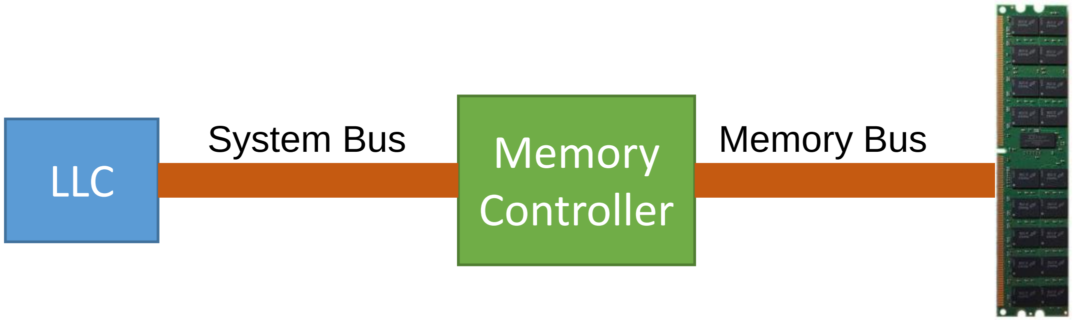
Main Memory & DRAM

Nima Honarmand

Main Memory — Big Picture

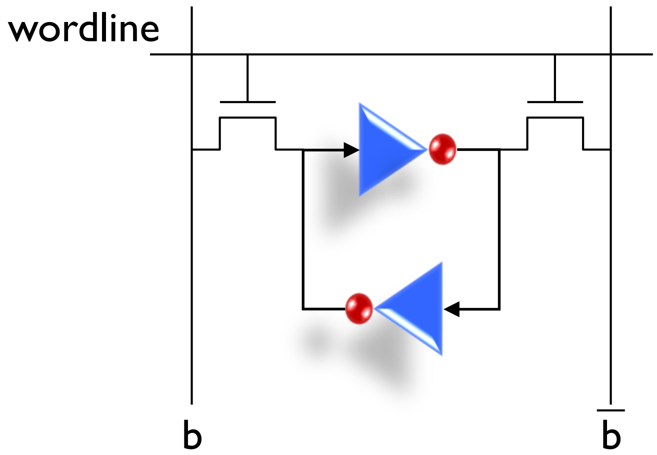
1) Last-level cache sends its memory requests to a Memory Controller
   - Over a system bus of other types of interconnect
2) Memory controller translates this request to a bunch of commands and sends them to DRAM devices
3) DRAM devices perform the operation (read or write) and return the results (if read) to memory controller
4) Memory controller returns the results to LLC



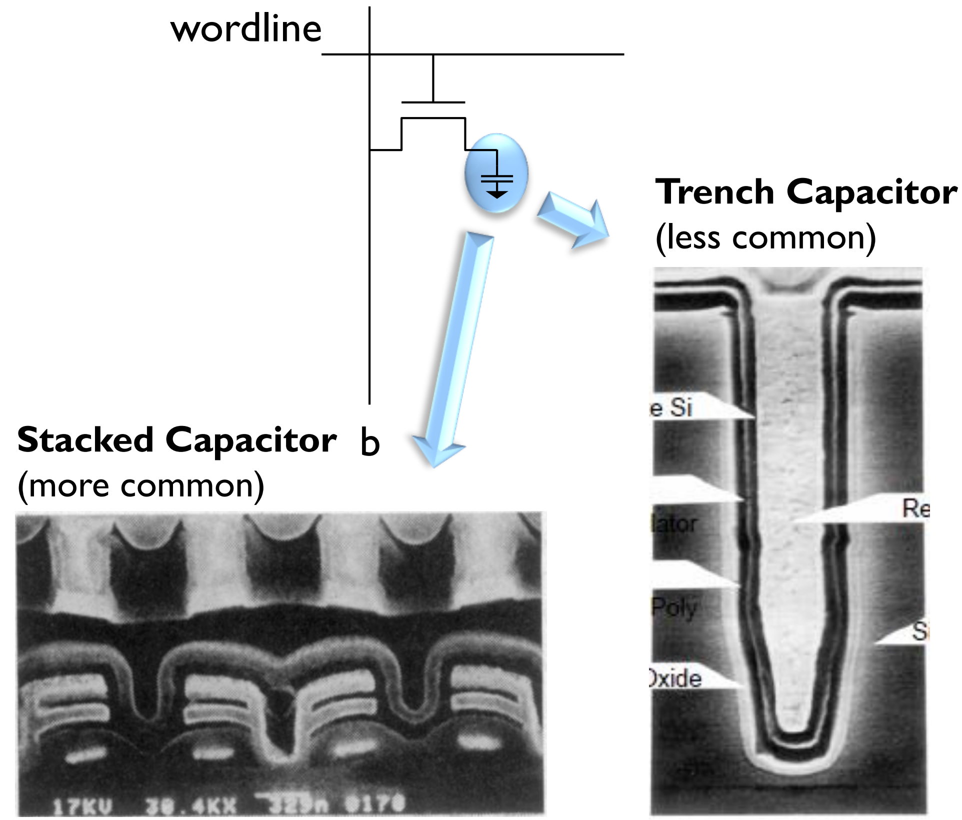
SRAM vs. DRAM

- SRAM = Static RAM
  - As long as power is present, data is retained
- DRAM = Dynamic RAM
  - If you don't refresh, you lose the data even with power connected
- SRAM: 6T per bit
  - built with normal high-speed VLSI technology
- DRAM: 1T per bit + 1 capacitor
  - built with special VLSI process optimized for density

Memory Cell Structures

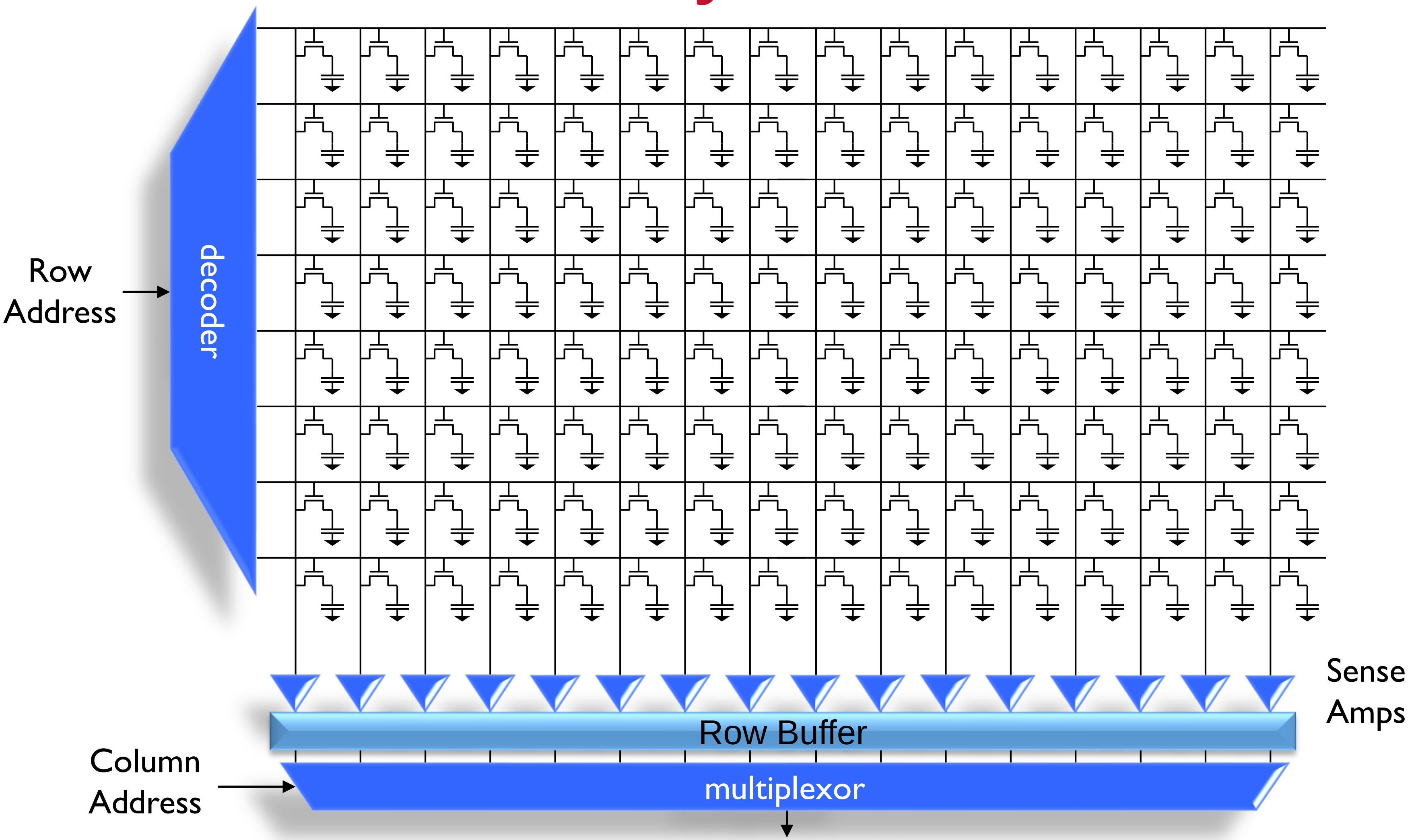


SRAM



DRAM

DRAM Cell Array

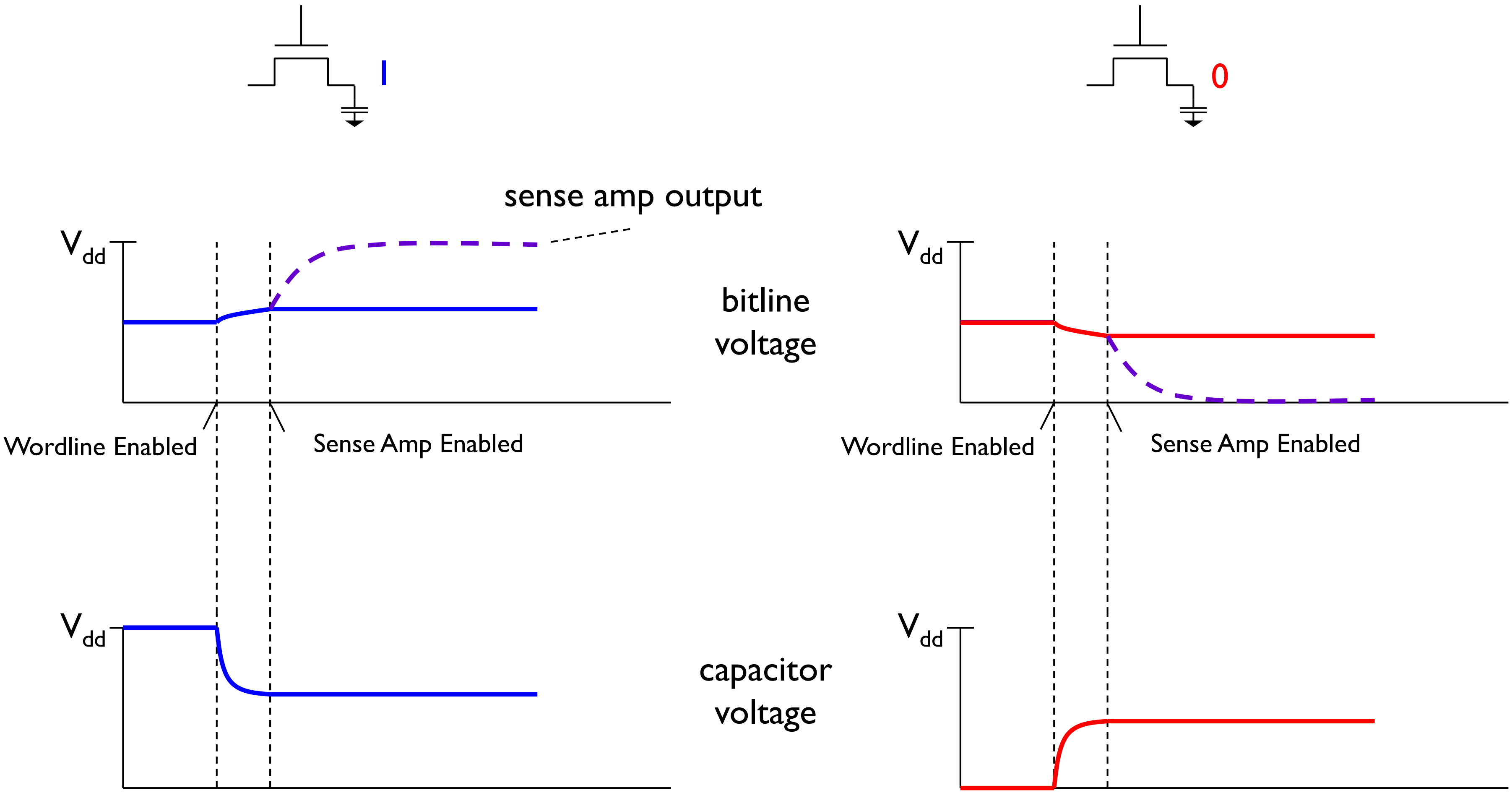


DRAM is much denser than SRAM

DRAM Array Operation

- Low-Level organization is very similar to SRAM
- Reads are *destructive*: contents are erased by reading
- *Row buffer* holds read data
  - Data in row buffer is called a *DRAM row*
    - Often called “page” – do not confuse with virtual memory page
  - Read gets entire row into the buffer
  - Block reads always performed out of the row buffer
    - Reading a whole row, but accessing one block
    - Similar to reading a cache line, but accessing one word

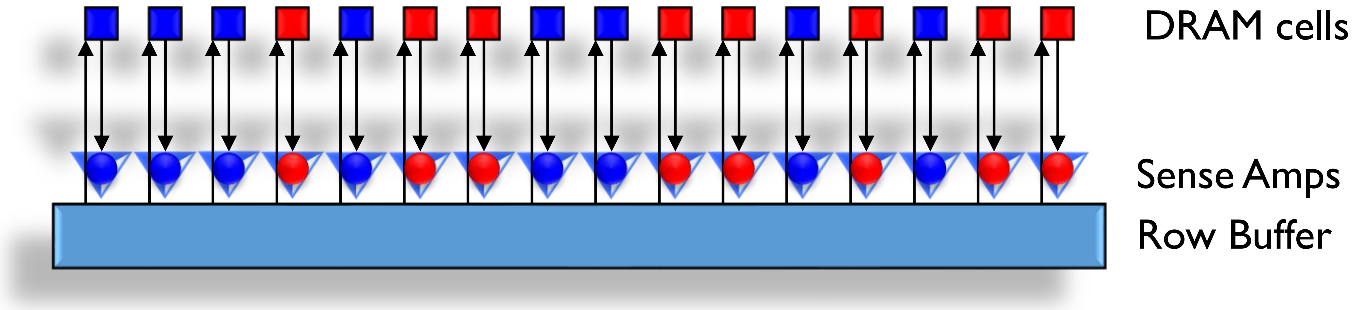
Destructive Read



After read of 0 or 1, cell contents close to ½

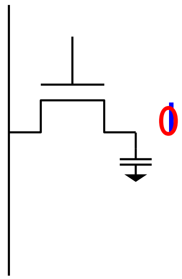
DRAM Read

- After a read, the contents of the DRAM cell are gone
  - But still “safe” in the row buffer
- Write bits back before doing another read
- Reading into buffer is slow, but reading buffer is fast
  - Try reading multiple lines from buffer *(row-buffer hit)*



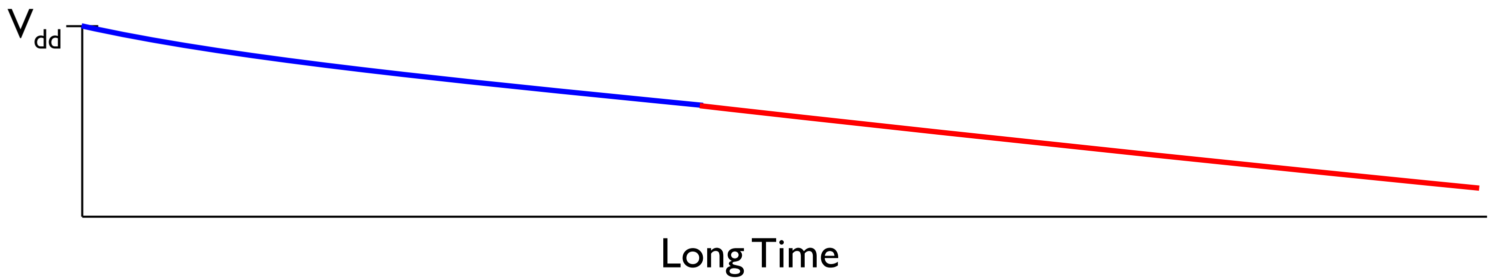
Process is called *opening* or *closing* a row

DRAM Refresh (1)



- Gradually, DRAM cell loses contents
  - Even if it's not accessed
  - This is why it's called "dynamic"

- DRAM must be regularly read and re-written
  - What to do if no read/write to row for long time?

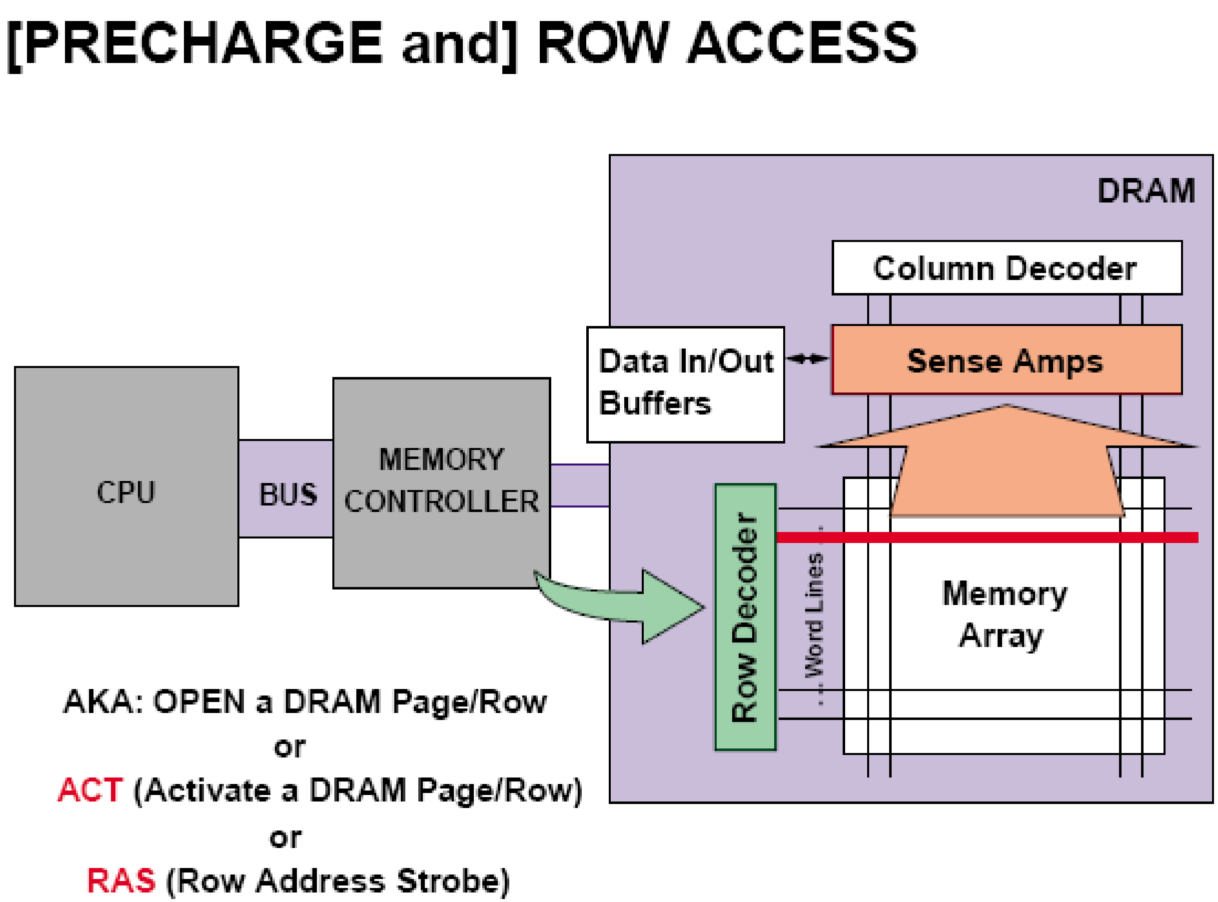


Must periodically *refresh* all contents

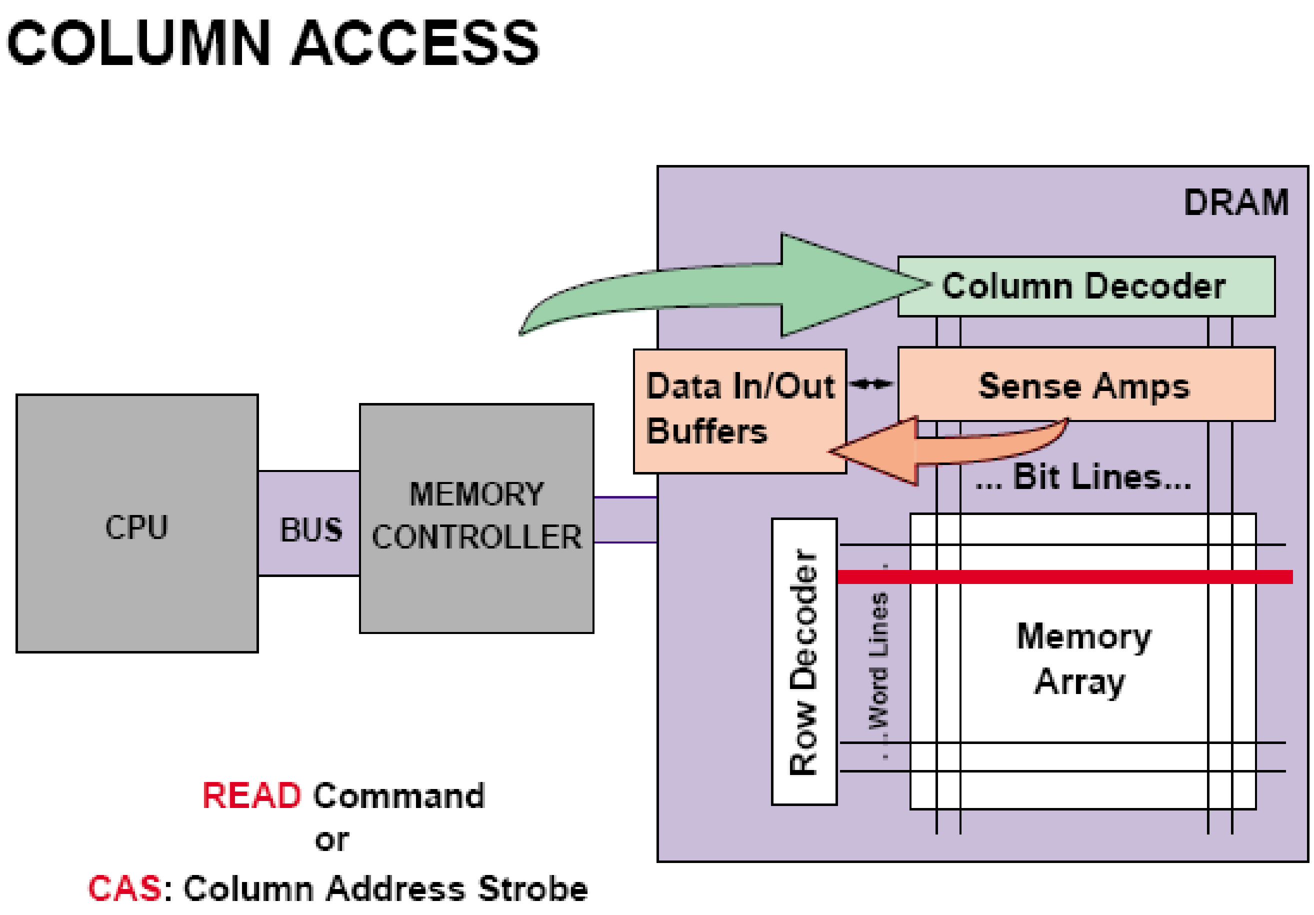
DRAM Refresh (2)

- Burst Refresh
  - Stop the world, refresh all memory
- Distributed refresh
  - Space out refresh one (or a few) row(s) at a time
  - Avoids blocking memory for a long time
- Self-refresh (low-power mode)
  - Tell DRAM to refresh itself
  - Turn off memory controller
  - Takes some time to exit self-refresh

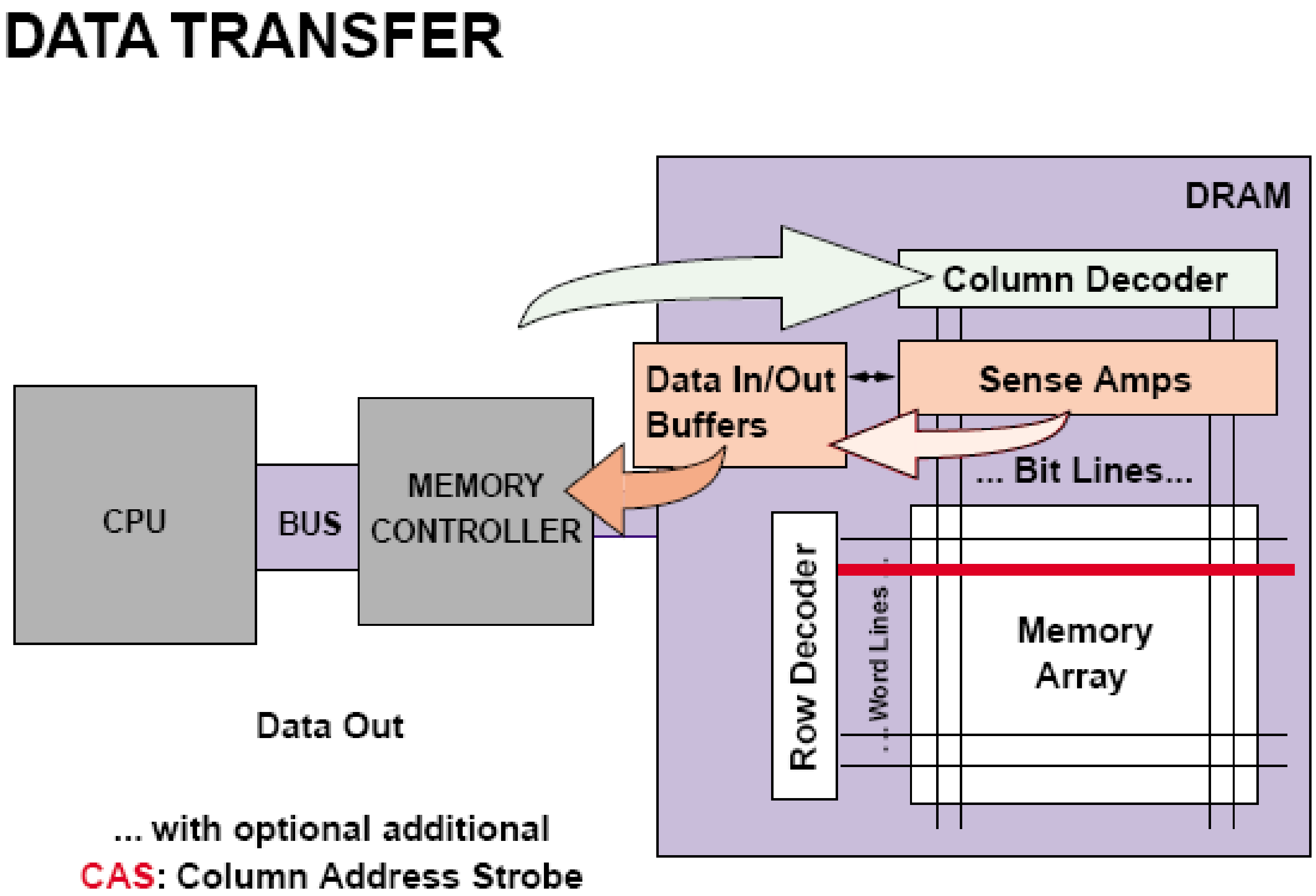
Typical DRAM Access Sequence (1)



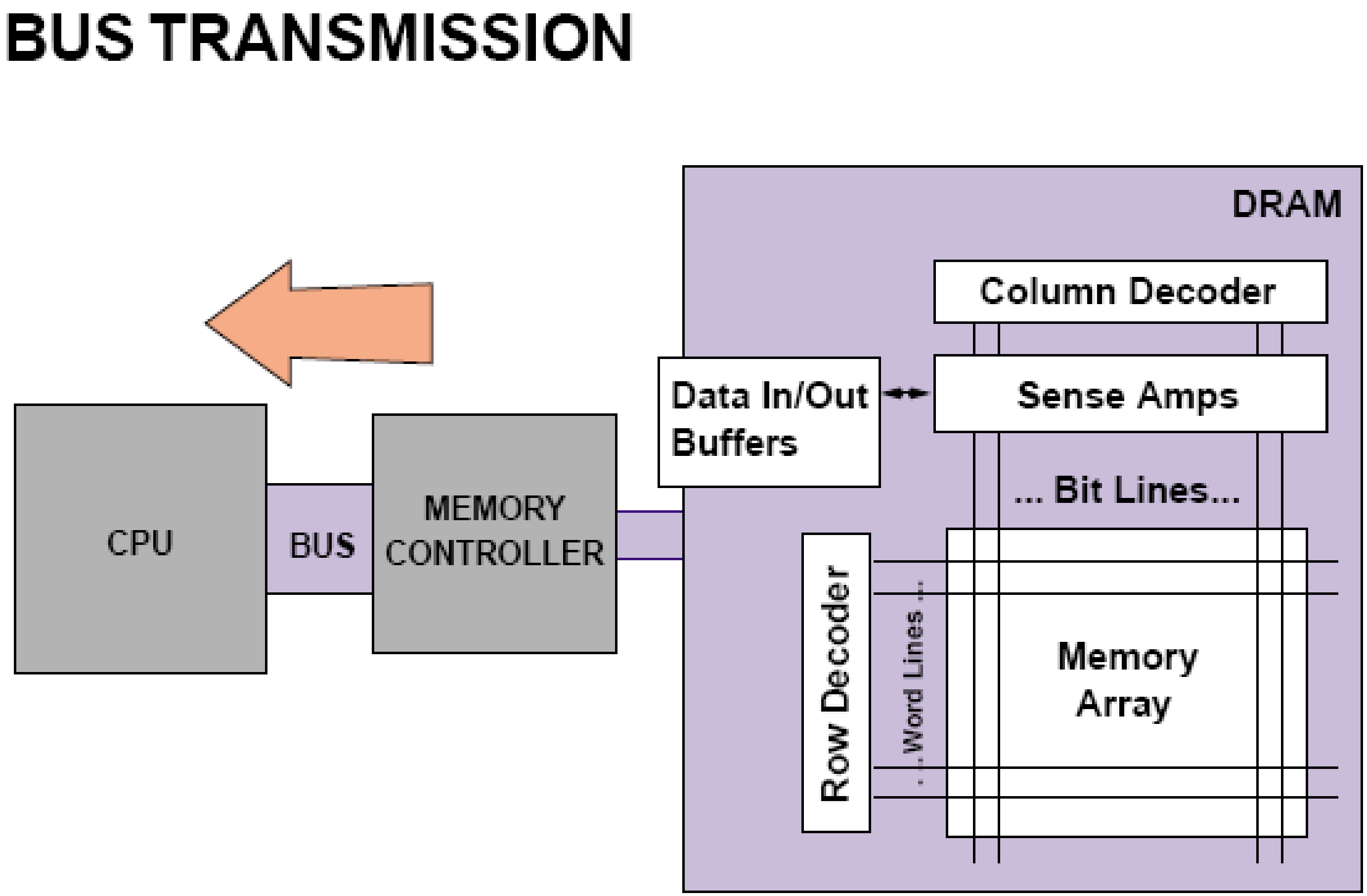
Typical DRAM Access Sequence (2)



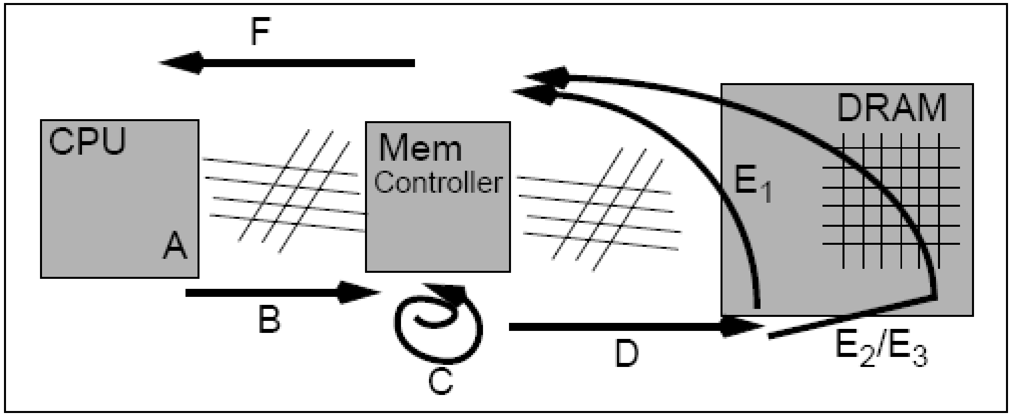
Typical DRAM Access Sequence (3)



Typical DRAM Access Sequence (4)



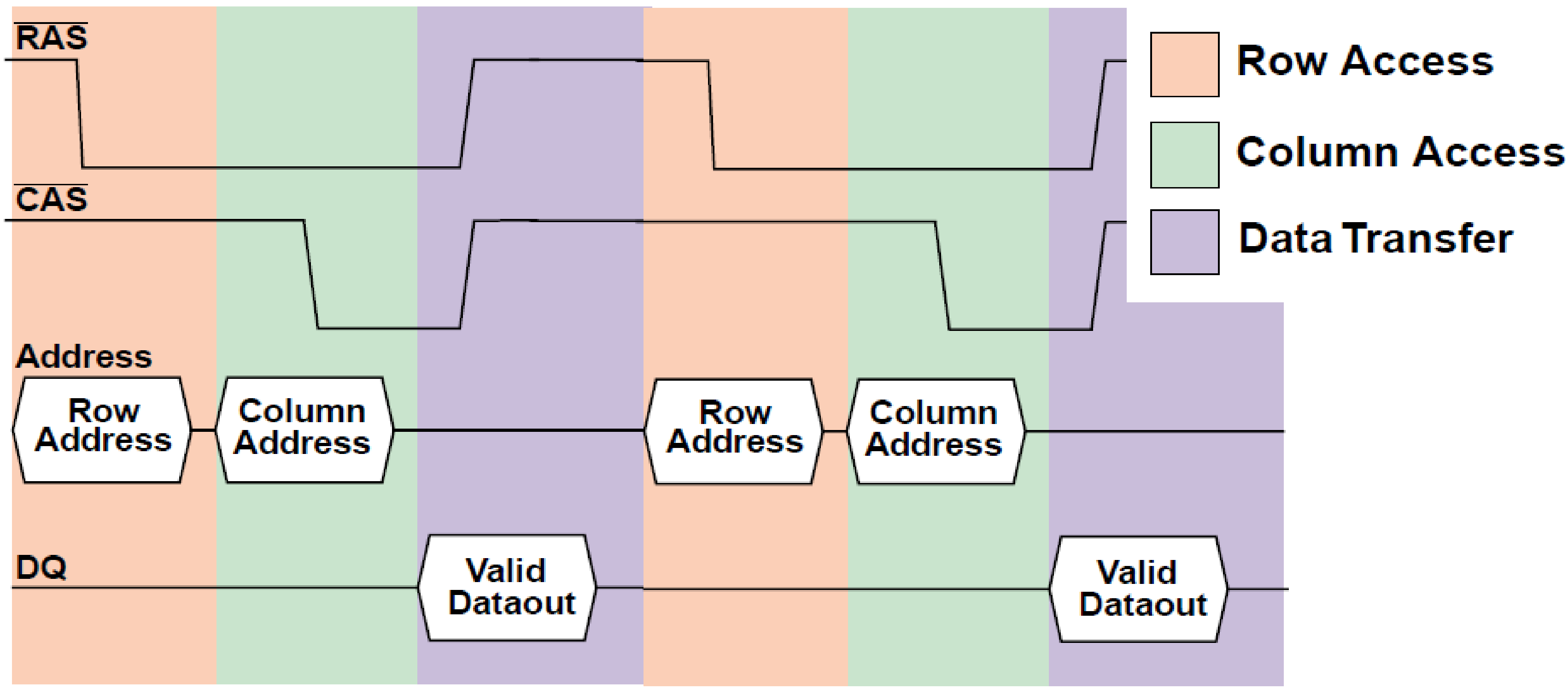
Typical DRAM Access Sequence (5)



A: Transaction request may be delayed in Queue
B: Transaction request sent to Memory Controller
C: Transaction converted to Command Sequences (may be queued)
D: Command/s Sent to DRAM
E_1: Requires only a **CAS** or
E_2: Requires **RAS + CAS** or
E_3: Requires **PRE + RAS + CAS**
F: Transaction sent back to CPU

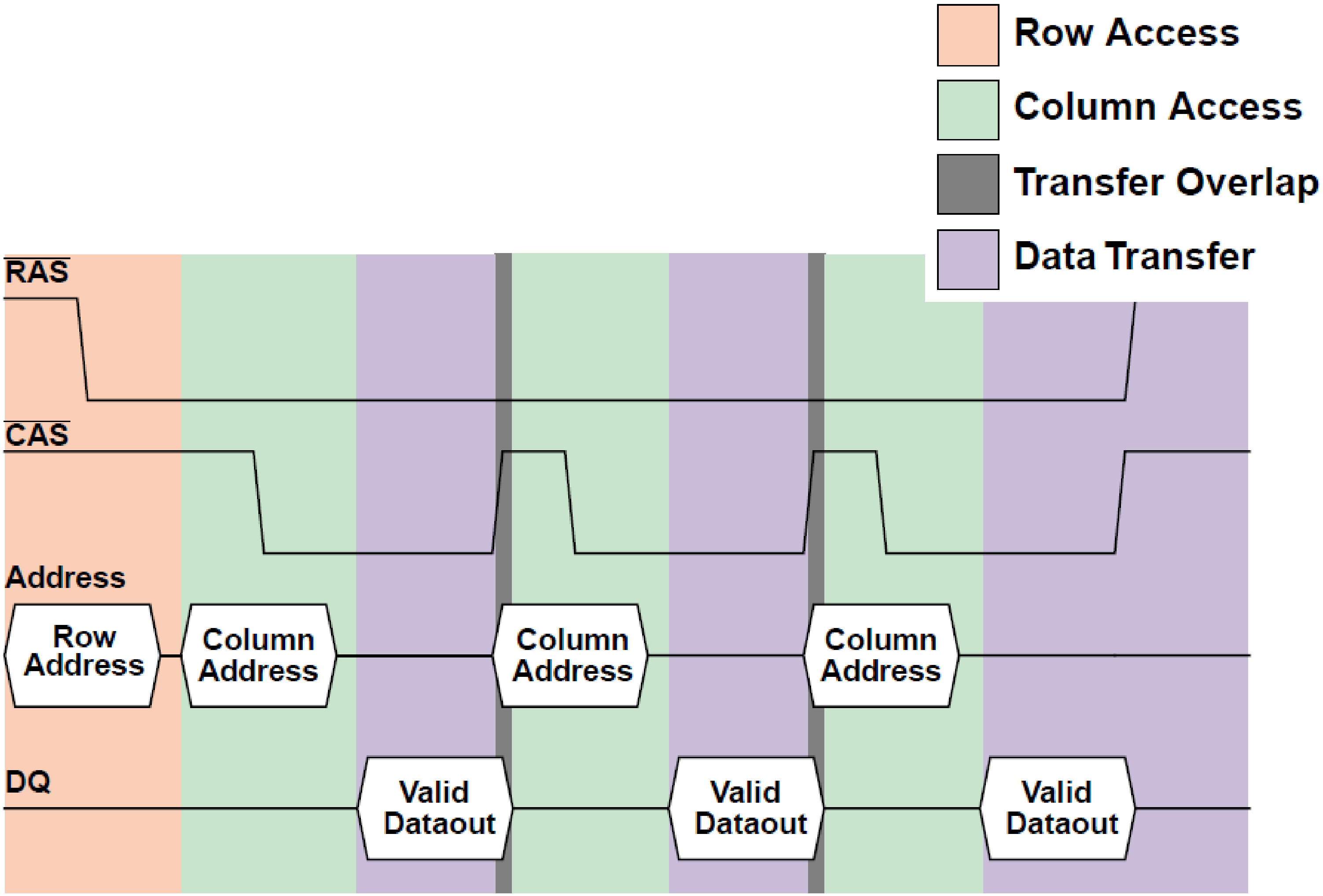
"DRAM Latency" = A + B + C + D + E + F

(Very Old) DRAM Read Timing



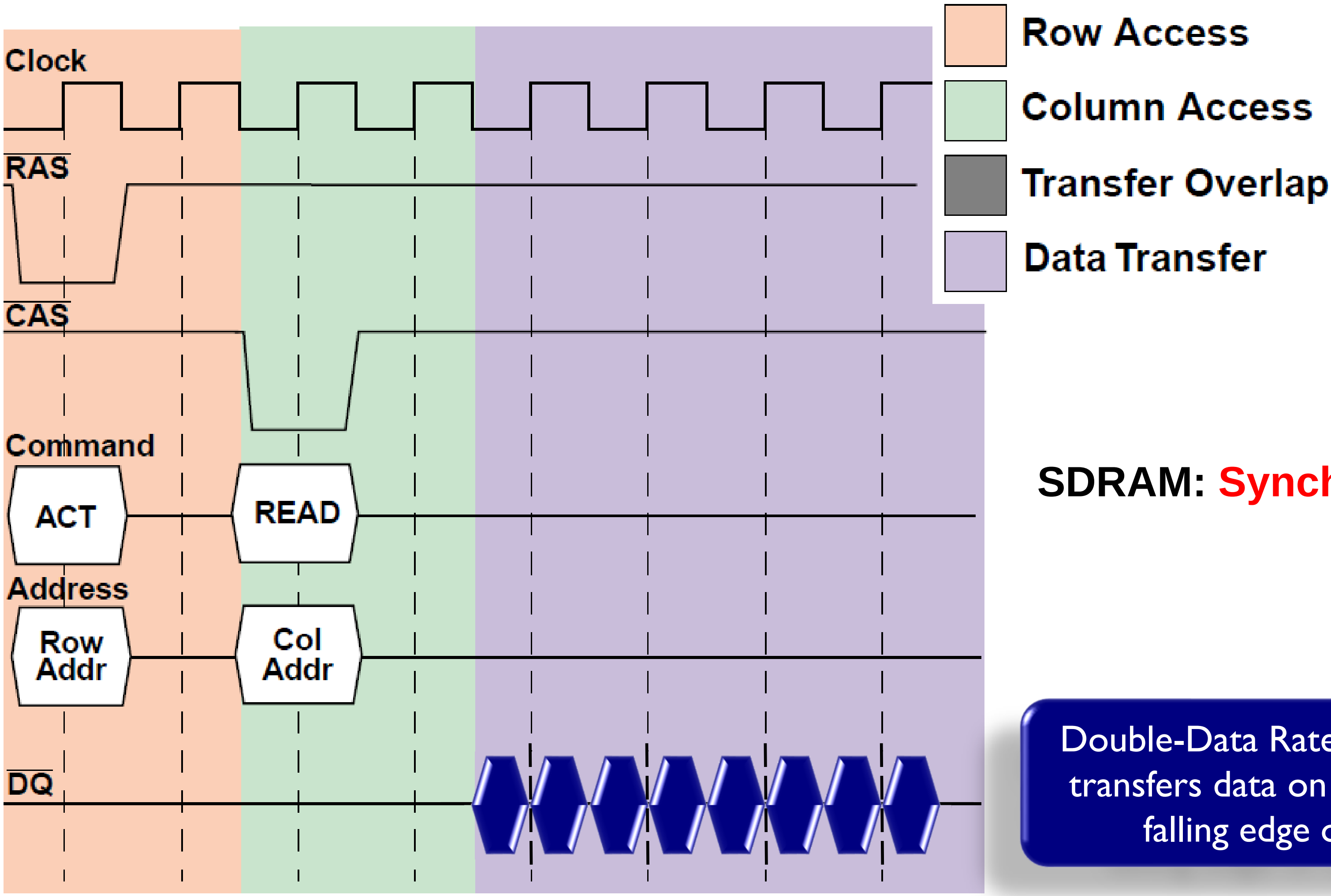
Original DRAM specified Row & Column every time

(Old) DRAM Read Timing w/ Fast-Page Mode



FPM enables multiple reads from page without RAS

(Newer) SDRAM Read Timing



SDRAM: Synchronous DRAM

Double-Data Rate (DDR) SDRAM transfers data on **both** rising and falling edge of the clock

SDRAM uses clock, supports bursts

From DRAM Array to DRAM Chip (1)

- A DRAM chip is one of the ICs you see on a **DIMM**
  - **DIMM** = Dual Inline Memory Module

=

- Typical DIMMs read/write memory in 64-bit (dword) ***beats***
- Each DRAM chip is responsible for a subset of bits in each beat
  - All DRAM chips on a DIMM are identical and work in lockstep
- The data width of a DRAM chip is the number of bits it reads/writes in a beat
  - Common examples: x4 and x8

From DRAM Array to DRAM Chip (2)

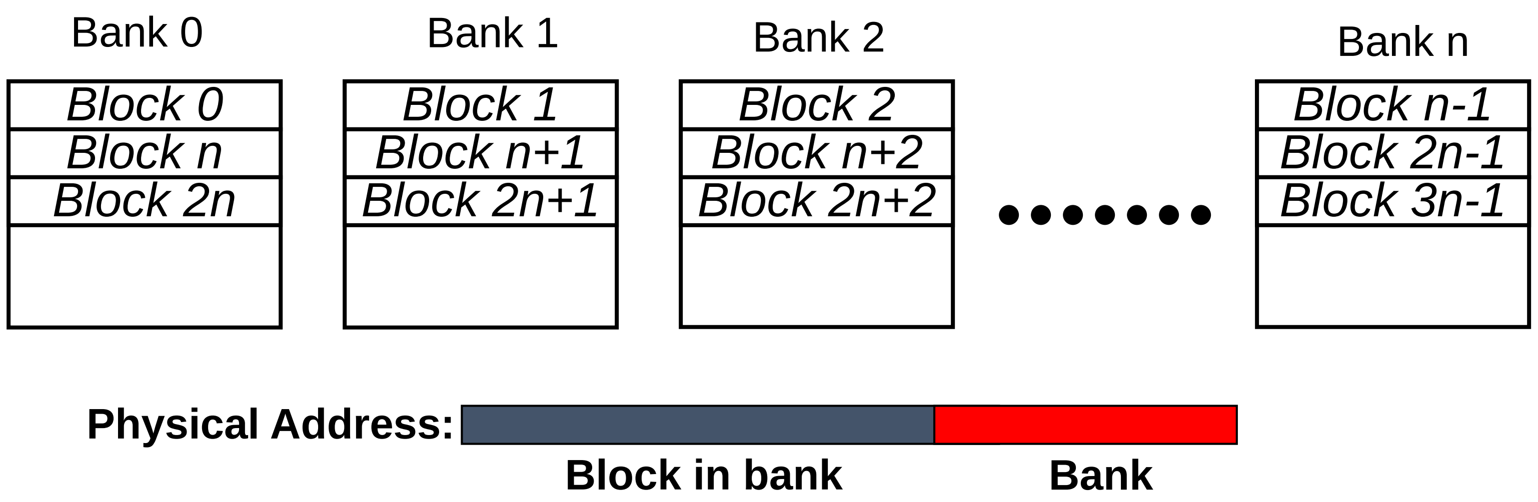
- Each DRAM Chip is internally divided into a number of **Banks**
  - Each bank is basically a fat DRAM array, i.e., columns are more than one bit (4-16 are typical)
- Each bank operates independently from other banks in the same device
- Memory controller sends the Bank ID as the higher-order bits of the row address

Banking to Improve BW

- DRAM access takes multiple cycles
- What is the miss penalty for 8 cache blocks?
  - Consider these parameters:
    - 1 cycle to send address
    - 10 cycle to read the row containing the cache block
    - 4 cycles to send-out the data (assume DDR)
  - (1 + 10 + 4) x 8 = 120
- *How can we speed this up?*

Simple Interleaved Main Memory

- Divide memory into n banks, “interleave” addresses across them so that cache-block A is
  - in bank “A mod n”
  - at block “A div n”



- Can access one bank while another one is busy

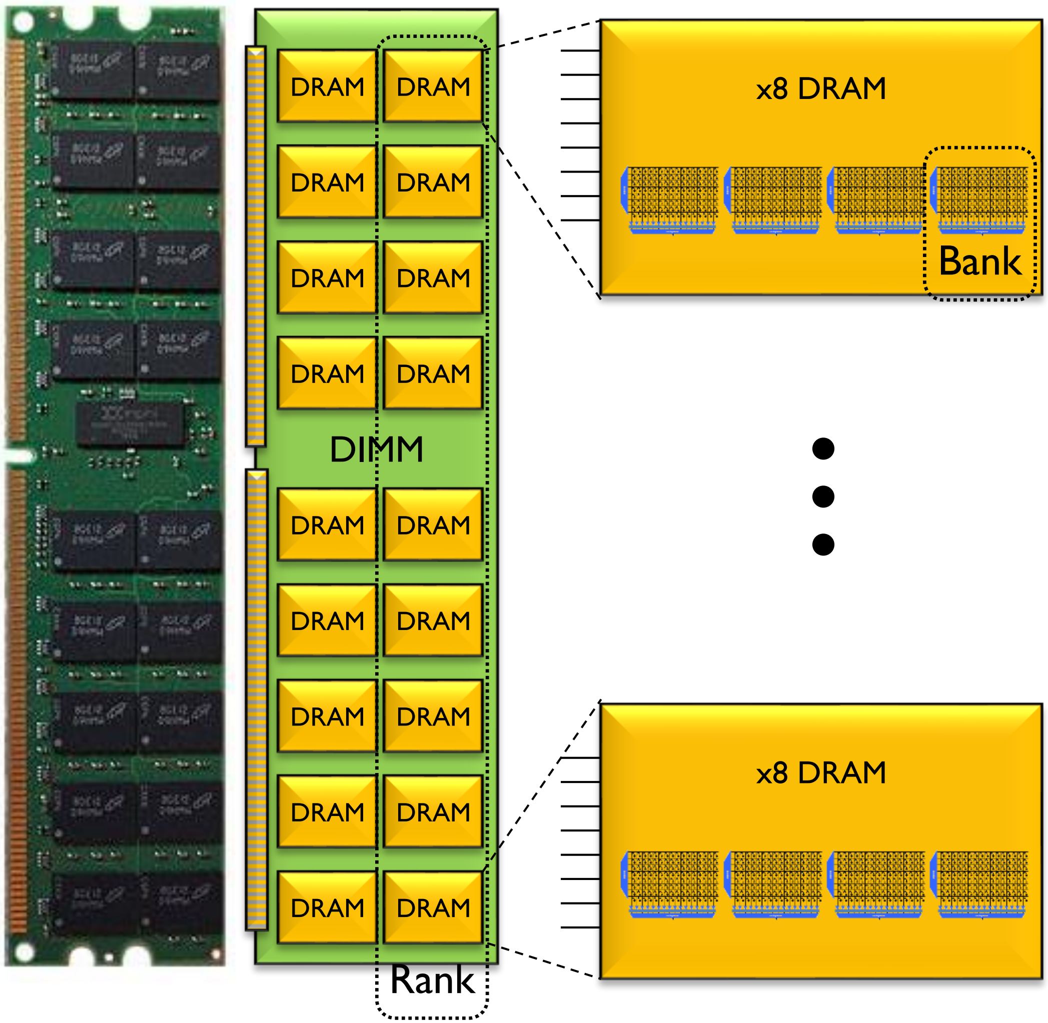
Banking to Improve BW

- In previous example, if we had 8 banks, how long would it take to receive all 8 blocks?
  - $(1 + 10 + 4) + 7 \times 4 = 43$ cycles

→ Interleaving increases memory bandwidth w/o a wider bus

Use parallelism in memory banks to hide memory latency

DRAM Organization



All banks within the rank share all address and control pins

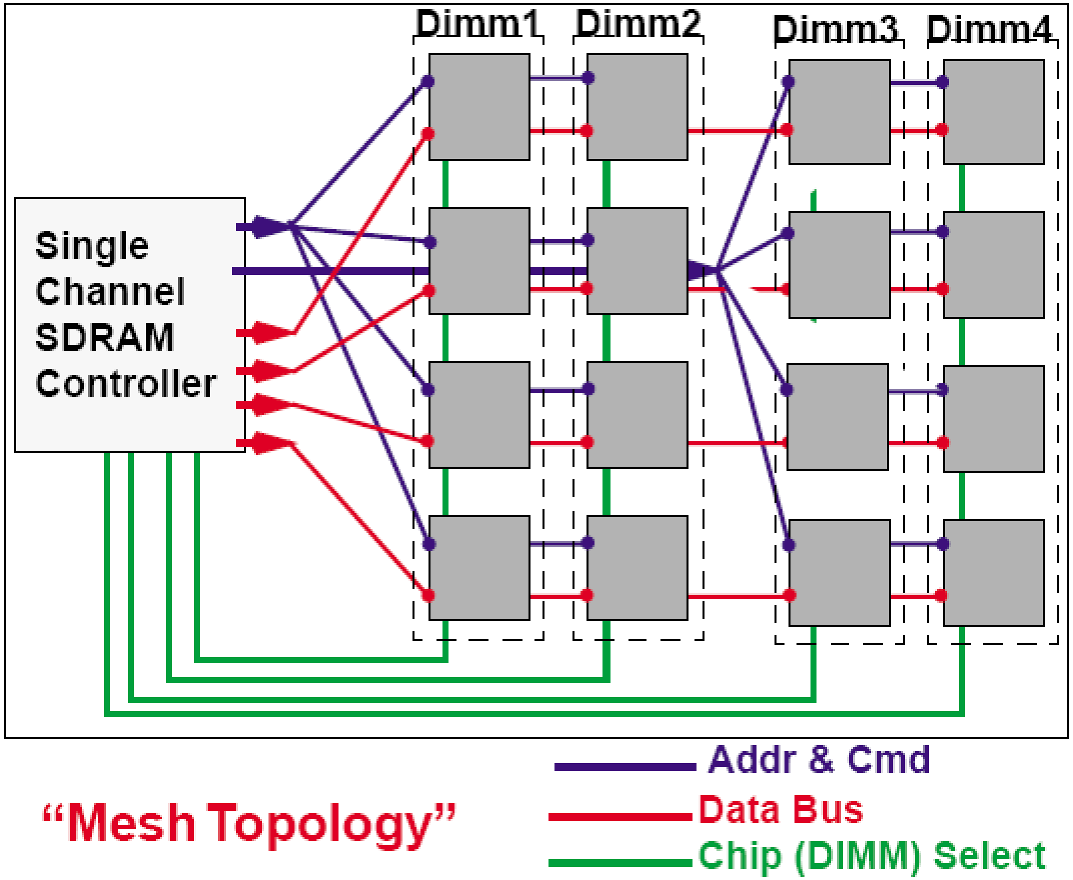
All banks are independent, but can only talk to one bank at a time

x8 means each DRAM outputs 8 bits, need 8 chips for DDRx (64-bit)

Why 9 chips per rank? 64 bits data, 8 bits ECC

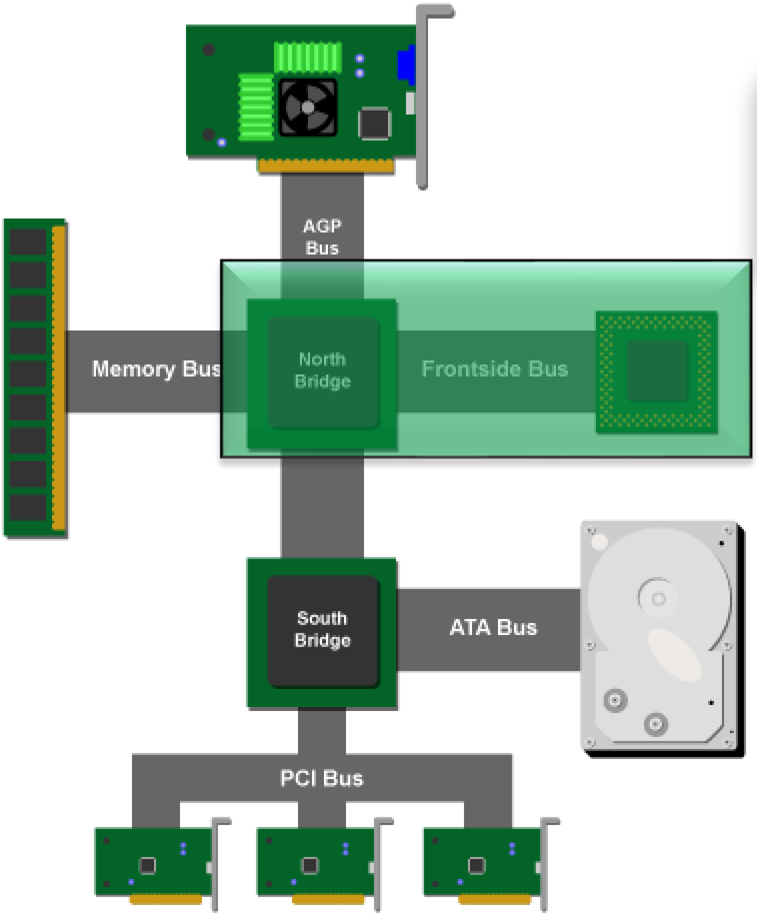
Dual-rank x8 (2Rx8) DIMM

SDRAM Topology



"Mesh Topology"

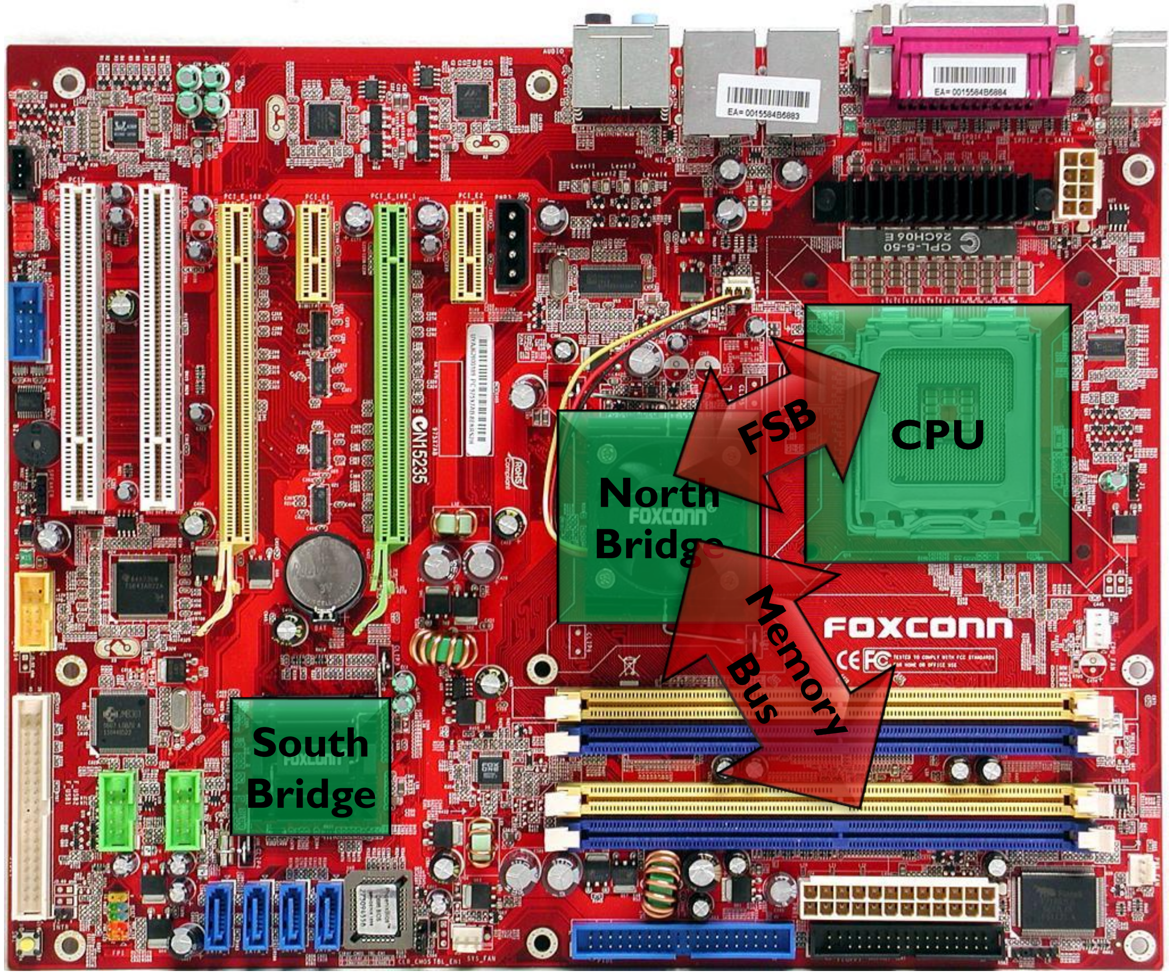
CPU-to-Memory Interconnect (1)



North Bridge can be Integrated onto CPU chip to reduce latency

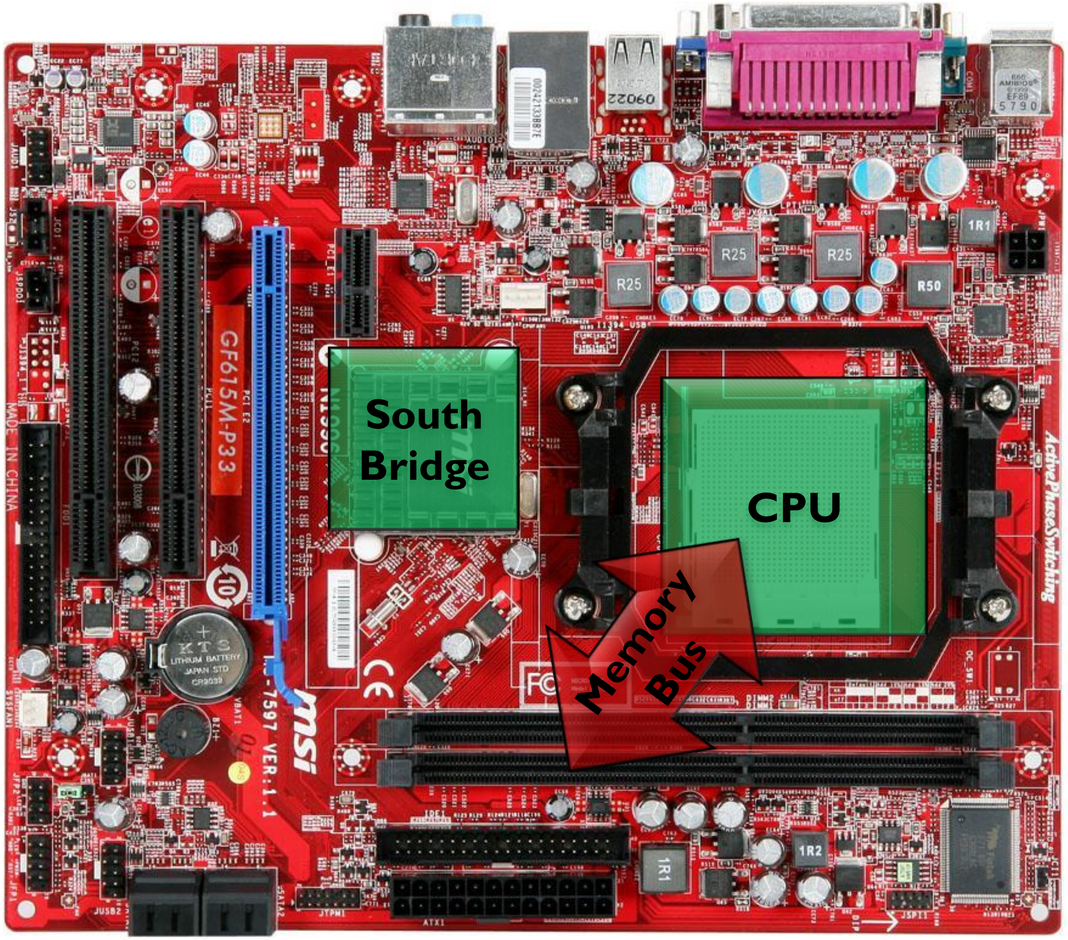
Figure from ArsTechnica

CPU-to-Memory Interconnect (2)



Discrete North and South Bridge chips (Old)

CPU-to-Memory Interconnect (3)



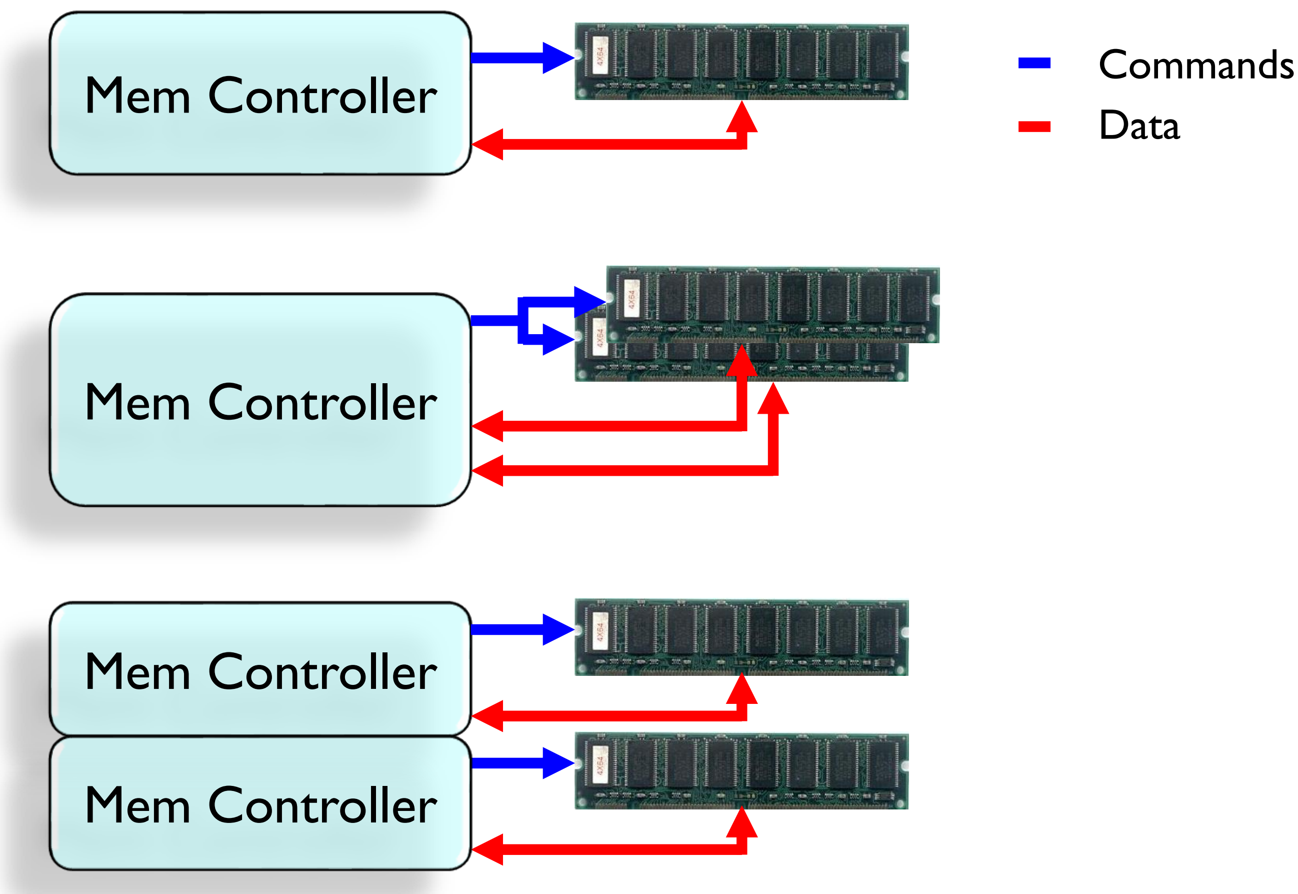
Integrated North Bridge (Modern Day)

Memory Channels

One controller
One 64-bit channel

One controller
Two 64-bit channels

Two controllers
Two 64-bit channels



Use multiple channels for more bandwidth

Memory-Level Parallelism (MLP)

- What if memory latency is 10000 cycles?
  - Runtime dominated by waiting for memory
  - What matters is ***overlapping memory accesses***
- *Memory-Level Parallelism (MLP)*:
  - "Average number of outstanding memory accesses when at least one memory access is outstanding."
- MLP is a metric
  - ***Not*** a fundamental property of workload
  - Dependent on the microarchitecture
- With high-enough MLP, you can hide arbitrarily large memory latencies

AMAT with MLP

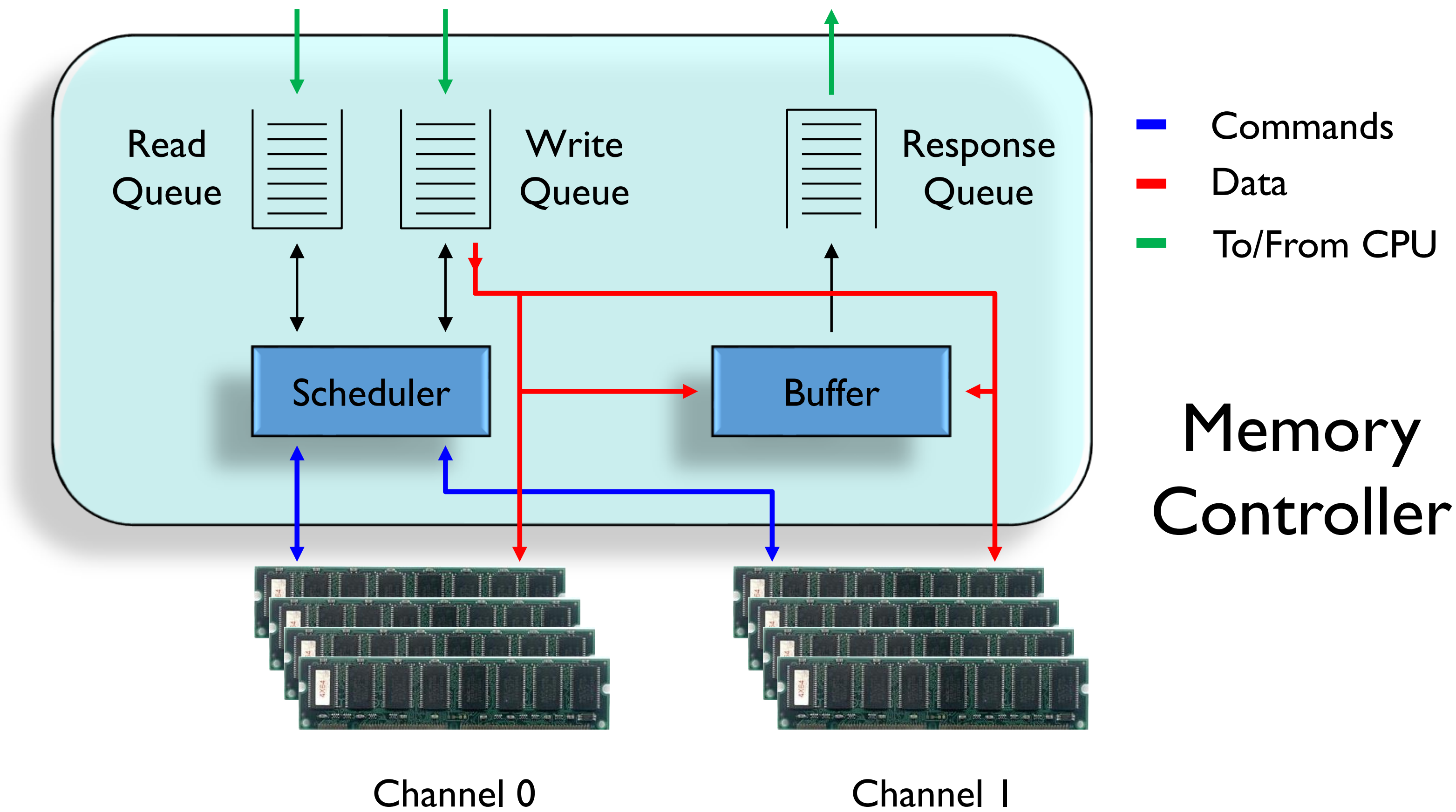
- If …
  cache hit is 10 cycles (core to L1 and back)
  memory access is 100 cycles (core to mem and back)

- Then …
  at 50% miss ratio: AMAT = 0.5×10+0.5×100 = 55

- Unless MLP is >1.0, then…
  at 50% mr, 1.5 MLP: AMAT = (0.5×10+0.5×100)/1.5 = 37
  at 50% mr, 4.0 MLP: AMAT = (0.5×10+0.5×100)/4.0 = 14

In many cases, MLP dictates performance

Memory Controller (1)



Memory Controller (2)

- Memory controller connects CPU and DRAM
- Receives requests after cache misses in LLC
  - Possibly originating from multiple cores
- Complicated piece of hardware, handles:
  - DRAM Refresh
  - Row-Buffer Management Policies
  - Address Mapping Schemes
  - Request Scheduling

Request Scheduling in MC (1)

- Write buffering
  - Writes can wait until reads are done
- Controller queues DRAM commands
  - Usually into per-bank queues
  - Allows easily reordering ops. meant for same bank
- Common policies:
  - *First-Come-First-Served (FCFS)*
  - *First-Ready—First-Come-First-Served (FR-FCFS)*

Request Scheduling in MC (2)

- First-Come-First-Served
  - Oldest request first
- First-Ready—First-Come-First-Served
  - *Prioritize column changes over row changes*
  - *Skip over older conflicting requests*
  - Find row hits (on queued requests)
    - Find oldest
    - If no conflicts with in-progress request → good
    - Otherwise (if conflicts), try next oldest

Stony Brook University

Request Scheduling in MC (3)

- Why is it hard?
- Tons of timing constraints in DRAM
  - t_{WTR}: Min. cycles before read after a write
  - t_{RC}: Min. cycles between consecutive open in bank
  - …
- Simultaneously track resources to prevent conflicts
  - Channels, banks, ranks, data bus, address bus, row buffers
  - Do it for many queued requests at the same time
    … while not forgetting to do refresh

Row-Buffer Management Policies

- **Open-page Policy**
  – After access, keep page in DRAM row buffer
  – Next access to same page → lower latency
  – If access to different page, must close old one first
    - Good if lots of spatial locality

- **Close-page Policy**
  – After access, immediately close page in DRAM row buffer
  – Next access to different page → lower latency
  – If access to different page, old one already closed
    - Good if no locality (random access)

Address Mapping Schemes (1)

- Question: How to map a physical addr to <channel ID, rank ID, bank ID, row ID, column ID>?
  - Goal: efficiently exploit channel/rank/bank level parallelism
- Multiple ***independent*** channels → max parallelism
  - Map consecutive cache lines to different channels
- Single channel, Multiple ranks/banks → OK parallelism
  - Limited by shared address and/or data pins
  - Map consecutive cache lines to banks within same rank
    - ***Reads*** from same rank are faster than from different ranks
- Accessing different rows from one bank is slowest
  - All requests serialized, regardless of row-buffer mgmt. policies
  - Rows mapped to same bank should avoid spatial locality
- Column mapping depends on row-buffer mgmt (why?)

Address Mapping Schemes (2)

[… … … … bank column …]

0x00000	0x00400	0x00800	0x00C00
0x00100	0x00500	0x00900	0x00D00
0x00200	0x00600	0x00A00	0x00E00
0x00300	0x00700	0x00B00	0x00F00

[… … … … column bank …]

0x00000	0x00100	0x00200	0x00300
0x00400	0x00500	0x00600	0x00700
0x00800	0x00900	0x00A00	0x00B00
0x00C00	0x00D00	0x00E00	0x00F00

Address Mapping Schemes (3)

- Example Open-page Mapping Scheme:

  High Parallelism: [row rank bank column channel offset]

  Easy Expandability: [channel rank row bank column offset]

- Example Close-page Mapping Scheme:

  High Parallelism: [row column rank bank channel offset]

  Easy Expandability: [channel rank row column bank offset]

Overcoming Memory Latency

- Caching
  - Reduce average latency by avoiding DRAM altogether
  - Limitations
    - Capacity (programs keep increasing in size)
    - Compulsory misses
- Prefetching
  - Guess what will be accessed next
  - Bring it to the cache ahead of time