A Case for Subarray-Level Parallelism (SALP) in DRAM
Yoongu Kim, Vivek Seshadri, Donghyuk Lee, Jamie Liu, Onur Mutlu (CMU)

1. DRAM Bank Conflicts

![Diagram showing DRAM bank conflicts and serialization of requests]

2. Timeline of DRAM Bank Conflicts

- Four requests to the same DRAM bank
- WR WR RD RD
- Large latency due to 3 problems:
  1. Serialization of requests
  2. Write penalty after WR request
  3. “Thrashing” of row-buffer

3. Our Goal

- Goal: Cost-effectively mitigate the detrimental effects of bank conflicts
- Naïve Solution: Simply add more banks
  - Very expensive

4. Two Key Observations

1. A DRAM bank is divided into subarrays
   - Each subarray has a local row-buffer

2. Subarrays are mostly independent...
   - Except when sharing global structures

5. Key Idea

- Reduce the sharing of...
  1. Global decoder: enable parallel access to multiple subarrays
  2. Global row-buffer: utilize multiple local row-buffers concurrently

6. Mechanism: MASA

- Multitude of Activated Subarrays
- Add two latches to each subarray

- Baseline vs. MASA performance metrics:
  - Normalized Dynamic Energy
  - Row-Buffer Hit-Rate
  - IPC Increase
  - Die-Size

- SALP-1 vs. SALP-2 vs. MASA vs. "Ideal"