Exploiting Application Error Resilience for Energy Savings in Memories

Georgios Karakonstantis[1,2]

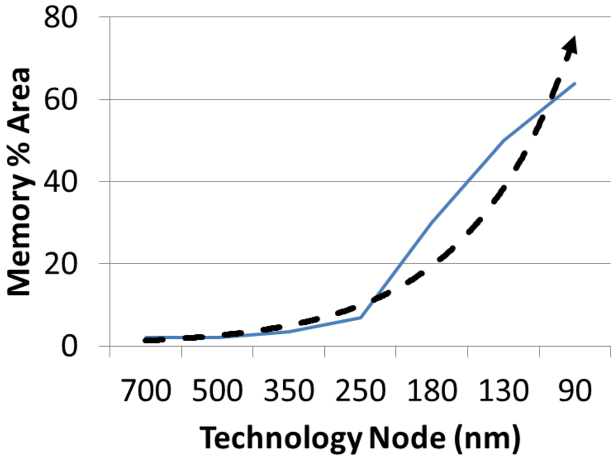
Adam Teman[1], Shrikanth Ganapathy[1], Andreas Burg[1]

[1]Swiss Federal Institute of Technology Lausanne (EPFL), Switzerland
[2]Queen's University Belfast, U.K.

WAPCO
Amsterdam 19/1/15

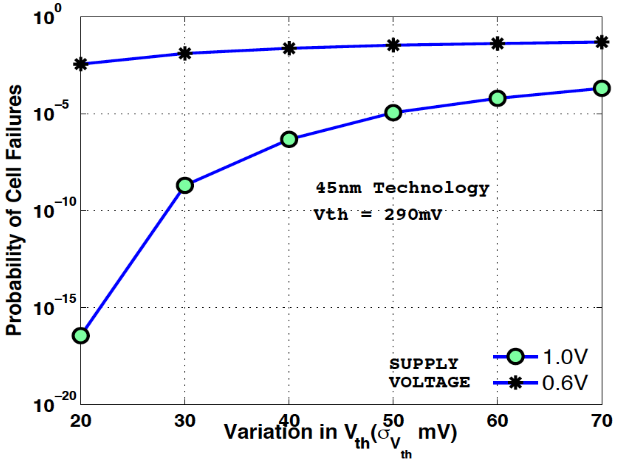
Memories in Nanometer Nodes

- The percentage of memories in today's systems is constantly increasing
  - Dynamic Memories – increased density
  - Static Memories - faster, no refresh power



- The high density requirements press for aggressive scaling of transistor sizes
  - Worsening of parametric variations
  - Worsening of retention time in DRAMs
    - More read, write, access failures

- The need for energy efficient and extended battery lifetime asks for scaled supply voltages
  - Memories become more prone to failures



Traditional Mechanisms for Robust Operation

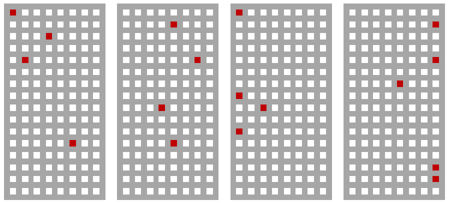
- Overdesign by adding **preventive guardbands** based on worst-case conditions assumed at design time
  - Up-scale voltage and/or size-up the transistors of all bit-cells
  - Refresh DRAM more frequently than required based on the worst case cell

- Add **redundant mechanisms** for detecting and correcting every single error
  - Error correcting codes

⚠ Power, performance and area overheads for all manufactured memory chips, even the good ones

- Each manufactured die is subject to different error pattern (number and location of errors)
- Worst case cell is used for guardbanding

Different instances of same designed memory



Need for alternative paradigms that relax the error-free requirements
=> Approximate computing exploiting application error-resilience

Outline

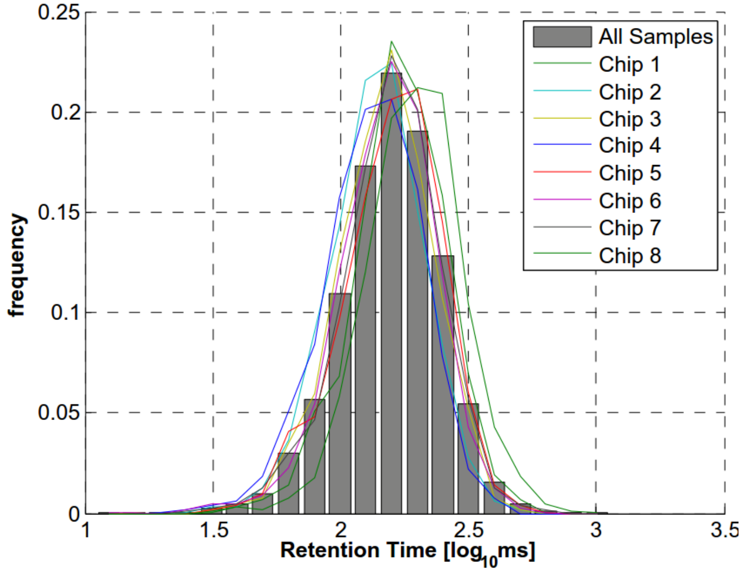
- **Potential for Energy Savings by Relaxing Worst-Case Guardbands**
  - Analysis of the DRAM retention time variability and traditional robust design
  - Energy savings by relaxing the worst-case and error-free requirements
  - Achieving graceful degradation for enabling and promoting approximate storage

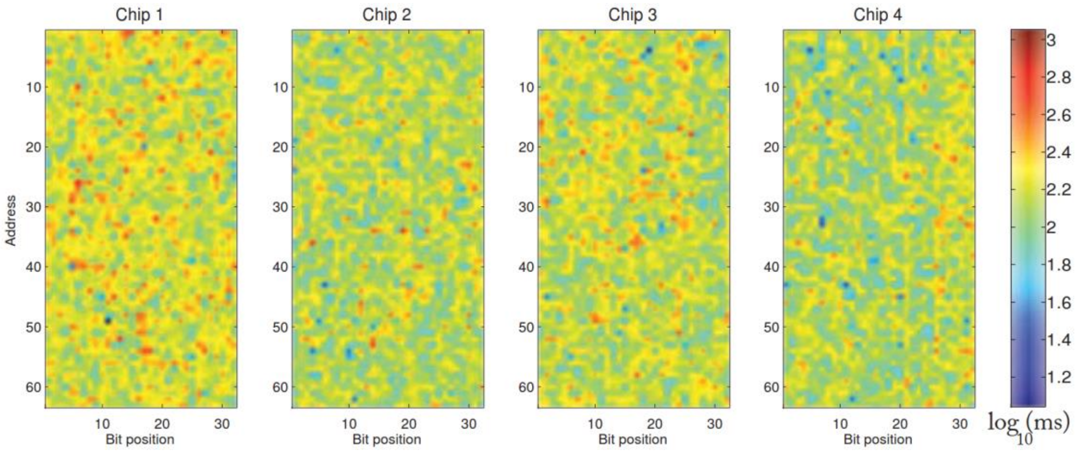
- **Alternative Error Mitigation Mechanisms**

- **Conclusion**

DRAM: Retention Time Variability & Conventional Robust Design

- ❑ Data integrity can be guaranteed for a limited time period
- ❑ Avoid retention-time violation by frequent power-hungry refresh cycles
- ❑ Silicon measurements indicated large variability (2 orders) of retention time across all manufactured dies of a 2kb array



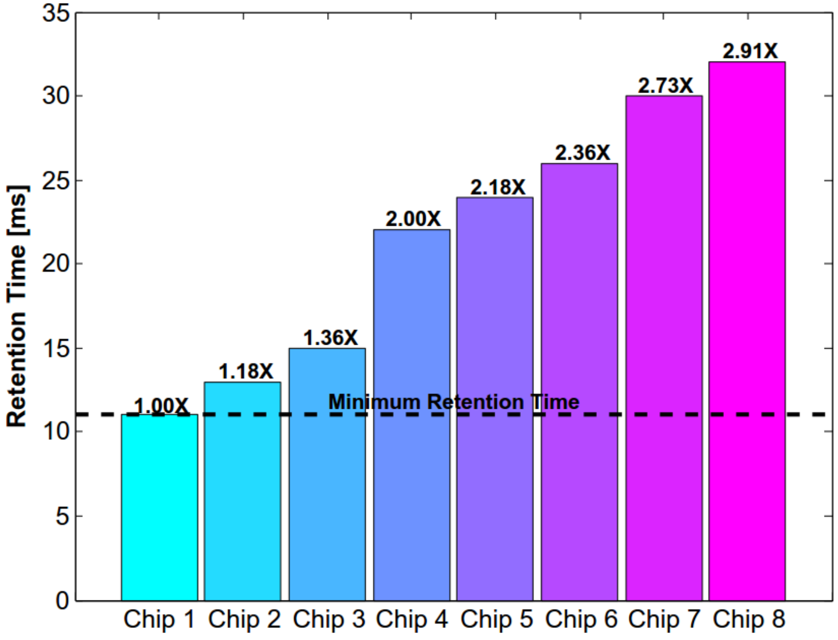


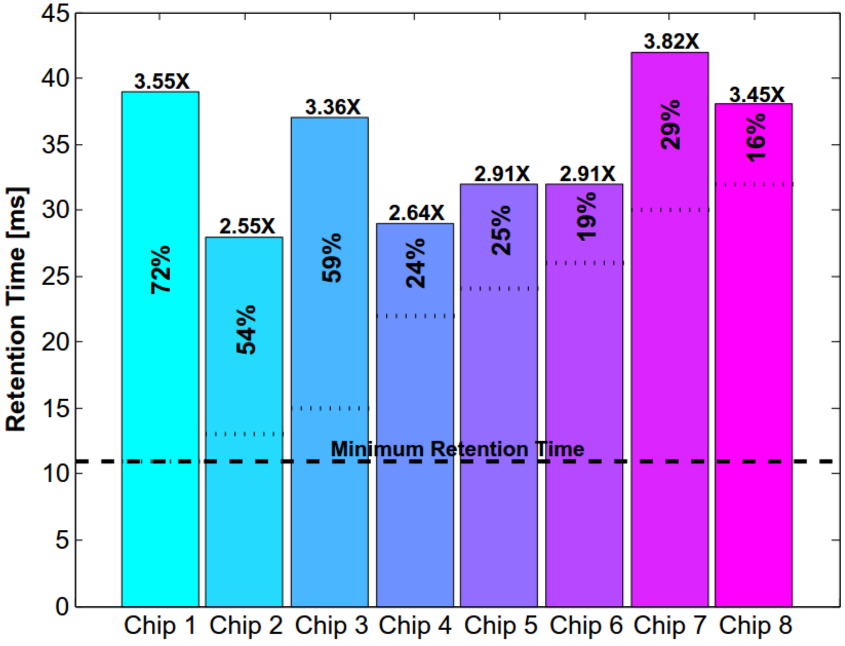
- ❑ Global refresh rate is determined by the WORST cell of all dies
  - ▪ Pessimistic performance
  - ▪ Large energy waste

Power Savings by Relaxing the Worst Case Assumptions and Error-Free Requirements

- 3x difference in refresh power
- Large energy savings by setting the refresh independently for each die
  - Extra cost for testing

- New criterion for setting the RT such that a limited number of errors is allowed
- Take advantage of the data integrity/refresh power trade offs

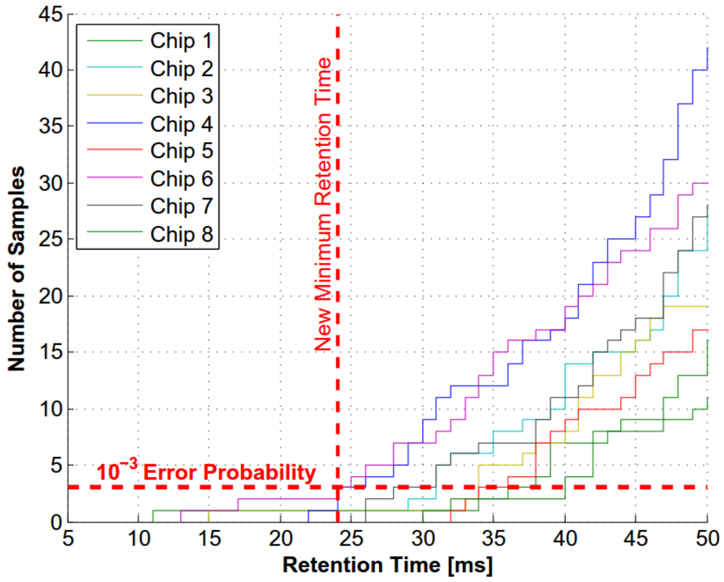


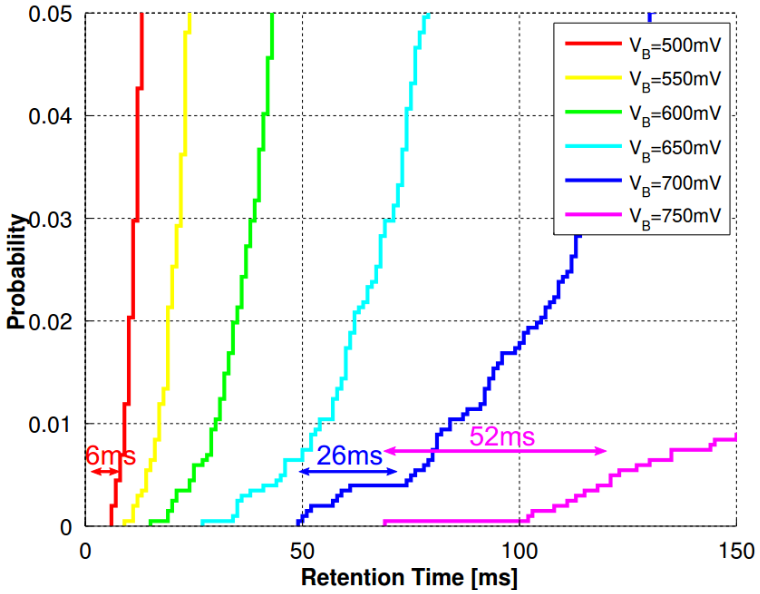


‘Approximate’ storage can lead to large power savings

Achieving Graceful Degradation for Enabling Approximate Storage

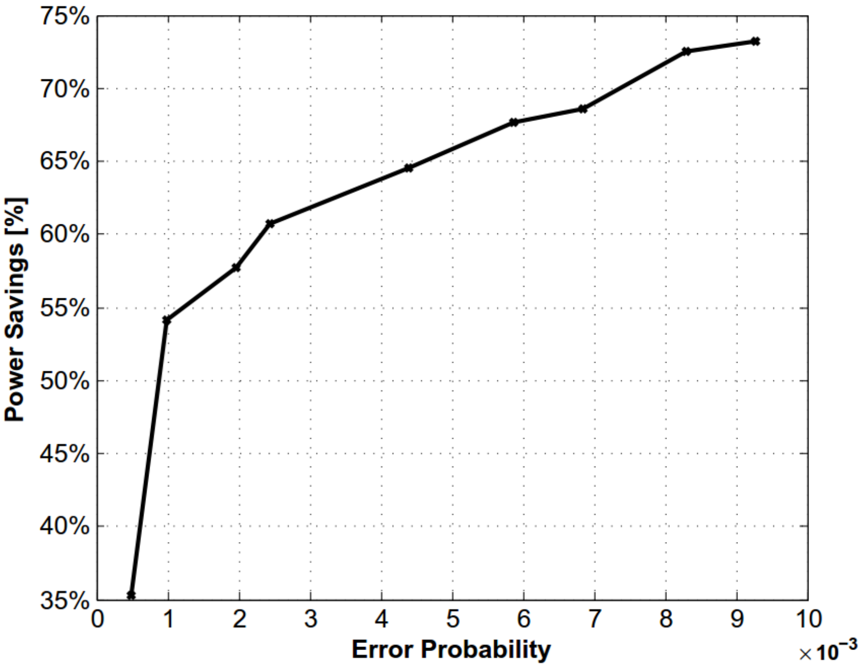
- Approximate storage can be useful only if the allowed number of errors are ensured to be low leading to minimum quality degradation
- The memory performance (retention time) need to scale gracefully rather than abruptly such that tolerating few errors lead to large savings
- Potential for shaping the distribution and achieving graceful degradation by using circuit level techniques (e.g. body biasing)

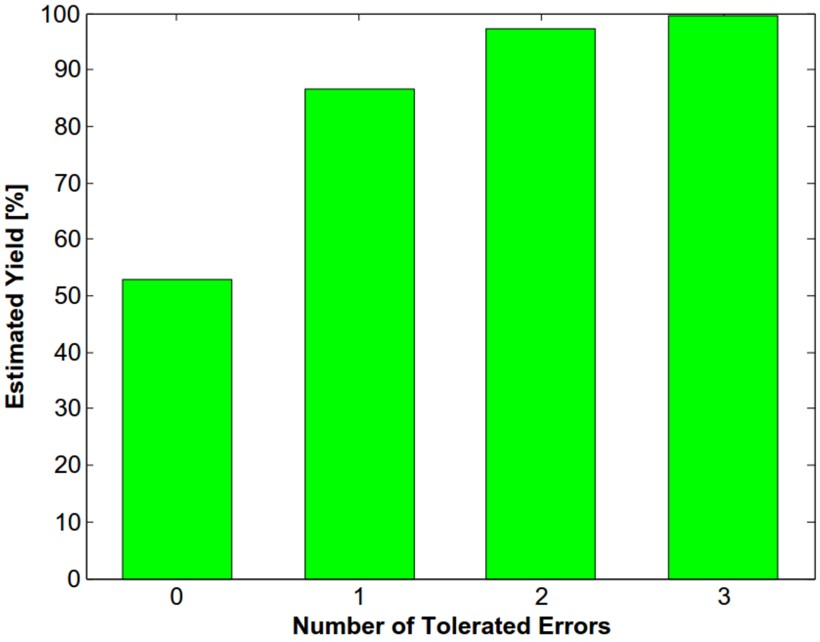




Power Savings and Yield Enhancement

- Utilizing a paradigm shift to approximate storage can lead to
  - Power savings by allowing less frequent refresh cycles, allowing the few resulting errors to be tolerated by the application
  - Yield enhancement by not discarding the dies that do not meet the minimum retention time and have few errors





Outline

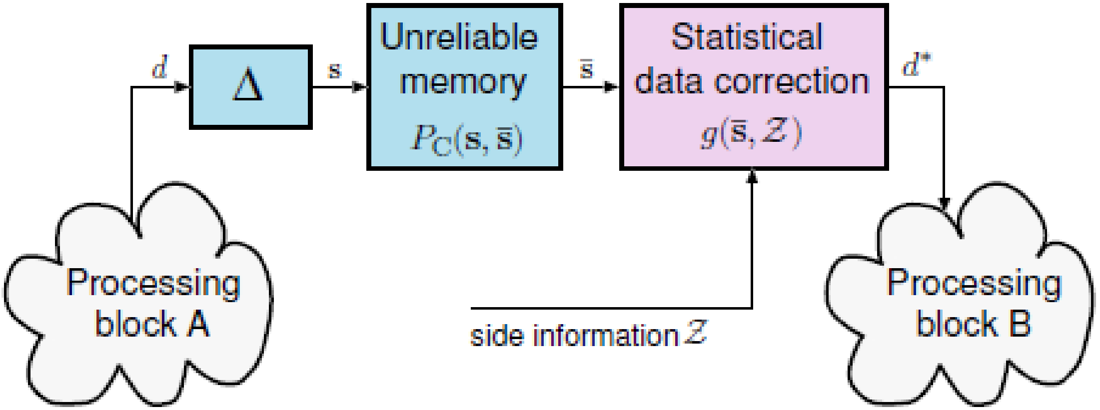
- ❑ **Potential for Energy Savings in DRAM**
  - Analysis of the retention time variability and conventional robust design
  - Energy savings by relaxing the worst-case and error-free requirements
  - Achieving graceful degradation for enabling and promoting approximate storage

- ❑ **Alternative Error Mitigation Mechanisms**
  - A Statistical correction scheme
  - Application to communication systems
  - Error Mitigation through But-Shuffling

- ❑ **Conclusion**

A Statistical Correction Scheme

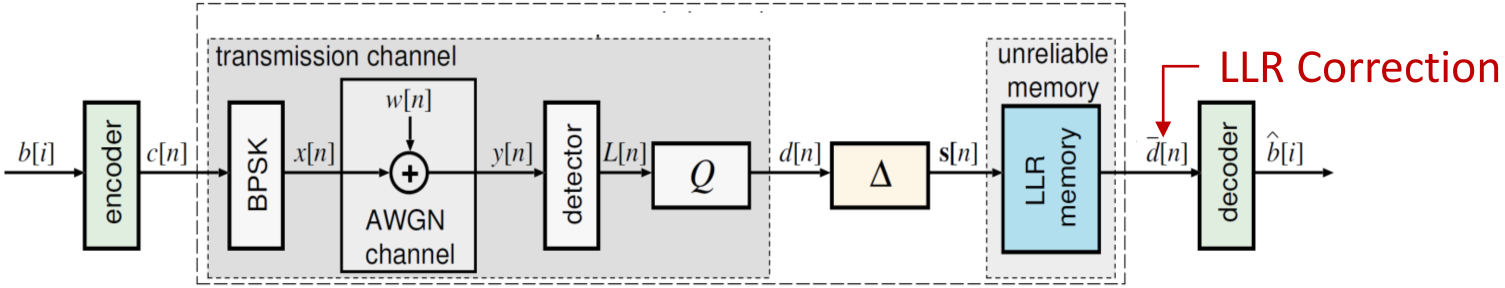
- Individual, single bit-flips can cause errors with very high magnitude
- **Traditional** schemes target the detection and correction of every single fault
- **Approximate Paradigm**: Graceful performance degradation
  - Requires confinement of errors (not necessarily correction)
- **Main idea:**
  - Detect errors (e.g., single-error detecting codes or sense amplifiers with marginal-read detection)
  - Substitute erroneous data with a “good estimate” -> based on data statistics



Examples for side information:

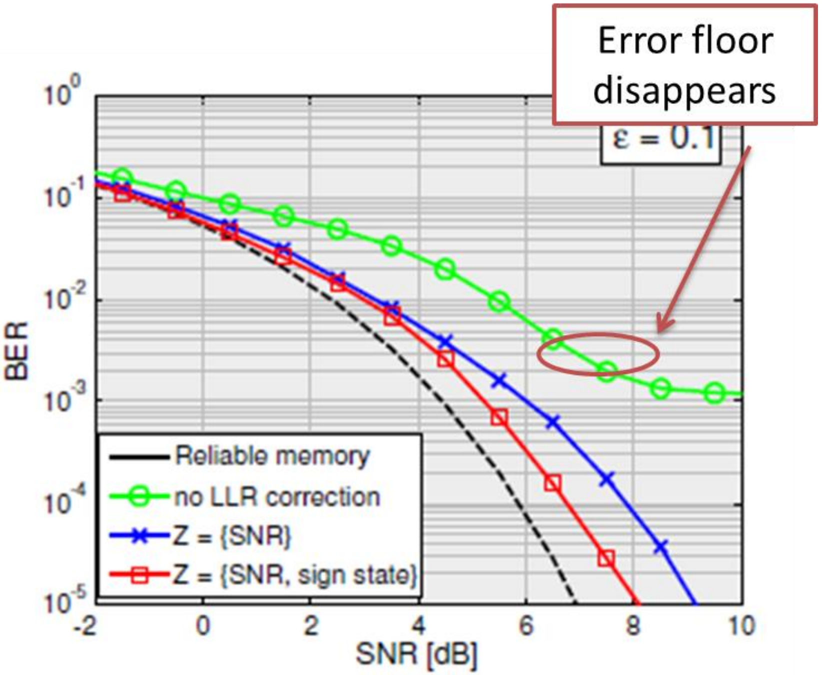
- Signal: mean, variance, PDF
- Hardware: basic ECC for error detection, tracking access time/retention time

Statistical Error Correction: Application to Communication Systems



LLR Correction

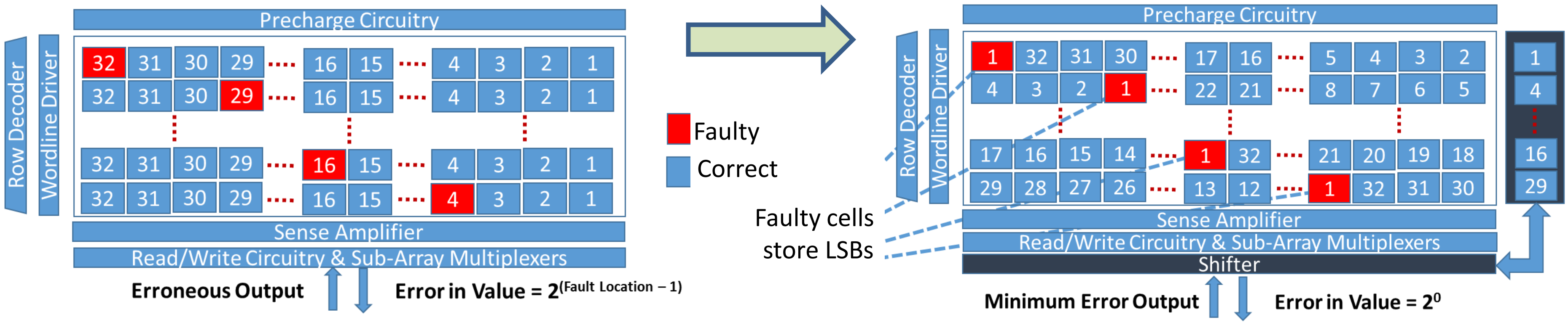
- ❑ Example: A coded communication system with 10% errors in the memory that stores the LLRs (reliability indicators)
- ❑ Faulty LLRs are corrected during read based on estimated mean values
- ❑ Two pieces of side information
  - ▪ Receive SNR (channel conditions)
  - ▪ Marginal-read for bit-cells containing the sign bit
- ❑ BER improves significantly
- ❑ The overhead of classical ECC can be reduced by 28% in a 9.6Kb array



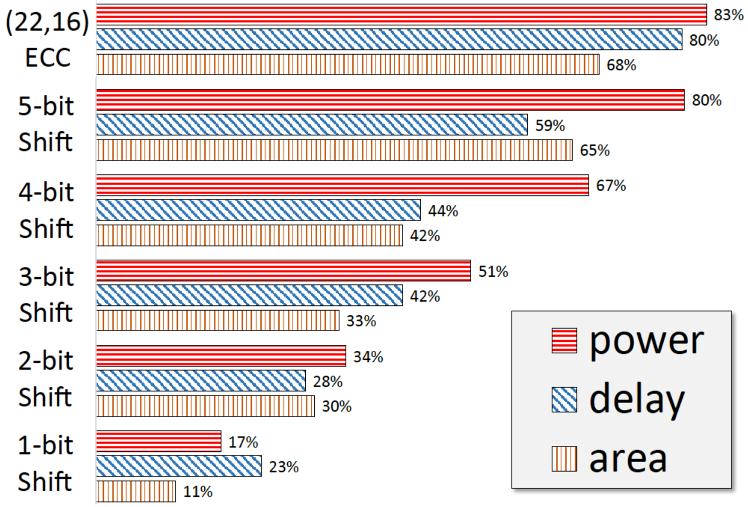
Error floor disappears

Error Mitigation through Bit-Shuffling

- ❑ **Main Idea:** Identify failing bit locations during runtime and store bits of lower significance (LSB) in those locations by shifting appropriately the bits



- ❑ Up-to 83% power, 89% area, 77% performance savings in 28nm vs a (39,32) SECDED ECC



- ❑ For 3 evaluated applications (Elasticnet, PCA and KNN) we observed 10%, 0.2% and 7% error in the output quality compared to the fault-free cases

Conclusion

- ❑ Application error resilience can be exploited in memories for limiting the overheads of traditional fault tolerant mechanisms

- ❑ Relaxing the worst case retention time assumptions and the error free requirements in DRAMs can lead to significant energy savings

- ❑ The benefits of approximate computing can increase by ensuring graceful quality/performance degradation

  => In DRAM the retention time distribution can be shaped appropriately through known circuit techniques such as body bias

  => The impact of allowed errors can be minimized through low cost error mitigation mechanisms that can exploit the statistical properties of various applications and help save considerable power

Thank you !

FP7 FET-OPEN SCoRPiO
http://www.scorpio-project.eu

FP7 Marie-Curie Fellowship Grant DARE
https://sites.google.com/site/daresystemeu