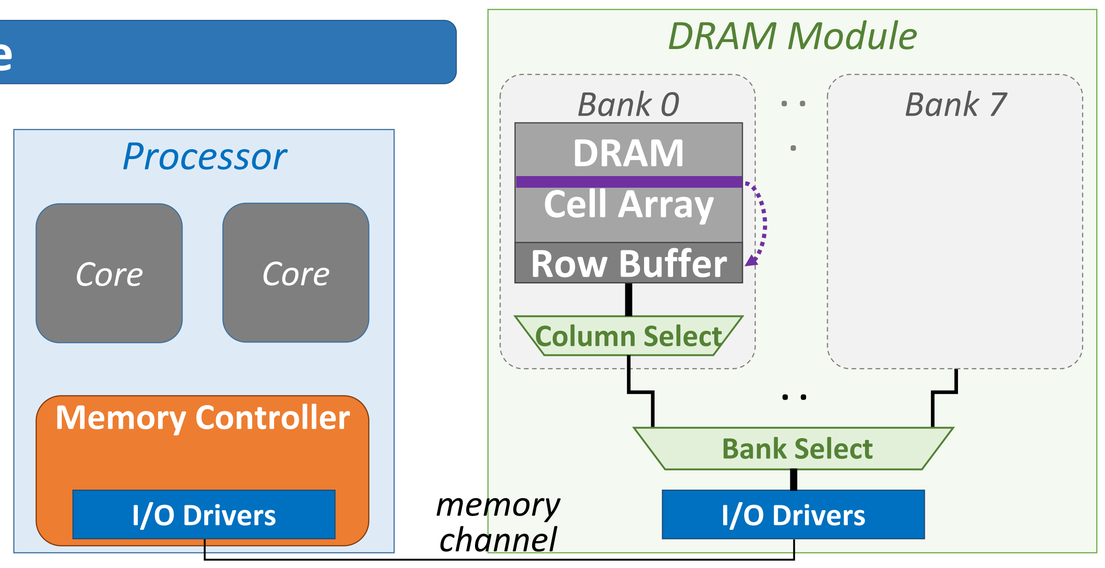
Reducing DRAM Power Consumption by Exploring and Modeling Memory Access Scheduling Policies

Sumit Kumar Yadav, Suyash Mahar, Naveen Kakarla *(advisor: Saugata Ghose)*

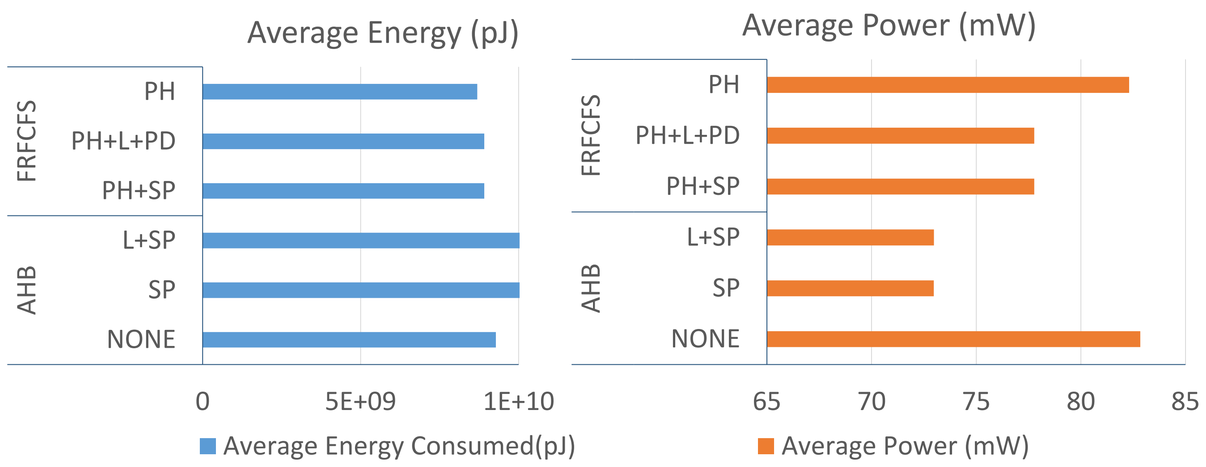
DRAM Power Consumption : A Major Issue

- Due to DRAM architecture, different locations take different time to access depending on the most recent access
- READ and WRITE requests are reordered by the memory controller instead of doing FCFS to exploit locality – known as memory access scheduling
- State-of-the-art DRAM schedulers exploit performance but waste power, which is important for embedded devices especially ones with battery.

▪ **DRAM now consumes up to half of the total system power**



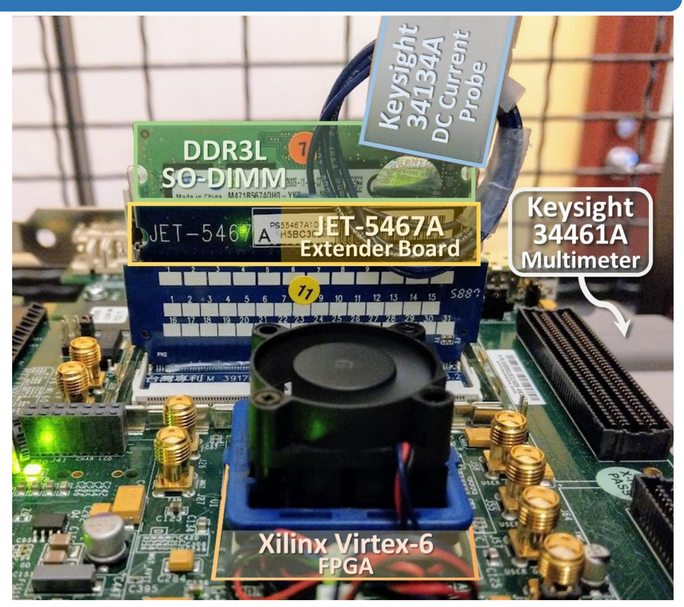
Modeling Power with state-of-the-art Tools: DRAMPower



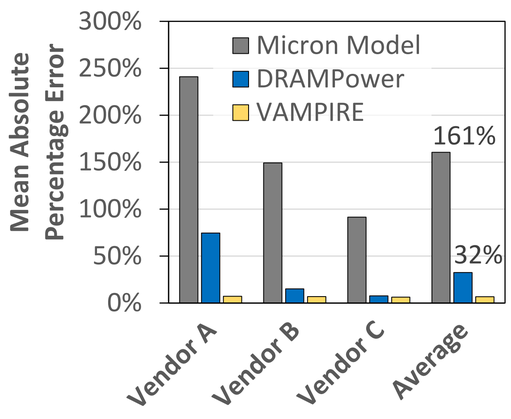
- DDRx type DRAMs provide control to power down sections (ranks) of the DRAM not in use. This can be exploited to save power.
- Simple power saving schemes interfere with locality exploitation and degrade performance.
- As a result, net energy therefore may increase.

Evaluating Power on Real Hardware

- Wrote microbenchmarks showing the benefits of a particular scheduler over others
- The corresponding power is measured on REAL hardware
- Helps pin-pointing what's exploited by a given scheduler over others



VAMPIRE: Variation-Aware model of Memory Power Informed by Real Experiments



- **VAMPIRE error (MAPE): 6.75%**
- **Open-source model: will be released in August 2018**
- https://github.com/CMU-SAFARI/VAMPIRE

How VAMPIRE Works

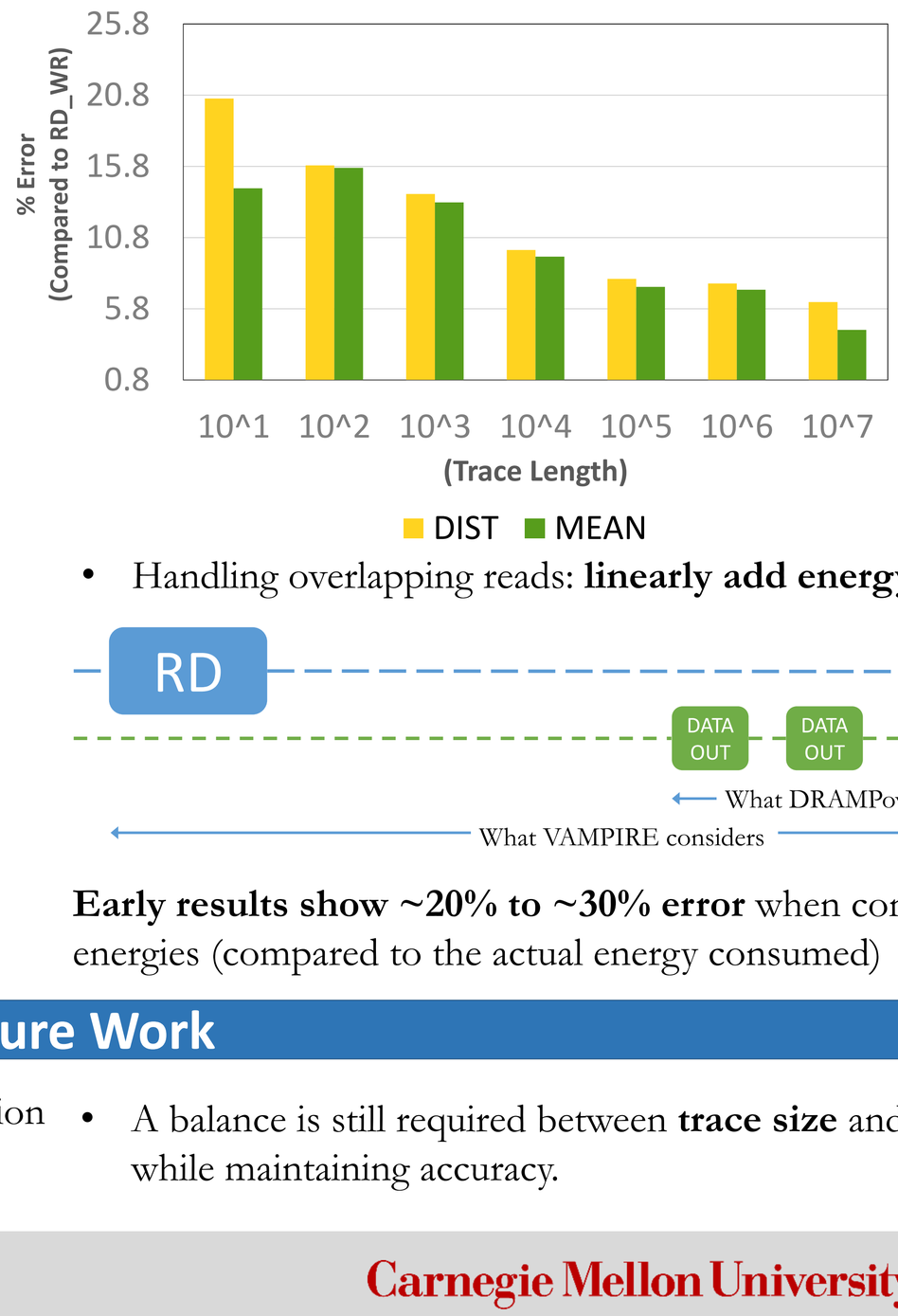
- Power of READ/WRITE depends on the data being read/written.
- It uses **real measured current data** to predict READ/WRITE power depending on the number of 1's and number of toggles.
- It also takes into account **structural variations** within a chip.

Limitations:

- It simulates DRAM's complete memory capacity
- Trace size is large, (maybe several GBs)
- Does not handle pipelined commands or idle energy

Solutions:

Use **distribution** or **mean** of the **# of 1s in the cache line** (instead of exact values) and **trade some accuracy**:



Error in VAMPIRE due to use of distribution or mean is a fundamental limitation of the equations used.

- Handling overlapping reads: **linearly add energy** of each request

Early results show ~20% to ~30% error when considering individual energies (compared to the actual energy consumed)

Future Work

- Additional effort is required to accurately predict energy consumption in case of **pipelined requests** and **idle energy**.
- A balance is still required between **trace size** and **simulation time** while maintaining accuracy.