Reducing DRAM Latency via Charge-Level-Aware Look-Ahead Partial Restoration

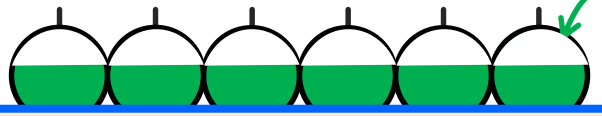
Yaohua Wang, Arash Tavakkol, Lois Orosa, Saugata Ghose,
Nika Mansouri Ghiasi, Minesh Patel, Jeremie S. Kim,
Hasan Hassan, Mohammad Sadrosadati, Onur Mutlu

SAFARI

Carnegie Mellon

Problem

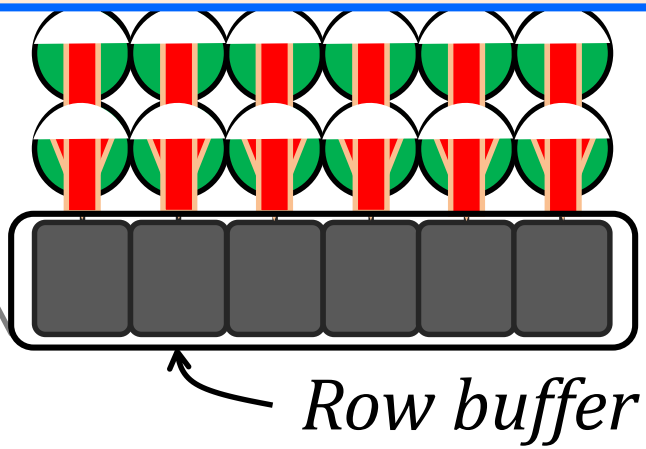
- DRAM access latency is a major bottleneck for system performance
- Fundamental operations when accessing DRAM



Restoration latency takes up to 43.6% of DRAM access latency

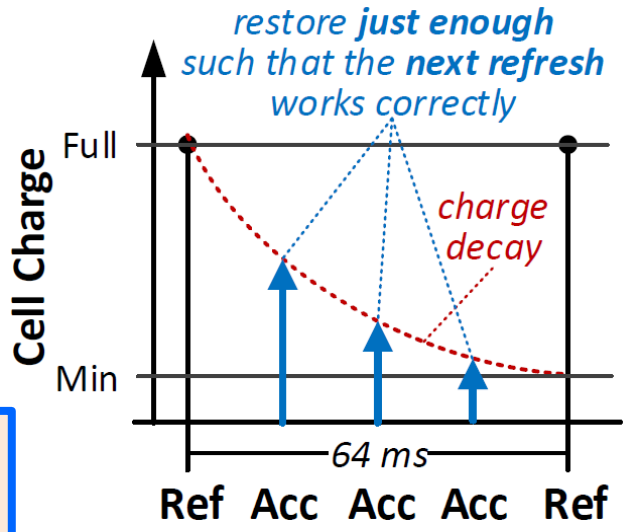
Restoration

Refresh



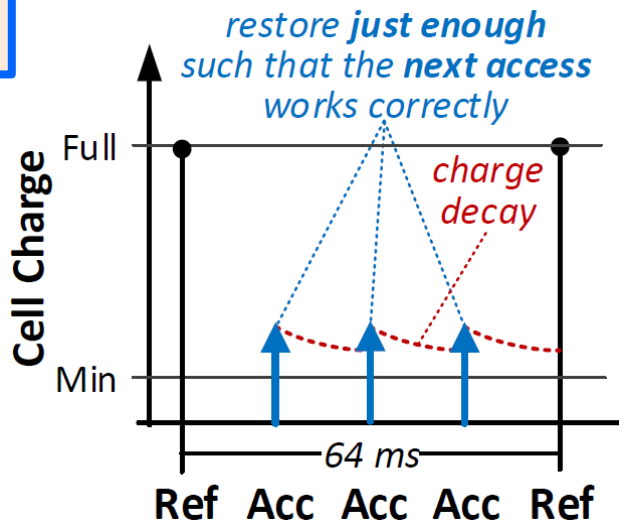
Motivation

- Prior work applies partial restoration to soon-to-be-refreshed DRAM rows



Observation:
a recently-accessed row is likely to be accessed again soon

- Partial restoration can be applied to soon-to-be-reactivated DRAM rows



Our Proposal

- We propose charge-level-aware look-ahead partial restoration (CAL)

1. CAL predicts the next access time at high accuracy (i.e., 98%)
2. CAL applies partial restoration
   a) based on the predicted next access time
   b) ensuring high enough restoration level
   c) maintaining the benefits of latency reduction mechanisms for highly-charged rows

Results

- We comprehensively evaluate CAL using a wide variety of workloads and across many system and mechanism parameters

14.7% performance improvement
11.3% energy reduction

- CAL is implemented fully within the memory controller without any changes to the DRAM module

Reducing DRAM Latency via Charge-Level-Aware Look-Ahead Partial Restoration

Yaohua Wang, Arash Tavakkol, Lois Orosa, Saugata Ghose,
Nika Mansouri Ghiasi, Minesh Patel, Jeremie S. Kim,
Hasan Hassan, Mohammad Sadrosadati, Onur Mutlu

SAFARI

Carnegie Mellon

ETH zürich

UNICAMP

Session 3-A, Oct 22