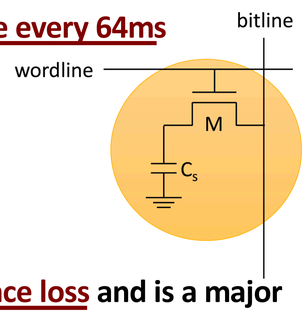
VRL-DRAM: Improving DRAM Performance via Variable Refresh Latency

Anup Das, Hasan Hassan and Onur Mutlu

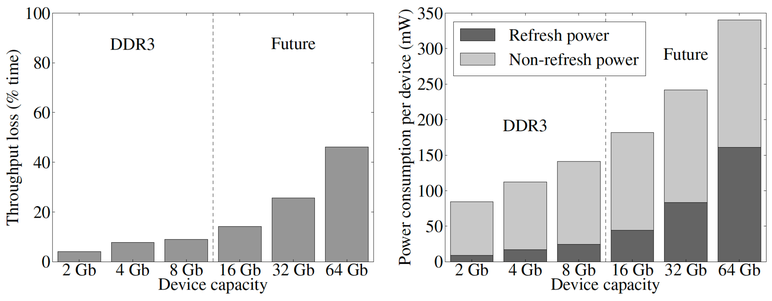
Drexel University ETH zürich

Background: DRAM Refresh

- A DRAM cell is composed of a capacitor and an access transistor
- A DRAM capacitor loses charge over time
- Needs periodic refresh e.g., once every 64ms



DRAM refresh causes performance loss and is a major source of power consumption

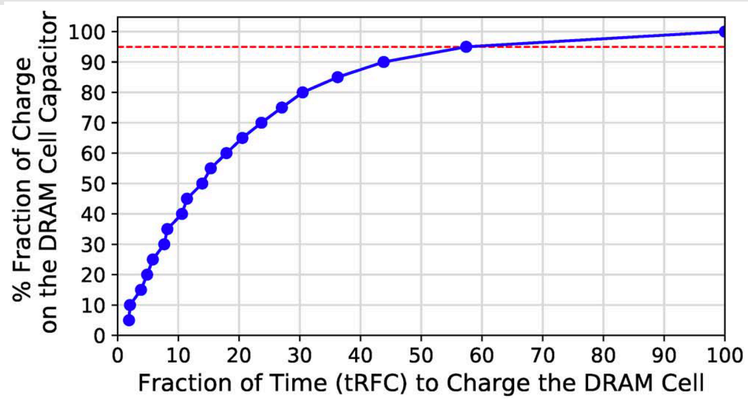


Source: Liu et al., RAIDR, ISCA 2012

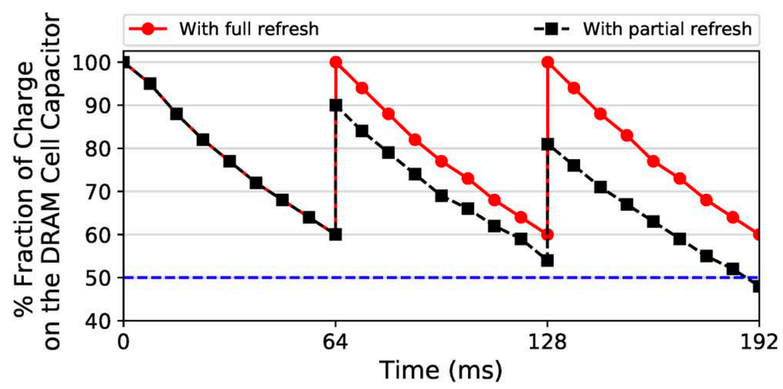
- Throughput loss due to DRAM refresh is nearly 50% at the 64 Gb DRAM density node
- Refresh power is more than 50% of the total memory power

Key Observations

1st key observation: Almost 50% of the refresh time is spent in injecting the last 5% of charge of a fully charged cell



2nd key observation: Once fully restored, a DRAM cell can sustain multiple partial refreshes without sacrificing data integrity



Goals

1. Understand refresh timings in DRAM through detailed circuit analysis
2. Analyze the potential of lowering refresh timing parameters using partial refresh operations
3. Investigate full and partial refresh scheduling to mitigate refresh overhead in DRAM

Key Idea

Use two timing parameters for DRAM refresh
- Long latency full refresh
- Short latency partial refresh

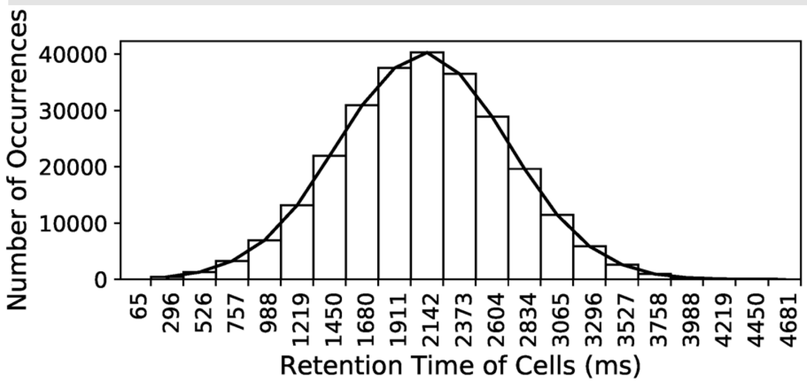
Variable Refresh Latency

Characterization to determine number of partial refreshes that a DRAM cell can reliably sustain

Schedule full refresh only when necessary, and otherwise ensure data integrity by issuing low-latency partial refresh operations.

DRAM Characterization

1st characterization: Retention characterization of DRAM



Retention time varies widely across DRAM cells

Refresh period (ms)	Number of rows in a bank
64	68
128	101
192	145
256	7878

Retention time of a DRAM row is the minimum of the retention time of any of the cells in the row

2nd characterization: Estimation of number of partial refreshes of DRAM rows that can be reliably sustained

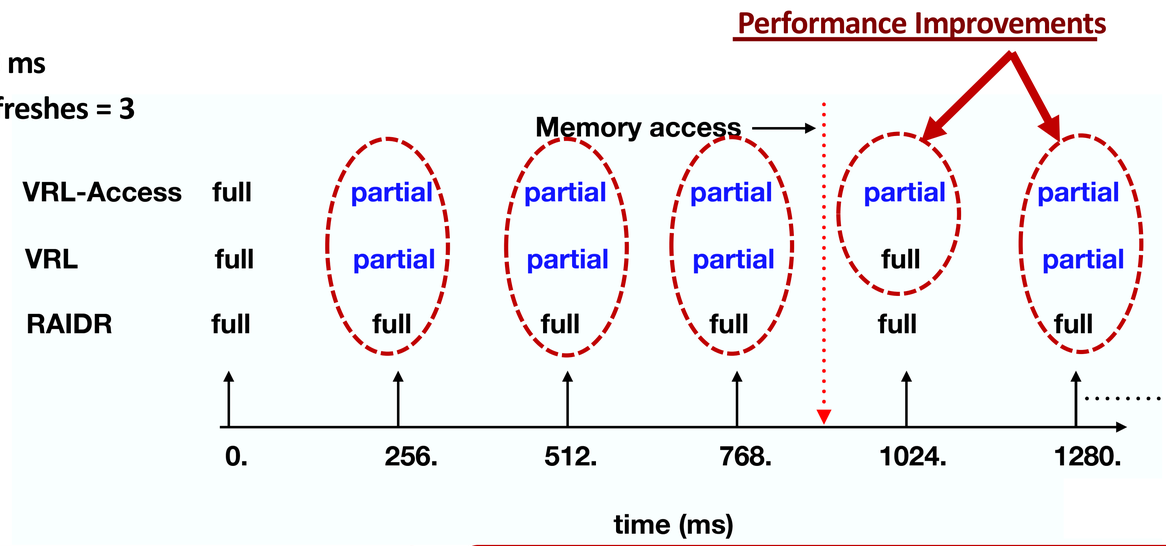
Row ID	Number of partial refreshes
0	3
1	1
. . .	. . .
127	5

Number of partial refreshes of a DRAM row is the minimum of the number of partial refreshes that can be sustained by all of the cells in the row

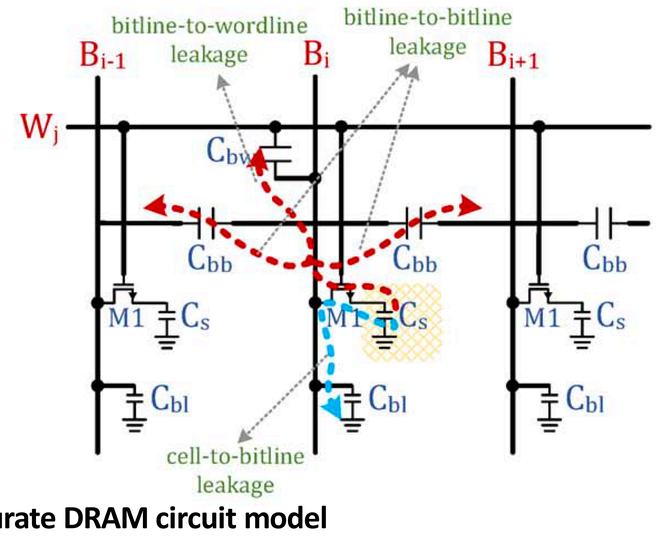
DRAM Refresh Scheduling

Baseline VRL: schedule partial and full refreshes to minimize refresh overhead
VRL-Access: use partial refresh in place of full refresh if there is a memory read or write access

Example: Row 0
Retention time = 256 ms
Numbet of partial refreshes = 3



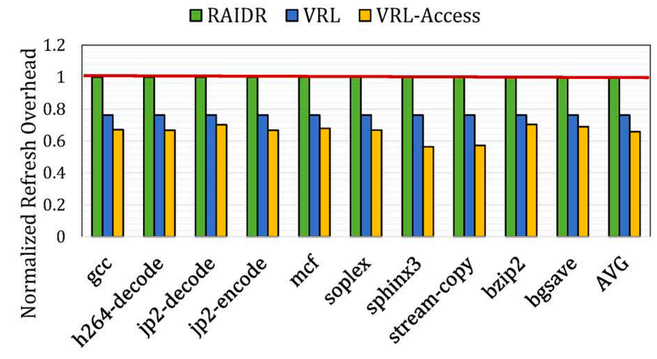
DRAM Analytical Model



Accurate DRAM circuit model
- Data pattern in DRAM
- Sneak paths
- Bitline/wordline parasitics

Open source tool
https://github.com/anupkdas-nus/VRL-DRAM

Evaluation



On average, refresh performance overhead reduces by 23% using VRL and 34% using VRL-Access

Bank Size	Pre-sensing time (cycles)			Simulation time		
	SPICE Sim.	Single cell Model	Our Model	SPICE Sim.	Single cell Model	Our Model
2048x32	7	6	7	1h 36m	5ms	10s
2048x128	8	6	8	1h 57m	5ms	51s
8192x32	9	6	9	4h 23m	21ms	20s
8192x128	11	6	10	5h 17m	21ms	108s
16384x32	14	6	12	17h 12m	44ms	90s
16384x128	16	6	14	21h 1m	44ms	209s

Analytical model accurate with respect to SPICE simulation with significant reduction of analysis time (over 10x)