CAMELab.org
Computer Architecture and Memory Systems Lab.

UT DALLAS

Half-DRAM: a High-bandwidth and Low-power DRAM Architecture from the Rethinking of Fine-grained Activation

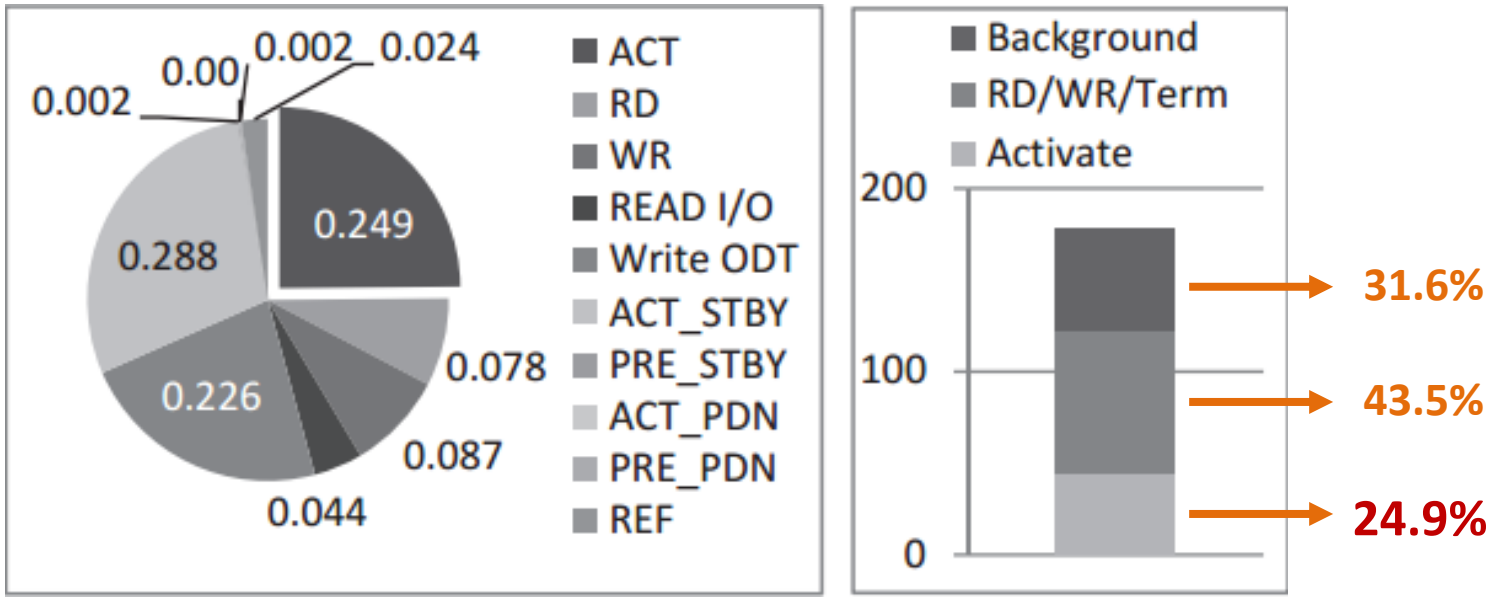
Tao Zhang, Ke Chen, Cong Xu, Guangyu Sun,
Tao Wang, Yuan Xie
Pennsylvania State University

ISCA - 2014

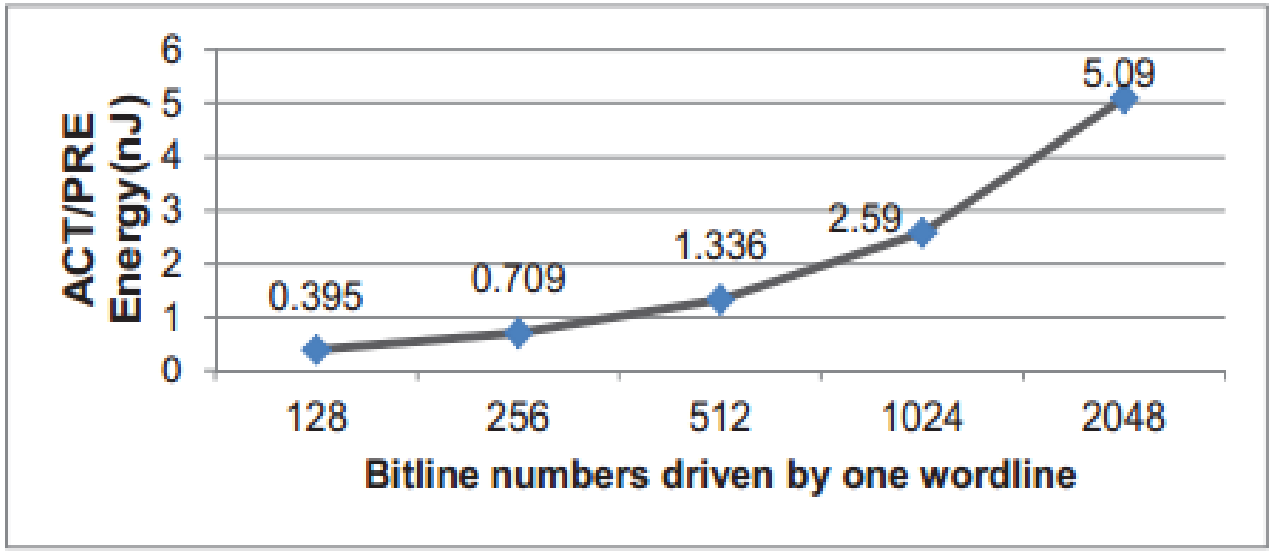
Motivation

- DRAM can consume more than 25% of the total power in a datacenter
  - => improving the power efficiency of DRAM one of the major challenges in the memory architecture design

- Activation and Precharge power can be around 25% of the total DRAM power



Motivation

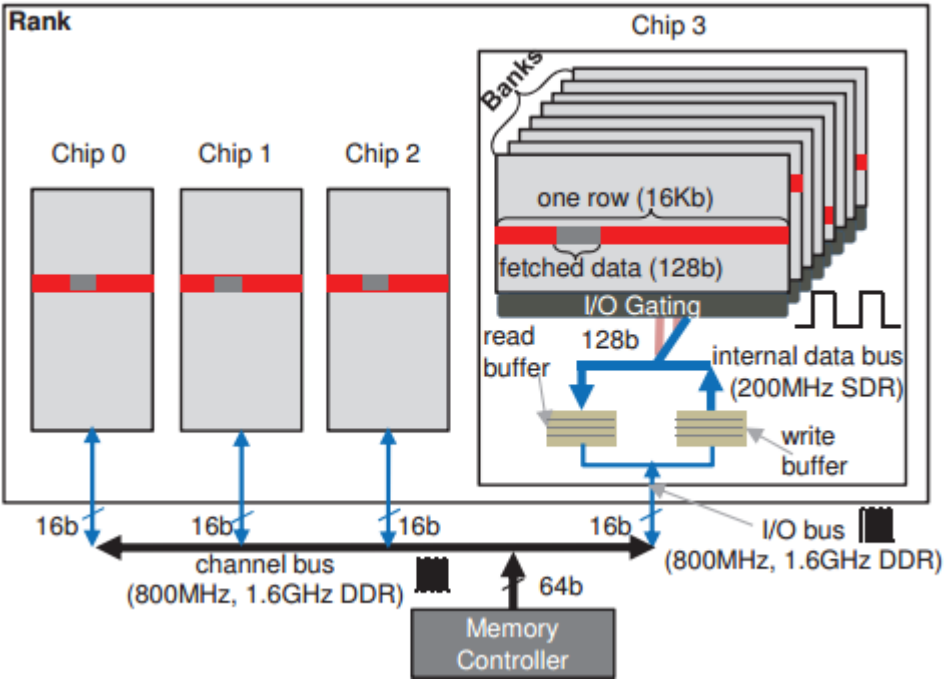


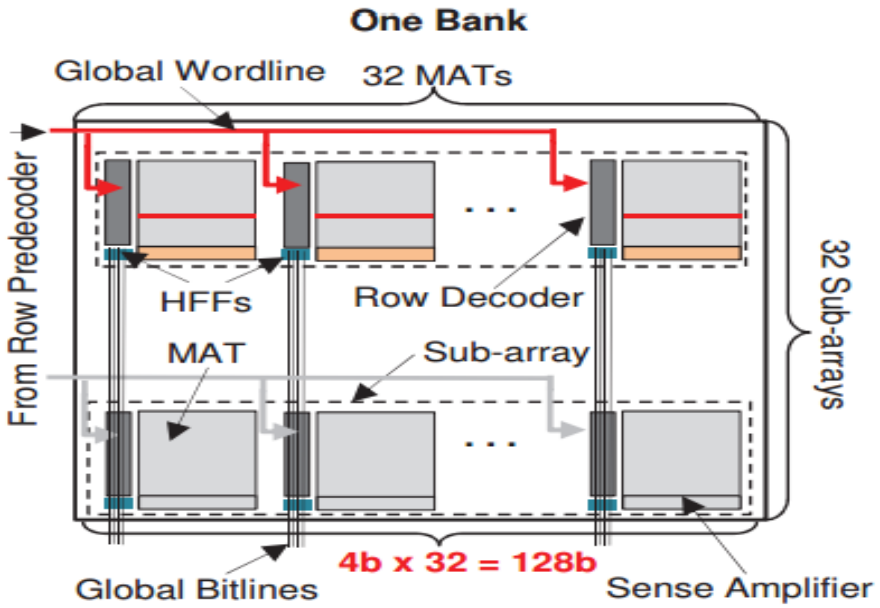
The activation power is *proportional* to the number of bitlines being activated during a memory access

- **For future memory chips with larger capacity and more bitlines, the power efficiency problem of row activation will become even worse**

The goal of this work is to minimize the Activation Power

DRAM Row Access and Activation Power



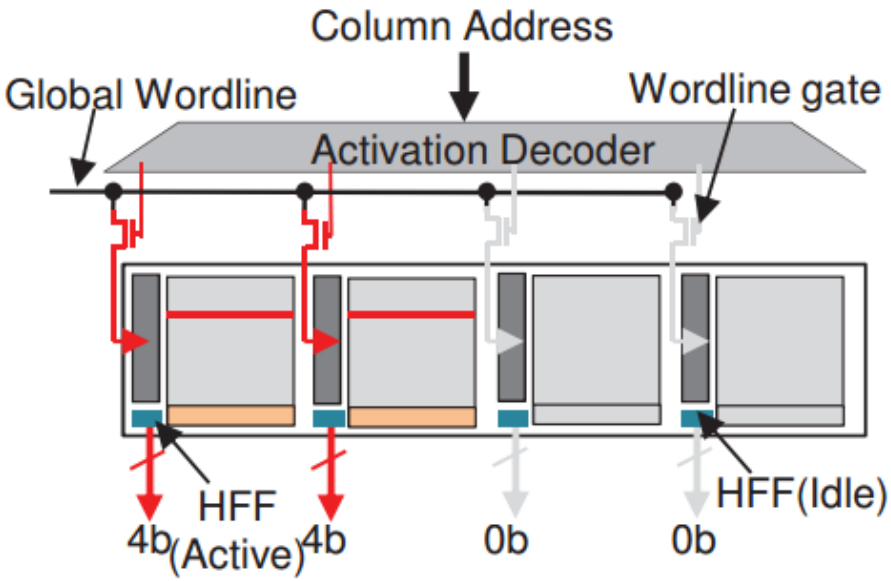


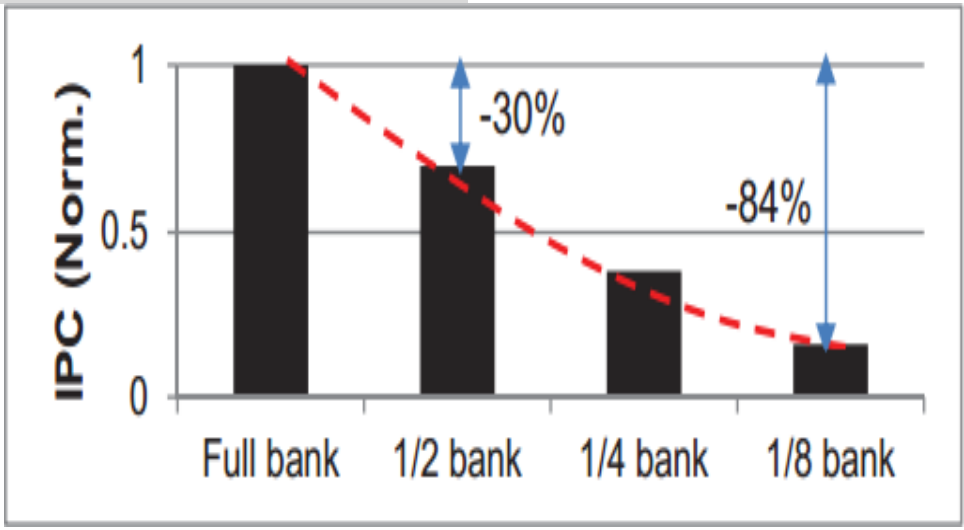
Row Overfetching: Entire row to be activated (red block) even though only a small portion of data are fetched at a time (gray block)

$$I_{ACT} = IDD0 - \frac{IDD3N \times tRAS + IDD2N \times (tRC - tRAS)}{tRC}$$

Reducing activated row size (i.e., number of bitlines) can significantly reduce the activation power

Prior Fine-Granule Row Access and the BW dilemma





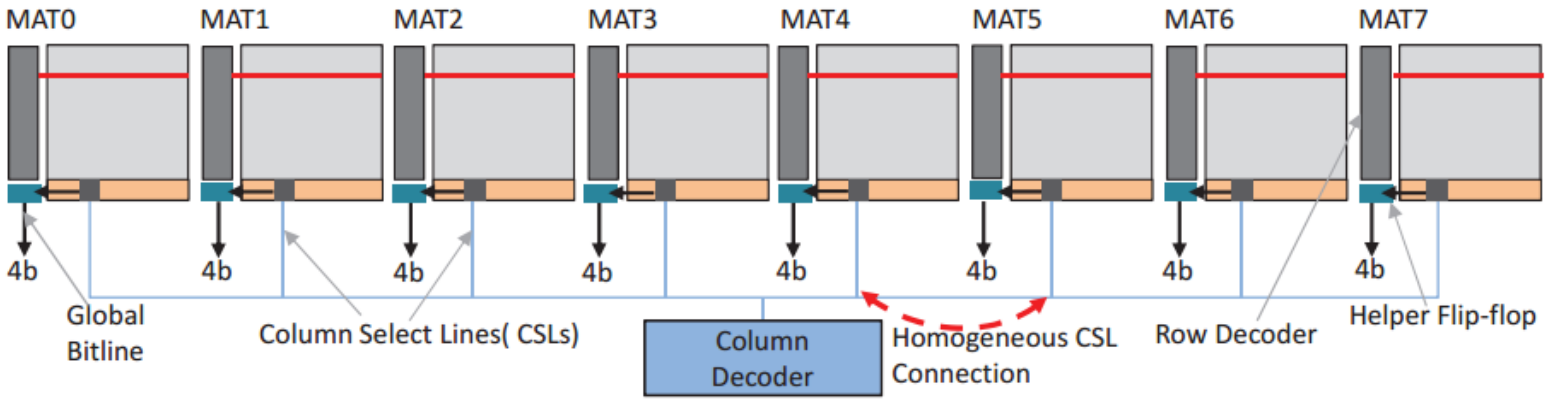
HFF: Helper Flip-Flop > latches the selected data and relay them on the internal data bus

- **An activation decoder is introduced to control the number of active MATs**

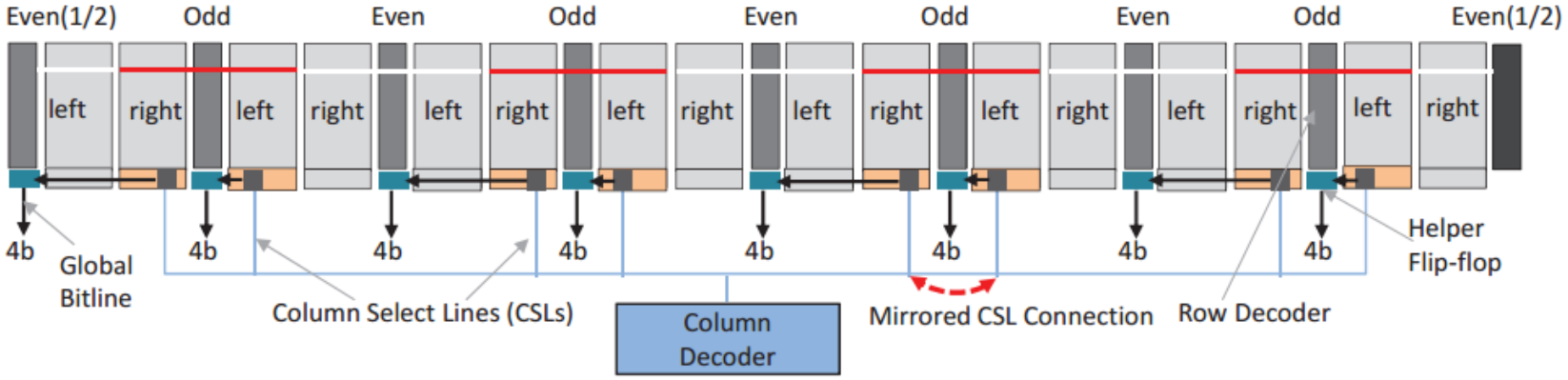
However… $\mathbf{B}_{DataBandwidth} = \mathbf{W}_{DataWidth} \times \mathbf{F}_{DataFrequency}$

Fine-grained activation techniques can result in significant memory performance degradation

Half-DRAM



(a) 1RD-1HFF: traditional structure



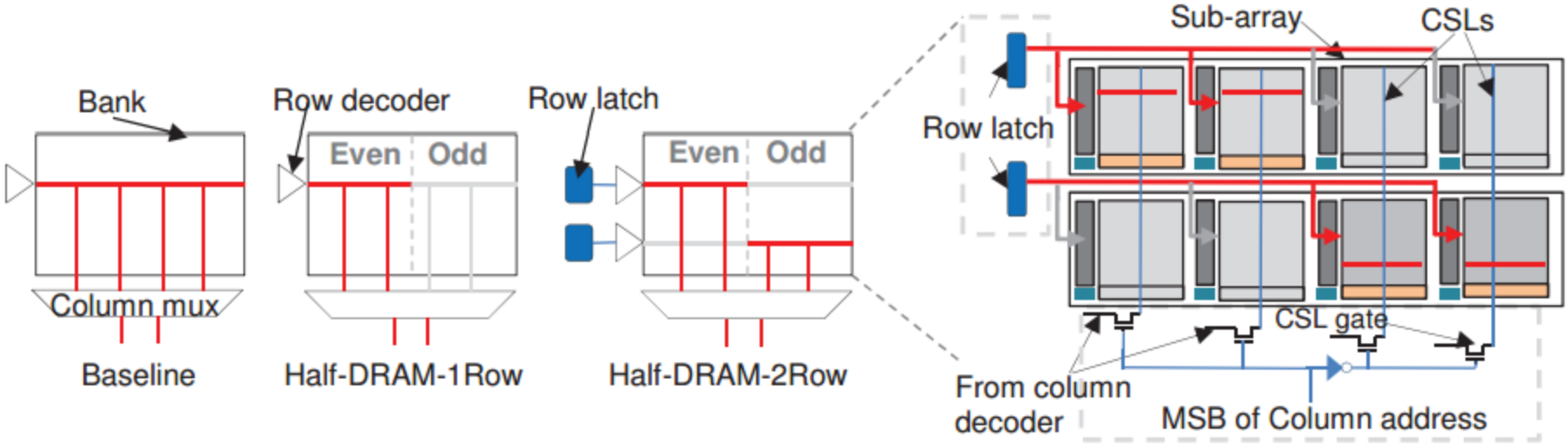
(b) 1RD-2HFF: Half-DRAM structure (proposed)

Traditional DRAM: One-to-one relationship exists between row decoder and HFF.

1RD-1HFF: One HFF group is dedicated to one row decoder and vice versa

Half-DRAM: MAT is split into "left" and "right" block - driven by different row address decoders. The sub-array is further divided into Odd and Even groups

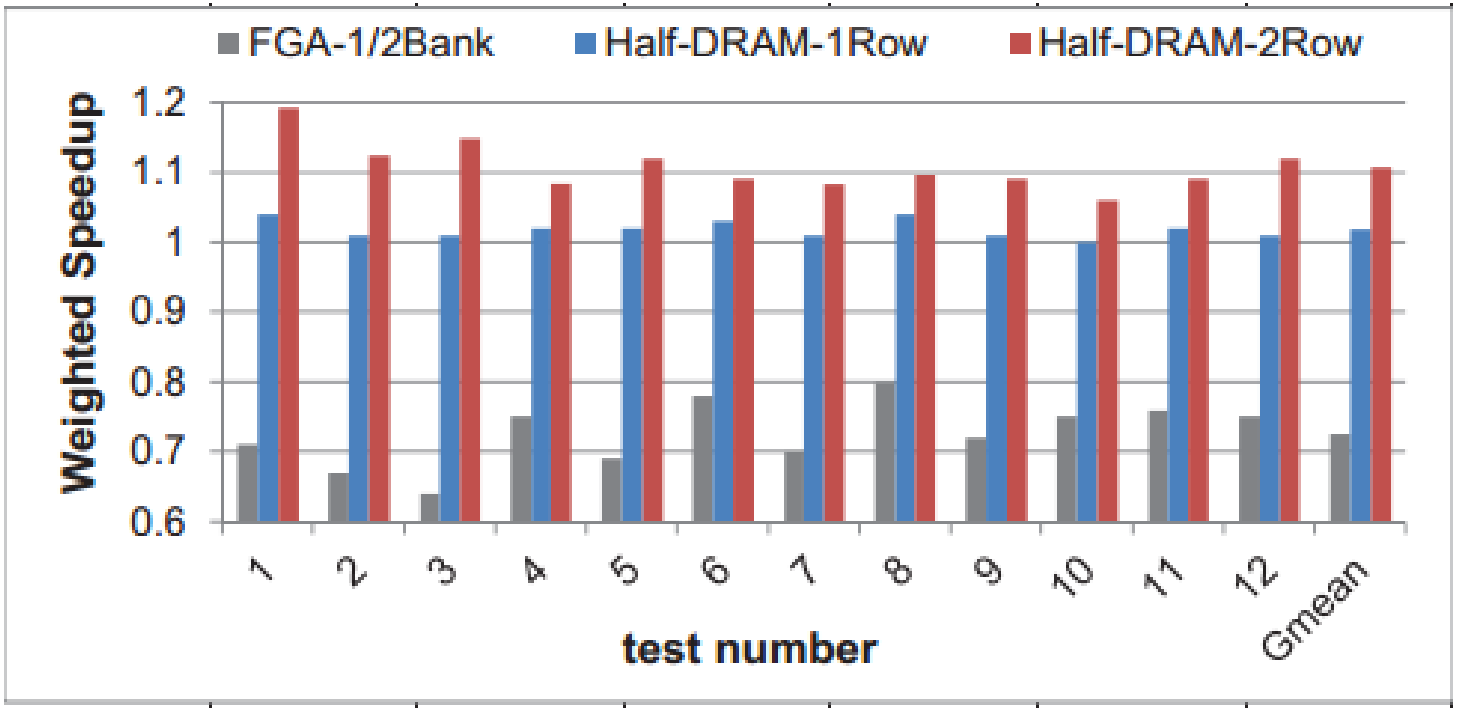
Half-DRAM



- Here, once a row decoder selects a wordline, both MATs are activated but each with half a row.
- Even if every other row decoder is disabled, all HFFs can still be active

- Though Half-DRAM mitigates the bandwidth reduction problem, it only has half of the row buffer size, which may degrade row buffer hit rate.
- Half-DRAM can be extended to the traditional full bank activation to provide full row buffer size …. *Half-DRAM-2Row*

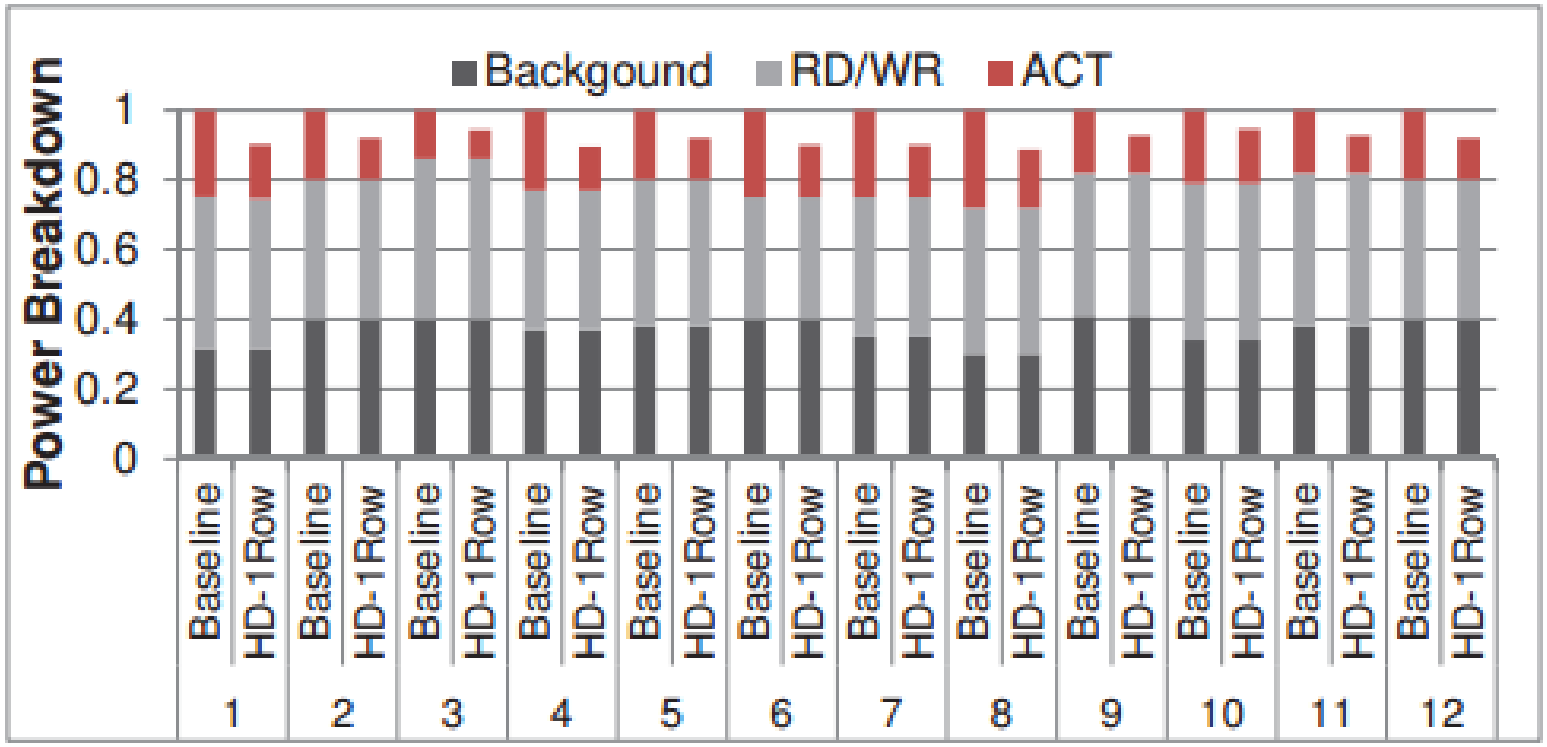
Performance Evaluation



Avg. performance improvement in Half-DRAM-2Row is 10.7%

test# **Benchmarks (SPEC2006+STRAM)**
[1]STREAM×4, [2]bwaves×4, [3]gobmk×4, [4]leslie3d×4, [5]libquantum×4,, [6]lbm×4, [7]mcf×4, [8]milc×4, [9]STREAM-gobmk-lbm-libquantum, [10]bwaves-leslie3d-mcf-milc, [11]lbm-libquantum-bwaves-leslie3d, [12]STREAM-gobmk-mcf-milc

Power Evaluation



On average, Half-DRAM-1Row can achieve 8.4% improvement on power efficiency over the baseline

What about Half-DRAM-2Row?

Thank you