Lecture 12: DRAM Basics

• Today: DRAM terminology and basics, energy innovations
DRAM Main Memory

- Main memory is stored in DRAM cells that have much higher storage density.

- DRAM cells lose their state over time – must be refreshed periodically, hence the name *Dynamic*.

- DRAM access suffers from long access time and high energy overhead.

- Since the pins on a processor chip are expected to not increase much, we will hit a memory bandwidth wall.
DRAM Organization

- Array
- 1/8th of the row buffer
- One word of data output
- Bank
- DRAM chip or device
- Rank
- DIMM
- Memory bus or channel
- On-chip Memory Controller
DRAM Array Access

16Mb DRAM array = 4096 x 4096 array of bits

12 row address bits arrive first
Row Access Strobe (RAS)

12 column address bits arrive next
Column Access Strobe (CAS)

4096 bits are read out
Column decoder

Eight bits returned to CPU, one per cycle
Row Buffer
Salient Points I

• DIMM, rank, bank, array → form a hierarchy in the storage organization

• Because of electrical constraints, only a few DIMMs can be attached to a bus

• Ranks help increase the capacity on a DIMM

• Multiple DRAM chips are used for every access to improve data transfer bandwidth

• Multiple banks are provided so we can be simultaneously working on different requests
Salient Points II

• To maximize density, arrays within a bank are made large → rows are wide → row buffers are wide (8KB read for a 64B request)

• Each array provides a single bit to the output pin in a cycle (for high density and because there are few pins)

• DRAM chips are described as xN, where N refers to the number of output pins; one rank may be composed of eight x8 DRAM chips (the data bus is 64 bits)

• The memory controller schedules memory accesses to maximize row buffer hit rates and bank/rank parallelism
Salient Points III

- Banks and ranks offer memory parallelism

- Row buffers act as a cache within DRAM
  - Row buffer hit: ~20 ns access time (must only move data from row buffer to pins)
  - Empty row buffer access: ~40 ns (must first read arrays, then move data from row buffer to pins)
  - Row buffer conflict: ~60 ns (must first writeback the existing row, then read new row, then move data to pins)

- In addition, must wait in the queue (tens of nano-seconds) and incur address/cmd/data transfer delays (~10 ns)
Technology Trends

• Improvements in technology (smaller devices) → DRAM capacities double every two years, but latency does not change much

• Power wall: 25-40% of datacenter power can be attributed to the DRAM system

• Will soon hit a density wall; may have to be replaced by other technologies (phase change memory, STT-RAM)

• Interconnects may have to be photonic to overcome the bandwidth limitation imposed by pins on the chip
Latency and Power Wall

- Latency and power can be both improved by employing smaller arrays; incurs a penalty in density and cost

- Latency and power can be both improved by increasing the row buffer hit rate; requires intelligent mapping of data to rows, clever scheduling of requests, etc.

- Power can be reduced by minimizing overfetch – either read fewer chips or read parts of a row; incur penalties in area or bandwidth
Overfetch

- Overfetch caused by multiple factors:

  - Each array is large (fewer peripherals → more density)
  - Involving more chips per access → more data transfer pin bandwidth
  - More overfetch → more prefetch; helps apps with locality
  - Involving more chips per access → less data loss when a chip fails → lower overhead for reliability
Re-Designing Arrays

Udipi et al., ISCA’10
Selective Bitline Activation

• Additional logic per array so that only relevant bitlines are read out

• Essentially results in finer-grain partitioning of the DRAM arrays

• Two papers in 2010: Udipi et al., ISCA’10, Cooper-Balis and Jacob, IEEE Micro
Rank Subsetting

• Instead of using all chips in a rank to read out 64-bit words every cycle, form smaller parallel ranks

• Increases data transfer time; reduces the size of the row buffer

• But, lower energy per row read and compatible with modern DRAM chips

• Increases the number of banks and hence promotes parallelism (reduces queuing delays)
Title

• Bullet