Lecture 16:
Main Memory

Vassilis Papaefstathiou
Iakovos Mavroidis

Computer Science Department
University of Crete

Memory Hierarchy

Capacity
Access Time
Cost

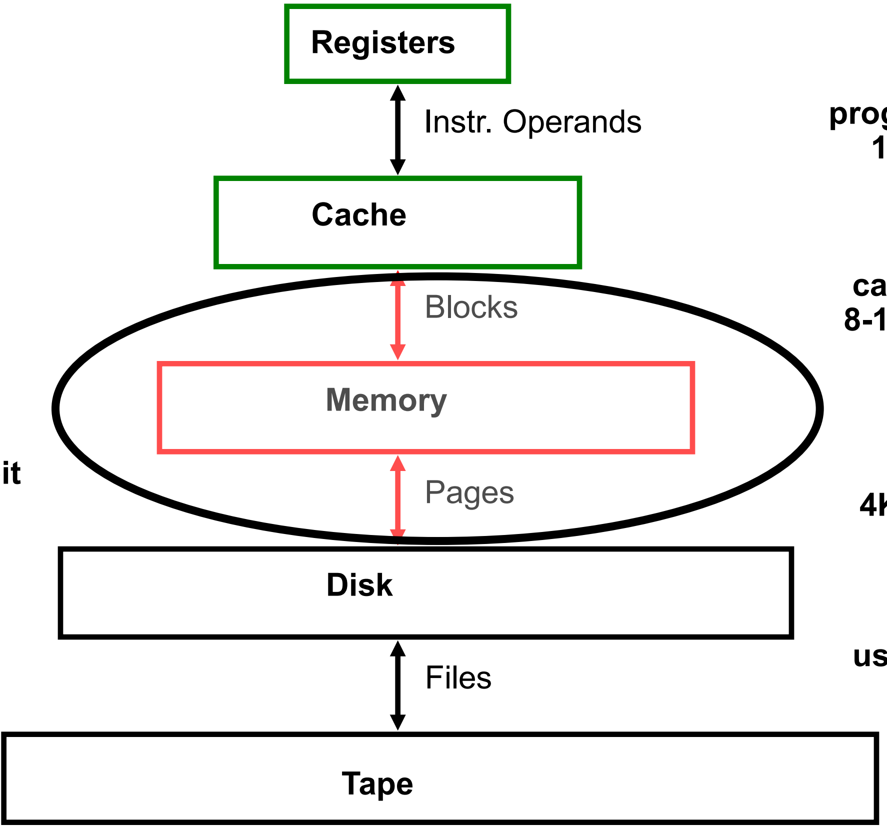
CPU Registers
100s Bytes
<10s ns

Cache
K Bytes
10-100 ns
1-0.1 cents/bit

Main Memory
M Bytes
200ns- 500ns
$.0001-.00001 cents /bit

Disk
G Bytes, 10 ms
(10,000,000 ns)
10^{-5} - 10^{-6} cents/bit

Tape
infinite
sec-min
10^{-8}



Registers

Instr. Operands

Cache

Blocks

Memory

Pages

Disk

Files

Tape

Staging
Xfer Unit

prog./compiler
1-8 bytes

cache cntl
8-128 bytes

OS
4K-16K bytes

user/operator
Mbytes

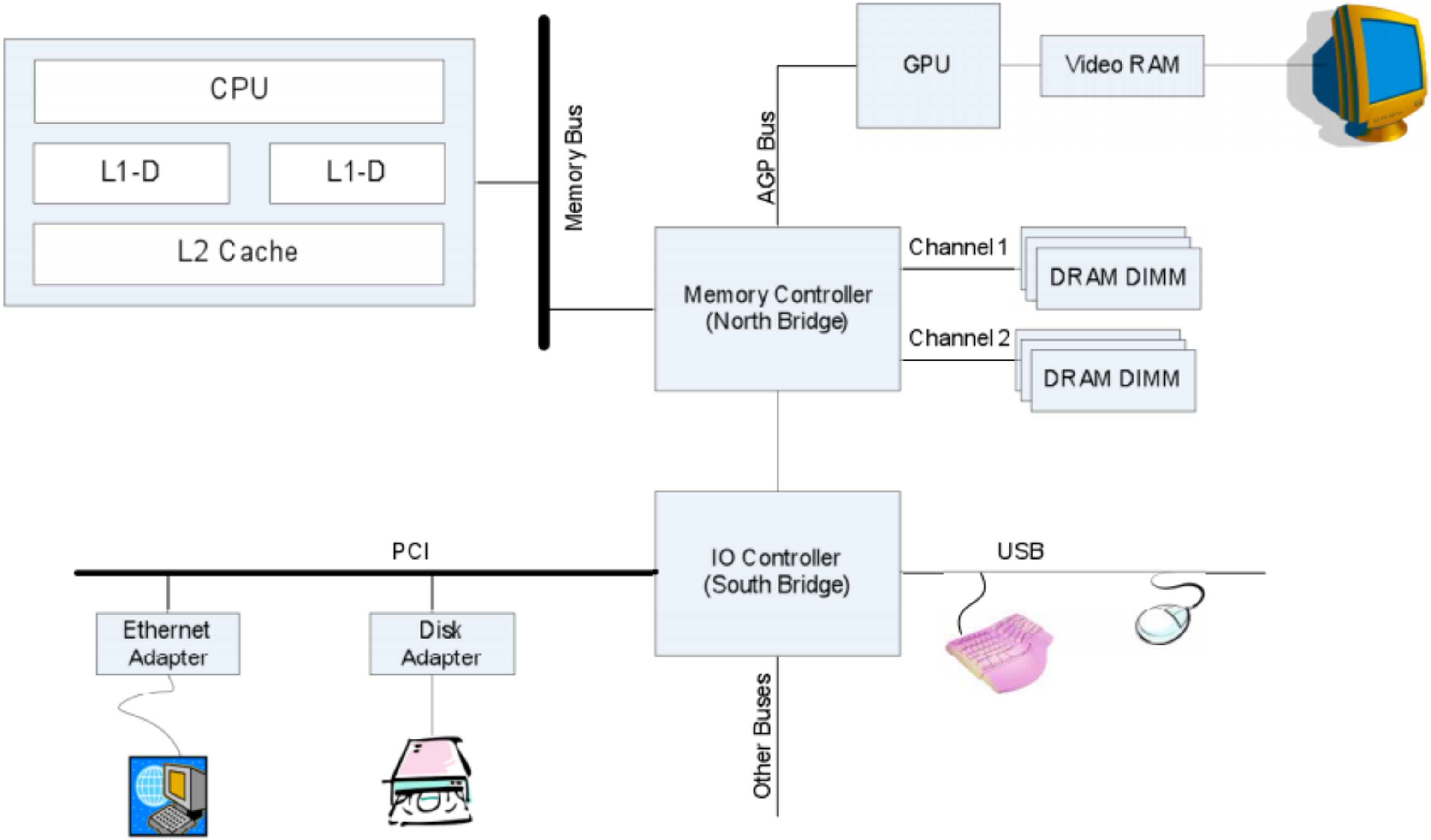
Upper Level

faster

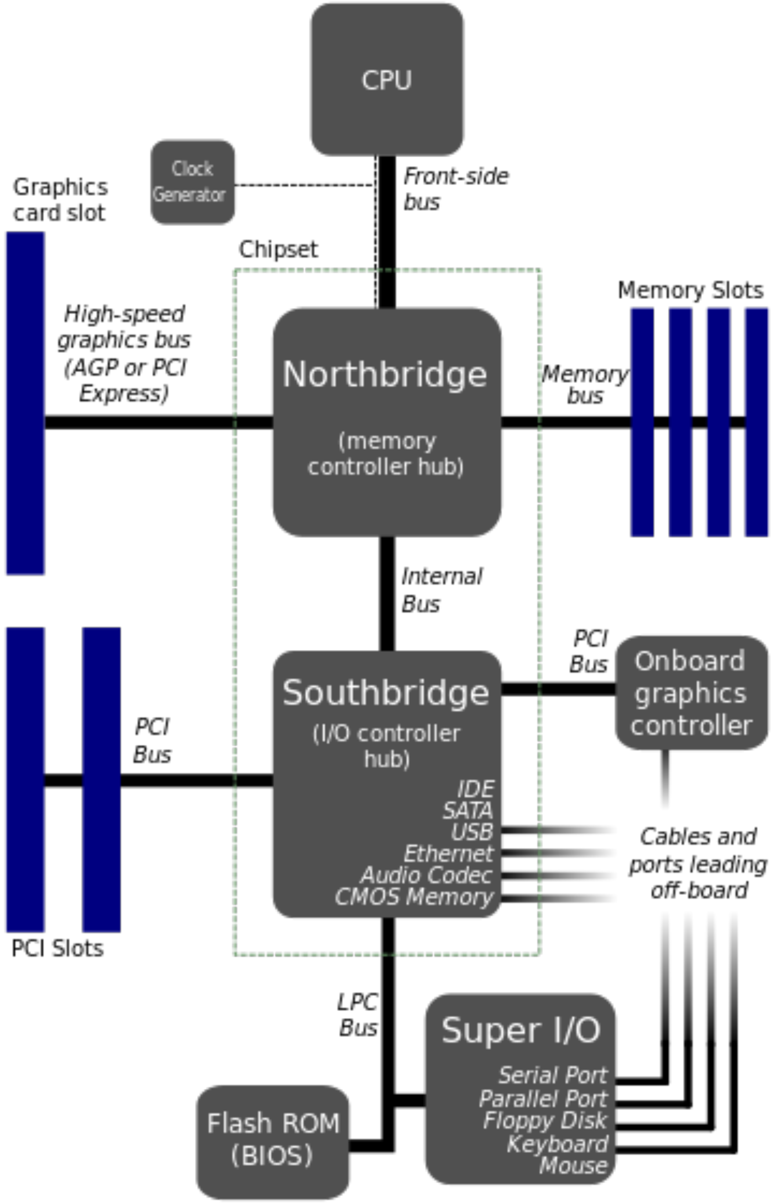
Larger

Lower Level

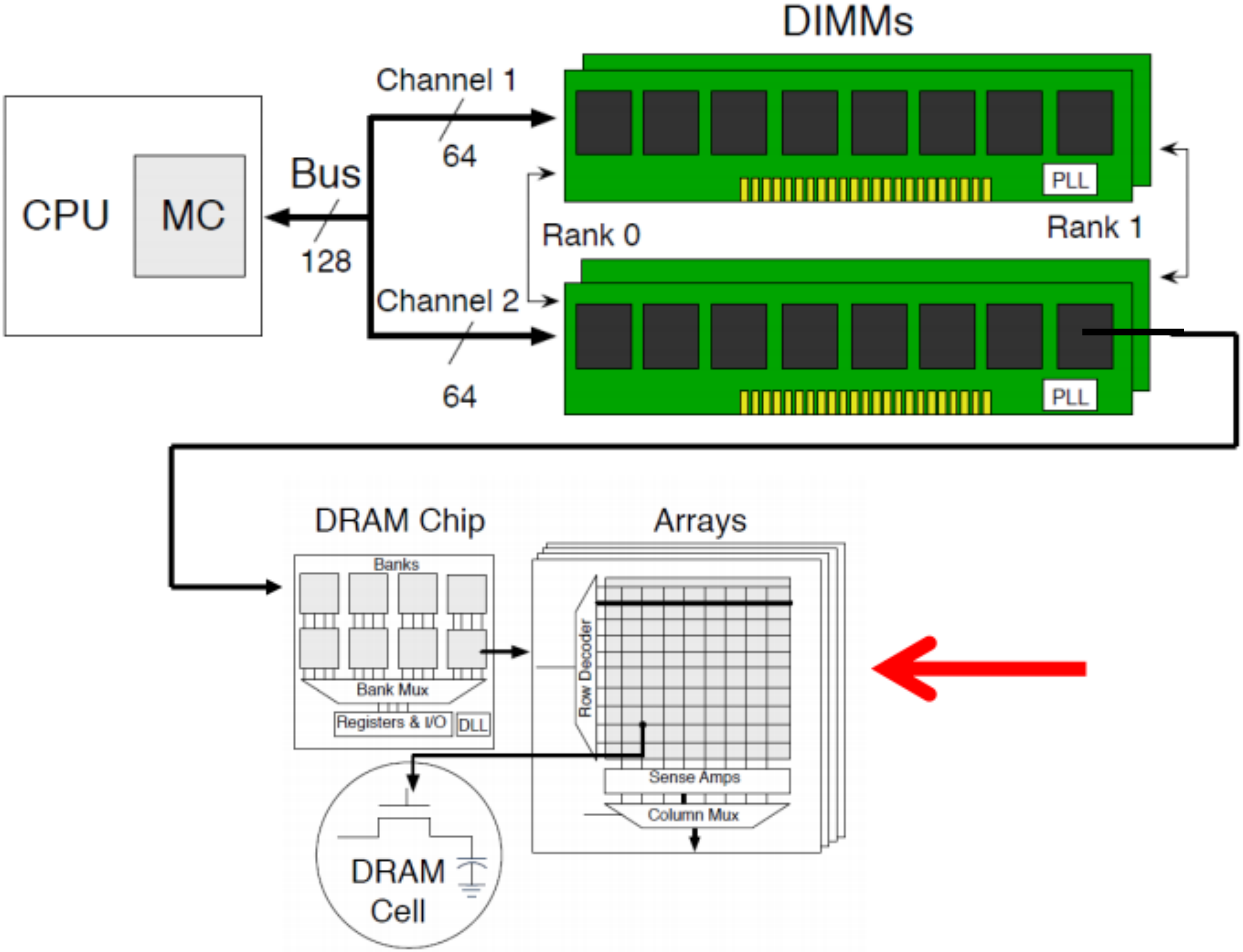
Computer System Overview



Typical Chipset Layout

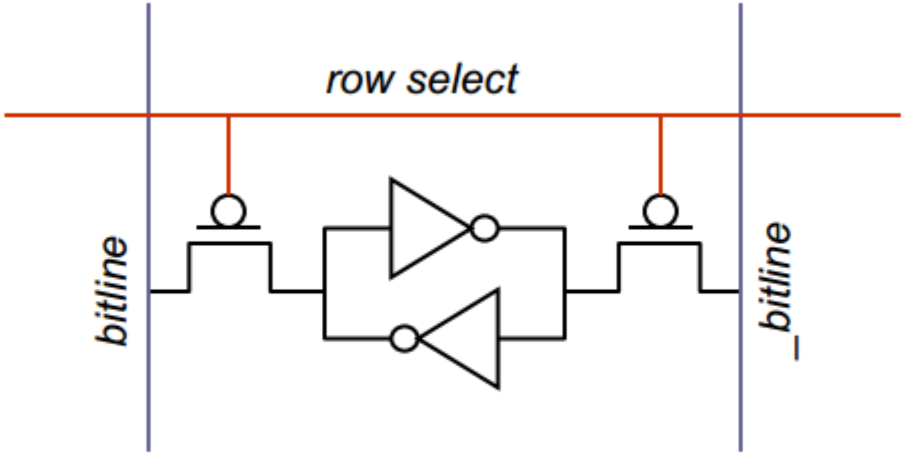


Main Memory Overview



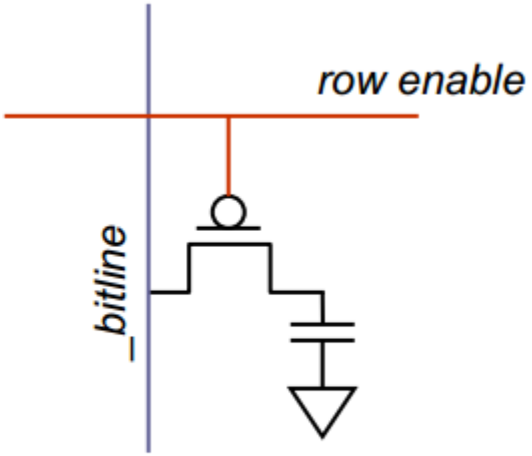
SRAM vs. DRAM

Static Random Access Mem.



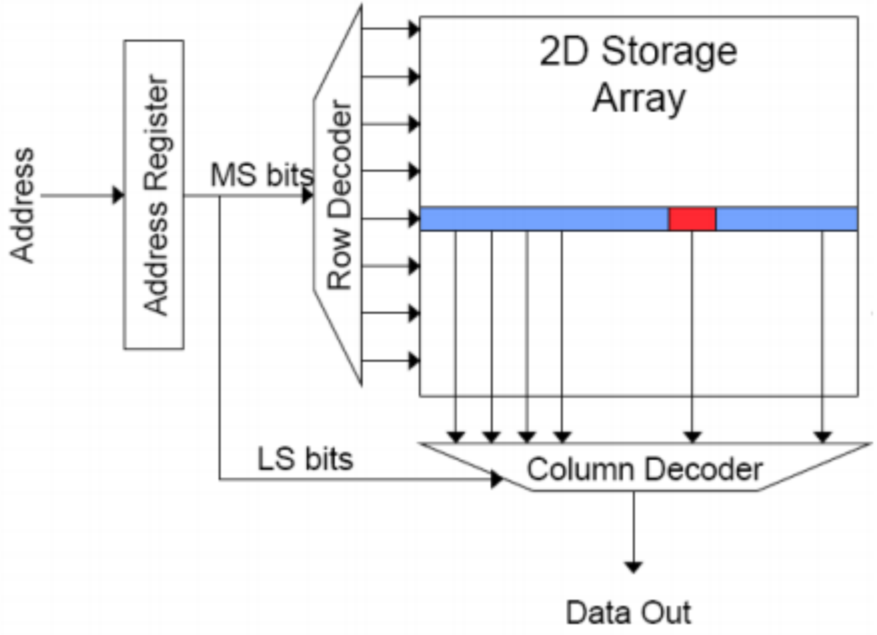
- 6T vs. 1T1C
  - Large (~6-10x)
- Bitlines driven by transistors
  - Fast (~10x)

Dynamic Random Access Mem.



- Bits stored as charges on node capacitance (non-restorative)
  - Bit cell loses charge when read
  - Bit cell loses charge over time
- Must periodically refresh
  - Once every 10s of ms

Memory Bank Organization

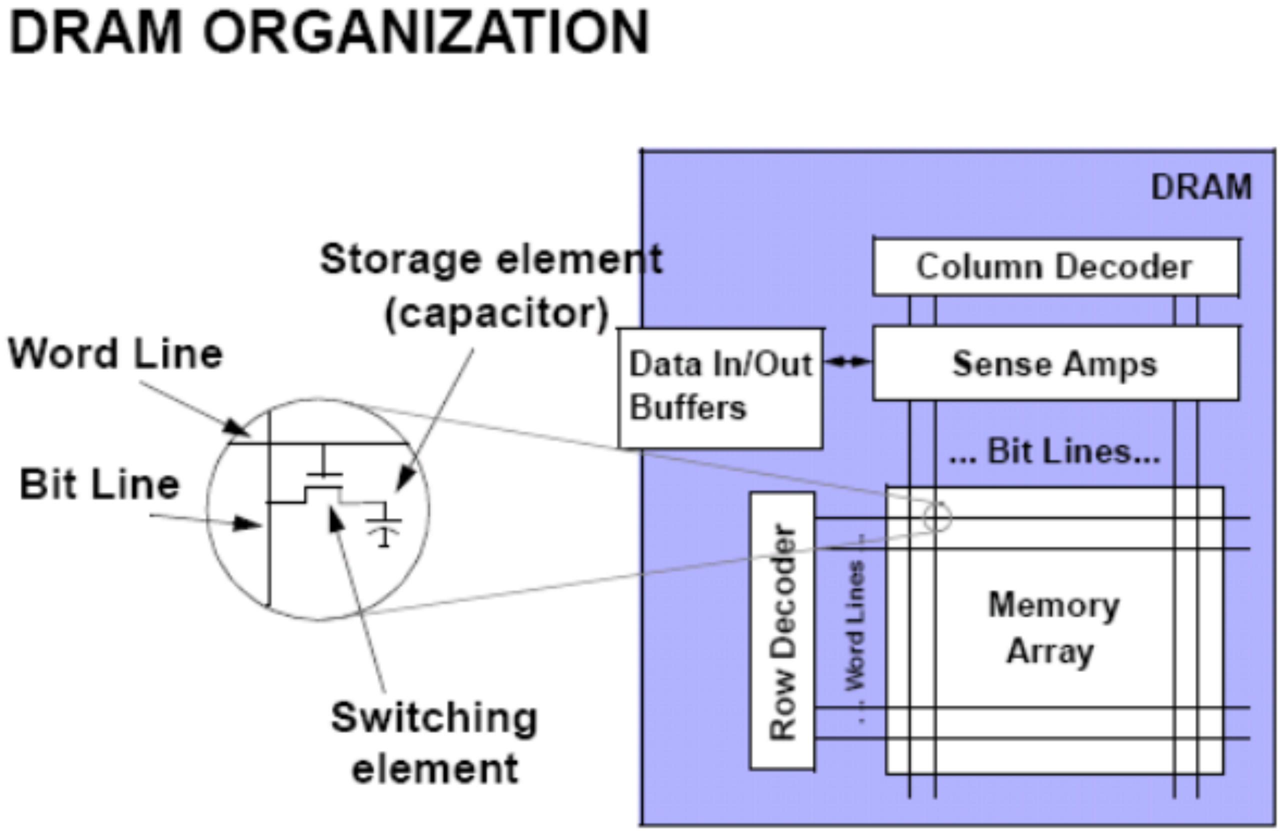


- Read access sequence
  - Decode row address & drive word-lines
  - Selected bits drive bit-lines
    - Entire row read
  - Amplify row data
  - Decode column address & select subset of row
  - Send to output
  - Precharge bit-lines for next access

Memory Terminology

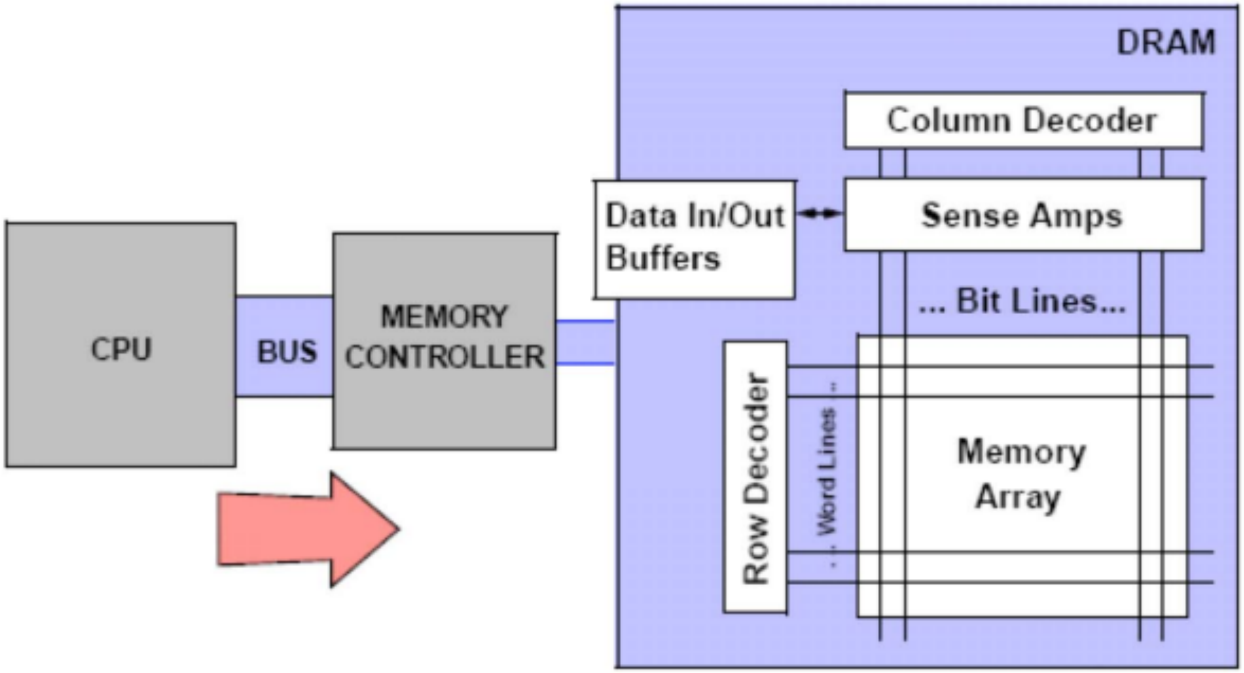
- Access time (latency)
  - Time from issuing and address to data out
- Cycle time
  - Minimum time between two request (repeat rate)
- Bandwidth
  - Bytes/unit of time we can extract from the memory
    - Peak: ignore initial latency
    - Sustained: include initial latency
- Concurrency
  - Number of accesses executing in parallel or overlapped manner
  - Can help increase bandwidth or improve latency

DRAM Basic Operation



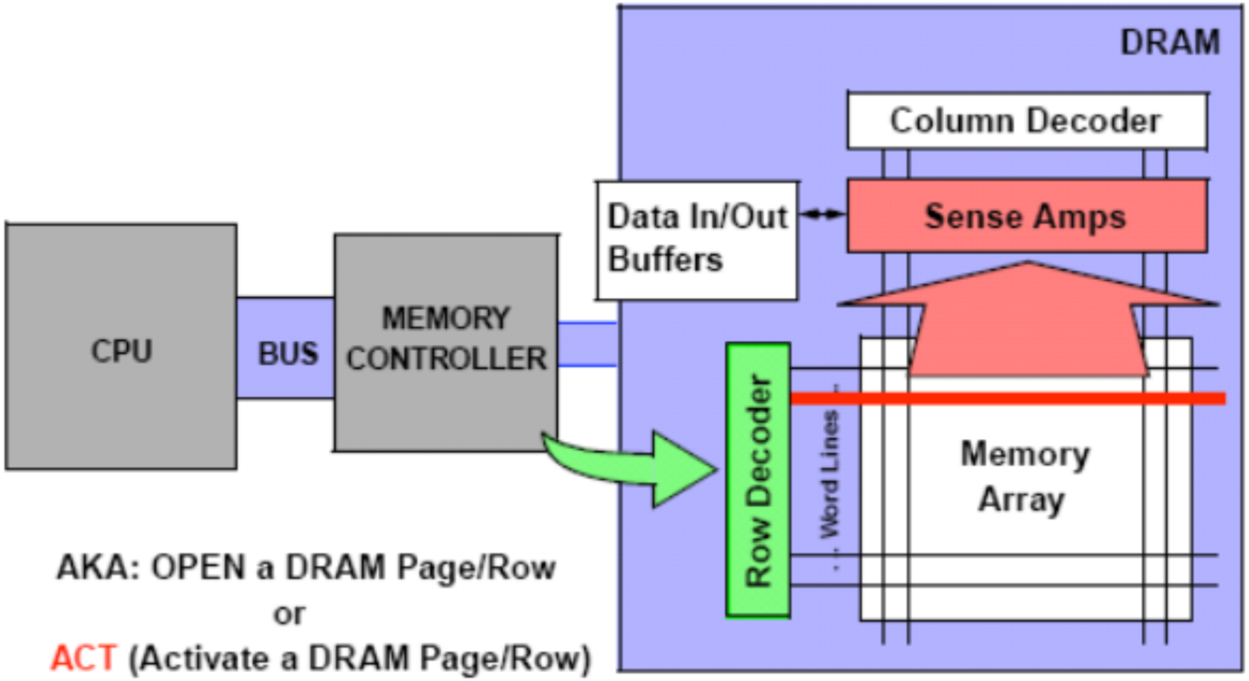
Basic DRAM operation (1)

BUS TRANSMISSION



Basic DRAM Operation (2)

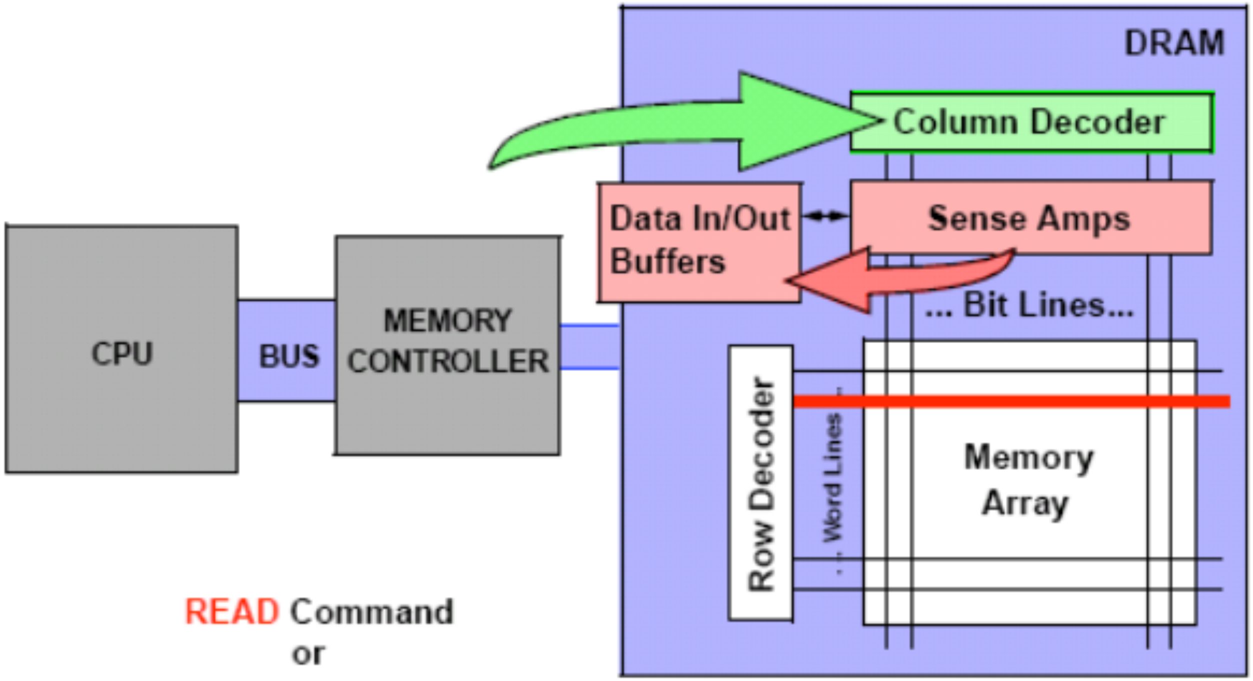
[PRECHARGE and] ROW ACCESS



AKA: OPEN a DRAM Page/Row
or
ACT (Activate a DRAM Page/Row)
or
RAS (Row Address Strobe)

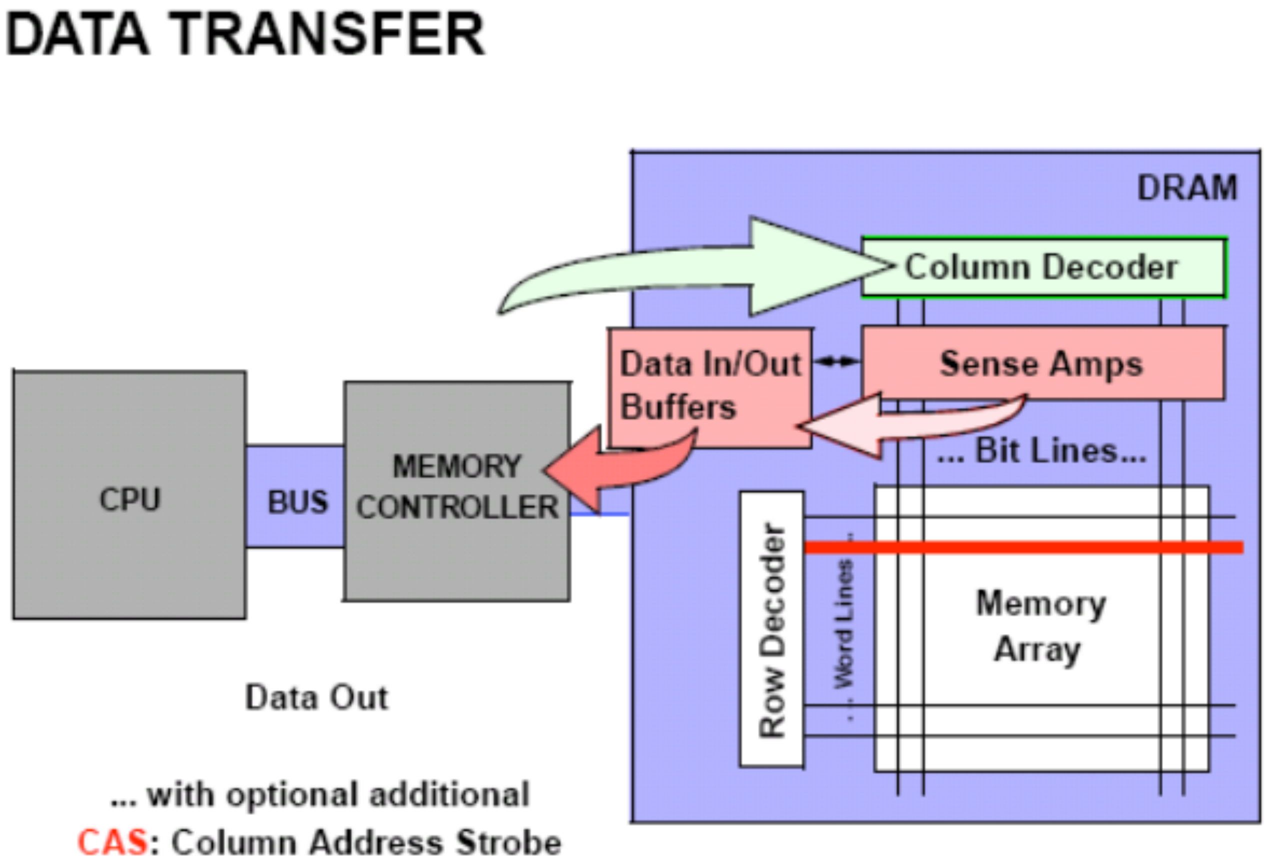
Basic DRAM Operation (3)

COLUMN ACCESS



READ Command
or
CAS: Column Address Strobe

Basic DRAM Operation (4)



- Not shown: precharge time, refresh time

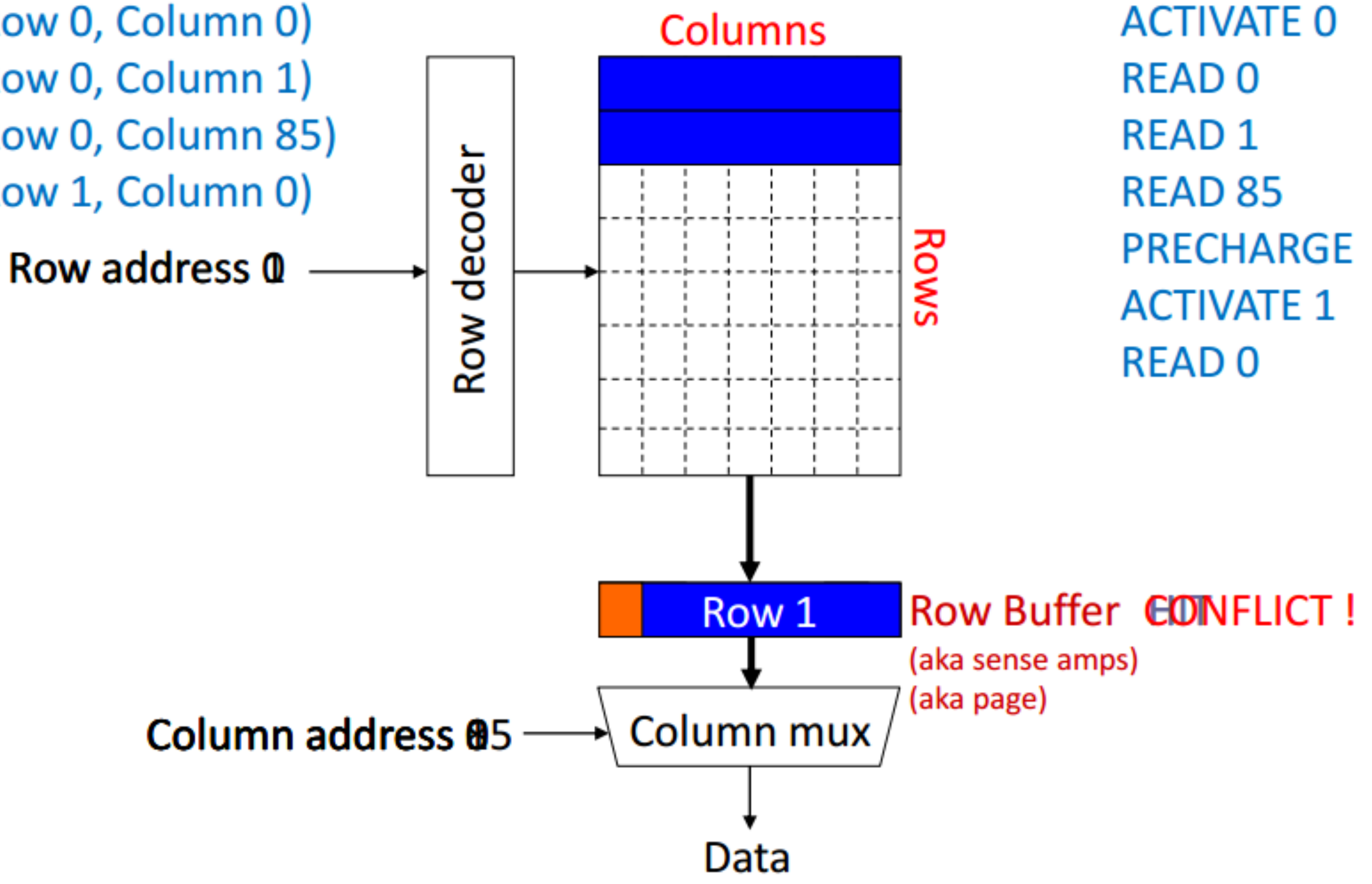
DRAM: Basic Operation

Addresses

(Row 0, Column 0)
(Row 0, Column 1)
(Row 0, Column 85)
(Row 1, Column 0)

Commands

ACTIVATE 0
READ 0
READ 1
READ 85
PRECHARGE
ACTIVATE 1
READ 0



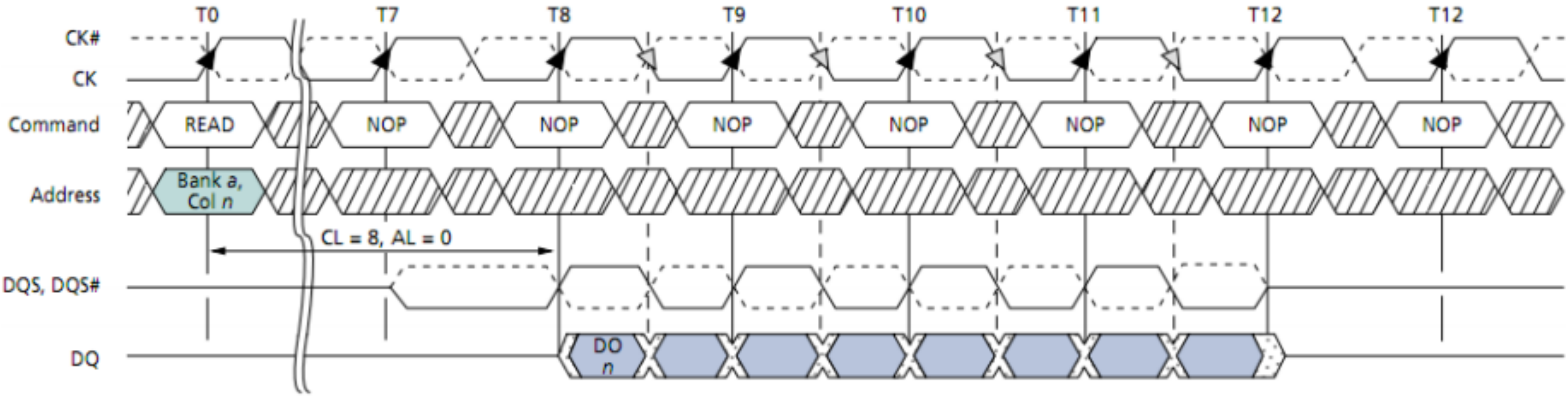
DRAM: Basic Operation

- Access to an “open row”
  - No need for ACTIVATE command
  - READ/WRITE to access row buffer

- Access to a “closed row”
  - If another row already active, must first issue PRECHARGE
  - ACTIVATE to open new row
  - READ/WRITE to access row buffer
  - Optional: PRECHARGE after READ/WRITEs finished

DRAM: Burst

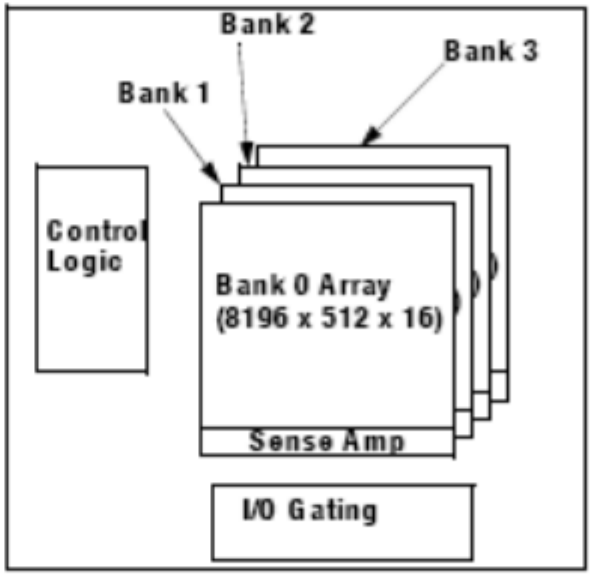
- Each READ/WRITE command can transfer multiple words (8 in DDR3)
- DRAM channel clocked faster than DRAM core



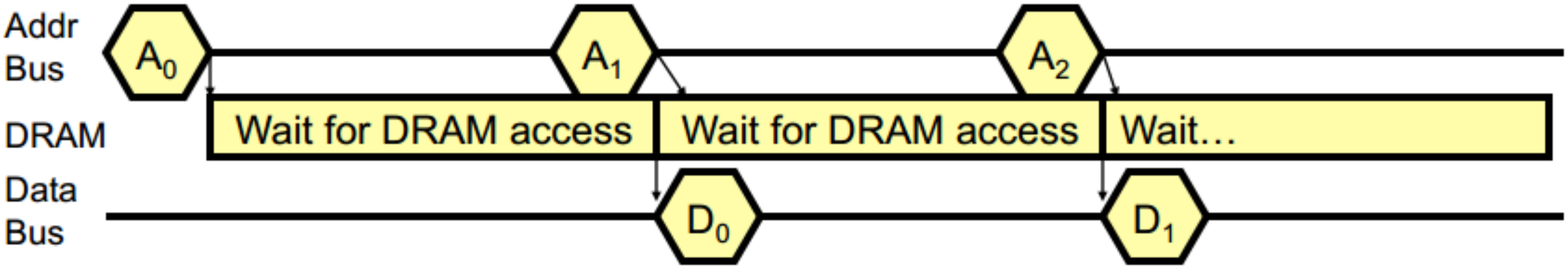
- Critical word first?

DRAM: Banks

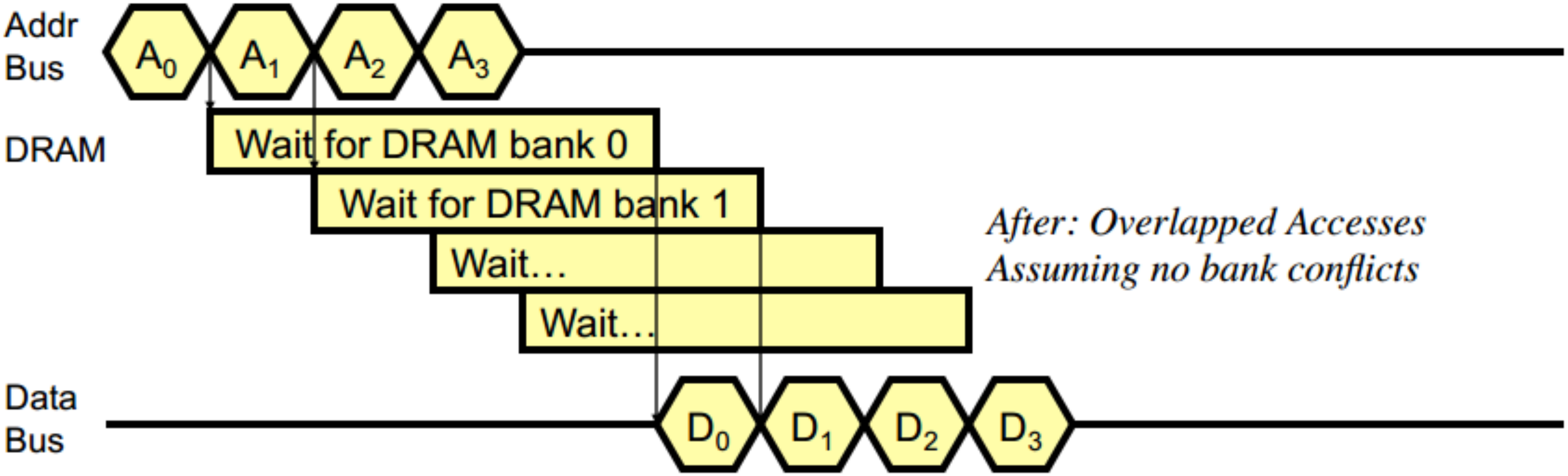
- Banks are independent arrays **WITHIN** a chip
  - DRAMs today have 4 to 32 banks
    - SDR/DDR SDRAM system: 4 banks
    - RDRAM system: 16-32 banks
- Advantages
  - Lower latency
  - Higher bandwidth by overlapping
  - Finer-grain power management
- Disadvantages
  - Bank area overhead
  - More complicated control



How Do Banks Help ?

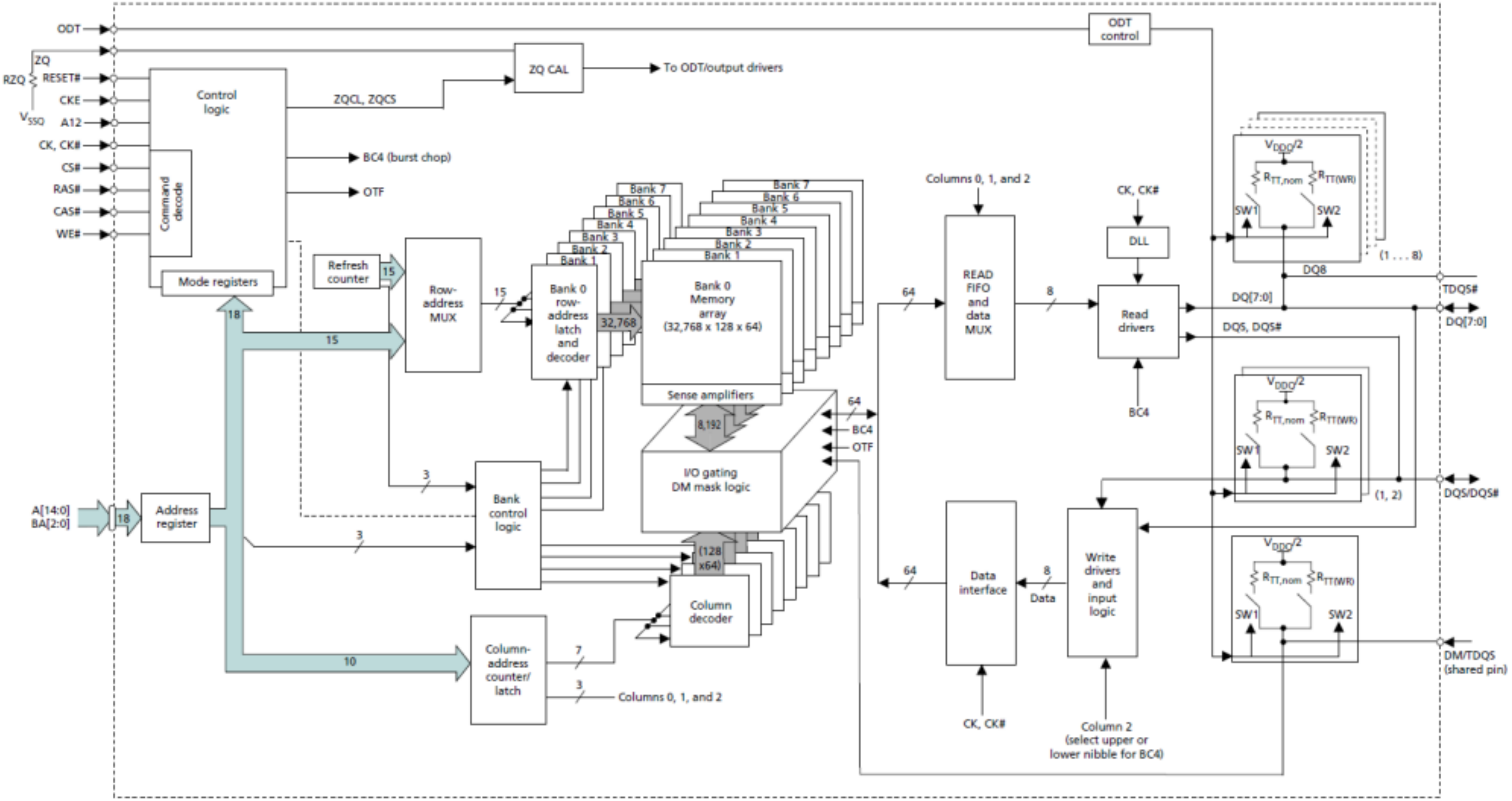


Before: No Overlapping
Assuming accesses to different DRAM rows



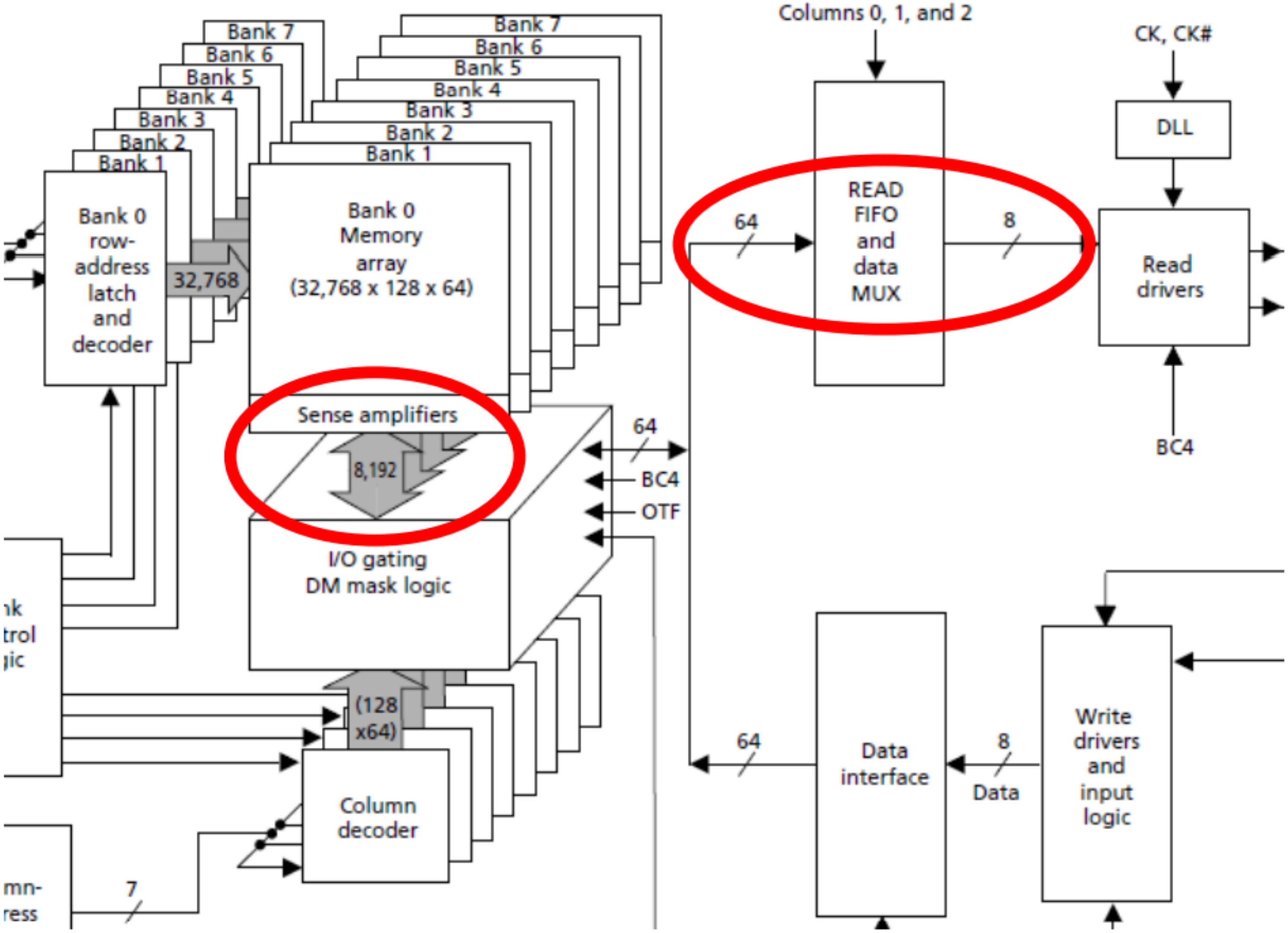
After: Overlapped Accesses
Assuming no bank conflicts

2Gb x 8 DDR3 Chip (Micron)



Observe: bank organization

2Gb x 8 DDR3 Chip (Micron)

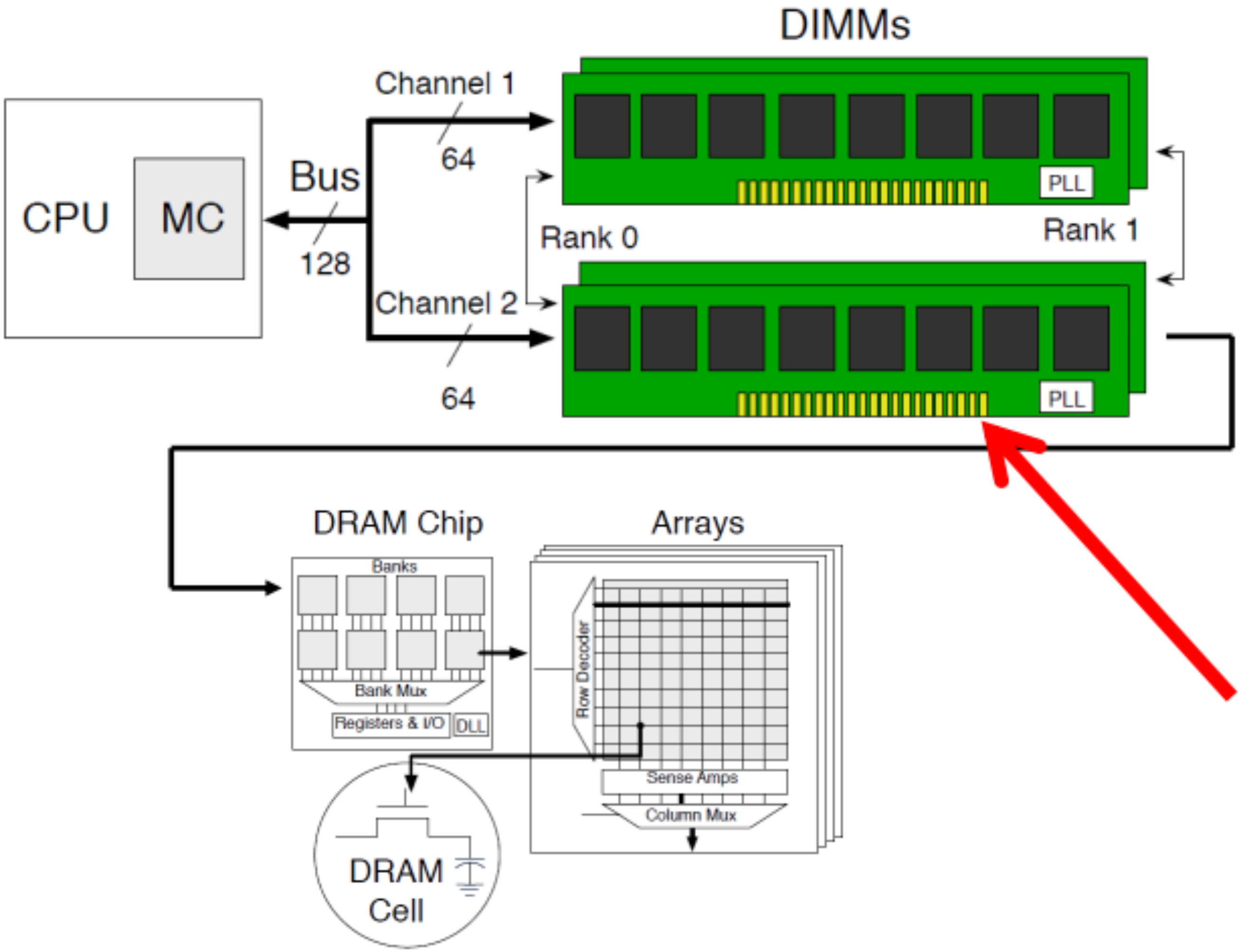


Observe: row width, 64 → 8 bit datapath

DDR3 SDRAM: Current Standard

- Introduced in 2007
- SDRAM = Synchronous DRAM = **Clocked**
- DDR = Double Data Rate
  - Data transferred on both clock edges
  - 400 MHz = 800 MT/s
- x4, x8, x16 datapath widths
- Minimum burst length of 8
- 8 banks
- 1Gb, 2Gb, 4Gb capacity common
- Relative to SDR/DDR/DDR2: + bandwidth, ~ latency

DRAM DIMM

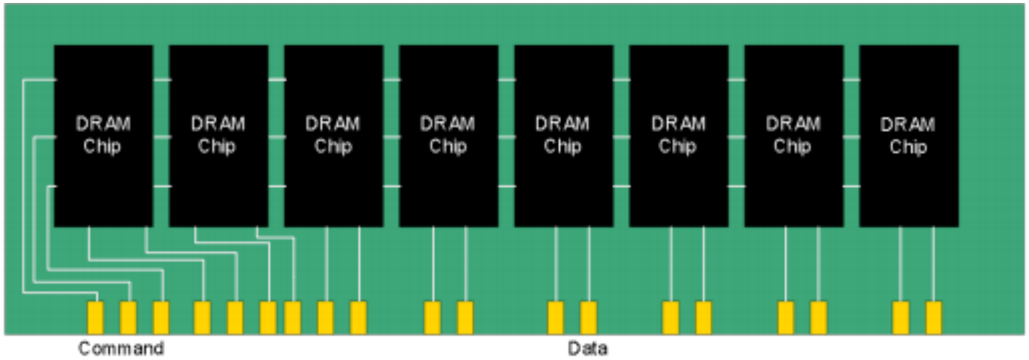

DIMMs
CPU
MC
Bus
128
Channel 1
64
Channel 2
64
Rank 0
Rank 1
PLL
PLL
DRAM Chip
Banks
Bank Mux
Registers & I/O
DLL
Arrays
Row Decoder
Sense Amps
Column Mux
DRAM
Cell

DRAM Modules

- DRAM chips have narrow interface (typically x4, x8, x16)
- Multiple chips are put together to form a wide interface
    - DIMM: Dual Inline Memory Module
    - To get a 64-bit DIMM, we need to access 8 chips with 8-bit interfaces
    - Share command/address lines, but not data

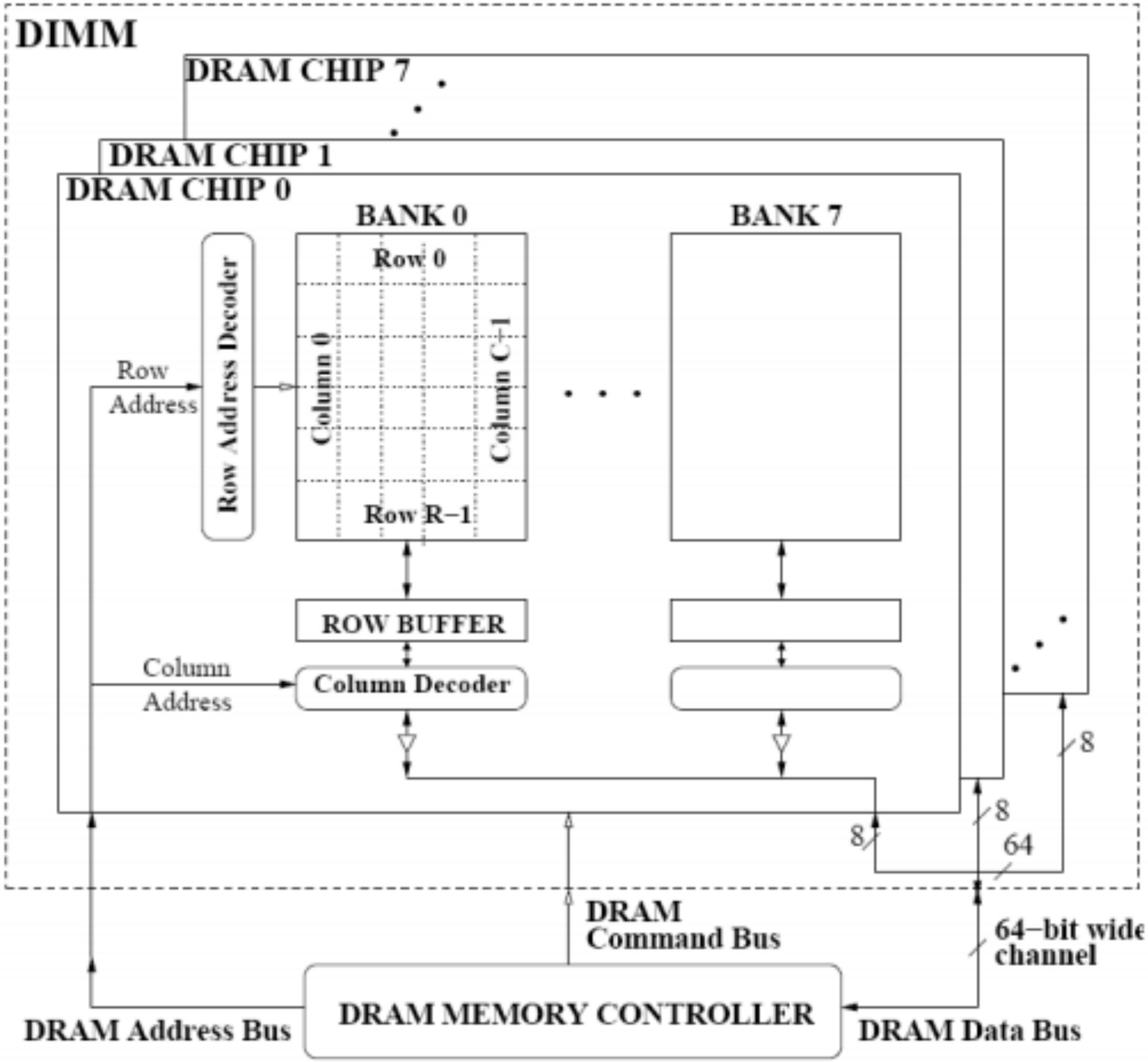
- Advantages
    - Acts like a high-capacity DRAM chip with a wide interface
        - 8x capacity, 8x bandwidth, same latency
- Disadvantages
    - Granularity: Accesses cannot be smaller than the interface width
        - 8x power

DRAM DIMMs

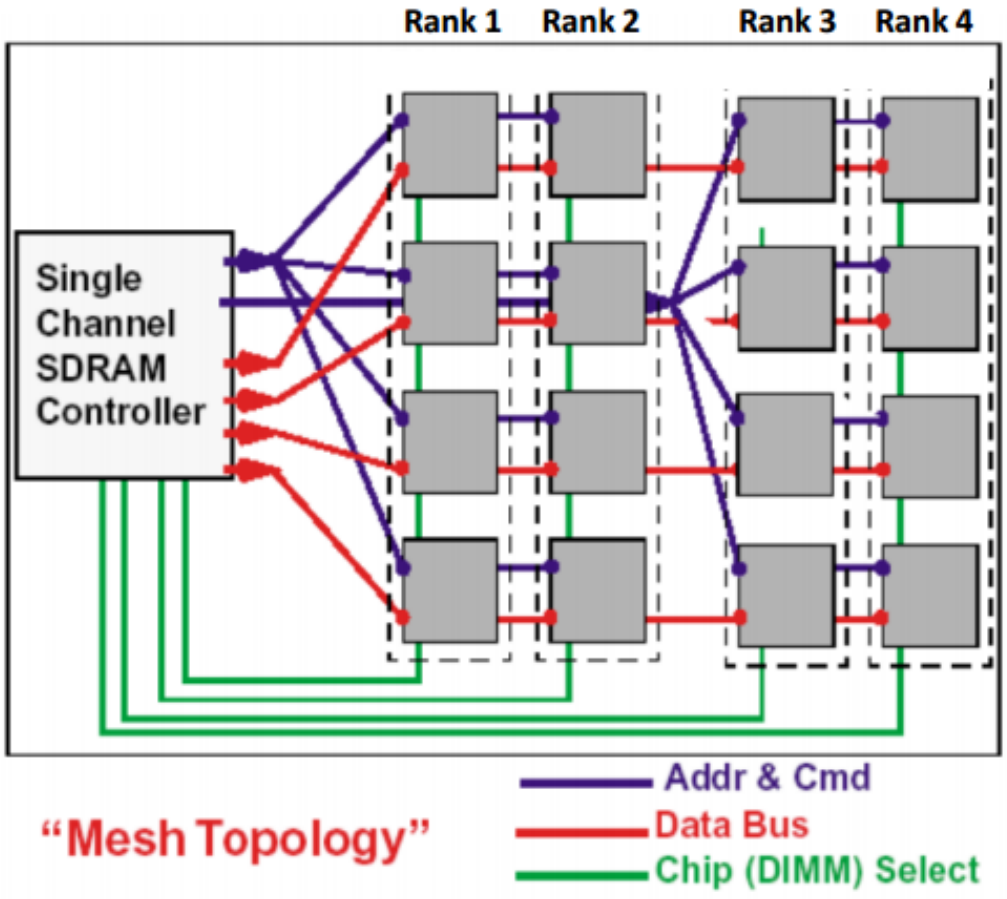


- Dual Inline Memory Module (DIMM)
  - A PCB with 8 to 16 DRAM chips
  - All chips receive identical control and addresses
  - Data pins from all chips are directly connected to PCB pins
- Advantages:
  - A DIMM acts like a high-capacity DRAM chip with a wide interface
    - E.g. use 8 chips with 8-bit interfaces to connect to a 64-bit memory bus
  - Easier to replace/add memory in a system
    - No need to solder/remove individual chips
- Disadvantage: memory granularity problem

64-bit Wide DIMM



Multiple DIMMs on a Channel



"Mesh Topology"

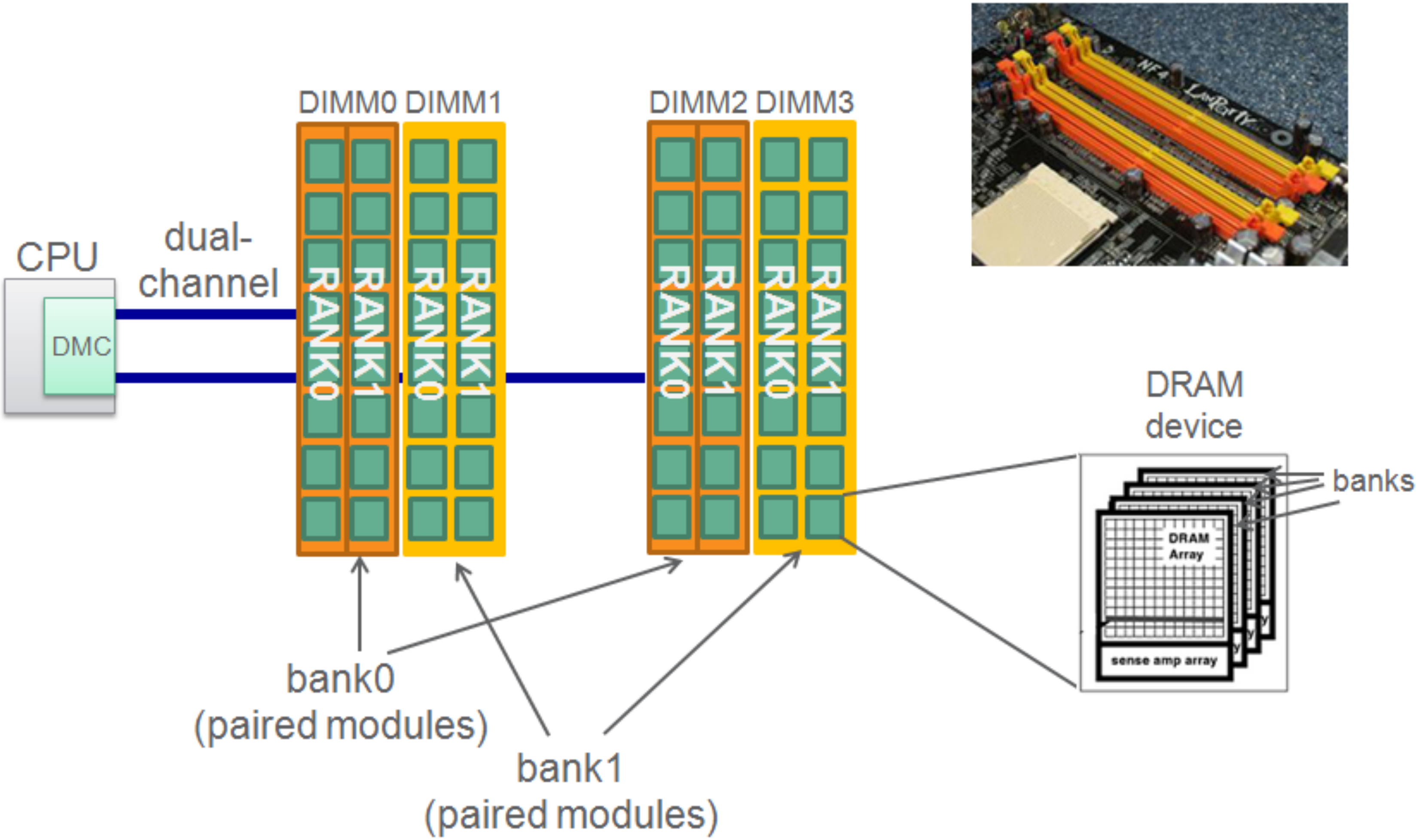
- Advantages:
  - Enables even higher capacity
- Disadvantages:
  - Interconnect latency, complexity, and energy get higher
  - Addr/Cmd signal integrity is a challenge

DRAM Ranks

- A DIMM may include multiple Ranks
    - A 64-bit DIMM using 8 chips with x16 interfaces has 2 ranks

- Each 64-bit group of chips is called a rank
    - All chips in a rank respond to a single command
    - Different ranks share command/address/data lines
        - Select between ranks with “Chip Select” signal
    - Ranks provide more “banks” across multiple chips (but don’t confuse rank and bank!)

Traditional Memory Hierarchy



State of the art

- DDR3
  - Transfer data at rising and falling edge
  - Regular DRAM – 200MHz (or 800MHz IO bus), 8byte width,6.4GBytes/sec
  - Double data rate 12.8GBytes/sec
  - 8-burst-deep prefetch buffer
- GDDR5 (Graphics Double Data Rate)
  - High performance designed for high bandwidth.
  - Based on DDR3 double data lines
  - GDDR5 has 8-bit wide prefetch buffers
- RAMBUS (RDRAM)
  - Split transaction bus, byte wide
  - More complicated electrical interface on DRAM and CPU
  - 800 MHz, 18 bits, 1.6GB/sec per chip
- DDR4
  - (1600-3200MHz IO bus), 8 byte width, 17-25GBytes/sec
  - 16 banks
- Hybrid Memory Cube (HMC)

DDR vs Rambus

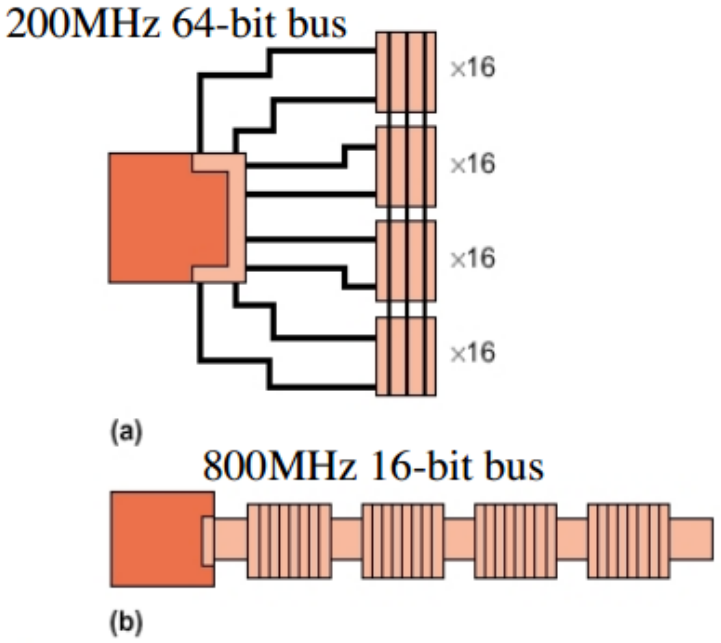
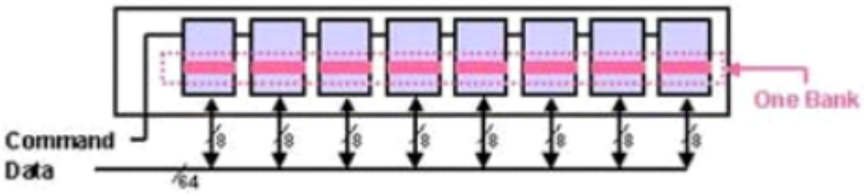
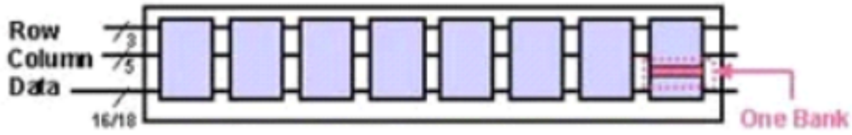


Figure 8. Bank counts: a 32-Mbyte, 64M SDRAM system with four large banks (a) versus a 32-Mbyte, 64M Direct RDRAM system with 32 small banks (b).

DIMM Modules



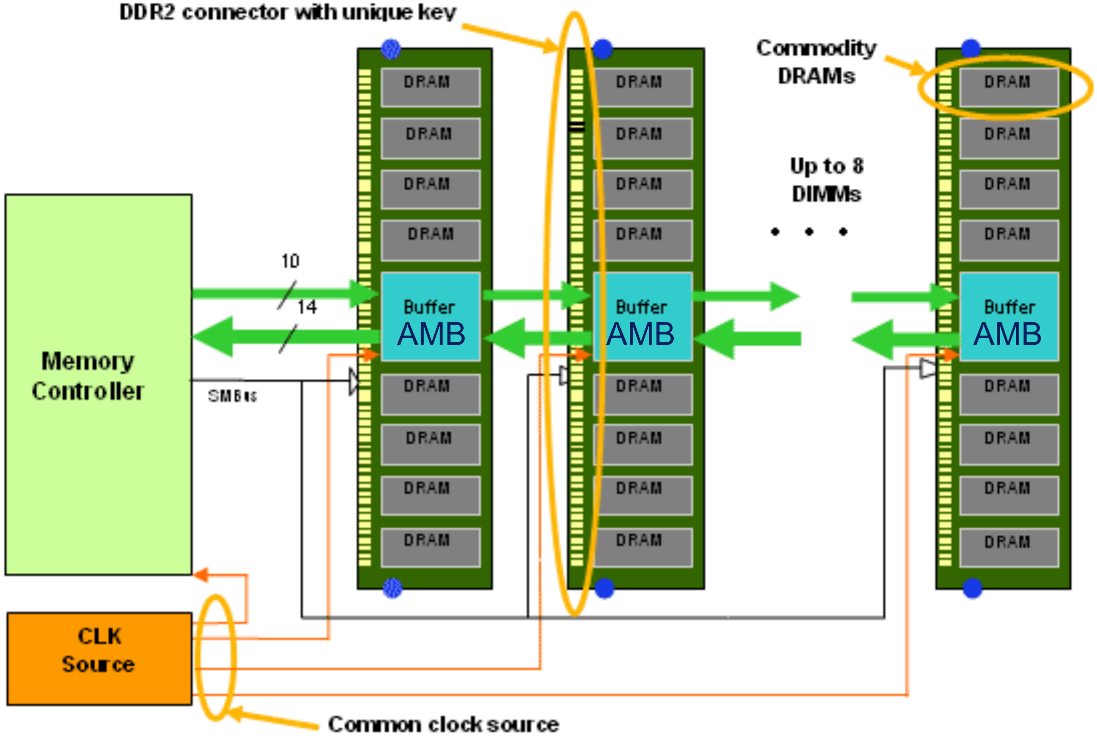
RIMM Modules



- Many banks/chip (4-32)
- Narrow fast interconnect (pipelined)
- High bandwidth
- Latency & area penalty

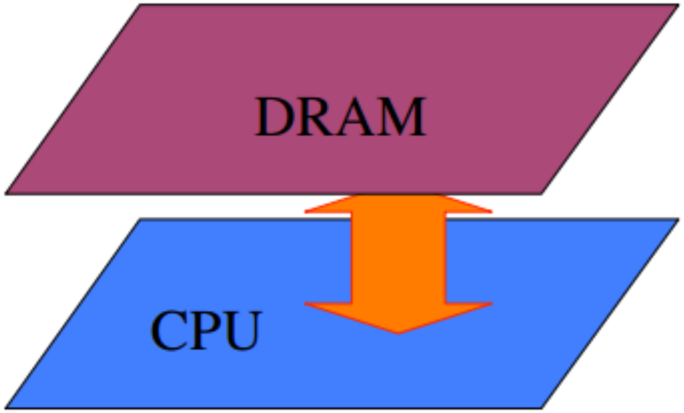
Fully Buffered DIMM (FB-DIMM)

- The DDR problem
  - Higher capacity → more DIMMs → lower data-rate (multidrop bus)
- FBDIMM approach: use point-to-point links
  - While still using commodity DRAM chips
  - Network with 12-beat packages, separate up/downstream wires

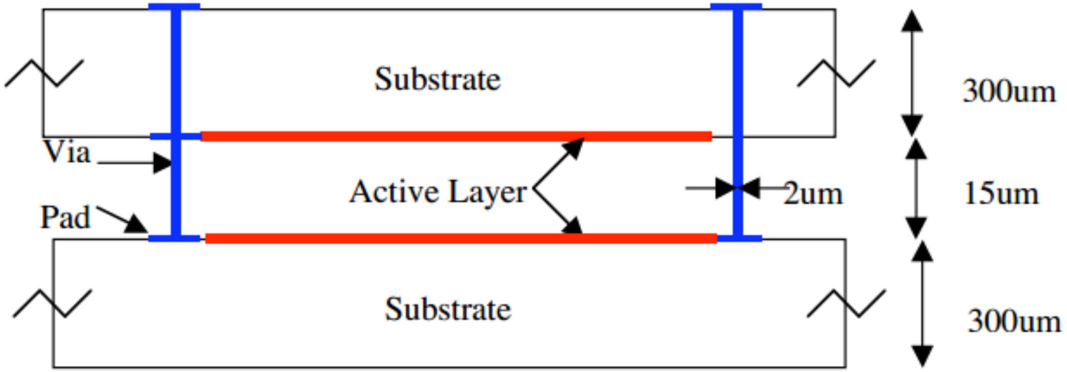


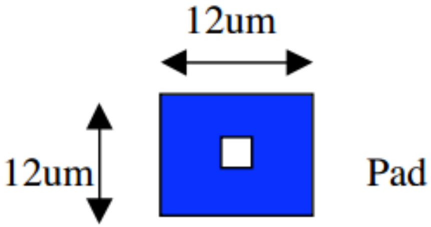
3D-Stacked DRAM

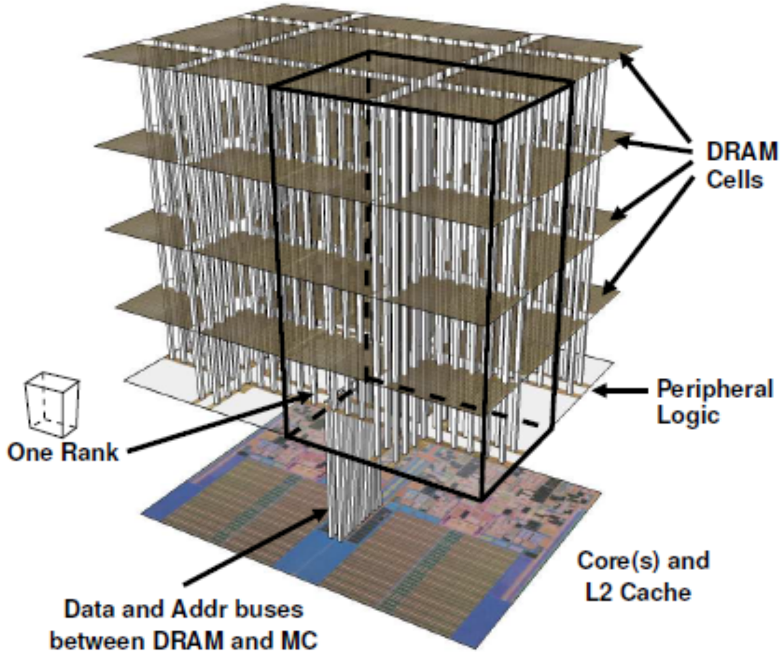
- Place wafers on top of one another
- Via^s complete paths between different wafers through small pads on the wafers



3D IC Structure

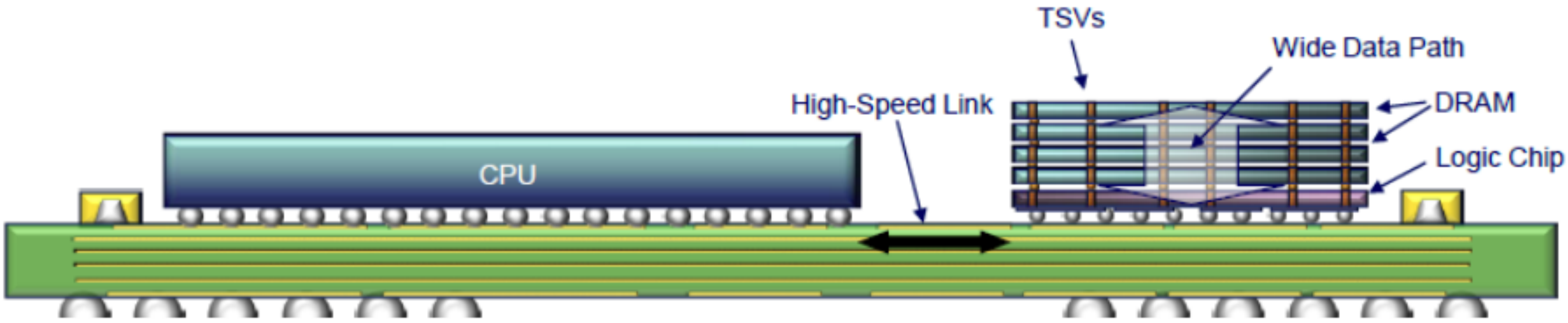






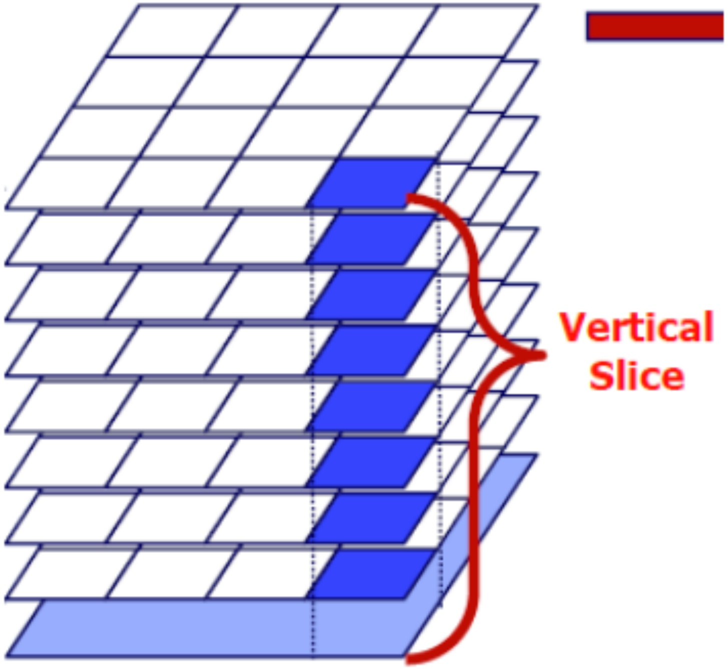
Micron HMC
"Hybrid Memory Cube"

- 3D-stacked device with memory+logic
- High capacity, low power, high bandwidth
- Can move functionalities to the memory package



HMC Details

- 32 banks per die x 8 dies = 256 banks per package
- 2 banks x 8 dies form 1 vertical slice (shared data bus)
- High internal data bandwidth (TSVs) ➔ entire cache line from a single array (2 banks) that is 256 bytes wide
- Future generations: eight links that can connect to the processor or other HMCs – each link (40 GBps) has 16 up and 16 down lanes (each lane has 2 differential wires)
- 1866 TSVs at 60 um pitch and 2 Gb/s (50 nm 1Gb DRAMs)
- 3.7 pJ/bit for DRAM layers and 6.78 pJ/bit for logic layer (existing DDR3 modules are 65 pJ/bit)


Vertical Slice