FIGARO: Improving System Performance via Fine-Grained In-DRAM Data Relocation and Caching

Yaohua Wang[1], **Lois Orosa**[2], Xiangjun Peng[3,1], Yang Guo[1], Saugata Ghose[4,5], Minesh Patel[2], Jeremie S. Kim[2], Juan Gómez Luna[2], Mohammad Sadrosadati[6], Nika Mansouri Ghiasi[2], Onur Mutlu[2,5]

1 National University of Defense Technology

2 ETH zürich

3 香港中文大學 The Chinese University of Hong Kong

4 University of Illinois Urbana-Champaign

5 Carnegie Mellon University

6 IPM Institute for Research in Fundamental Sciences

SAFARI

MICRO 2020

Motivation and Goal

- **Problem:** DRAM latency is a performance bottleneck for many applications

- In-DRAM caches mitigate this latency
  - by augmenting regular-latency DRAM with small-but-fast regions of DRAM that serve as a cache

- **Existing in-DRAM** caches have mechanisms for relocating data that have two main **inefficiencies**:
  1) Coarse-grained (i.e., multi-kilobyte) in-DRAM data relocation
  2) Relocation latency increases with the physical distance between the slow and fast regions

- **Goal:** reduce DRAM latency via an in-DRAM cache that provides
  1) Fine-grained (i.e., multi-byte) data relocation
  2) Distance-independent relocation latency

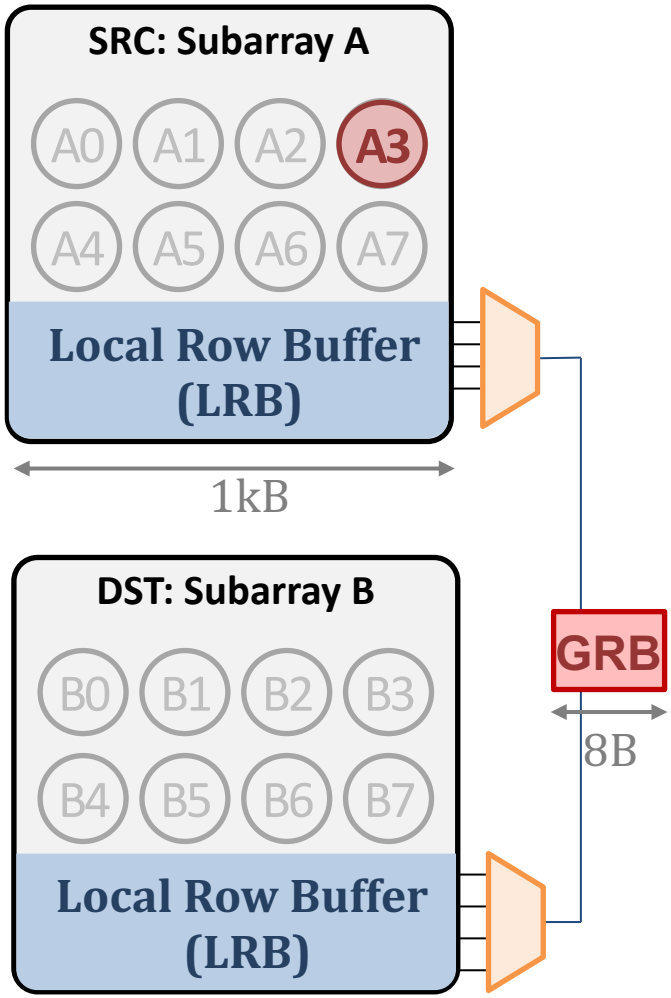
FIGARO Substrate

- **FIGARO** leverages existing shared structures within a modern DRAM device to perform data relocation
- **Observations:**

  1) All local row buffers (LRBs) in a bank are connected to a single shared global row buffer (GRB)

  2) The GRB has smaller width (e.g., 8B) than the LRBs (e.g., 1kB)

- **Key Idea:** use the existing shared GRB among subarrays within a DRAM bank to perform fine-grained in-DRAM data relocation



FIGCache (Fine-Grained In-DRAM Cache)

- **Key idea:** cache only small, frequently-accessed portions of different DRAM rows in a designated region of DRAM

- FIGCache uses FIGARO to relocate data into and out of the cache at fine granularity

- **Results:**
  - Improves system performance by 16.3% on average
  - Reduces DRAM energy by 7.8% on average
  - Outperforms a state-of-the-art coarse-grained in-DRAM cache
  - Performs close to ideal low-latency DRAM

FIGARO: Improving System Performance via Fine-Grained In-DRAM Data Relocation and Caching

Yaohua Wang[1], **Lois Orosa**[2], Xiangjun Peng[3,1], Yang Guo[1], Saugata Ghose[4,5], Minesh Patel[2], Jeremie S. Kim[2], Juan Gómez Luna[2], Mohammad Sadrosadati[6], Nika Mansouri Ghiasi[2], Onur Mutlu[2,5]

[1] [2] [3] [4] [5] [6]

SAFARI

MICRO 2020