GPU Teaching Kit

Accelerated Computing

Module 6 – Memory Access Performance

Lecture 6.1 - DRAM Bandwidth

Objective

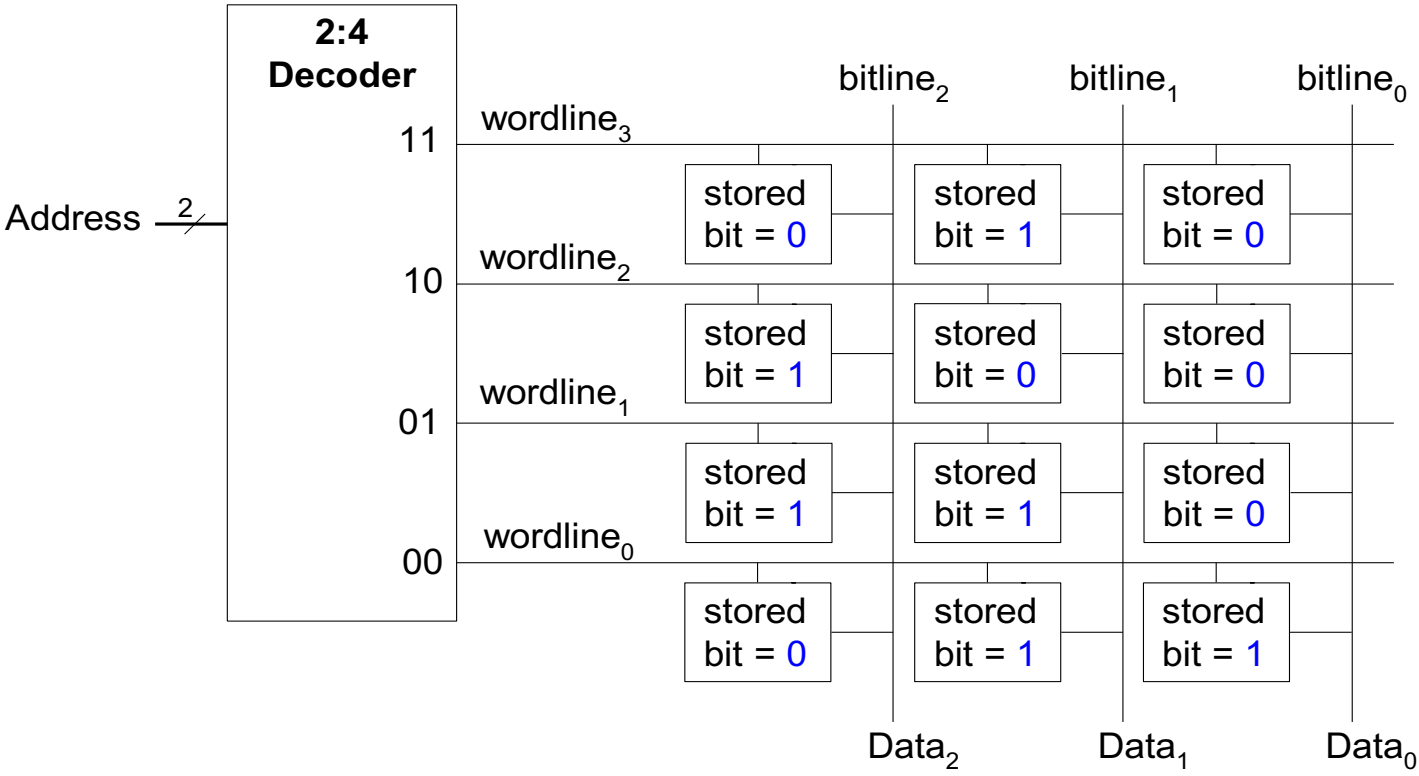
- To learn that memory bandwidth is a first-order performance factor in a massively parallel processor
  - DRAM bursts, banks, and channels
  - All concepts are also applicable to modern multicore processors

Global Memory (DRAM) Bandwidth

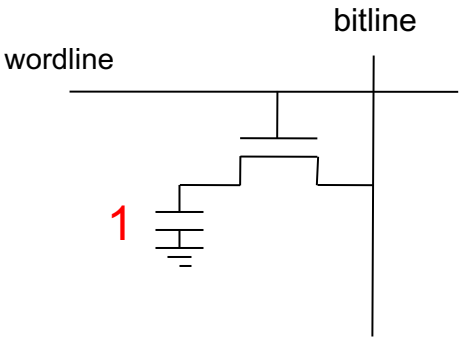
- Ideal

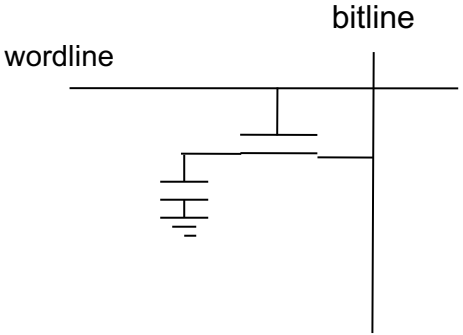
- Reality

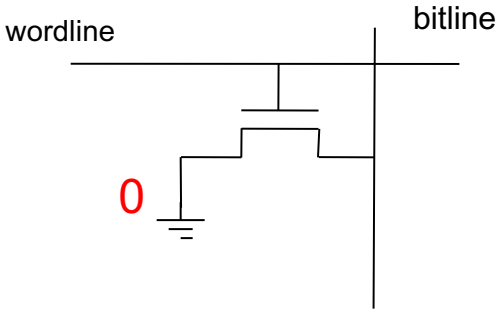
DRAM Core Array Organization

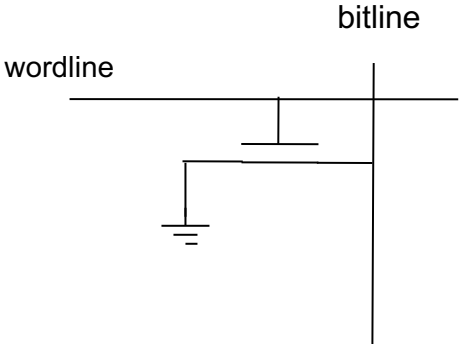


Bit Cell

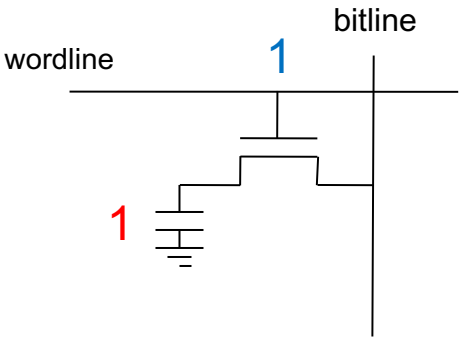

bitline
wordline
1

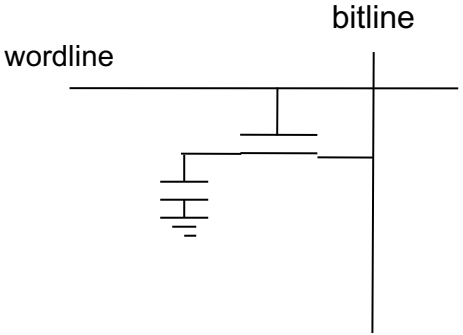

bitline
wordline

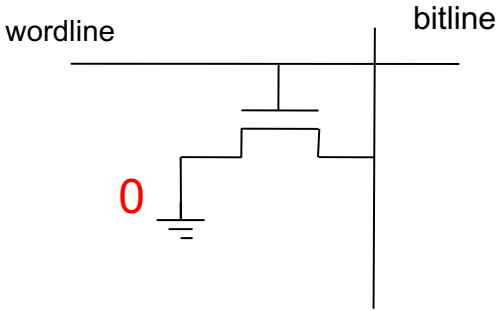

wordline
bitline
0

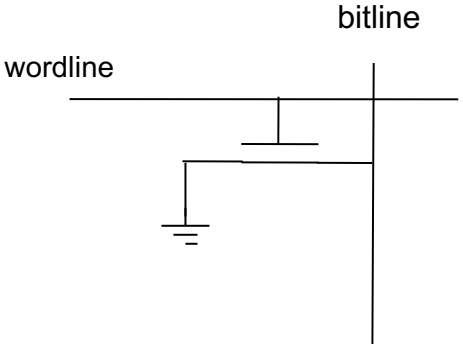

bitline
wordline

Bit Cell

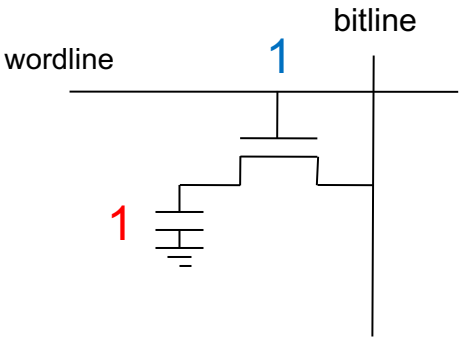

bitline
wordline
1
1

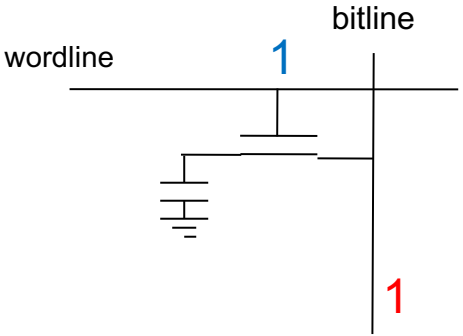

bitline
wordline

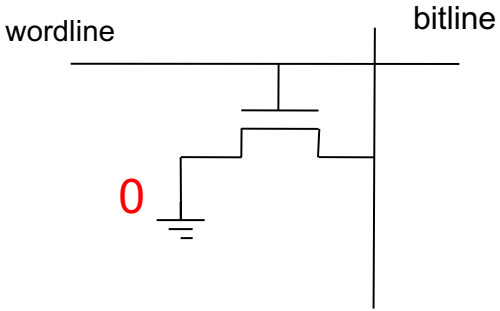

wordline
bitline
0

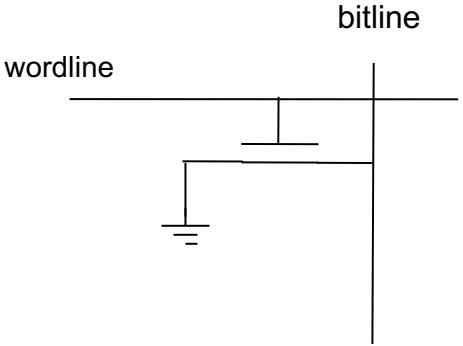

bitline
wordline

Bit Cell

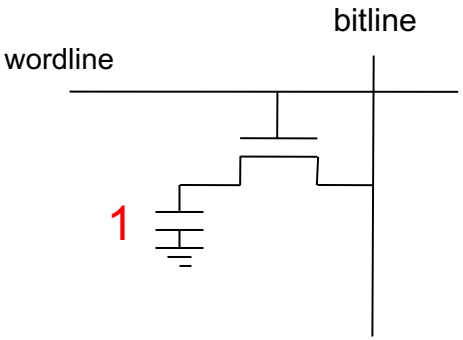

bitline
wordline
1
1

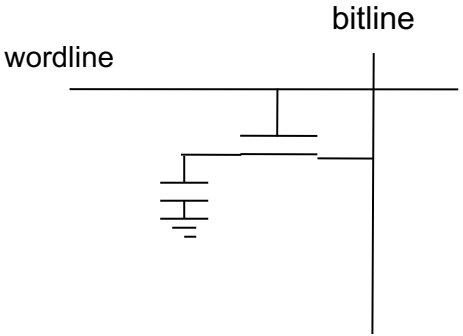

bitline
wordline
1
1

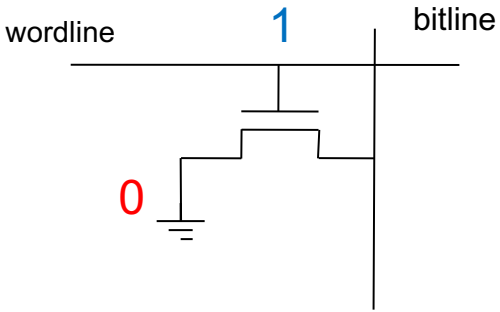

wordline
bitline
0

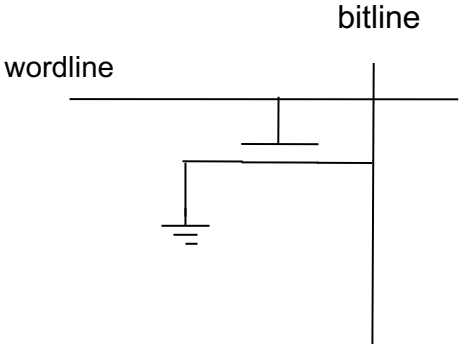

bitline
wordline

Bit Cell

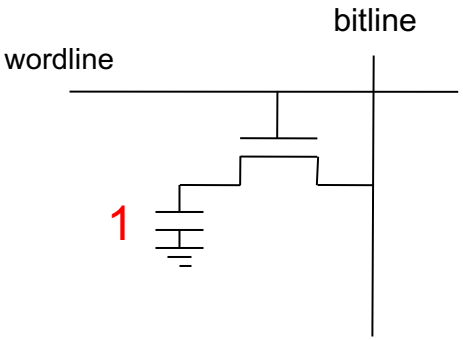

bitline
wordline
1

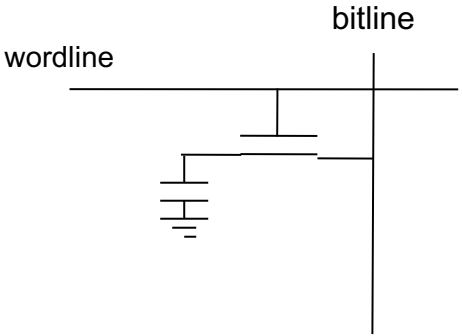

bitline
wordline

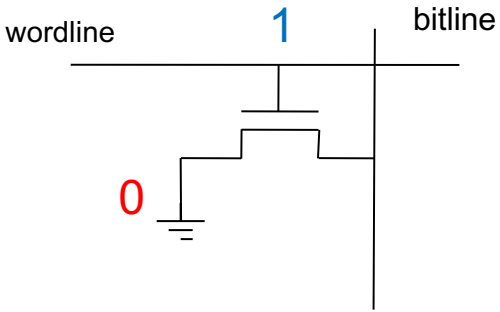

wordline
1
bitline
0

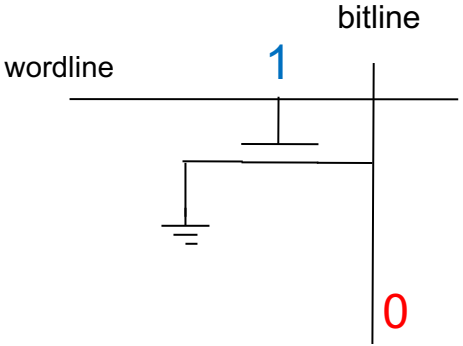

bitline
wordline

Bit Cell


bitline
wordline
1

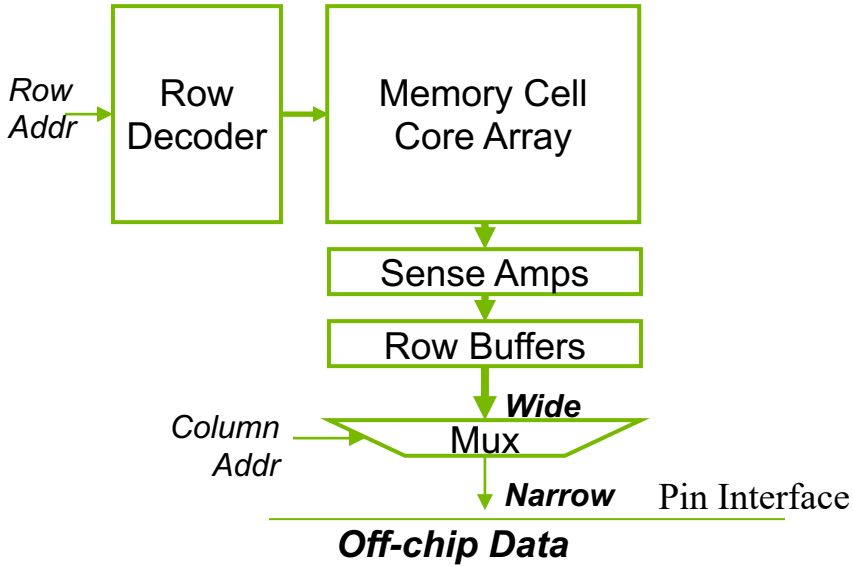

bitline
wordline


wordline
1
bitline
0


bitline
wordline
1
0

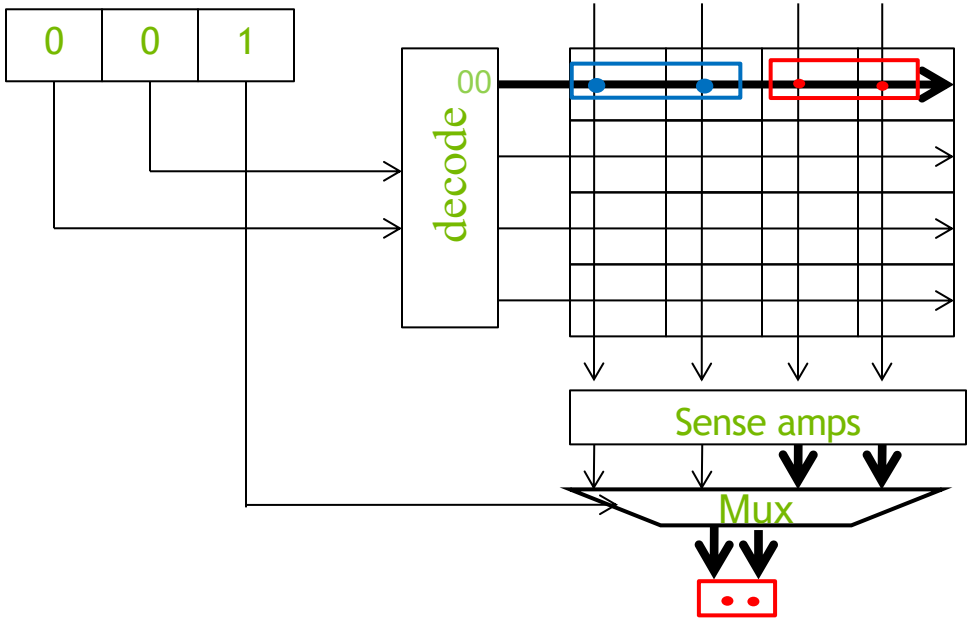
DRAM Core Array Organization

- Each DRAM core array has about 16M bits
- Each bit is stored in a tiny capacitor made of one transistor



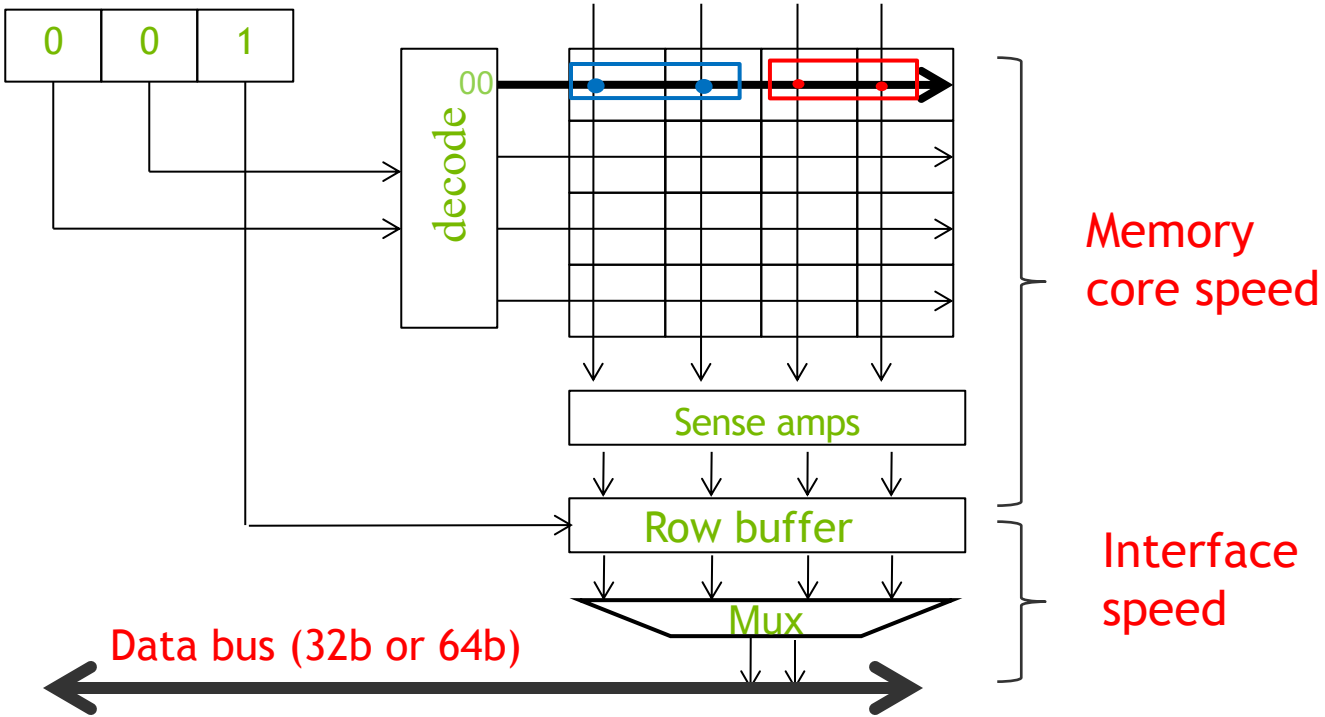
A very small (8x2-bit) DRAM Core Array

– Assume that you read only what you need



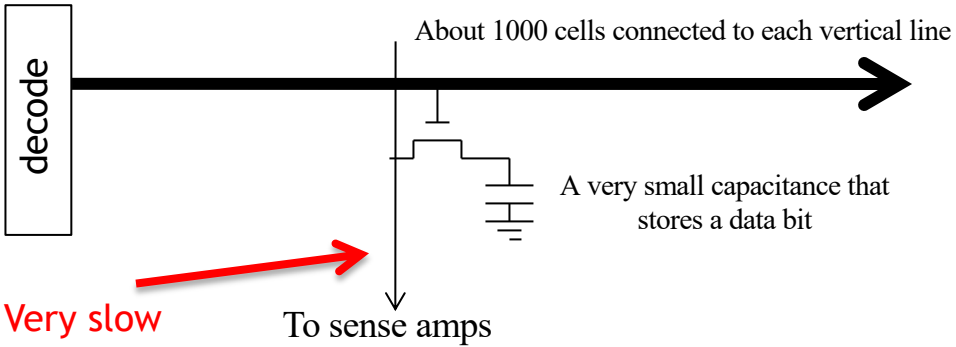
A very small (8x2-bit) DRAM Core Array

- How fast to move from memory core matrix to row buffer?
- How fast to move from row buffer to data bus (interface speed)



DRAM Core Arrays are Slow

- Reading from a cell in the core array is a very slow process
  - DDR: Memory core speed = ½ bus interface speed
  - DDR2: Memory core speed = ¼ bus interface speed
  - DDR3: Memory core speed = ⅛ bus interface speed
  - … likely to be worse in the future

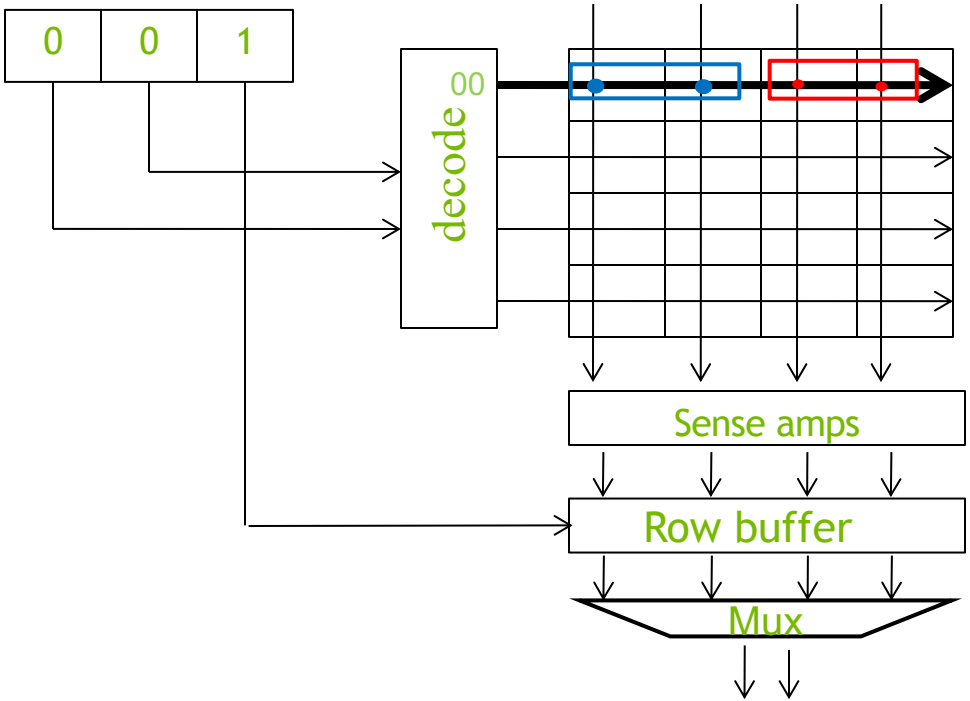


DRAM with Bursting

- Assume interface width: 32b (4B)
- Assume DDR3:
  - Memory core speed is 1/8 of bus interface speed
  - Row length is 8x bus interface width to compensate the speed unbalance
  - Row length is thus 8 x 4B = 32B = 256b
- How it works:
  - Loads 32B at once into row buffer
  - Transfers a burst of 8 bus transactions of 32b each

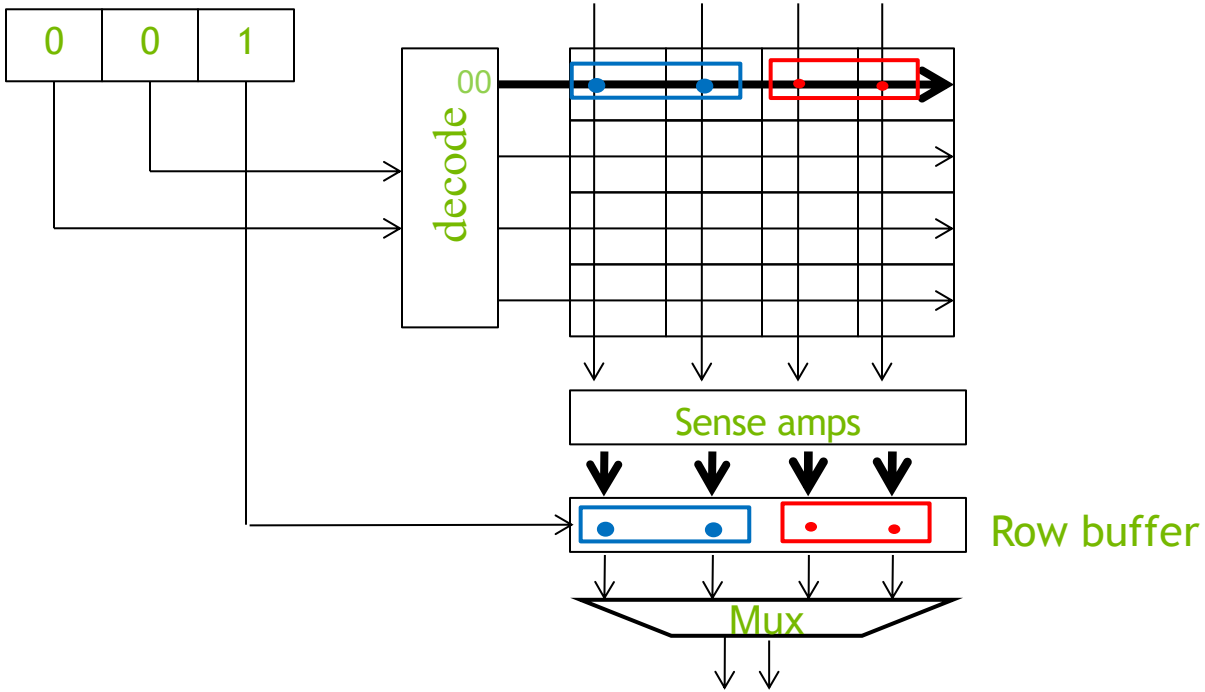
A very small (8x2-bit) DRAM Core Array

- Assume interface width: 2b
- Assume DDR: Memory core speed is 1/2 of bus interface speed



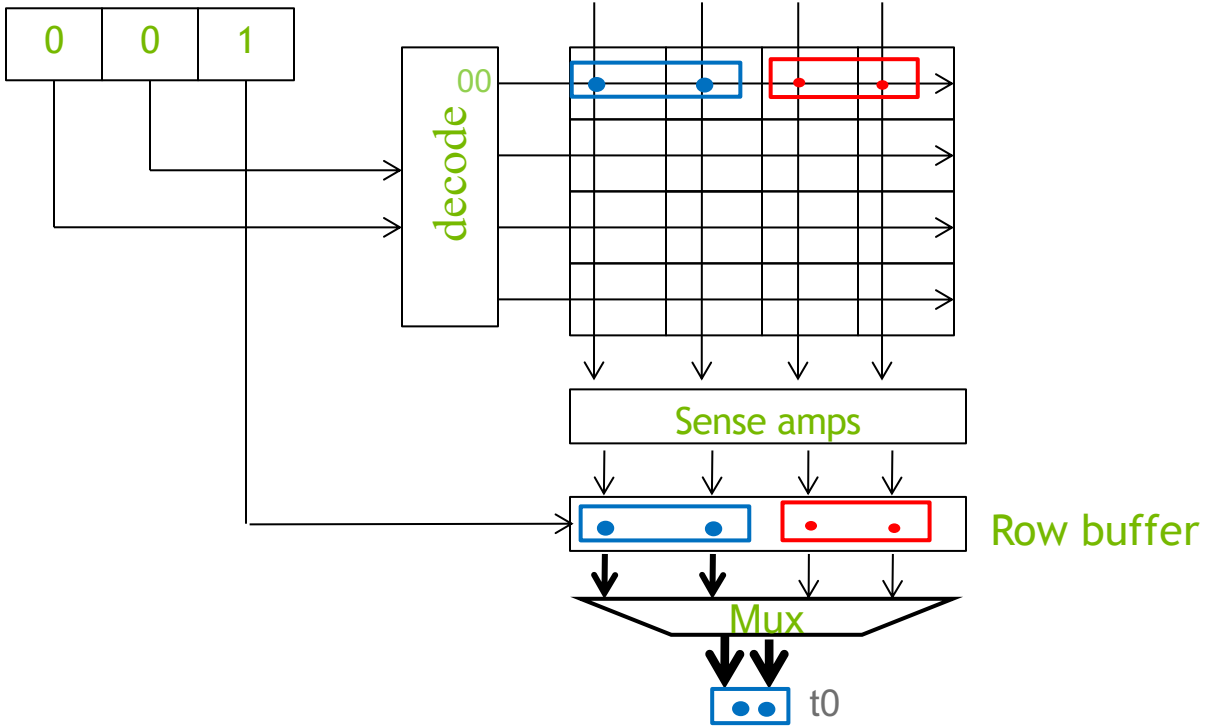
A very small (8x2-bit) DRAM Core Array

- Assume interface width: 2b
- Assume DDR: Memory core speed is 1/2 of bus interface speed



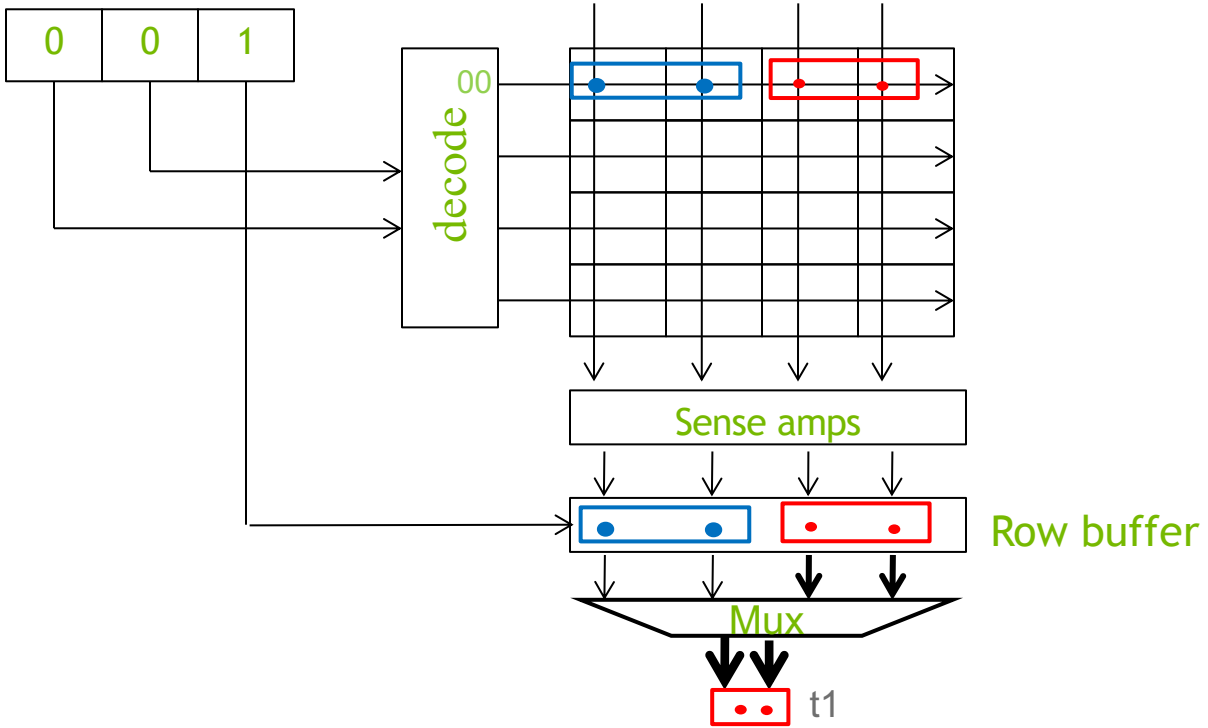
A very small (8x2-bit) DRAM Core Array

- Assume interface width: 2b
- Assume DDR: Memory core speed is 1/2 of bus interface speed

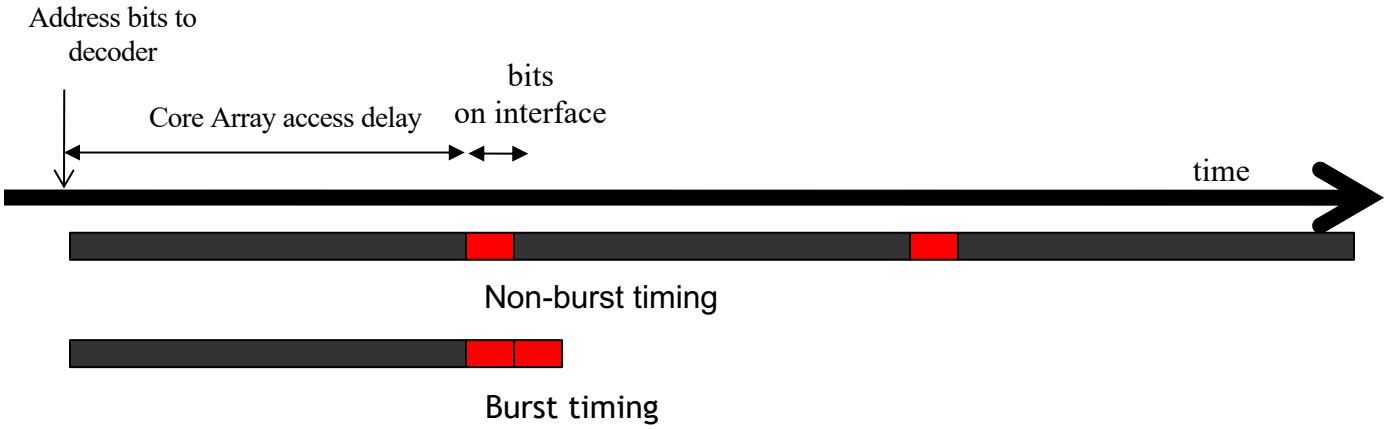


A very small (8x2-bit) DRAM Core Array

- Assume interface width: 2b
- Assume DDR: Memory core speed is 1/2 of bus interface speed



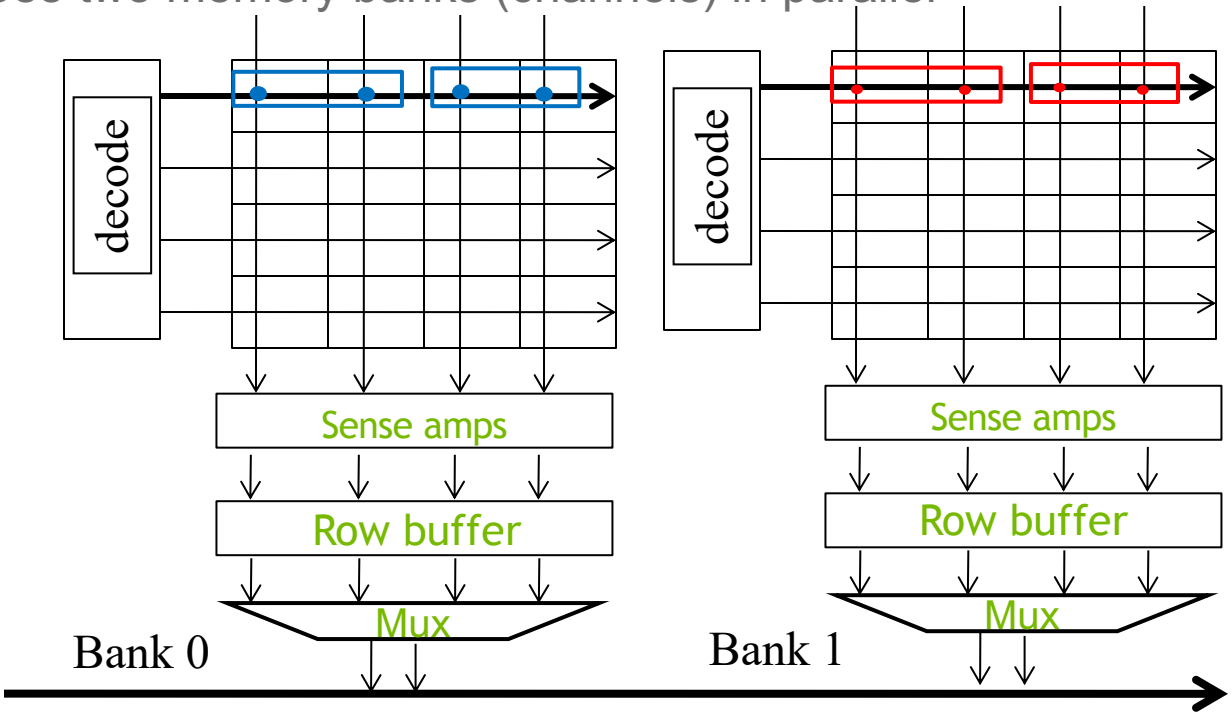
DRAM Bursting Timing Example



Modern DRAM systems are designed to always be accessed in burst mode. Burst bytes are transferred to the processor but discarded when accesses are not to sequential locations.

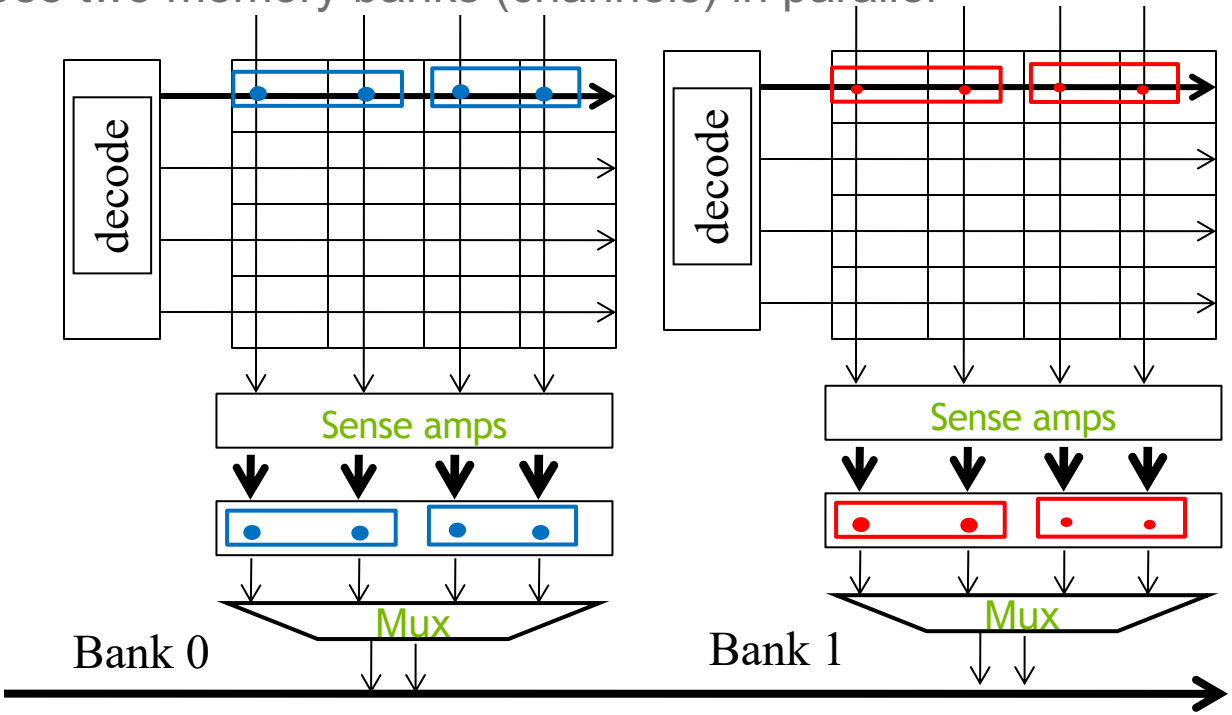
Multiple DRAM Banks (or Channels)

- Sometimes cannot increase memory core row anymore
- Use two memory banks (channels) in parallel



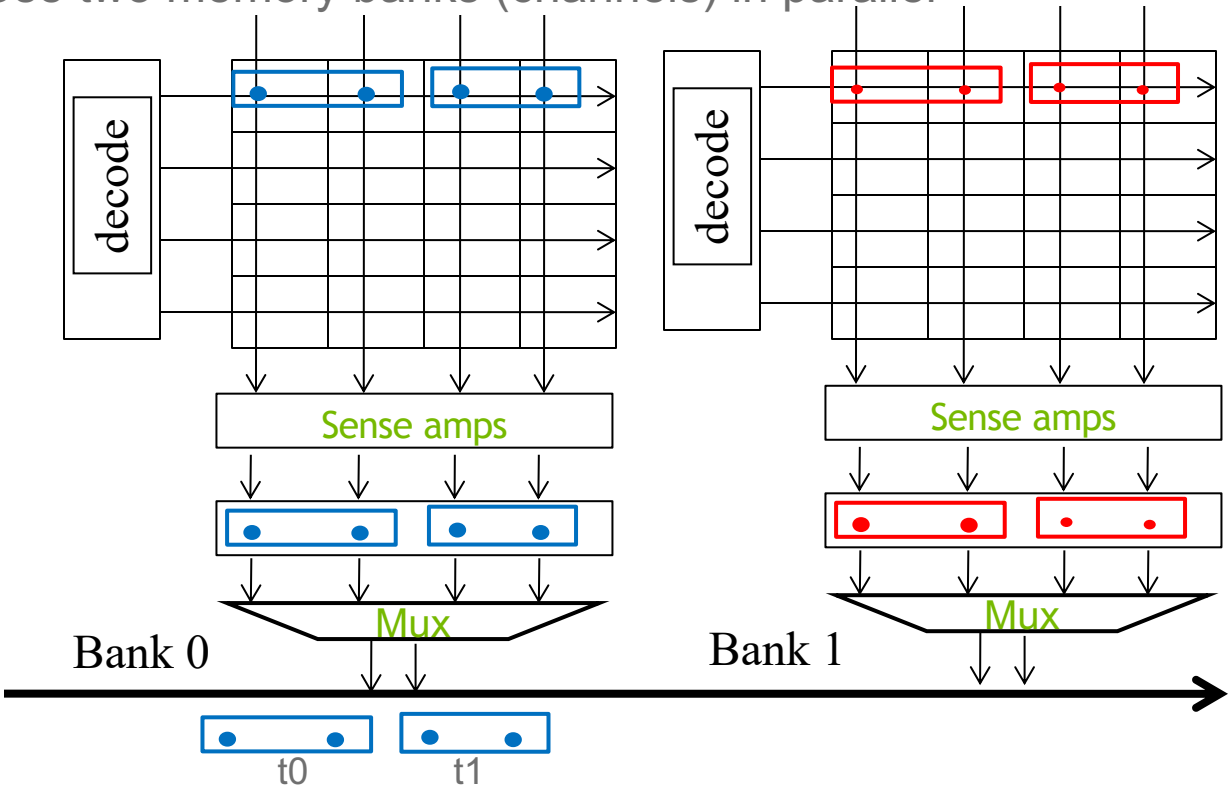
Multiple DRAM Banks (or Channels)

- Sometimes cannot increase memory core row anymore
- Use two memory banks (channels) in parallel



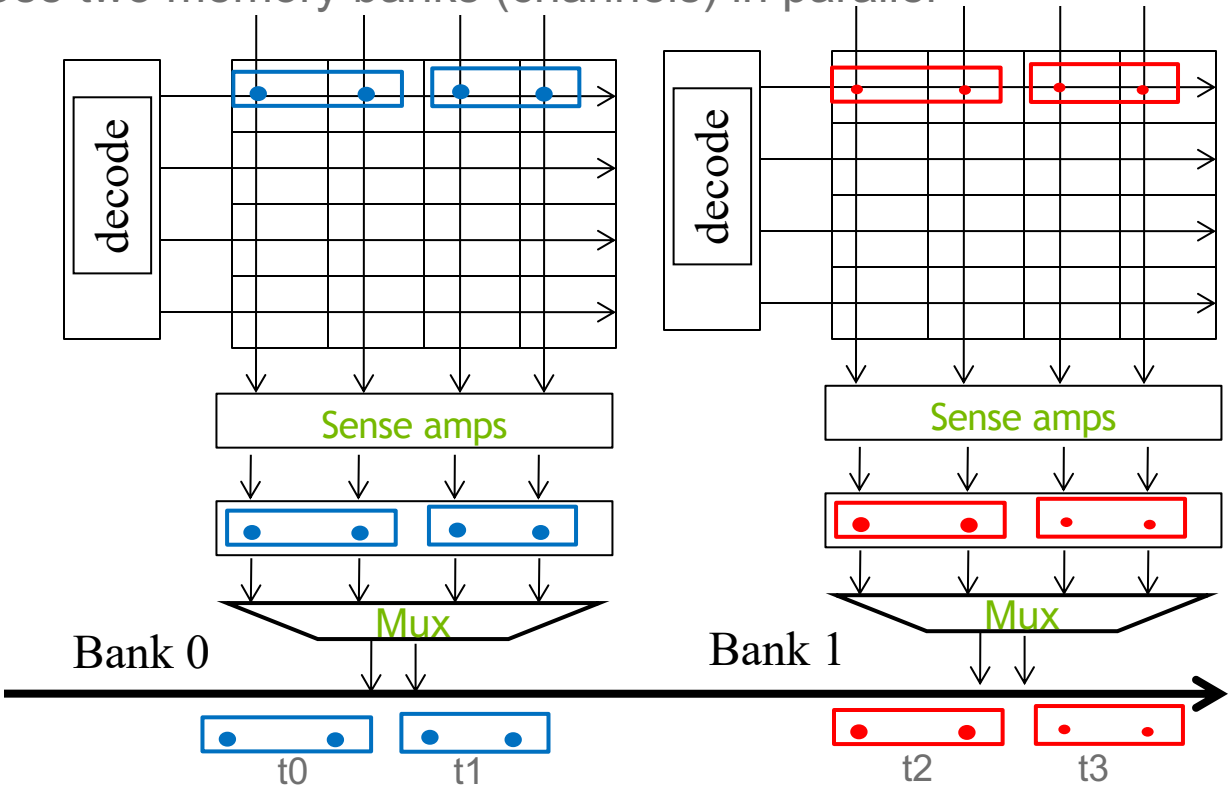
Multiple DRAM Banks (or Channels)

- Sometimes cannot increase memory core row anymore
- Use two memory banks (channels) in parallel



Multiple DRAM Banks (or Channels)

- Sometimes cannot increase memory core row anymore
- Use two memory banks (channels) in parallel



DRAM Bursting with Banking

Single bank

Single-Bank burst timing, dead time on interface

Dual bank

Multi-Bank burst timing, reduced dead time

GPU off-chip memory subsystem

- Assume DDR2 global memory
  - Consider bus width 64b (8B), and a transfer at each cycle
  - Consider interface speed @ 1.1GHz
  - Consider memory core speed @ 276MHz
  - 1.1GHz / 276Mhz = 4x slower
  - Thus memory row needs to be 4x 8B = 32B to compensate
- Assume NVIDIA GTX280 GPU with the memory above:
  - This GPU demands a peak global memory bandwidth of 141.7 GB/s, or 141.7 / 1.1 = 128 B/cycle
  - But memory provides only 8 B/cycle
  - To feed the total peak bandwidth needs 128 / 8 = 16 banks
- Need to re-use burst section data!!