Reliability of Emerging Memory Devices

Asif Ali Khan

Fall Semester 2024

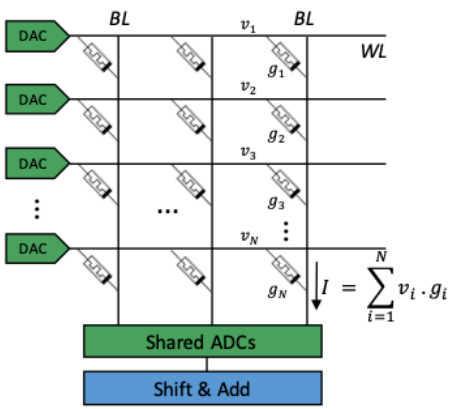
Department of Computer Systems Engineering

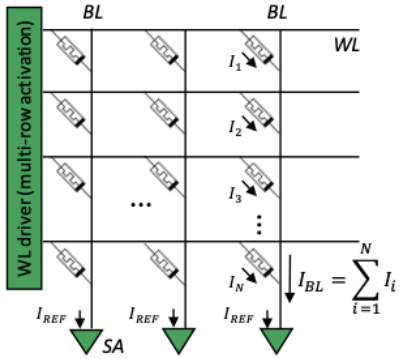
UET Peshawar, Pakistan

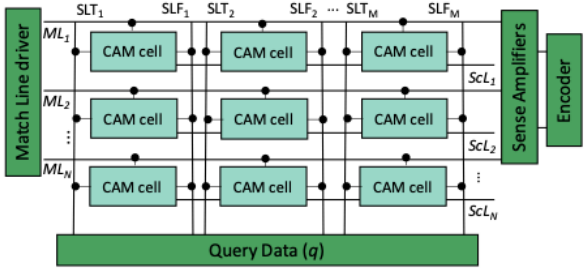
Dec 12, 2024

Recap: Compute-in-memory (CIM)

- The CIM paradigm aims to completely eliminate the data movement
- The fundamental idea is to exploit the physical properties of the memory devices to perform computations
- Not every computation can be performed with every technology







Khan et al., Arxiv 2024

Memory Reliability

What?

- Ability of memory devices to perform correctly over time under different conditions

Memory Reliability

What?

- Ability of memory devices to perform correctly over time under different conditions

Why?

- Critical for data integrity, system stability, and performance

Key reliability metrics

- Retention time: Ability to store data without corruption

Key reliability metrics

- Retention time: Ability to store data without corruption
- Endurance: Number of read/write cycles before failure

Key reliability metrics

- Retention time: Ability to store data without corruption
- Endurance: Number of read/write cycles before failure
- Error rate: Frequency of error during read/write operation

Key reliability metrics

- Retention time: Ability to store data without corruption
- Endurance: Number of read/write cycles before failure
- Error rate: Frequency of error during read/write operation
- Power cycling: Stability under repeated on/off cycles

Key reliability metrics

- Retention time: Ability to store data without corruption
- Endurance: Number of read/write cycles before failure
- Error rate: Frequency of error during read/write operation
- Power cycling: Stability under repeated on/off cycles
- Temperature stability: Resilience under varying thermal condition

Conventional memories

- SRAM
  - Soft errors due to cosmic rays
  - Power consumption in standby mode

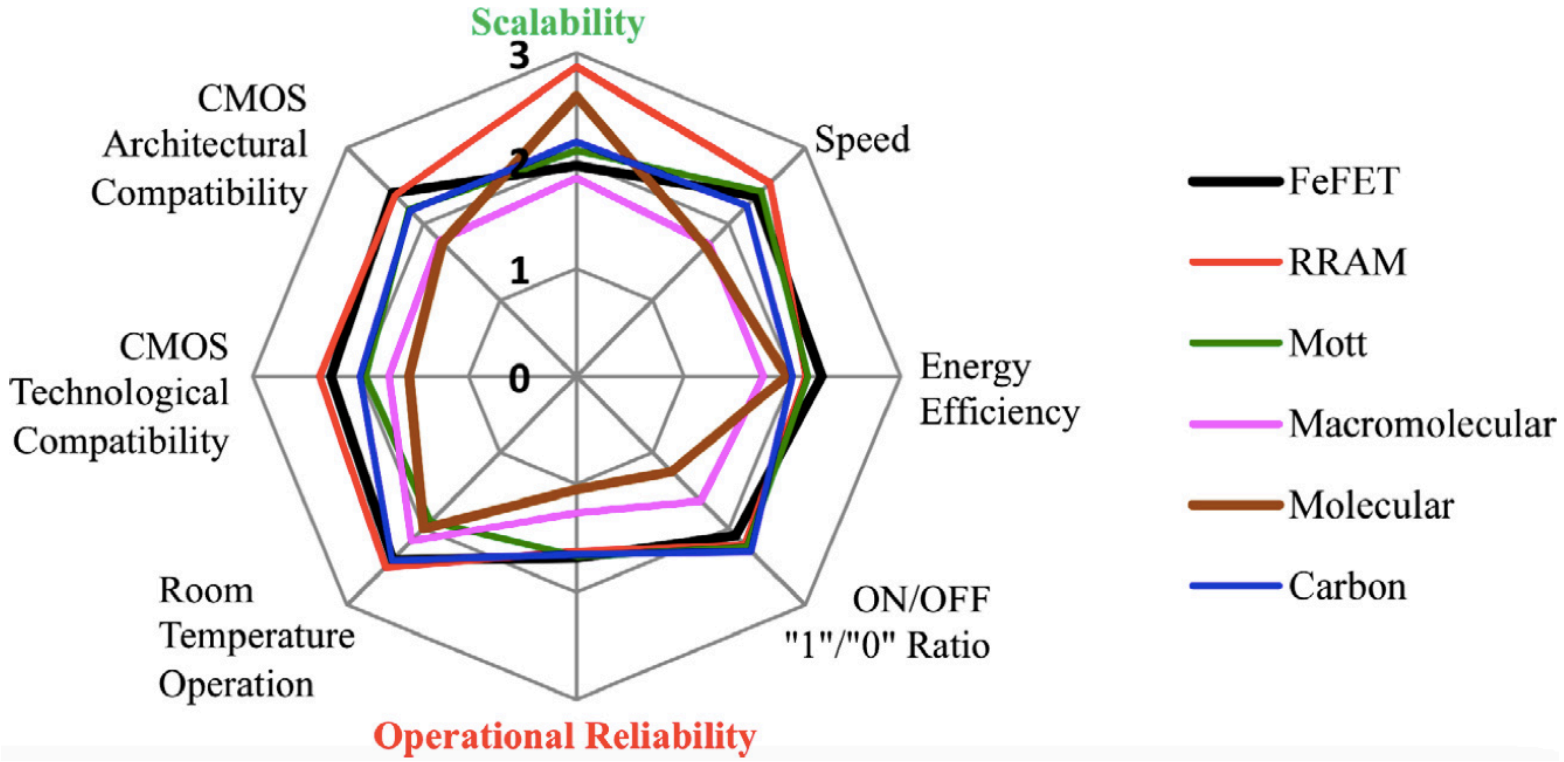
Conventional memories

- SRAM
  - Soft errors due to cosmic rays
  - Power consumption in standby mode
- DRAM:
  - Row hammer effect causing bit flips
  - Data retention issues with scaling

Conventional memories

- SRAM
  - Soft errors due to cosmic rays
  - Power consumption in standby mode
- DRAM:
  - Row hammer effect causing bit flips
  - Data retention issues with scaling
- Flash Memory:
  - Limited write endurance
  - Retention degradation with smaller feature sizes
  - Charge leakage in floating gate cells

Emerging NVMs



An Chen, Solid State Electronics, 2016

Reliability challenges in NVMs

- RRAM
  - Variability in resistance states
  - Retention times
  - Endurance issues

Reliability challenges in NVMs

- RRAM
  - Variability in resistance states
  - Retention times
  - Endurance issues
- PCM
  - Drift in resistance over time
  - Cycling induced wear and tear

Reliability challenges in NVMs

- RRAM
  - Variability in resistance states
  - Retention times
  - Endurance issues
- PCM
  - Drift in resistance over time
  - Cycling induced wear and tear
- MRAM
  - Write-energy and variability trade-off
  - Thermal stability of magnetic layers

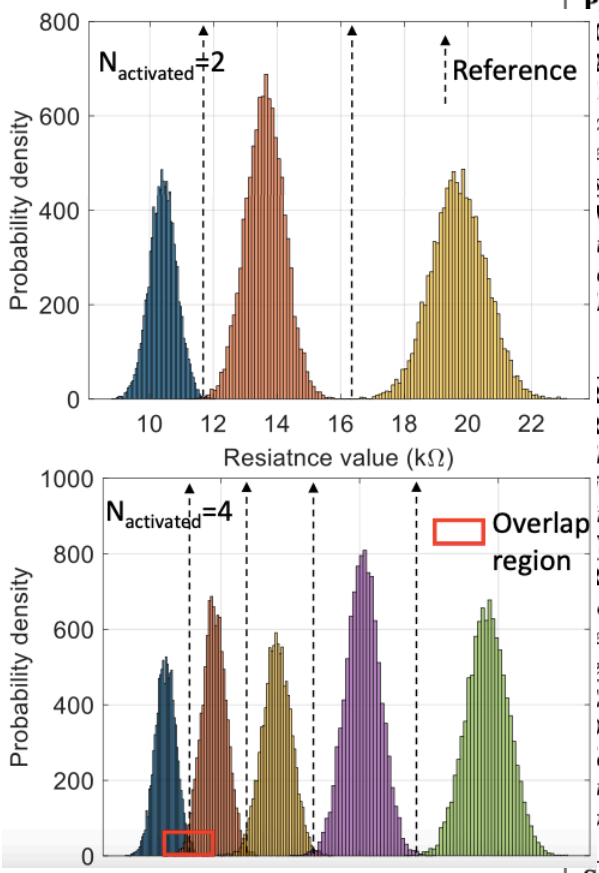
Reliability challenges in NVMs

- RRAM
  - Variability in resistance states
  - Retention times
  - Endurance issues
- PCM
  - Drift in resistance over time
  - Cycling induced wear and tear
- MRAM
  - Write-energy and variability trade-off
  - Thermal stability of magnetic layers
- RTMs: Misalignment issues (position errors)

CIM reliability

Reliability issue/*description*	Effect on logic gates	Statistical behavior	Implementation in proposed framework
Program variability (cycle-to-cycle (C2C) and device-to-device (D2D)) [20,21]: *At each programming (even with same applied voltage/time), the resultant resistance state will be (slightly) different*	Variations in device resistance lead to voltage variations in resistive voltage divider	The read-out resistance state after programming follows a statistical distribution, where the programming sets the distribution mean. Mean is statistically distributed, following either a C2C distribution (1 device): or a D2D distribution (multiple devices)	In MC-runs: draw devices parameters (filament length etc.) out of statistical distributions. Evaluate device resistance after switching with JART-model, fit to distribution and end resistance states.
Write Failures [23,24]: *at any fixed programming conditions, not all devices will switch (both for SET and RESET)*	Switching of memristive devices is not guaranteed for a certain (V,t) pulse. Devices have no fixed switching threshold.	Switching follows a stochastic process.	Obtain mean fitted switching probability function using MC-analysis with varied device parameters.
Read Noise [25] (also known as RTN or program instability): *Short-time current fluctuations (jumps) during device read-out over time, caused by resistance changes. Fluctuations increase with resistance.*	Variations in device resistance lead to voltage variations in resistive voltage divider (similar as Program variability)	At each device read, actual resistance values are varying over a statistical distribution. Mean is determined by programming. Distribution width increases with resistance.	Modeled as random walk with changes in oxygen vacancy concentration (only applied for Scouting)
Retention/State Drift [26] *Long-time changes of device resistance , effect is typically temperature accelerated*	Device resistance drifts over time and may lead to increased number of failures	Effect can be deterministically described on level of distributions as shift and tilt of read resistance distribution	-
Endurance *Device resistance window typically decreases with increasing number of program cycles, at the end devices no longer switch (stuck at 0 or 1)*	Change of device resistance may cause voltage divider and output stage errors, while stuck-at devices may cause write failures	Effect is also deterministic on distribution level: drift of C2C distributions and eventually occurrence of write failures	-
Sudden bit flips *Radiation-caused perturbation of CMOS based logic gates.*	Logic error	Potentially Erratic	Not present in ReRAM devices (but may affect transistors)

CIM reliability



Reliability issue/*description*	Effect on logic gates	Statistical behavior	Implementation in proposed framework
Program variability (cycle-to-cycle C2C) and device-to-device (D2D)) 20,21]: *t each programming (even with ame applied voltage/time), the esultant resistance state will be lightly) different*	Variations in device resistance lead to voltage variations in resistive voltage divider	The read-out resistance state after programming follows a statistical distribution, where the programming sets the distribution mean. Mean is statistically distributed, following either a C2C distribution (1 device): or a D2D distribution (multiple devices)	In MC-runs: draw devices parameters (filament length etc.) out of statistical distributions. Evaluate device resistance after switching with JART-model, fit to distribution and end resistance states.
Vrite Failures [23,24]: *t any fixed programming conditions, ot all devices will switch (both for ET and RESET)*	Switching of memristive devices is not guaranteed for a certain (V,t) pulse. Devices have no fixed switching threshold.	Switching follows a stochastic process.	Obtain mean fitted switching probability function using MC-analysis with varied device parameters.
ead Noise [25] (also known as TN or program instability): *hort-time current fluctuations umps) during device read-out over me, caused by resistance changes. luctuations increase with resistance.*	Variations in device resistance lead to voltage variations in resistive voltage divider (similar as Program variability)	At each device read, actual resistance values are varying over a statistical distribution. Mean is determined by programming. Distribution width increases with resistance.	Modeled as random walk with changes in oxygen vacancy concentration (only applied for Scouting)
etention/State Drift [26] *ong-time changes of device esistance , effect is typically mperature accelerated*	Device resistance drifts over time and may lead to increased number of failures	Effect can be deterministically described on level of distributions as shift and tilt of read resistance distribution	-
ndurance *evice resistance window typically ecreases with increasing number of rogram cycles, at the end devices no nger switch (stuck at 0 or 1)*	Change of device resistance may cause voltage divider and output stage errors, while stuck-at devices may cause write failures	Effect is also deterministic on distribution level: drift of C2C distributions and eventually occurrence of write failures	-
Sudden bit flips *Radiation-caused perturbation of CMOS based logic gates.*	Logic error	Potentially Erratic	Not present in ReRAM devices (but may affect transistors)

Techniques to improve reliability

- Error correction codes (ECC)
  - Single error correction, double error detection codes (SECDED)

Techniques to improve reliability

- Error correction codes (ECC)
  - Single error correction, double error detection codes (SECDED)
- Redundancy
  - Triple modular redundancy (TMR) or higher

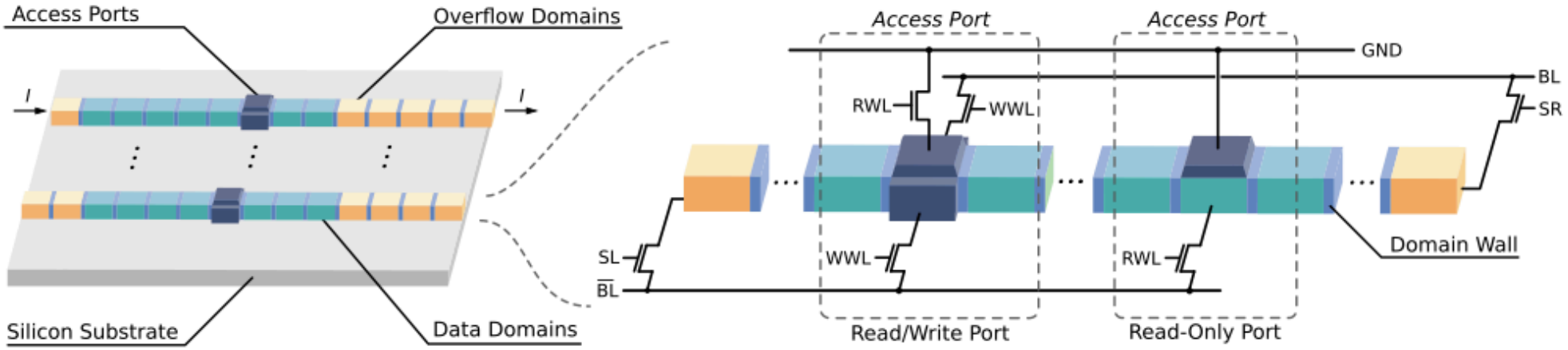
Techniques to improve reliability

- Error correction codes (ECC)
  - Single error correction, double error detection codes (SECDED)
- Redundancy
  - Triple modular redundancy (TMR) or higher
- Wear levelling
  - A standard techniques to prevent premature wear in NVMs/Flash

Techniques to improve reliability

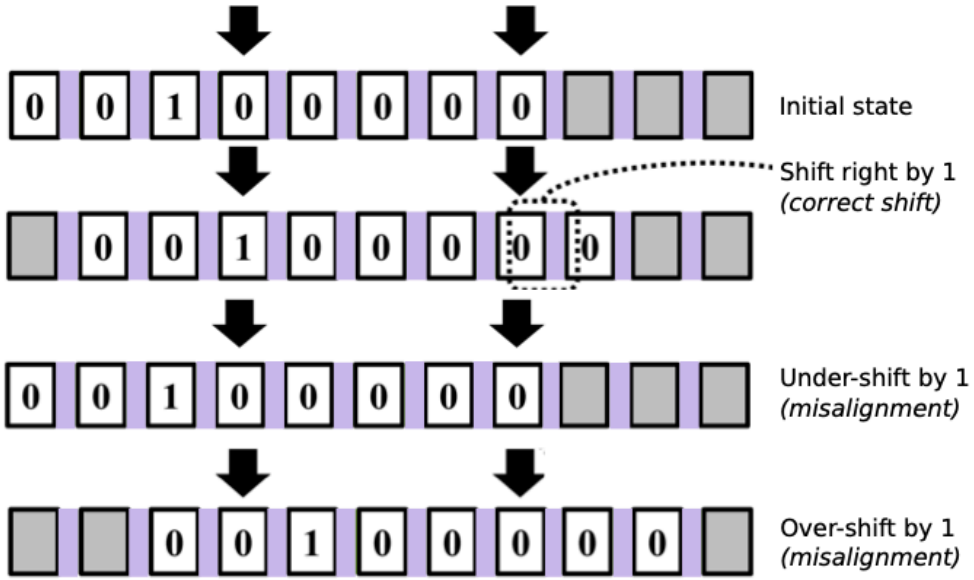
- Error correction codes (ECC)
  - Single error correction, double error detection codes (SECDED)
- Redundancy
  - Triple modular redundancy (TMR) or higher
- Wear levelling
  - A standard techniques to prevent premature wear in NVMs/Flash
- Circuit level techniques
  - Write verification for Flash and RRAM
  - Adaptive compression etc.

ECC codes for racetrack memory



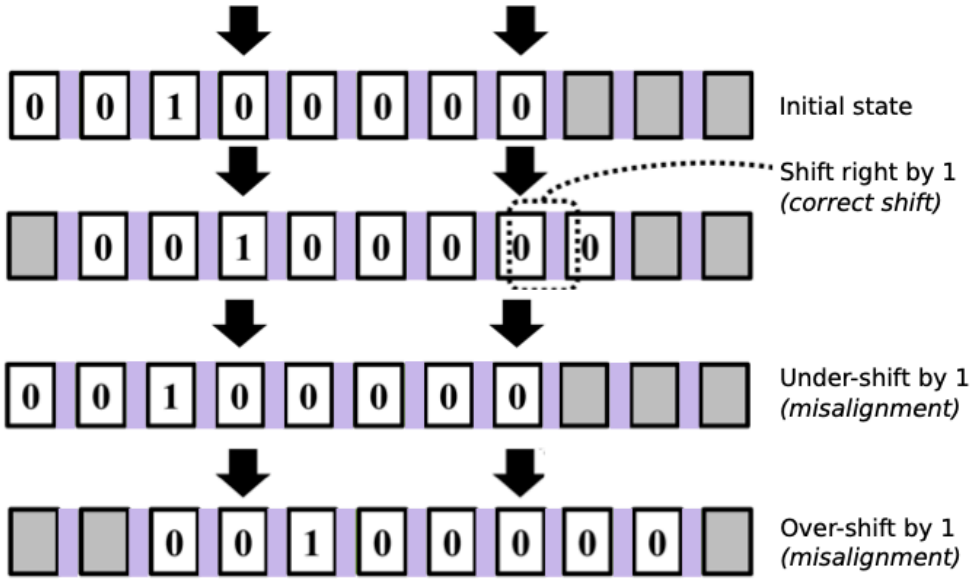
Khan et al., IEEE TC 2023

ECC codes for racetrack memory



Zhang et al., ISCA 2015

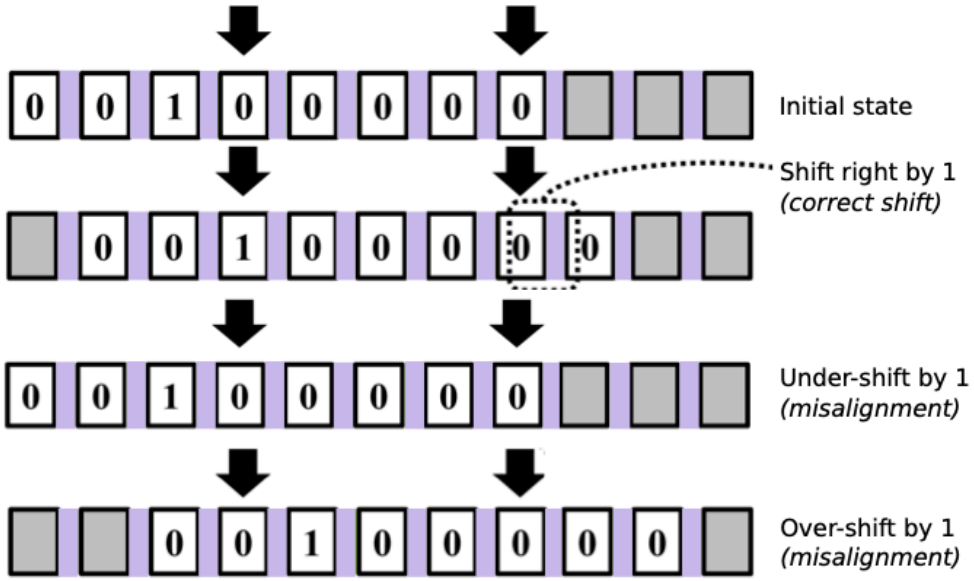
ECC codes for racetrack memory



Under- or over-shift can occur

Zhang et al., ISCA 2015

ECC codes for racetrack memory

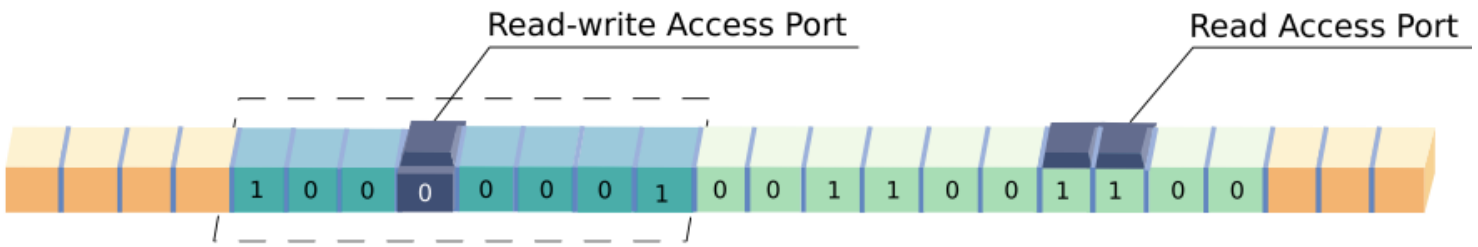


Under- or over-shift can occur

Distance	$\pm k$ Step Error Rate		
	$k = 1$	$k = 2$	$k \geq 3$
1	4.55×10^{-5}	1.37×10^{-21}	too small
2	9.95×10^{-5}	1.19×10^{-20}	too small
3	2.07×10^{-4}	5.59×10^{-20}	too small
4	3.76×10^{-4}	1.80×10^{-19}	too small
5	5.94×10^{-4}	4.47×10^{-19}	too small
6	8.43×10^{-4}	9.96×10^{-18}	too small
7	1.10×10^{-3}	7.57×10^{-15}	too small

Zhang et al., ISCA 2015

ECC codes for racetrack memory



- There are other proposals that reduces the overhead and improve performance

Khan et al., IEEE TC 2023

Reliability optimizations: Flip-N-write

suppress unnecessary bit programming actions by inspecting the old data word before writing the new data word and to opportunistically re-encode the new data word to further minimize bit programming. The following pseudo-code cap-

- Write as few bits as possible
- Only replace data bits if needed (compare new data to the old data and only write the different bits)

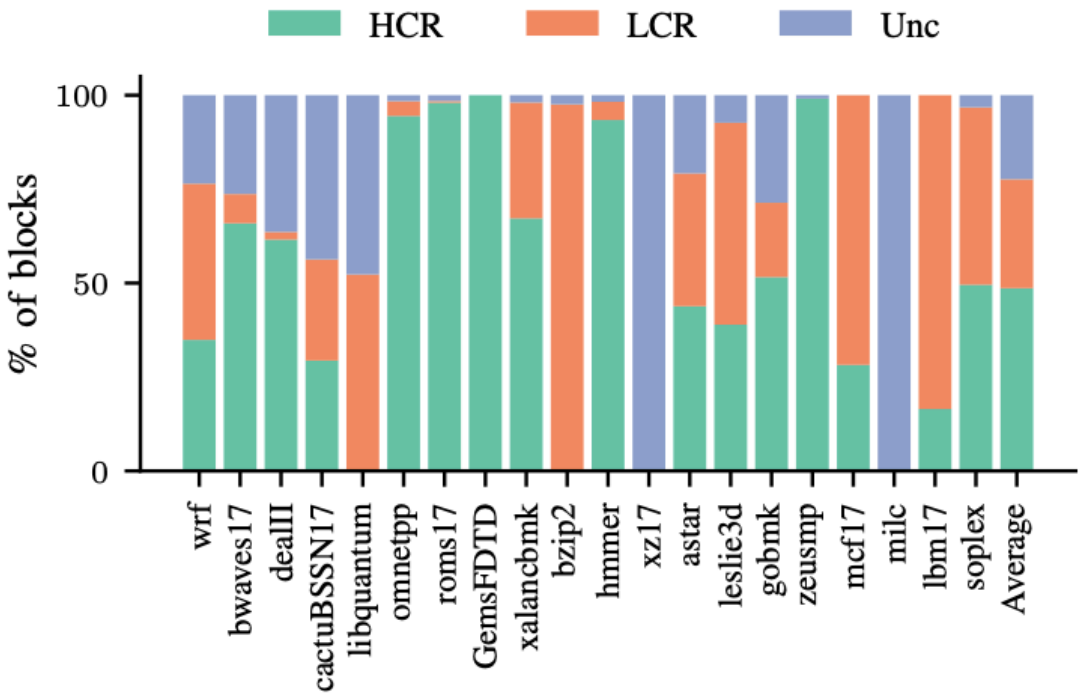
Cho et al., MICRO 2009

Compression-aware cache insertion

- In the case of hard faults, cache lines are disabled

Compression-aware cache insertion

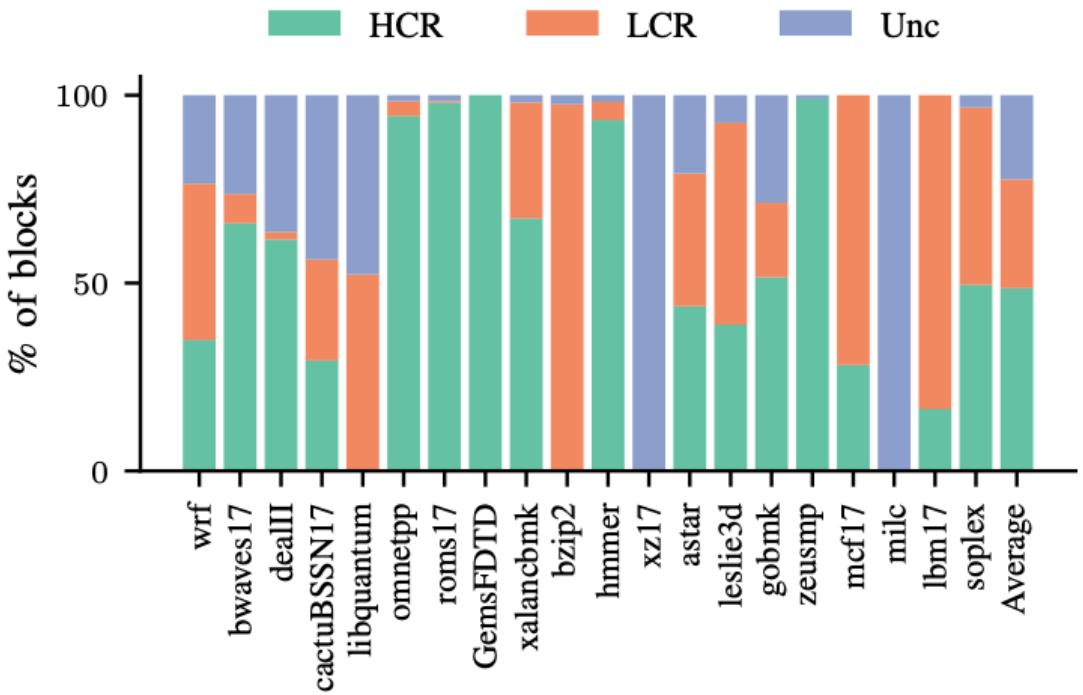
- In the case of hard faults, cache lines are disabled
- Not all cache lines require 100% capacity



Escuin et al., HPCA 2023

Compression-aware cache insertion

- In the case of hard faults, cache lines are disabled
- Not all cache lines require 100% capacity
- Map cache lines to NVM cache blocks, depending on the compression ratio of the cache line and faulty cells in the cache block



Escuin et al., HPCA 2023

Collaboratively optimizing for reliability and performance

- ECC is not for free
- The performance and energy overhead can be easily over 50%
- Therefore, it makes sense to jointly optimize for performance/reliability
- For instance, for reliable CIM-logic:
  - Reduce data movement by mapping dependent contents to the same column
  - Replace multi-operands operations with 2-operands ops

Bereholschi et al., SAMOS 2024

Optimizing for reliability overhead

- The expensive ECCs might not be equally effective in all cases

Layer	10^{-4}						4.55×10^{-5}						10^{-5}					
	64	32	16	8	4	2	64	32	16	8	4	2	64	32	16	8	4	2
1	-0.53	0.02	-0.23	-0.60	0.00	0.11	-0.51	-0.46	-0.79	-0.63	-0.63	0.00	-0.71	-0.46	-0.63	-0.58	0.00	0.00
2	-10.4	-9.68	-8.26	-6.50	-3.62	-1.73	-9.47	-9.17	-6.75	-3.67	-1.78	-0.12	-5.60	-3.08	-2.29	-1.07	-1.04	-0.35
3	-12.0	-11.7	-11.3	-7.80	-2.63	-1.86	-9.73	-6.29	-4.20	-4.28	-1.60	-1.27	-6.26	-2.82	-0.61	-0.86	-0.96	-0.58
4	-30.5	-23.0	-22.5	-22.6	-5.07	-3.49	-29.7	-24.9	-4.74	-9.60	-1.50	-0.96	-11.9	-9.78	-2.65	-1.68	-0.58	-0.61
5	-12.9	-16.6	-13.4	-5.86	-5.07	-0.97	-20.7	-13.1	-6.32	-7.13	-5.12	-0.96	-3.28	-1.88	-1.30	-0.99	-0.74	-0.23
6	-57.0	-48.1	-47.7	-46.5	-16.9	-3.31	-48.6	-41.3	-33.6	-14.4	-5.27	-0.91	-40.5	-20.8	-7.13	-2.09	-0.71	-0.61
7	-52.1	-49.3	-45.2	-40.7	-13.7	-3.11	-51.4	-46.8	-29.8	-20.5	-4.94	-0.89	-32.2	-27.7	-6.70	-2.06	-1.47	-0.81
8	-19.1	-18.0	-16.3	-8.69	-3.47	-0.97	-11.9	-10.0	-7.59	-2.34	-0.99	-0.28	-8.68	-2.95	-1.45	-1.45	-0.33	0.00
9	-18.1	-17.1	-11.5	-6.78	-2.50	-1.07	-20.4	-9.83	-9.37	-1.98	-0.89	-0.30	-6.29	-2.29	-1.02	-0.12	-0.48	-0.18
10	-10.9	-7.19	-5.27	-4.87	-1.53	-0.13	-7.61	-7.00	-6.14	-1.93	-0.84	-0.23	-3.23	-2.60	-0.99	-0.89	-0.81	-0.38
11	-46.5	-42.7	-33.5	-19.6	-9.45	-1.15	-35.0	-28.3	-18.6	-6.82	-2.90	-0.51	-17.9	-7.33	-4.30	-1.35	0.05	-0.07
12	-12.5	-11.6	-9.27	-5.27	-1.53	-0.61	-12.4	-8.79	-5.30	-1.98	-0.74	-0.05	-5.17	-2.19	-1.30	-0.91	-0.35	-0.10
13	-8.53	-5.35	-4.77	-3.64	-1.38	-0.56	-6.04	-4.86	-2.42	-1.63	-0.71	-0.40	-2.88	-1.58	-0.28	-0.58	-0.33	-0.76
14	-9.56	-5.89	-3.80	-2.70	-0.33	0.08	-6.49	-5.88	-2.72	-0.61	-0.18	-0.48	-2.29	-1.07	-0.20	0.13	-0.43	-0.46
15	-7.82	-5.40	-5.00	-3.62	-1.28	-0.23	-5.78	-5.37	-2.95	-2.37	-0.96	-0.33	-1.81	-0.63	-0.56	-0.66	-0.74	-0.10
16	-52.8	-46.0	-36.2	-7.84	-5.73	-1.61	-44.7	-31.0	-17.6	-7.26	-2.72	-0.66	-18.8	-5.88	-2.67	-0.79	-0.79	-0.46
17	-45.2	-35.4	-25.6	-12.2	-3.82	-0.48	-34.9	-24.7	-13.7	-4.63	-1.63	-0.58	-11.6	-4.25	-1.42	-1.27	0.05	-0.33
18	-3.21	-2.06	-2.29	-1.38	-0.33	-0.41	-1.98	-1.75	-1.09	-0.48	-0.25	-0.23	-0.84	-0.18	-0.05	-0.53	-0.28	-0.07
19	-51.6	-40.3	-22.8	-9.30	-2.27	-0.46	-31.1	-21.5	-8.71	-2.06	-0.99	-0.18	-7.54	-2.77	-1.37	-0.02	-0.05	-0.10
20	-56.8	-51.3	-30.2	-7.31	-0.46	-0.20	-45.0	-31.8	-7.31	-0.18	-0.07	-0.02	-6.70	-0.48	0.00	0.05	-0.07	-0.15
21	-61.4	-62.1	-43.2	-25.0	-21.4	-2.93	-59.5	-55.5	-32.5	-14.2	-3.74	-1.07	-38.8	-3.95	-4.15	-2.04	-0.89	-0.76

Bereholschi et al., SAMOS 2024

Optimizing for reliability overhead

- The expensive ECCs might not be equally effective in all cases

- Balancing performance and accuracy tradeoff by selectively protecting only important layers/ regions.

Layer	10^{-4}						4.55×10^{-5}						10^{-5}					
	64	32	16	8	4	2	64	32	16	8	4	2	64	32	16	8	4	2
1	-0.53	0.02	-0.23	-0.60	0.00	0.11	-0.51	-0.46	-0.79	-0.63	-0.63	0.00	-0.71	-0.46	-0.63	-0.58	0.00	0.00
2	-10.4	-9.68	-8.26	-6.50	-3.62	-1.73	-9.47	-9.17	-6.75	-3.67	-1.78	-0.12	-5.60	-3.08	-2.29	-1.07	-1.04	-0.35
3	-12.0	-11.7	-11.3	-7.80	-2.63	-1.86	-9.73	-6.29	-4.20	-4.28	-1.60	-1.27	-6.26	-2.82	-0.61	-0.86	-0.96	-0.58
4	-30.5	-23.0	-22.5	-22.6	-5.07	-3.49	-29.7	-24.9	-4.74	-9.60	-1.50	-0.96	-11.9	-9.78	-2.65	-1.68	-0.58	-0.61
5	-12.9	-16.6	-13.4	-5.86	-5.07	-0.97	-20.7	-13.1	-6.32	-7.13	-5.12	-0.96	-3.28	-1.88	-1.30	-0.99	-0.74	-0.23
6	-57.0	-48.1	-47.7	-46.5	-16.9	-3.31	-48.6	-41.3	-33.6	-14.4	-5.27	-0.91	-40.5	-20.8	-7.13	-2.09	-0.71	-0.61
7	-52.1	-49.3	-45.2	-40.7	-13.7	-3.11	-51.4	-46.8	-29.8	-20.5	-4.94	-0.89	-32.2	-27.7	-6.70	-2.06	-1.47	-0.81
8	-19.1	-18.0	-16.3	-8.69	-3.47	-0.97	-11.9	-10.0	-7.59	-2.34	-0.99	-0.28	-8.68	-2.95	-1.45	-1.45	-0.33	0.00
9	-18.1	-17.1	-11.5	-6.78	-2.50	-1.07	-20.4	-9.83	-9.37	-1.98	-0.89	-0.30	-6.29	-2.29	-1.02	-0.12	-0.48	-0.18
10	-10.9	-7.19	-5.27	-4.87	-1.53	-0.13	-7.61	-7.00	-6.14	-1.93	-0.84	-0.23	-3.23	-2.60	-0.99	-0.89	-0.81	-0.38
11	-46.5	-42.7	-33.5	-19.6	-9.45	-1.15	-35.0	-28.3	-18.6	-6.82	-2.90	-0.51	-17.9	-7.33	-4.30	-1.35	0.05	-0.07
12	-12.5	-11.6	-9.27	-5.27	-1.53	-0.61	-12.4	-8.79	-5.30	-1.98	-0.74	-0.05	-5.17	-2.19	-1.30	-0.91	-0.35	-0.10
13	-8.53	-5.35	-4.77	-3.64	-1.38	-0.56	-6.04	-4.86	-2.42	-1.63	-0.71	-0.40	-2.88	-1.58	-0.28	-0.58	-0.33	-0.76
14	-9.56	-5.89	-3.80	-2.70	-0.33	0.08	-6.49	-5.88	-2.72	-0.61	-0.18	-0.48	-2.29	-1.07	-0.20	0.13	-0.43	-0.46
15	-7.82	-5.40	-5.00	-3.62	-1.28	-0.23	-5.78	-5.37	-2.95	-2.37	-0.96	-0.33	-1.81	-0.63	-0.56	-0.66	-0.74	-0.10
16	-52.8	-46.0	-36.2	-7.84	-5.73	-1.61	-44.7	-31.0	-17.6	-7.26	-2.72	-0.66	-18.8	-5.88	-2.67	-0.79	-0.79	-0.46
17	-45.2	-35.4	-25.6	-12.2	-3.82	-0.48	-34.9	-24.7	-13.7	-4.63	-1.63	-0.58	-11.6	-4.25	-1.42	-1.27	0.05	-0.33
18	-3.21	-2.06	-2.29	-1.38	-0.33	-0.41	-1.98	-1.75	-1.09	-0.48	-0.25	-0.23	-0.84	-0.18	-0.05	-0.53	-0.28	-0.07
19	-51.6	-40.3	-22.8	-9.30	-2.27	-0.46	-31.1	-21.5	-8.71	-2.06	-0.99	-0.18	-7.54	-2.77	-1.37	-0.02	-0.05	-0.10
20	-56.8	-51.3	-30.2	-7.31	-0.46	-0.20	-45.0	-31.8	-7.31	-0.18	-0.07	-0.02	-6.70	-0.48	0.00	0.05	-0.07	-0.15
21	-61.4	-62.1	-43.2	-25.0	-21.4	-2.93	-59.5	-55.5	-32.5	-14.2	-3.74	-1.07	-38.8	-3.95	-4.15	-2.04	-0.89	-0.76

Bereholschi et al., SAMOS 2024

Thank you!

asif.ali@uetpeshawar.edu.pk