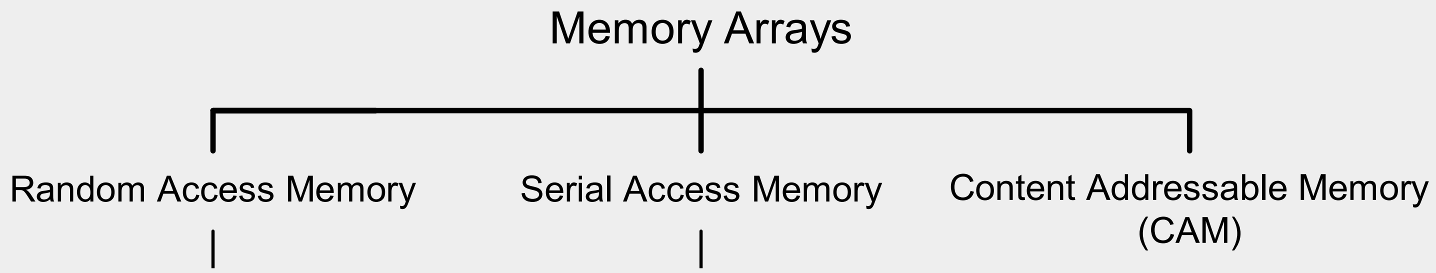
M2 – Memory Systems

M2 – Outline

- Memory Hierarchy
- Cache Blocking – Cache Aware Programming
- SRAM, DRAM
- Virtual Memory
- Virtual Machines
- Non-volatile Memory, Persistent NVM

Memory Technology



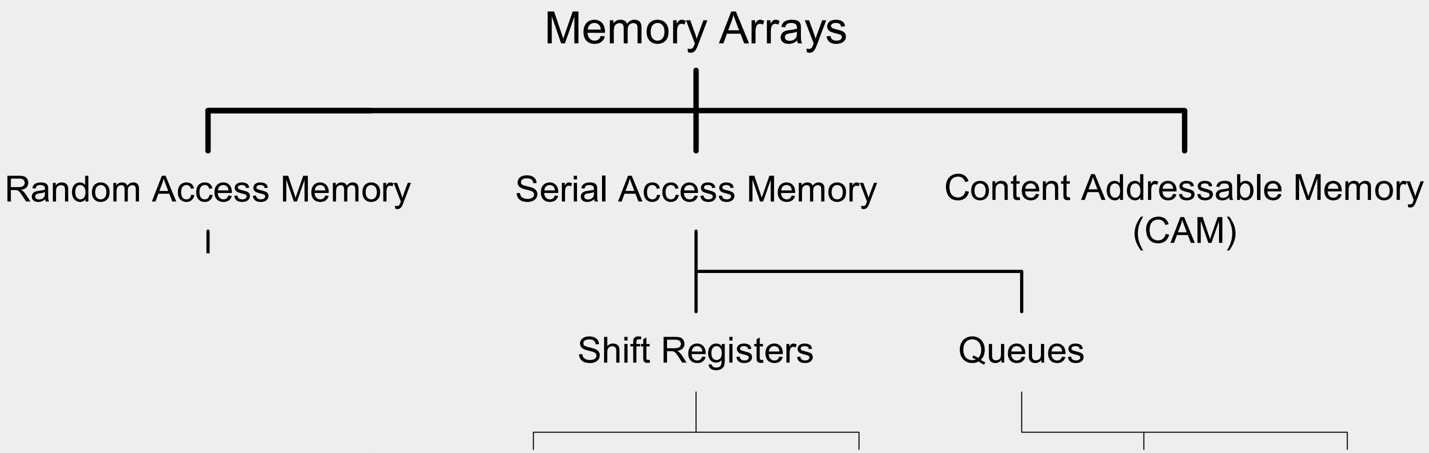
Memory Arrays

Random Access Memory

Serial Access Memory

Content Addressable Memory (CAM)

Memory Technology



Memory Arrays

Random Access Memory

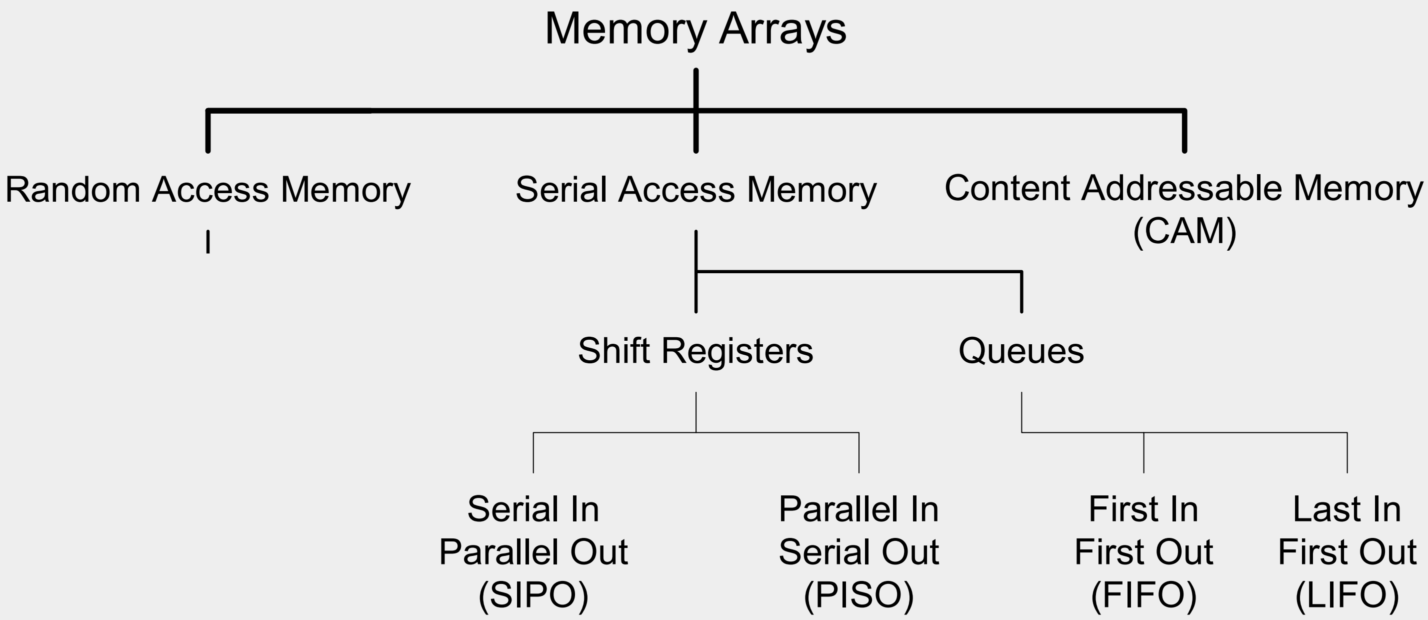
Serial Access Memory

Content Addressable Memory (CAM)

Shift Registers

Queues

Memory Technology



Memory Arrays

Random Access Memory

Serial Access Memory

Content Addressable Memory
(CAM)

Shift Registers

Queues

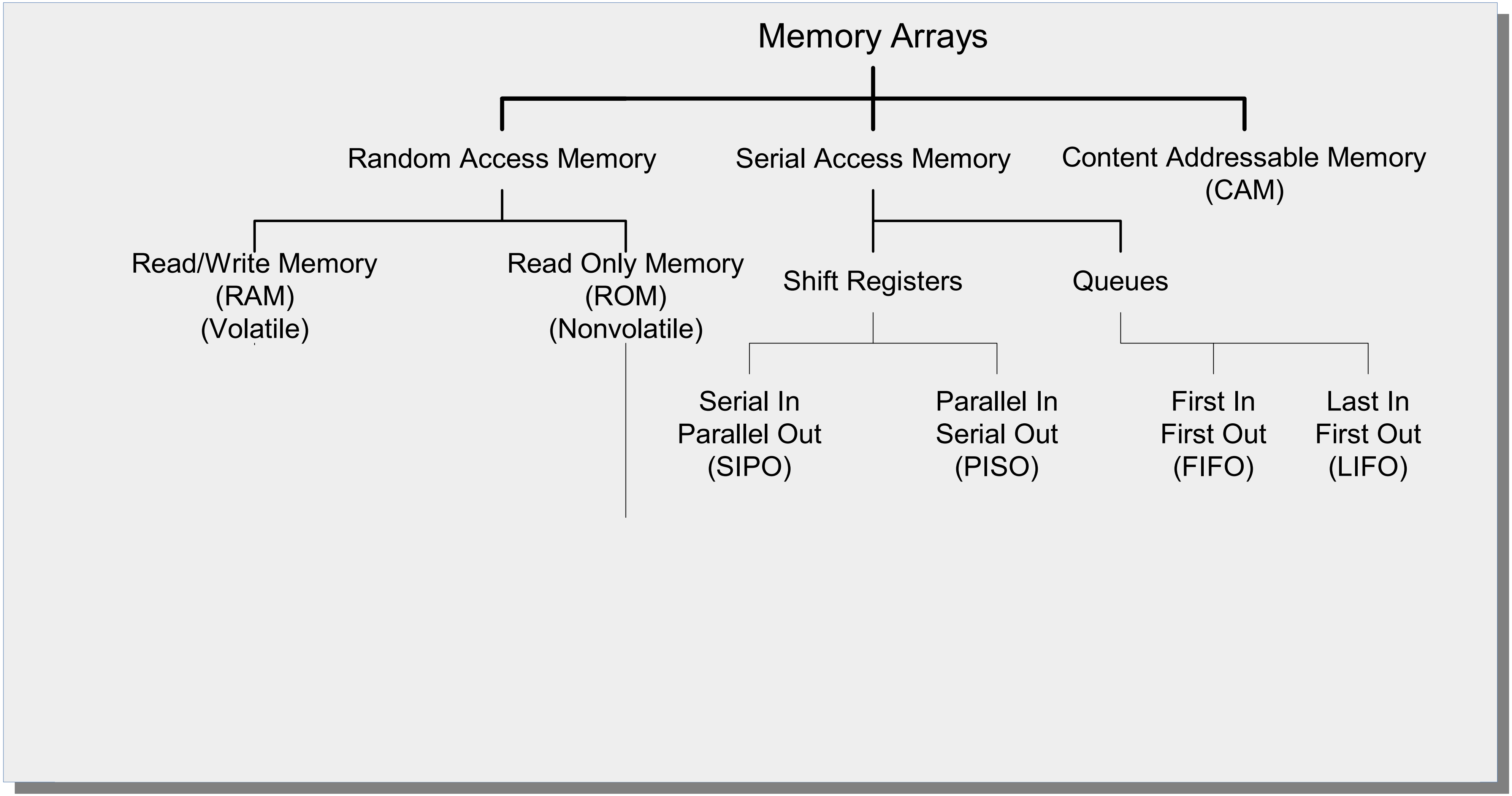
Serial In
Parallel Out
(SIPO)

Parallel In
Serial Out
(PISO)

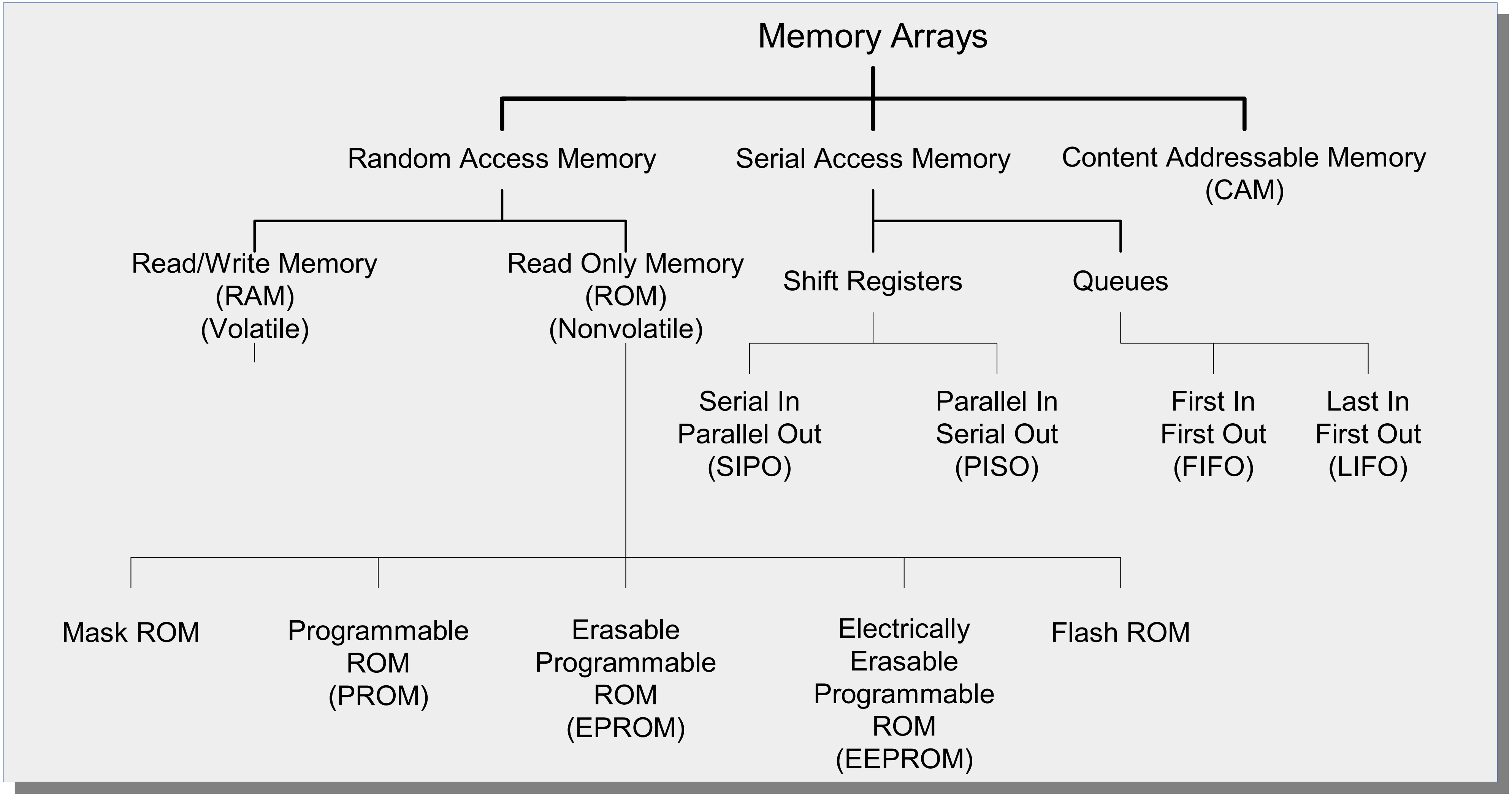
First In
First Out
(FIFO)

Last In
First Out
(LIFO)

Memory Technology



Memory Technology



Memory Arrays

Random Access Memory

Serial Access Memory

Content Addressable Memory
(CAM)

Read/Write Memory
(RAM)
(Volatile)

Read Only Memory
(ROM)
(Nonvolatile)

Shift Registers

Queues

Serial In
Parallel Out
(SIPO)

Parallel In
Serial Out
(PISO)

First In
First Out
(FIFO)

Last In
First Out
(LIFO)

Mask ROM

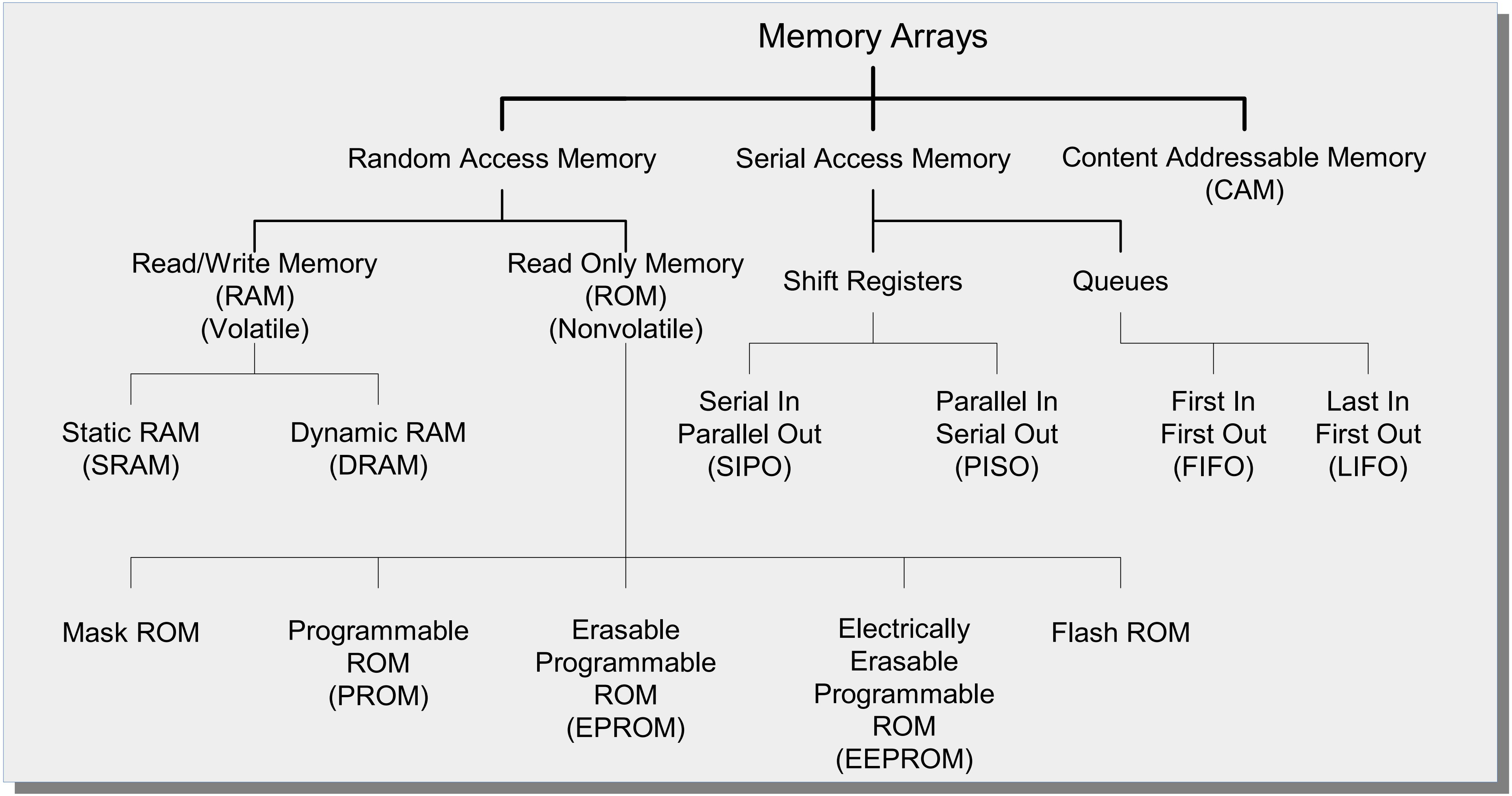
Programmable
ROM
(PROM)

Erasable
Programmable
ROM
(EPROM)

Electrically
Erasable
Programmable
ROM
(EEPROM)

Flash ROM

Memory Technology



Memory Arrays

Random Access Memory

Serial Access Memory

Content Addressable Memory (CAM)

Read/Write Memory (RAM) (Volatile)

Read Only Memory (ROM) (Nonvolatile)

Shift Registers

Queues

Static RAM (SRAM)

Dynamic RAM (DRAM)

Serial In Parallel Out (SIPO)

Parallel In Serial Out (PISO)

First In First Out (FIFO)

Last In First Out (LIFO)

Mask ROM

Programmable ROM (PROM)

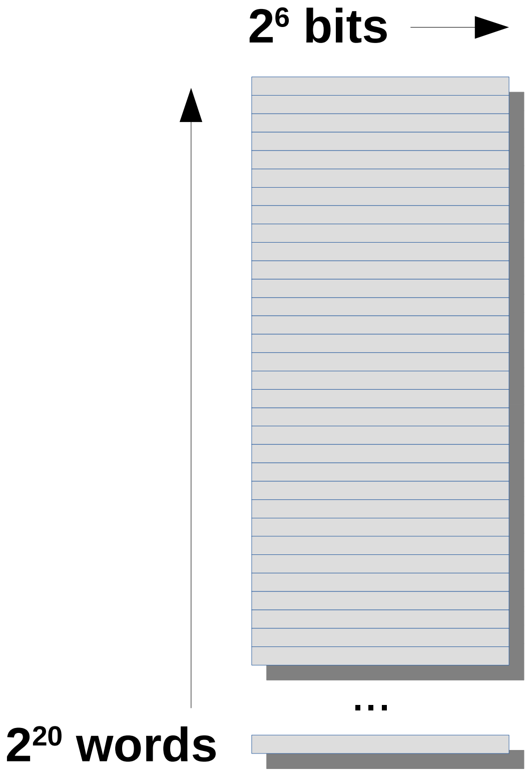
Erasable Programmable ROM (EPROM)

Electrically Erasable Programmable ROM (EEPROM)

Flash ROM

Memory Array

- Organized as 2^n words of 2^m bits each
  - Usually n >> m (1M vs. 64)
  - n = 20; m = 6

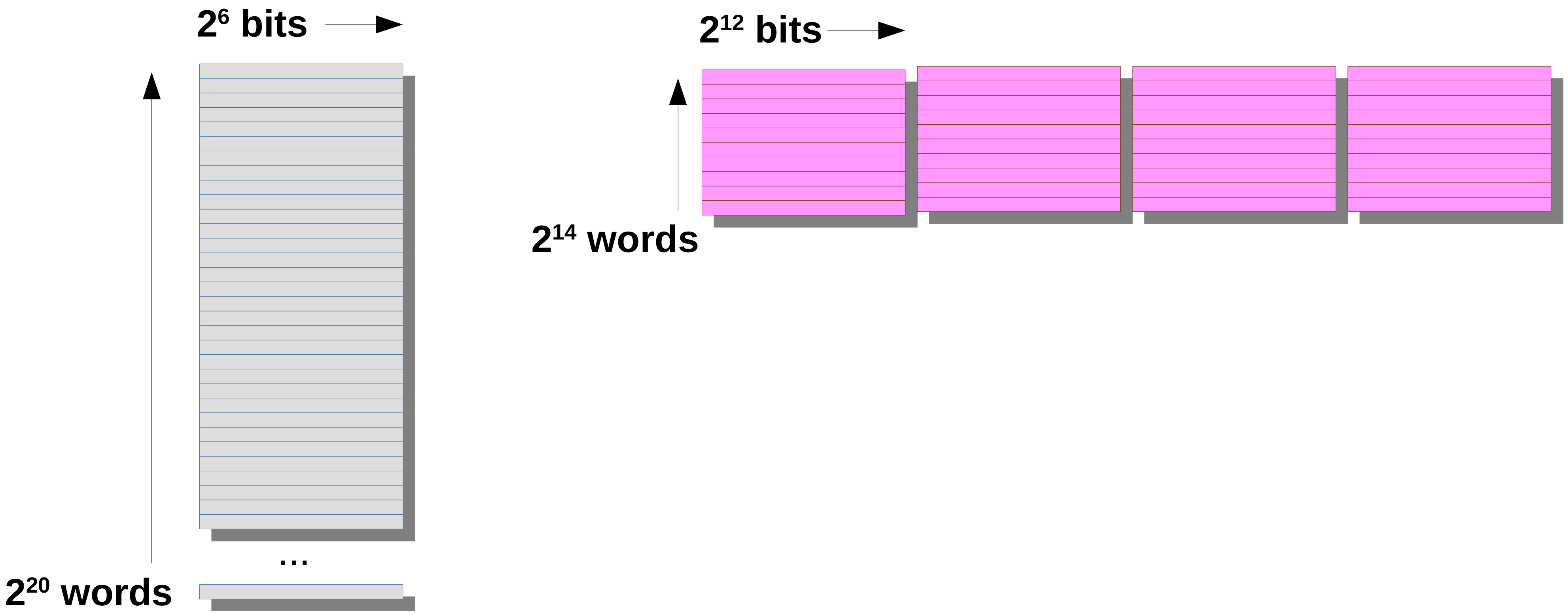


Memory Array

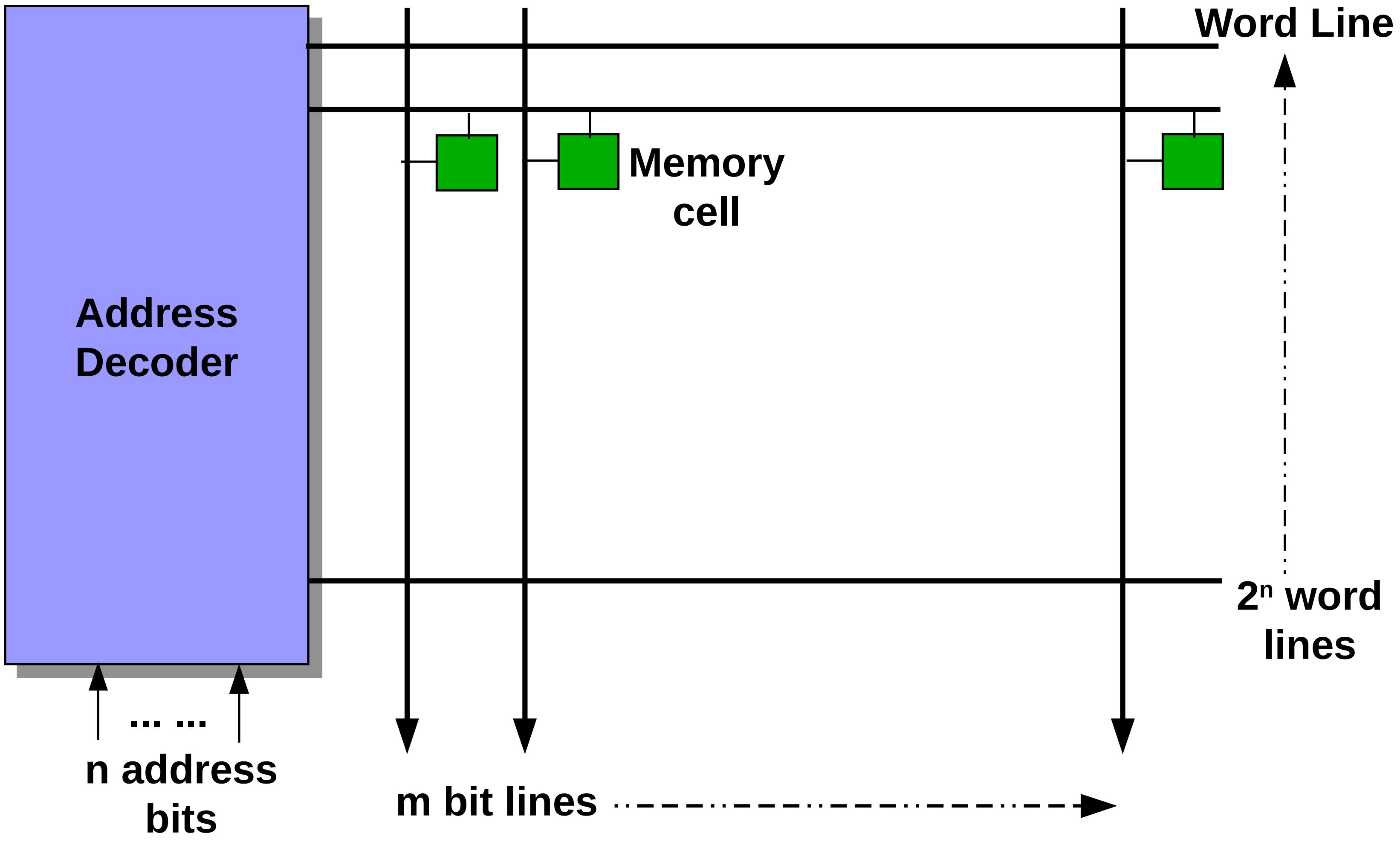
- Organized as 2^n words of 2^m bits each
  - Usually n >> m (1M vs. 64)
  -

Memory Array

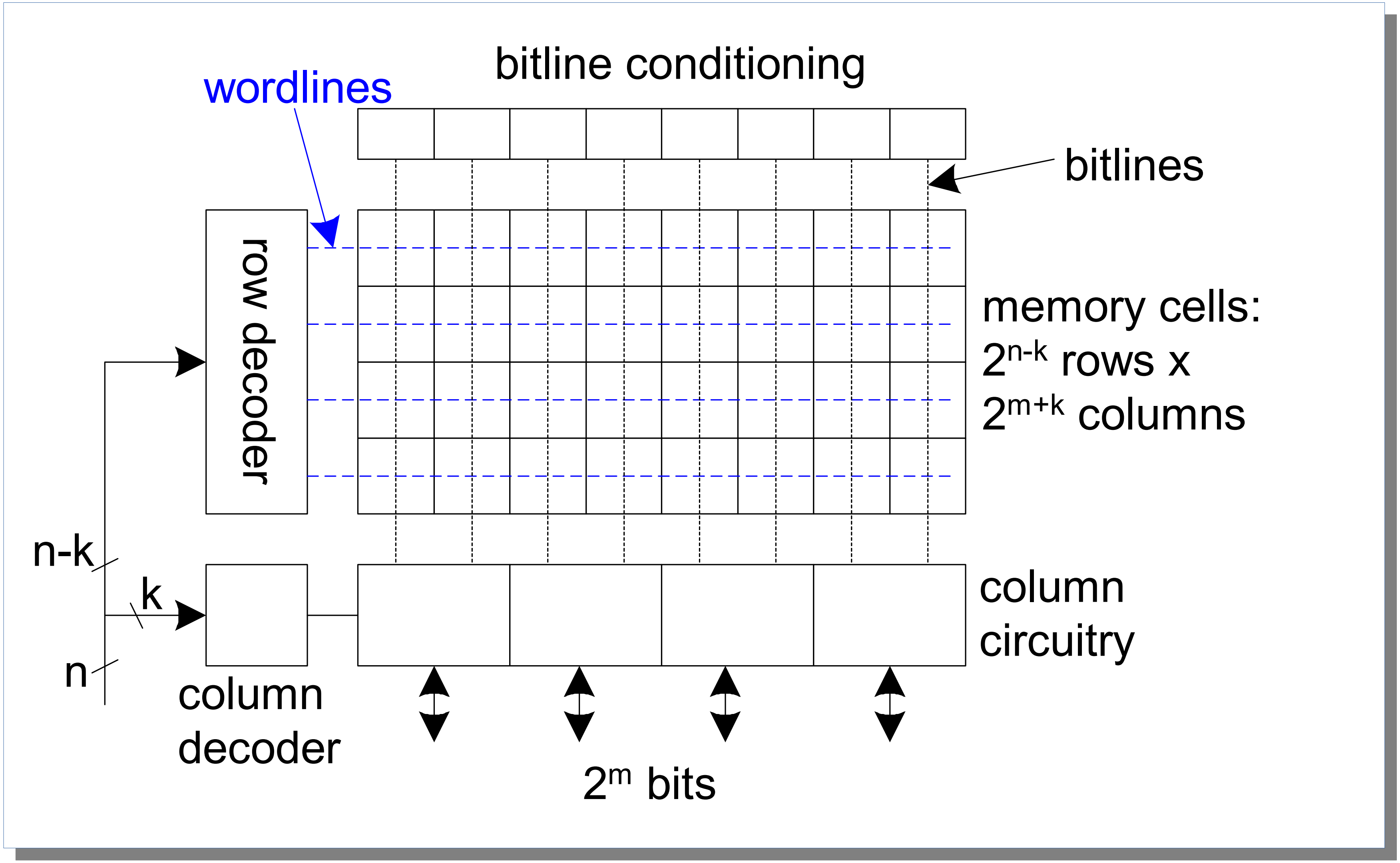
- Organized as 2^n words of 2^m bits each
  - Usually n >> m (1M vs. 64)
- Fold array to 2^{n-k} rows x 2^{m+k} columns



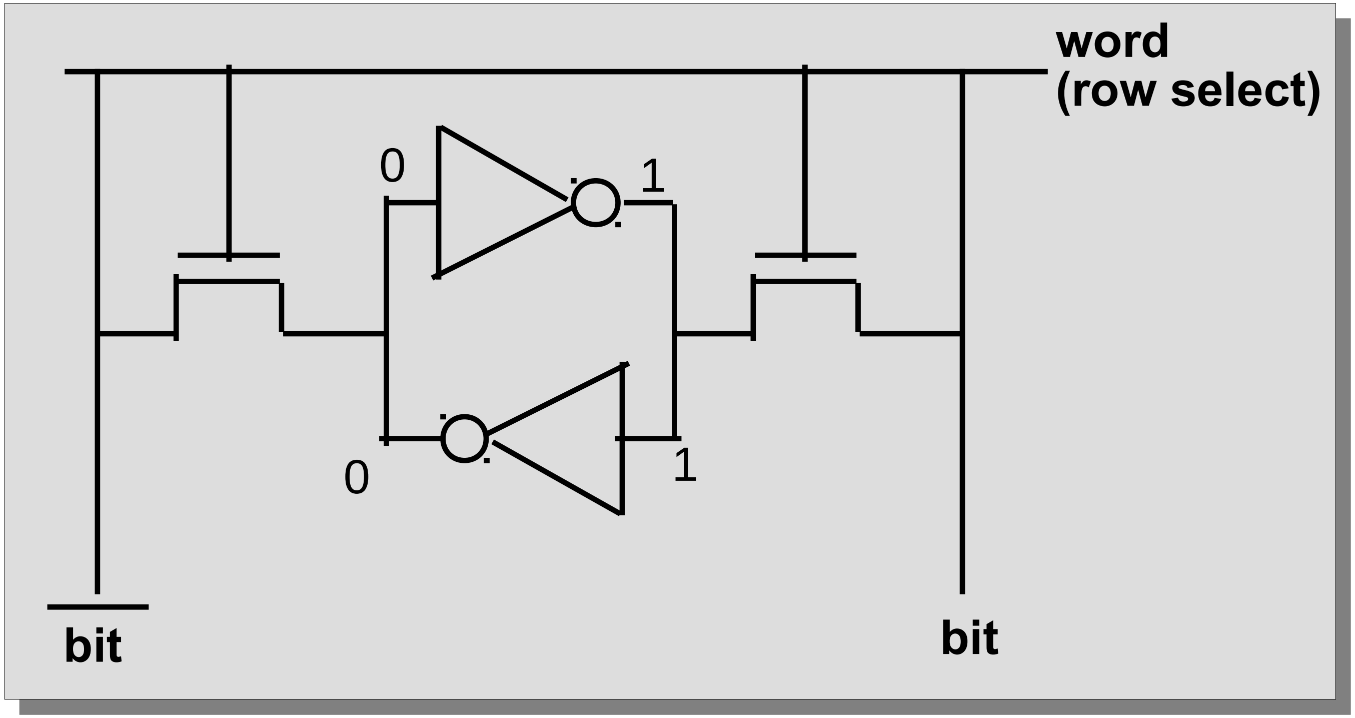
Memory Array



Memory Array



6T Static RAM Cell



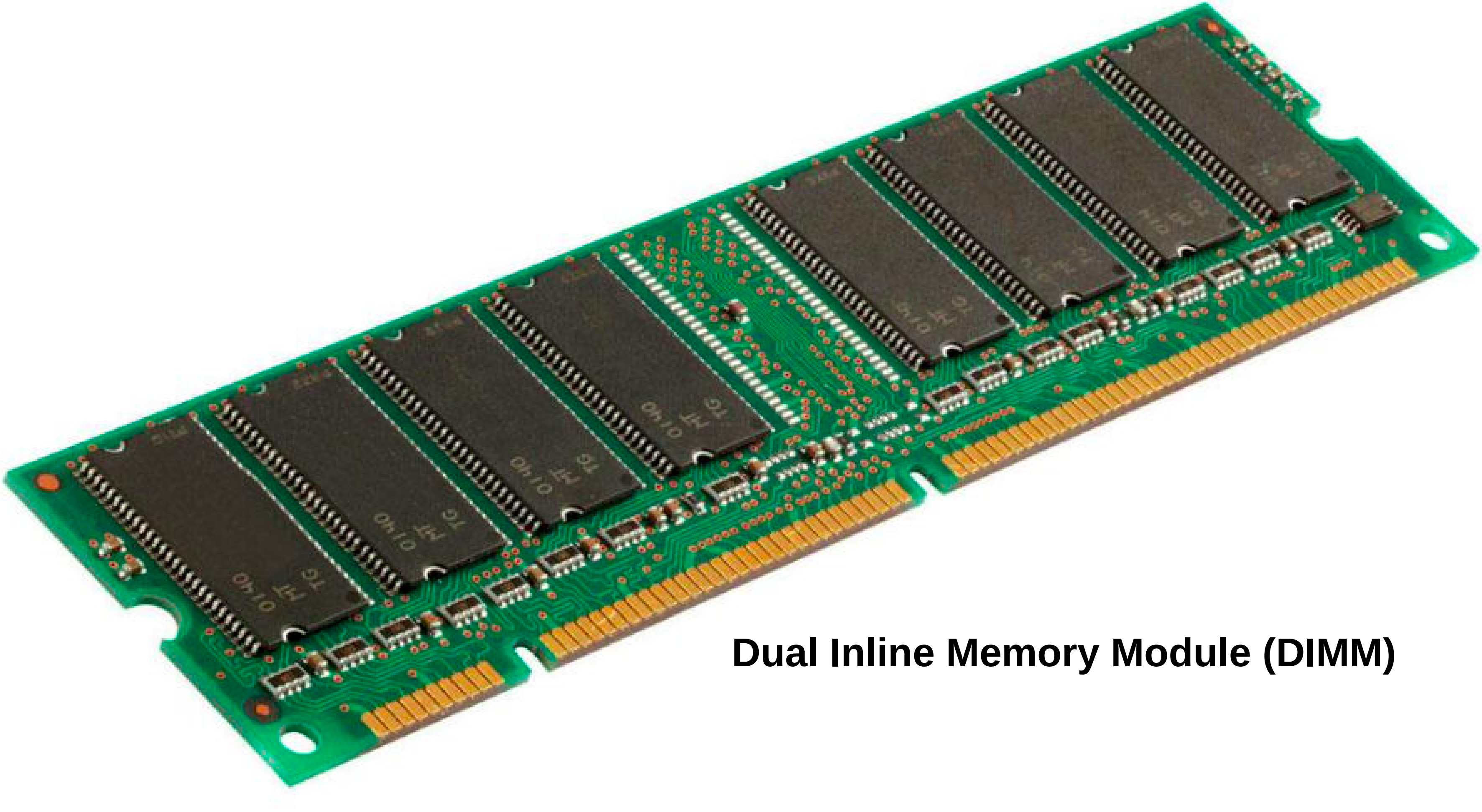
6T SRAM Cell Operation

- Read:
  - Precharge bit, bit_b
  - Raise wordline
  - Cell puts value into bit and its complement in bit_b
  - Sense amplifiers sense difference between bit and bit_b

6T SRAM Cell Operation

- Write:
  - Drive data onto bit, bit_b
  - Raise wordline
  - Access transistors set the cell to new state

DRAM – Main Memory



Dual Inline Memory Module (DIMM)

Memory Technology

- Main memory serves as input and output to I/O interfaces and the processor.
- DRAMs for main memory, SRAM for caches

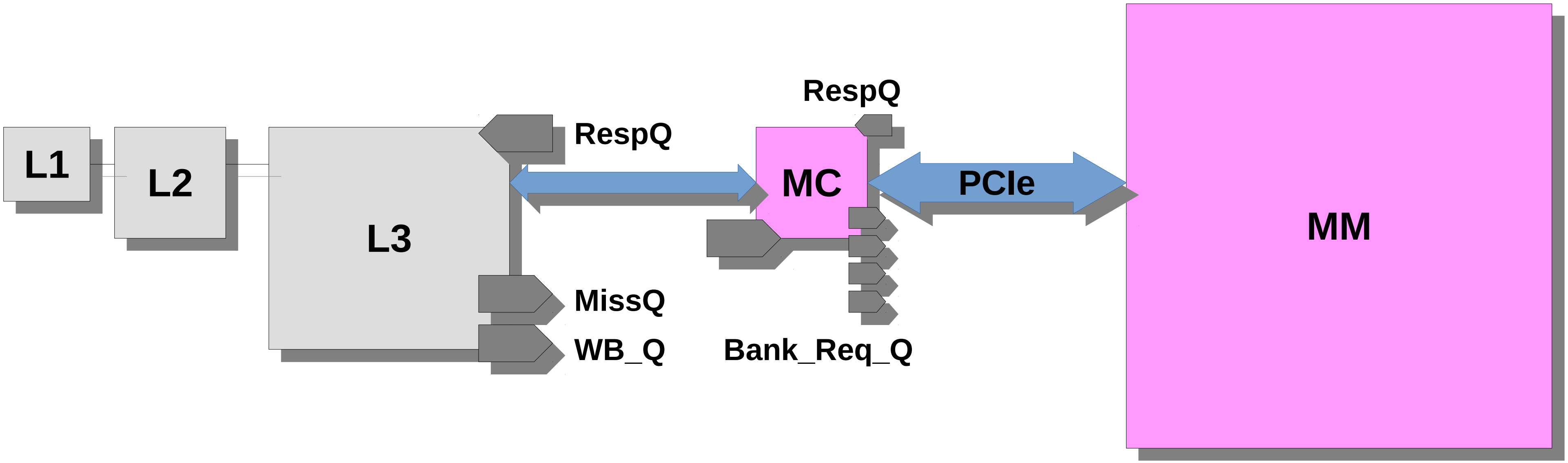
DRAM

- Dynamic Random Access Memory (DRAM)
  - 8x more dense than SRAM
  -

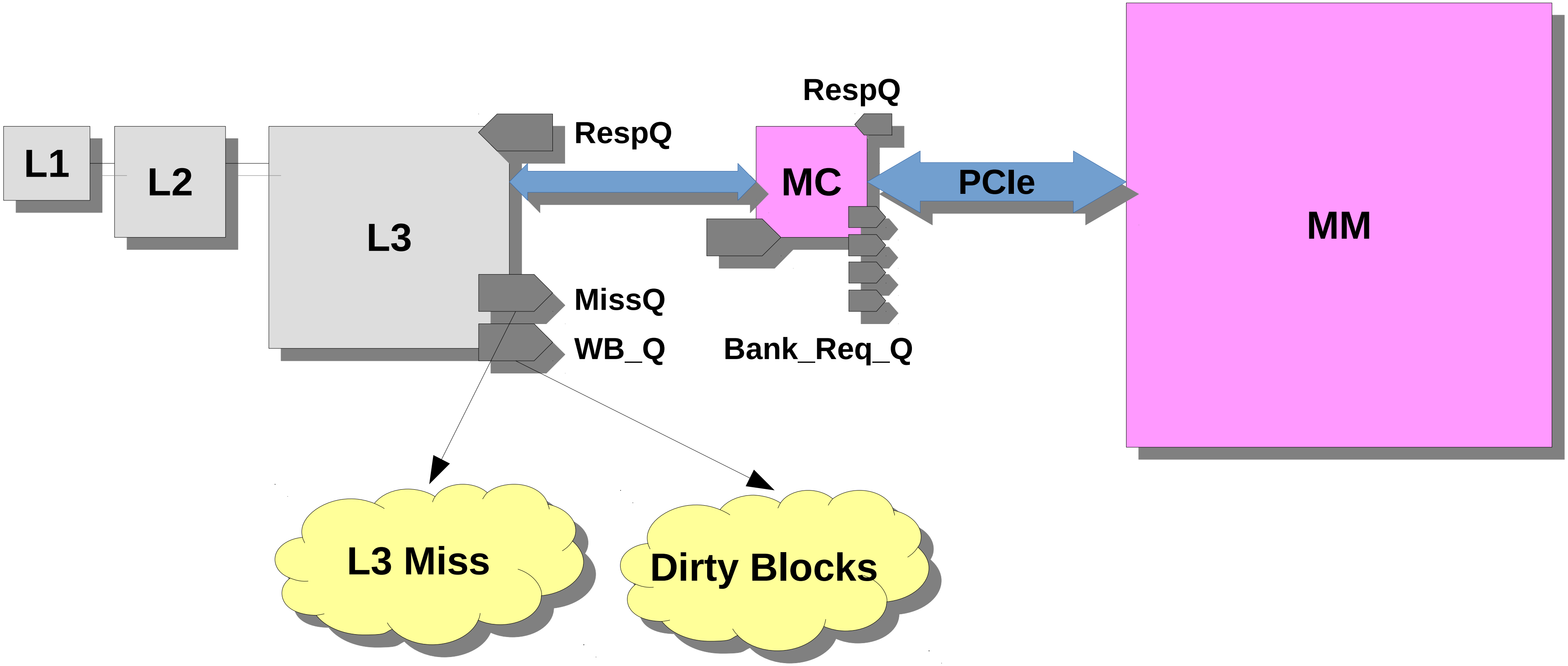
DRAM

- Dynamic Random Access Memory (DRAM)
  - 8x more dense than SRAM
  - Dynamic: Charge leak
  - Must be re-written after being read
  - Must be periodically refreshed

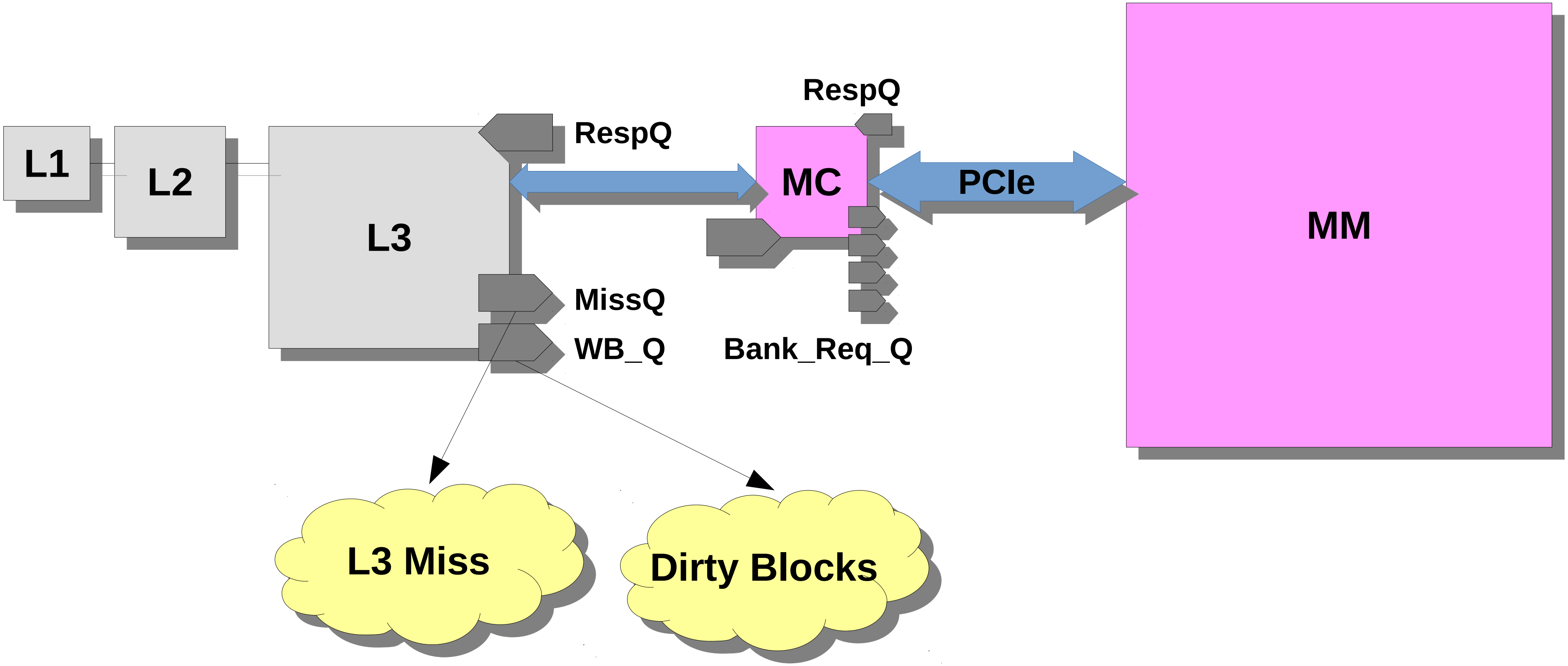
Main Memory



Main Memory


L1
L2
L3
RespQ
MissQ
WB_Q
RespQ
MC
PCIe
MM
Bank_Req_Q
L3 Miss
Dirty Blocks

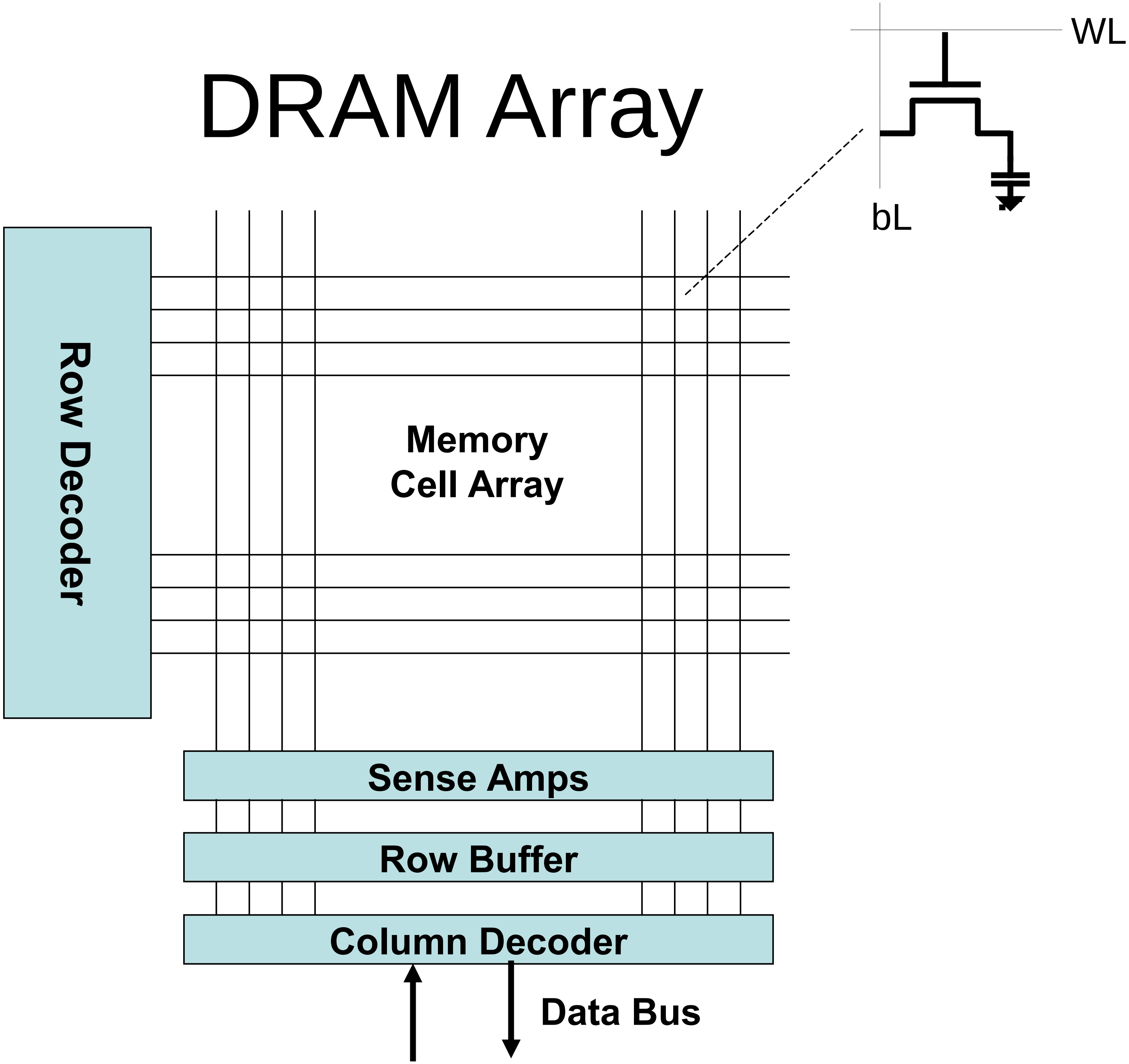
Main Memory


L1
L2
L3
RespQ
MissQ
WB_Q
RespQ
MC
PCIe
MM
Bank_Req_Q
L3 Miss
Dirty Blocks

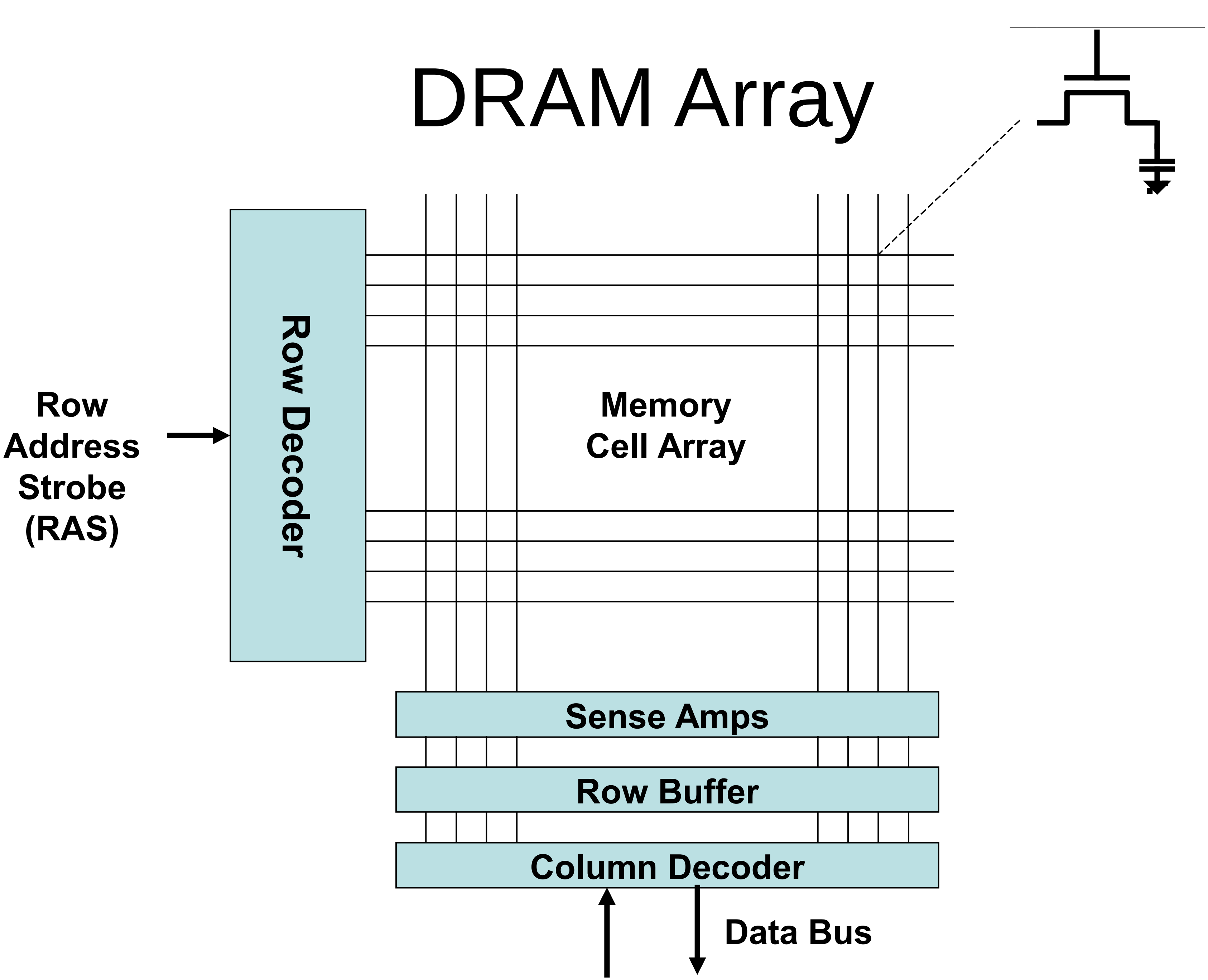
Memory Controller

- Schedule access requests
- Decodes the bank address(es) from the head(s) of the Q
- Transfers the address(es) into corresponding bank(s)
- Respond to the L3 Response queue

DRAM Array

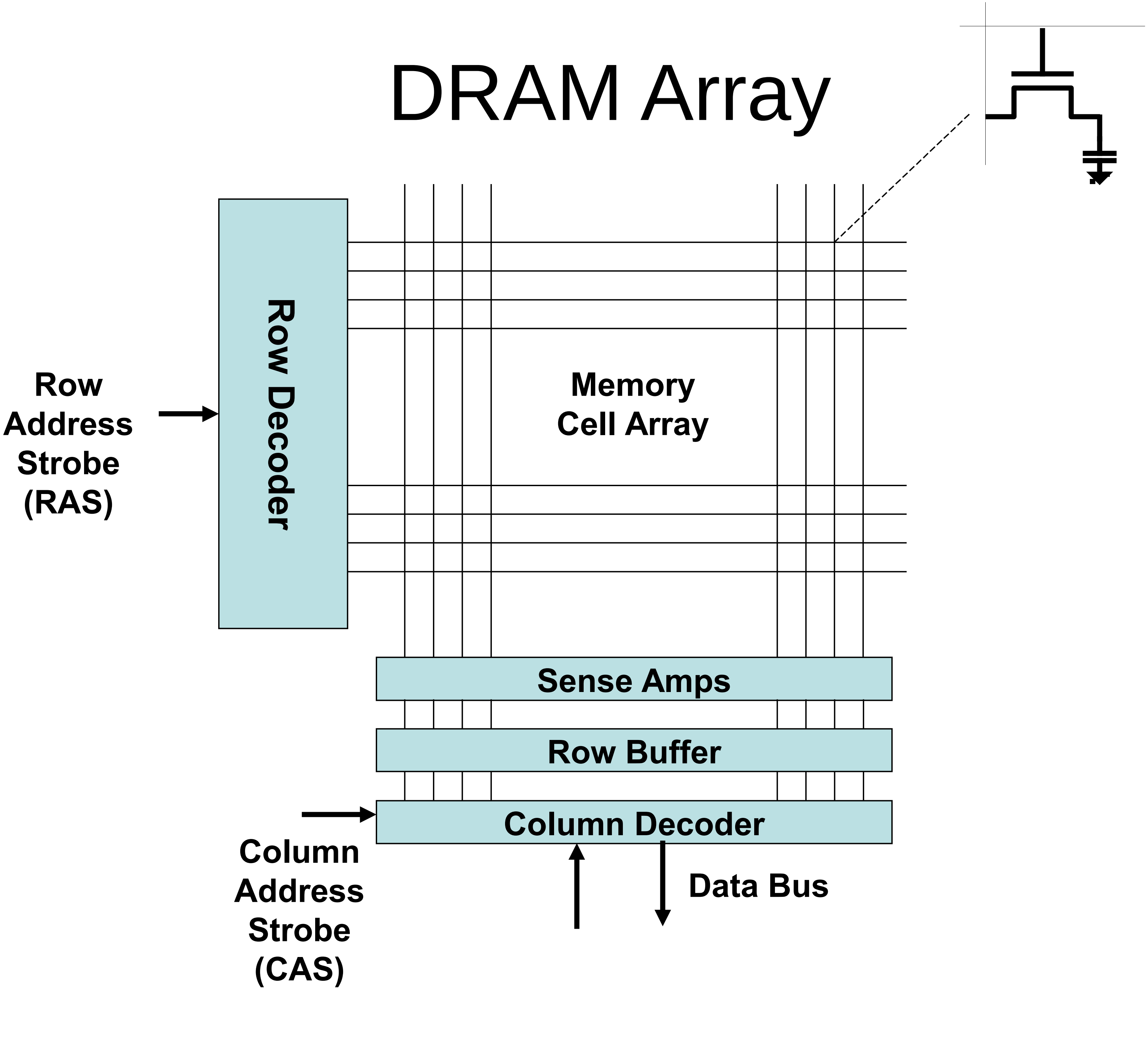

WL
bL
Row Decoder
Memory
Cell Array
Sense Amps
Row Buffer
Column Decoder
Data Bus

DRAM Array

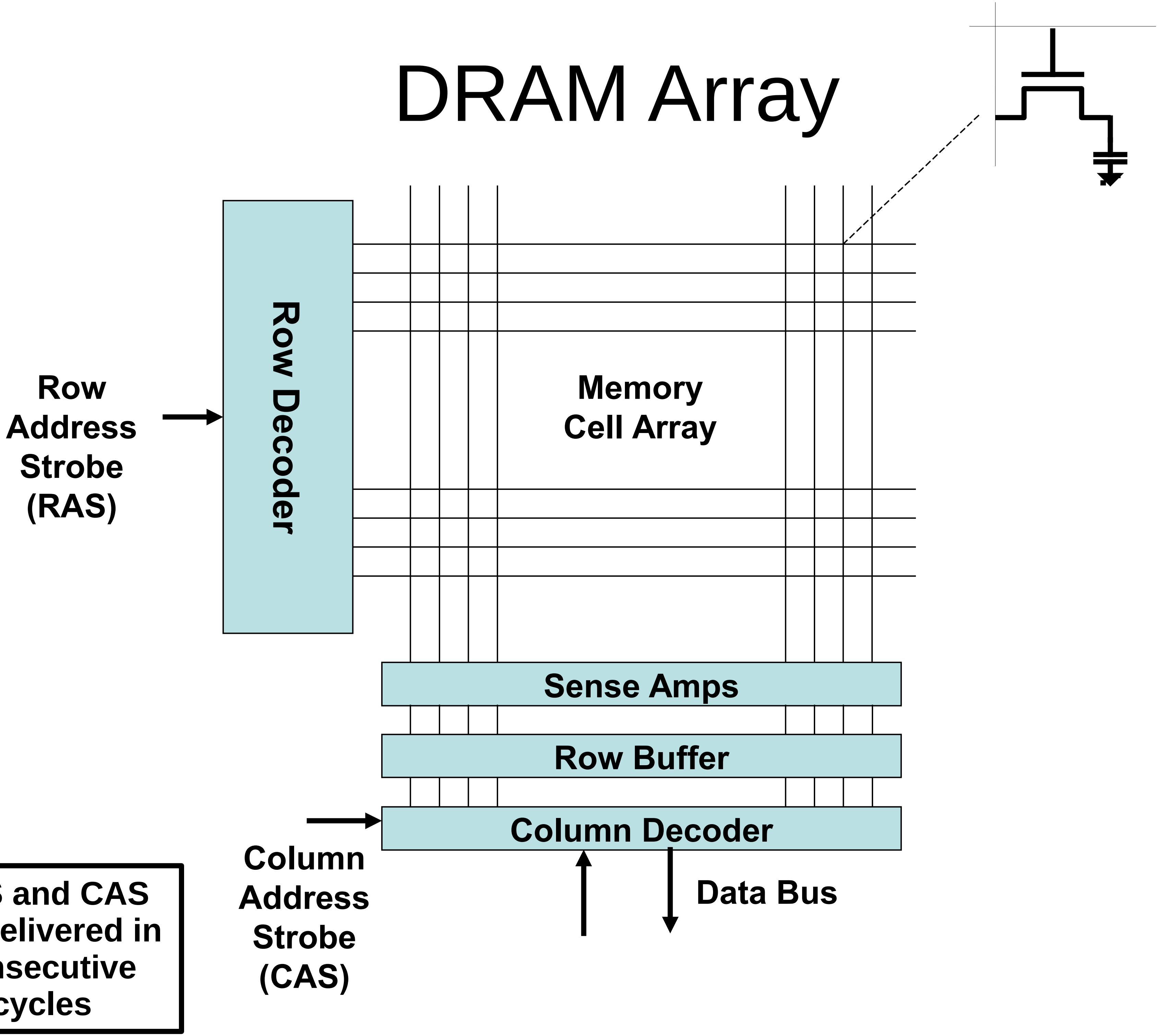


Row Address Strobe (RAS)
Row Decoder
Memory Cell Array
Sense Amps
Row Buffer
Column Decoder
Data Bus

DRAM Array

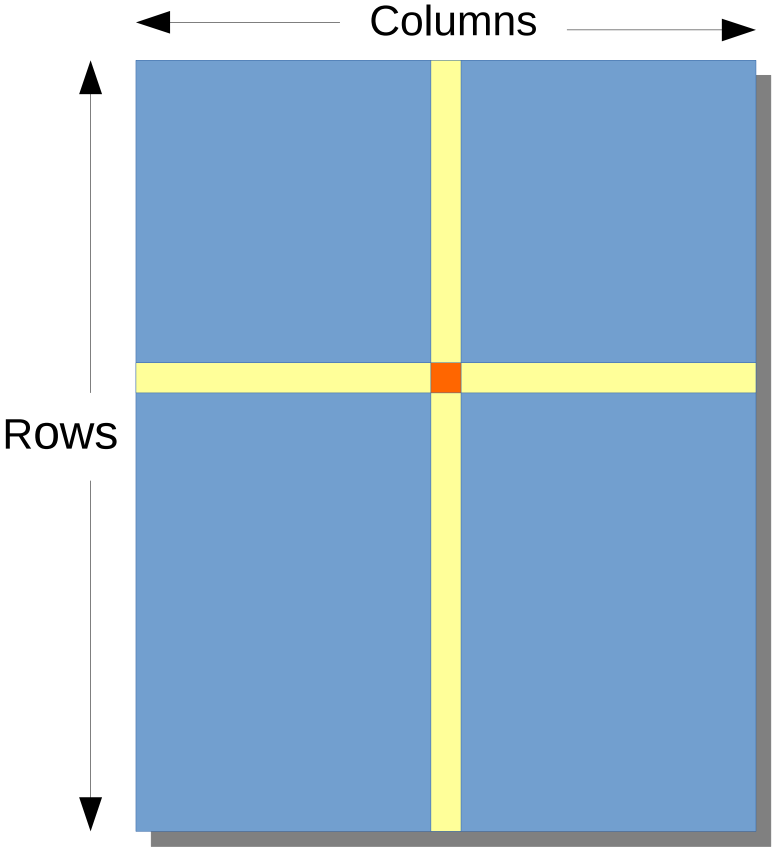

Row Decoder
Row
Address
Strobe
(RAS)
Memory
Cell Array
Sense Amps
Row Buffer
Column Decoder
Column
Address
Strobe
(CAS)
Data Bus

DRAM Array

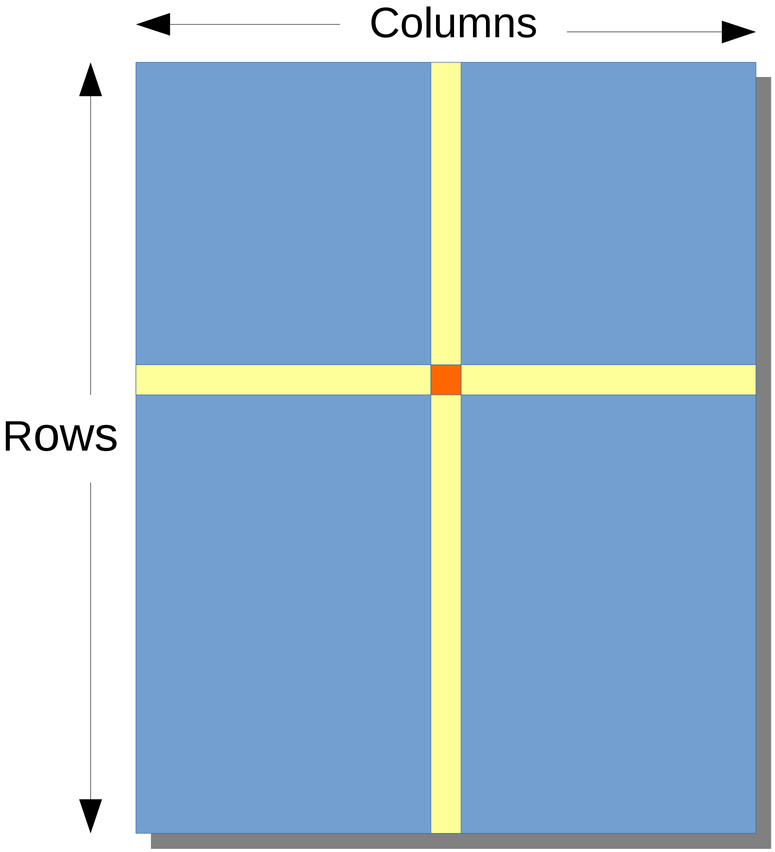


RAS and CAS are delivered in consecutive cycles

DRAM Bank

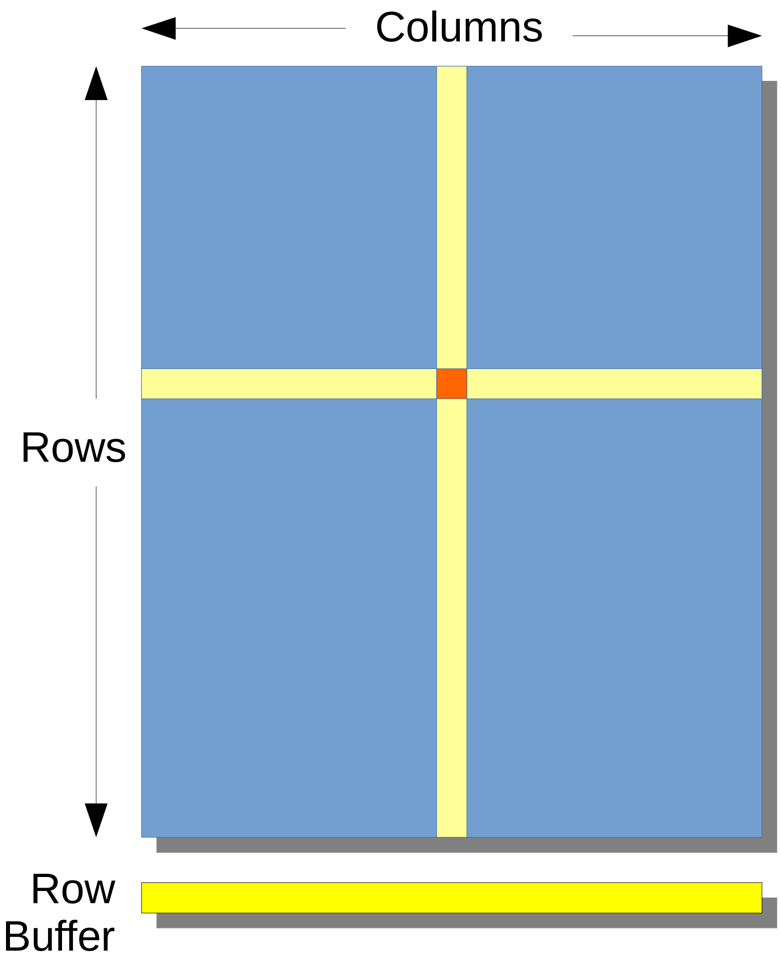


DRAM Bank



Row Address Strobe selects a Row

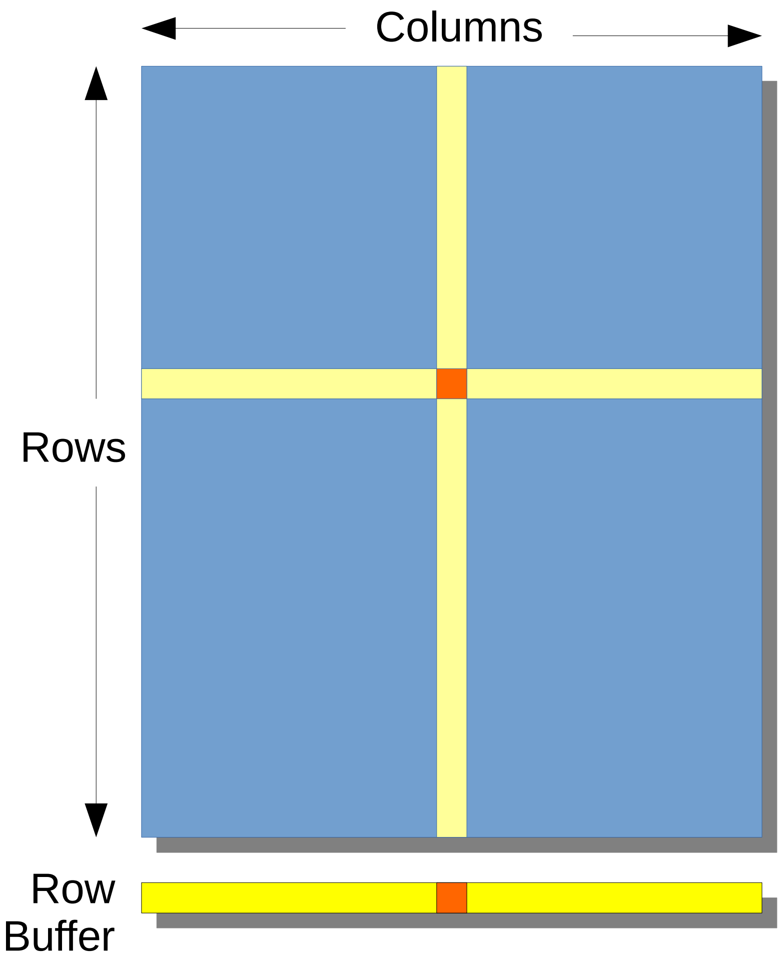
DRAM Bank



Row Address Strobe selects a Row

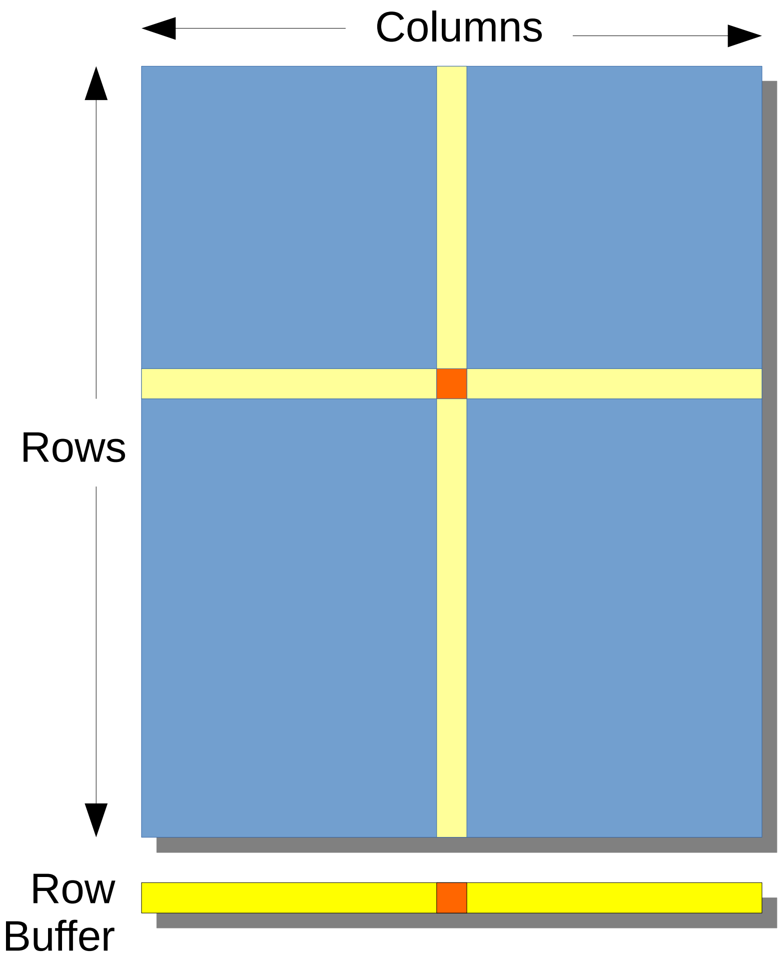
Row is read into the **Row Buffer**

DRAM Bank



Column Address Strobe

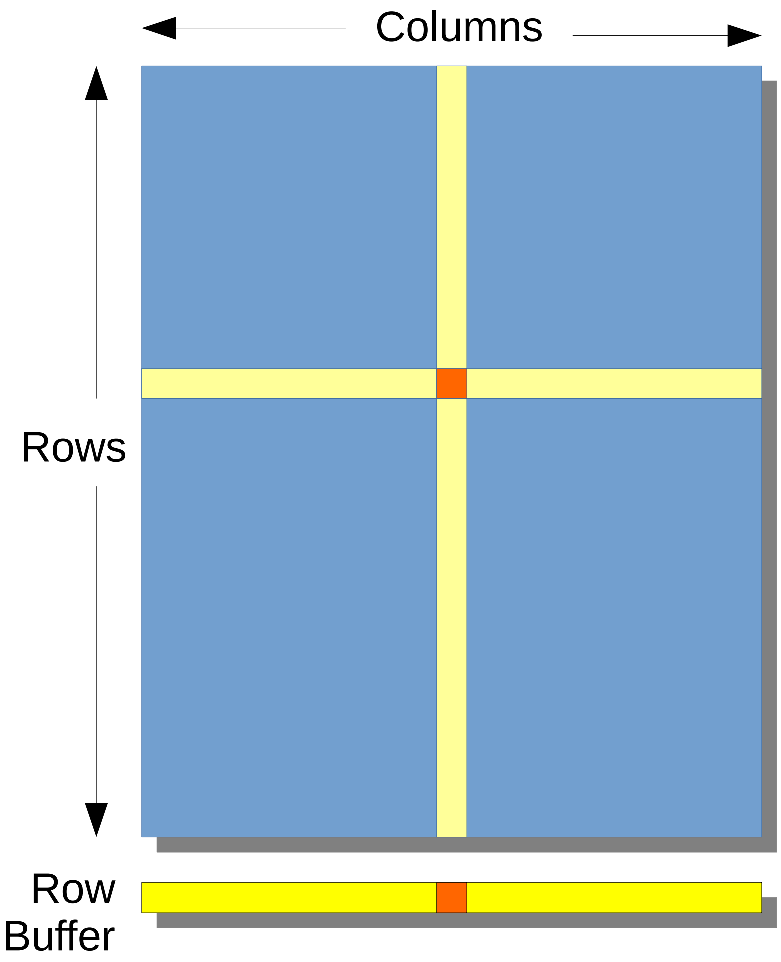
DRAM Bank



Column Address Strobe

Row Buffer is not changed
till a new Row is read

DRAM Bank

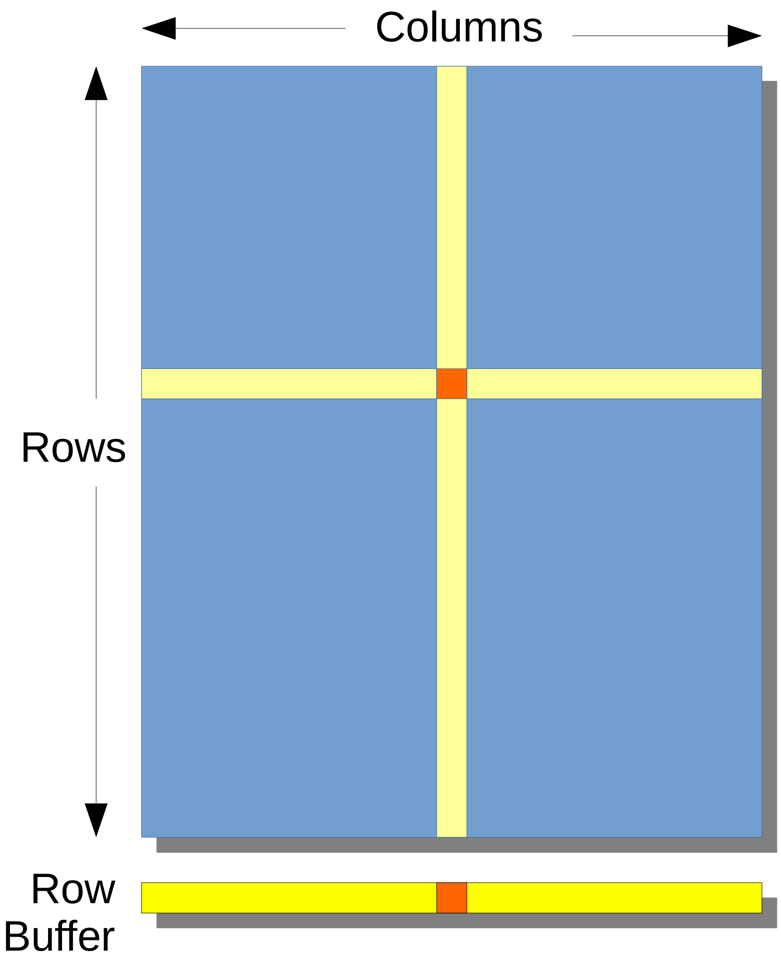


Column Address Strobe

Row Buffer is not changed till a new Row is read

Open Row Policy

DRAM Bank



Column Address Strobe

Row Buffer is not changed till a new Row is read

Open Row Policy

Maximize row buffer hits

DRAM Access

- Activate RAS
  - Precharge Row
- Row Access
- Turn on CAS

DRAM Access

- Activate RAS
  - Precharge Row
- Row Access
- Turn on CAS
- Access Latency:

$$t_{Access} = t_{Precharge} + t_{Row\ Access} + t_{CAS}$$

Row Buffer

- Exploits spatial locality of reference
- The most recent row read from a bank
- Acts like a cache

Rank, Bank

- Bank: a subset of a rank that is busy during one request
  - 4, 8 or 16 in one chip

Rank, Bank

- Bank: a subset of a rank that is busy during one request
  - 4, 8 or 16 in one chip
- Rank: a collection of DRAM chips that work together to respond to a request and keep the data bus full

Row Buffers

- Each bank has a single row buffer
- Row buffers act as a cache within DRAM

Row Buffers

- **Row buffer hit**: ~20 ns access time (time to move data from row buffer to pins)

Row Buffers

- **Row buffer hit**: ~20 ns access time (time to move data from row buffer to pins)
- **Empty row buffer access**: ~40 ns (read arrays + move data from row buffer to pins)

Row Buffers

- **Row buffer hit**: ~20 ns access time (time to move data from row buffer to pins)
- **Empty row buffer access**: ~40 ns (read arrays + move data from row buffer to pins)
- **Row buffer conflict**: ~60 ns (precharge bitlines + read new row + move data to pins)

Row Buffers

- **Row buffer hit**: ~20 ns access time (time to move data from row buffer to pins)
- **Empty row buffer access**: ~40 ns (read arrays + move data from row buffer to pins)
- **Row buffer conflict**: ~60 ns (precharge bitlines + read new row + move data to pins)
- Waiting time in the Queue (tens of nano-seconds) and incur address/cmd/data transfer delays (~10 ns)

DRAM Refresh

- Every DRAM cell must be refreshed within a 64 ms window
- A row read/write automatically refreshes the row
- Every refresh command performs refresh on a number of rows, the memory system is unavailable during that time
- A refresh command is issued by the memory controller once every 7.8µs on average
  - 8192 rows in RAM. 64ms/8192 = 7.8µs

Error Correction

- SECDED – single error correct double error detect
  - 8b code for every 64-bit word

Error Correction

- SECDED – single error correct double error detect
  - 8b code for every 64-bit word
- A rank is now made up of 9 x8 chips, instead of 8 x8 chips

Error Correction

- SECDED – single error correct double error detect
  - 8b code for every 64-bit word
- A rank is now made up of 9 x8 chips, instead of 8 x8 chips
- Stronger forms of error protection exist: a system is **chipkill correct** if it can handle an entire DRAM chip failure

Future Memory Trends

- Processor pin count is not increasing
- High memory bandwidth requires high pin frequency
- 3D stacking can enable high memory capacity and high channel frequency (e.g., Micron HMC)
- Phase Change Memory cells
- Silicon Photonics

References

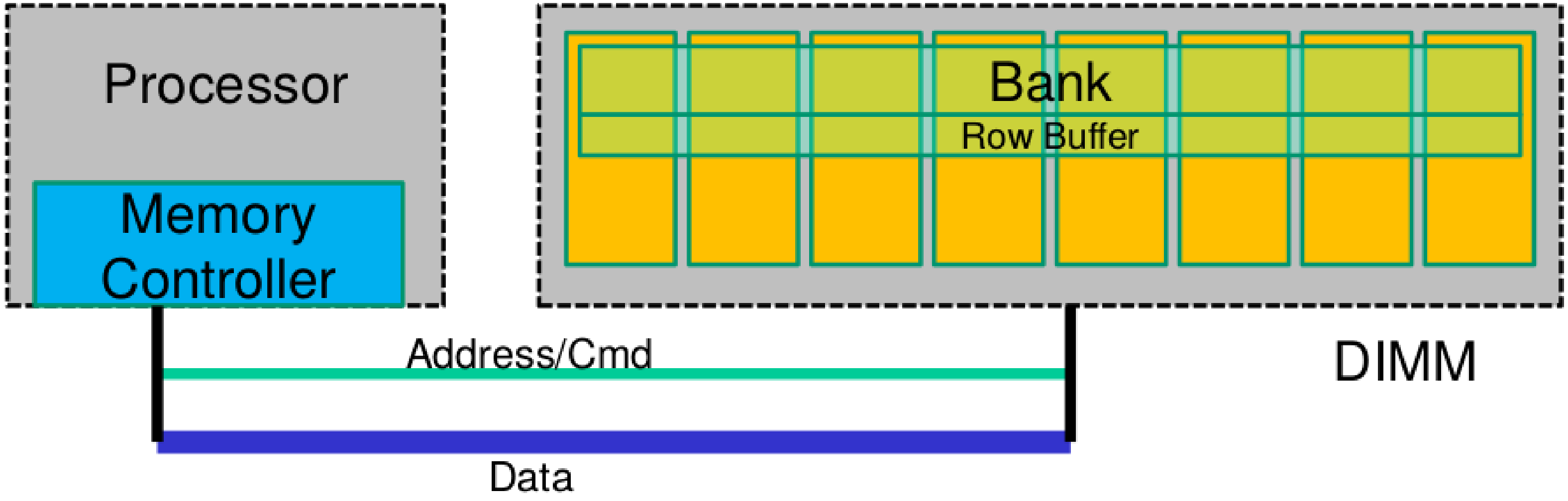
- Rajeev Balasubramonian, CS6810 – Computer Architecture. University of Utah.
- Hennessy and Patterson. Computer Architecture. 5e. MK. Appendix B, Chapter 2.
- Bruce Jacob, Spencer Ng, David Wang. Memory Systems: Cache, DRAM. Elsevier, 2007.

M2 – Outline

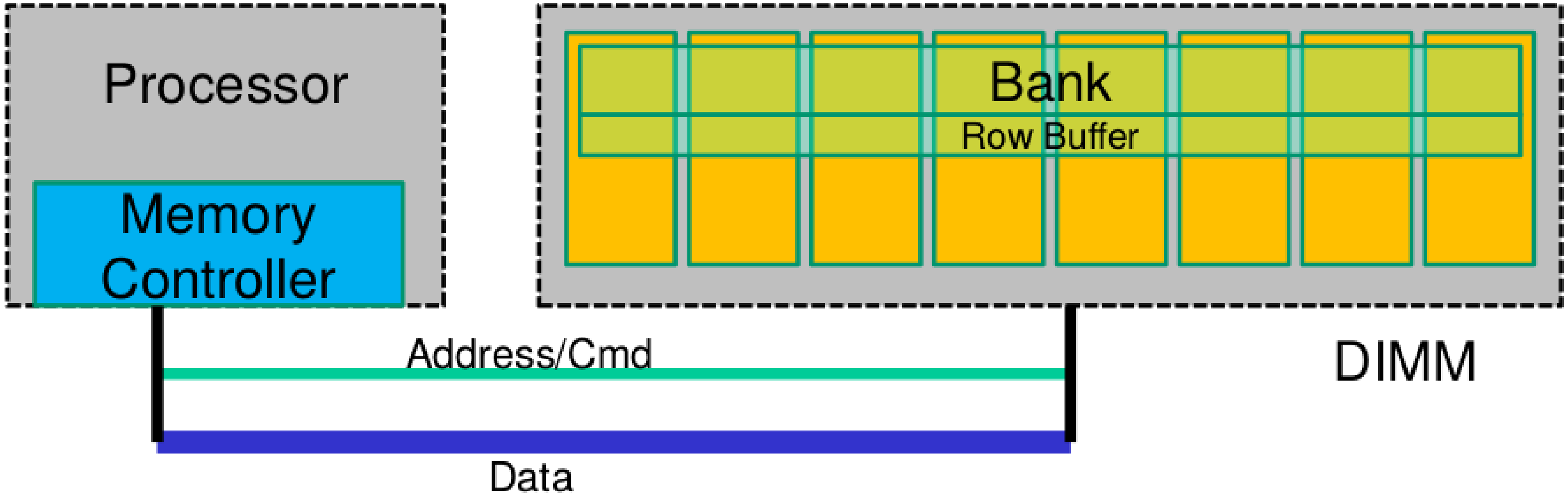
- Memory Hierarchy
- Cache Blocking – Cache Aware Programming
- SRAM, DRAM
- Virtual Memory
- Virtual Machines
- Non-volatile Memory, Persistent NVM

Extra

Main Memory

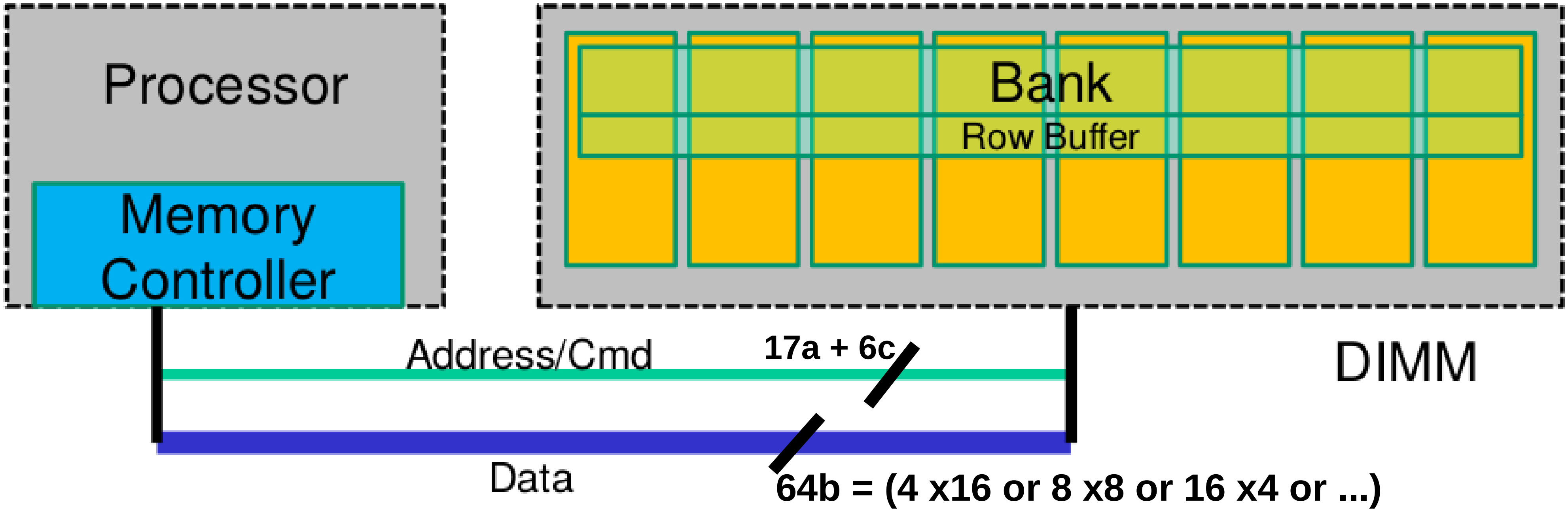

Processor
Memory Controller
Bank
Row Buffer
Address/Cmd
DIMM
Data

Main Memory

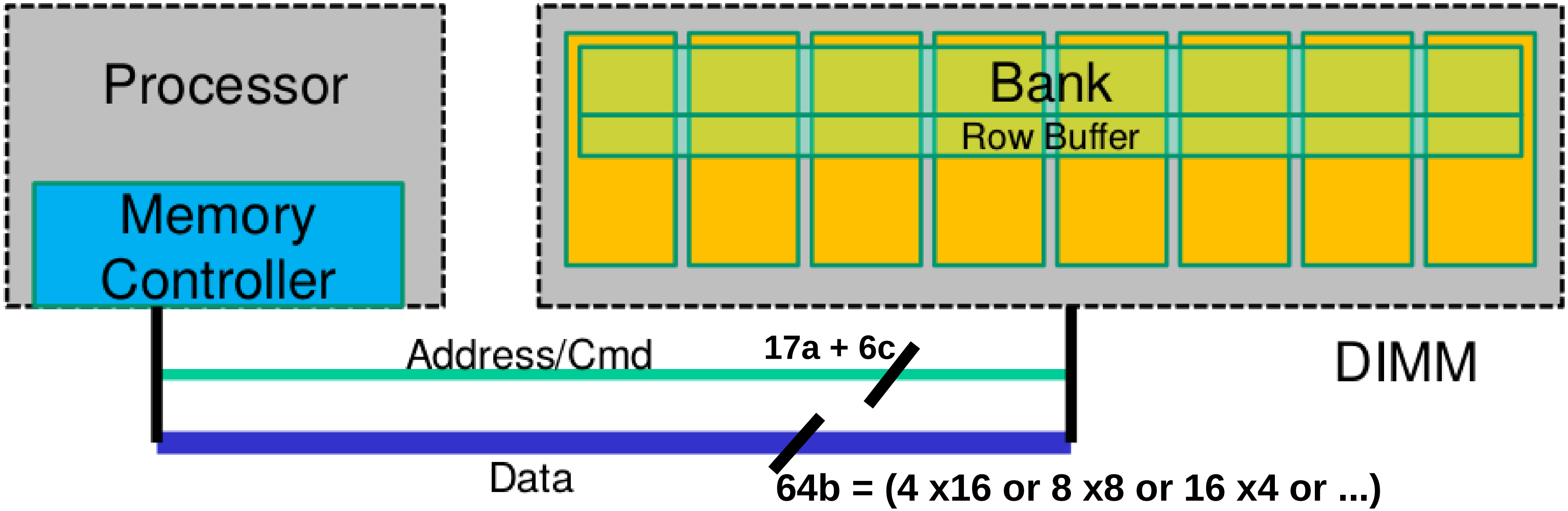


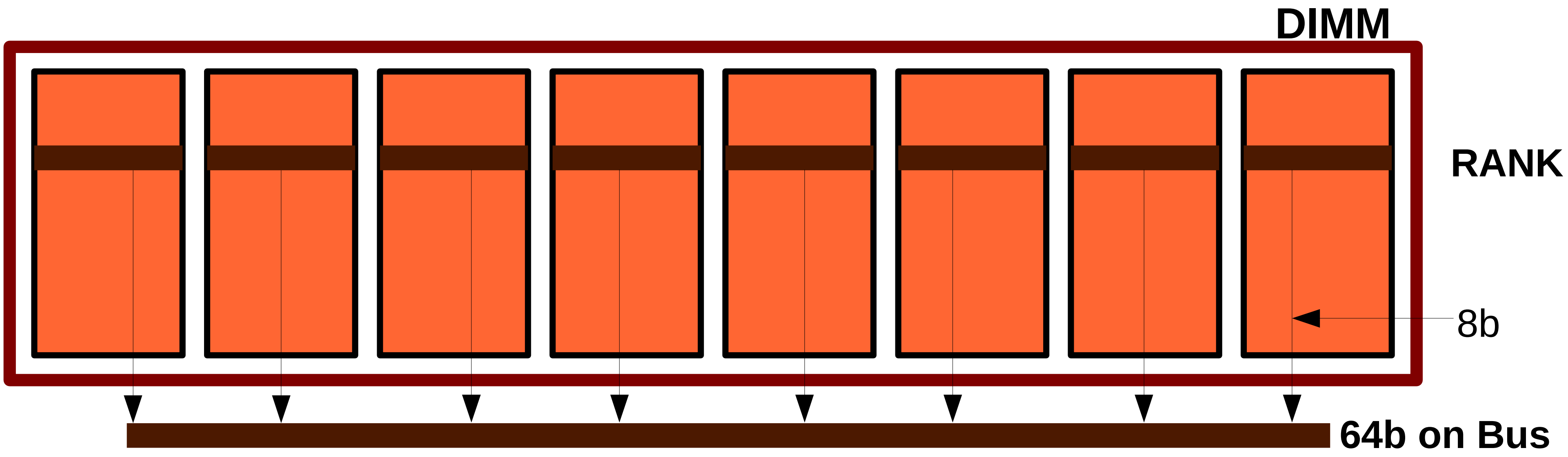
- Memory Channel = Data (64b) + Address/Cmd (23b = 17a + 6c)
- DIMM: a PCB with DRAM chips on the back and front
- Transfers one cache line size (64B) per address

Main Memory


Processor
Memory Controller
Bank
Row Buffer
DIMM
Address/Cmd
17a + 6c
Data
64b = (4 x16 or 8 x8 or 16 x4 or ...)

Main Memory


Processor
Memory Controller
Bank
Row Buffer
DIMM
Address/Cmd
17a + 6c
Data
64b = (4 x16 or 8 x8 or 16 x4 or ...)


DIMM
RANK
8b
64b on Bus

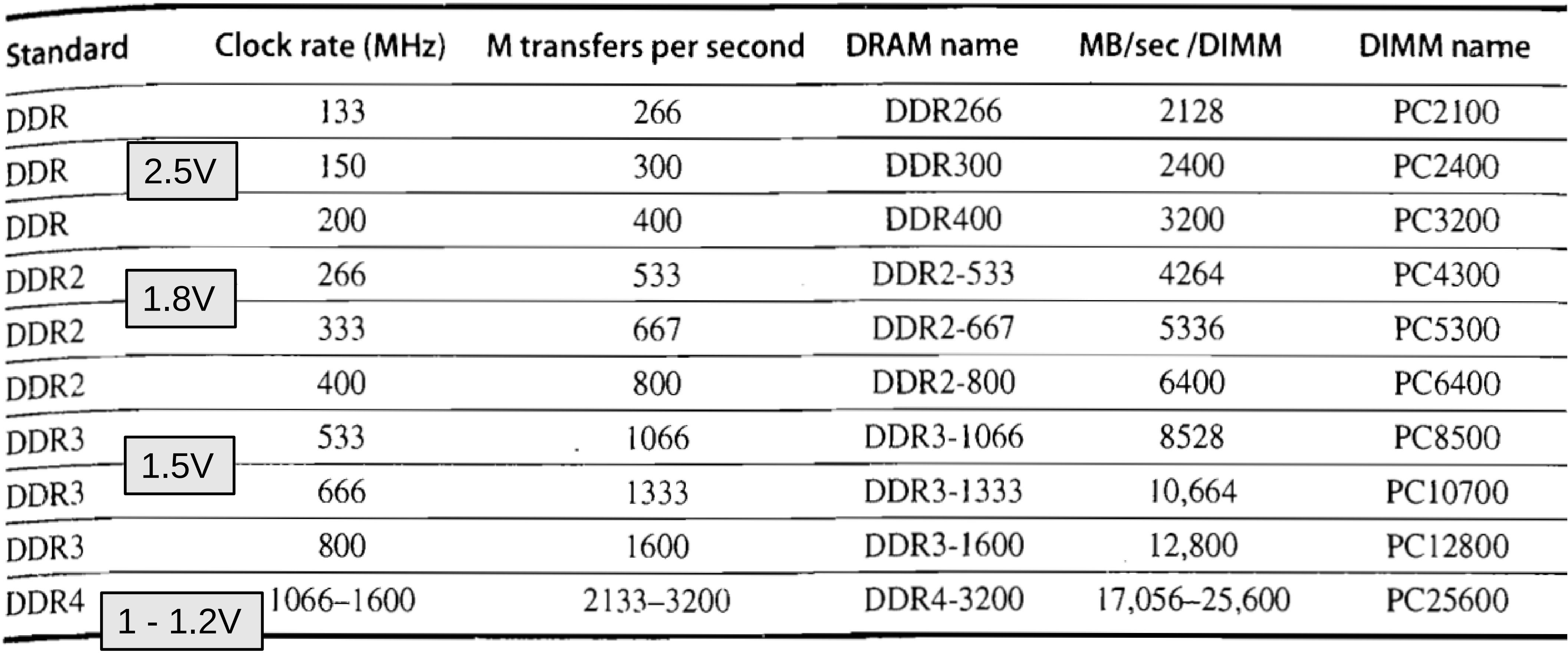
Memory Technology - Optimizations

- Multiple accesses to same row
- Synchronous DRAM
  - Clocked operation, Burst mode
- Wider interfaces
- Double data rate
- Multiple banks on each DRAM device

Scheduling Policies

- FCFS: Issue the first read or write in the queue that is ready for issue
  - RoB commits loads and stores in program order
- First Ready – FCFS: Issue loads that result in row buffer hits.
- Stall Time Fair: First issue row buffer hits, unless other threads are being neglected

Clock rates, Bandwidth and Names



Standard		Clock rate (MHz)	M transfers per second	DRAM name	MB/sec /DIMM	DIMM name
DDR		133	266	DDR266	2128	PC2100
DDR	2.5V	150	300	DDR300	2400	PC2400
DDR		200	400	DDR400	3200	PC3200
DDR2	1.8V	266	533	DDR2-533	4264	PC4300
DDR2		333	667	DDR2-667	5336	PC5300
DDR2		400	800	DDR2-800	6400	PC6400
DDR3	1.5V	533	1066	DDR3-1066	8528	PC8500
DDR3		666	1333	DDR3-1333	10,664	PC10700
DDR3		800	1600	DDR3-1600	12,800	PC12800
DDR4	1 - 1.2V	1066–1600	2133–3200	DDR4-3200	17,056–25,600	PC25600

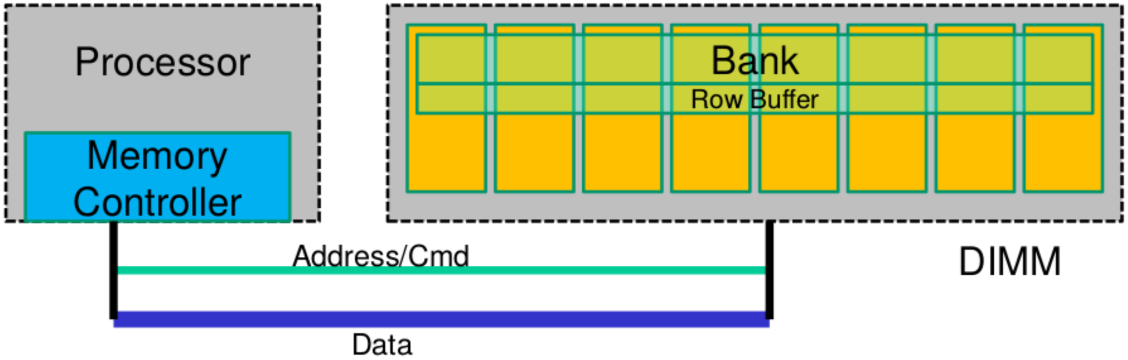
- GDDR5 – Graphics memory based on DDR3
  - 2x – 5x bandwidth per DRAM vs. DDR3
  - Wider interface, higher clockrate
  - Attached via soldering instead of socketted DIMM

Open/Closed Page Policies

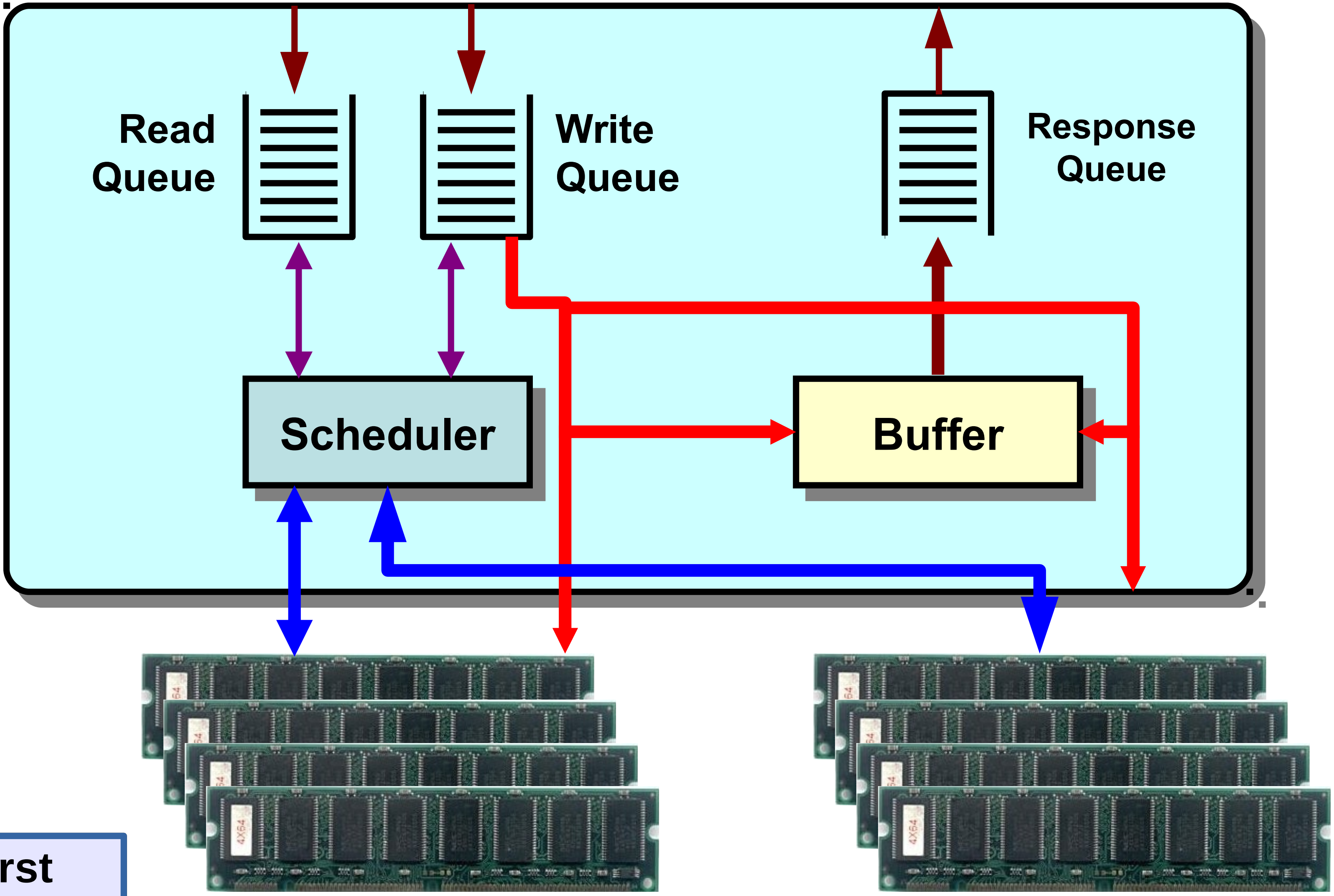
- **Open Page Policy**: Row buffers are kept open
  - Useful when access stream has locality
  - Row buffer hits are cheap (20ns)
  - Row buffer miss is a bank conflict and expensive (60ns)
- **Closed Page Policy**: Bitlines are precharged immediately after access
  - Useful when access stream has little locality
  - Nearly every access is a row buffer miss (40ns)
  - The precharge is usually not on the critical path
- Modern memory controller policies lie somewhere between these two extremes (usually proprietary)

Reads and Writes

- A single bus is used for reads and writes
- Bus direction must be reversed when switching between reads and writes
  - Takes time and leads to bus idling
- Writes are performed in bursts
- Write queue stores pending writes until a high watermark is reached



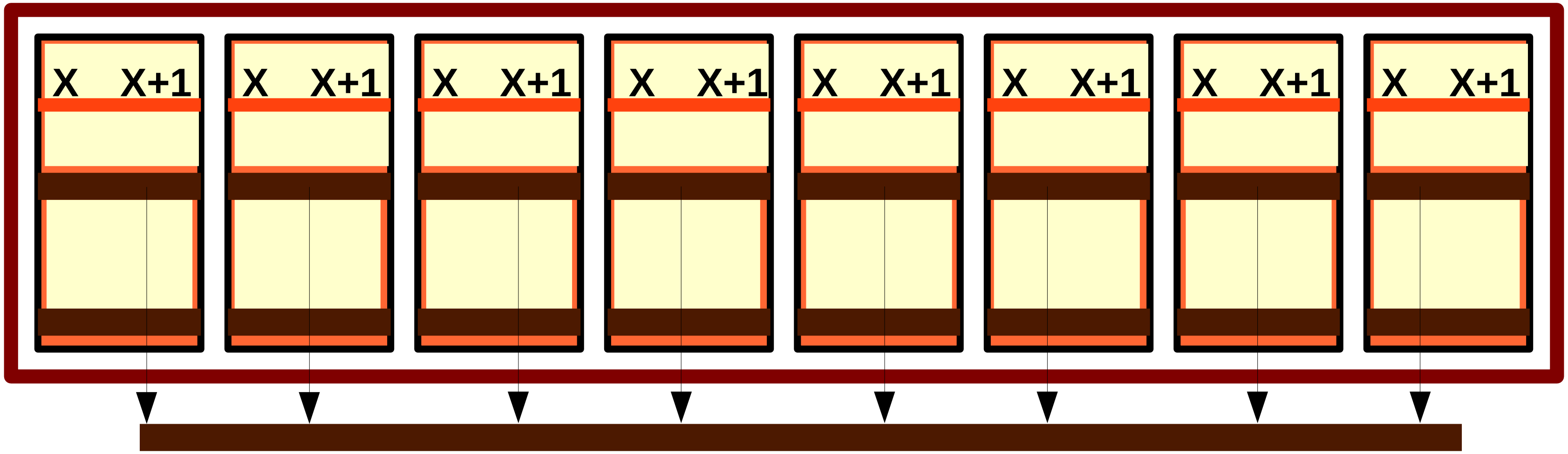
Memory Controller



FCFS, First Ready-FCFS, Stall Time Fair

Address Mapping Policies

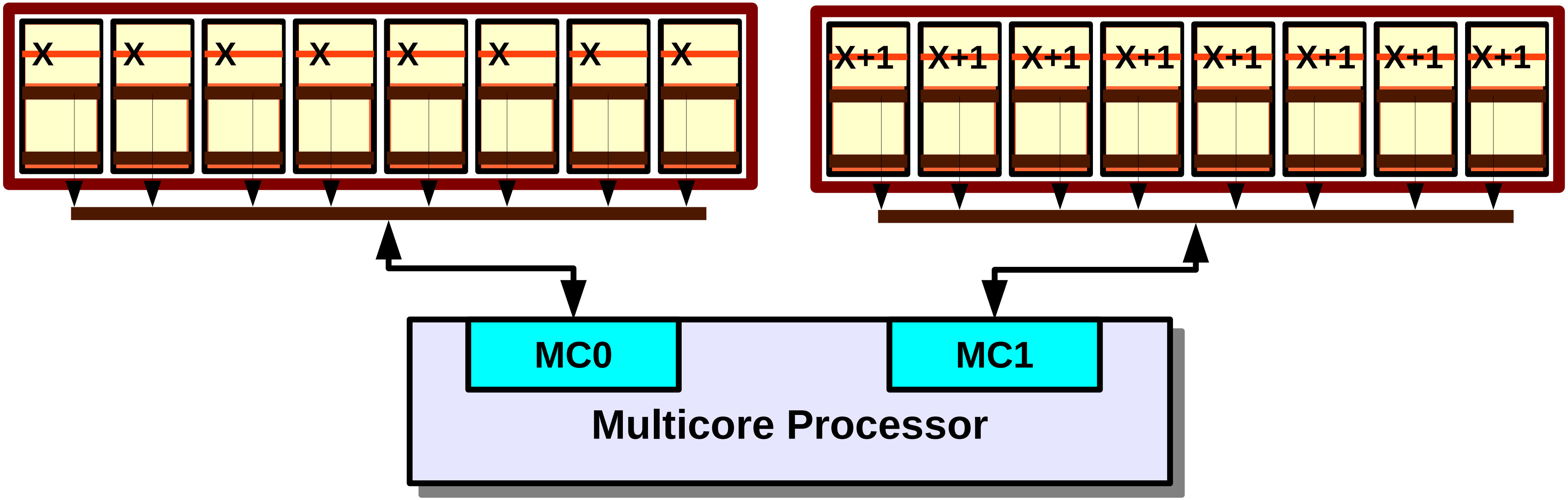
- Consecutive cache lines can be placed in the same row to boost row buffer hit rates
  - row:rank:bank:channel:column:blkoffset



- Time between access to cache block X and X+1 = 20ns (row buffer hit)

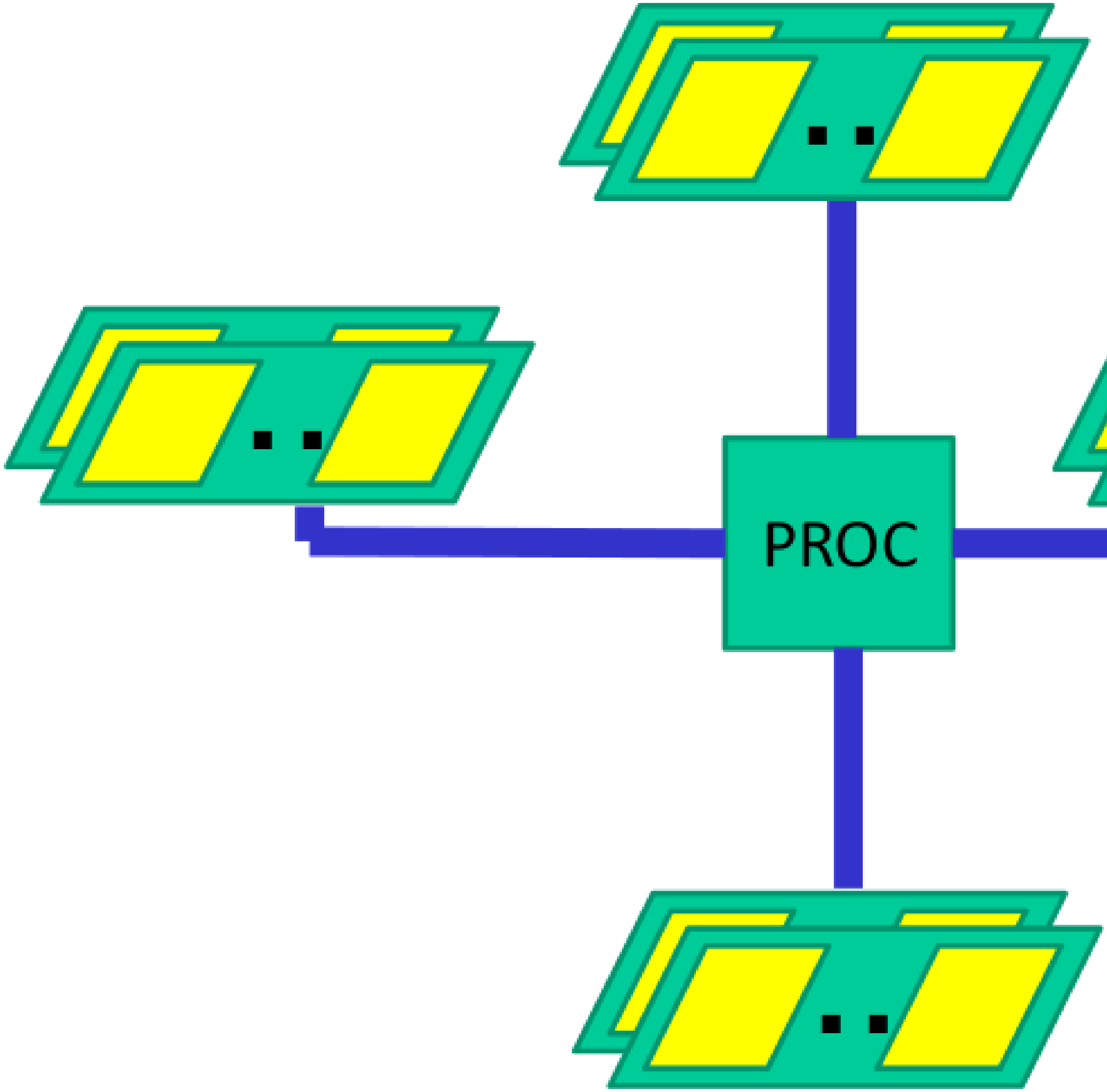
Address Mapping Policies

- Consecutive cache lines can be placed in different ranks to boost parallelism
  - row:column:rank:bank:channel:blkoffset



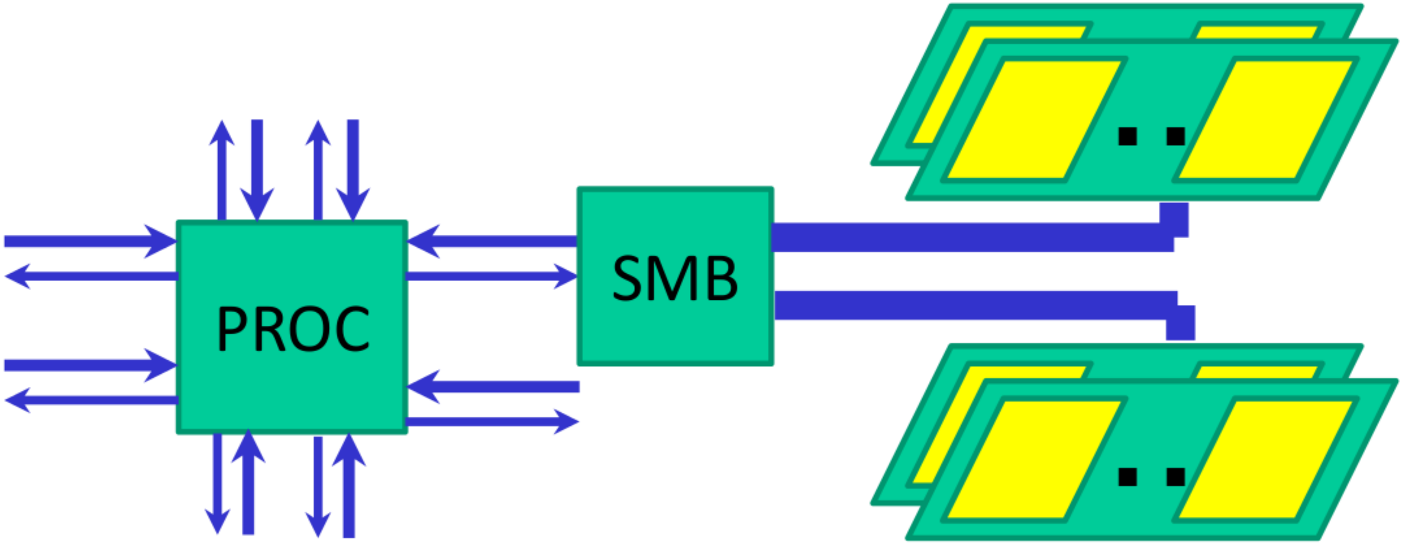
- Cache blocks X and X+1 can be accessed simultaneously

Modern Memory System



- 4 DDR3 channels
- 64-bit data channels
- 800 MHz channels
- 1-2 DIMMs/channel
- 1-4 ranks/channel

Modern Memory System

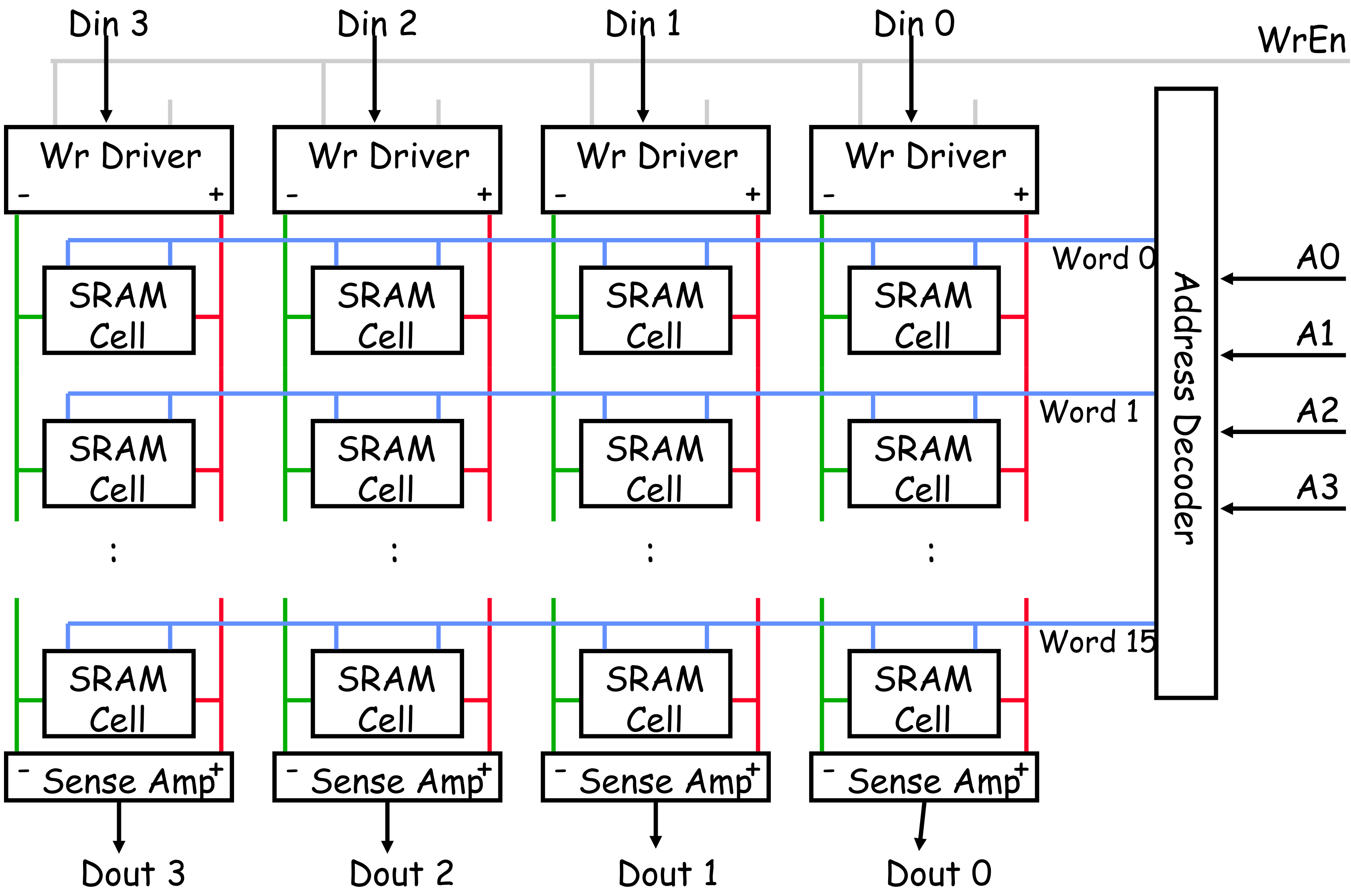


- The link into the processor is narrow and high frequency
- The Scalable Memory Buffer chip is a “router” that connects to multiple DDR3 channels (wide and slow)
- Boosts processor pin bandwidth and memory capacity
- More expensive, high power

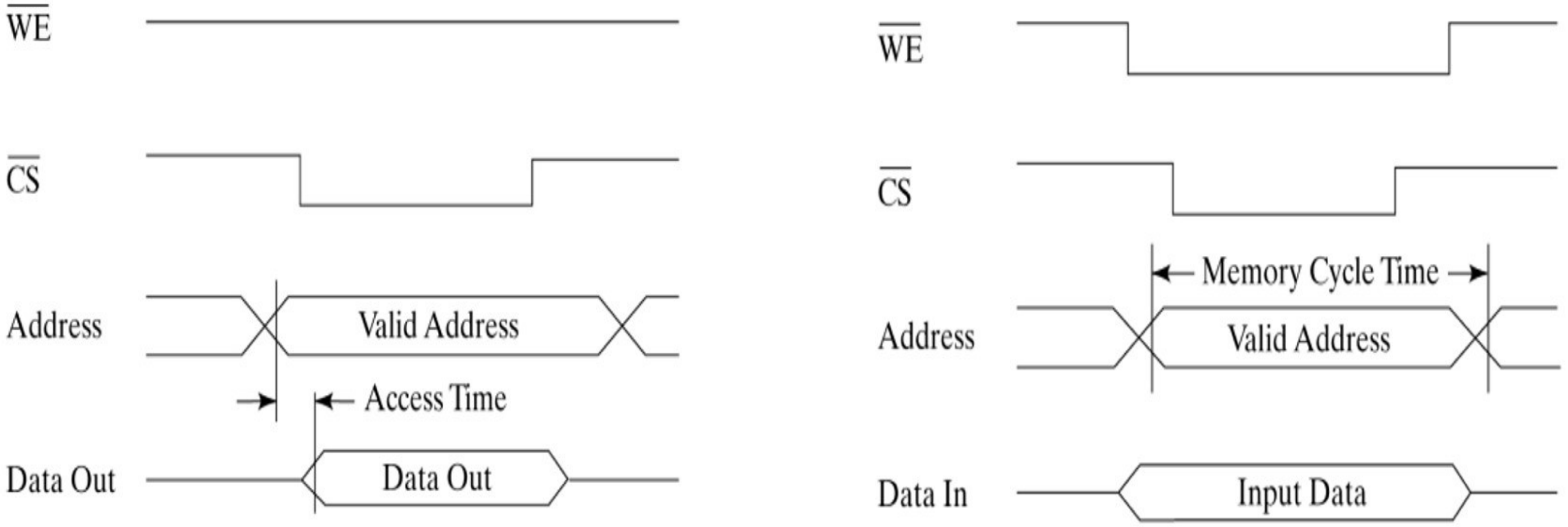
Memory Array

- 2^n *words* of 2^m *bits* each
- If n >> m, fold by 2^k into fewer *rows* of more *columns*
- *Good regularity – easy to design*
- Very high density if good cells are used

16-Word x 4-bit SRAM Organization

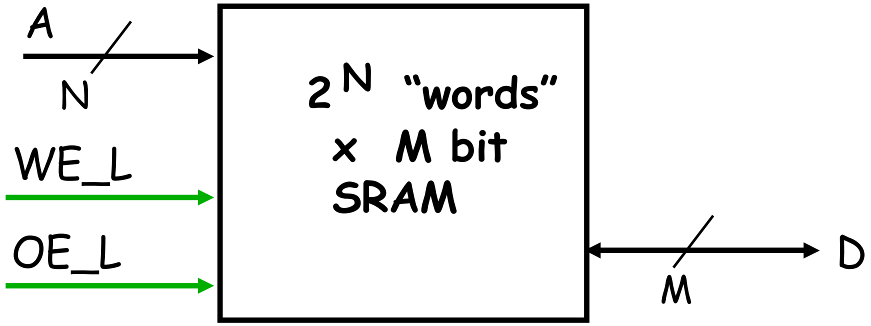


Simplified SRAM timing diagram



- Read: Valid address, then Chip Select
- Access Time: address good to data valid
  - even if not visible on out
- Cycle Time: min between subsequent mem operations
- Write: Valid address and data with WE_l, then CS
  - Address must be stable a setup time before WE and CS go low
  - And hold time after one goes high
- When do you drive, sample, or Z the data bus?

Logic Diagram of a Typical SRAM

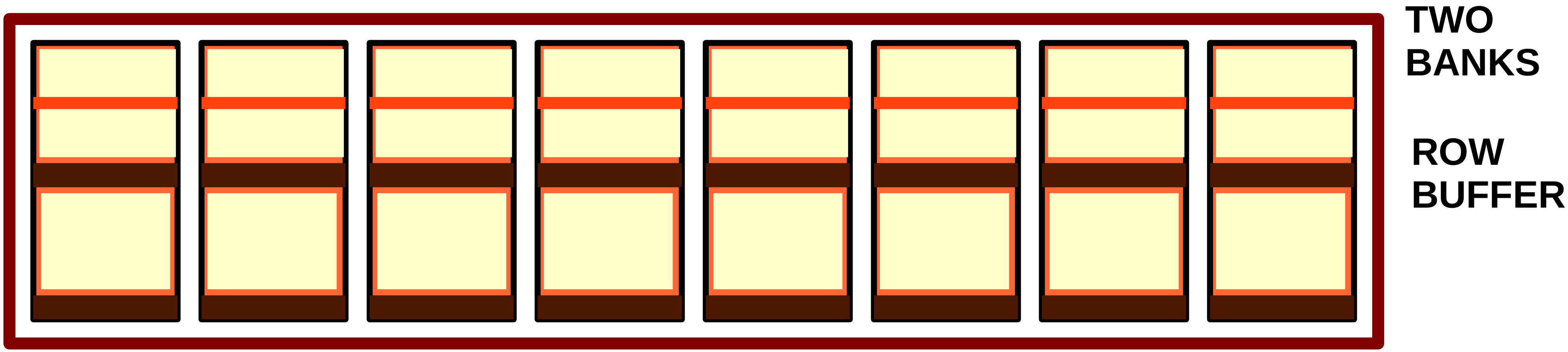


- Write Enable is usually active low (WE_L)
- Din and Dout are combined to save pins:
- A new control signal, Output Enable (OE_L)
  - WE_L is asserted (Low), OE_L is unasserted (High)
    - » D serves as the data input pin
  - WE_L is unasserted (High), OE_L is asserted (Low)
    - » D is the data output pin
  - Neither WE_L and OE_L are asserted?
    - » Chip is disconnected

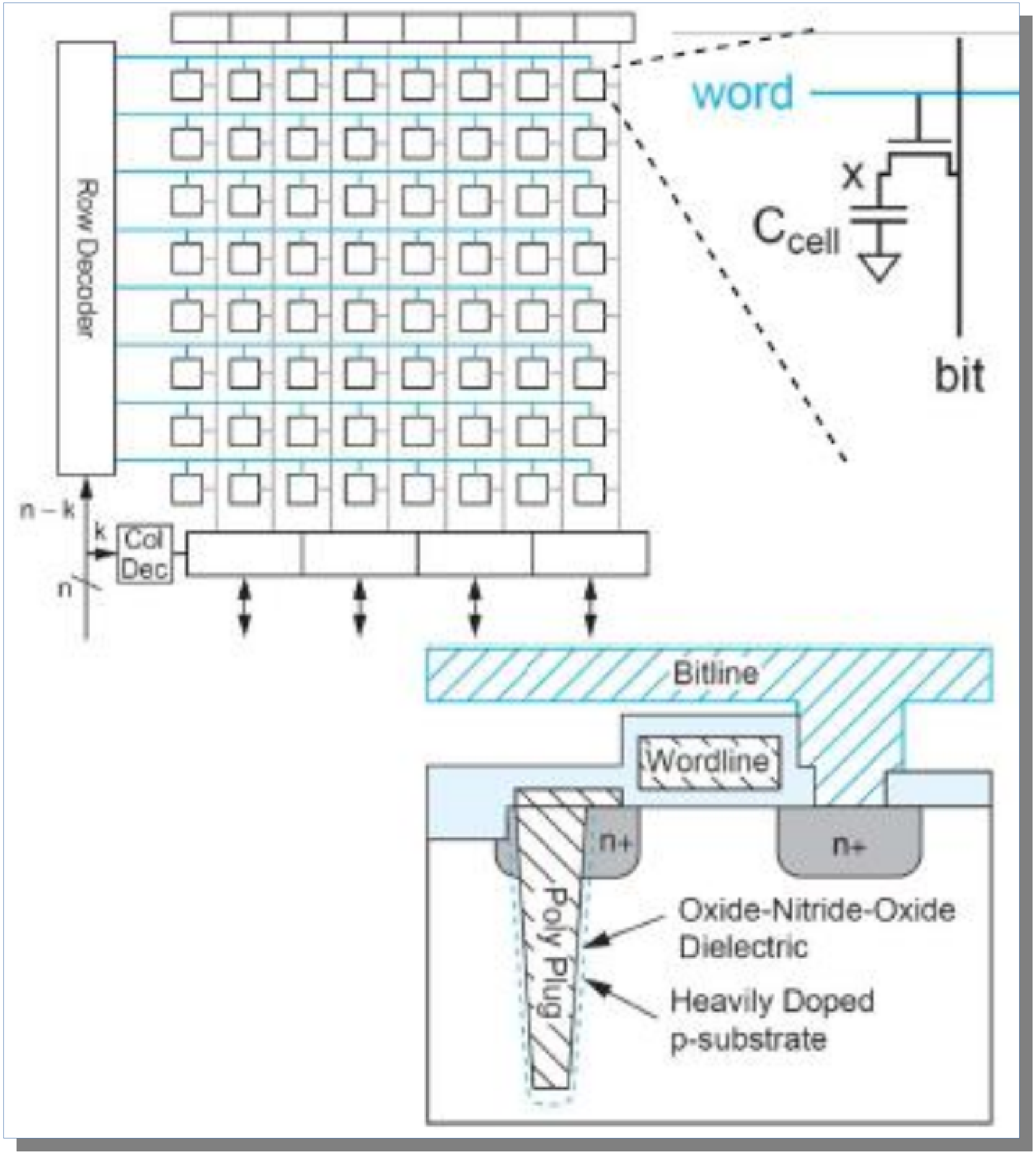
or chipSelect (CS) + WE

Main Memory

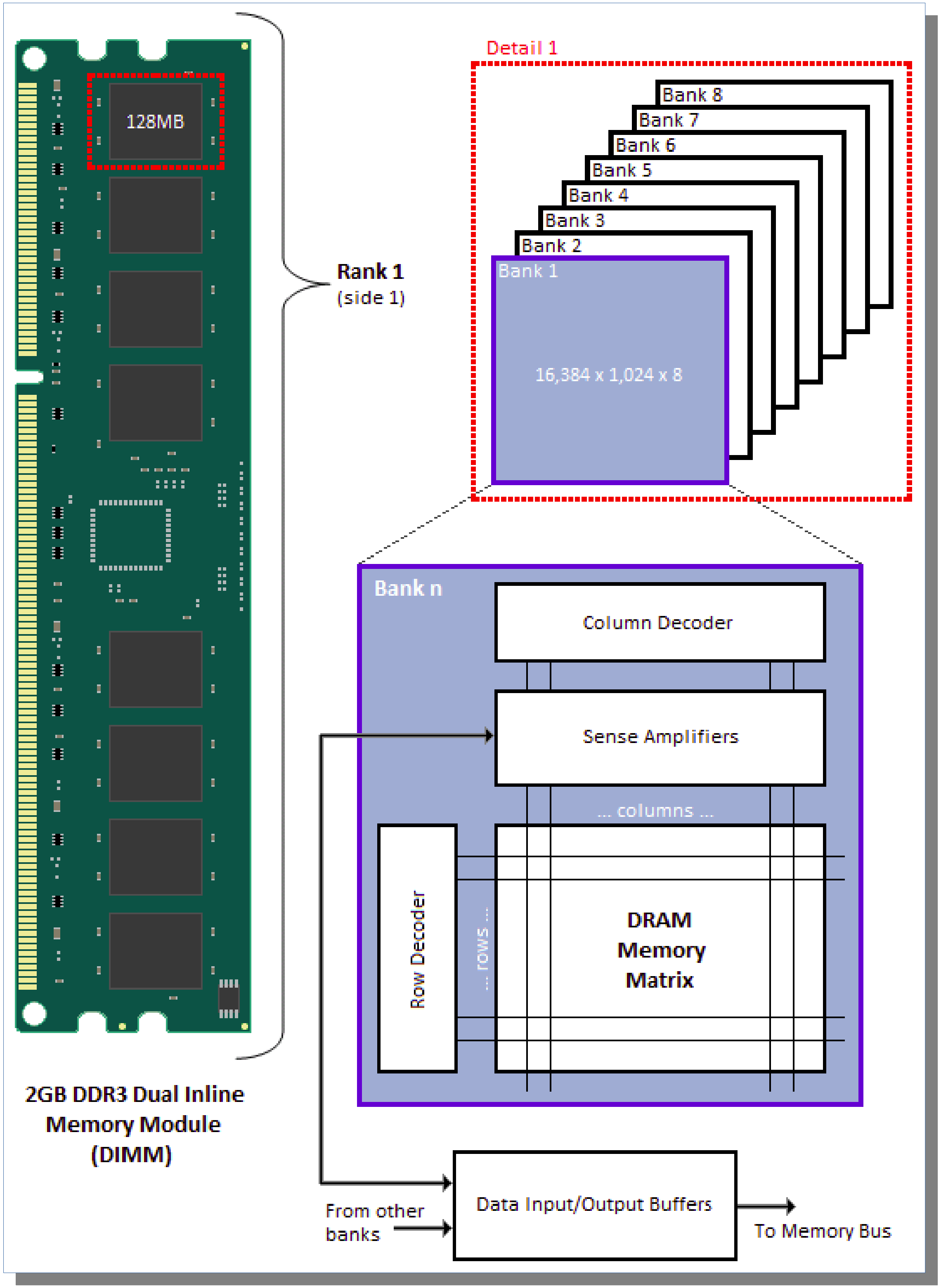
- Row buffer: the last row (say, 8 KB) read from a bank, acts like a cache
- Bank: a subset of a rank that is busy during one request
  - 4, 8 or 16 in one chip
- Rank: a collection of DRAM chips that work together to respond to a request and keep the data bus full



DRAM


Row Decoder
n – k
k
Col Dec
n
word
x
C_{cell}
bit
Bitline
Wordline
n+
n+
Poly Plug
Oxide-Nitride-Oxide Dielectric
Heavily Doped p-substrate

DRAM Hierarchy



DRAM Hierarchy

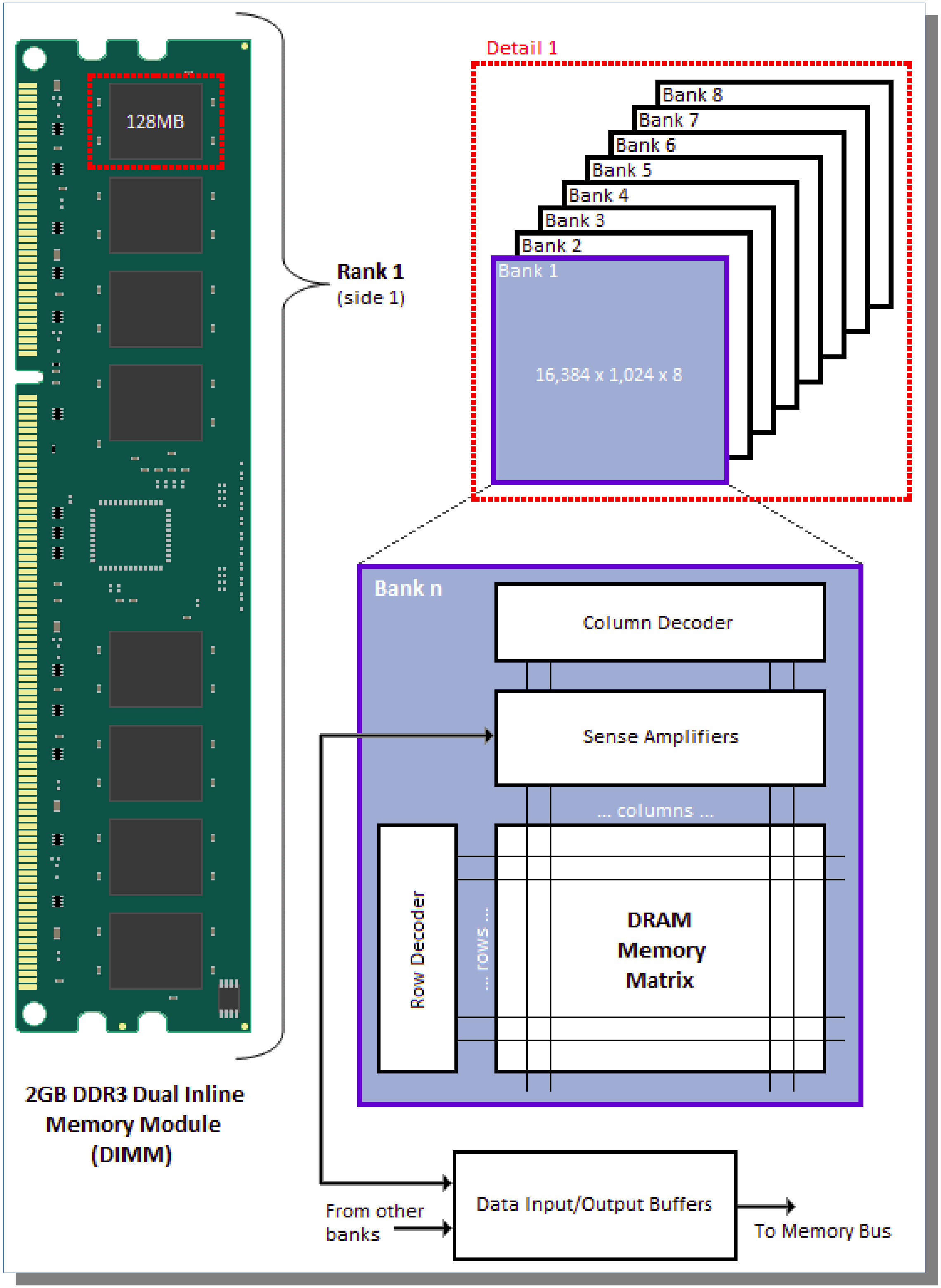
128MB =

16,384 rows/bank

x 1,024 columns addresses/row

x 1 byte/column address

x 8 stacked banks per IC.



2GB DDR3 Dual Inline Memory Module (DIMM)

http://www.anandtech.com/

DRAM Hierarchy

128MB =

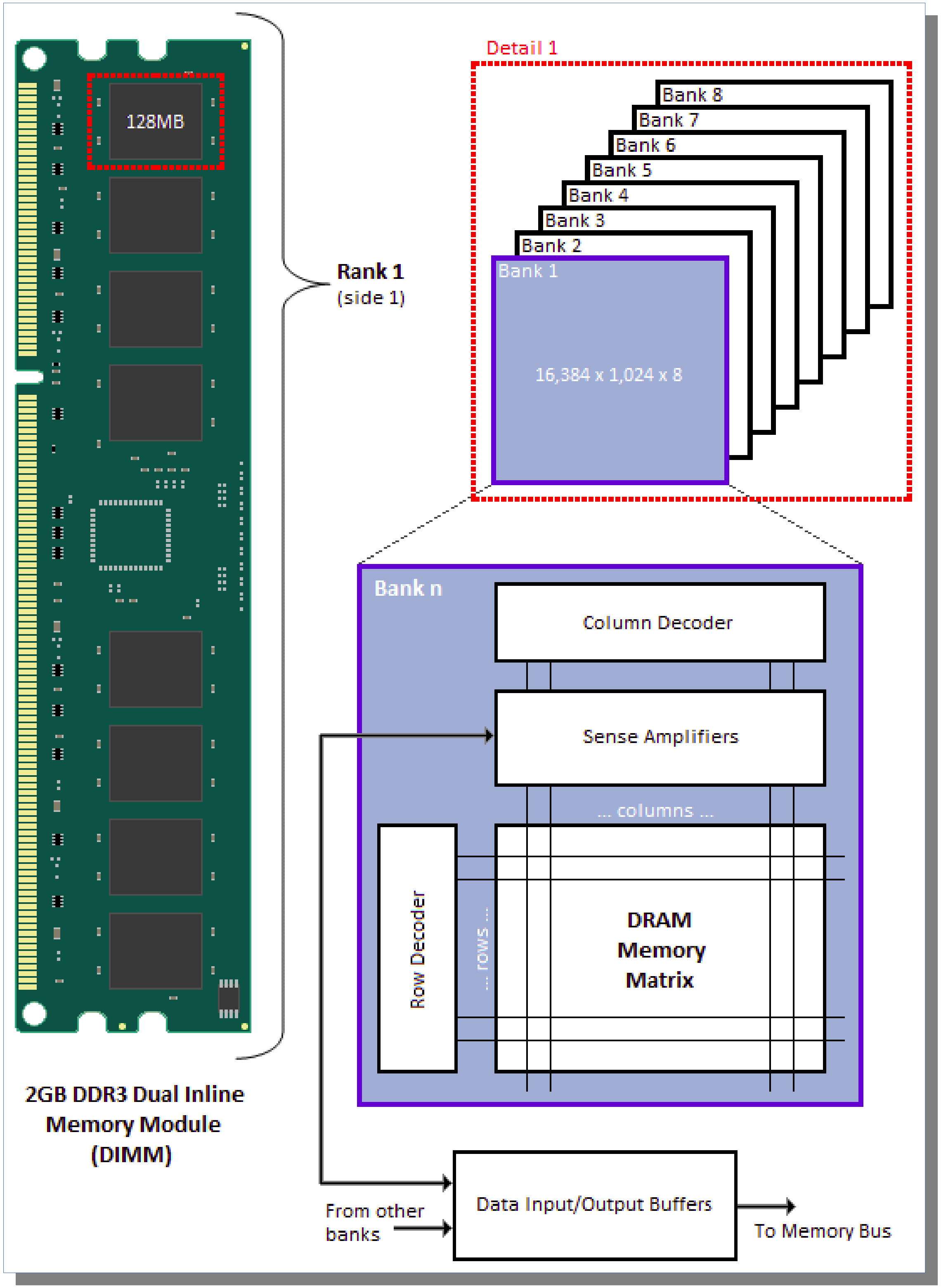
16,384 rows/bank

x 1,024 columns addresses/row

x 1 byte/column address

x 8 stacked banks per IC.

128MB x 8 ICs per rank = 1GB in Rank 1.



http://www.anandtech.com/

DRAM Hierarchy

128MB =

16,384 rows/bank

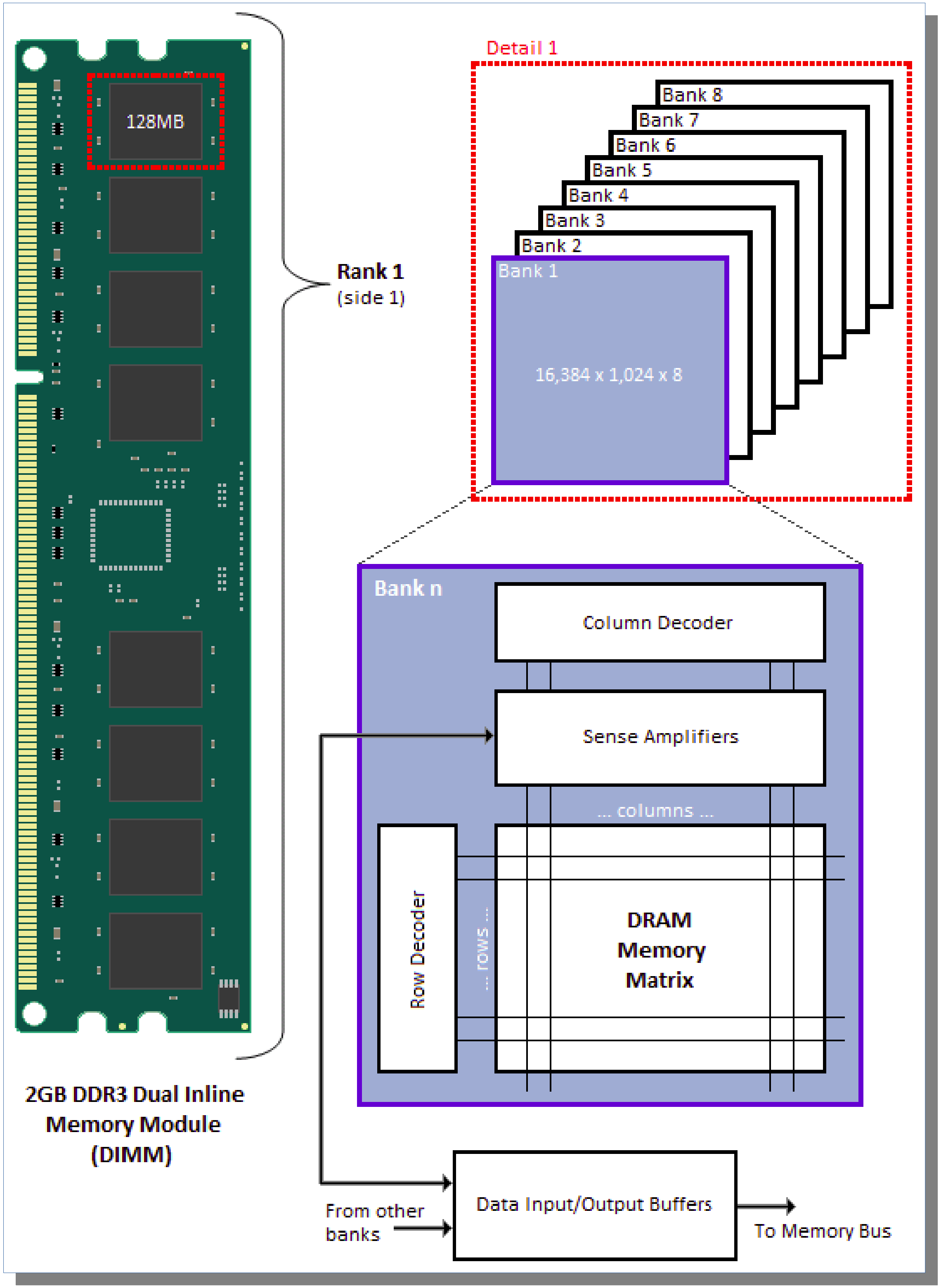
x 1,024 columns addresses/row

x 1 byte/column address

x 8 stacked banks per IC.

128MB x 8 ICs per rank = 1GB in Rank 1.

1GB (Rank 1) + 1GB (Rank 2) = 2GB per module.

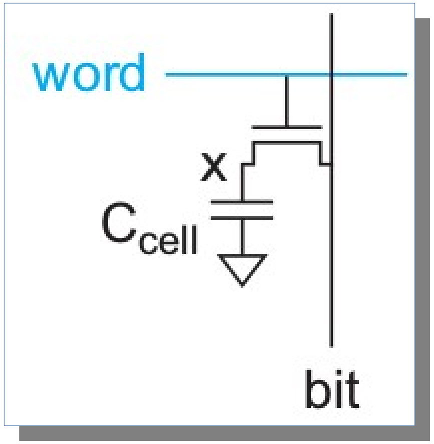


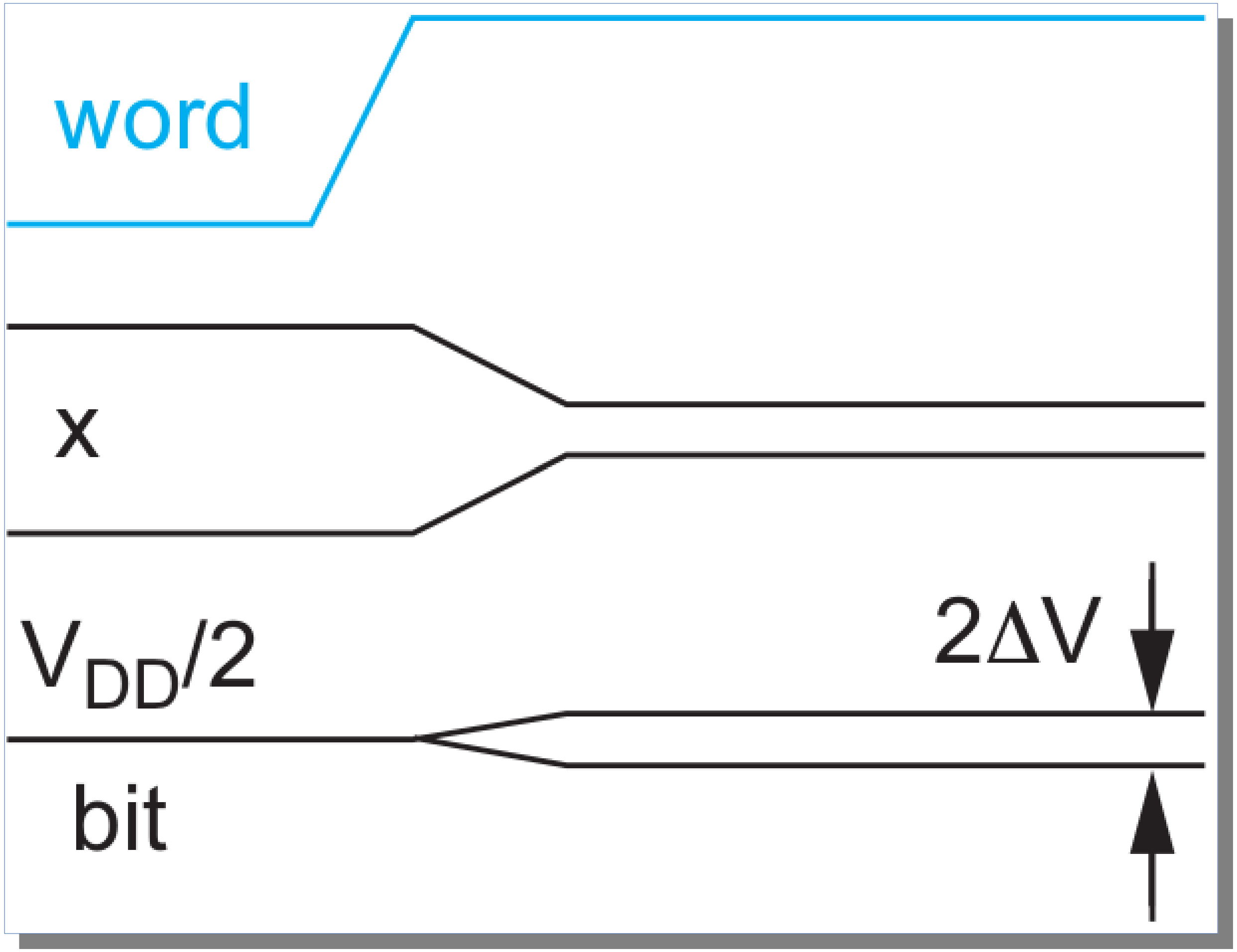
http://www.anandtech.com/

Memory Technology

- Bandwidth
- Access Time
  - Time between read request and when desired word arrives

DRAM Cell Read


word
x
C_{cell}
bit


word
x
$V_{DD}/2$
$2\Delta V$
bit