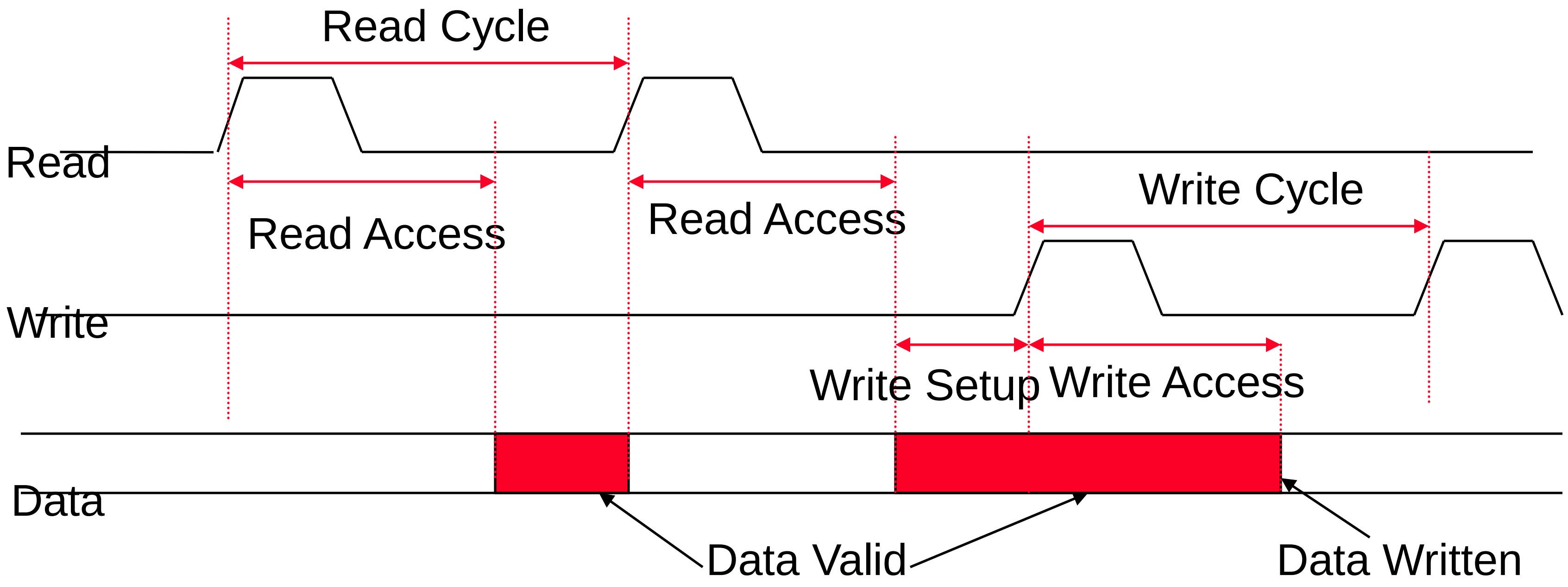
VLSI Digital Circuits

Memory Cell Designs

Link: https://uweb.engr.arizona.edu/~ece507/lecture16.ppt

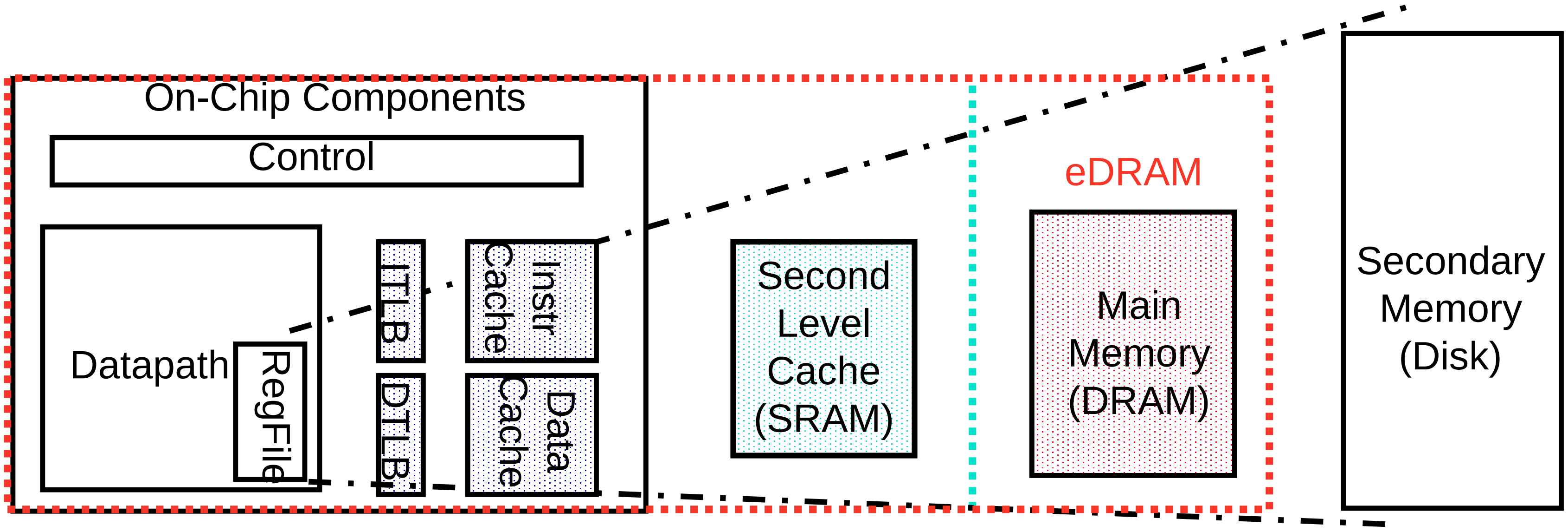
Memory Definitions

- Size – Kbytes, Mbytes, Gbytes, Tbytes
- Speed
  - Read Access – delay between read request and the data available
  - Write Access – delay between write request and the writing of the data into the memory
  - (Read or Write) Cycle - minimum time required between successive reads or writes



A Typical Memory Hierarchy

- By taking advantage of the principle of locality, we can
  - present the user with as much memory as is available in the cheapest technology
  - at the speed offered by the fastest technology.



Speed (ns):	.1's	1's	10's	100's	1,000's
Size (bytes):	100's	K's	10K's	M's	T's
Cost:	highest				lowest

More Memory Definitions

- Function – functionality, nature of the storage mechanism
  - static and dynamic; volatile and nonvolatile (NV); read only (ROM)
- Access pattern – random, serial, content addressable

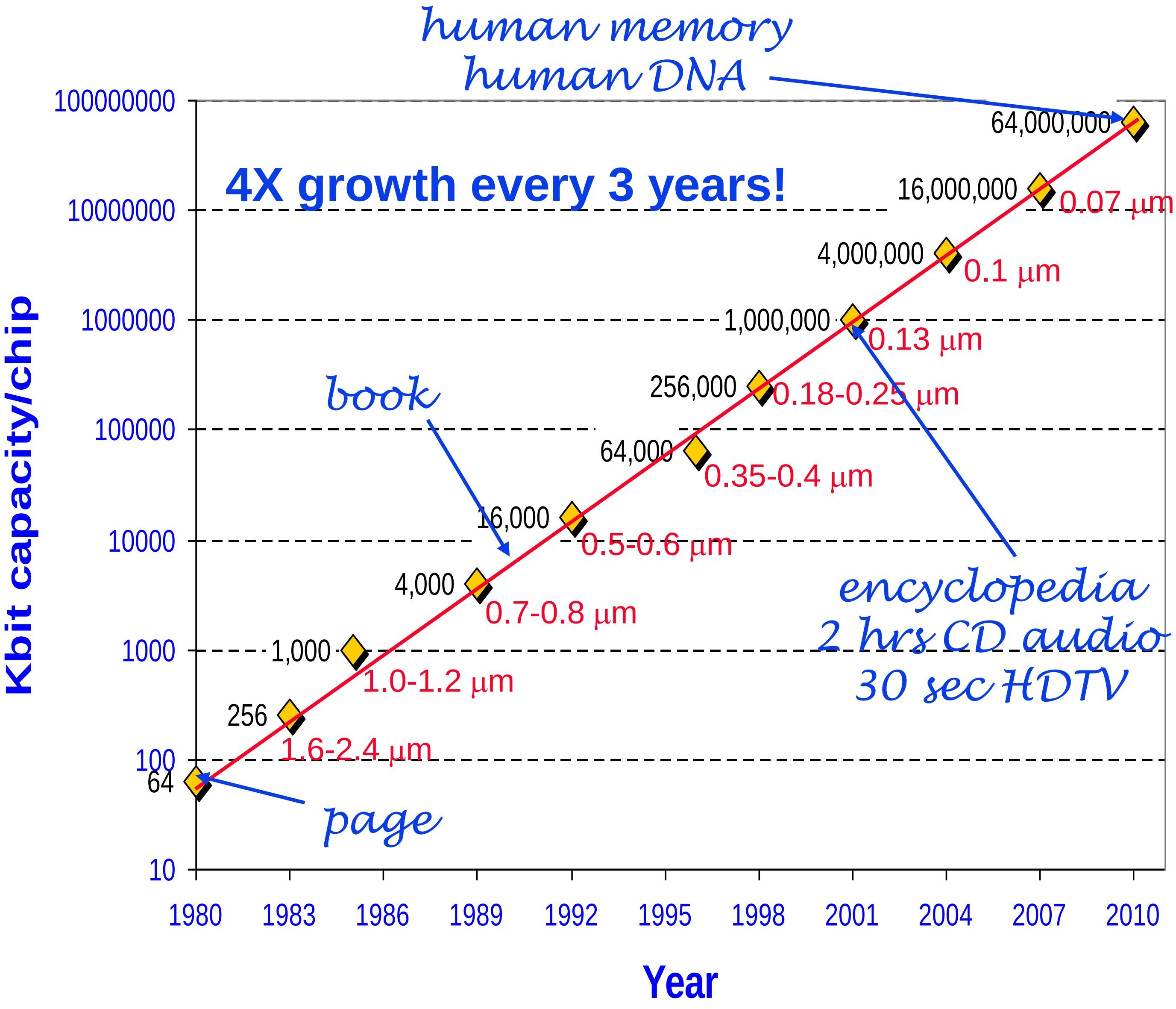
Read Write Memories (RWM)		**NVRWM**	**ROM**
Random Access	Non-Random Access	EPROM	Mask-prog. ROM
SRAM (cache, register file)	FIFO, LIFO	EEPROM	
DRAM (main memory)	Shift Register	FLASH	Electrically-prog. PROM
CAM			

- Input-output architecture – number of data input and output ports (multiported memories)
- Application – embedded, secondary, tertiary

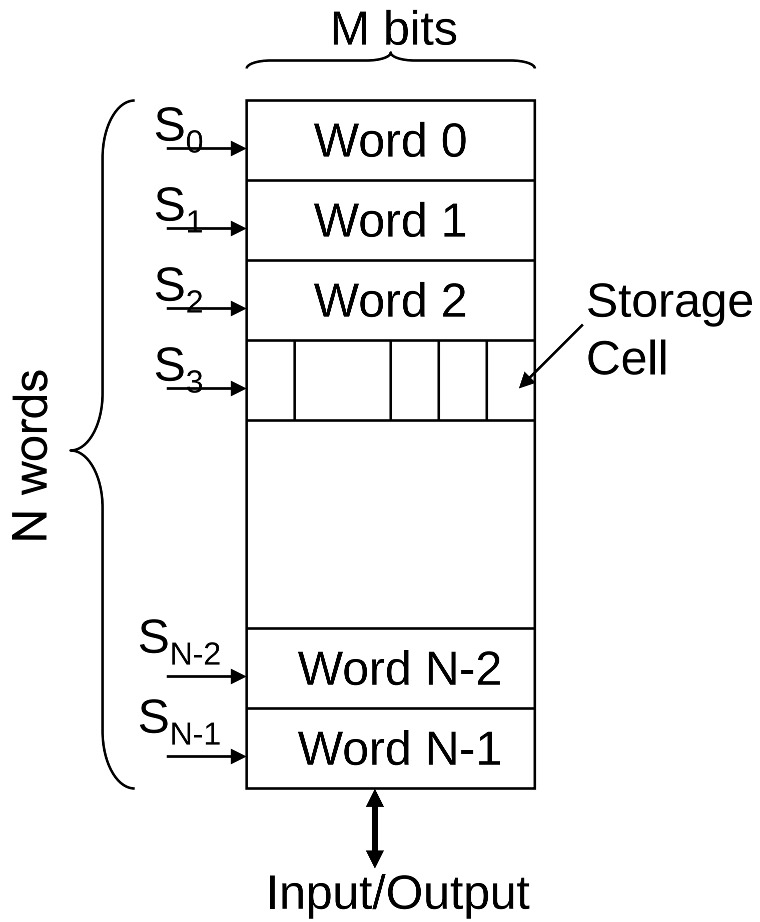
Random Access Read Write Memories (WRMs)

- SRAM – Static Random Access Memory
  - data is stored as long as supply is applied
  - large cells (6 fets/cell) – so fewer bits/chip
  - fast – so used where speed is important (e.g., caches)
  - differential outputs (output BL and !BL)
  - use sense amps for performance
  - compatible with CMOS technology
- DRAM - Dynamic Random Access Memory
  - periodic refresh required (every 1 to 4 ms) to compensate for the charge loss caused by leakage
  - small cells (1 to 3 fets/cell) – so more bits/chip
  - slower – so used for main memories
  - single ended output (output BL only)
  - need sense amps for correct operation
  - not typically compatible with CMOS technology

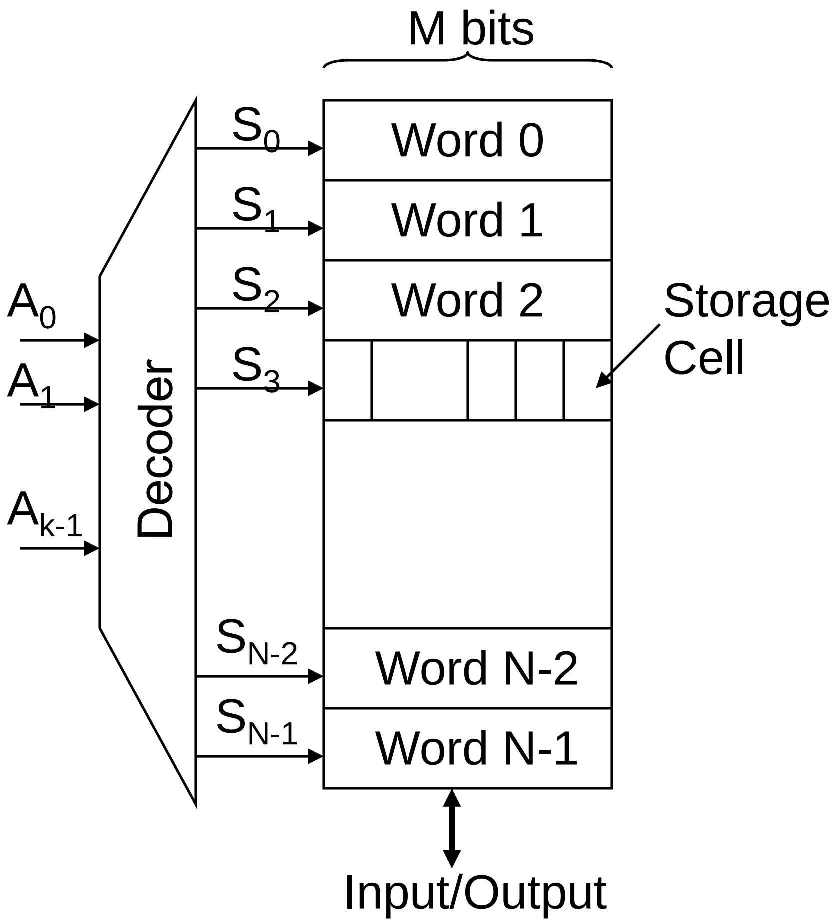
Evolution in DRAM Chip Capacity



1D Memory Architecture



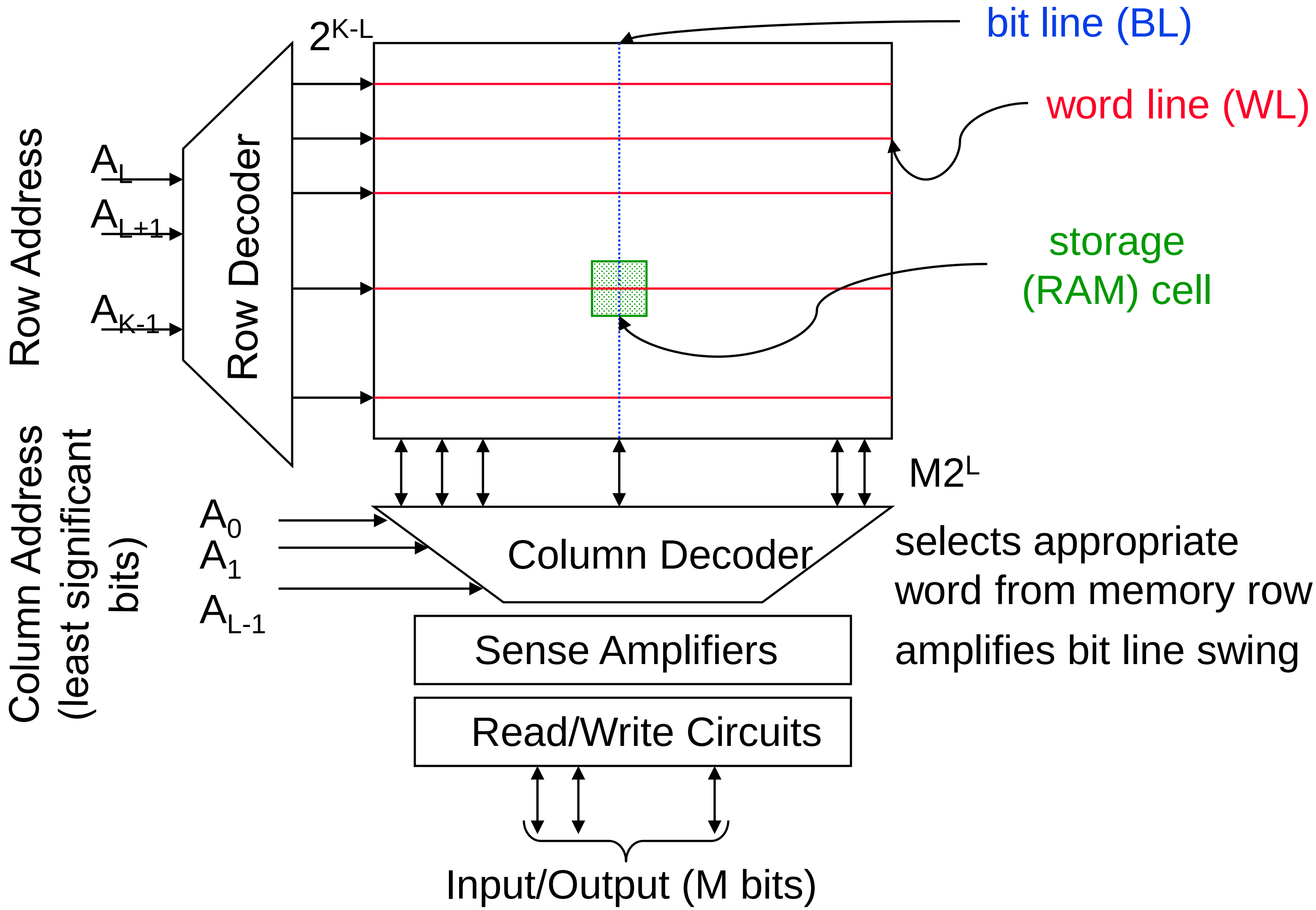
N words → N select signals



Decoder reduces # of inputs

$K = \log_2 N$

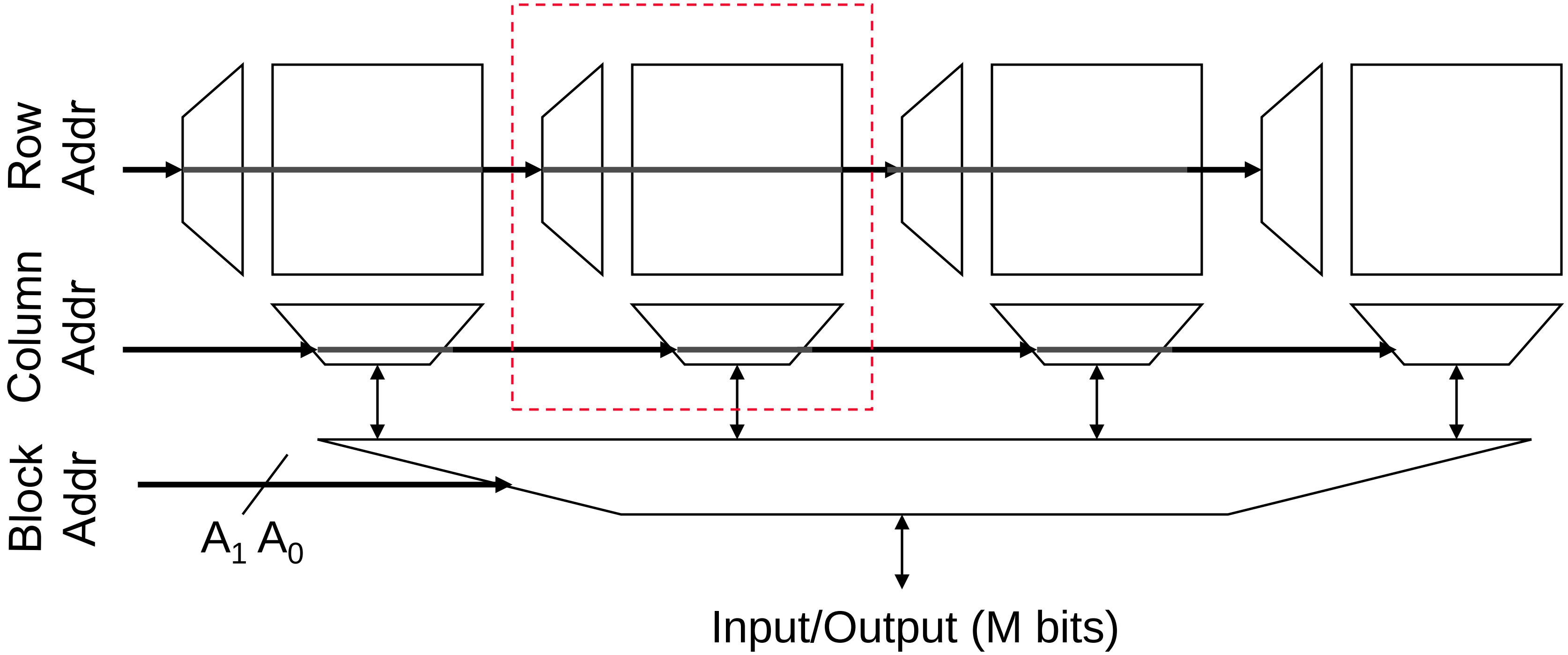
2D Memory Architecture



selects appropriate word from memory row

amplifies bit line swing

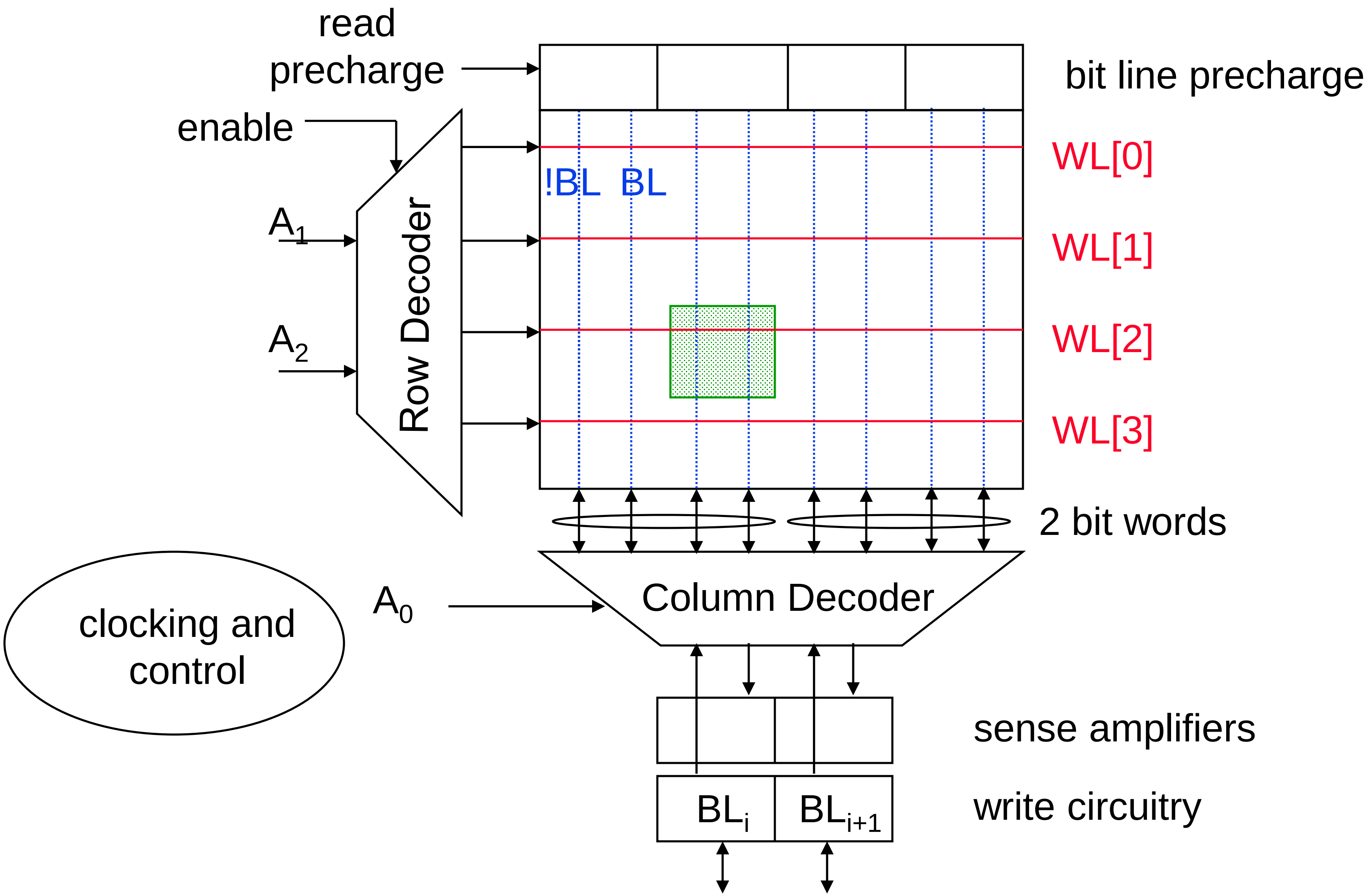
3D (or Banked) Memory Architecture



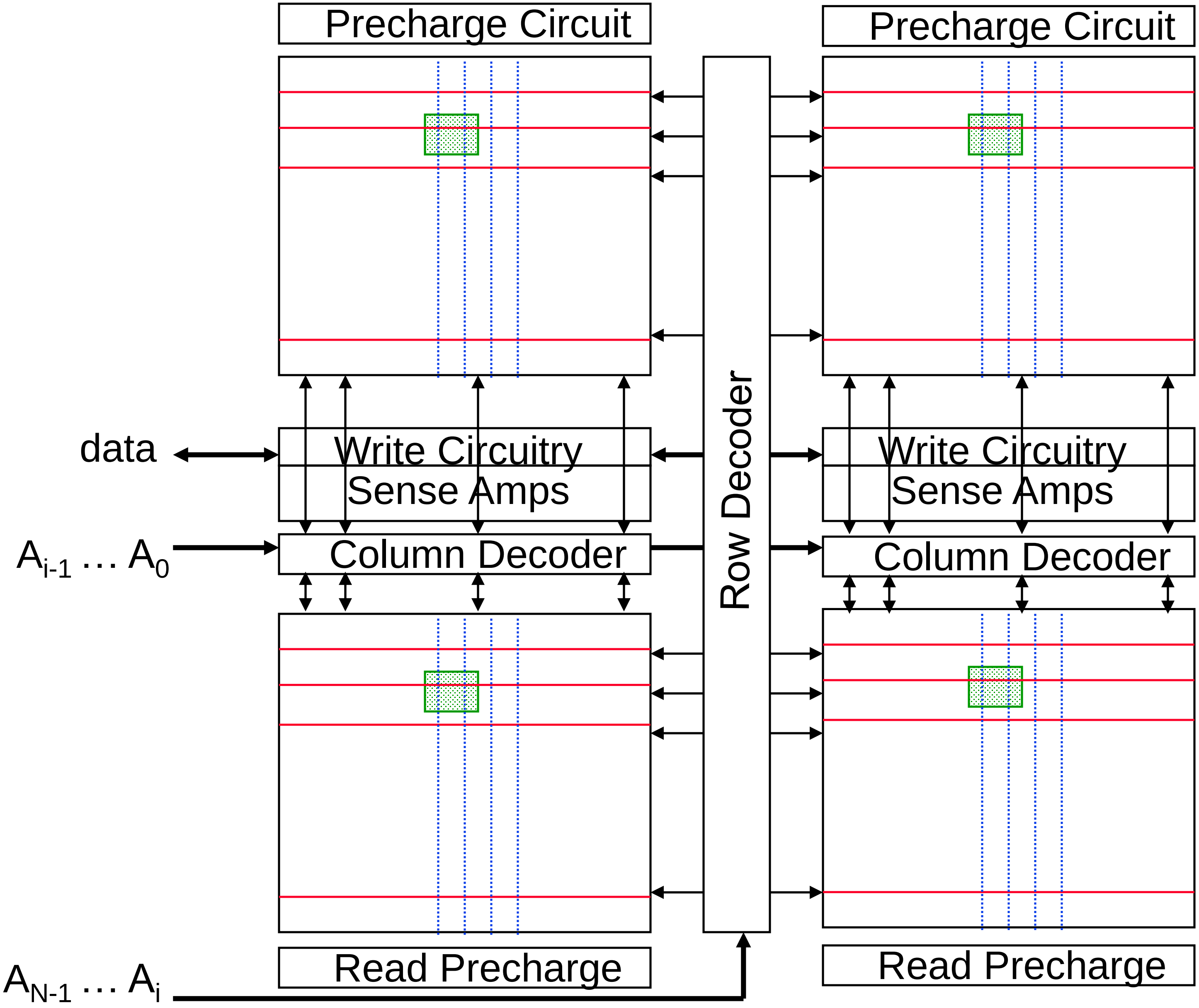
Advantages:

1. Shorter word and bit lines so faster access
2. Block addr activates only 1 block saving power

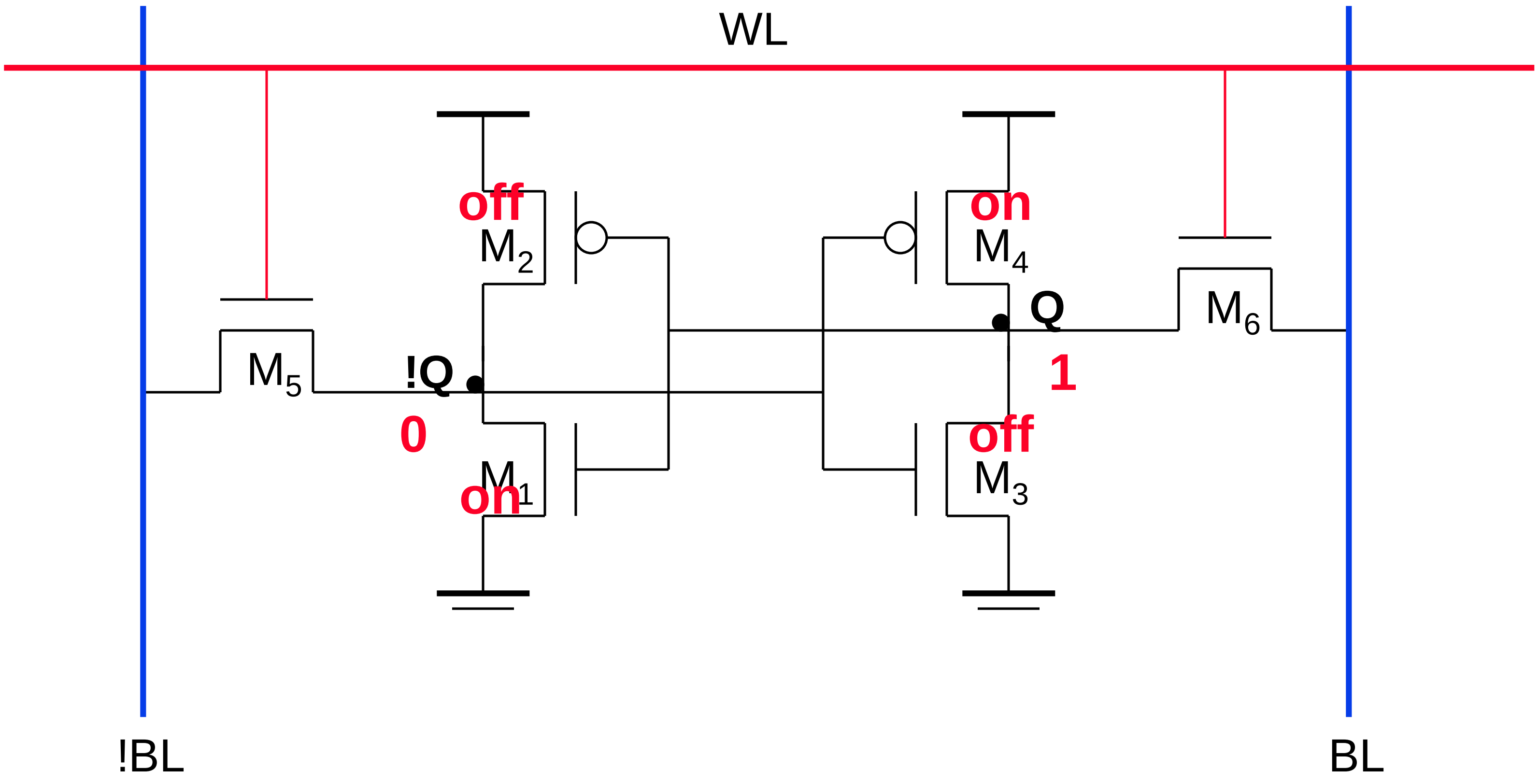
2D 4x4 SRAM Memory Bank



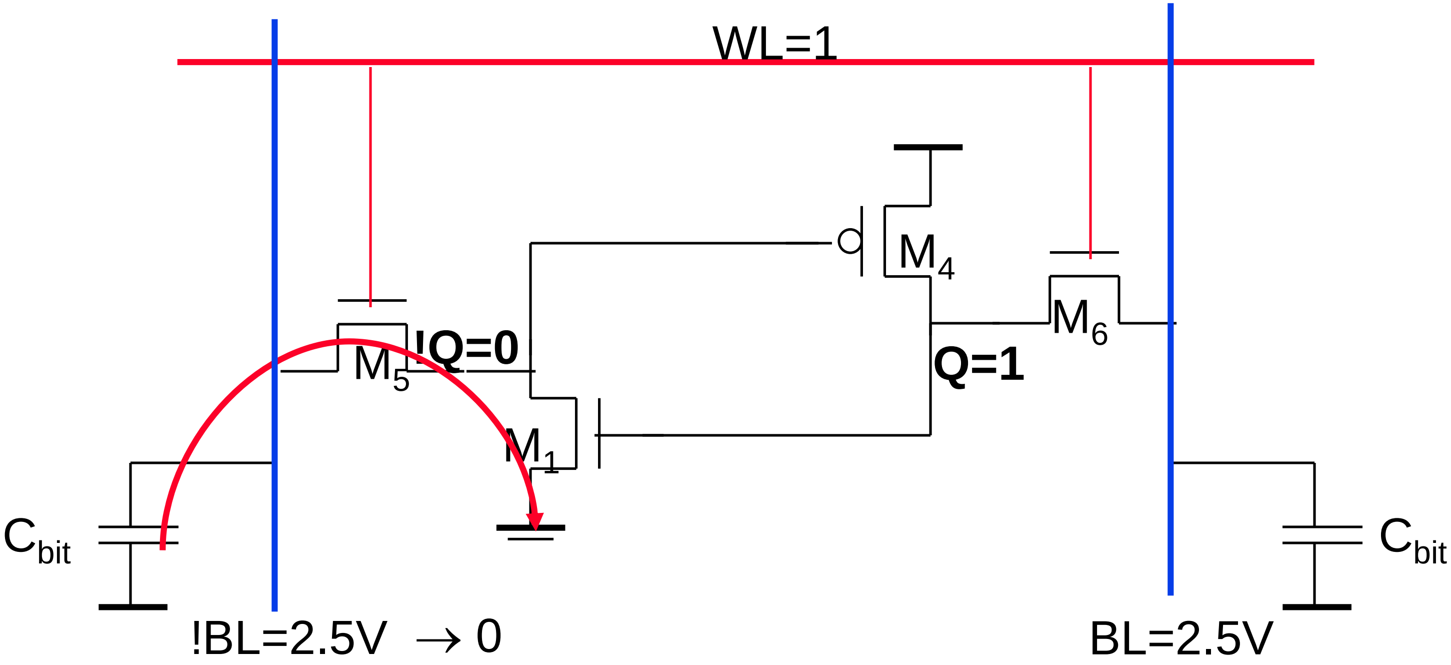
Quartering Gives Shorter WLs and BLs



6-Transistor SRAM Storage Cell



SRAM Cell Analysis (Read)



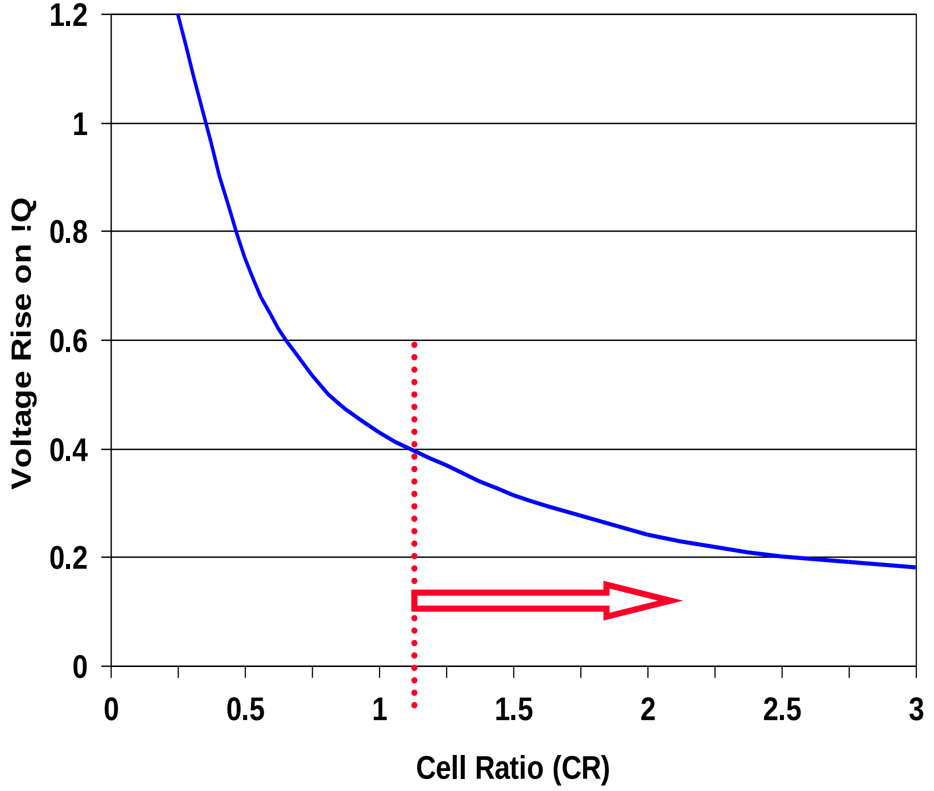
- Read-disturb (read-upset): must limit the voltage rise on !Q to prevent read-upsets from occurring while simultaneously maintaining acceptable circuit speed and area
  - M_1 must be stronger than M_5 when storing a 1 (as shown)
  - M_3 must be stronger than M_6 when storing a 0

Read Voltage Ratios

$$\Delta V_{!Q} = [V_{DSATn} + CR(V_{DD} - V_{Tn}) - \sqrt{(V_{DSATn}{}^2(1+CR) + CR^2(V_{DD} - V_{Tn})^2)}]/CR$$

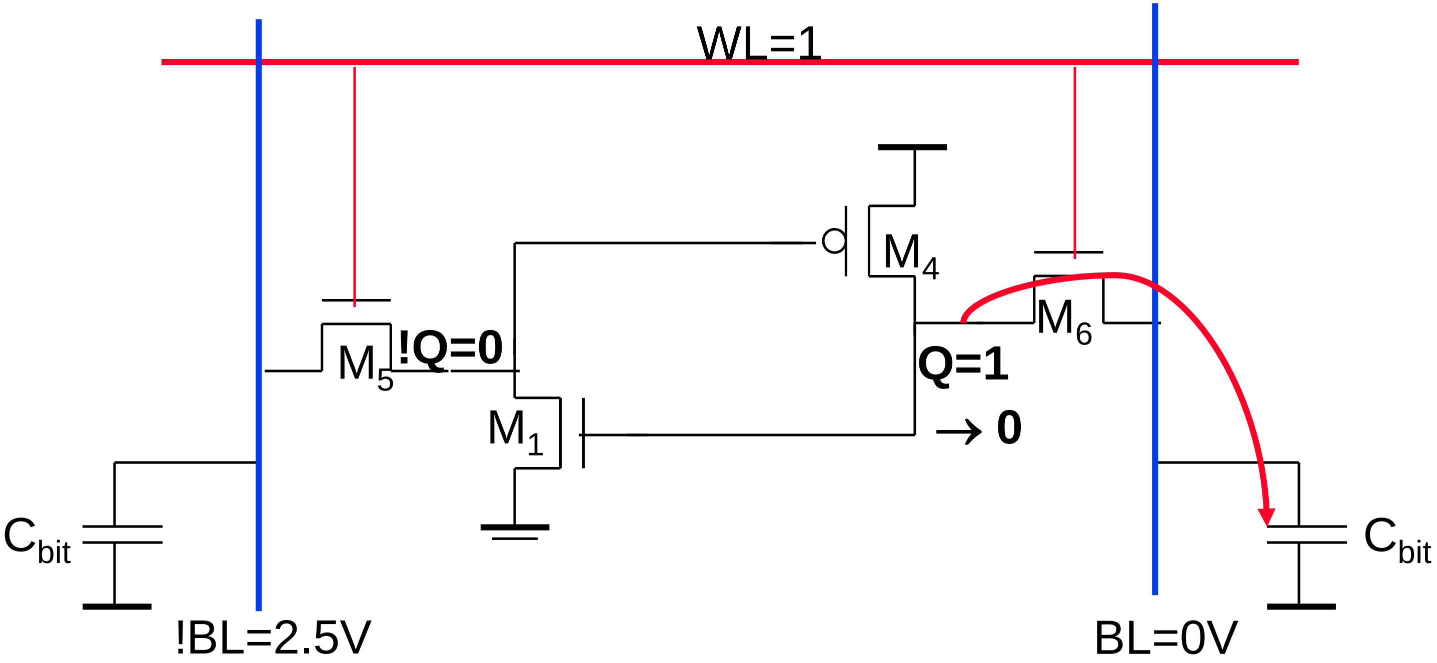
where CR is the Cell Ratio = $(W_1/L_1)/(W_5/L_5)$ ≥ **1.2**

V_{DD} = 2.5V
V_{Tn} = 0.4V



- Keep cell size minimal while maintaining read stability
  - Make M_1 minimum size and increase the L of M_5 (to make it weaker)
    - increases load on WL
  - Make M_5 minimum size and increase the W of M_1 (to make it stronger)
- Similar constraints on $(W_3/L_3)/(W_6/L_6)$ when storing a 0

SRAM Cell Analysis (Write)



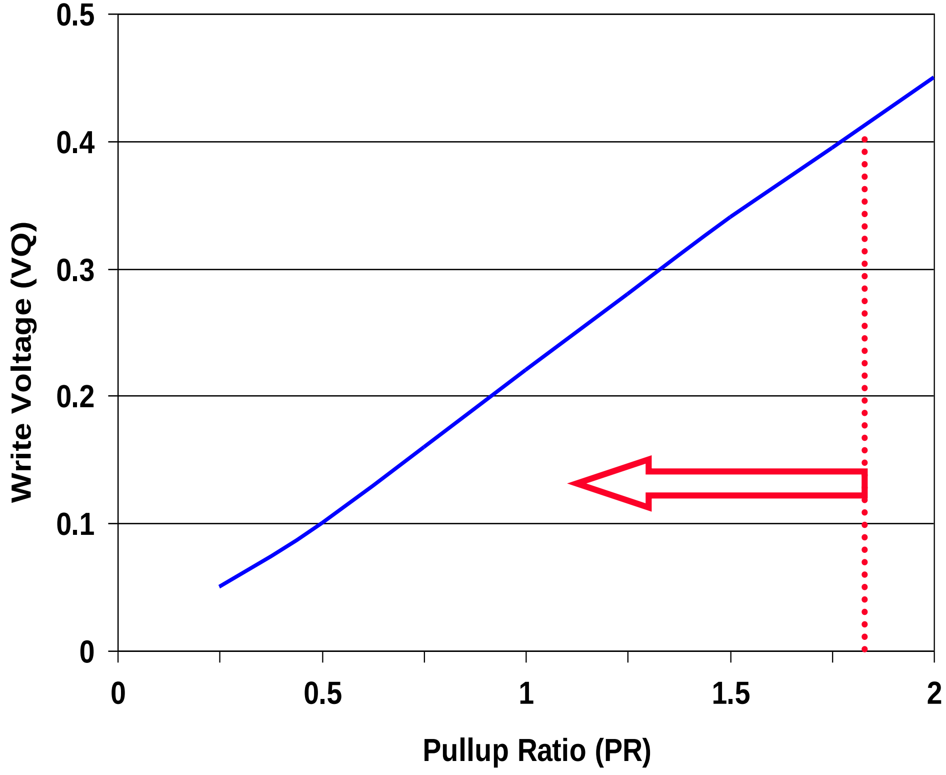
- The !Q side of the cell cannot be pulled high enough to ensure writing of 0 (because M_1 is on and sized to protect against read upset). So, the new value of the cell has to be written through M_6.
  - M_6 must be able to overpower M_4 when storing a 1 and writing a 0
  - M_5 must be able to overpower M_2 when storing a 0 and writing a 1

Write Voltage Ratios

$$V_Q = (V_{DD} - V_{Tn}) - \sqrt{((V_{DD} - V_{Tn})^2 - 2(\mu_p/\mu_n)(PR)((V_{DD} - |V_{Tp}|)V_{DSATp} - V_{DSATp}^2/2))}$$

where PR is the Pull-up Ratio = $(W_4/L_4)/(W_6/L_6)$ **$\leq$ 1.8**

$V_{DD} = 2.5V$
$|V_{Tp}| = 0.4V$
$\mu_p/\mu_n = 0.5$



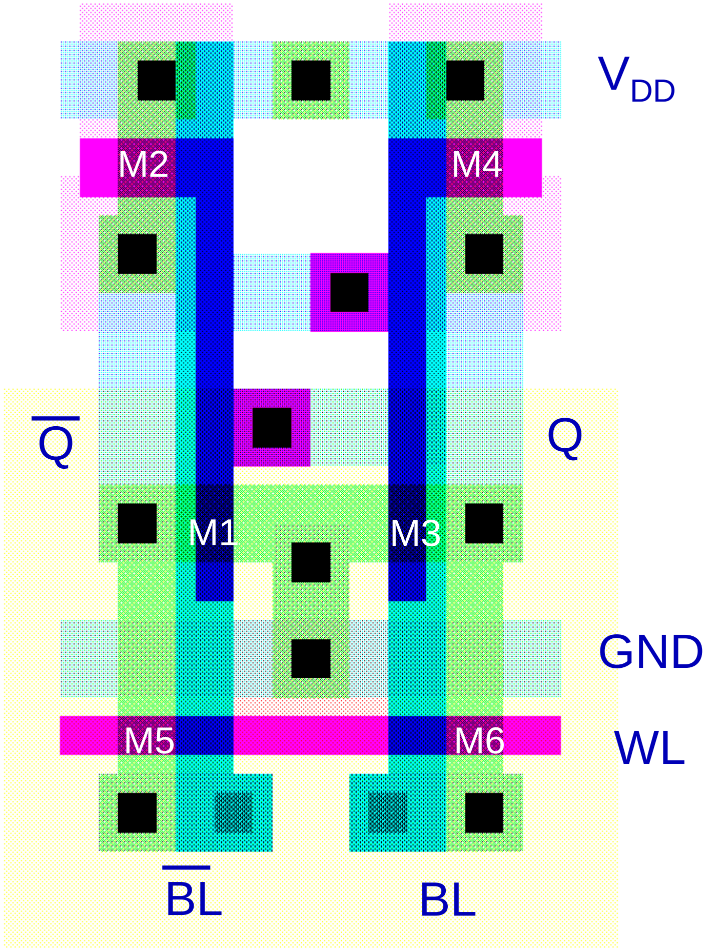
- Keep cell size minimal while allowing writes
  - Make M_4 and M_6 minimum size
- Be sure to consider worst case process corners (strong PMOS, weak NMOS, high V_{DD})

Cell Sizing and Performance

- Keeping cell size minimal is critical for large SRAMs
  - Minimum sized pull down fets (M_1 and M_3)
    - Requires longer than minimum channel length, L, pass transistors (M_5 and M_6) to ensure proper CR
    - But up-sizing L of the pass transistors increases capacitive load on the word lines *and* limits the current discharged on the bit lines both of which can adversely affect the speed of the read cycle
  - Minimum width and length pass transistors
    - Boost the width of the pull downs (M_1 and M_3)
    - Reduces the loading on the word lines and increases the storage capacitance in the cell – both are good! – but cell size may be slightly larger

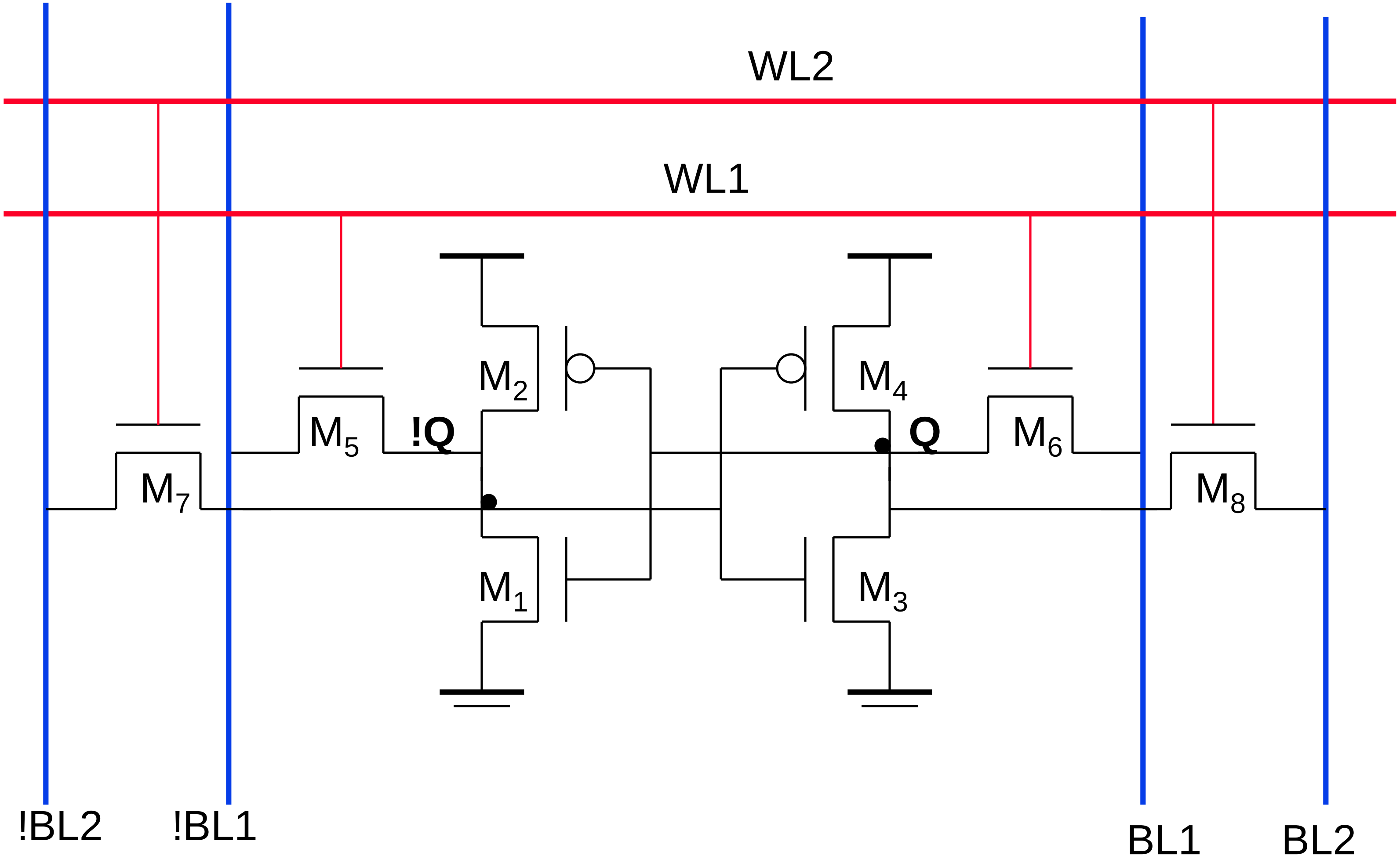
- Performance is determined by the read operation
  - To accelerate the read time, SRAMs use sense amplifiers (so that the bit line doesn’t have to make a full swing)

6-T SRAM Layout



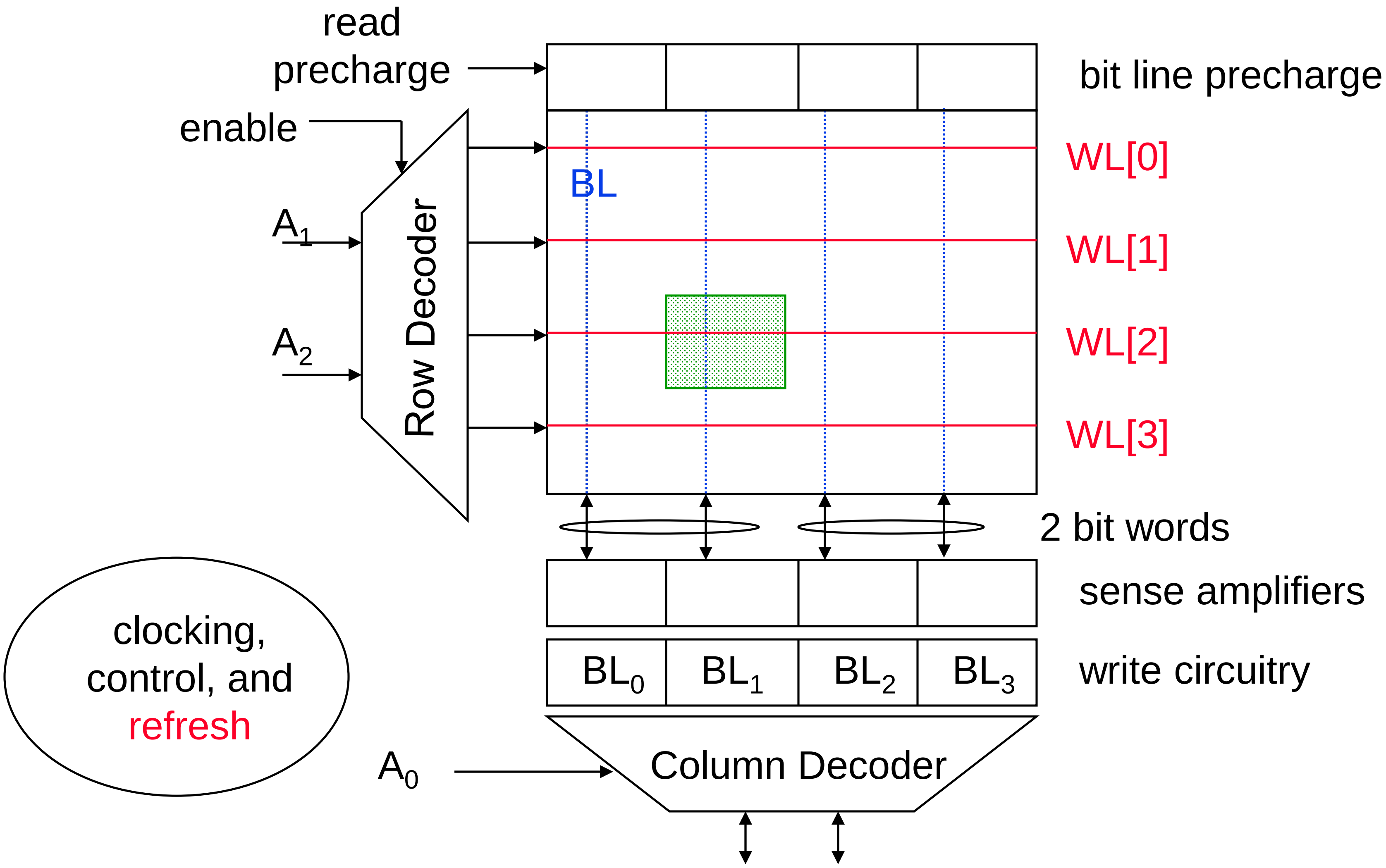
- Simple and reliable, but big
  - signal routing and connections to two bit lines, a word line, and both supply rails
- Area is dominated by the wiring and contacts (11.5 of them)
- Other alternatives to the 6-T cell include the resistive load 4-T cell and the TFT cell neither of which are available in a standard CMOS logic process

Multiple Read/Write Port Storage Cell

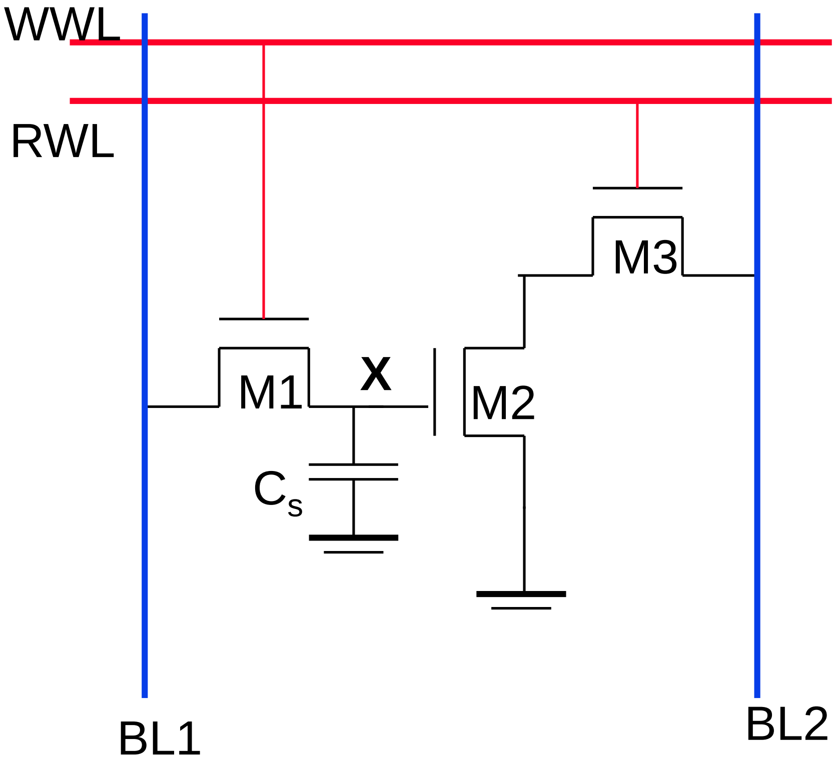


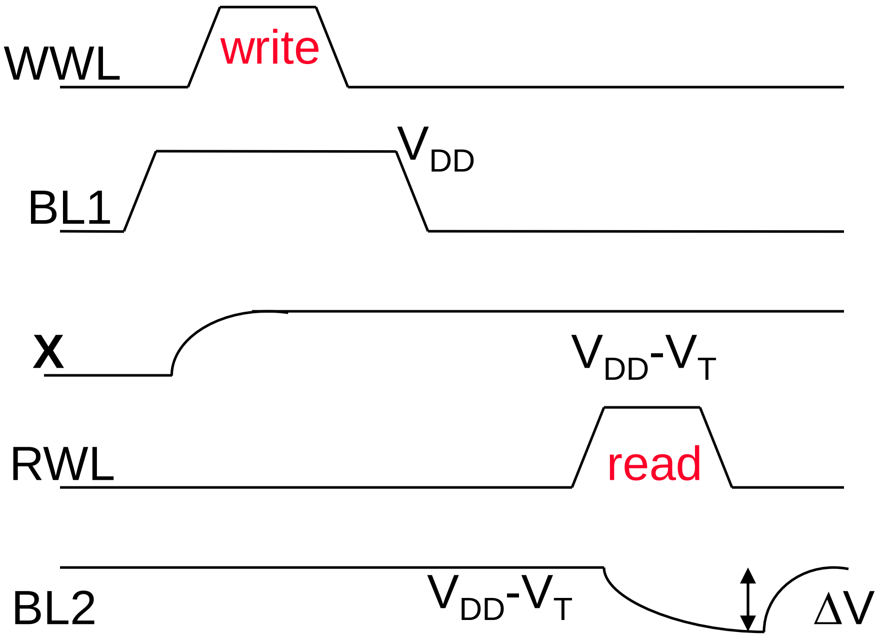
- To avoid read upset, the widths of M_1 and M_3 will have to be sized up by a factor equal to the number of simultaneously open read ports

2D 4x4 DRAM Memory



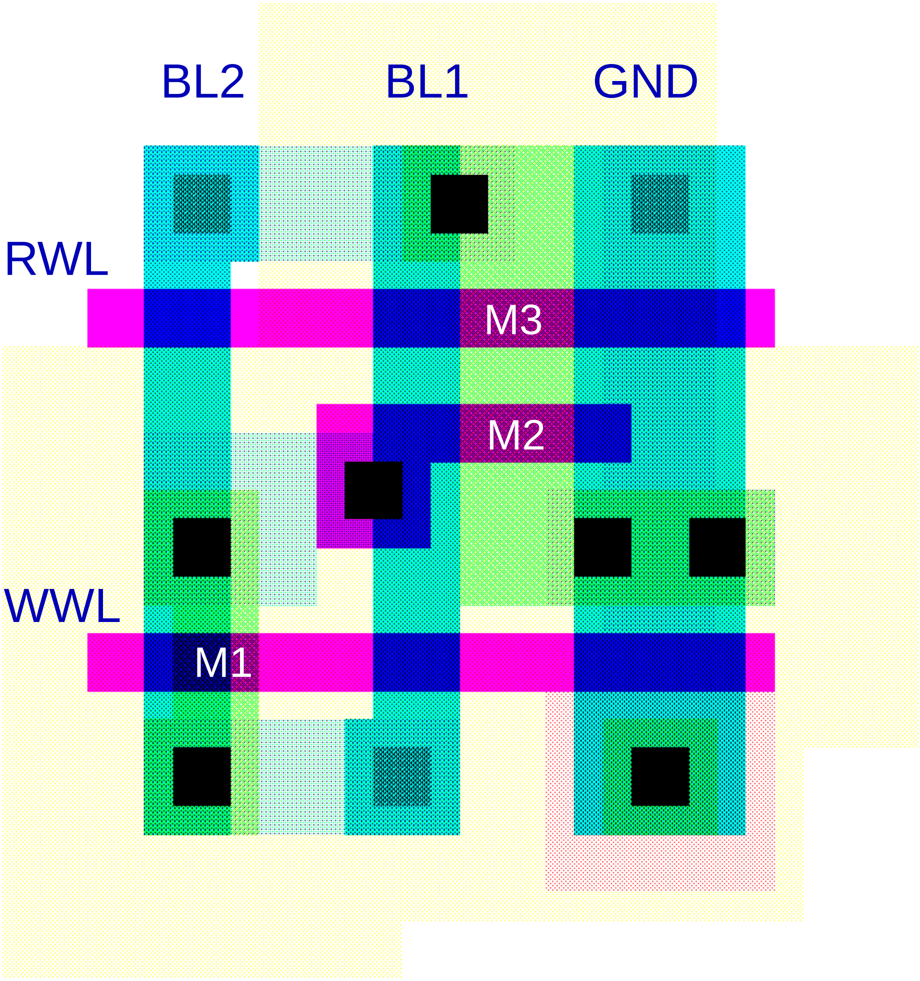
3-Transistor DRAM Cell





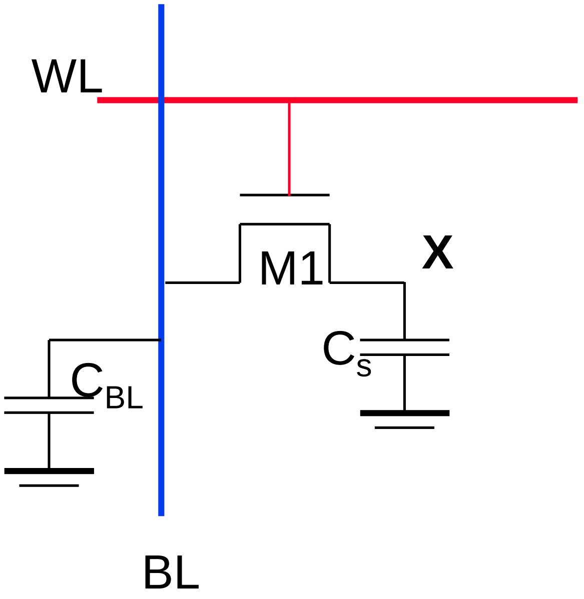
- Write: C_s is charged (or discharged) by asserting WWL and BL1
  - Value stored at node X when writing a 1 is V_{WWL} - V_{Tn}
- Read: C_s is “sensed” by asserting RWL and observing BL2
  - Read is non-destructive and inverting

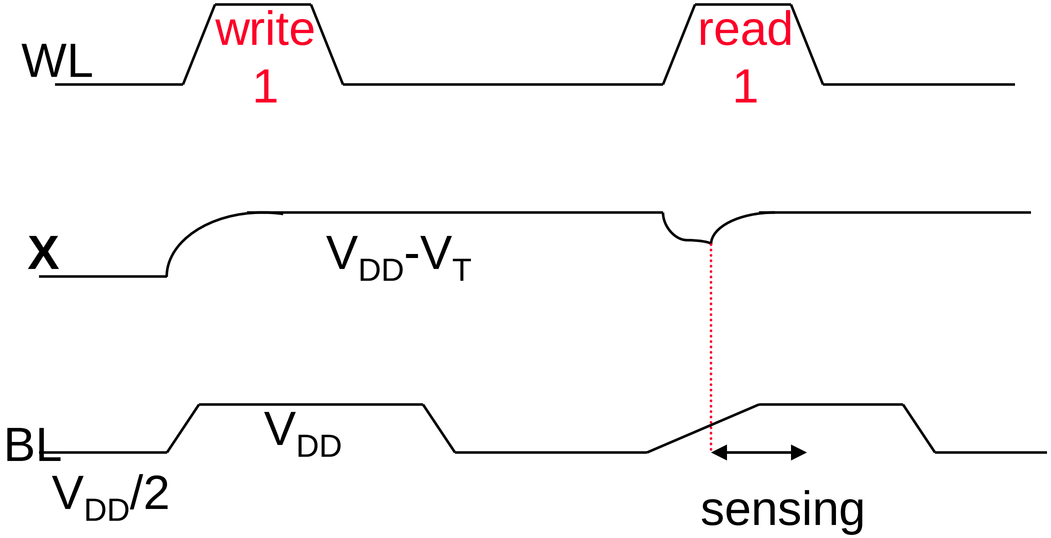
3-T DRAM Layout



- Total cell area is 576 λ^2 (compared to 1,092 λ^2 for the 6-T SRAM cell)
- No special processing steps are needed (so compatible with logic CMOS process)
- Can use bootstrapping (raise V_{WWL} to a value higher than V_{DD}) to eliminate threshold drop when storing a “1”

1-Transistor DRAM Cell





- Write: C_s is charged (or discharged) by asserting WL and BL
- Read: Charge redistribution occurs between C_{BL} and C_s
    - Read is destructive, so must refresh after read

1-T DRAM Cell Observations

- Cell is single ended (complicates the design of the sense amp)
- Cell requires a sense amp for each bit line due to charge redistribution based read
  - BL's precharged to $V_{DD}/2$ (not V_{DD} as with SRAM design)
  - all previous designs used SAs for speed, not functionality
- Cell read is destructive; refresh must follow to restore data
- Cell requires an extra capacitor (C_S) that must be explicitly included in the design
  - not compatible with logic CMOS process
- A threshold voltage is lost when writing a 1 (can be circumvented by bootstrapping the word lines to a higher value than V_{DD})

Peripheral Memory Circuitry

- Row and column decoders
- Read bit line precharge logic
- Sense amplifiers
- Read/write circuitry
- Timing and control

- Speed
- Power consumption
- Area – pitch matching

Row Decoders

- Collection of 2^M complex logic gates organized in a regular, dense fashion

- (N)AND decoder for 8 address bits

$$WL(0) = !A_7 \,\&\, !A_6 \,\&\, !A_5 \,\&\, !A_4 \,\&\, !A_3 \,\&\, !A_2 \,\&\, !A_1 \,\&\, !A_0$$

$$\ldots$$

$$WL(255) = A_7 \,\&\, A_6 \,\&\, A_5 \,\&\, A_4 \,\&\, A_3 \,\&\, A_2 \,\&\, A_1 \,\&\, A_0$$

- NOR decoder for 8 address bits

$$WL(0) = !(A_7 \mid A_6 \mid A_5 \mid A_4 \mid A_3 \mid A_2 \mid A_1 \mid A_0)$$

$$\ldots$$

$$WL(255) = !(!A_7 \mid !A_6 \mid !A_5 \mid !A_4 \mid !A_3 \mid !A_2 \mid !A_1 \mid !A_0)$$

- Goals: Pitch matched, fast, low power

Implementing a Wide NOR Function

- Single stage 8x256 bit decoder (as in Lecture 22)
  - One 8 input NOR gate per row x 256 rows = 256 x (8+8) = 4,096
  - Pitch match and speed/power issues

- Decompose logic into multiple levels

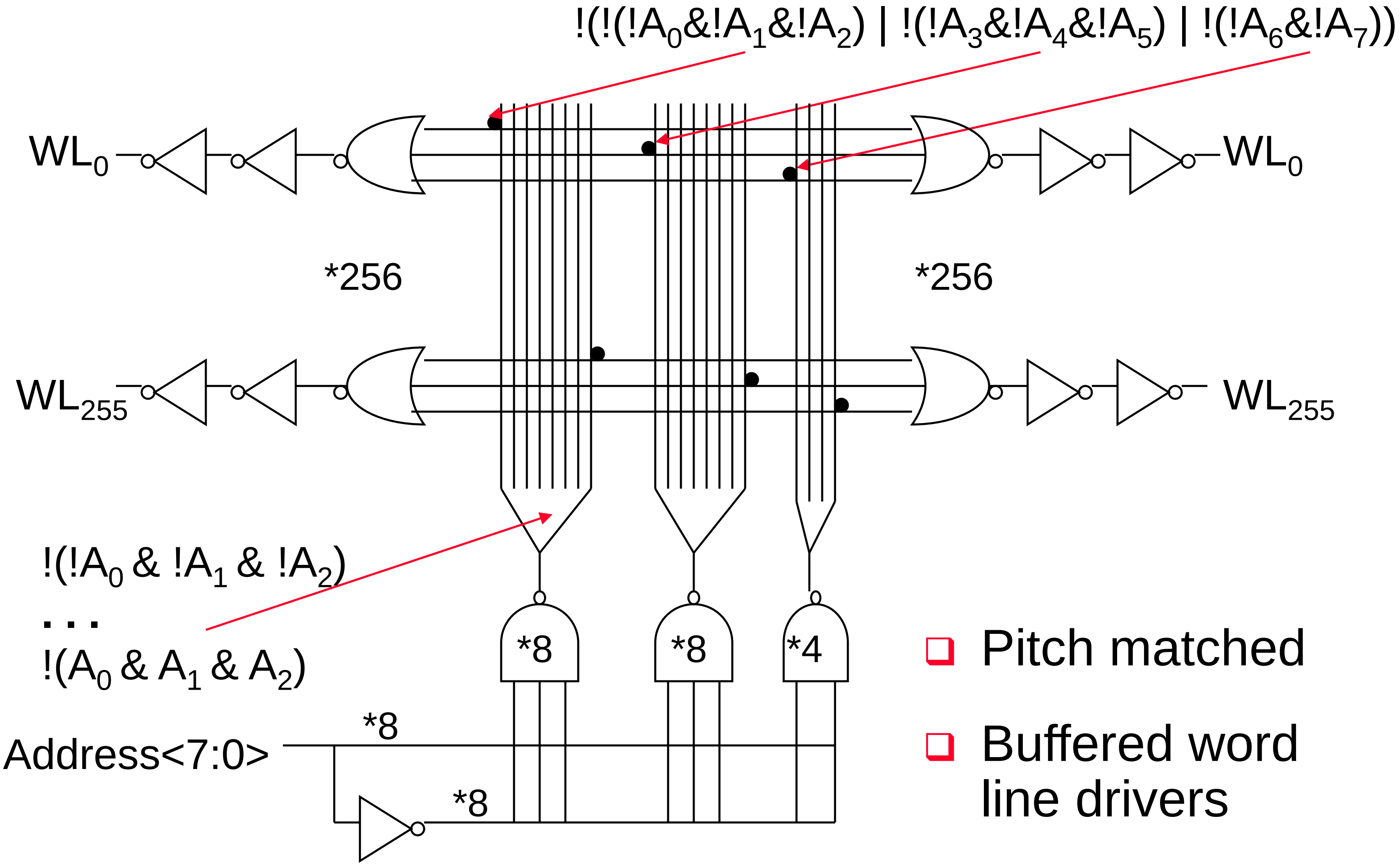
$$!WL(0) = !(!(A_7 \mid A_6) \,\&\, !(A_5 \mid A_4) \,\&\, !(A_3 \mid A_2) \,\&\, !(A_1 \mid A_0))$$

  - First level is the predecoder (for each pair of address bits, form $A_i|A_{i-1}$, $A_i|!A_{i-1}$, $!A_i|A_{i-1}$, and $!A_i|!A_{i-1}$)
  - Second level is the word line driver

- Predecoders reduce the number of transistors required
  - Four sets of four 2-bit NOR predecoders = 4 x 4 x (2+2) = 64
  - 256 word line drivers, each a four input NAND – 256 x (4+4) = 2,048
    - 4,096 vs 2,112 = almost a 50% savings

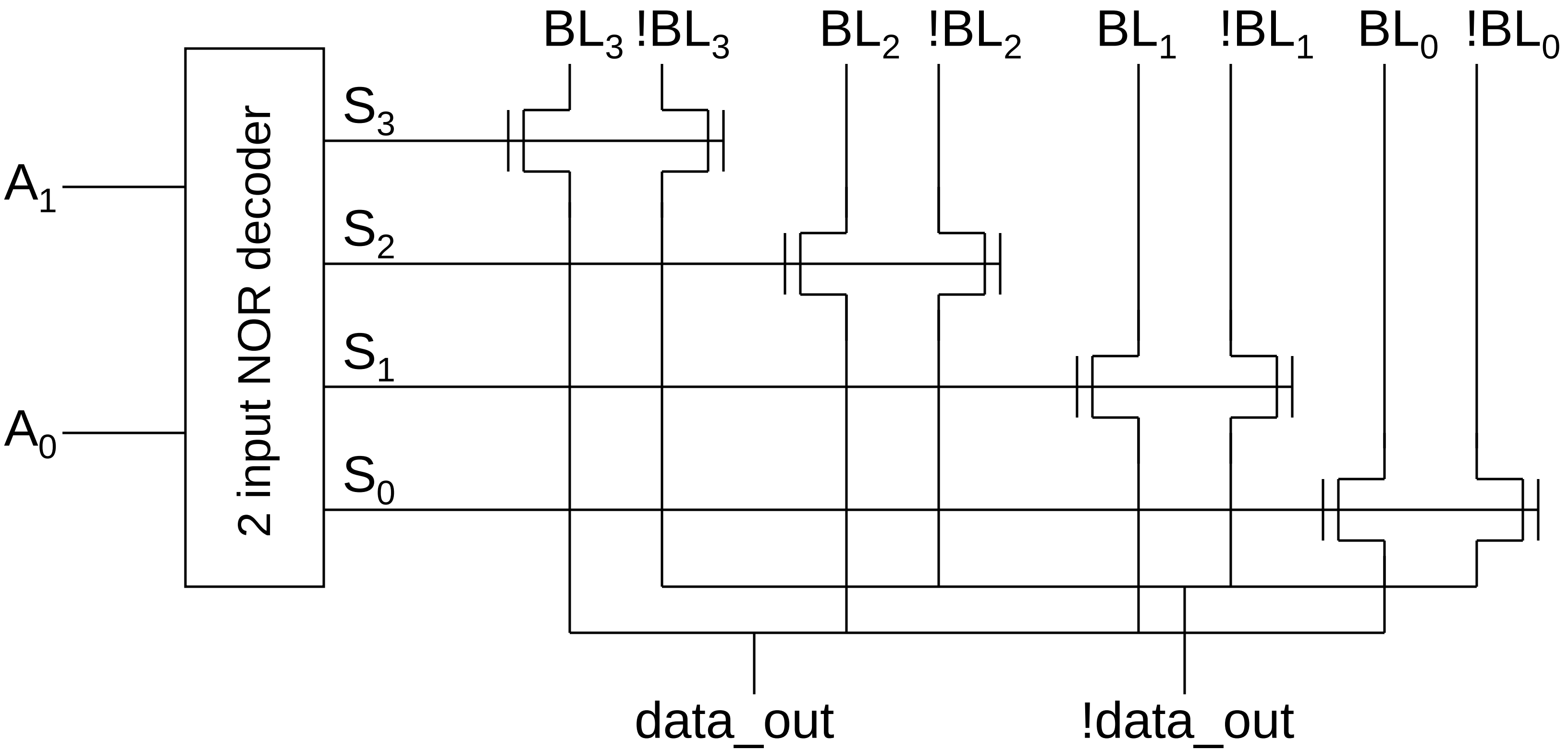
- Number of inputs to the gates driving the WLs is halved, so the propagation delay is reduced by a factor of ~4

Split Row Two-Level 8x256 Decoder


$!(!(!A_0\&!A_1\&!A_2) \mid !(!A_3\&!A_4\&!A_5) \mid !(!A_6\&!A_7))$
WL_0
WL_0
*256
*256
WL_{255}
WL_{255}
$!(!A_0 \& !A_1 \& !A_2)$
. . .
$!(A_0 \& A_1 \& A_2)$
*8
*8
*4
Address<7:0>
*8
*8

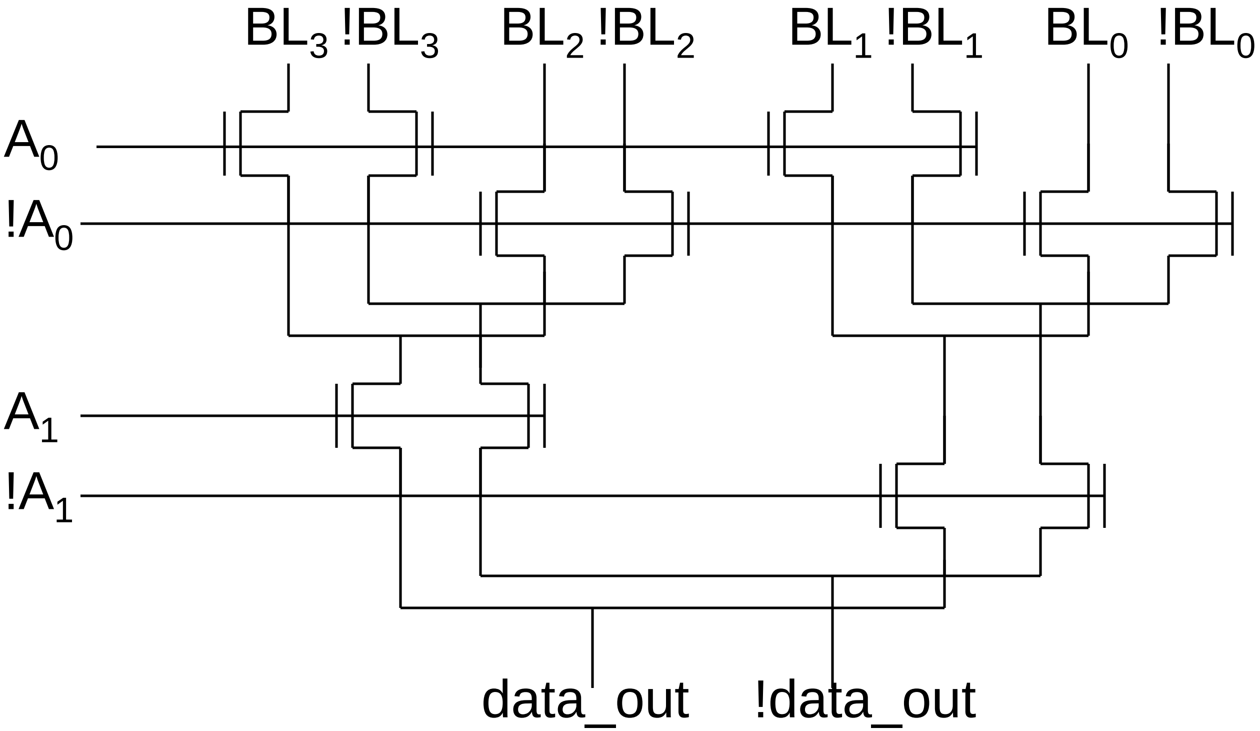
- Pitch matched
- Buffered word line drivers

Pass Transistor Based Column Decoder



- Read: connect BLs to the Sense Amps (SA) Writes: drive one of the BLs low to write a 0 into the cell
  - Fast since there is only one transistor in the signal path. However, there is a large transistor count ($(K+1)2^K + 2 \times 2^K$)
  - For K = 2 → 3×2^2 (decoder) + 2×2^2 (PTs) = 12 + 8 = 20

Tree Based Column Decoder



- ❑ Number of transistors reduced to (2 x 2 x (2^K -1))
  - for K = 2 → 2 x 2 x (2^2 – 1) = 4 x 3 = 12
- ❑ Delay increases quadratically with the number of sections (K) (so prohibitive for large decoders)
  - can fix with buffers, progressive sizing, combination of tree and pass transistor approaches

Decoder Complexity Comparisons

- Consider a memory with 10b address and 8b data

Conf.	Data/Row	Row Decoder	Column Decoder
1D	8b	10b = a 10x2^{10} decoder Single stage = 20,480 Two stage = 10,320	

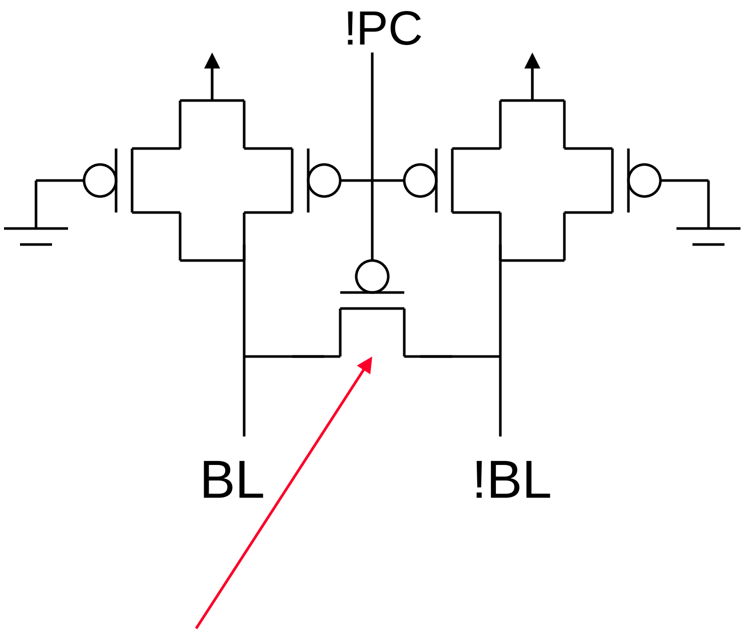
2D	32b
	(32x256 core)
		8b = 8x2^{8} decoder
		Single stage = 4,096 T
		Two stage = 2,112 T
			2b = 2x2^{2} decoder
			PT = 76 T
			Tree = 96 T

2D	64b
	(64x128 core)
		7b = 7x2^{7} decoder
		Single stage = 1,792 T
		Two stage = 1,072 T
			3b = 3x2^{3} decoder
			PT = 160 T
			Tree = 224 T

2D	128b
	(128x64 core)
		6b = 6x2^{6} decoder
		Single stage = 768 T
		Two stage = 432 T
			4b = 4x2^{4} decoder
			PT = 336 T
			Tree = 480 T

Bit Line Precharge Logic

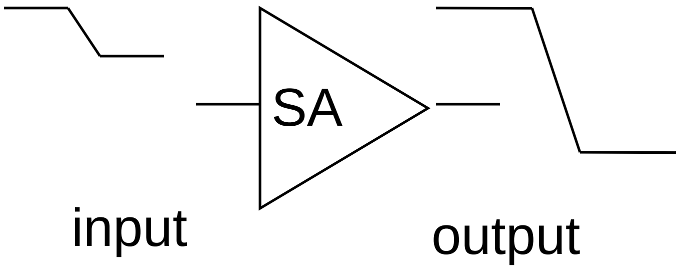
- First step of a Read cycle is to precharge (PC) the bit lines to V_{DD}
  - every differential signal in the memory must be equalized to the same voltage level before Read
- Turn off PC and enable the WL
  - the grounded PMOS load limits the bit line swing (speeding up the next precharge cycle)



equalization transistor - speeds up equalization of the two bit lines by allowing the capacitance and pull-up device of the nondischarged bit line to assist in precharging the discharged line

Sense Amplifiers

- Amplification – resolves data with small bit line swings (in some DRAMs required for proper functionality)



- Delay reduction – compensates for the limited drive capability of the memory cell to accelerate BL transition

small

$$t_p = (C * \Delta V) / I_{av}$$

large

make Δ V as small as possible

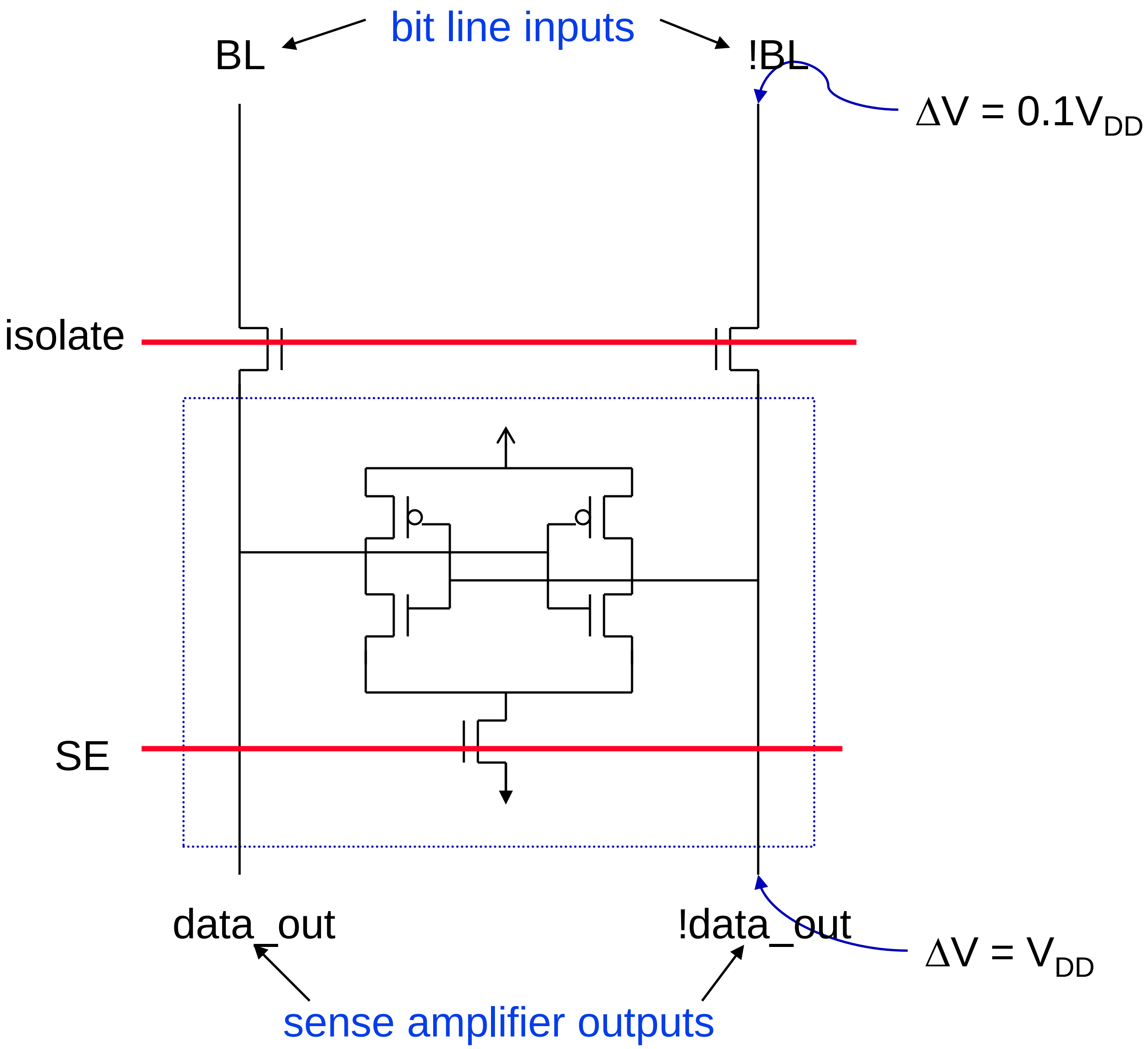
- Power reduction – eliminates a large part of the power dissipation due to charging and discharging bit lines

- Signal restoration – for DRAMs, need to drive the bit lines full swing after sensing (read) to do data refresh

Classes of Sense Amplifiers

- Differential SA – takes small signal differential inputs (BL and !BL) and amplifies them to a large signal single-ended output
  - common-mode rejection – rejects noise that is equally injected to both inputs
- Only suitable for SRAMs (with BL and !BL)
- Types
  - Current mirroring
  - Two-stage
  - Latch based
- Single-ended SA – needed for DRAMs

Latch Based Sense Amplifier



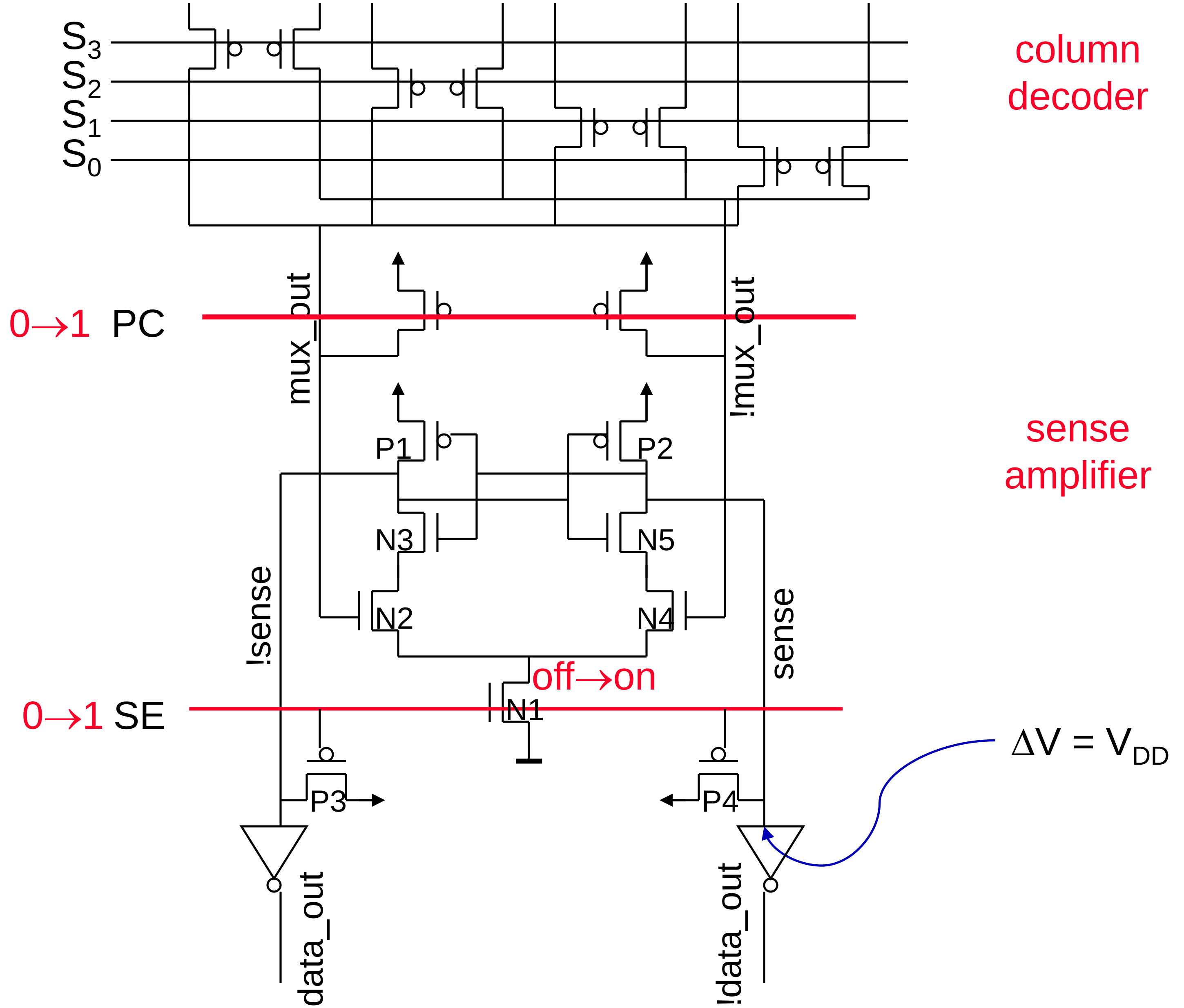
bit line inputs

$\Delta V = 0.1V_{DD}$

$\Delta V = V_{DD}$

sense amplifier outputs

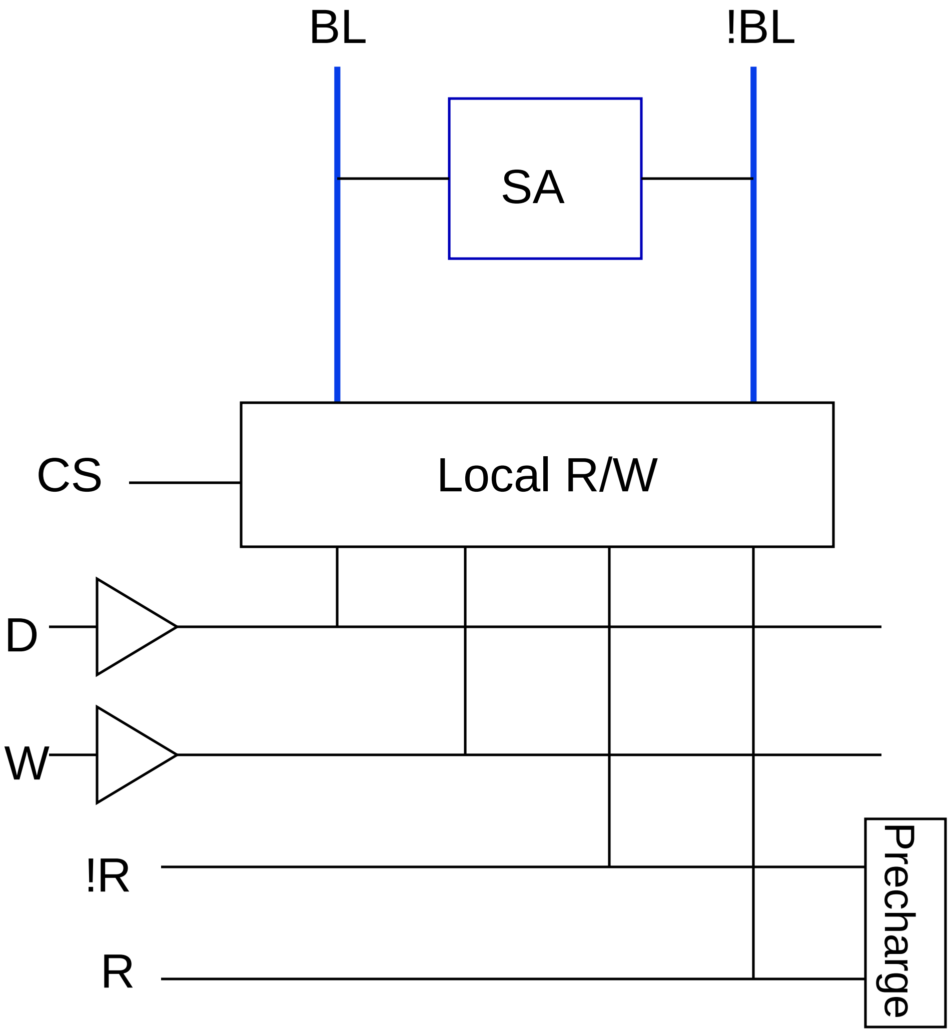
Alpha Differential Amplifier/Latch



column decoder

sense amplifier

Read/Write Circuitry

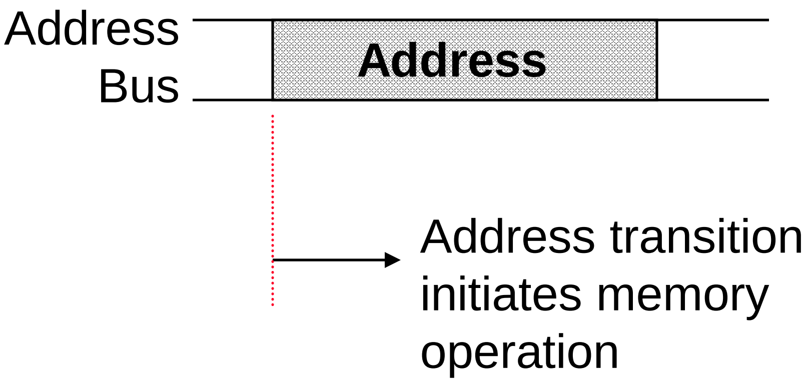


D: data (write) bus
R: read bus
W: write signal
CS: column select
(column decoder)

Local W (write):
BL = D, !BL = !D
enabled by W & CS
Local R (read):
R = BL, !R = !BL
enabled by !W & CS

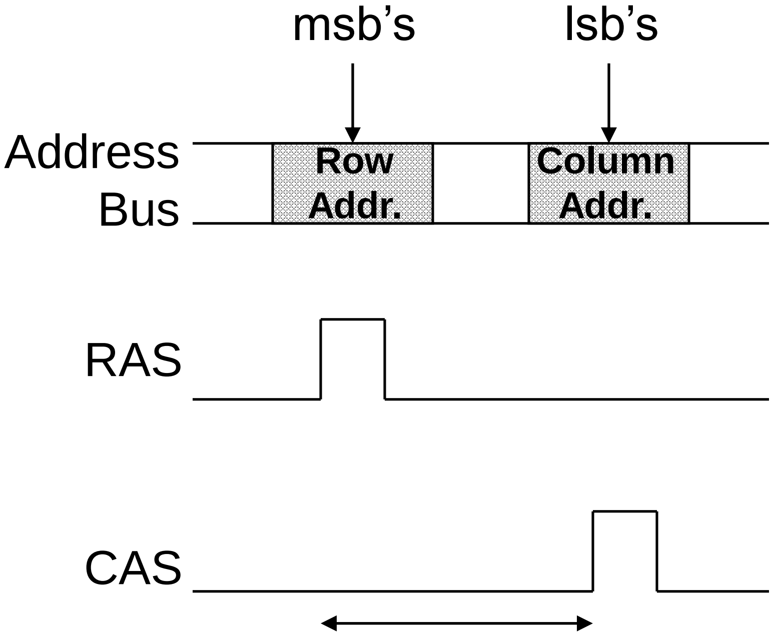
Approaches to Memory Timing

SRAM Timing
Self-Timed



Address transition initiates memory operation

DRAM Timing
Multiplexed Addressing



RAS-CAS timing