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

LEVERAGING TL-DRAM: CACHING

Hardware-Managed Cache

Inter-Segment Migration

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