Main Memory

Moving further away from the CPU…..

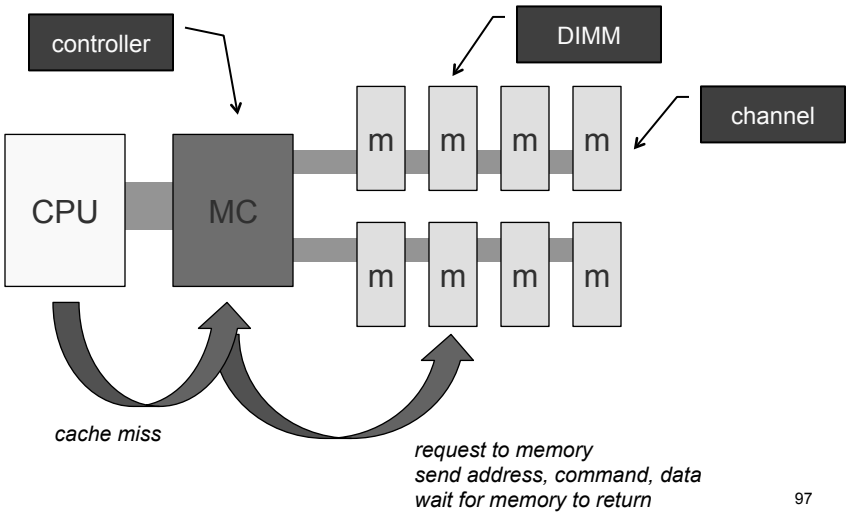
95

Main Memory

- Performance measurement
  - Latency - cache miss penalty
  - Bandwidth - large block sizes of L2 argue for B/W
- Memory latency
  - *Access time*: Time between when a read is requested and when the data arrives
  - *Cycle time*: Minimum time between requests to memory
  - Cycle time > Access time: Address lines must be stable between successive accesses

96

Main Memory



97

Hierarchical Organization

System

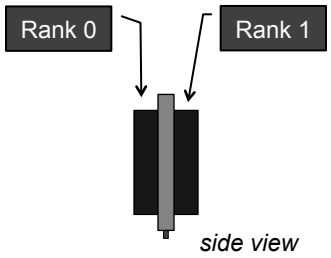
Internal

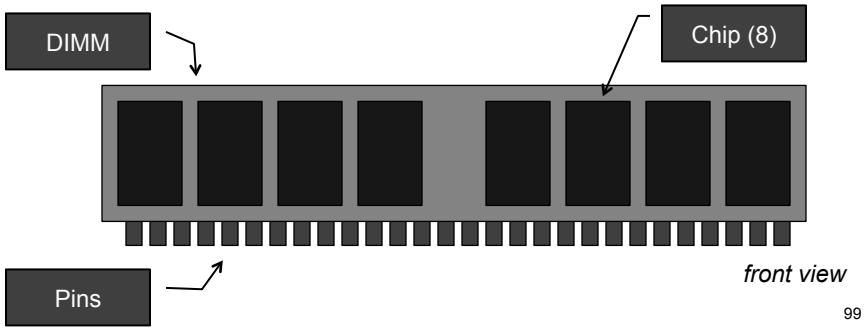
1. Channel – independent connection to DIMMs
2. DIMM – independent modules of memory chips
3. Rank – independent set of chips on each DIMM
4. Chip – individual memory chip of Rank/DIMM
5. Bank – internal independent memory partition
6. Row – internally cached row of a bank

98

DIMM organization

- Dual Inline Memory Module
  - Two-sided group of memory chips
  - Connected to channel
  - Receives addresses, commands, data
  - Each side is rank of multiple (4,8) chips

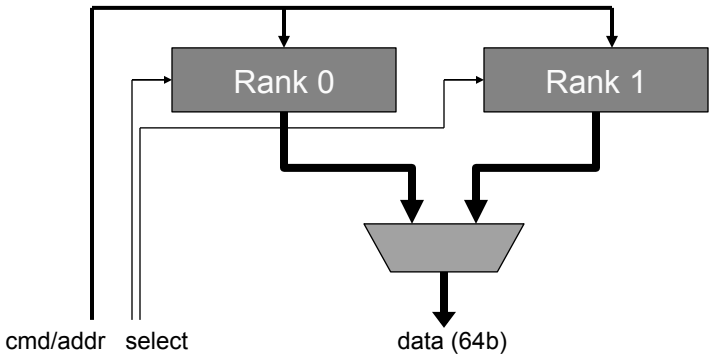




99

Rank organization

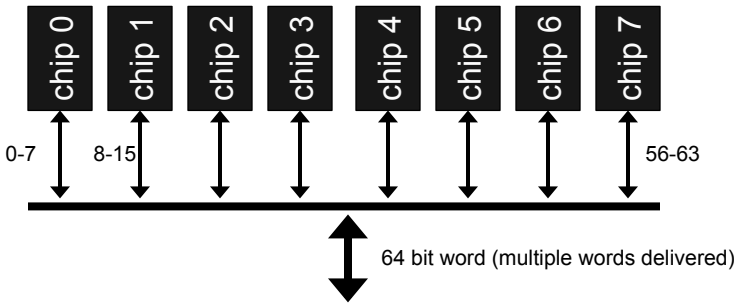
- Independent group of chips on front/back
- Connected to the channel



100

Rank organization

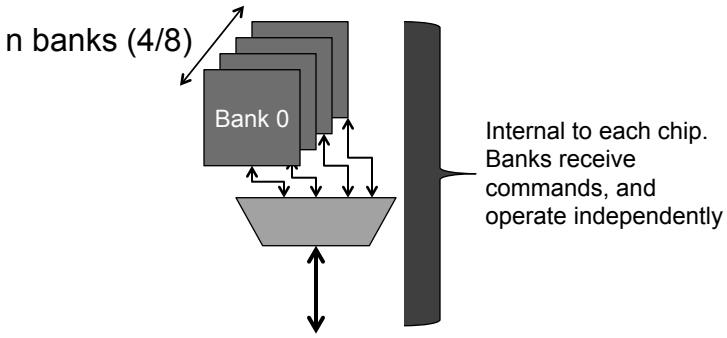
- Multiple memory chips per rank
- Each chip provides part of data
- Data size is typically 64 bits



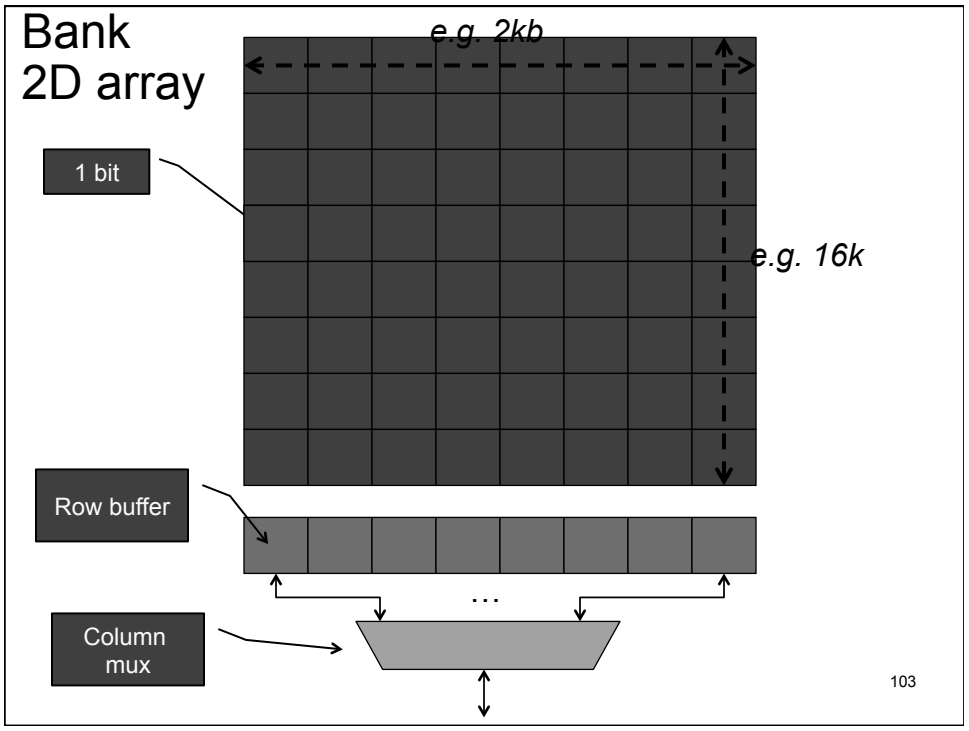
101

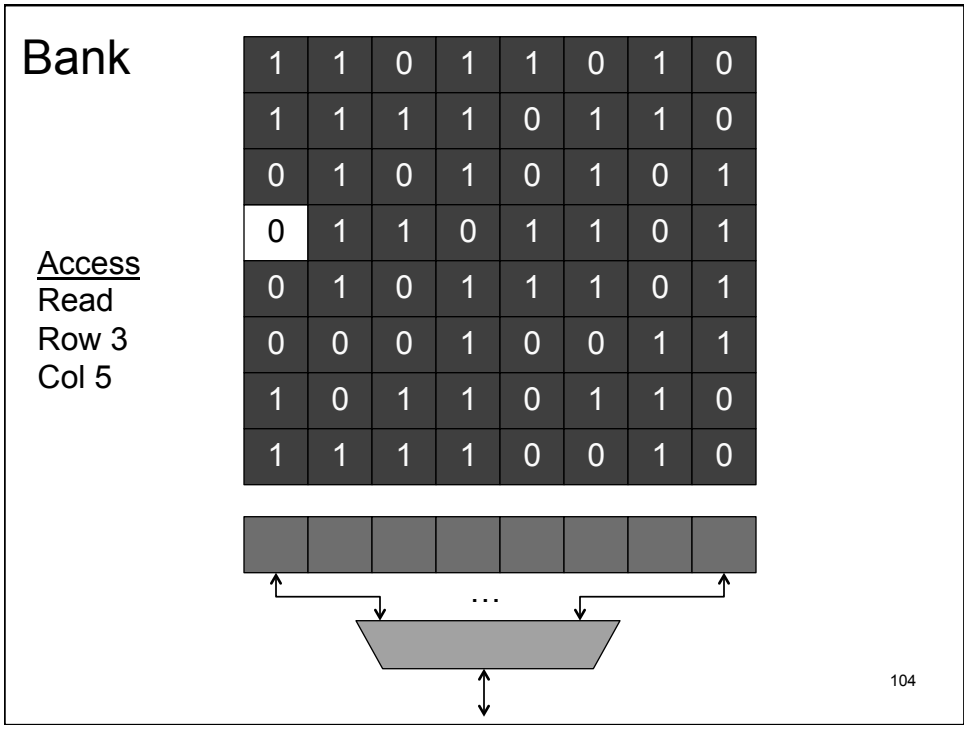
Bank

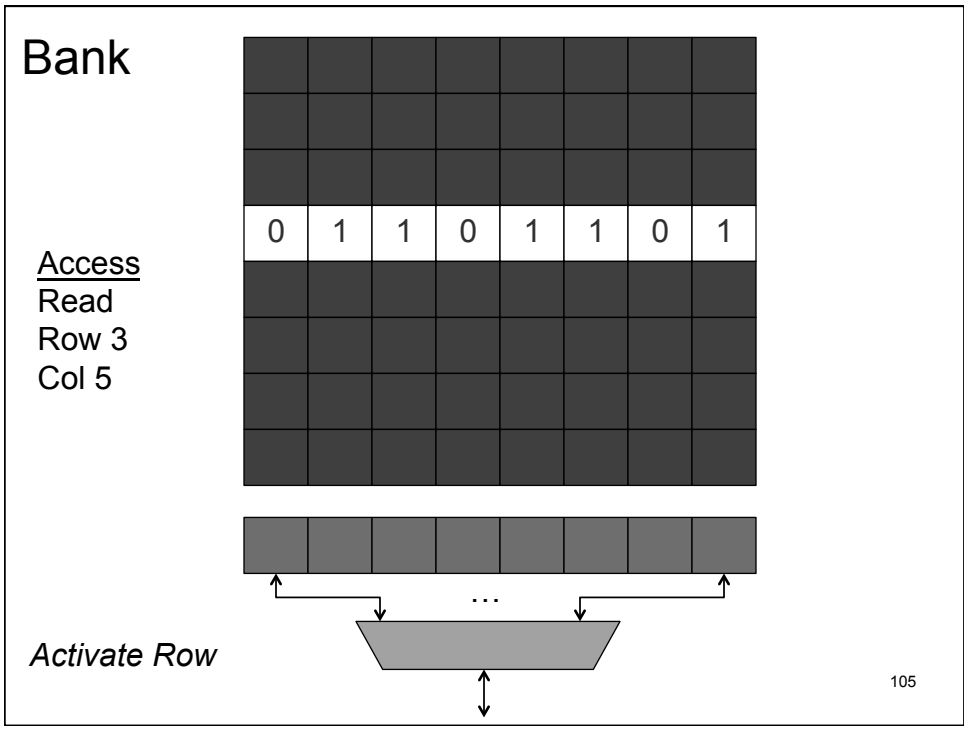
- Internal to each chip
- Partition of bits accessed independently

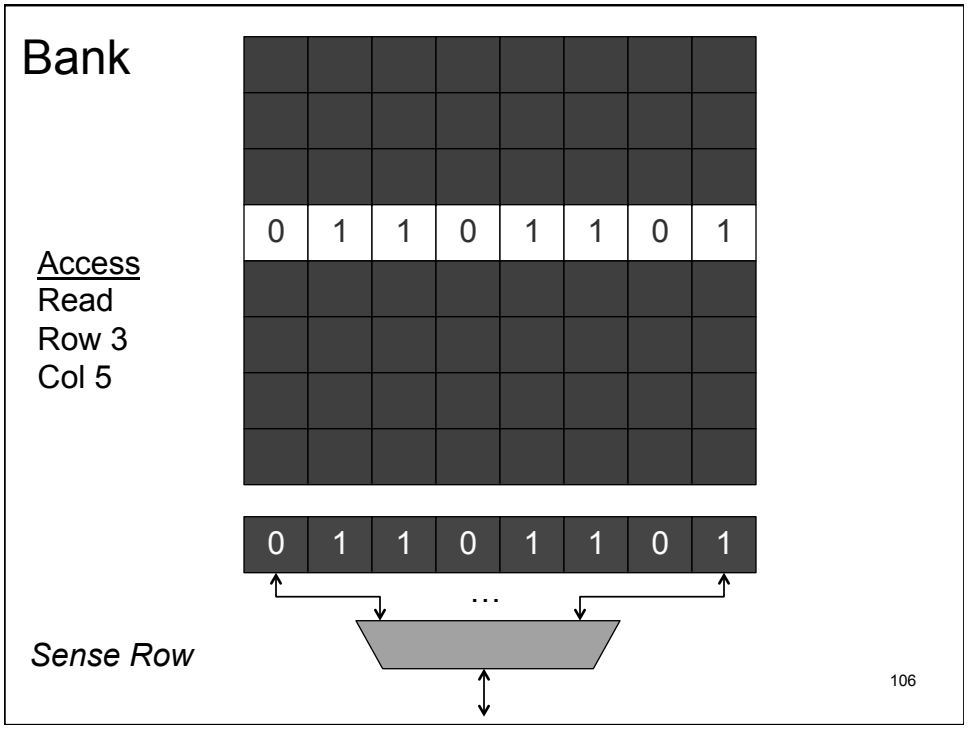


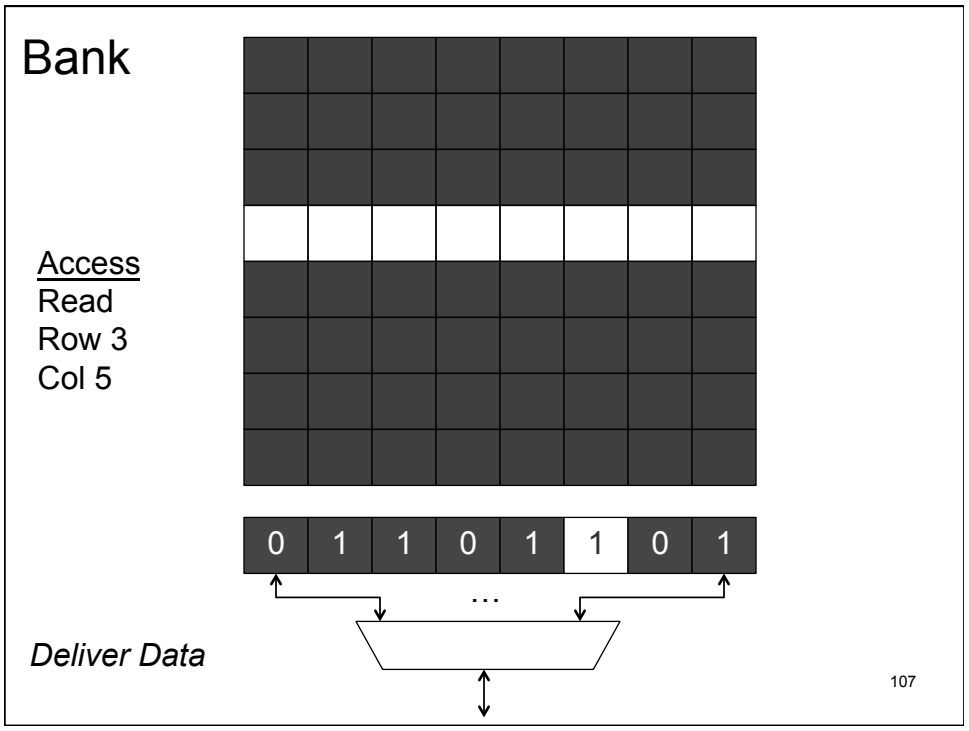
102

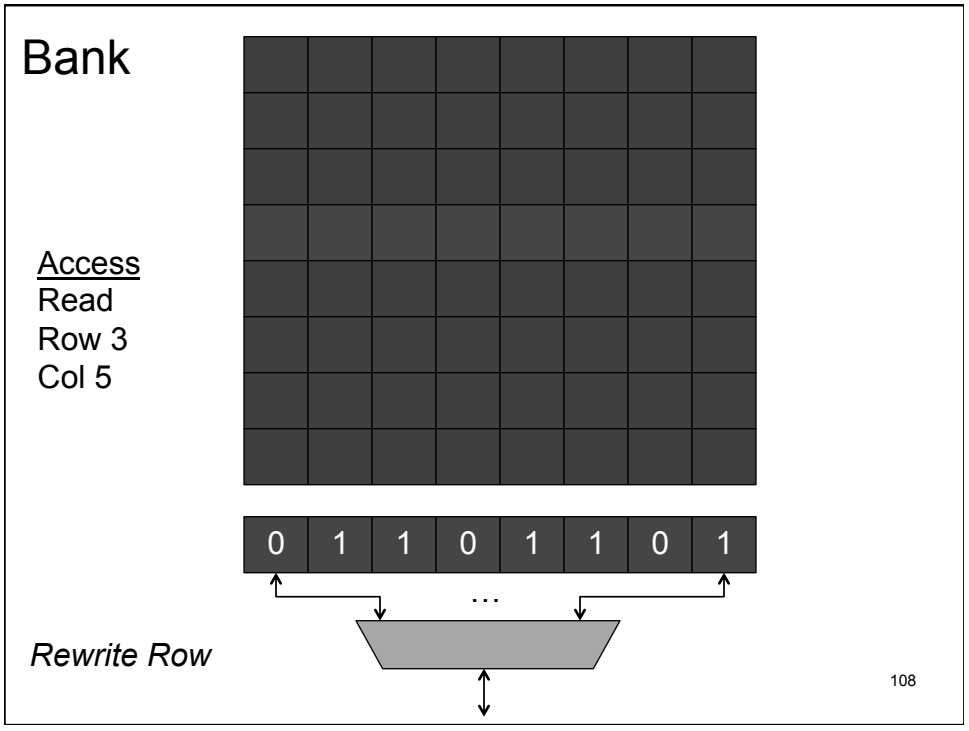

Bank
2D array
e.g. 2kb
1 bit
e.g. 16k
Row buffer
…
Column
mux
103

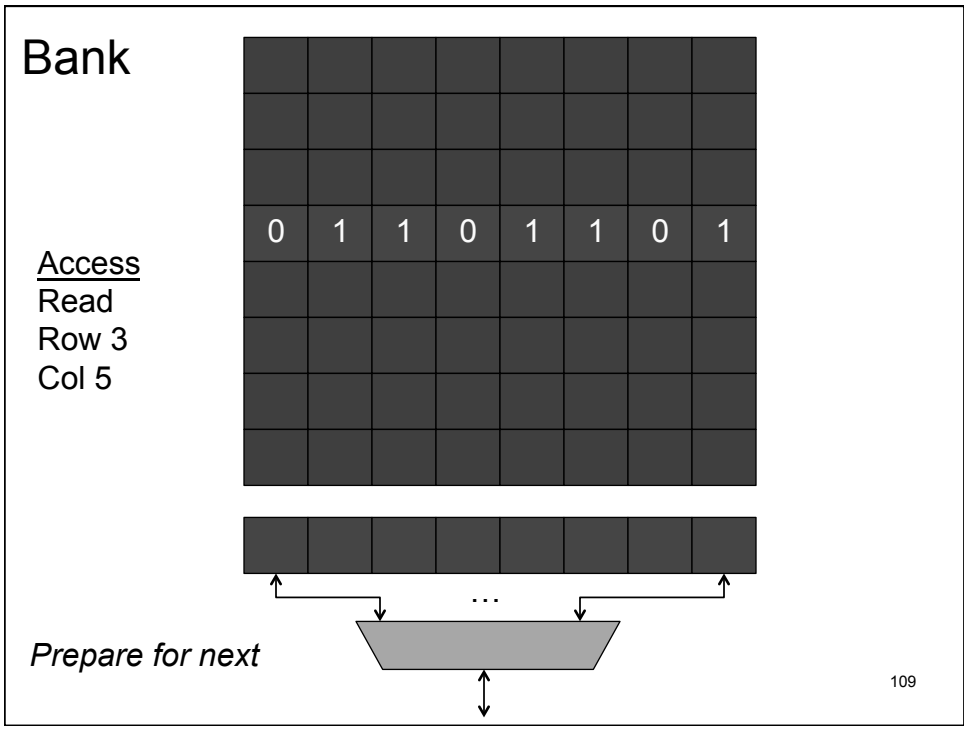

Bank
1 1 0 1 1 0 1 0
1 1 1 1 0 1 1 0
0 1 0 1 0 1 0 1
0 1 1 0 1 1 0 1
0 1 0 1 1 1 0 1
0 0 0 1 0 0 1 1
1 0 1 1 0 1 1 0
1 1 1 1 0 0 1 0
Access
Read
Row 3
Col 5
…
104

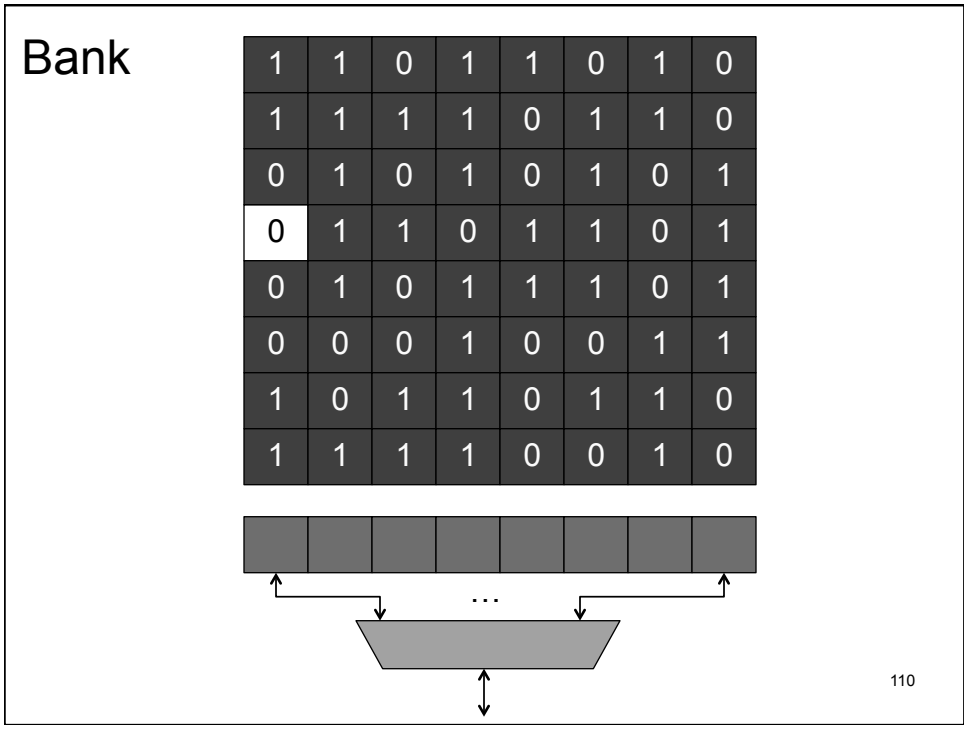

Bank
0 1 1 0 1 1 0 1
Access
Read
Row 3
Col 5
…
Activate Row
105


Bank
0 1 1 0 1 1 0 1
Access
Read
Row 3
Col 5
0 1 1 0 1 1 0 1
…
Sense Row
106


Bank
Access
Read
Row 3
Col 5
0 1 1 0 1 1 0 1
…
Deliver Data
107

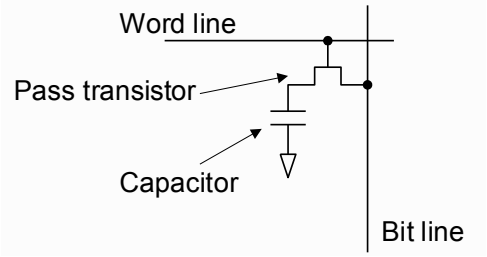

Bank
Access
Read
Row 3
Col 5
0 1 1 0 1 1 0 1
…
Rewrite Row
108


Bank
0 1 1 0 1 1 0 1
Access
Read
Row 3
Col 5
…
Prepare for next
109

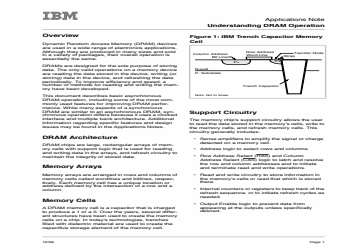

Bank
1 1 0 1 1 0 1 0
1 1 1 1 0 1 1 0
0 1 0 1 0 1 0 1
0 1 1 0 1 1 0 1
0 1 0 1 1 1 0 1
0 0 0 1 0 0 1 1
1 0 1 1 0 1 1 0
1 1 1 1 0 0 1 0
…
110

Bit Cell

- Structure used to store logical 0 or 1
- Stored as a charge



1 transistor (access) + 1 capacitor (storage)


IBM

Applications Note

Understanding DRAM Operation

Overview

Figure 1. IBM Trench Capacitor Memory Cell

Support Circuitry

DRAM Architecture

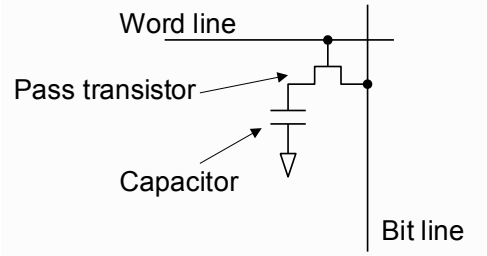
Memory Arrays

Memory Cells

physical implementation (from IBM)

Bit Cell

- Structure used to store logical 0 or 1
- Stored as a charge



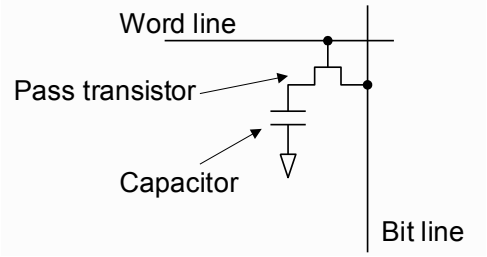
1 transistor (access) + 1 capacitor (storage)

WRITE bit cell

1. Load value into row buffer
2. Enable word line
3. If 1, capacitor is charged
4. If 0, capacitor is discharged

Bit Cell

- Structure used to store logical 0 or 1
- Stored as a charge



1 transistor (access) + 1 capacitor (storage)

READ bit cell

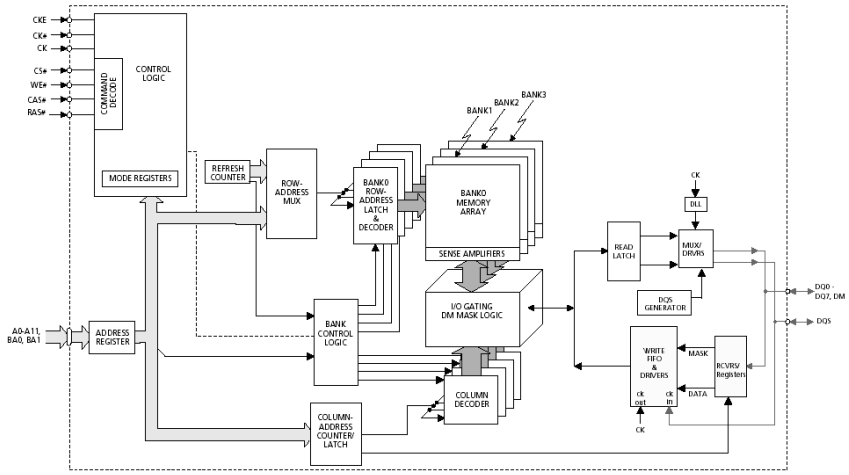
1. Bit line charged 1/2
2. Enable word line
3. Value in cap read onto bit line
4. Bit line swings high/low
5. Sense amp detects swing
6. Value is "latched" in row buffer
7. Restore row

Sense amp part of row buffer
Read is destructive

113

Overall DRAM chip organization

Figure 1: 128Mb DDR SDRAM Functional Block Diagram



DRAM chip operation

- Addresses are <row, column> pairs
- Limited address signals (bits) in channel bus
- Address sent as Row, then Col
  - Multiplex address pins to reduce number of pins
  - Column Address Strobe (*CAS*) and Row Address Strobe (*RAS*)

Closed Page Mode

- Send Row address (RAS) – Open the row buffer (read it)
- Send Col address (CAS)
- Deliver data
- Prepare for next <row, column> command (*PRECHARGE*)
- **Suppose**: R:<10,8>, R<10,9>, R<10,10> ….

115

DRAM chip operation

- Accesses exhibit locality
- Row buffer can act as a “little” cache in DRAM
- *Deliver data from same row for different columns!*

Open Page Mode

- Leave row buffer “open” to serve further column accesses
- So called column hits (aka “row buffer hits”)
- Send only the column address (RAS, CAS, CAS….CAS)
  - » E.g. R:<10>,<8>,<9>,<10>
- Memory can also “burst” open data from a row
- Must close row when complete, or conflicting access to it
  - » PRECHARE for next Open (RAS)

116

DRAM latency

- Several components affect DRAM latency
- Latency can be variable as well
- Primary components are:
  1. Cache controller (from CPU to memory controller)
  2. Controller latency
  3. Controller to DRAM transfer time (bus management)
  4. DRAM bank latency
  5. DRAM to CPU transfer time (via the controller)

117

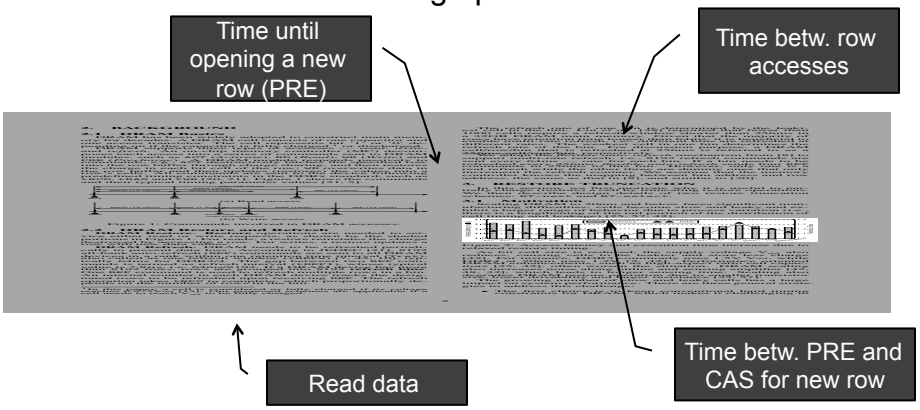
DRAM latency

- Controller Latency
  - Intelligent scheduling: Maximize row buffer hits
  - Queuing and scheduling delay
  - Low-level commands (PRE,ACT,R/W)
- DRAM Latency
  - Depends on the state of the DRAM
  - Best case: CAS latency (row is open)
  - Medium case: RAS + CAS (bitlines are precharged)
  - Worst case: RAS + CAS + PRECHARGE
  - *Note, can have conflicts in banks – scheduling important*
- **Sequence: (1) PRE, (2) ACT, (3) R/W**

118

DRAM timing

- Driven by specifications – JEDEC
- Controls when/how long operations take

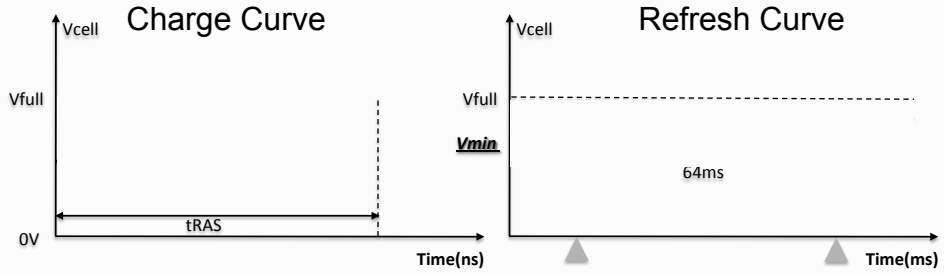


DRAM refresh

- Capacitor loses charge over time
- Refresh: Restore charge before lost
  - ACTIVATE + PRECHARGE to access the row, restoring it
  - Periodic refresh – often 64 or 128 ms
  - Refresh done before too much charge is lost

DRAM refresh

- Capacitor loses charge over time
- Refresh: Restore charge before lost
  - ACTIVATE + PRECHARGE to access the row, restoring it
  - Periodic refresh – often 64 or 128 ms
  - Refresh done before too much charge is lost



121