nano-tera.ch
swiss scientific initiative in health / security / environment systems

ICYSOC RTD 2013

FNSNF

Approximate Computing with Unreliable Dynamic Memories

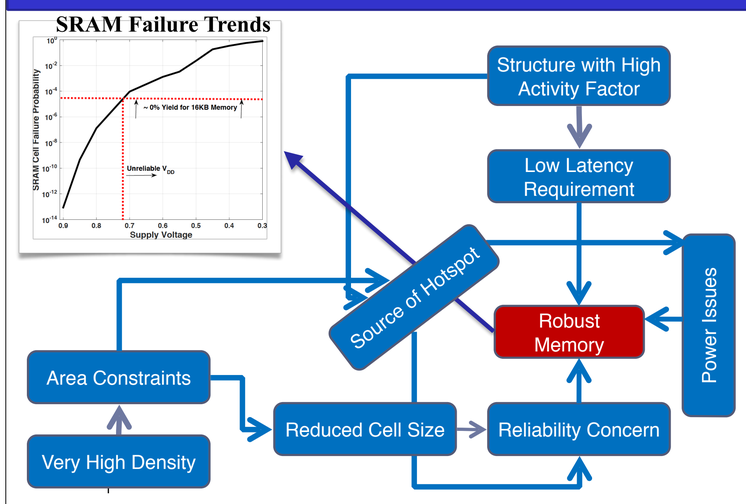
S.Ganapathy*, A.Teman*, R.Giterman†, A.Burg*, and G Karakonstantis‡

*Telecommunications Circuits Lab (TCL), Ecole Polytechnique Fédérale de Lausanne, Switzerland

†ENICS Lab, Faculty of Engineering, Bar Ilan University, Israel

‡High Performance and Distributed Computing, Queen's University Belfast, United Kingdom

ROBUST SRAM MEMORY DESIGN CHALLENGES



High Memory Density

- Large amount of on-chip memory required for enhancing system performance
- Reduced SRAM cell dimensions for higher density causes increase in parametric variations

Low Vdd Operation

- Tremendous power reduction
- Reduced Vdd lowering SRAM cell noise margin increasing susceptibility to failures

Low-Latency Access

- Faster access improves system performance
- Requires reduced timing margins increasing susceptibility to timing failures

GAIN-CELL (GC) BASED MEMORY DESIGNS

Comparison of Various Embedded Memory Options

	1TIC eDRAM	6T SRAM	Gain Cell eDRAM
Low-V_{dd} Margin	Poor	Poor (ratioed)	Good (ratioed, gain function)
Cell cap	<10 fF	Irrelevant	~ 1fF
Cell Size	$15F^2$	$130F^2$	$65F^2$
Process	Trench cap	Logic Compatible	Logic Compatible
Access	2 ns	< 1ns	2 ns

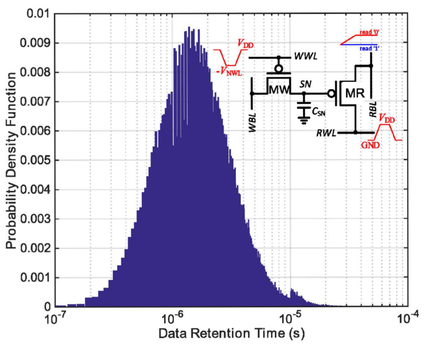
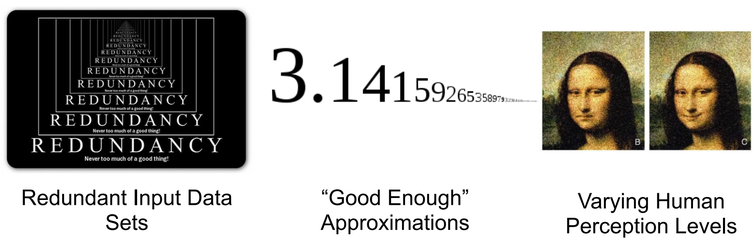


Figure: Probability Density Function of Data Retention Time of 2T GC-eDRAM Bitcell in a 28 nm FD-SOI process. The plot was extracted for 100k Monte Carlo samples at 25C.(inset): Schematic and typical operational waveforms of the 2T Gain Cell.

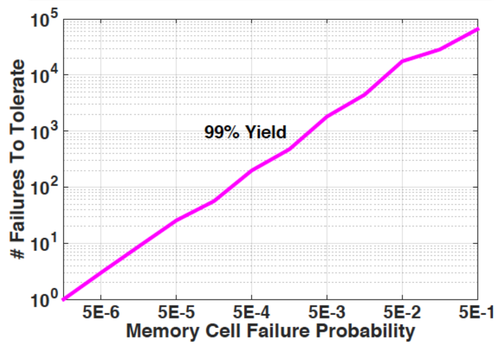
MEMORIES FOR APPROXIMATE COMPUTING PARADIGM

- Exploiting the inherent error-resilience of certain class of applications to trade-off quality of result for power and performance
- Allows for imprecise computations as long as quality is acceptable

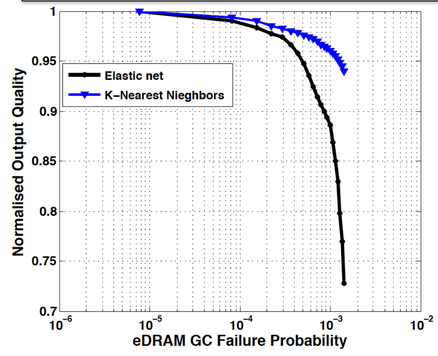
Tolerable Imprecision Exploited by:



- Relax reliability requirements of memories and provide only sufficient amount of protection based on application requirements
- Tolerating failures also helps improve memory parametric yield



- Impact of memory failures studied for 2 machine learning benchmarks - K-Nearest Neighbours (KNN) and ElasticNet.
- Measured R^2 and Score for ElasticNet and KNN respectively.



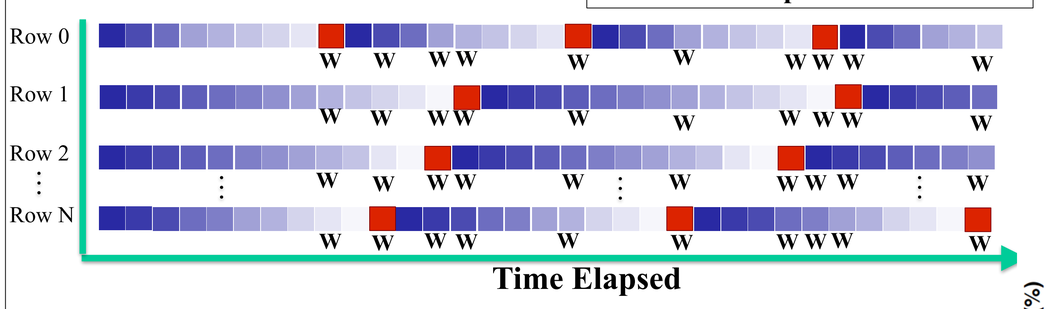
- Output quality normalised to value obtained when executing benchmarks on fault-free memory
- Normalised **quality 97%** and **90%** for KNN and ElasticNet for failure probability - 1E-3
- Tolerating **177 failures** in a 16kB memory provides "good-enough" quality

MINIMUM REFRESH GC eDRAM FOR APPROXIMATE COMPUTING

eDRAM Refresh Strategies

Perfect Non-Conflicting Line-Level Refresh Scheme

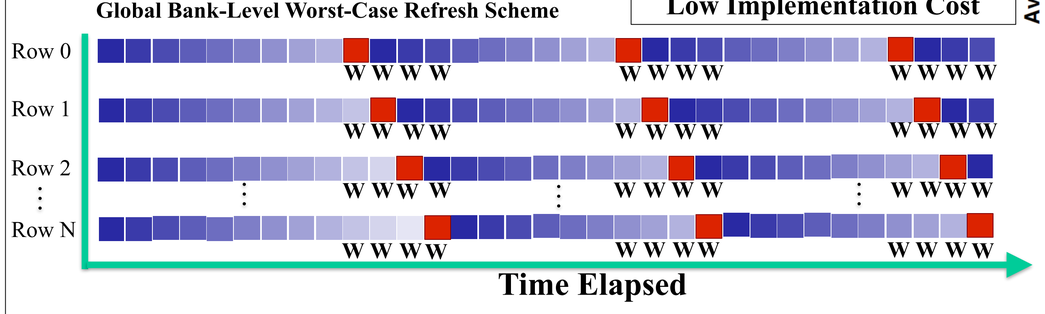
Maximum Availability
Minimum Refresh Power Overhead
Maximum Implementation Cost



Refresh Operation **W**-Wait ! Memory Unavailable

Global Bank-Level Worst-Case Refresh Scheme

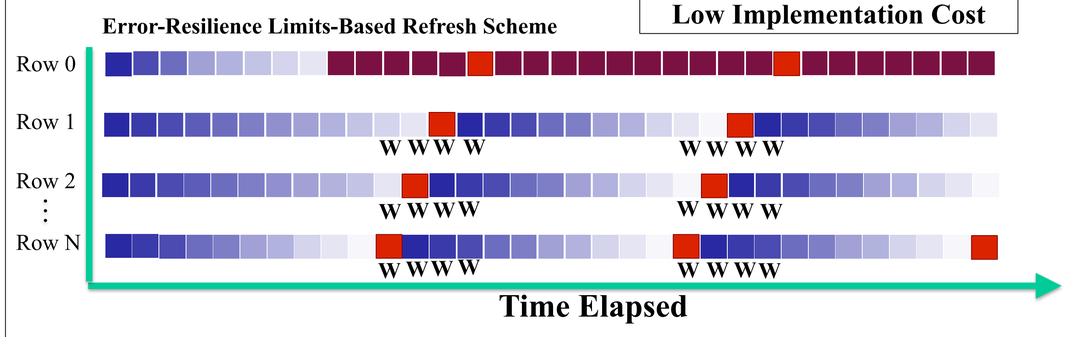
Low Availability
High Refresh Power Overhead
Low Implementation Cost



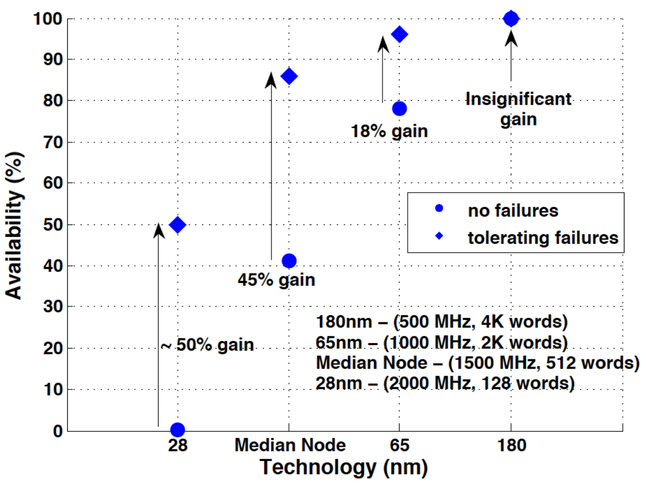
Some Cells Failed Before Refresh

Error-Resilience Limits-Based Refresh Scheme

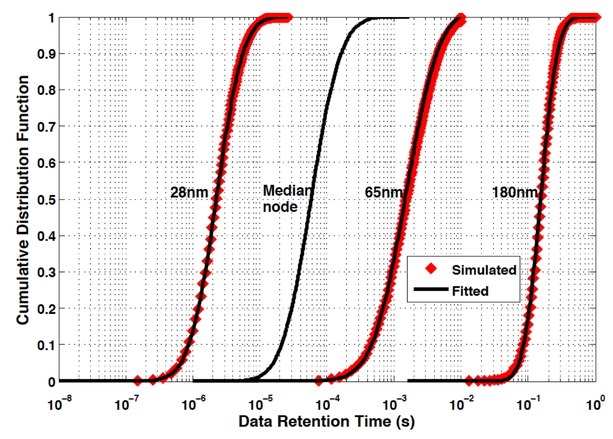
High Availability
Low Refresh Power Overhead
Low Implementation Cost



- eDRAM Refresh Rate selected based on acceptable failure rate estimated for a given Data Retention Time (DRT).
- Median node DRT estimated by fitting experimental data approximated to lognormal distributions



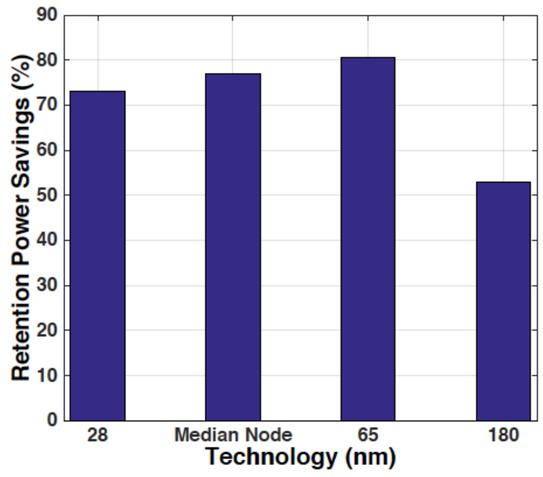
Memory availability of GC-eDRAM arrays in various technology nodes: 180 nm CMOS, 65 nm CMOS, and 28 nm FD-SOI, and an estimated median node. Availability is shown for traditional worst-case design ("No Failures") and for error-tolerance of Perr < 10^{-3} for memory bank sizes and operating frequencies typical for each node.



Cumulative Distribution Functions of the data retention times of 2T GC-eDRAM bitcells for various process technologies: 180 nm CMOS, 65 nm CMOS, 28 nm FD-SOI, and a estimated median node.

- Availability does not have much variation at mature nodes compared to advanced nodes.
- By tolerating as much as 177 errors in 16kB memory, availability can be increases between **18% and 50%** when moving from **65nm to 28nm** node

- Since failures can be tolerated, the refresh rate is adjusted to satisfy specific yield criterion.
- For 65nm node, this potentially reduces the refresh power by as much as **5X** and for other technologies, there is a minimum **2X** savings in power.



Retention power savings during low activity periods for GC-eDRAM rrays at different nodes, operated with an error tolerance of Perr < 10^{-3}

"Variation Aware Performance Analysis of Gain Cell Embedded DRAMs" W.Zhang *et.al.* ISLPED'10

"Quality Programmable Vector Processors for Approximate Computing" S.Venkataramani *et.al.* MICRO'13

"Replica technique for adaptive refresh timing of gain-cell-embedded DRAM" A.Teman e*t.al.* TCS II'14

"Approximate Computing with Unreliable Dynamic Memories" S.Ganapathy *et.al.* NEWCAS'15

"Energy versus data integrity trade-offs in embedded high-density logic compatible dynamic memories" A.Teman *et.al.* DATE'15