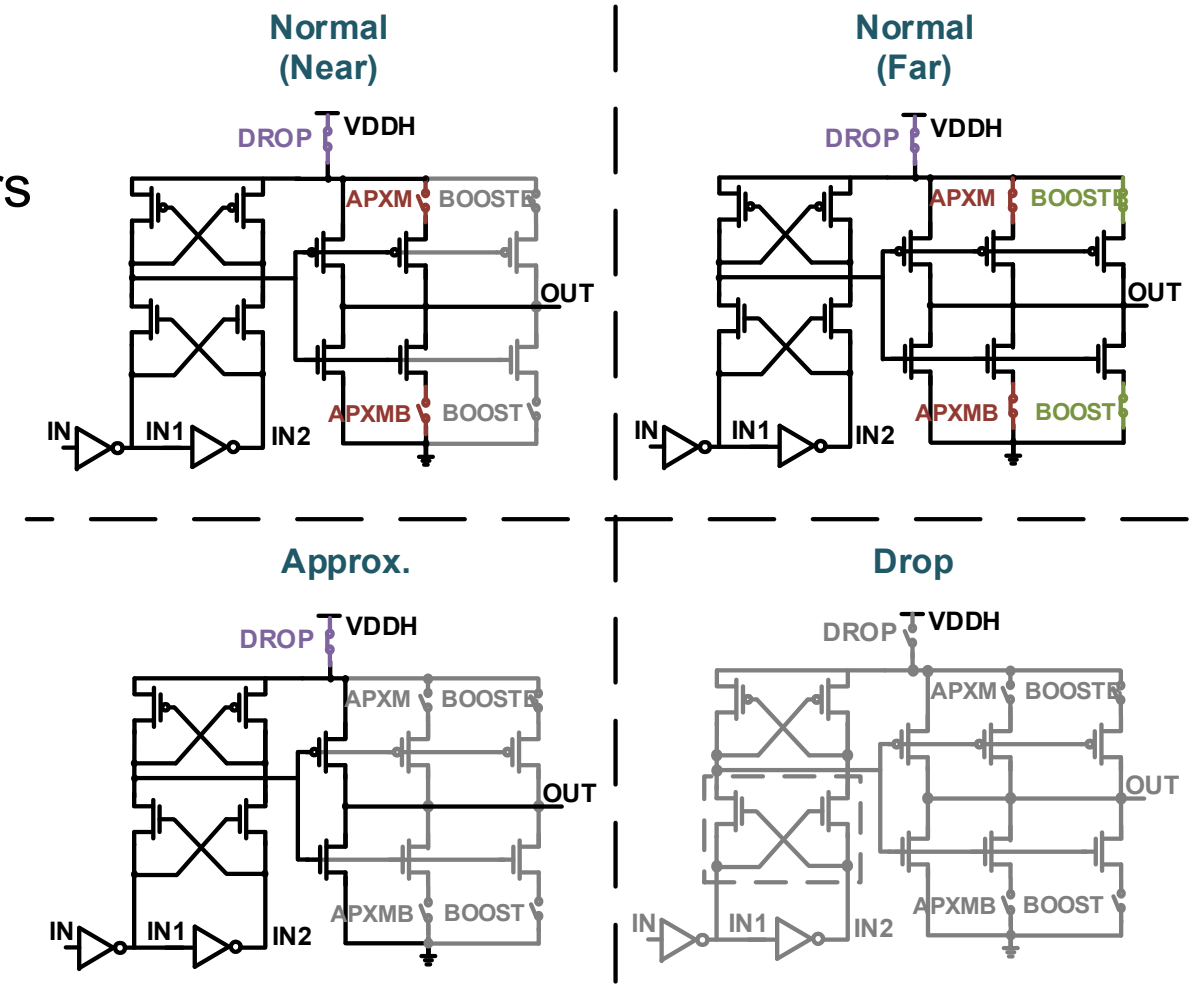
☐ Approximate Write Scheme

➢ **Mode selection**

- Normal : Near: 2 sets of transistors
  Far: 3 sets of transistors
- Approx. : 1 set of transistors
- Drop : no power supply

➢ IR-drop on BL/SL → Far boost



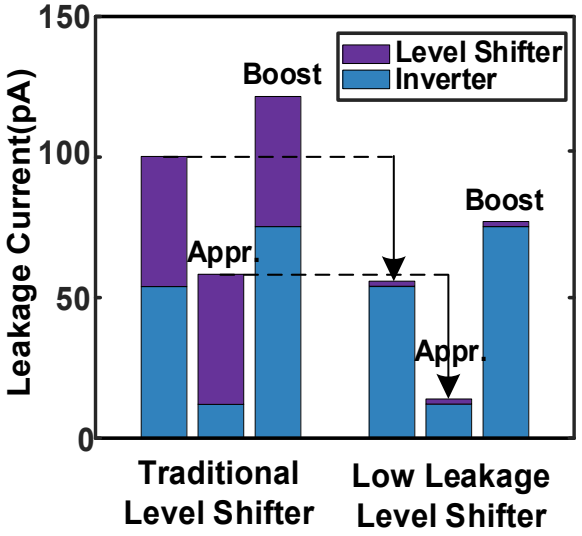


2 Proposed Approximate Scheme

☐ Approximate Write Scheme

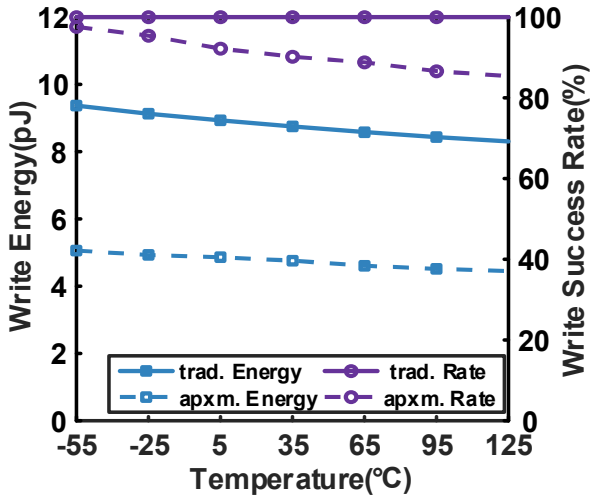
➢ Static Leakage current

- 44.12%↓ leakage current
- 74.78%↓ static power



➢ Single write access

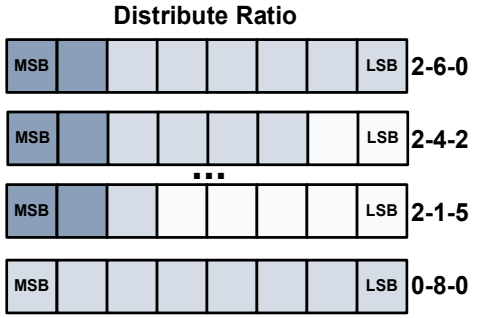
- 3.99pJ ↓ 45.6% ↓ energy (27℃)
- 95.1% write success rate

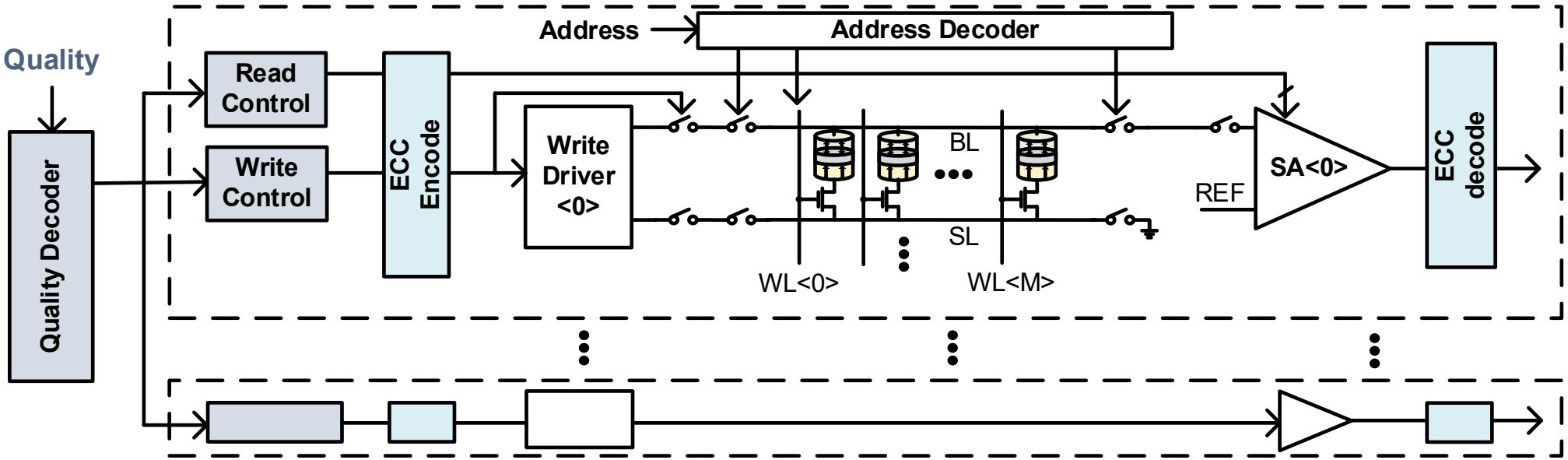


2 Proposed Approximate Scheme

☐ Bit-level Data Mapping

- Fine-grained: MSB precise, LSB approximate
- Distribute ratio： 3 modes distribute in 8 bits
- Quality determines the distribution schemes





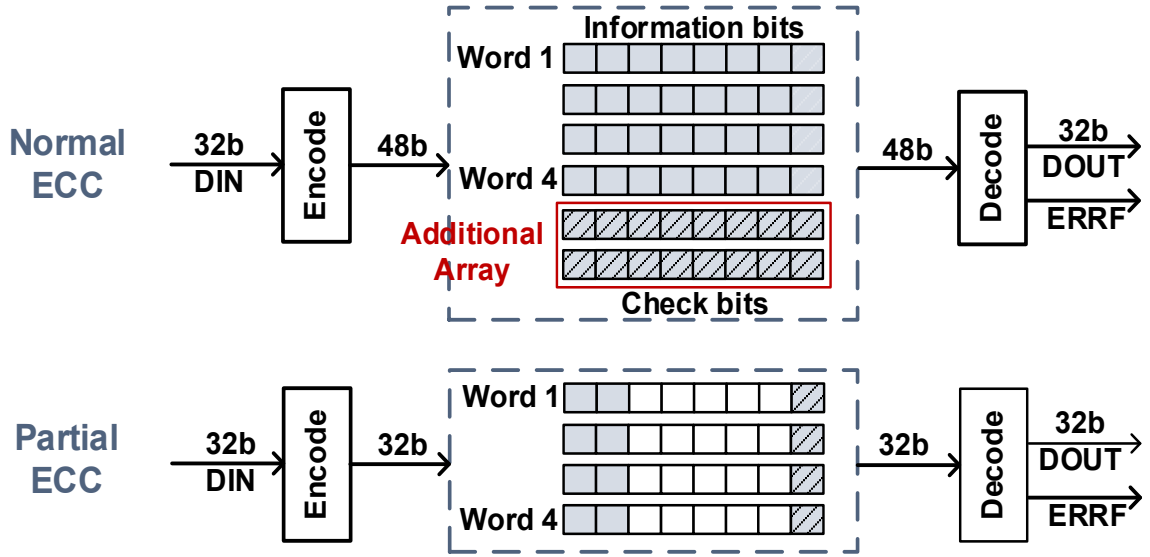
2 Proposed Approximate Scheme

☐ Partial Error Correction

Hamming	Correct single-bit error	Low hardware complexity
Reed-Solomon	Correct multi-bit errors	High complexity High power consumption
BCH	Correct multi-bit errors	
…	…	…

Suitable for fault-tolerant applications

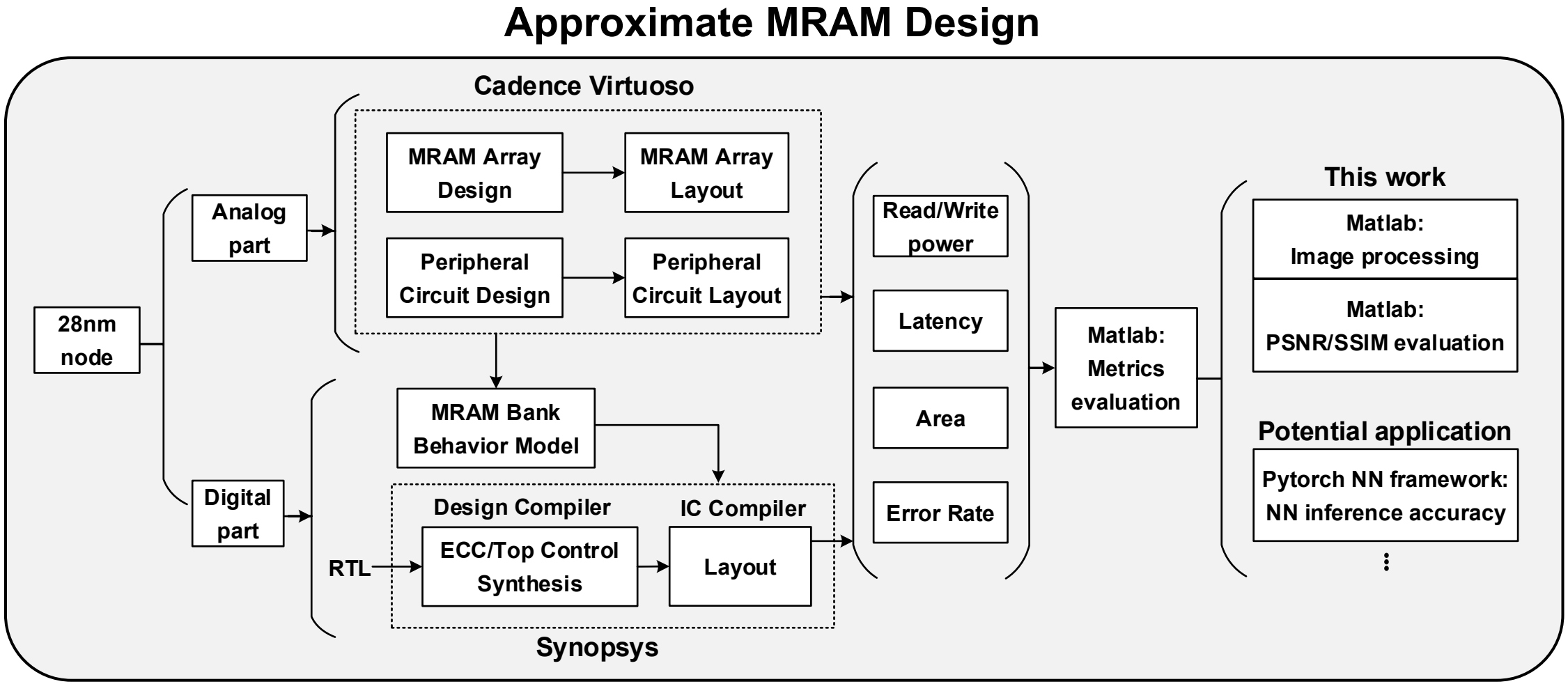
- Hamming(12,8):
  8 information bits, 4 check bits
- Use LSBs to store check bits:
  no additional arrays



Outline

3 Evaluation

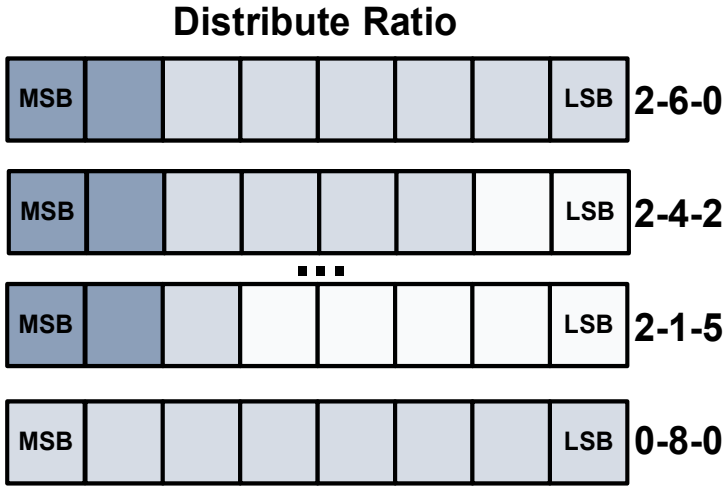
☐ Simulation Settings

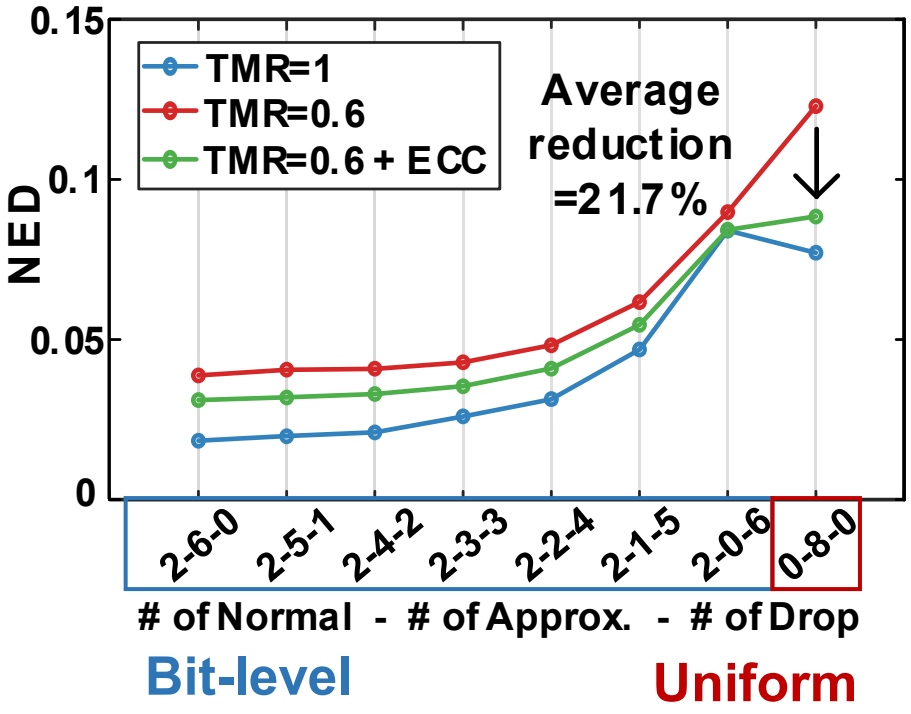


3 Evaluation

☐ Analysis Metrics: Normalized Error Distance (NED)

- $NED(ab) = \frac{ED(a,b)}{D} = \frac{\left|\Sigma_i a[\mathrm{i}]*2^{\mathrm{i}} - \Sigma_j b[j]*2^{j}\right|}{D}$
- measure the error distance after approximation
- Reliable: regardless of the size of the word





3 Evaluation

Analysis Metrics

Trade-off

- $power * NED$ & $powersaving/NED$
- Evaluate the energy-quality trade-off.
- A smaller $power * NED$, or larger $powersaving/NED$, means a better trade-off.

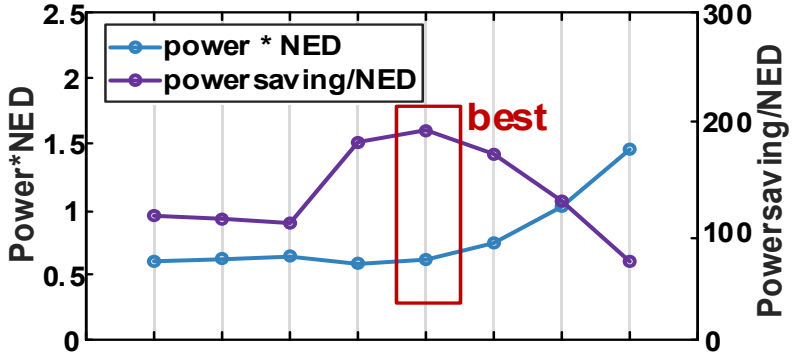
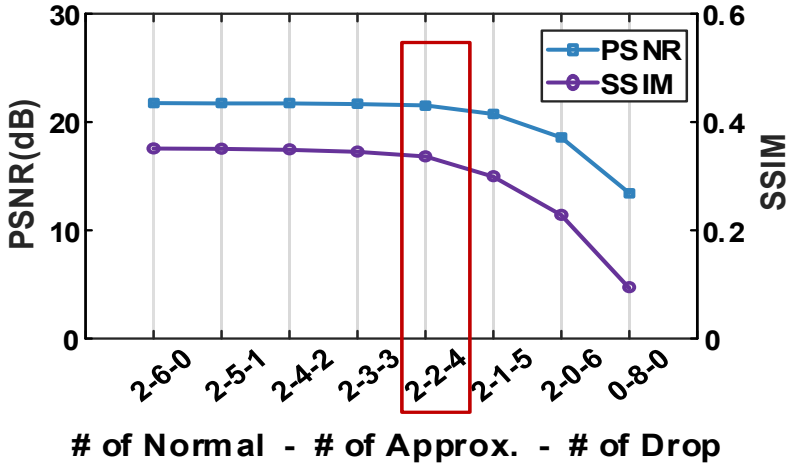


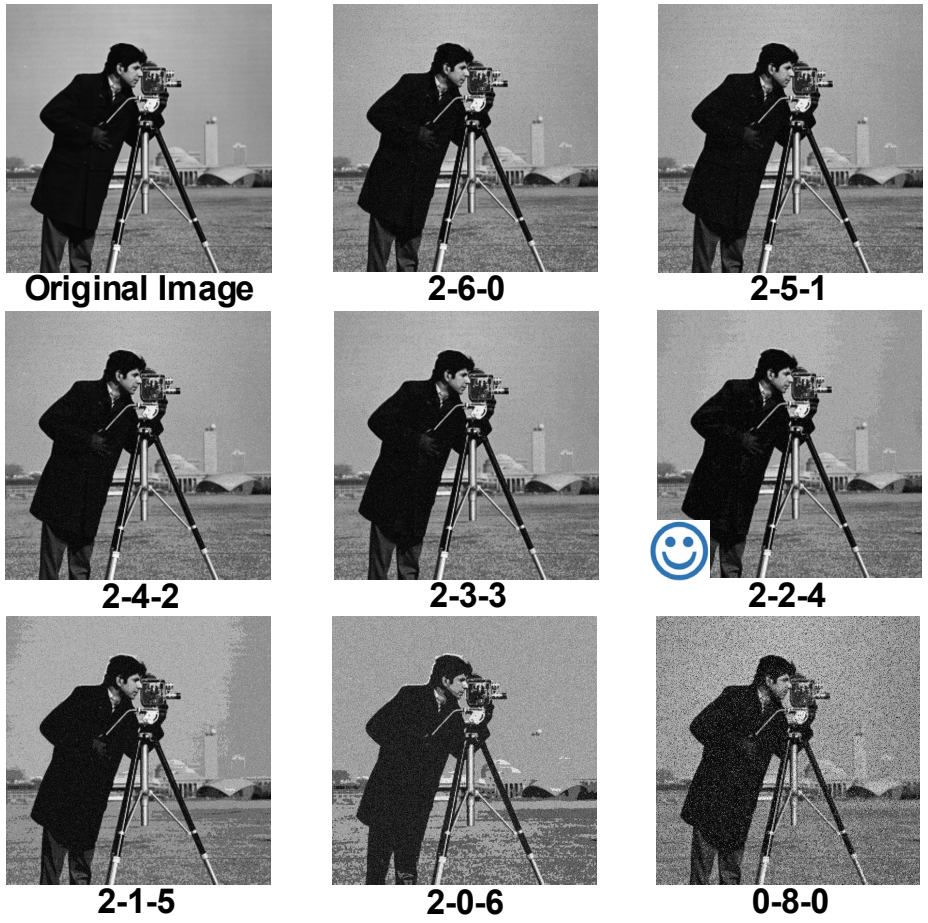
Image Processing

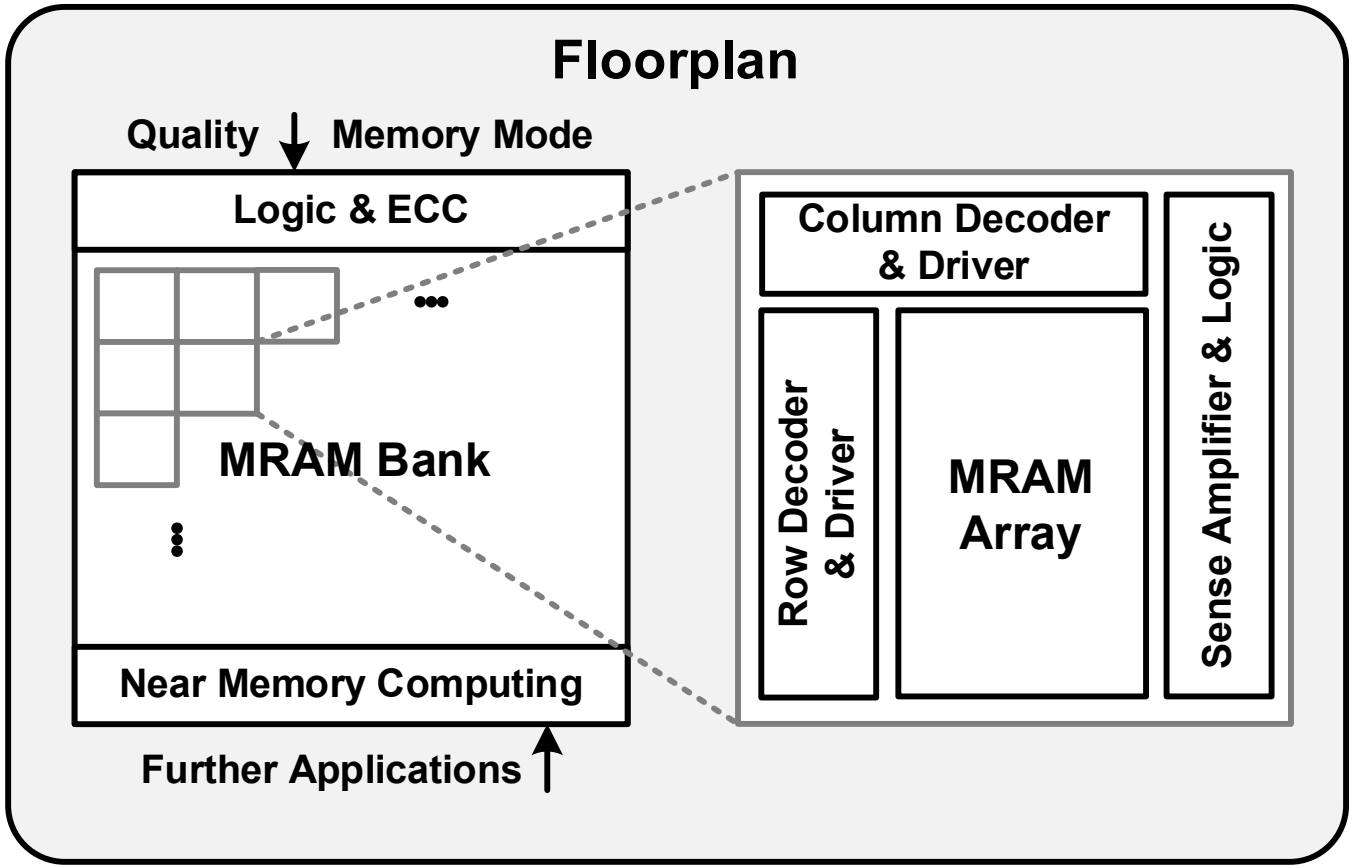
- Peak Signal-to-Noise Ratio(PSNR)
  Higher PSNR means less distortion in the image.
- Structural Similarity(SSIM)
  The closer SSIM is to 1, the more similar the images are.



3 Evaluation

- Applications: Image Processing and Potential Silicon Demonstration





3 Evaluation

- ✓ Both read and write access approximation
- ✓ Power reduction of up to 49.5%
- ✓ Better energy-quality trade-off
- ✓ Negligible area overhead

	TOC'23[15]	DAC'15[16], DATE'17[17]	JETC'20[18]	Design& Test'23[19]	ICCAD'17[20]	This Work
Application	image	cache	cache	image	image	image
Approximate Method	reduce t_W	reduce t_W, additional SA	I_R, I_R, t_R, t_W, retention	reduce t_W, speculation	reduce I_W	read and write circuit structure
ΔRead Power(%)	N/A	28	22.5	N/A	N/A	55.6
ΔWrite Power(%)	49.5	22	54.9	21.3	20	46.1
ΔPower(%)	49.5	9-30	42.5	21.3	20	49.5
BER(%)[b]	5	1	12	6.4	4.8	6
NED[a]	0.0148	0.0073	0.0291	0.181	0.0153	0.0311
ΔPower/NED[a]	33.45	12.33-41.1	14.6	11.77	13.07	15.92
Area Overhead	Y	Y	N/A	Y	N	N

[a]Estimated from the data.

[b]Estimated BER with partial error correction.

Outline

4 Conclusion

☐ Contributions of this work:

- A fine-grained approximate scheme for STT-MRAM is presented, allowing for dynamic modification to quality to achieve a **better energy-quality trade-off**
- Bit-level **read and write approximate schemes** are implemented by simply changing the control signals, resulting in an **energy saving** with **negligible area overhead** and **minimal quality loss**
- A **partial error correction** and **bit-level data mapping** method is proposed to protect the quality against the loss from TMR reduction.

Thank You!

Q & A