ECE4740: Digital VLSI Design

Lecture 28: Memories

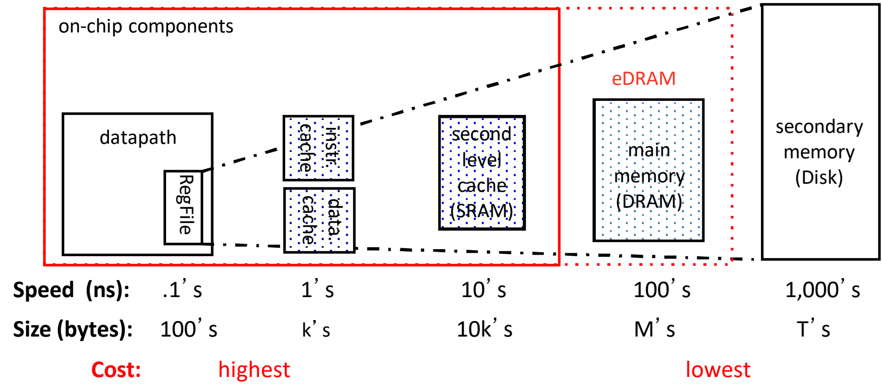
1024

Contribute significantly to area and power of VLSI circuits

Semiconductor memories

1025

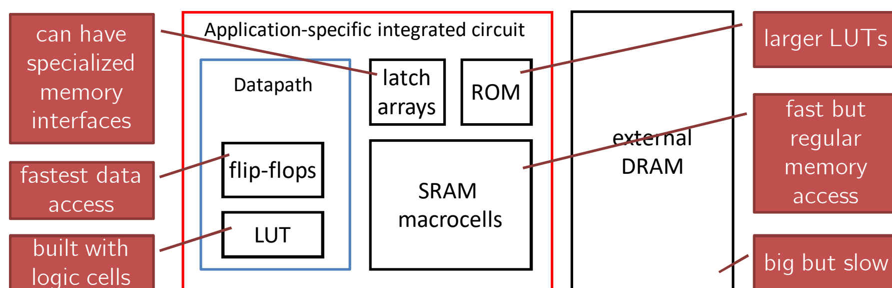
Memory hierarchy in a processor



- Exploit locality:
  - large amount of cheap memory
  - sufficient amount of fast, expensive memory

1026

Memories in an ASIC



- Critical data stored in flip-flops within datapath
- Small memories, typically build from standard-cell based static cells (latches or flip-flops)
- Larger amounts of data stored in SRAM macrocells
- Off-chip DRAMs store GBs of data

1027

Semiconductor memories

read/write memories

non-volatile

RWM		NVRWM	ROM
Random Access	Non-Random Access	EPROM	Mask-programmed
SRAM (cache, register file)	FIFO/LIFO	E^2PROM	
DRAM	Shift register CAM	FLASH	Electrically-programmed (PROM)

- 50% (growing) of silicon area is memory in most designs
- Often limits throughput or energy efficiency

1028

Also this: write-only memories

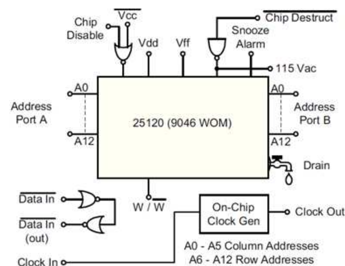
signetics FULLY ENCODED, 9046 X N, RANDOM ACCESS WRITE-ONLY-MEMORY

25120

FINAL SPECIFICATION[10]

DESCRIPTION

The Signetics 25000 Series 9C46XN Random Access Write-Only-Memory employs both enhancement and depletion mode P-Channel, N-Channel and Neu[1] channel MOS devices. Although a static device, a single TTL level clock phase is required to drive the on-board multi-port clock generator. Data refresh is accomplished during CB and LH periods [11]. Quadri-state outputs (when applicable) allow expansion in many directions, depending on organization.



INPUT PROTECTION

All terminals are provided with slip-on latex protectors for the prevention of Voltage Destruction. (PILL packaged devices do not require protection.)

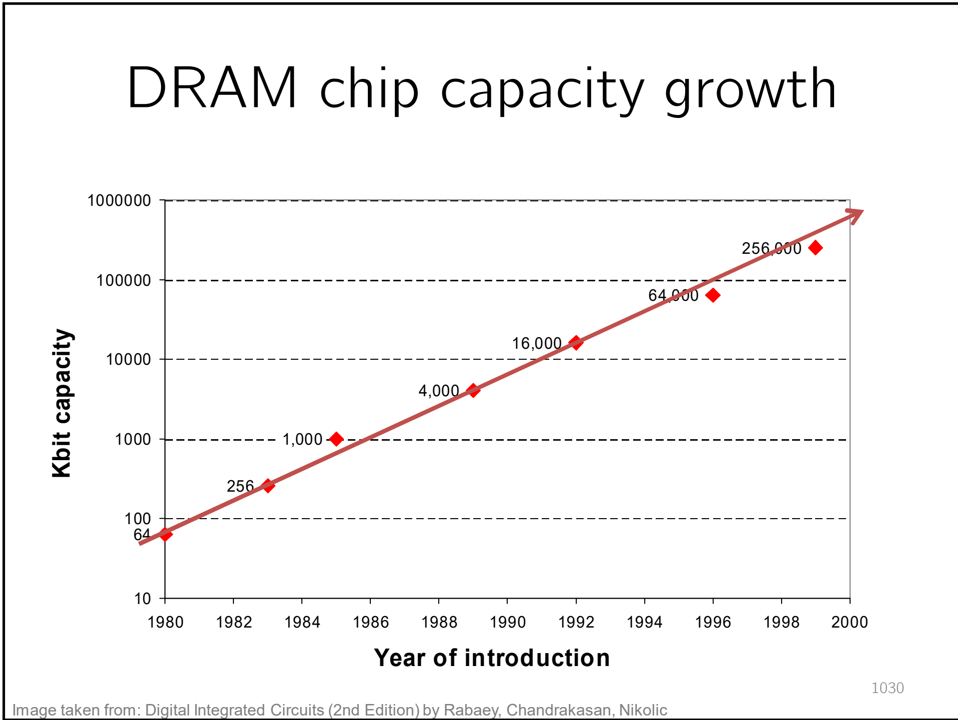
COOLING

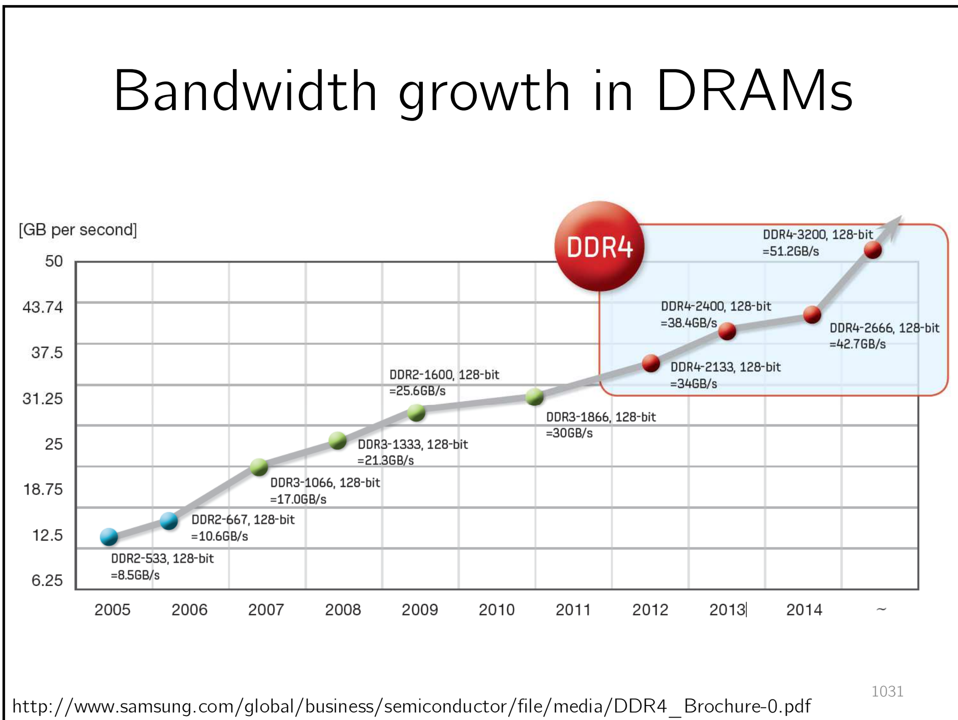
The 25120 is easily cooled by the employment of a six foot fan ⅓" ility is from the package. If the device fails you have exceeded the ragings. In such cases, more air is recommended.

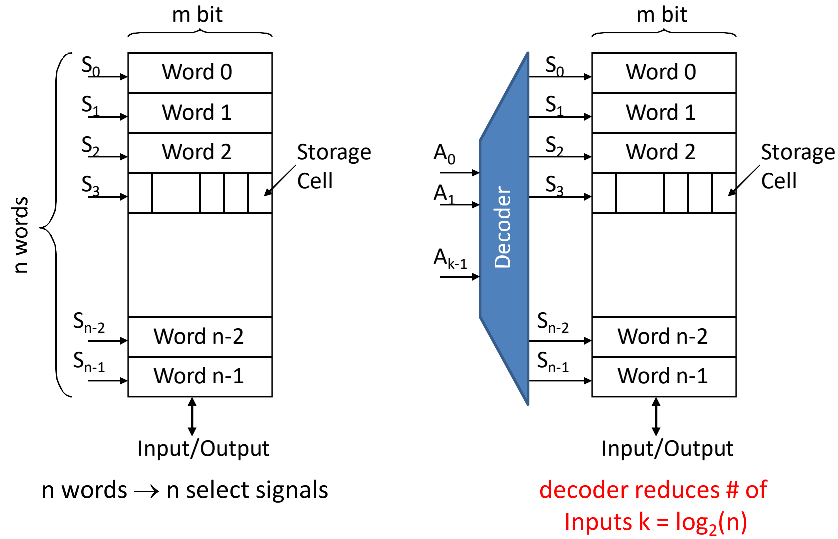
APPLICATIONS

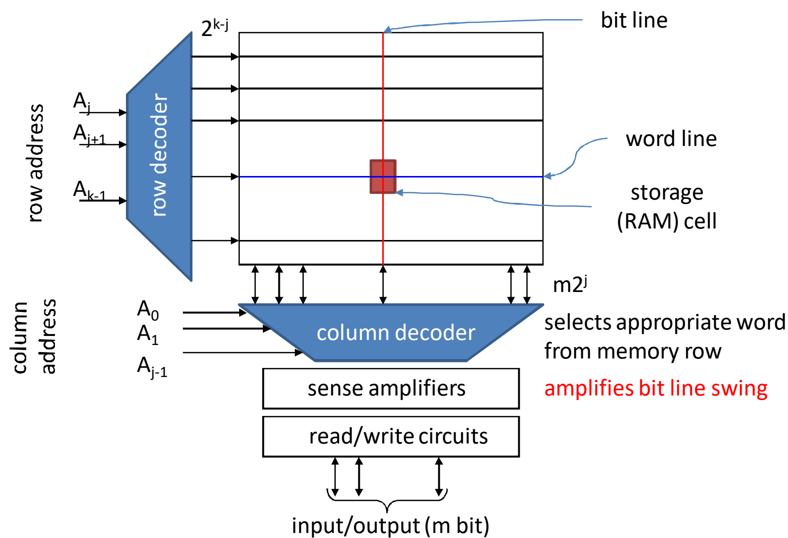
DON'T CARE BUFFER STORES
LEAST SIGNIFICANT CONTROL MEMORIES
POST MORTEM MEMORIES (WEAPON SYSTEMS)
ARTIFICIAL MEMORY SYSTEMS
NON-INTELLIGENT MICRO CONTROLLERS
FIRST-IN NEVER-OUT (FINO) ASYNCHRONOUS BUFFERS.
OVERFLOW REGISTER (BIT BUCKET)

1029

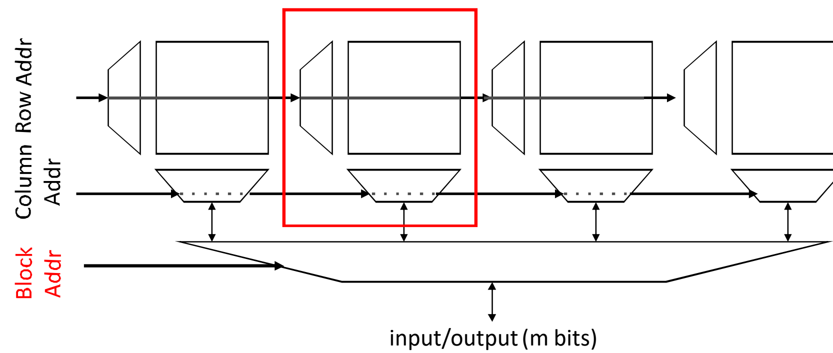

DRAM chip capacity growth
Kbit capacity
1000000
100000
10000
1000
100
64
10
256,000
64,000
16,000
4,000
1,000
256
1980
1982
1984
1986
1988
1990
1992
1994
1996
1998
2000
Year of introduction
1030
Image taken from: Digital Integrated Circuits (2nd Edition) by Rabaey, Chandrakasan, Nikolic


Bandwidth growth in DRAMs
[GB per second]
50
43.74
37.5
31.25
25
18.75
12.5
6.25
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
~
DDR4
DDR2-533, 128-bit =8.5GB/s
DDR2-667, 128-bit =10.6GB/s
DDR3-1066, 128-bit =17.0GB/s
DDR3-1333, 128-bit =21.3GB/s
DDR2-1600, 128-bit =25.6GB/s
DDR3-1866, 128-bit =30GB/s
DDR4-2133, 128-bit =34GB/s
DDR4-2400, 128-bit =38.4GB/s
DDR4-2666, 128-bit =42.7GB/s
DDR4-3200, 128-bit =51.2GB/s
1031
http://www.samsung.com/global/business/semiconductor/file/media/DDR4_Brochure-0.pdf


Memory architecture overview
m bit
n words
S_0
S_1
S_2
S_3
S_{n-2}
S_{n-1}
Word 0
Word 1
Word 2
Word n-2
Word n-1
Storage Cell
Input/Output
n words → n select signals
A_0
A_1
A_{k-1}
Decoder
decoder reduces # of
Inputs k = $\log_2(n)$
1032


Memory architecture details
2^{k-j}
bit line
word line
storage (RAM) cell
row address
row decoder
A_j
A_{j+1}
A_{k-1}
column address
A_0
A_1
A_{j-1}
$m2^j$
column decoder
selects appropriate word from memory row
sense amplifiers
amplifies bit line swing
read/write circuits
input/output (m bit)
1033

Split large memories in blocks/banks



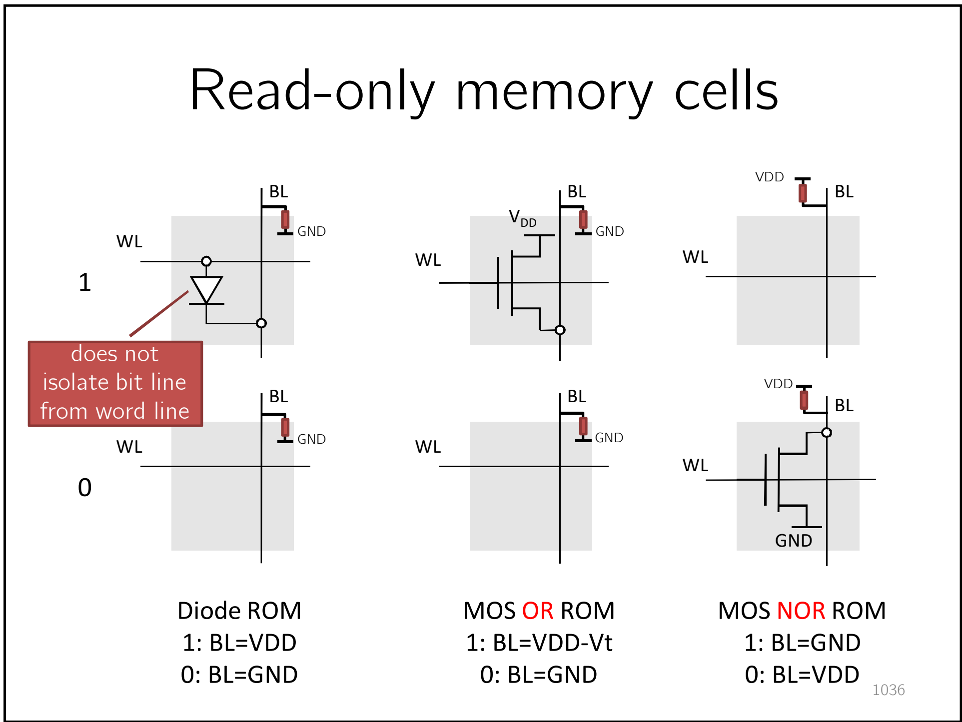
- Shorter word and bit lines → faster
- BlockAddr signal activates only 1 block → power reduction

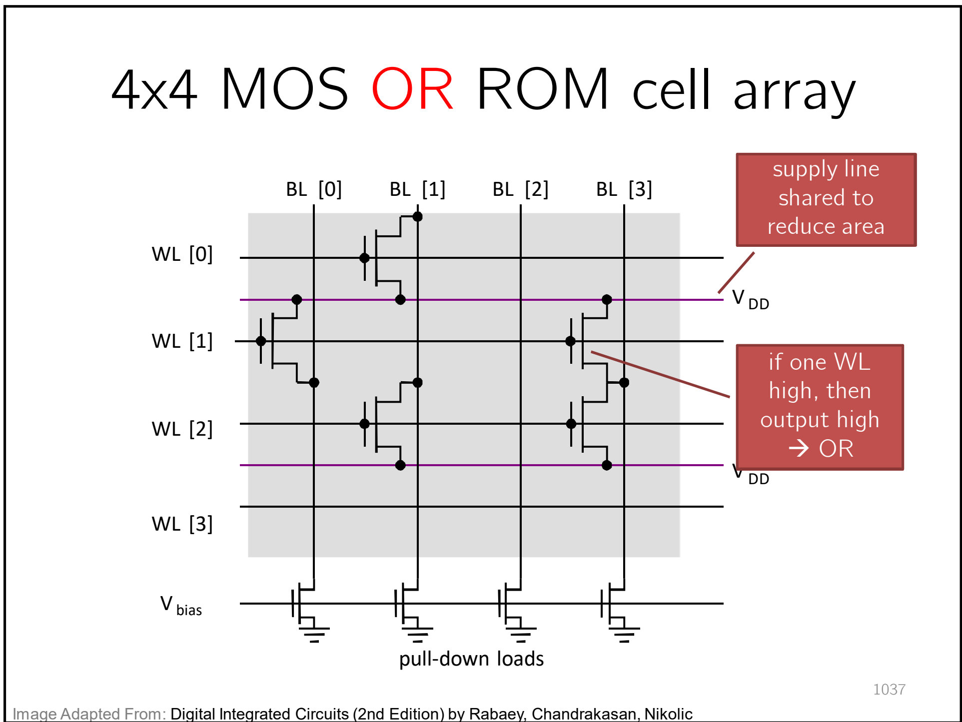
1034

Useful for large look-up tables (LUTs)

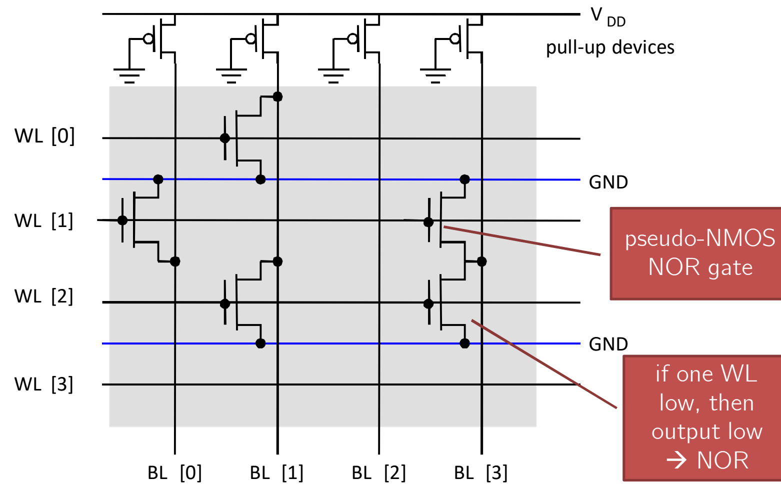
Read-only memories (ROMs)

1035


Read-only memory cells
BL
GND
WL
1
does not isolate bit line from word line
BL
GND
WL
0
BL
VDD
GND
WL
BL
GND
WL
VDD
BL
WL
VDD
BL
WL
GND
Diode ROM
1: BL=VDD
0: BL=GND
MOS OR ROM
1: BL=VDD-Vt
0: BL=GND
MOS NOR ROM
1: BL=GND
0: BL=VDD
1036

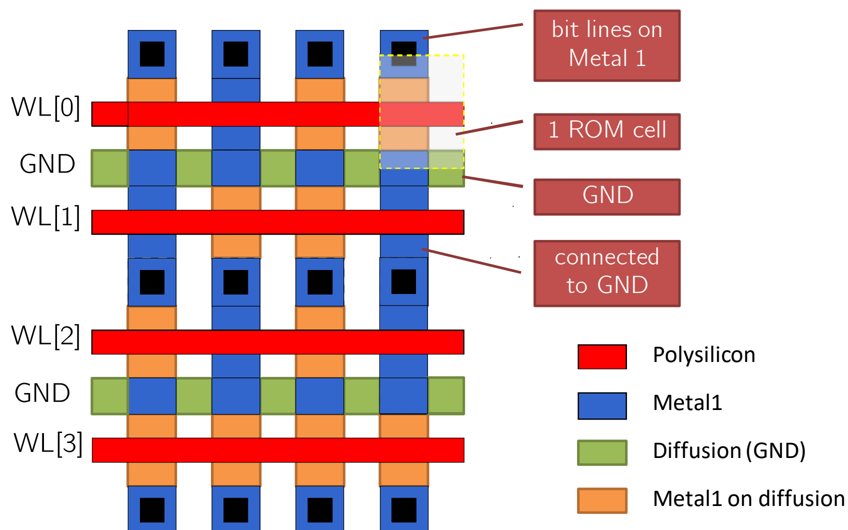

4x4 MOS OR ROM cell array
BL [0]
BL [1]
BL [2]
BL [3]
supply line shared to reduce area
WL [0]
V DD
WL [1]
if one WL high, then output high → OR
WL [2]
V DD
WL [3]
V bias
pull-down loads
1037
Image Adapted From: Digital Integrated Circuits (2nd Edition) by Rabaey, Chandrakasan, Nikolic

4x4 MOS NOR ROM cell array

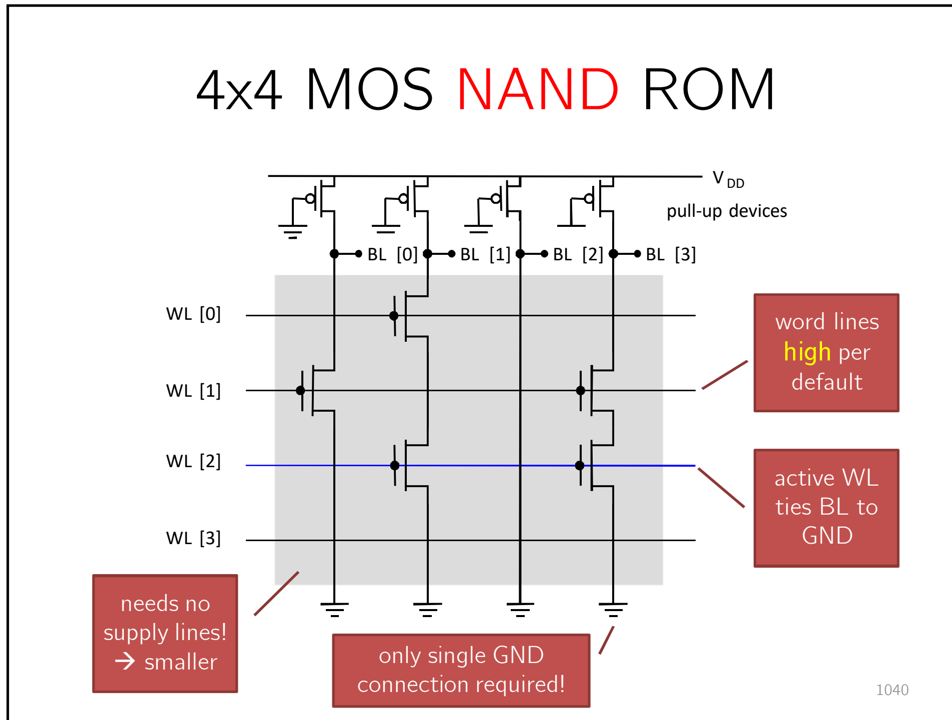


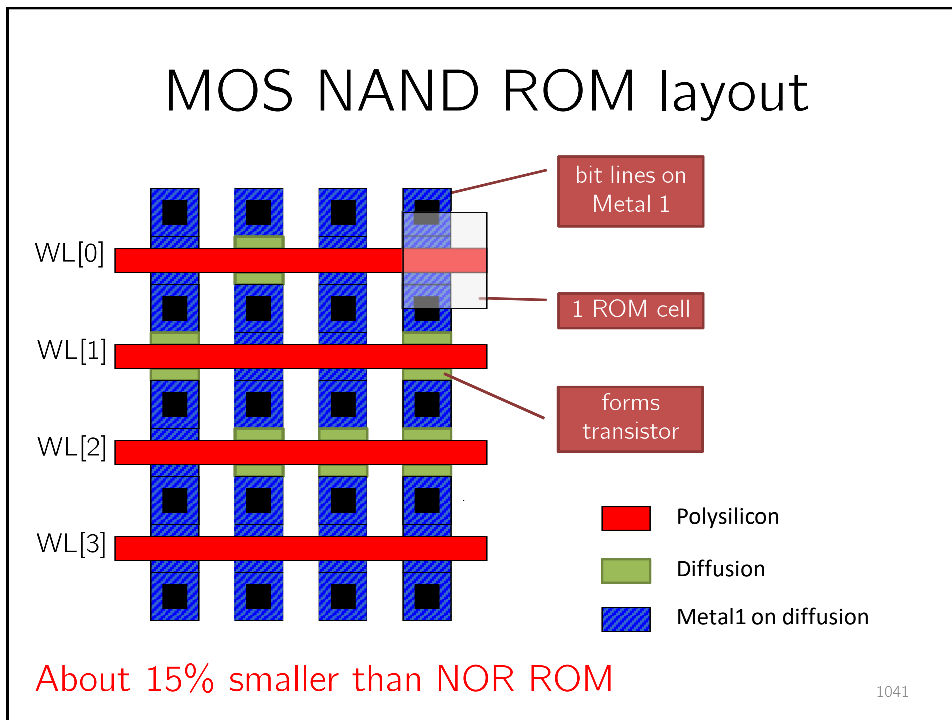
1038

MOS NOR ROM layout



1039

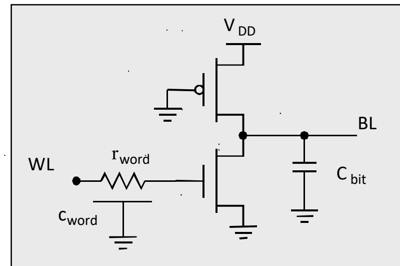

4x4 MOS NAND ROM
V DD
pull-up devices
BL [0]
BL [1]
BL [2]
BL [3]
WL [0]
WL [1]
WL [2]
WL [3]
word lines high per default
active WL ties BL to GND
needs no supply lines! → smaller
only single GND connection required!
1040


MOS NAND ROM layout
WL[0]
WL[1]
WL[2]
WL[3]
bit lines on Metal 1
1 ROM cell
forms transistor
Polysilicon
Diffusion
Metal1 on diffusion
About 15% smaller than NOR ROM
1041

(Model for NOR ROM)

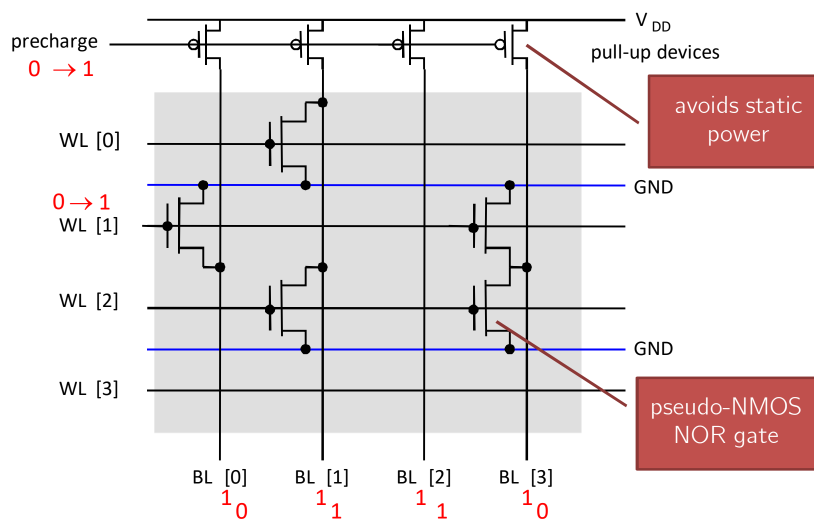
- Word-line parasitics
  - Wire capacitance and gate capacitance
  - Wire resistance (polysilicon)
- Bit-line parasitics
  - Resistance not dominant (metal)
  - Drain and gate-drain capacitance

NOR model



1042

Precharged MOS NOR ROM



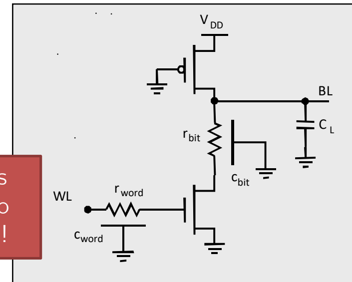
1043

Image adapted from: Digital Integrated Circuits (2nd Edition) by Rabaey, Chandrakasan, Nikolic

(Model for NAND ROM)

- Word-line parasitics
  - Wire capacitance and gate capacitance
  - Wire resistance (polysilicon)

NAND model



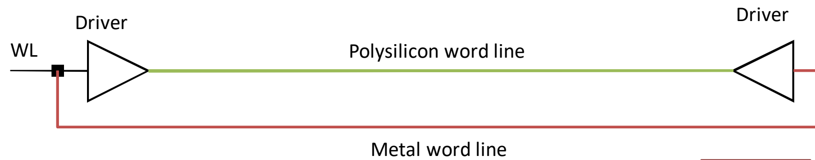
NAND ROM is smaller but also much slower!!!!

- Bit-line parasitics
  - Resistance of cascaded transistors dominates
  - Drain/source and complete gate capacitance

1044

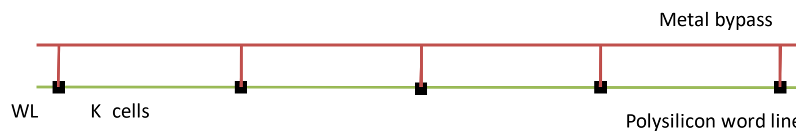
Remember: reducing word-line delay

- Use bypasses!
  - Drive word line from both sides



  - Use metal bypass (better materials)

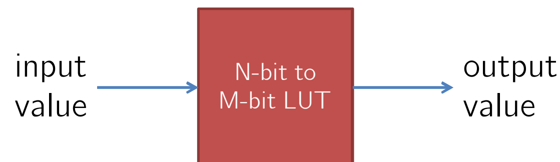
e.g., use silicides



1045

ROMs for look-up tables (LUTs)

- LUTs used in ASIC designs and some processors
  - Store fixed program (hard-coded software)
  - Fast division and square root units
  - Approximating arbitrary arithmetic functions



- Just set input value to word address and output value is ROM content
- ROM-based approach only useful for large LUTs

1046

RWM is hard to pronounce

Read-write memories (RAMs)

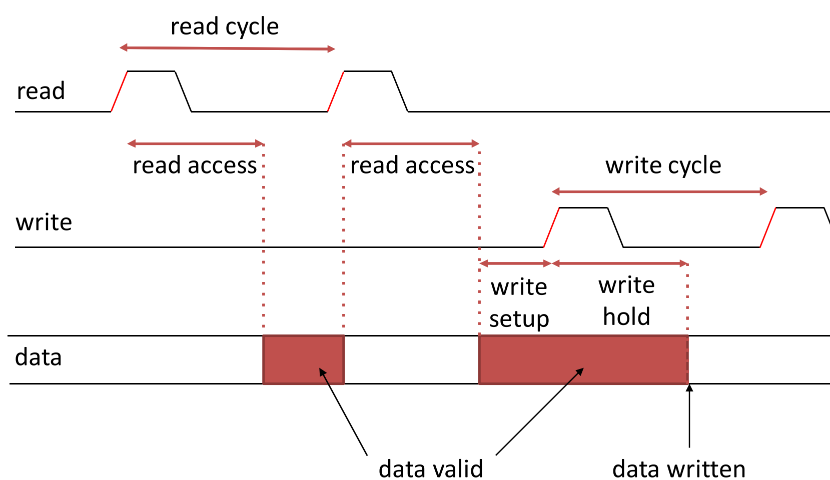
1047

Read-write memories (RAMs)

- Static: SRAM
  - Data is stored as long as supply is applied
  - Large cells (6T) → fewer bits per area
  - Fast (used where speed is important)
  - Differential outputs (BL and !BL)
  - Use sense amps for better performance
  - Compatible with CMOS technology
- Dynamic: DRAM
  - Requires periodic refresh
  - Small cells (1T to 3T) → more bits per area
  - Slower (used for large main memories)
  - Single-ended output (BL only)
  - Need sense amps for correct operation
  - Not typically compatible with regular CMOS technology

1048

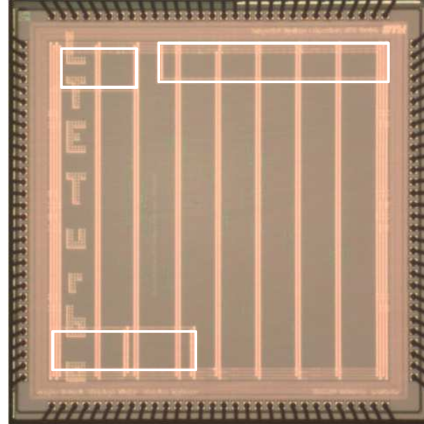
Memory timing



1049

SRAMs are used everywhere!

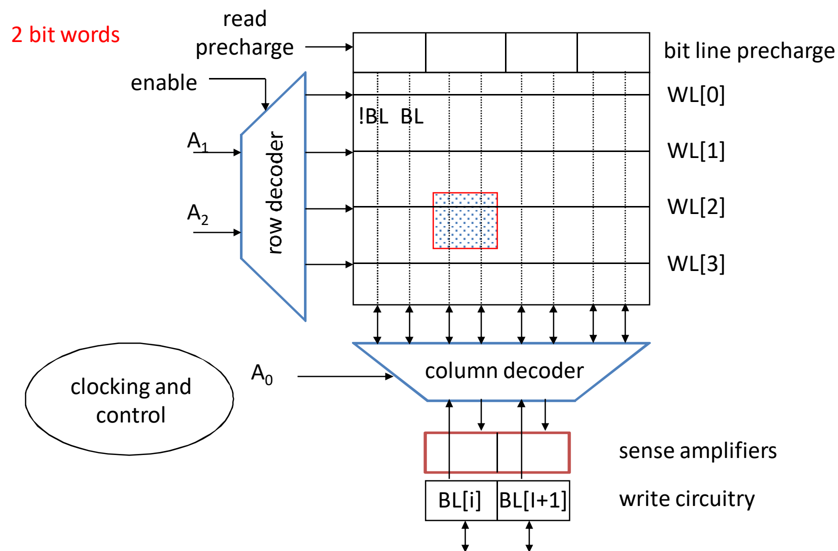
- Turbo-decoder ASIC for LTE
- 65nm CMOS
- 129kb SRAM
- Single-port only
- Runs at 300MHz



S, Benkeser, Belfanti, & Huang, 2011

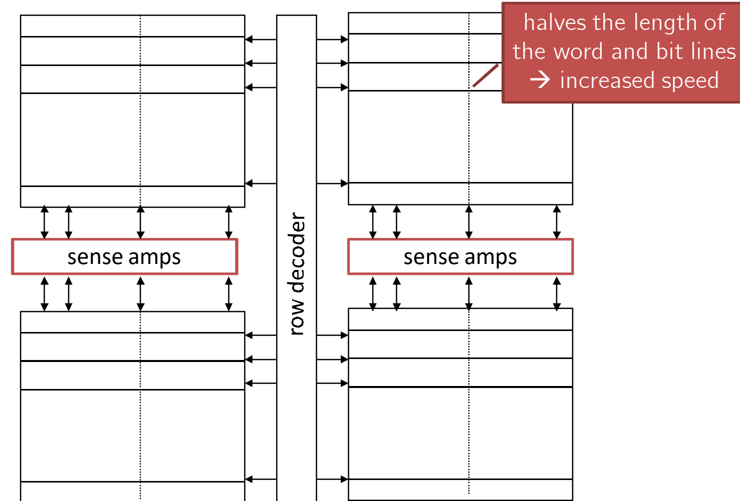
1050

4x4 static RAM (SRAM)



1051

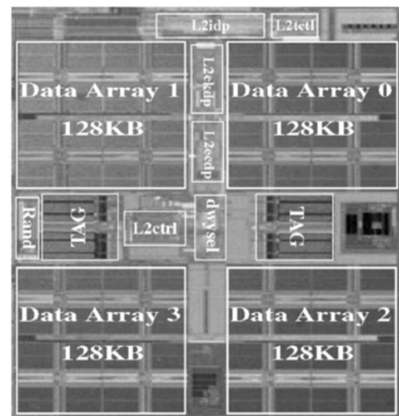
2D memory configuration



1052

Example: large SRAM

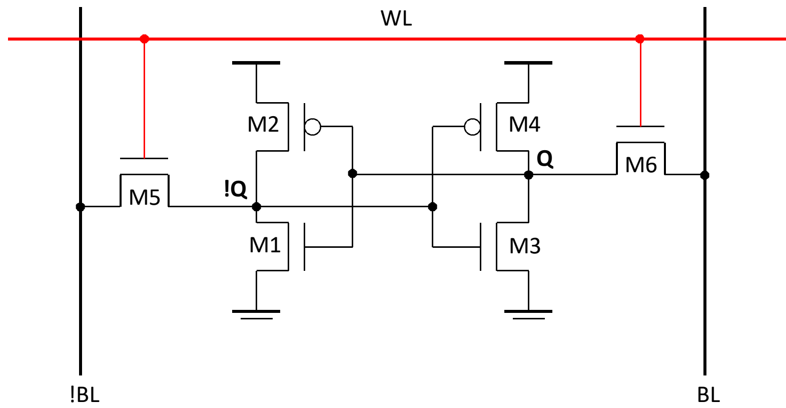
- UltraSparc 512KB cache SRAM
- Split into 4 subarrays
  - 4x 128KB subarrays
  - Each subarray consists of 16 8KB banks
  - Also includes control and cache circuitry



1053

Image taken from: CMOS VLSI Design: A Circuits and Systems Perspective by Weste, Harris

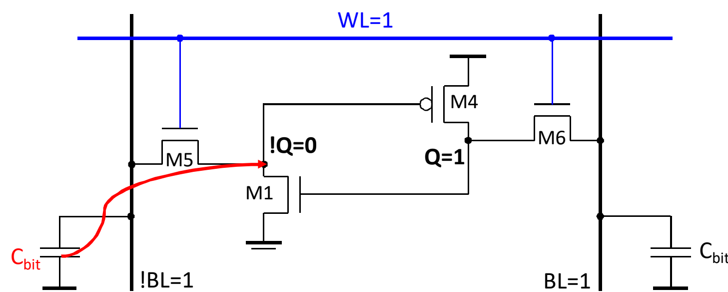
6T SRAM cell



- Cross-coupled inverters → sizing critical

1054

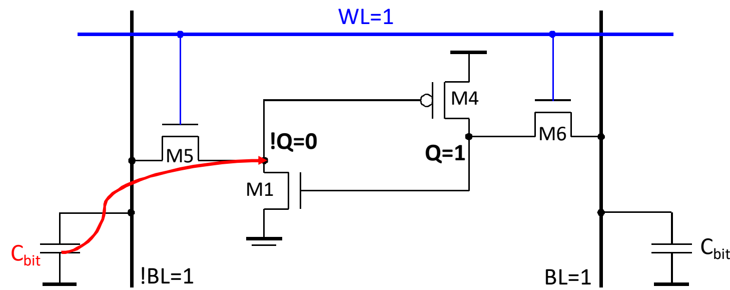
6T SRAM cell: read



- Read disturb (read upset):
  - Bit-line capacitance C_{bit} can be in pF range
  - Limit allowed voltage rise on !Q to not change SRAM cell state

1055

(6T SRAM cell: read)



- Cell ratio: CR = $(W_1/L_1)/(W_5/L_5)$
- ΔV = maximum allowed voltage ripple

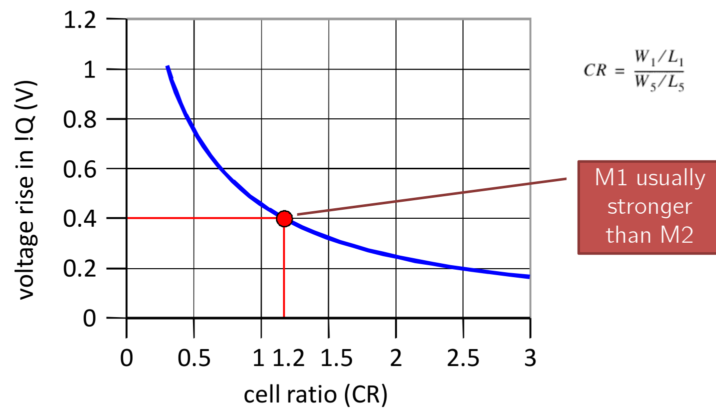
$$k_{n,M5}\left((V_{DD}-\Delta V-V_{Tn})V_{DSATn}-\frac{V_{DSATn}^2}{2}\right) = k_{n,M1}\left((V_{DD}-V_{Tn})\Delta V-\frac{\Delta V^2}{2}\right)$$

$$\Delta V = \frac{V_{DSATn}+CR(V_{DD}-V_{Tn})-\sqrt{V_{DSATn}^2(1+CR)+CR^2(V_{DD}-V_{Tn})^2}}{CR}$$

M5 = saturation
M1 = linear

1056

(6T SRAM cell: read)

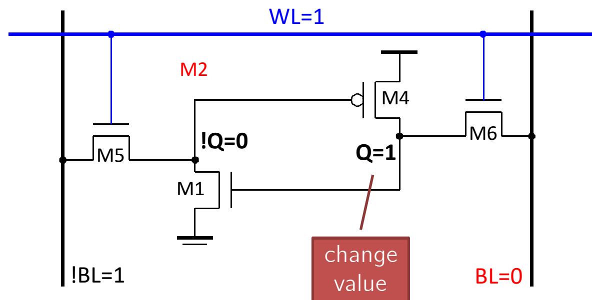


- Common cell ratios are 1.25 to 2

V_{dd} = 2.5V, V_{Tn} = 0.5V

1057

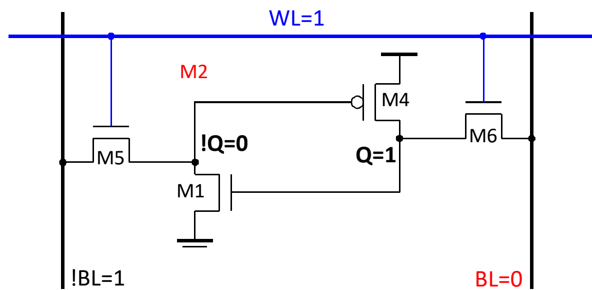
6T SRAM cell: write



- Q=1 stored in cell, trying to write a 0
- M6 must be more conductive than M4 to pull node Q low enough for M1 & M2

1058

(6T SRAM cell: write)



- Pullup ratio: PR= $(W_4/L_4)/(W_6/L_6)$

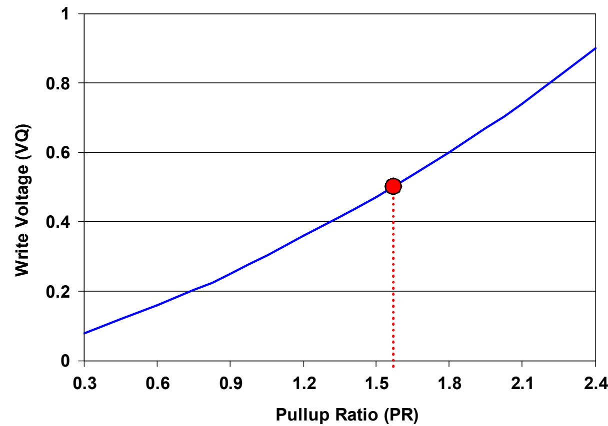
$$k_{n,M6}\left((V_{DD}-V_{Tn})V_Q-\frac{V_Q^2}{2}\right) = k_{p,M4}\left((V_{DD}-|V_{Tp}|)V_{DSATp}-\frac{V_{DSATp}^2}{2}\right)$$

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

M4 = saturation
M6 = linear

1059

(6T SRAM cell: write)



- Node Q must be pulled below V_{Tn}

V_{dd} = 2.5V, $|V_{Tp}|$ = 0.5V, μ_p/μ_n = 0.5

1060

Cell sizing

- Keep cell size minimized → max. density
- Min-sized pull-down FETs (M1 and M3)
  - Requires min-width and longer-than-min-length PTs (M5 and M6) to ensure proper CR
  - Sizing of PTs increases load on bit lines
- Min-sized PTs can be used
  - Increase width of pull-downs (M1 and M3)
  - Reduces load on word lines but increases size

1061

Cell layout of 6T SRAM

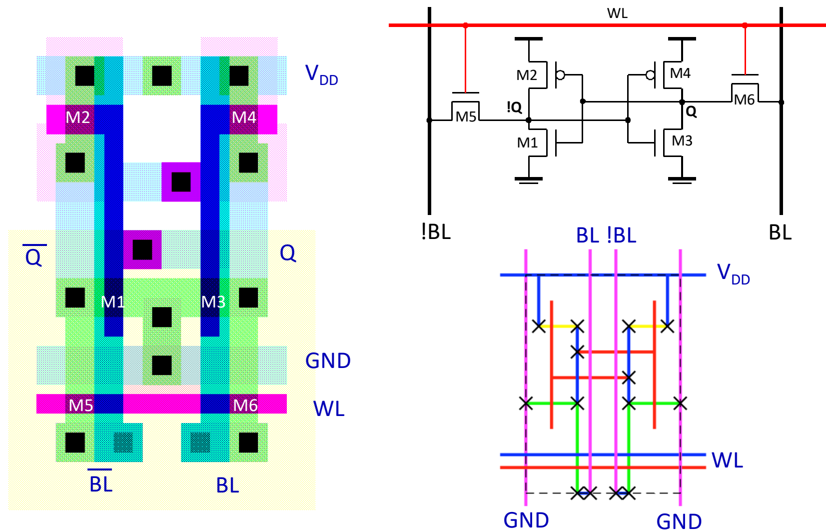


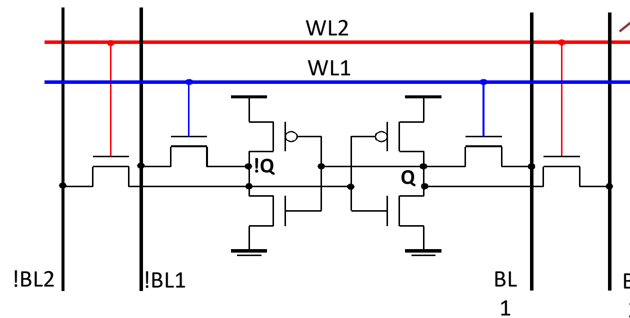
Image taken from: CMOS VLSI Design: A Circuits and Systems Perspective by Weste, Harris

1062

Memories with multiple ports

- So far: single-port memories:
  - either read or write
- Dual (or two) port memories:
  - read and write, 2x read, or 2x write

careful with sizing of PTs and also access contentions



1063

Reducing bit-line delay

- Reduce voltage swing
  - Needs sense amplifier to restore signal
- Isolate memory cells from bit lines after sensing → pulsed word line
- Isolate sense amplifiers from bit lines after sensing → bit line isolation

1064

Bit line isolation

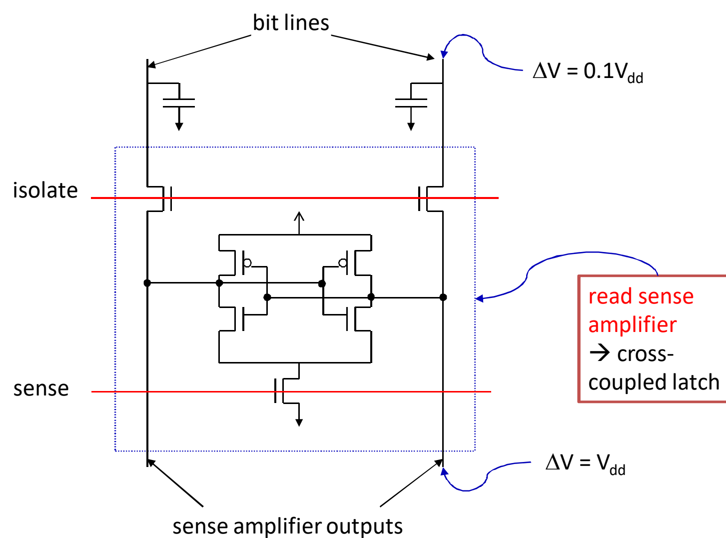


Image taken from: Digital Integrated Circuits (2nd Edition) by Rabaey, Chandrakasan, Nikolic

1065

Higher density than static RW memories (SRAMs)

Dynamic RAM (DRAM)

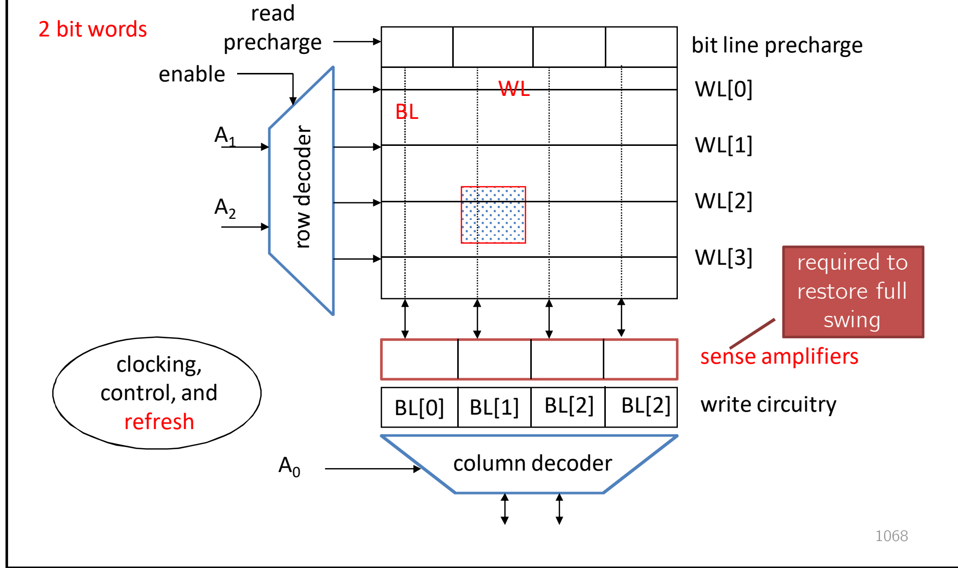
1066

DRAM properties

- Requires periodic refresh
- Small cells (1T to 3T) → more bits/area
- Slower (used for large main memories)
- Single-ended output (bit-line only)
- Need sense amps for correct operation
- Typically not compatible with regular CMOS technology ☹

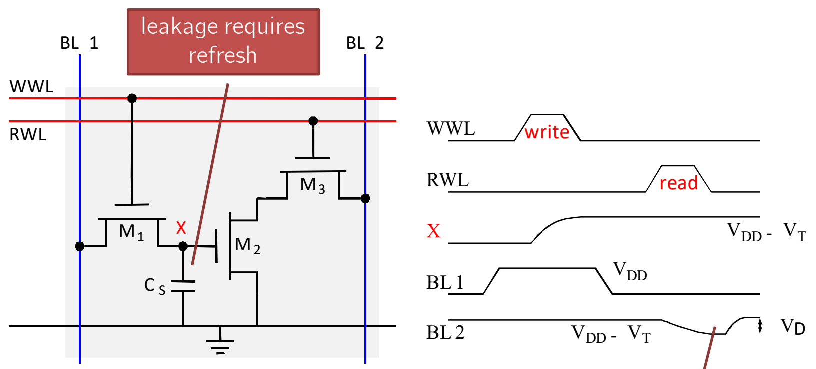
1067

4x4 dynamic RAM (DRAM)



1068

3T DRAM cell

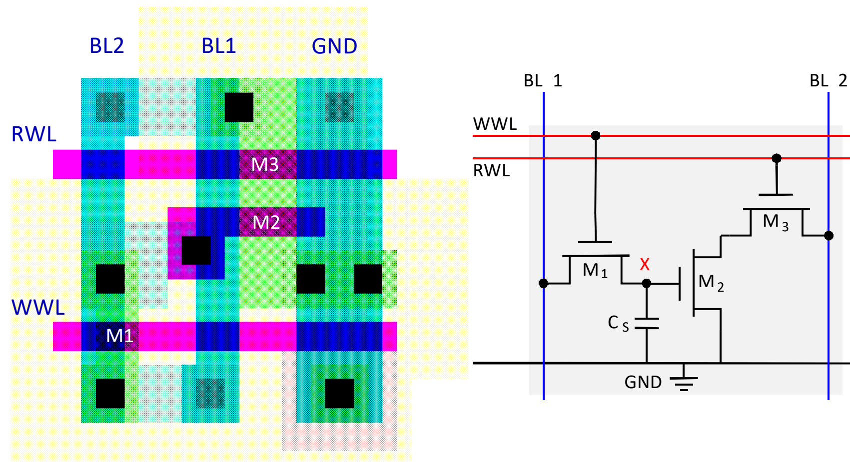


- No constraint on device ratios
- Reads are non-destructive
- Value stored at node X: "1"=V_{WWL}-V_{Tn}

1069

Image taken from: Digital Integrated Circuits (2nd Edition) by Rabaey, Chandrakasan, Nikolic

Layout of 3T DRAM cell

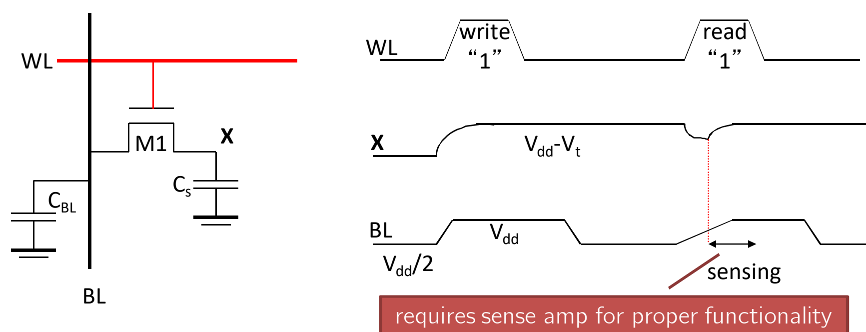


- About 2x smaller than SRAM cell

1070

Image taken from: Digital Integrated Circuits (2nd Edition) by Rabaey, Chandrakasan, Nikolic

1T DRAM cell

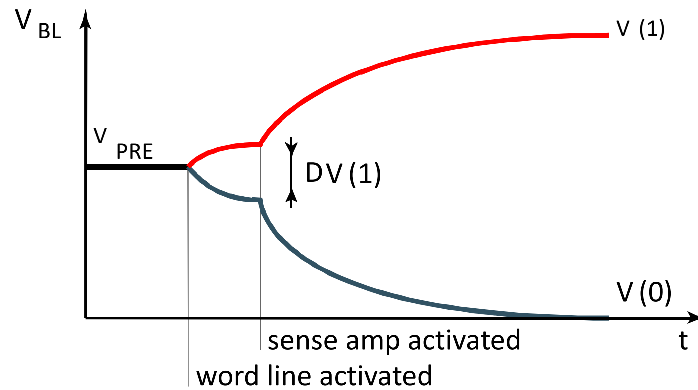


- Write: C_S is charged (or discharged) by asserting WL and asserting (or lowering) BL
- Read: Charge distribution between C_{BL} and C_S
- Read is destructive → refresh after read

1071

Image taken from: Digital Integrated Circuits (2nd Edition) by Rabaey, Chandrakasan, Nikolic

Sense amplifier operation



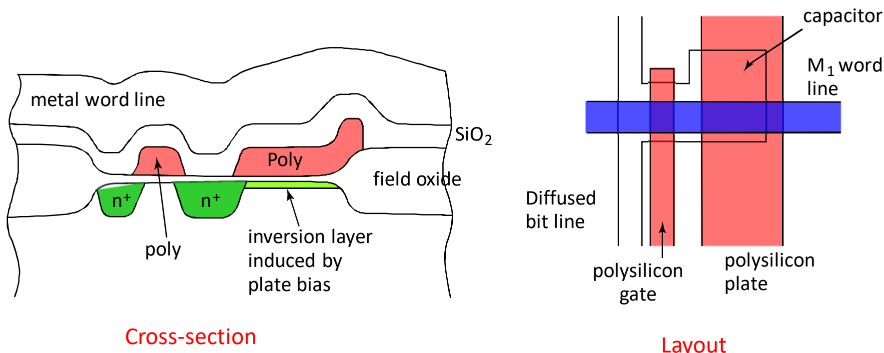
- Sense amplifier needed for functionality

1072

Image taken from: Digital Integrated Circuits (2nd Edition) by Rabaey, Chandrakasan, Nikolic

1T DRAM cell properties

- Output is single-ended (requires special sense amps)
- Requires extra capacitor to store state
- Not compatible with conventional CMOS processes



1073

Decoders, sense amps, etc.

Peripheral memory circuits

Row decoders (remember Lab 2)

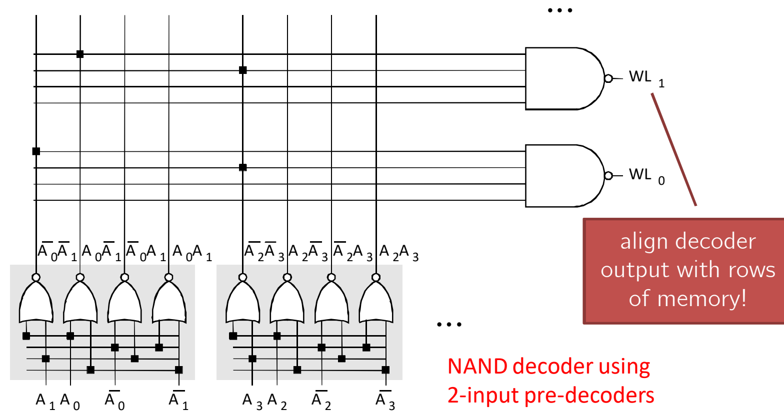
- M to 2^M decoder consists of 2^M logic gates
- Two main options:
  - NAND decoder

$$WL_0 = A_0 A_1 A_2 A_3 A_4 A_5 A_6 A_7 A_8 A_9$$
$$\cdots$$
$$WL_{511} = \bar{A}_0 A_1 A_2 A_3 A_4 A_5 A_6 A_7 A_8 A_9$$

  - NOR decoder

$$WL_0 = \overline{A_0 + A_1 + A_2 + A_3 + A_4 + A_5 + A_6 + A_7 + A_8 + A_9}$$
$$\cdots$$
$$WL_{511} = \overline{A_0 + \bar{A}_1 + \bar{A}_2 + \bar{A}_3 + \bar{A}_4 + \bar{A}_5 + \bar{A}_6 + \bar{A}_7 + \bar{A}_8 + \bar{A}_9}$$

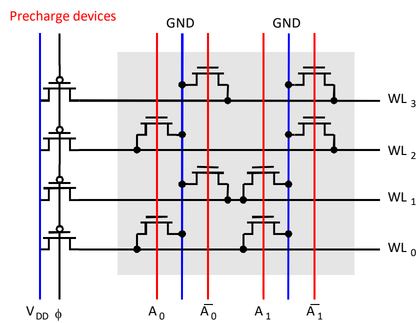
Hierarchical row decoding



- Multi-stage decoding improves performance

1076

Dynamic NOR row decoder

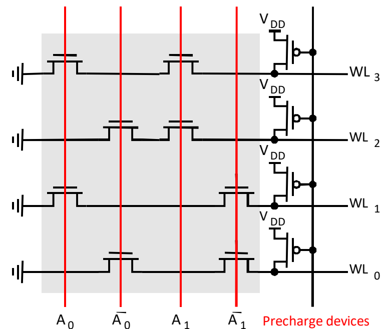


- Precharge all outputs high
- Then, GND inactive outputs
- Active "high" signals

- Signal goes through at most one FET
- Almost constant propagation delay

1077

Dynamic NAND row decoder

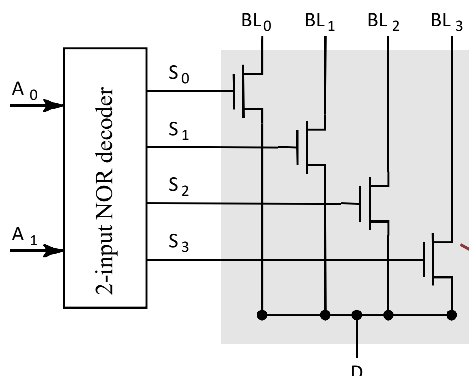


- Precharge all outputs high
- Then, discharge active output
- Active "low" signals

- All inputs must be low during precharge (prevent VDD and GND short)
- Slower than NOR implementation (3 FETs in series)

1078

PT-based column decoder



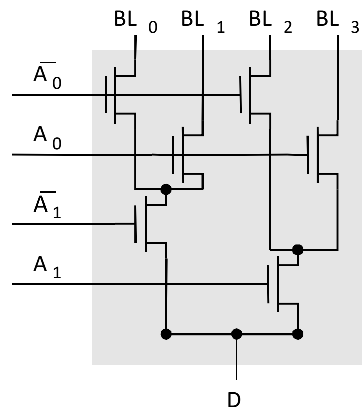
- One transistor per bit line
- $k*2^k$ transistors in total
- k=10 → 10240T

SRAMs require 2x the transistors for BL and !BL

- Advantage: speed (only 1 extra T in path)
- Disadvantage: large transistor count

1079

Tree-based column decoder



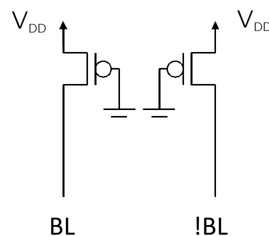
- One transistor per bit line
- $2^*(2^k-1)$ transistors in total
- k=10 → 2046T

- Advantage: number of T reduced significantly
- Disadvantage: delay increases quadratically with stages
  - prohibitive for large decoders
  - Add buffers, progressive sizing, combination of tree and PTs

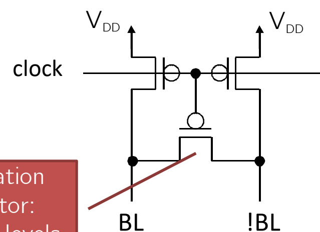
1080

Bit-line precharging: SRAM

static pull-up precharge



clocked precharge



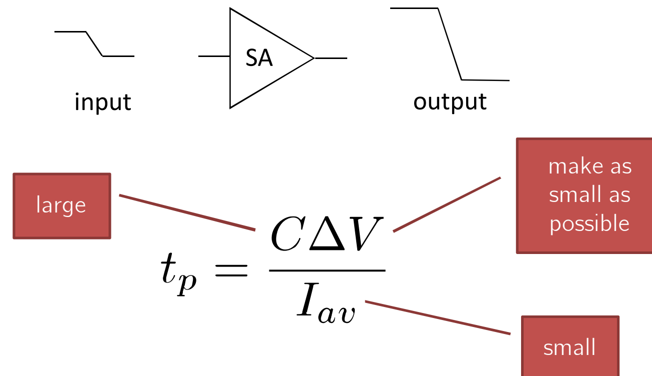
equalization transistor: equalizes levels on BL and !BL

- Static: no precharge clock required, but consumes static power (fight against bit-line discharge)
- Clocked: can use large precharge devices and bit line equalization much faster, but large clock load

1081

Sense amplifiers

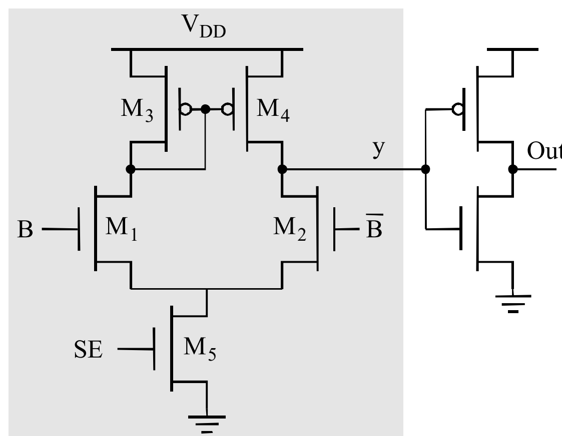
- Amplifies small swing on bit lines to full rail-to-rail swing needed at memory output



$$t_p = \frac{C\Delta V}{I_{av}}$$

1082

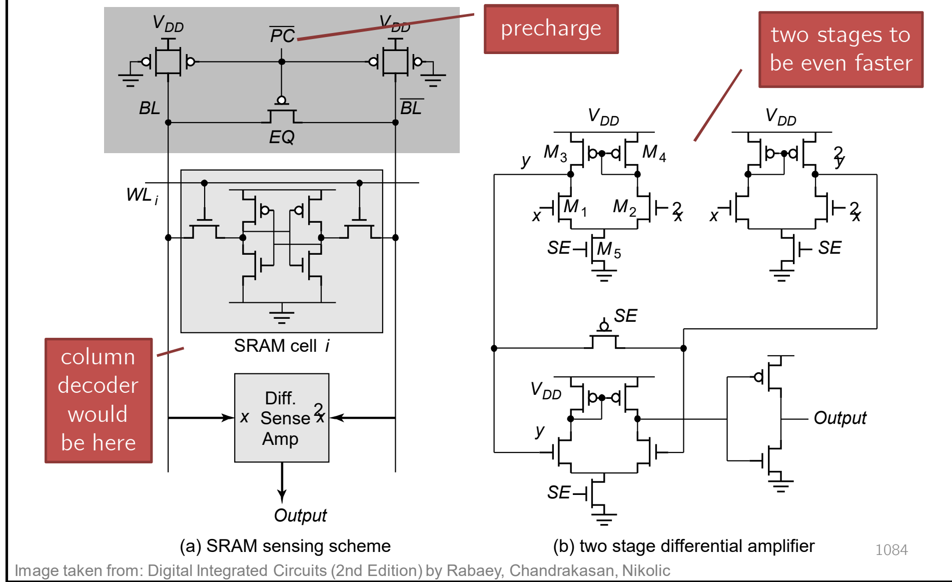
Differential sense amplifier



- Directly applicable to SRAMs

1083

Differential sensing (cont'd)

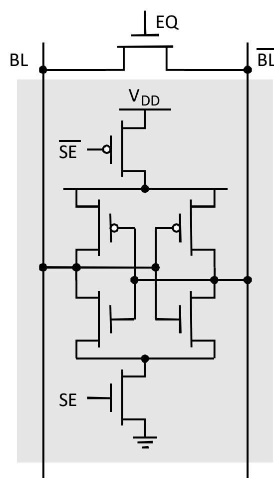


(a) SRAM sensing scheme (b) two stage differential amplifier

1084

Image taken from: Digital Integrated Circuits (2nd Edition) by Rabaey, Chandrakasan, Nikolic

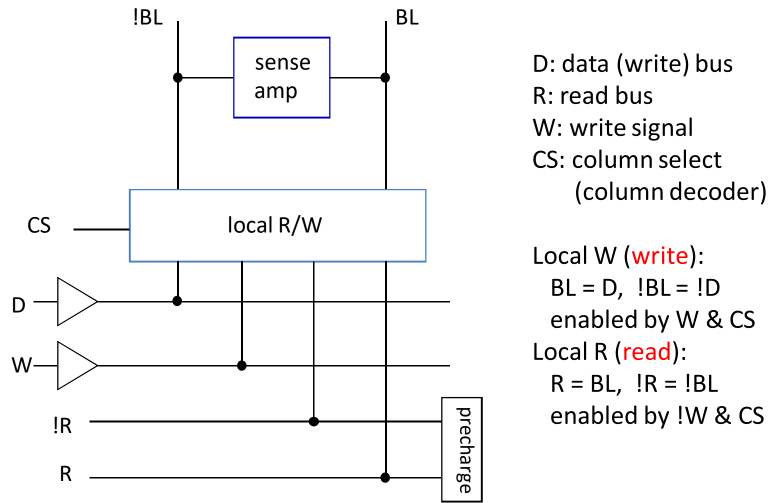
Latch-based sense amp



- EQ used to initialize latch in its meta-stable point
- Once adequate voltage established SE=1 enables sense amplifier
- Positive feedback quickly forces output to stable operating point

1085

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

1086

Trade-off noise for density and performance

Reliability and yield

1087

Reliability and yield

- Semiconductor memories trade-off noise margin for bit density and performance
  - Sensitive to noise (crosstalk, supply noise, etc.)
- High-density and large die size causes yield problems:

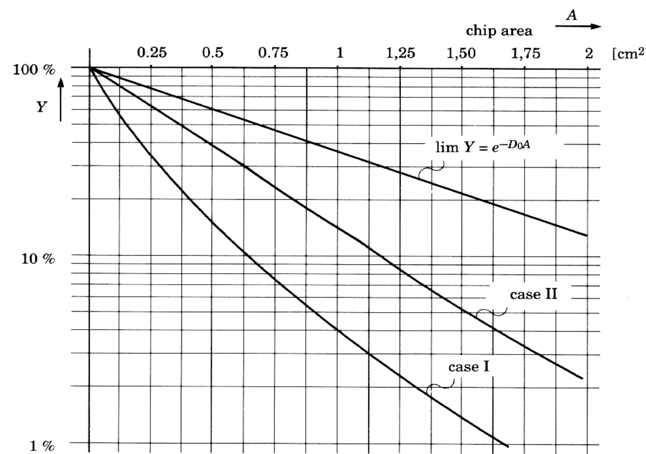
A = chip area D = defect density

$$\text{Yield} = 100 \frac{\#\text{ of good chips/waver}}{\#\text{ of chips/wafer}} \qquad Y = \left(\frac{1 - e^{-AD}}{AD}\right)^2$$

- Increase yield using error correction and redundancy

1088

Yield vs. chip area and process

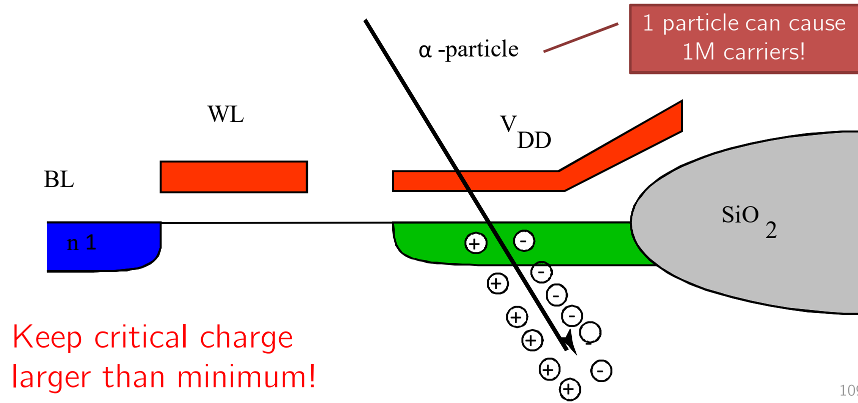


- Yield curves at different stages of process maturity [Veendricks92]

1089

Soft errors

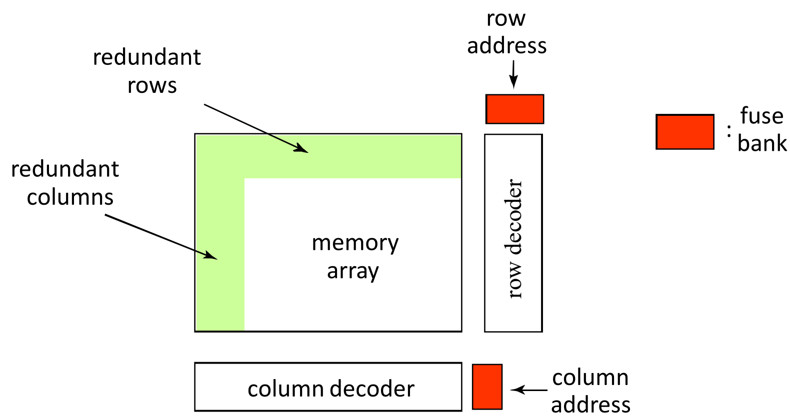
- Ionizing radiation can cause non-recurrent and nonpermanent errors in memories



Keep critical charge larger than minimum!

1090

Redundancy in memory structure



- Replace bad row or column with "spare" → set by fuse bank
- Solves problem at manufacturing time, but not soft errors

1091

Redundancy and error correction

- Suitable for preventing soft errors

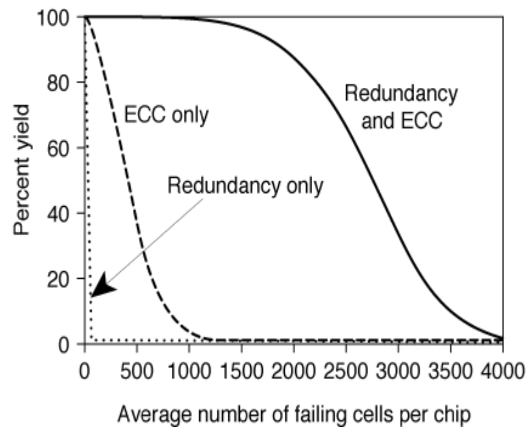


Image taken from: Digital Integrated Circuits (2nd Edition) by Rabaey, Chandrakasan, Nikolic

1092

Example: (7,4,3) Hamming code

- Consider 4-bit number $[B_3,B_5,B_6,B_7]$
- Add 3 parity check bits $[P_1,P_2,P_4]$
- Chose parity bits as follows:

$$P_1 \oplus B_3 \oplus B_5 \oplus B_7 = 0 \quad 1$$
$$P_2 \oplus B_3 \oplus B_6 \oplus B_7 = 0 \quad 1$$
$$P_4 \oplus B_5 \oplus B_6 \oplus B_7 = 0 \quad 0$$

011 = 3

- Error in bit B_3 would cause 1st and 2nd parity check to fail
- Implies that bit 3 is in error → can be corrected (flipped)

1093