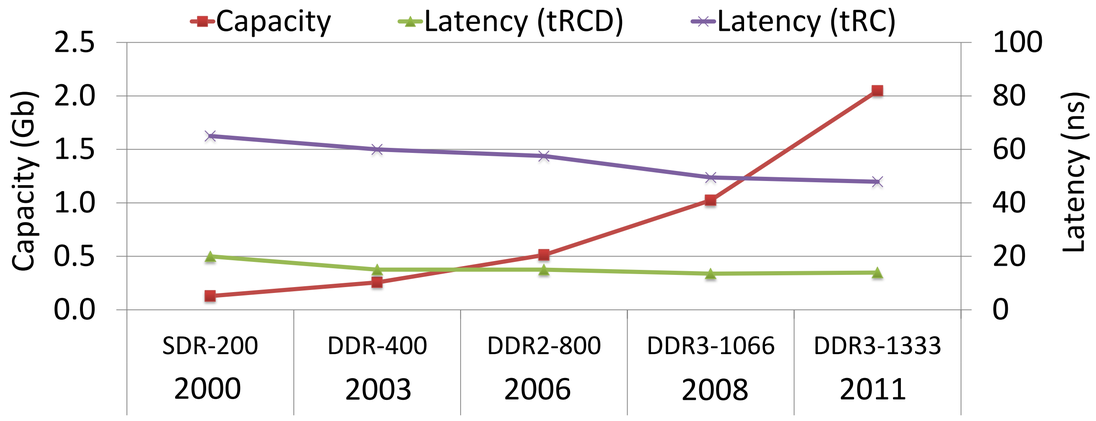
Tiered-Latency DRAM: A Low Latency and Low Cost DRAM

Donghyuk Lee, Yoongu Kim, Vivek Seshadri, Jamie Liu, Lavanya Subramanian, Onur Mutlu (Carnegie Mellon University)

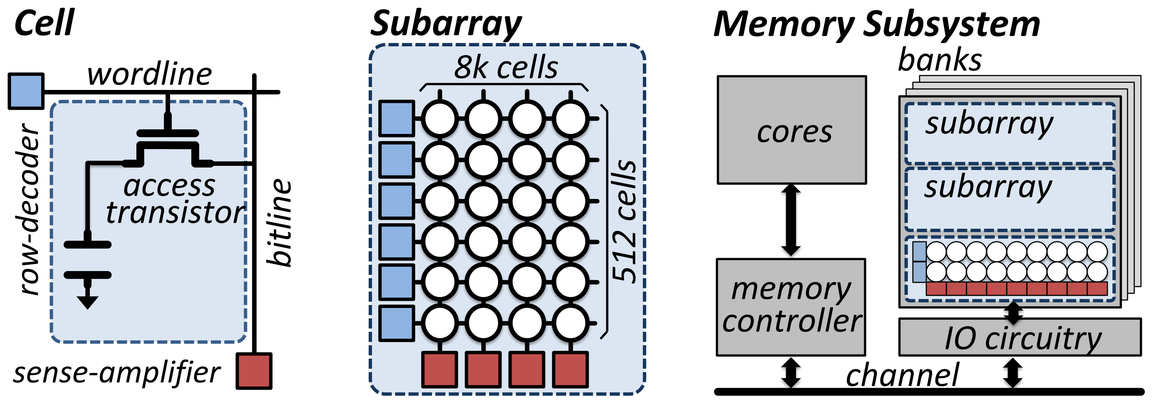
The Memory Latency Problem

- Commodity DRAM is optimized mainly for capacity, not latency
  - 16X increased capacity vs. 1.3X reduced latency

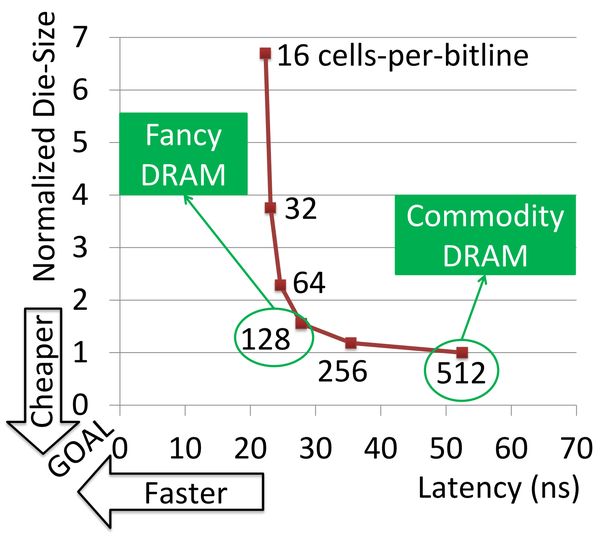


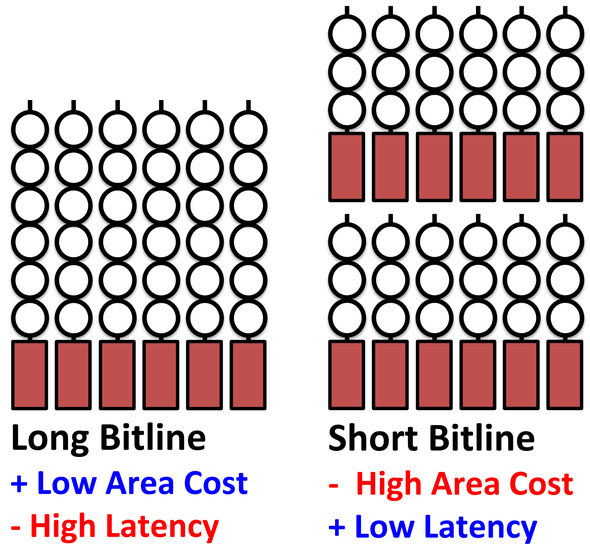
DRAM Architecture

- Long Bitline (512 cells)
- Large Bitline Capacitance: causes high access latency
  - 5X the Cell Capacitance

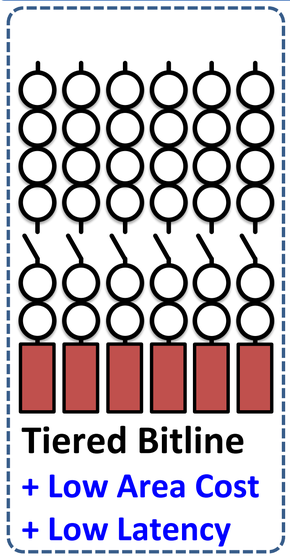


Latency-Capacity Tradeoff





TL-DRAM: ~ Best of Both Worlds



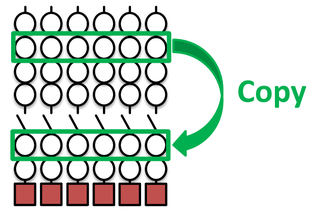
- **Idea**: **Divide a subarray into two portions** with an ***isolation transistor***
  - ***Near segment***: **fast access, low power**
  - ***Far segment***: **mostly slow access, high power**
- **Latency (tRC)**
  - Near segment: 53ns → 23ns (**57% ↓**)
  - Far segment: 53ns → 65ns (**23% ↑**)
- **Power**
  - Near segment: **51% ↓**
  - Far segment: **49% ↑**
- **Area cost : 3%** (due to isolation transistor)

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

Leveraging the TL-DRAM: Caching

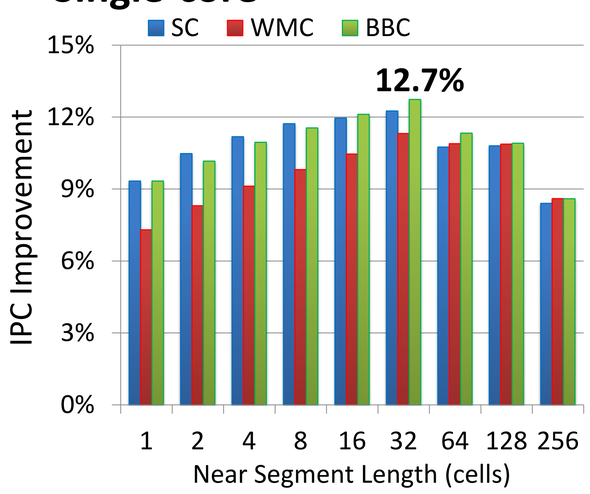
- **Caching**: Copy the row from far segment to near segment



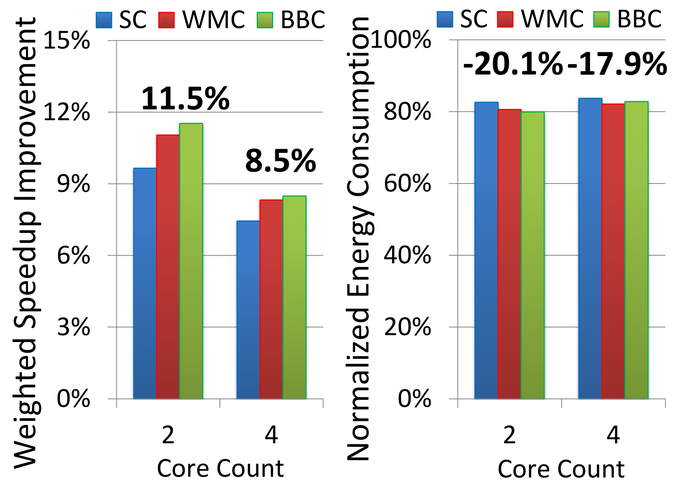
- **Simple LRU Caching (SC):** Cache a row on access
- **Wait-Minimized Caching (WMC):** Cache a row if another is waiting for the bank
- **Benefit-Based Caching (BBC):** Cache a row if it provides high latency savings
  - Keep track of **latency savings *(benefit)*** for each cached row in a table

Results

- **Single-core**



- **Multi-core**



- System: CPU:5.3GHz/LLC: 512KB (per core)
- Memory: DDR3-1066, Row-interleaved & Closed-row
- Benchmark: TPC, Stream, SPEC CPU2006, random-access
- Simulation: in-house x86 simulator with detailed memory model

Summary and 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 Science & Technology Center for Cloud Computing
Carnegie Mellon University
Georgia Tech
intel
PRINCETON UNIVERSITY
UC Berkeley