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

Yaohua Wang¹, Lois Orosa², Xiangjun Peng³,¹, Yang Guo¹, Saugata Ghose⁴,⁵, Minesh Patel², Jeremie S. Kim², Juan Gómez Luna², Mohammad Sadrosadati⁶, Nika Mansouri Ghiasi², Onur Mutlu²,⁵

1. 华中师范大学
2. ETH Zürich
3. 香港中文大学
4. University of Illinois
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 Wang¹, Lois Orosa², Xiangjun Peng³,¹, Yang Guo¹, Saugata Ghose⁴,⁵, Minesh Patel², Jeremie S. Kim², Juan Gómez Luna², Mohammad Sadrosadati⁶, Nika Mansouri Ghiasi², Onur Mutlu²,⁵

¹ Tsinghua University ⁵ The Chinese University of Hong Kong
² ETH Zürich ⁶ IPM
³ Carnegie Mellon University ⁷ IIT
⁴ University of Illinois Urbana-Champaign

MICRO 2020