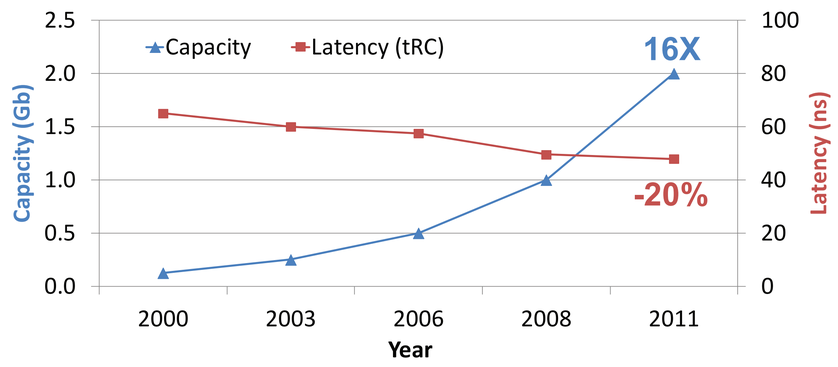
Tiered-Latency DRAM: A Low Latency and Low Cost DRAM Architecture

Donghyuk Lee, Yoongu Kim , Vivek Seshadri, Jamie Liu, Lavanya Subramanian, Onur Mutlu (CMU)

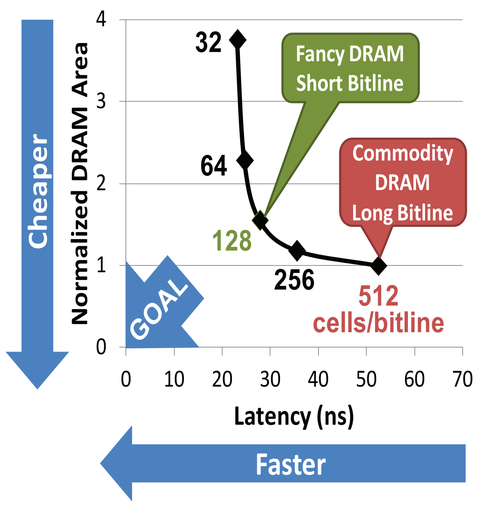
THE MEMORY LATENCY PROBLEM

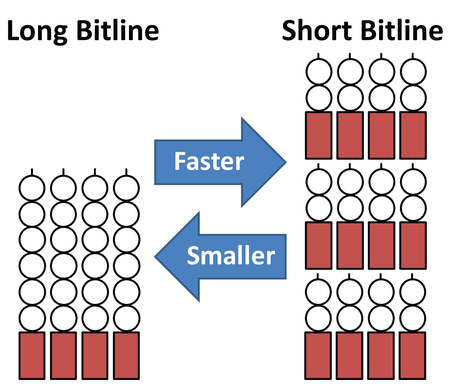
- **Commodity DRAM is optimized mainly for capacity, not latency**



- **Our Goal: Reduce DRAM latency with low area cost**

LATENCY-CAPACITY TRADEOFF



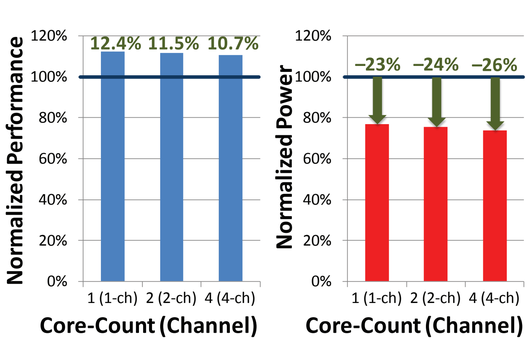


LEVERAGING THE TL-DRAM SUBSTRATE

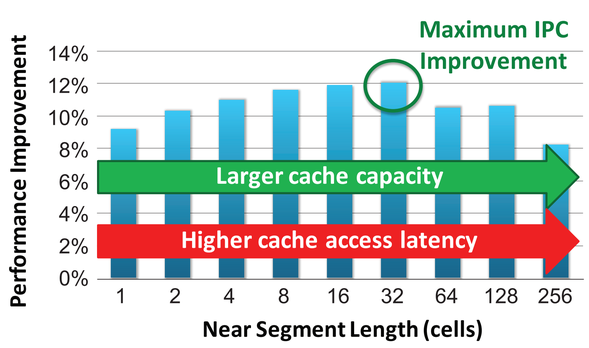
- **Fully transparent (no change to system)**
- **Use near-segment as hardware-managed cache**
  - **Far segment: Main memory**
  - **Near segment: Caches an accessed row**
  - **Memory controller manages the near segment**
- **Use near-segment as software-managed cache**
  - **OS/VMM manages the near segment**
- **Multi-level main memory**
  - **Allocate from fast vs. slow DRAM**
  - **Application or system software decides where a page goes**

RESULTS

Performance & Power Consumption



Varying Near Segment Length

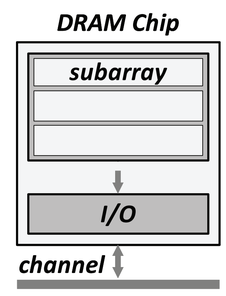


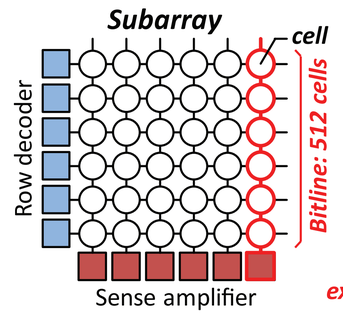
Intel Science & Technology Center for Cloud Computing

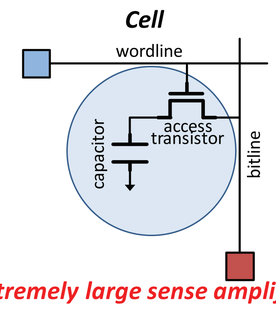
Carnegie Mellon University

Georgia Tech

DRAM ARCHITECTURE







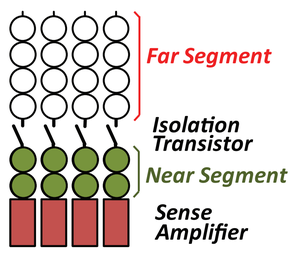
extremely large sense amplifier (≈100X the cell size)

Long Bitline: Amortizes sense amplifier's overhead → Small area
Long Bitline: Large bitline capacitance → High latency

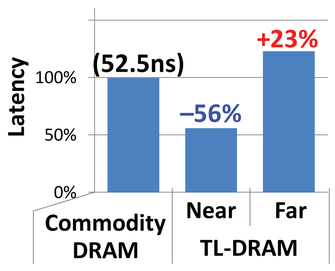
TL-DRAM: ~BEST OF BOTH WORLDS

- **Idea: Divide a subarray into two segments with an isolation transistor**
  - **Near segment: fast access, low power**
  - **Far segment: mostly slow access, high power**

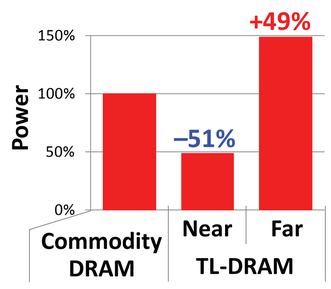
Tiered Bitline



DRAM Latency (tRC)



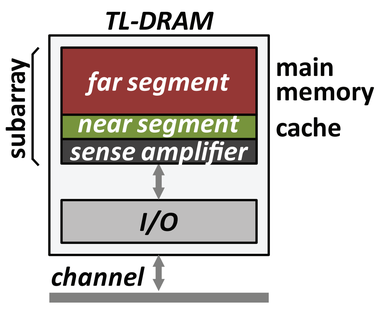
DRAM Power



- **Area cost: 3% (due to isolation transistor)**

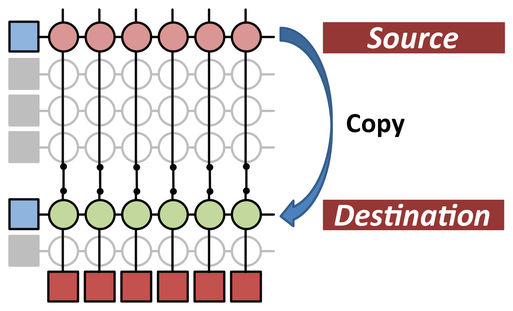
LEVERAGING TL-DRAM: CACHING

Hardware-Managed Cache



- Caching: Copy the row from far segment to near segment

Inter-Segment Migration



- Copy data from source to destination across ***shared bitlines*** concurrently

SUMMARY & ONGOING WORK

- **TL-DRAM: A new memory architecture that introduces latency heterogeneity by keeping technology homogeneity**
  - **Same chip, same technology: fast and slow portions**
- **Exposing TL-DRAM to system software**
  - **System software management algorithms**
- **Exploring Tiered Latency in NVM**
  - **Could be easier to adopt**
- **Fitting TL-DRAM into DRAM/NVM/Flash/Disk cooperative page management and allocation mechanisms**

intel

PRINCETON UNIVERSITY

UC Berkeley

UNIVERSITY of WASHINGTON