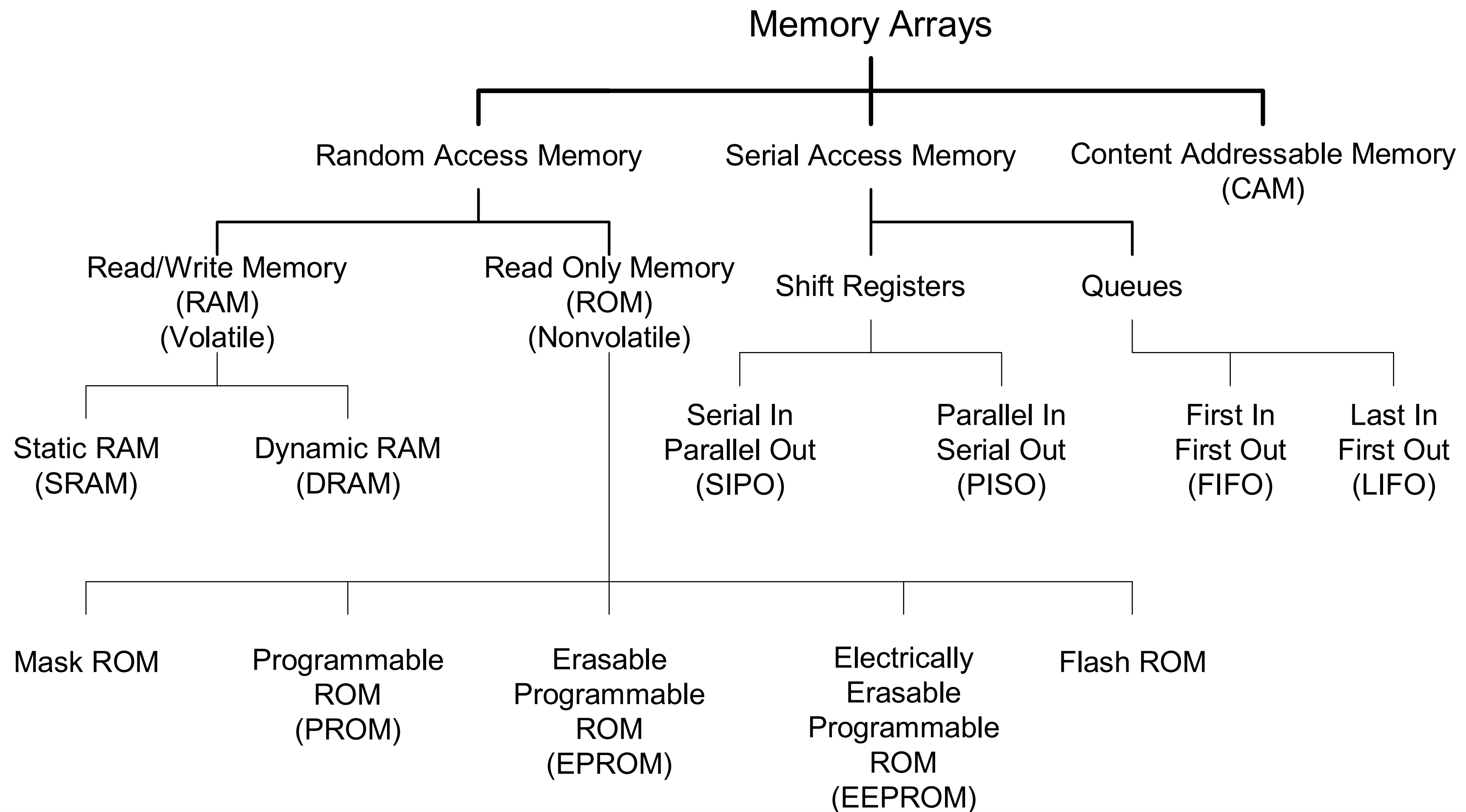
Array Structured Memories

STMicro/Intel

UCSD CAD LAB

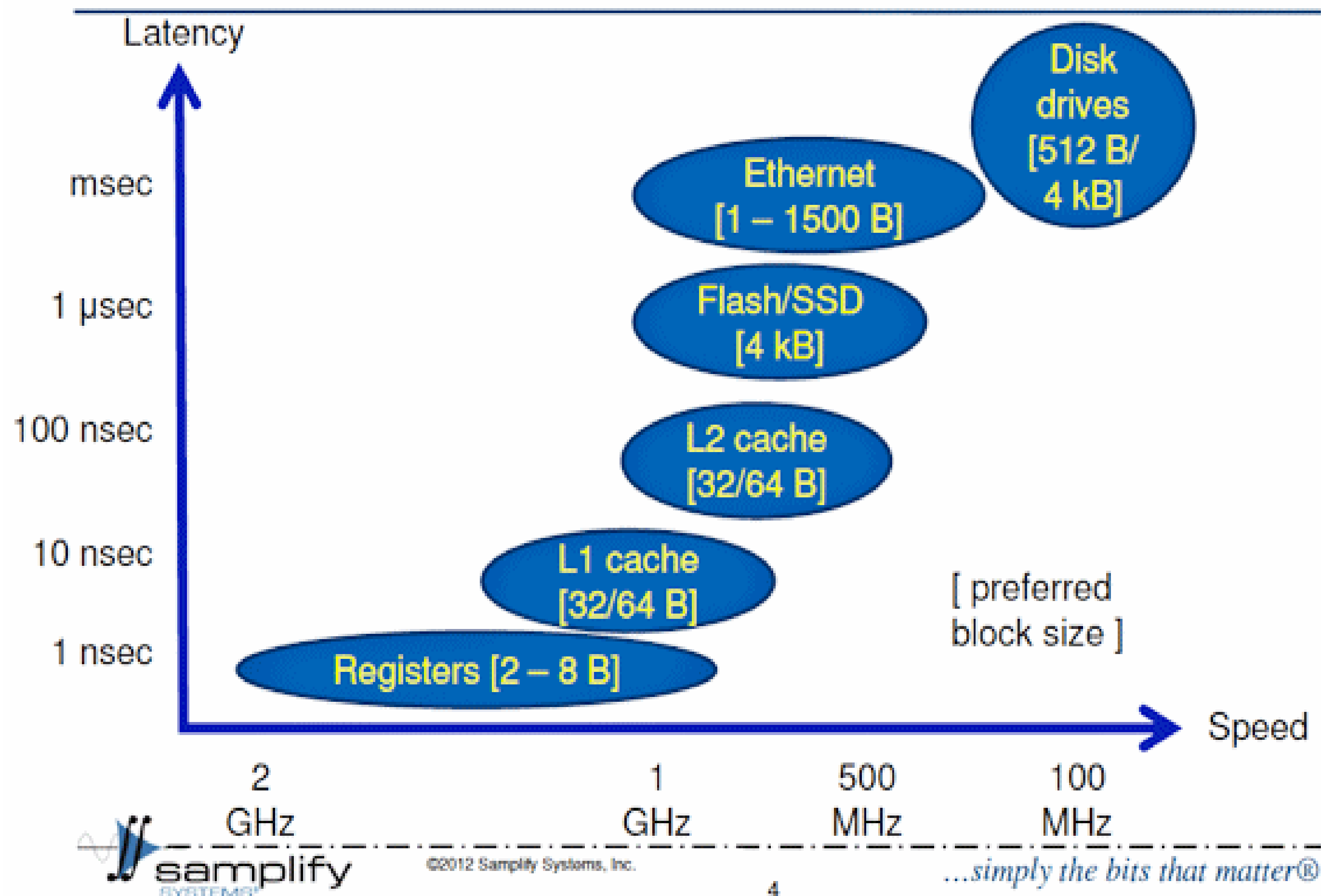
Weste Text

Memory Arrays



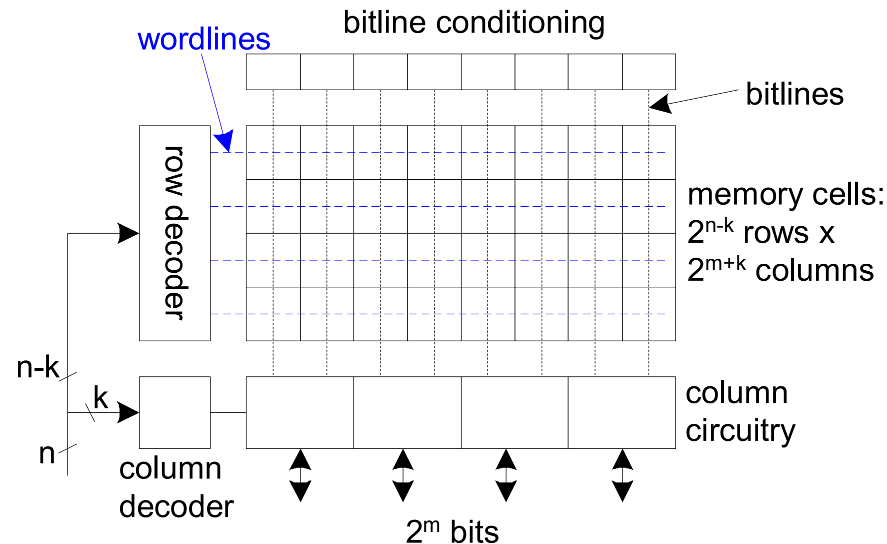
Feature Comparison Between Memory Types

Memory Type	SRAM	DRAM	Flash
Speed	Very fast	Fast	Very slow
Density	Low	High	Very high
Endurance	Better	Better	Poor
Power	Low	High	Very low
Refresh	No	Yes	No
Retention	Volatile	Volatile	Non-volatile
Scalable	Good	Bad	Good
Mechanism	Bi-stable latch	Capacitor	FN tunneling, HCI


The Memory Hierarchy
Latency
msec
1 µsec
100 nsec
10 nsec
1 nsec
Disk drives [512 B/ 4 kB]
Ethernet [1 – 1500 B]
Flash/SSD [4 kB]
L2 cache [32/64 B]
L1 cache [32/64 B]
Registers [2 – 8 B]
[preferred block size]
Speed
2 GHz
1 GHz
500 MHz
100 MHz
samplify SYSTEMS
©2012 Samplify Systems, Inc.
4
...simply the bits that matter®

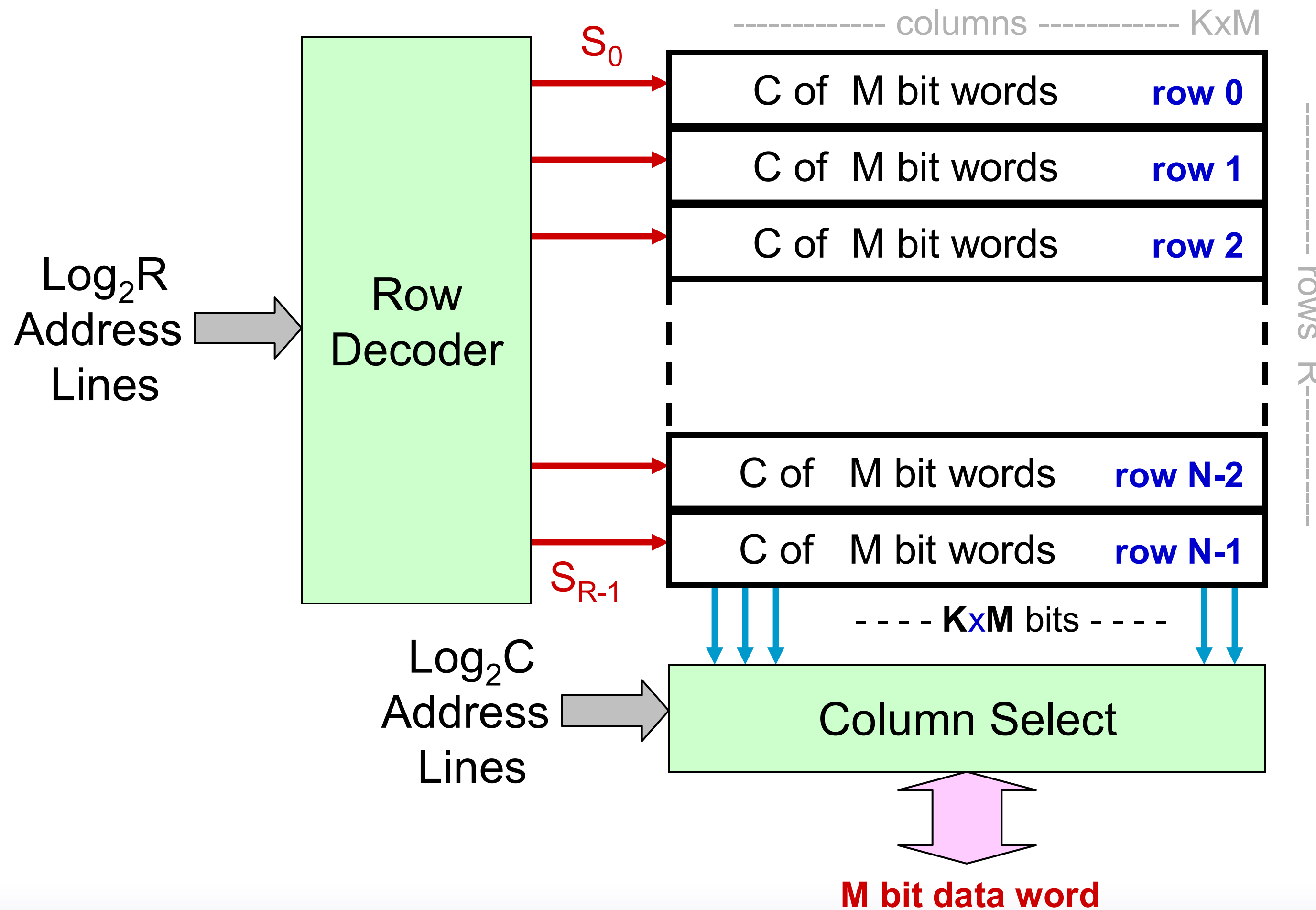
Array Architecture

- 2^n *words* of 2^m *bits* each
- If n >> m, fold by 2^k into fewer *rows* of more *columns*

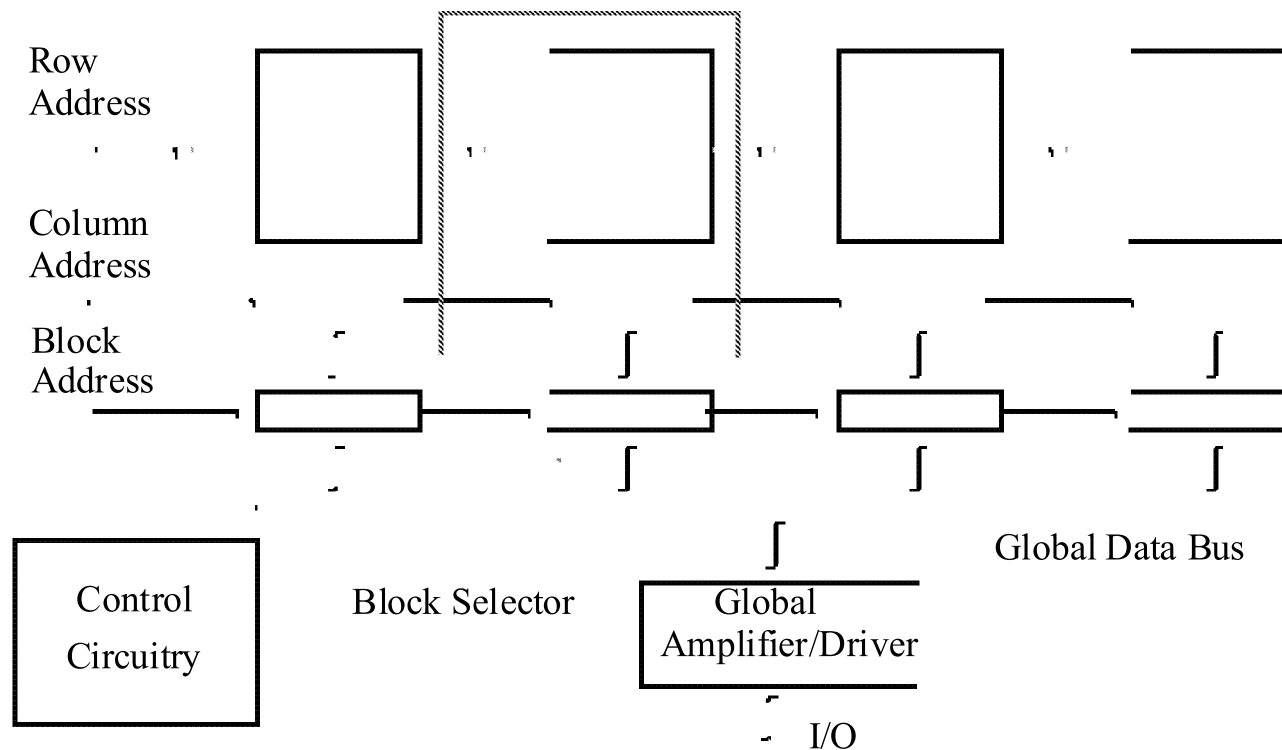


- Good regularity – easy to design
- Very high density if good cells are used

Memory - Real Organization



Hierarchical Memory Architecture

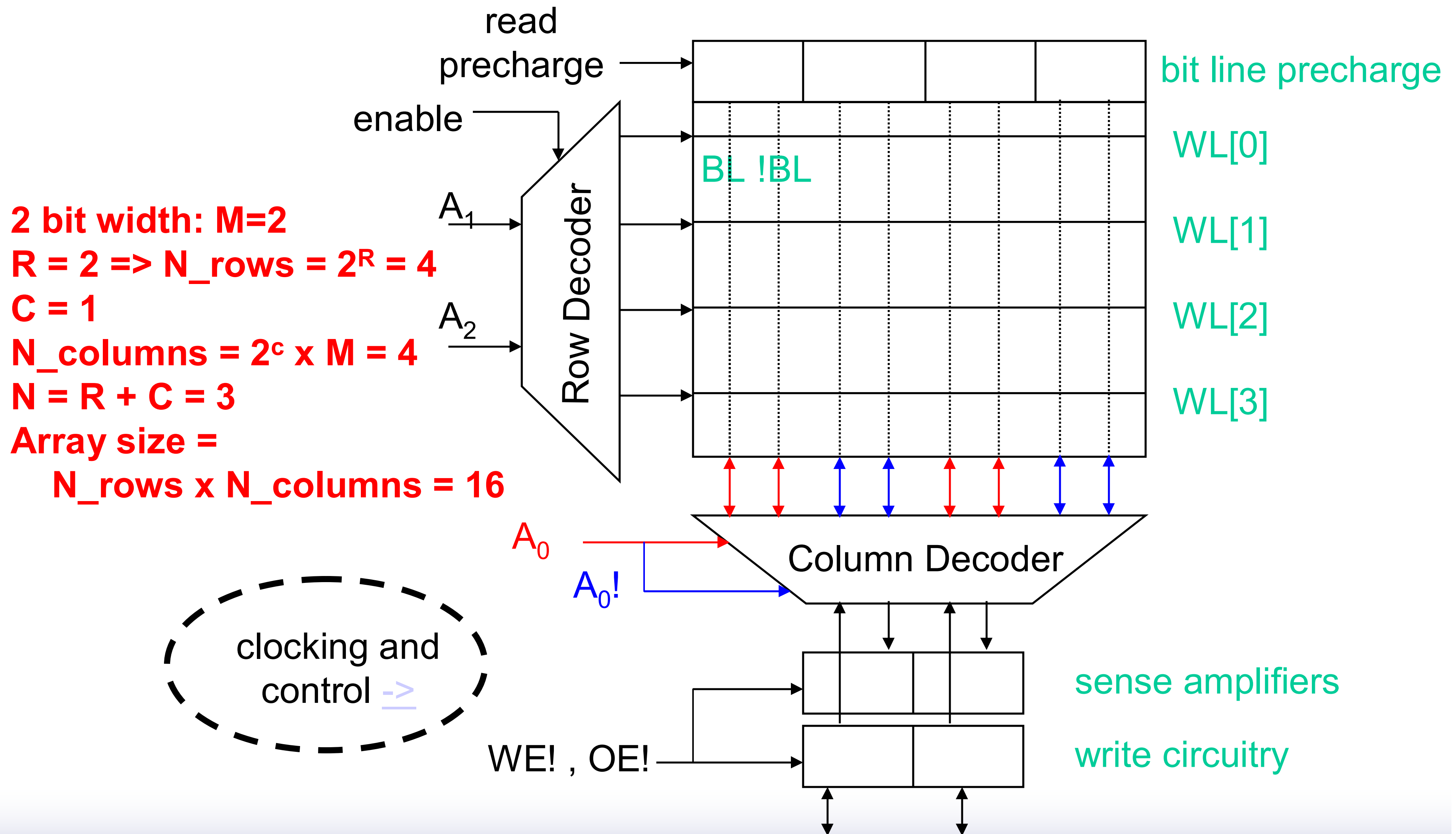


Advantages:
1. Shorter wires within blocks
2. Block address activates only 1 block => power savings

Array Organization Design Issues

- aspect ratio should be relative square
  - Row / Column organisation (matrix)
  - R = log2(N_rows); C = log2(N_columns)
  - R + C = N (N_address_bits)
- number of rows should be power of 2
  - number of bits in a row need not be…
- sense amplifiers to speed voltage swing
- 1 -> 2^R row decoder
- 1 -> 2^C column decoder
  - M column decoders (M bits, one per bit)
    - M = output word width

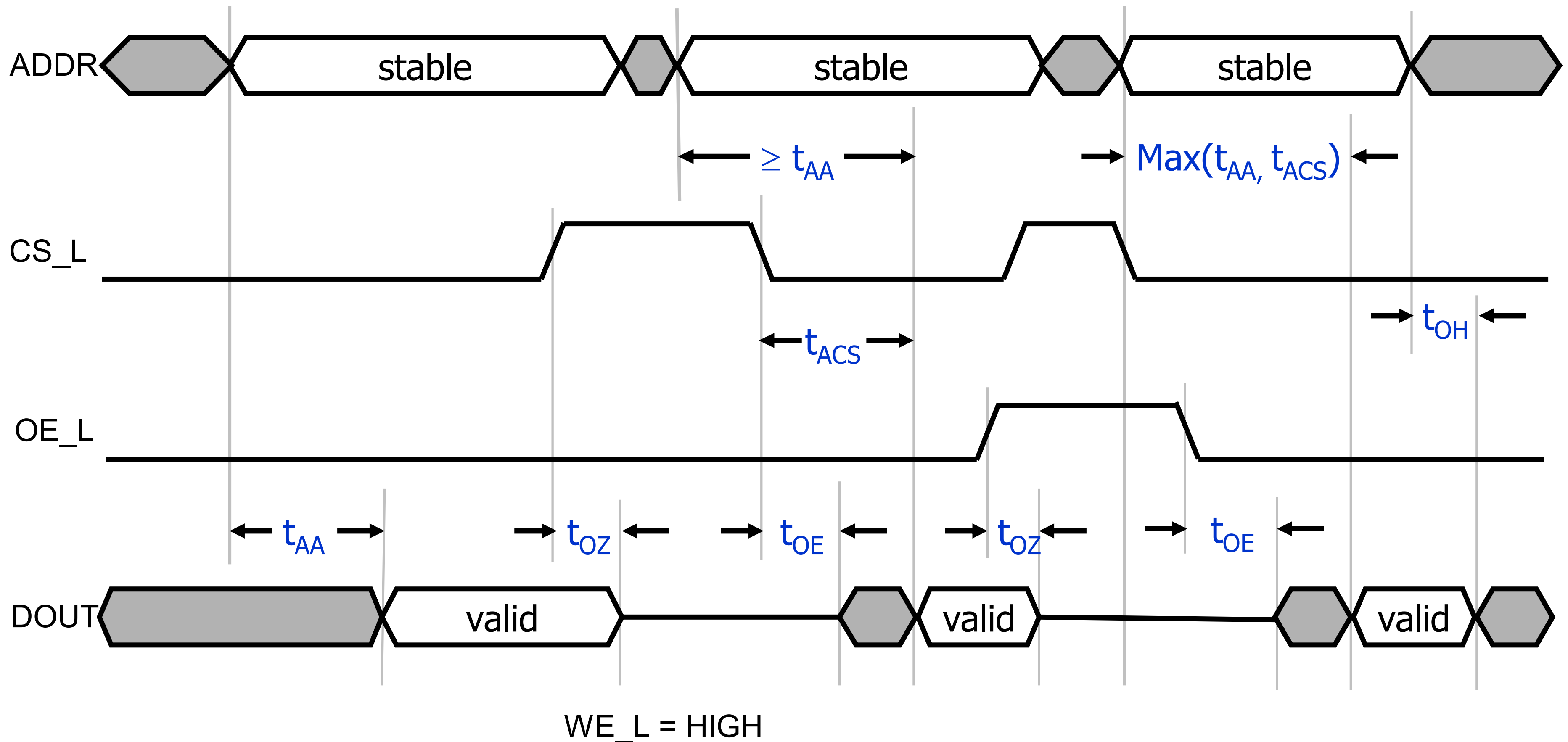
Simple 4x4 SRAM Memory



SRAM Read Timing (typical)

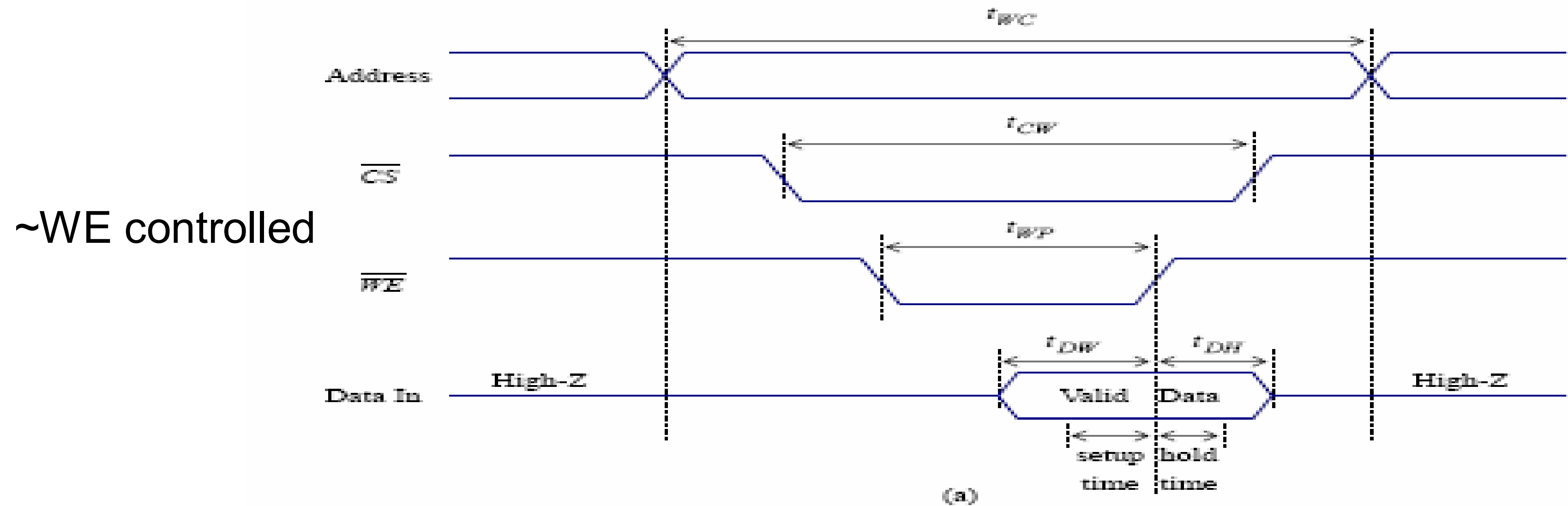
- **t_{AA} (access time for address):** time for stable output after a change in address.
- **t_{ACS} (access time for chip select):** time for stable output after CS is asserted.
- **t_{OE} (output enable time):** time for low impedance when OE and CS are both asserted.
- **t_{OZ} (output-disable time):** time to high-impedance state when OE or CS are negated.
- **t_{OH} (output-hold time):** time data remains valid after a change to the address inputs.

SRAM Read Timing (typical)

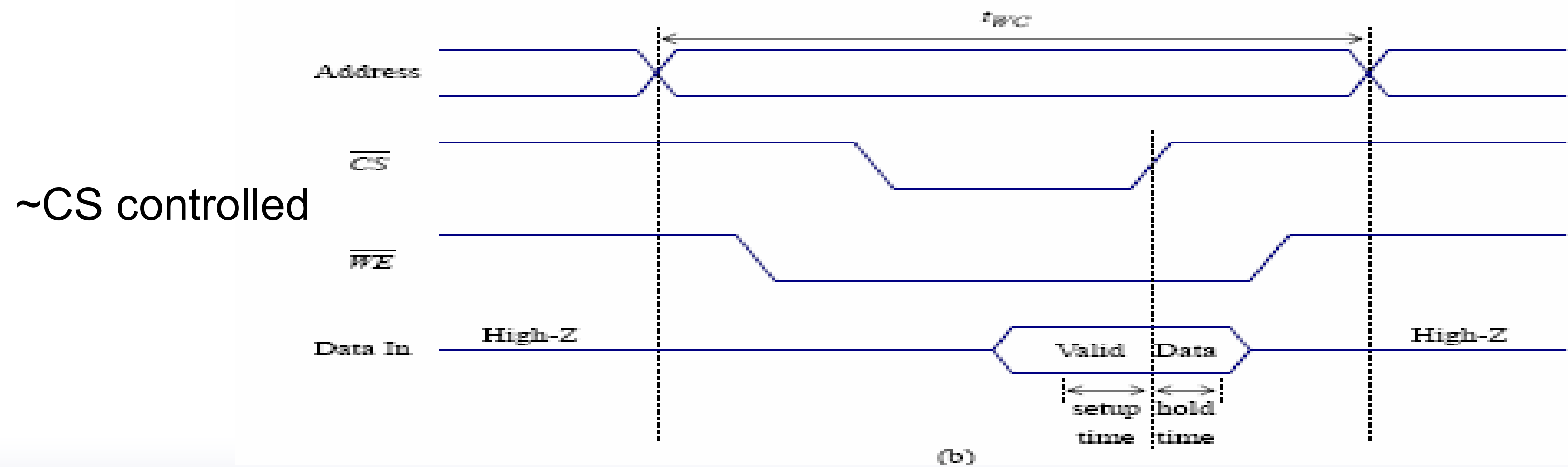


SRAM write cycle timing

~WE controlled



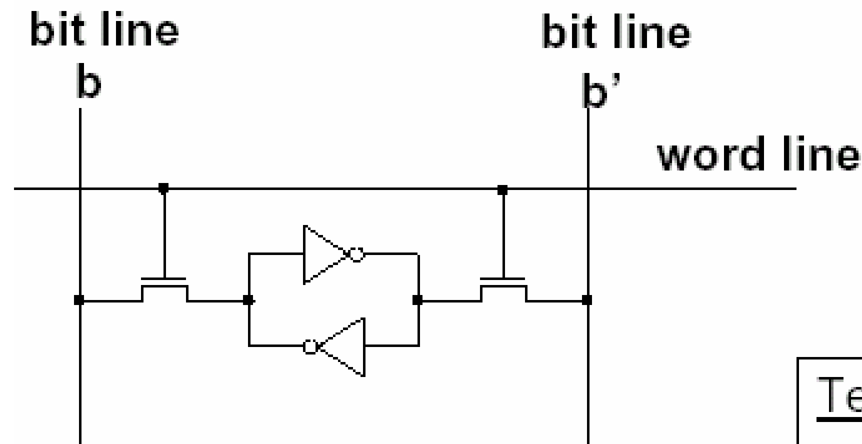
~CS controlled

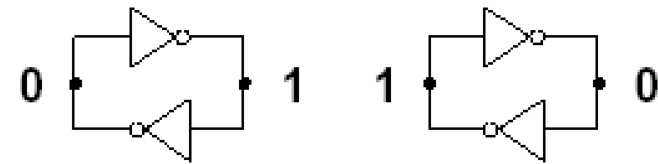


SRAM Cell Design

- Memory arrays are large
  - Need to optimize cell design for area and performance
  - Peripheral circuits can be complex
    - 60-80% area in array, 20-40% in periphery
- Classical Memory cell design
  - 6T cell full CMOS
  - 4T cell with high resistance poly load
  - TFT load cell

Anatomy of the SRAM Cell





Terminology:

bit line: carries data

word line: used for addressing

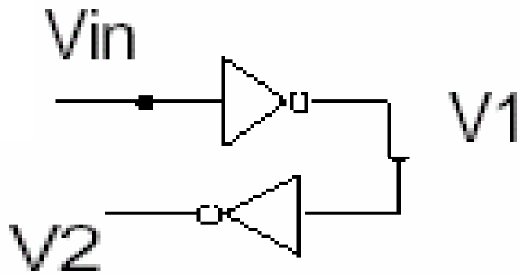
Write:

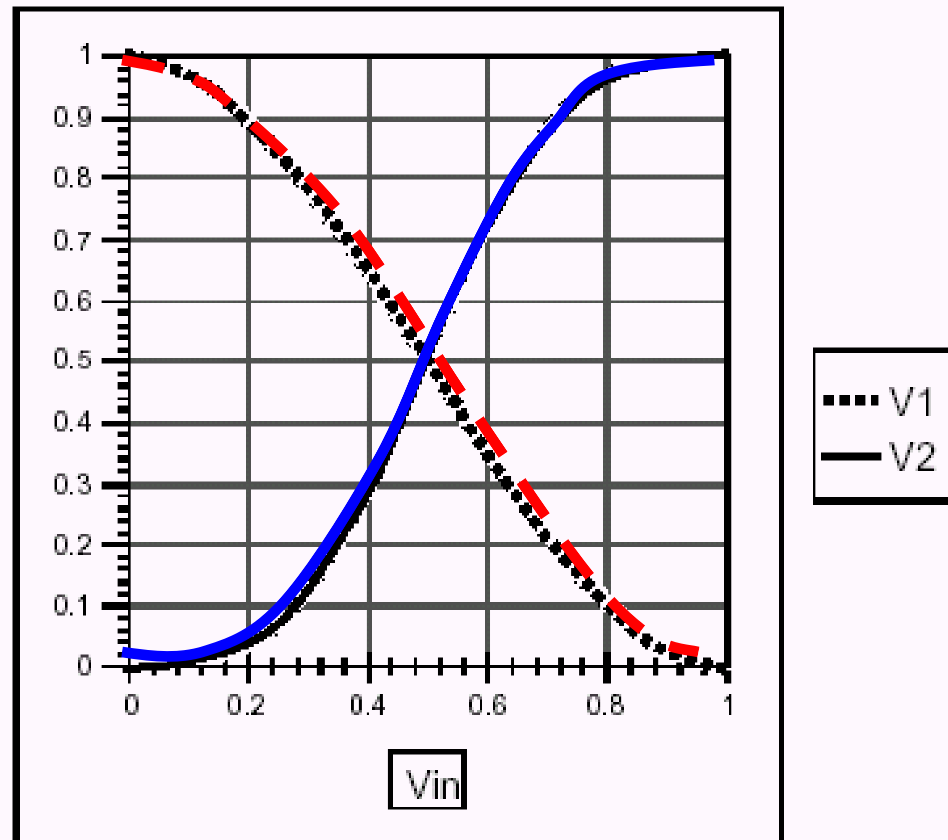
- set bit lines to new data value
  - b' = ~b
- raise word line to "high"
  - sets cell to new state
  - Low impedance bit-lines

Read:

- set bit lines high
- set word line high
- see which bit line goes low
- High impedance bit lines

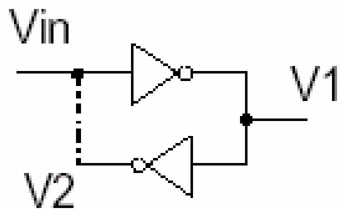
SRAM Cell Operating Principle

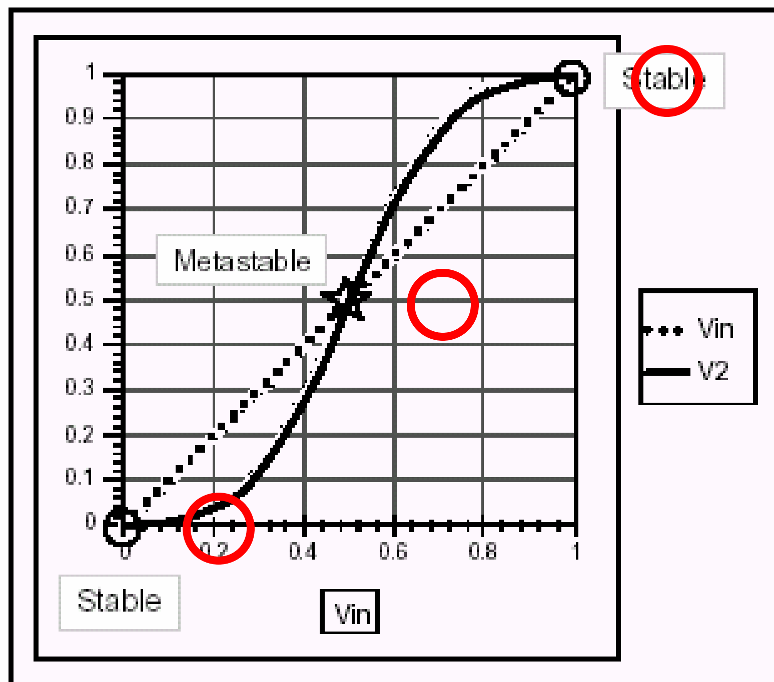




- Inverter Amplifies
  - Negative gain
  - Slope < –1 in middle
  - Saturates at ends
- Inverter Pair Amplifies
  - Positive gain
  - Slope > 1 in middle
  - Saturates at ends

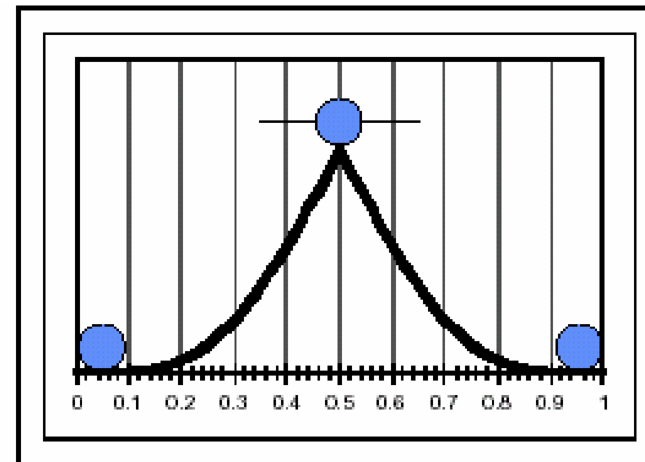
Bistable Element





Stability

- **Require Vin = V2**
- **Stable at endpoints**
  recover from pertubation
- **Metastable in middle**
  Fall out when perturbed

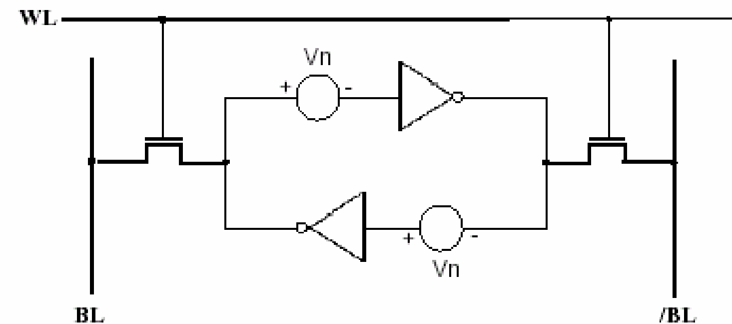


Ball on Ramp Analogy

Cell Static Noise Margin

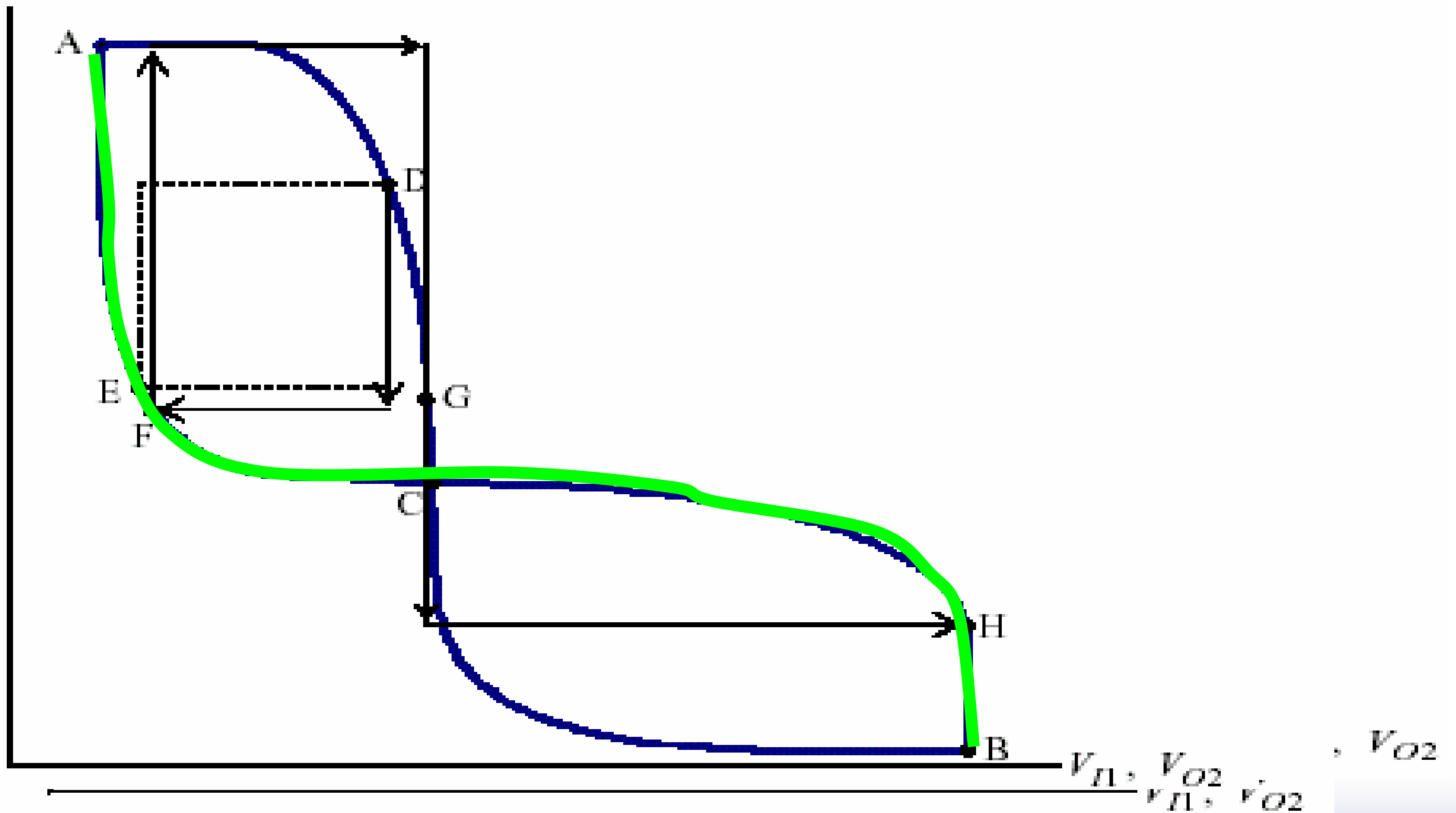
Cell state may be disturbed by

- **DC**
  - **Layout pattern offset**
  - **Process mismatches**
    - non-uniformity of implantation
    - gate pattern size errors
- **AC**
  - **Alpha particles**
  - **Crosstalk**
  - **Voltage supply ripple**
  - **Thermal noise**



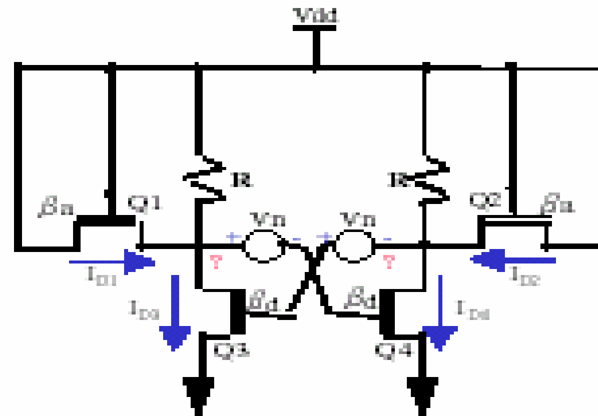
SNM (static noise margin) = Maximum Value of V_n *not* flipping cell state

SNM: Butterfly Curves



SNM for Poly Load Cell

- Q1, Q2, Q3 : Saturation Region
- Q4 : Linear Region



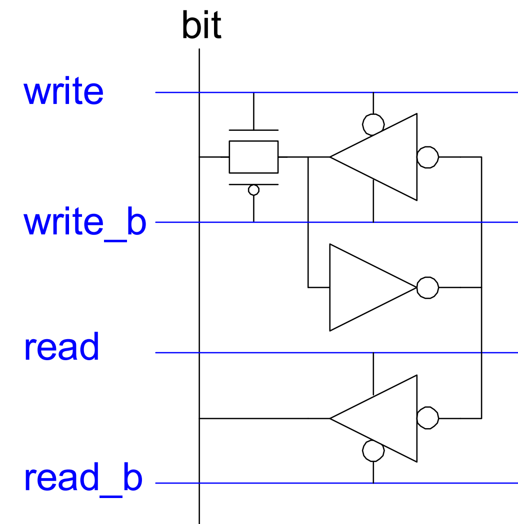
Results

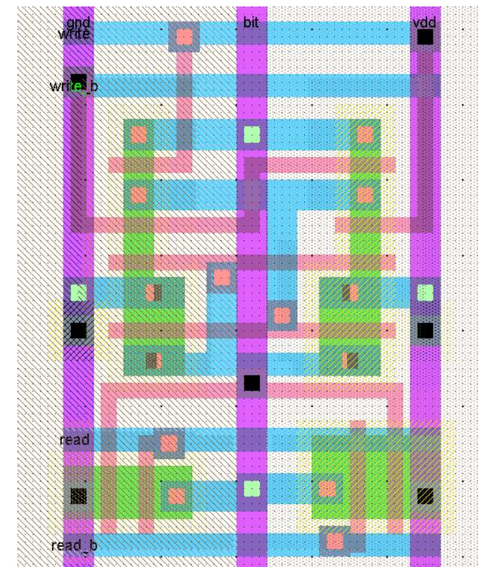
$$SNM\ (Vn_{(MAX)}) = \frac{\sqrt{\gamma} - 1}{\sqrt{\gamma} + 1} Vth + \frac{\gamma + 1 - \sqrt{2\gamma^{3/2} + \gamma + 1}}{\gamma(\sqrt{\gamma} + 1)} (Vdd - Vth)$$

where γ (Cell Ratio) $= \beta d / \beta a$

12T SRAM Cell

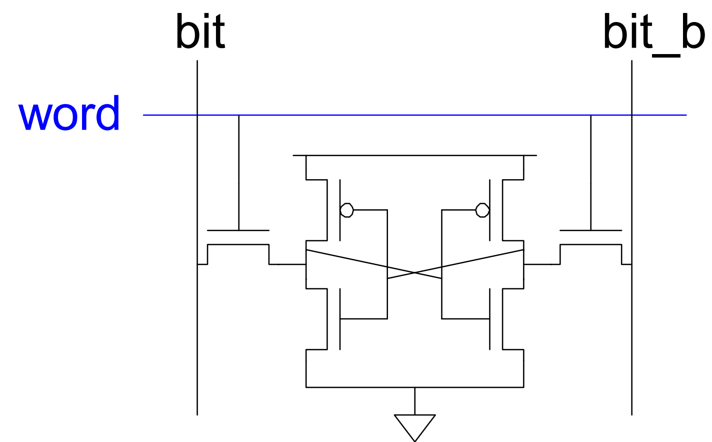
- Basic building block: SRAM Cell
  - 1-bit/cell (noise margin again)
- 12-transistor (12T) SRAM cell
  - Latch with TM-gate write
  - Separately buffered read



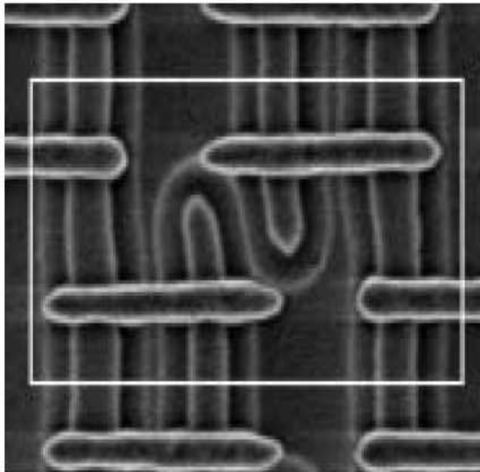


6T SRAM Cell

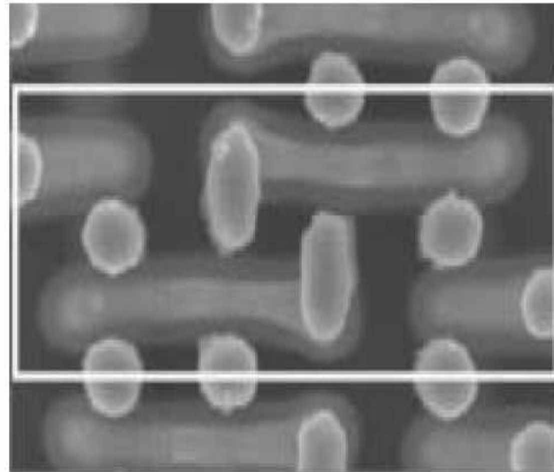
- Cell size accounts for most of array size
  - Reduce cell size at cost of complexity/margins
- 6T SRAM Cell
- Read:
  - Precharge bit, bit_b
  - Raise wordline
- Write:
  - Drive data onto bit, bit_b
  - Raise wordline



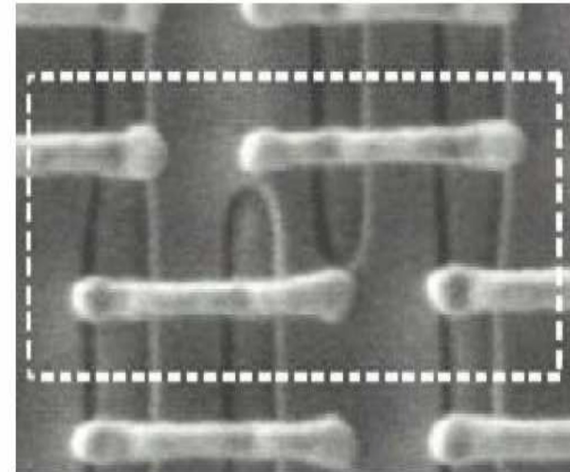
SRAM Design



TI 65nm:
$0.46 \times 1.06 um^2$



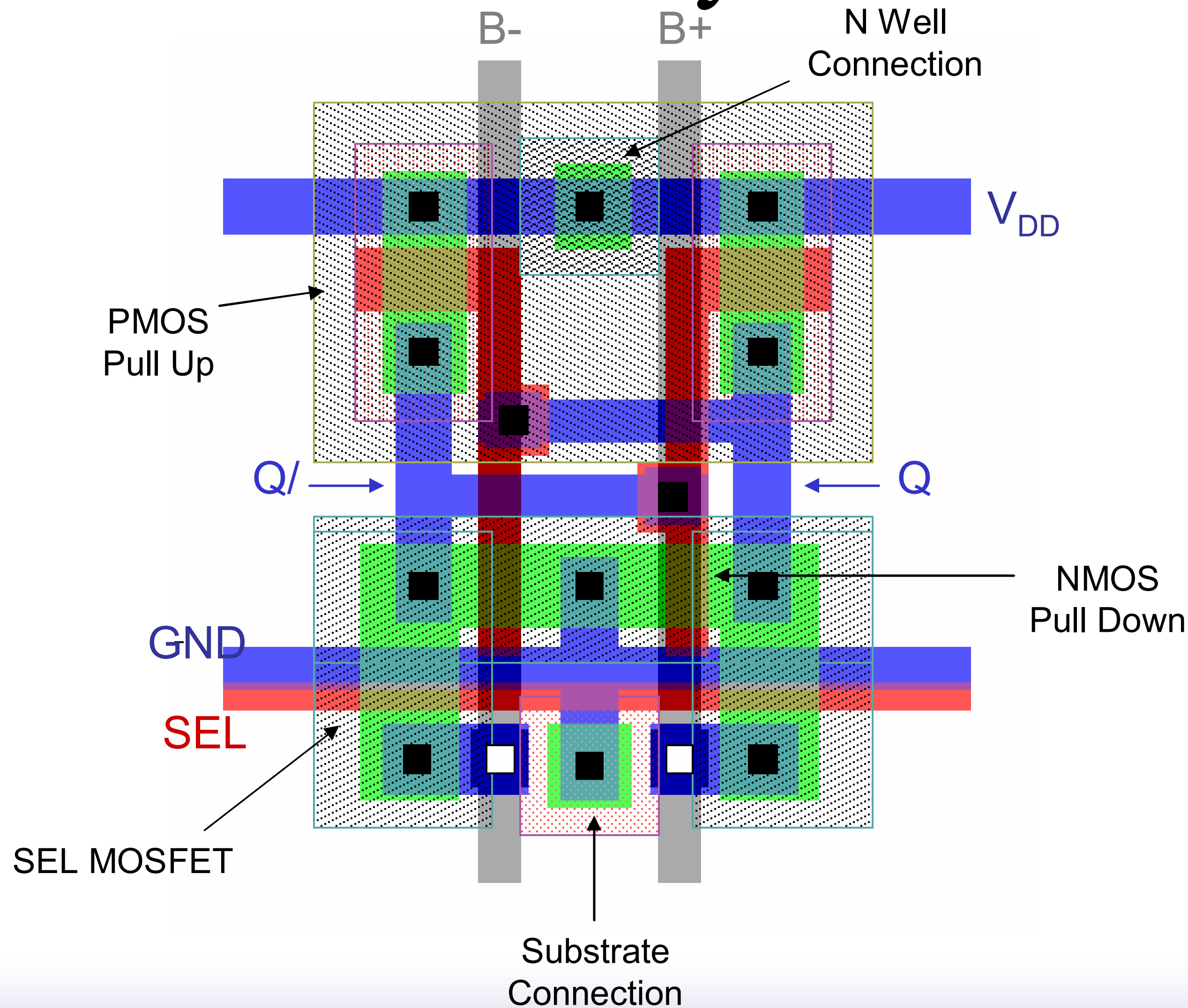
IBM 65nm:
$0.41 \times 1.25 um^2$



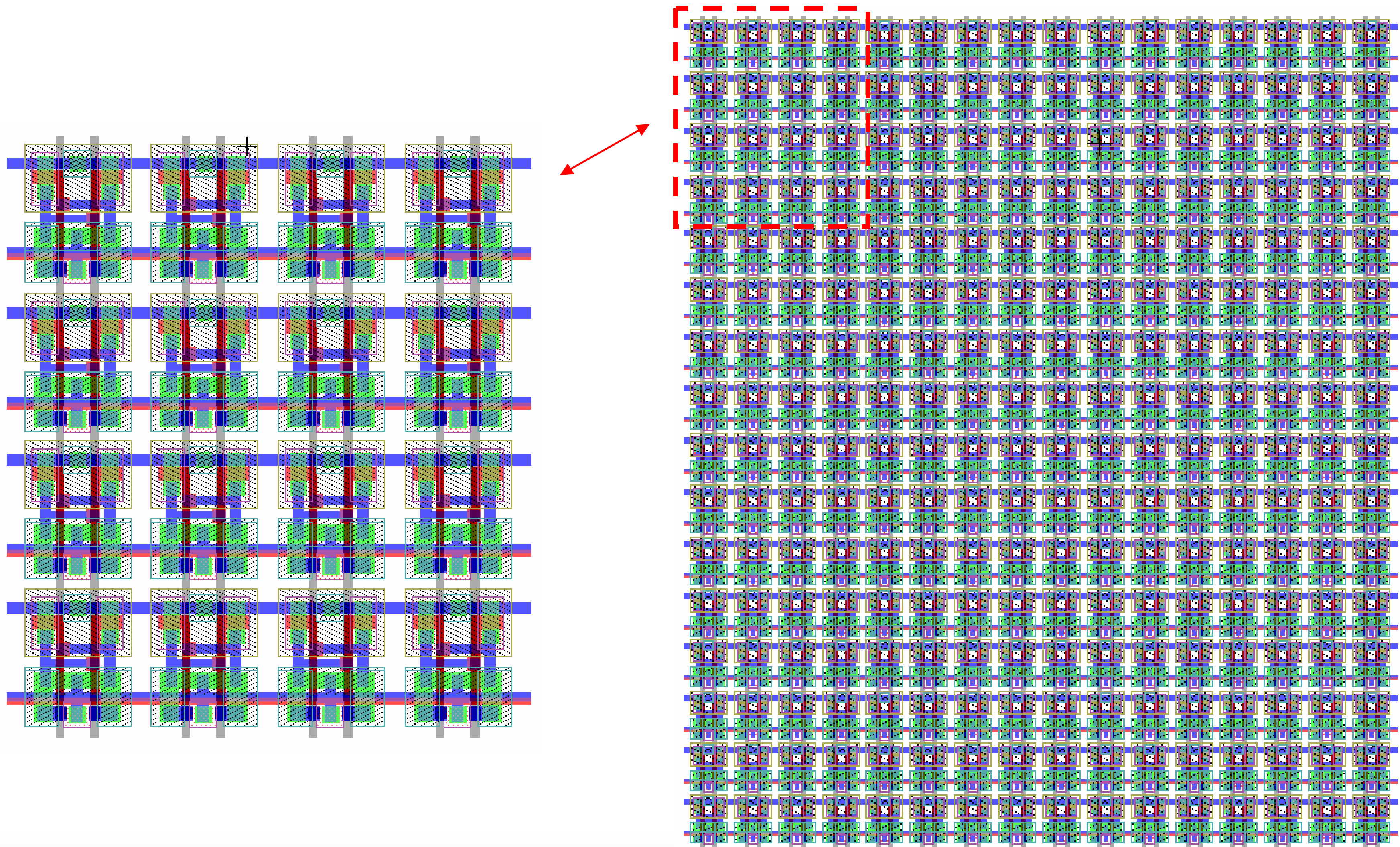
Intel 65nm:
$0.46 \times 1.24 um^2$

* Figures courtesy A. Chatterjee et al., P. Bai et al., and Z. Luo et al., Int. Electron Device Meeting Tech. Digest, 2004

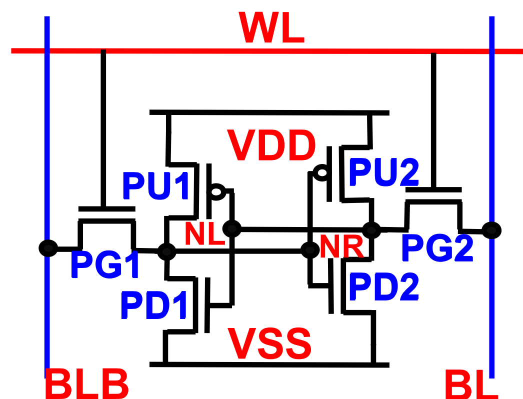
Vertical 6T Cell Layout



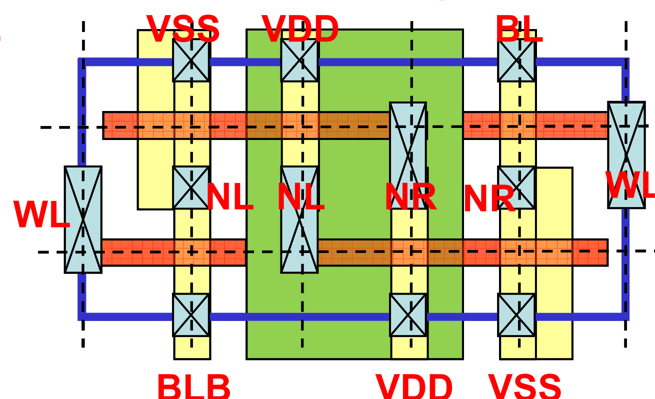
6T SRAM Array Layout



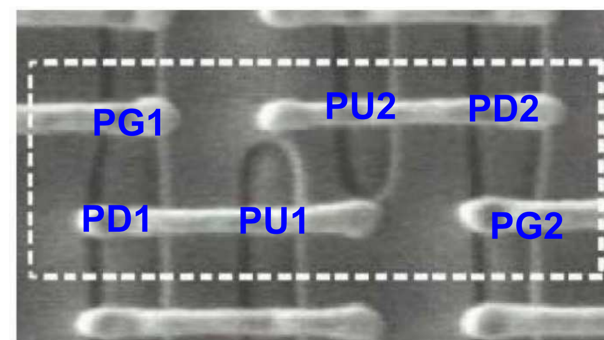
SRAM Bitcell Design



Schematic



Layout



Micrograph

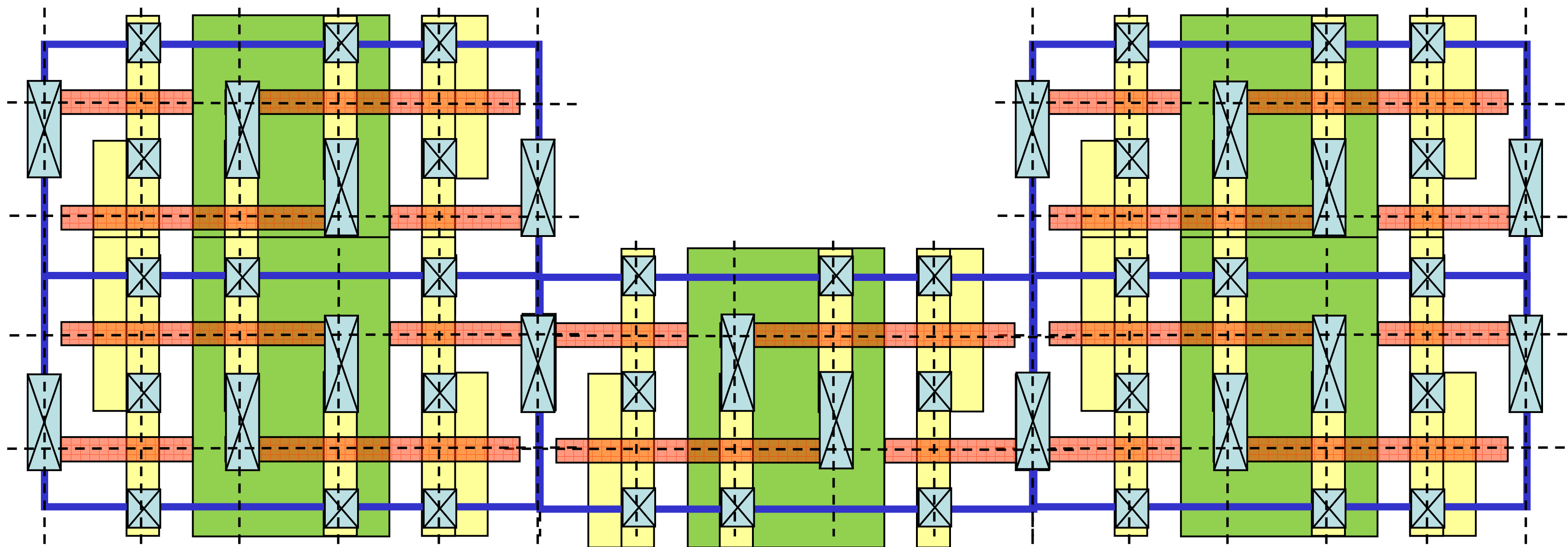
❑ Requirements of SRAM bitcell design

- **Stable read operation**: Do not disturb data when reading
- **Stable write operation**: Must write data within a specified time
- **Stable data retention**: Data should not be lost

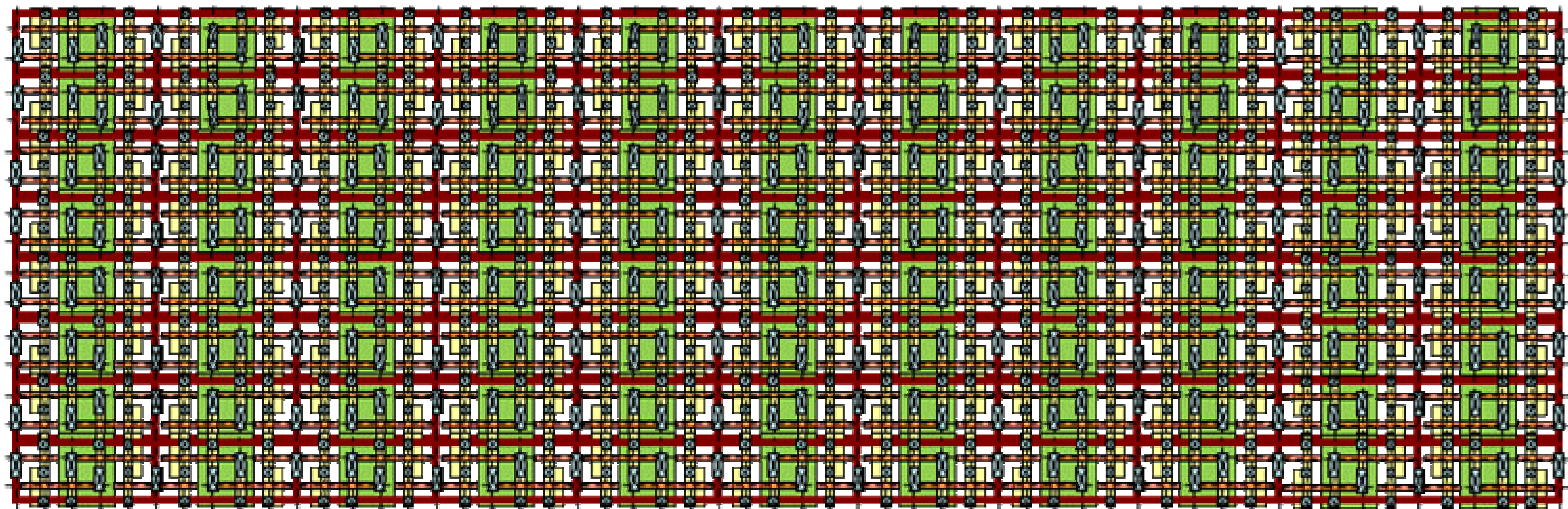
❑ Typical transistor sizing

- Cell ratio (= I(PD) / I(PG)) = 1.5 ~ 2.5
- Pull-up ratio (= I(PU) / I(PG)) = 0.5

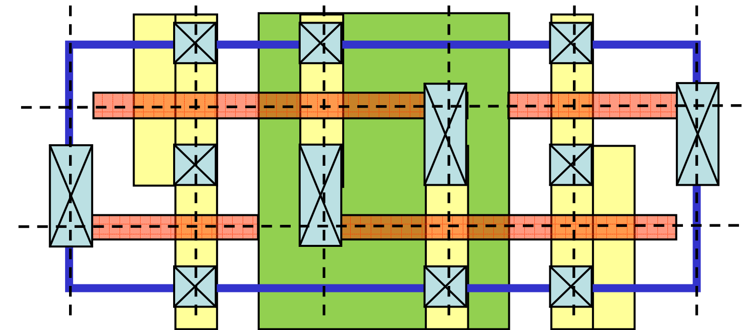
Bitcell Assembly



Bitcell Array



Detailed SRAM Bitcell Layout

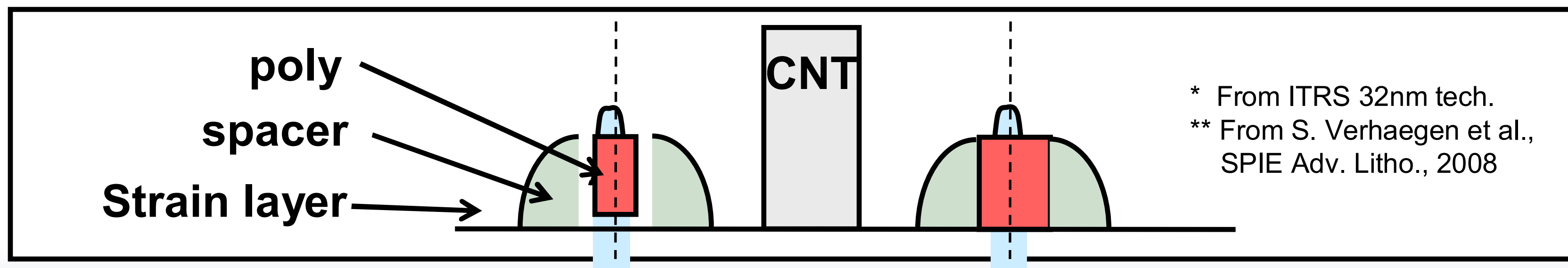


- Vertical: 2 poly pitch
- Horizontal: 5 contact pitch
- Poly-to-contact space > overlay + spacer + strain_layer + CD_control

  (6.4nm*) (≈ 8nm**) (≈10nm**) (≈ 2.6nm*) = 27nm
- **1 poly pitch** = 2 poly_to_contact + poly_width + contact_width

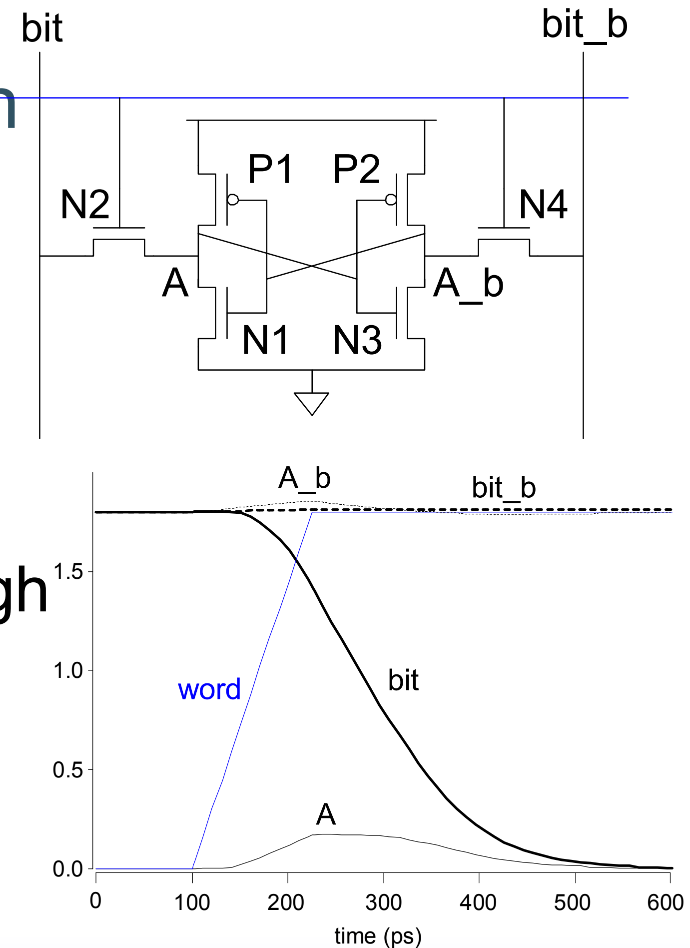
  ≈ 54 + 32 + 45** = 131 nm

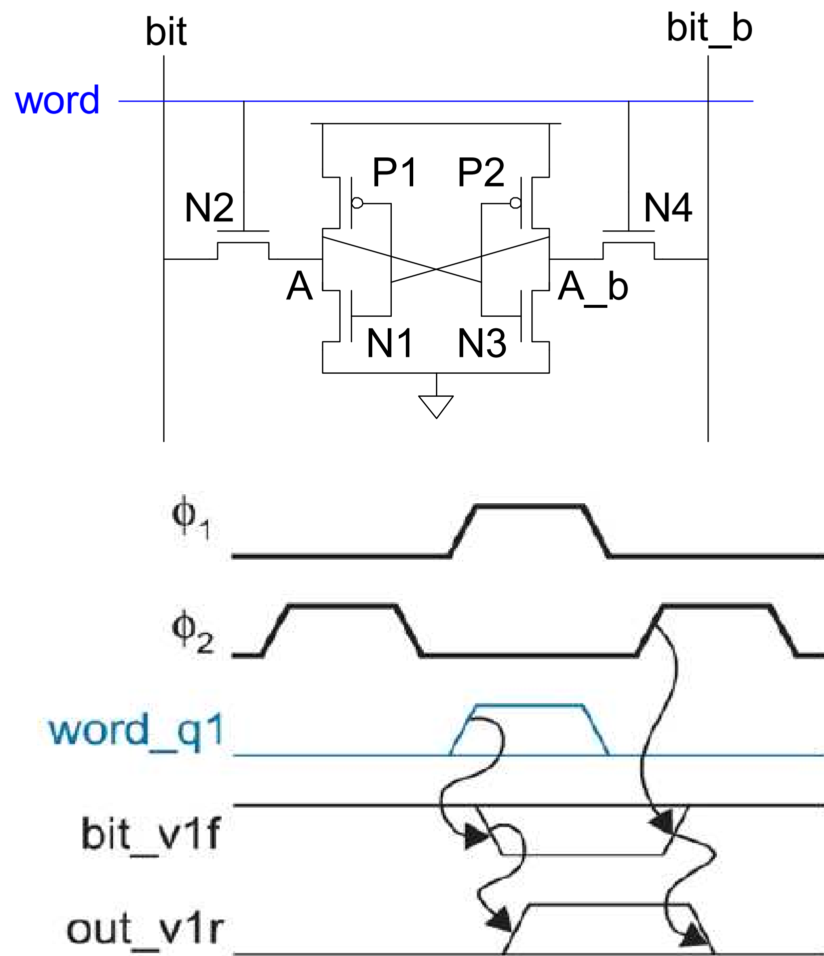
A pitch is a multiple of a **drawing grid** for fine-grain pattern placement

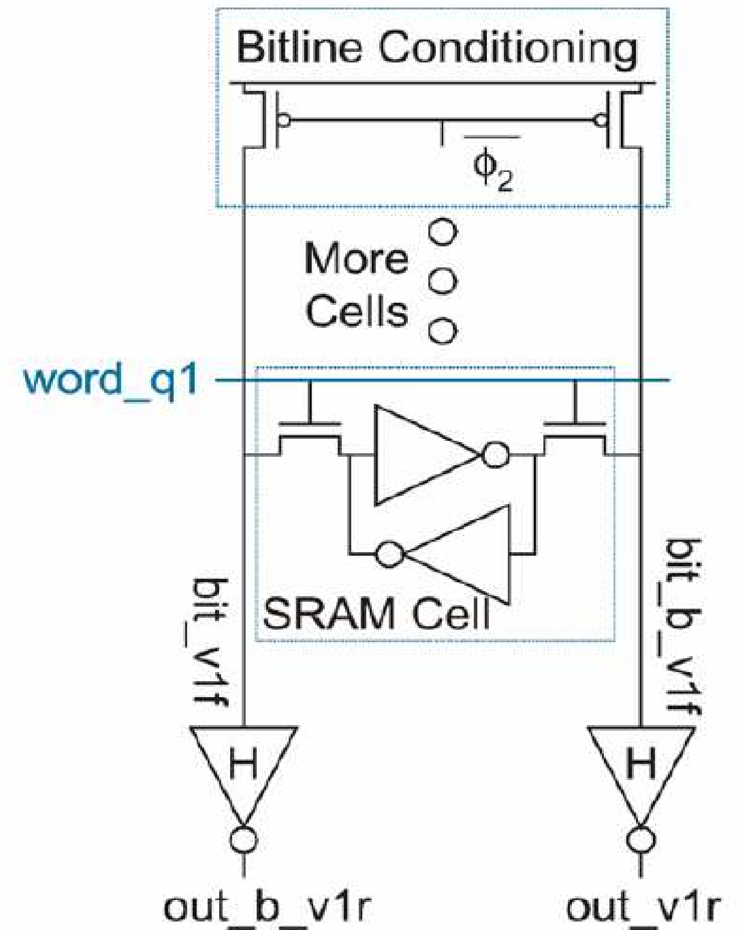


SRAM Read

- Precharge both bitlines high
- Then turn on wordline
- One of the two bitlines will
  - be pulled down by the cell
- Ex: A = 0, A_b = 1
  - bit discharges, bit_b stays high
  - But A bumps up slightly
- *Read stability*
  - A must not flip
  - N1 >> N2

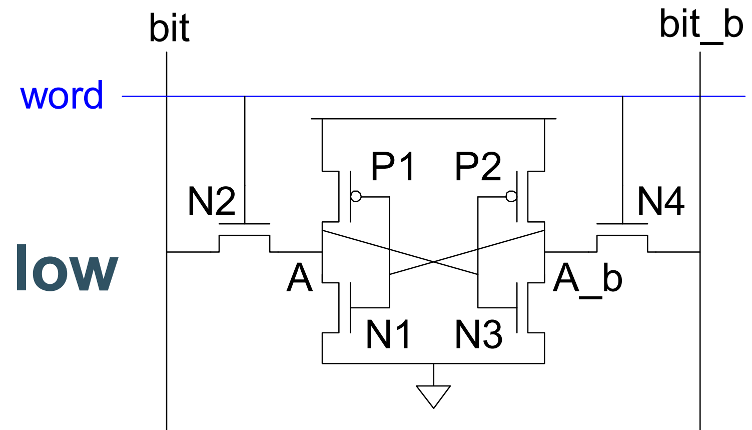




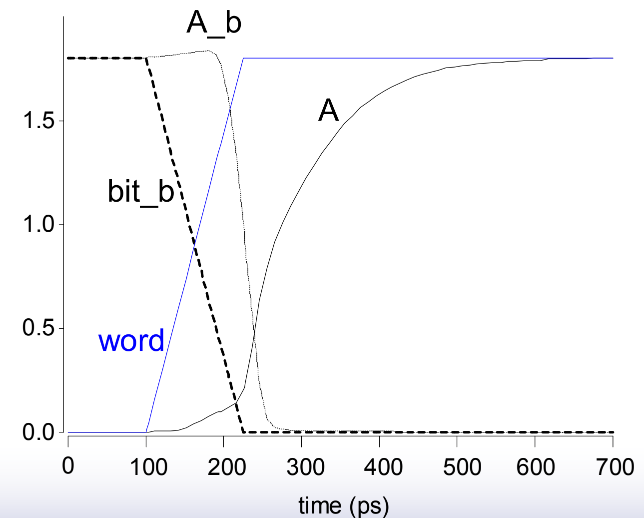


SRAM Read, 0 is stored in the cell

SRAM Write

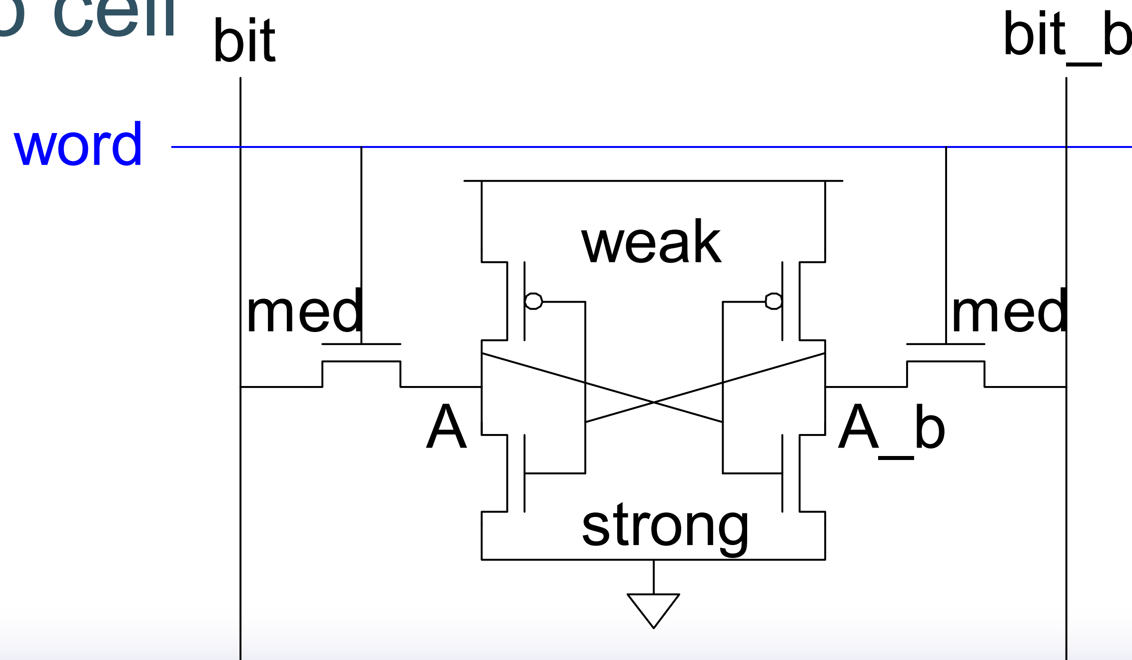


- Drive one bitline high, other low
- Then turn on wordline
- Bitlines overpower cell
- Ex: A = 0, A_b = 1, bit = 1, bit_b = 0
  - Force A_b low, then A rises high
- *Writability*
  - Must overpower feedback
  - P2 << N4 to force A_b low,
  - N1 turns off, P1 turns on,
  - raise A high as desired



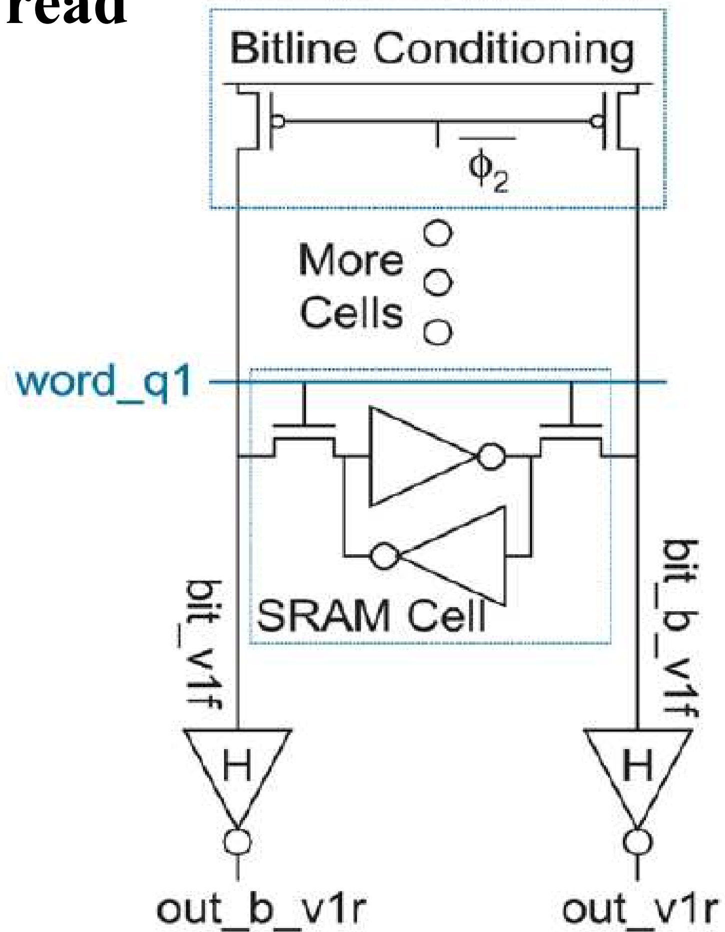
SRAM Sizing

- High bitlines must not overpower inverters during reads
- But low bitlines must write new value into cell

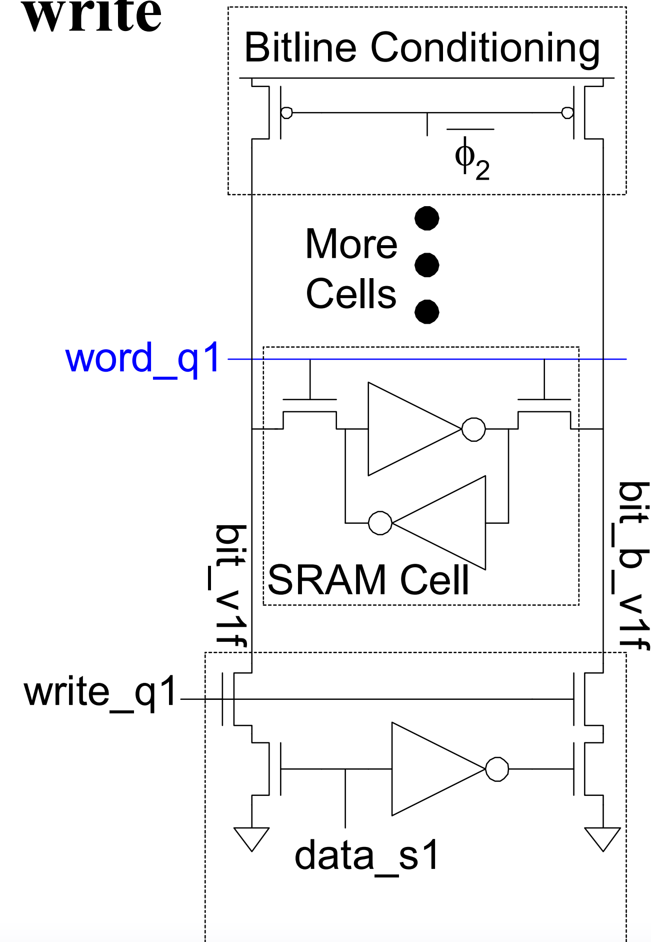


SRAM Column Example

read

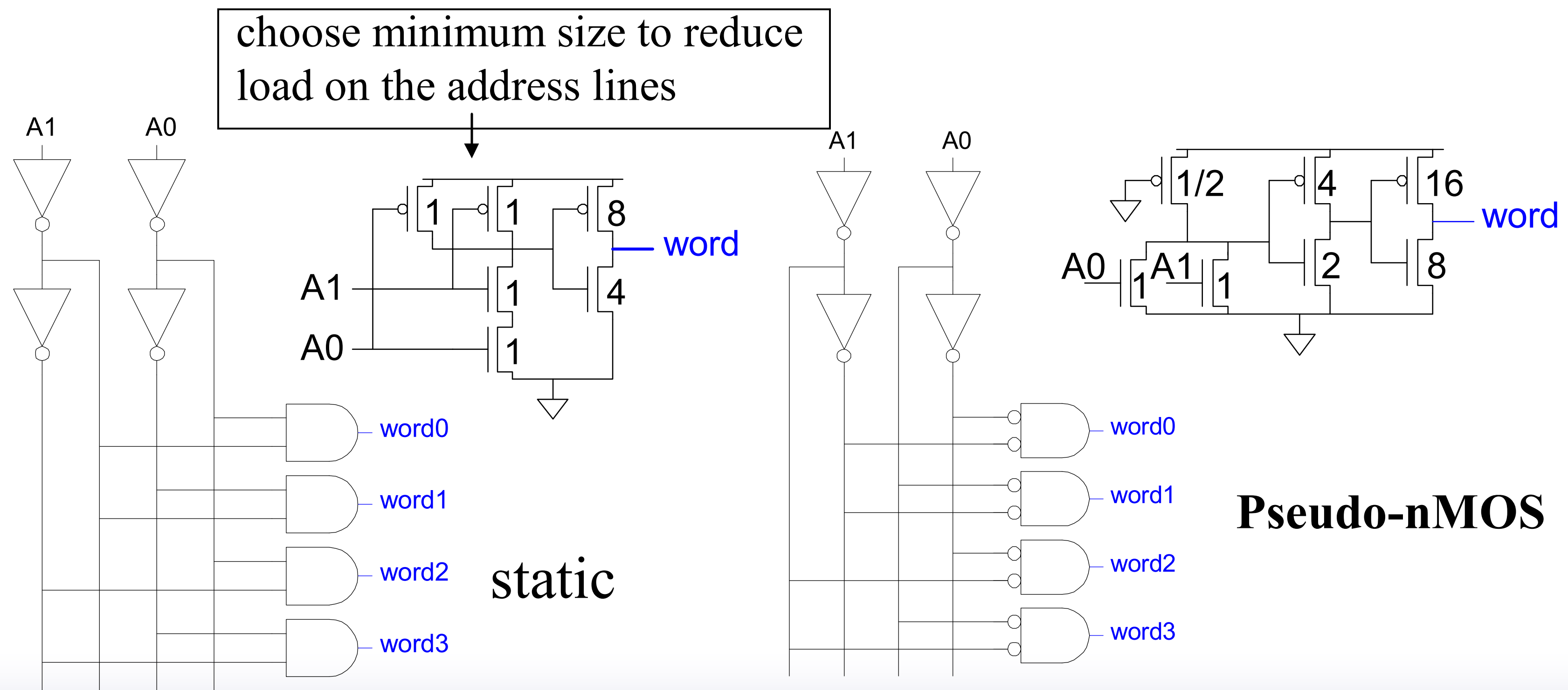


write

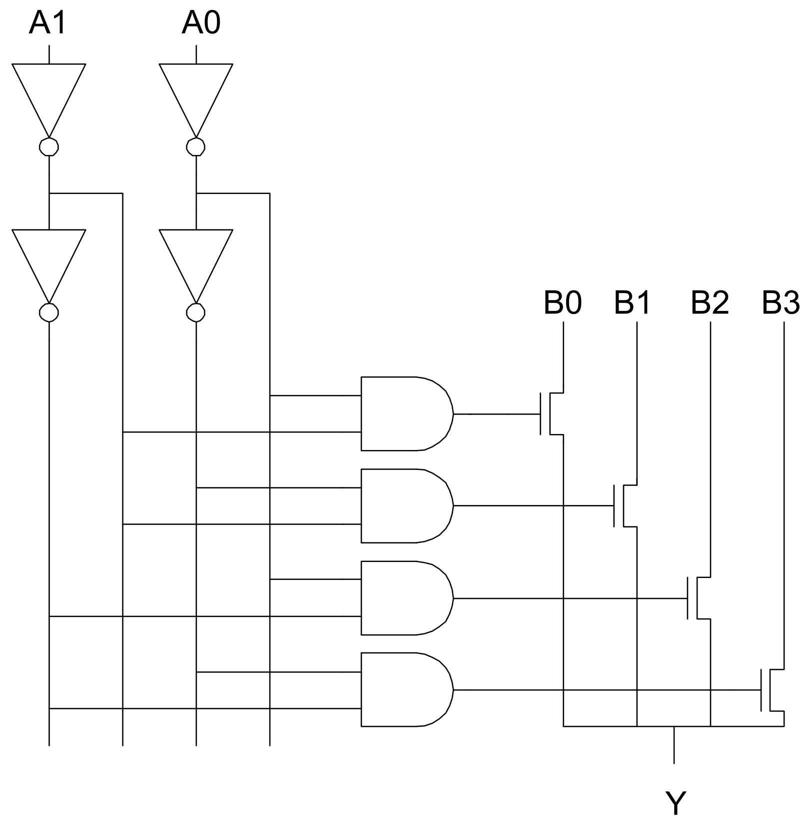


Decoders

- n:2^n decoder consists of 2^n n-input AND gates
  - One needed for each row of memory
  - Build AND from NAND or NOR gate



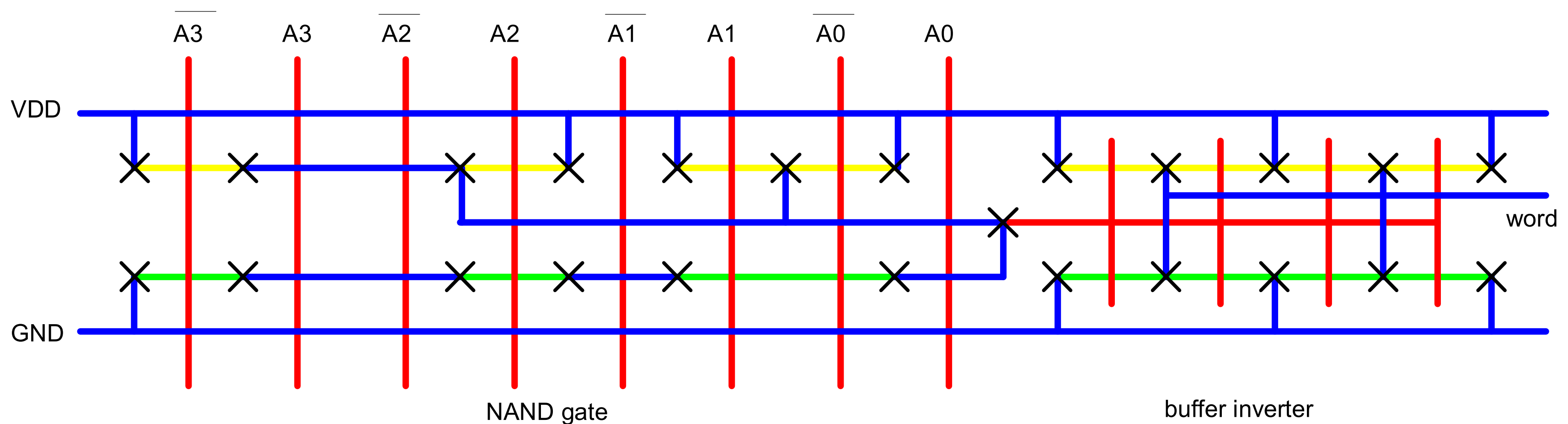
Single Pass-Gate Mux



bitlines propagate through 1 transistor

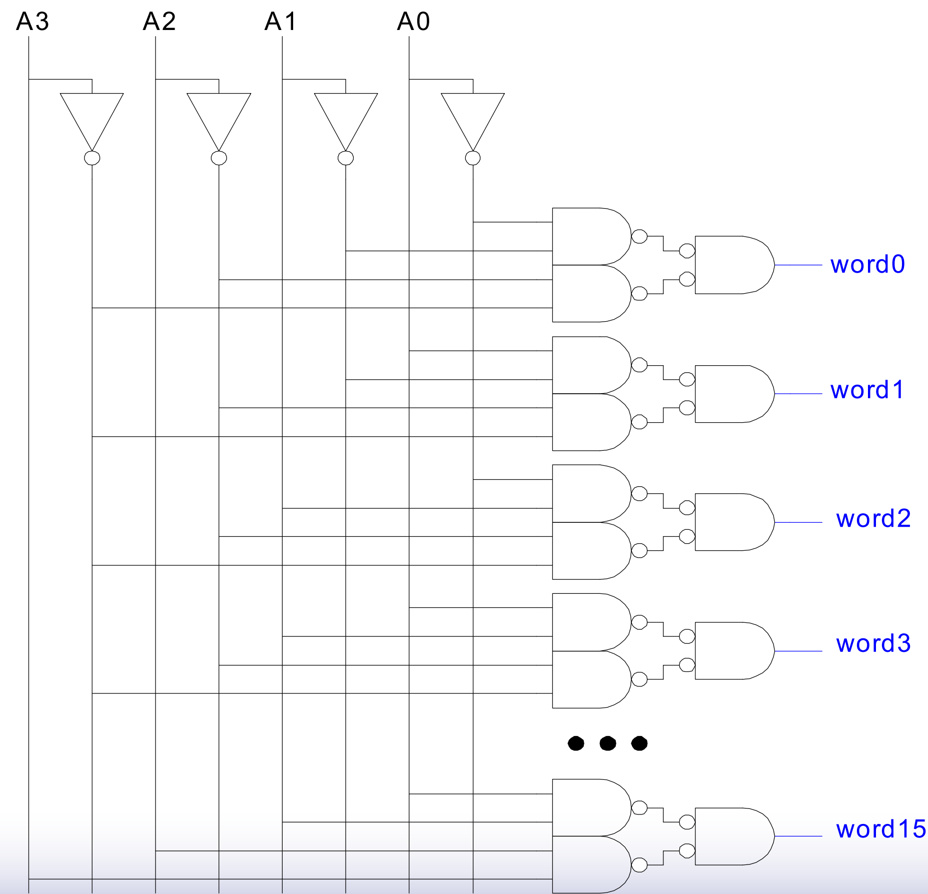
Decoder Layout

- Decoders must be pitch-matched to SRAM cell
  - Requires very skinny gates



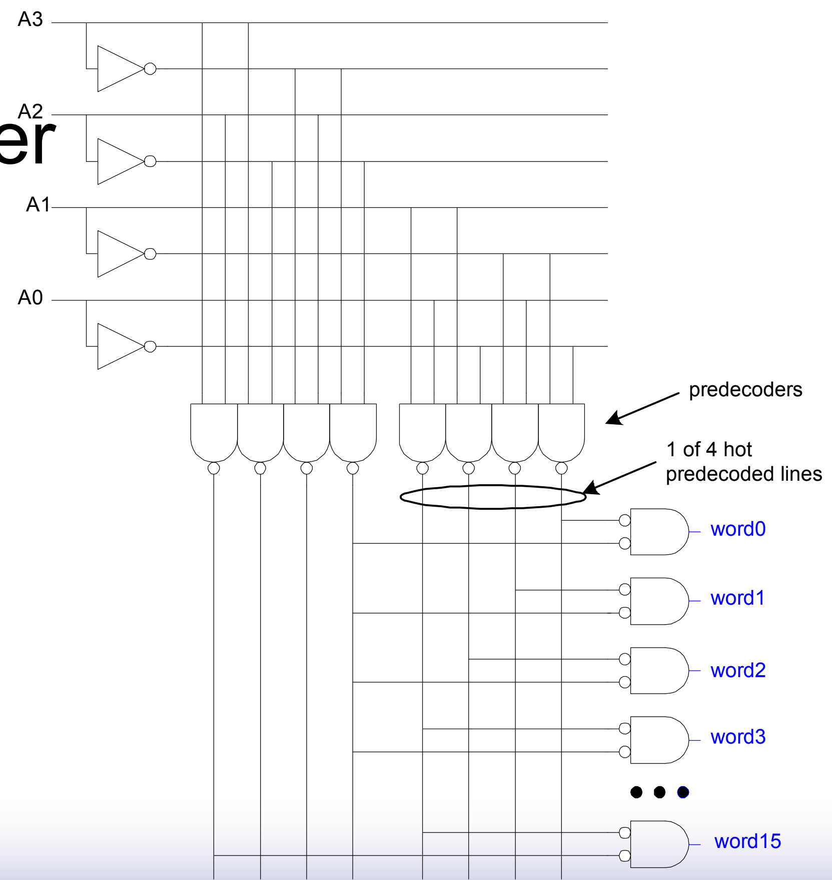
Large Decoders

- For n > 4, NAND gates become slow
  - Break large gates into multiple smaller gates



Predecoding

- Many of these gates are redundant
  - Factor out common gates into predecoder
  - Saves area
  - Same path effort



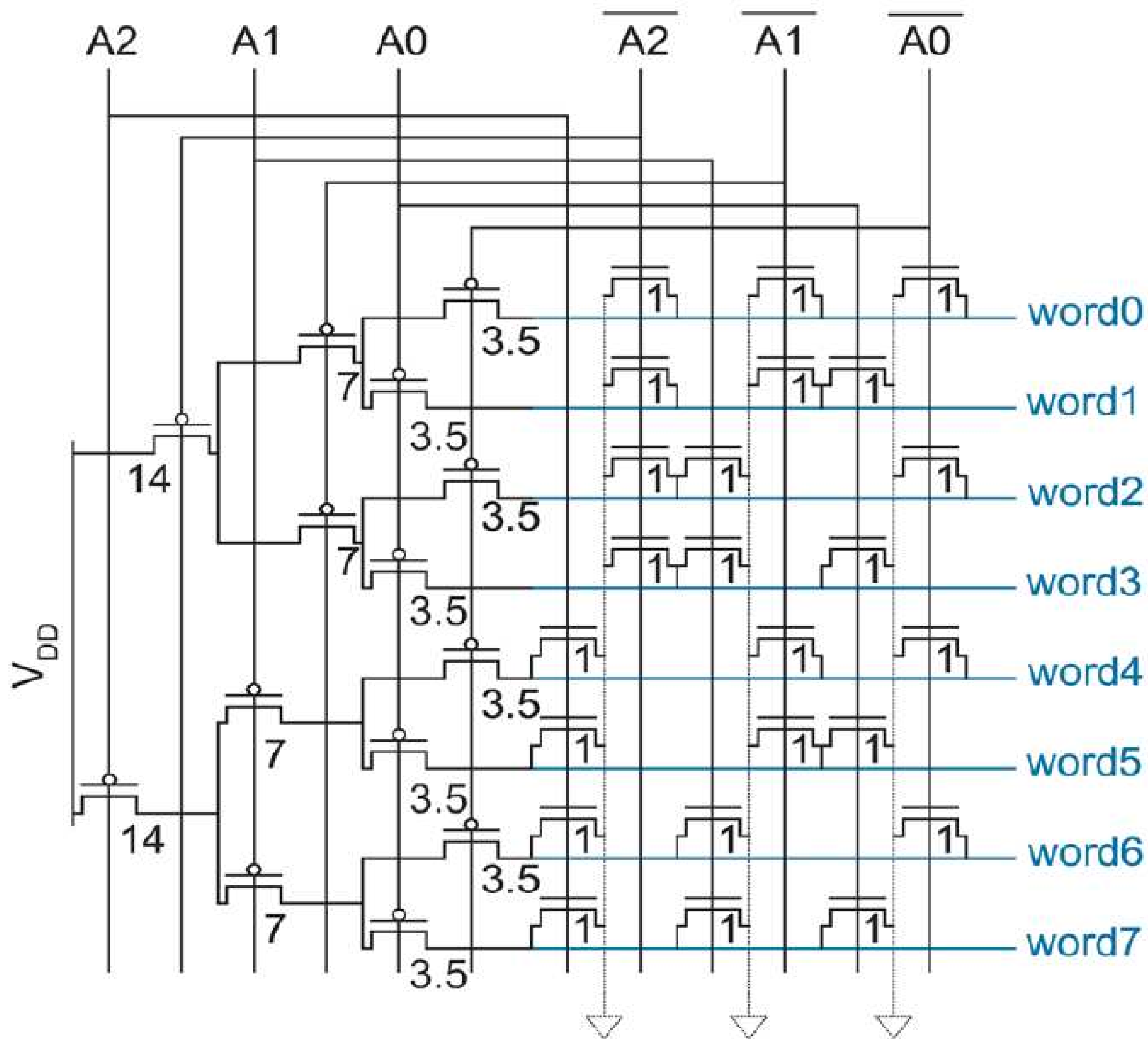


FIG 11.14 Lyon-Schediwy decoder

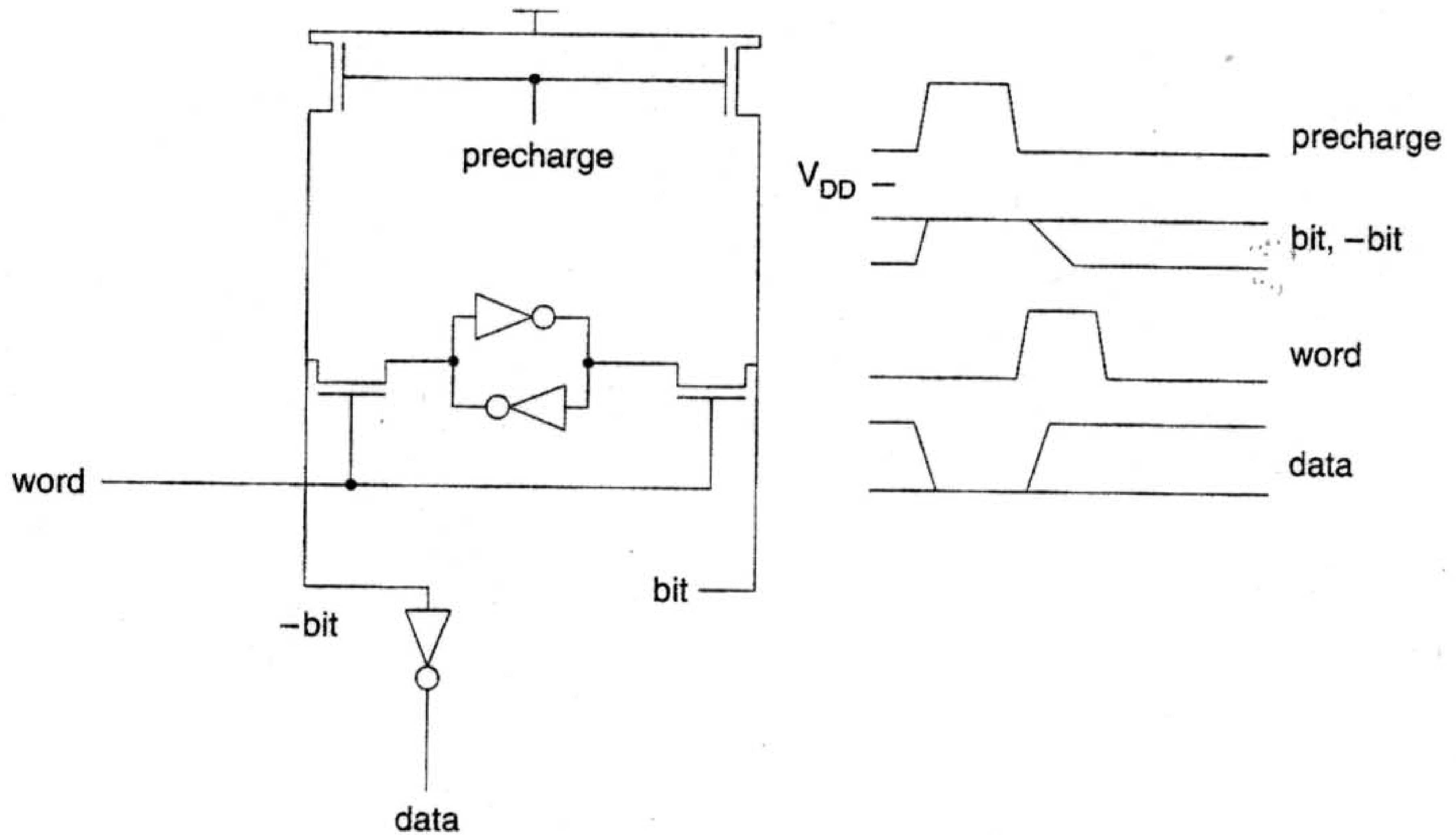
Column Circuitry

- Some circuitry is required for each column
  - Bitline conditioning
  - Sense amplifiers
  - Column multiplexing
- Each column must have write drivers and read sensing circuits

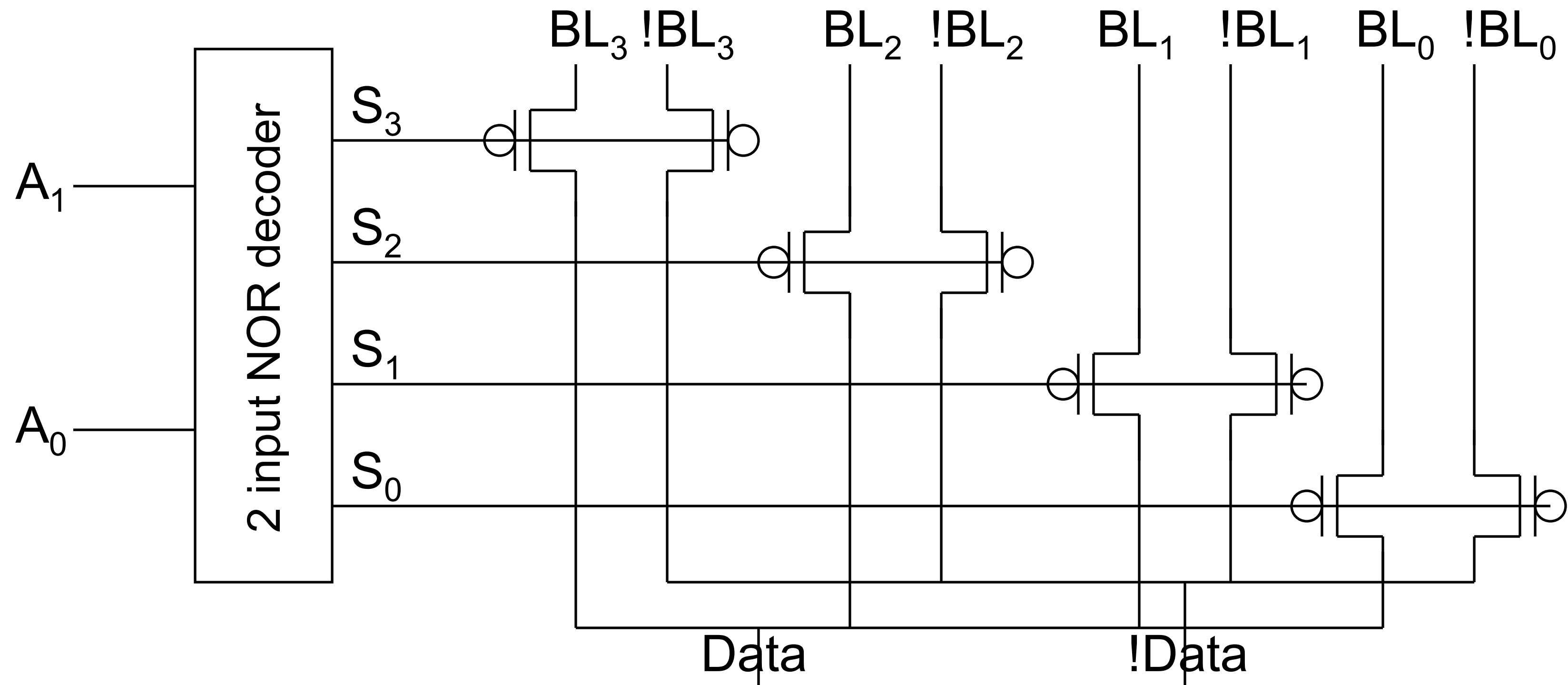
Column Multiplexing

- Recall that array may be folded for good aspect ratio
- Ex: 2k word x 16 folded into 256 rows x 128 columns
  - Must select 16 output bits from the 128 columns
  - Requires 16 8:1 column multiplexers

Typical Column Access



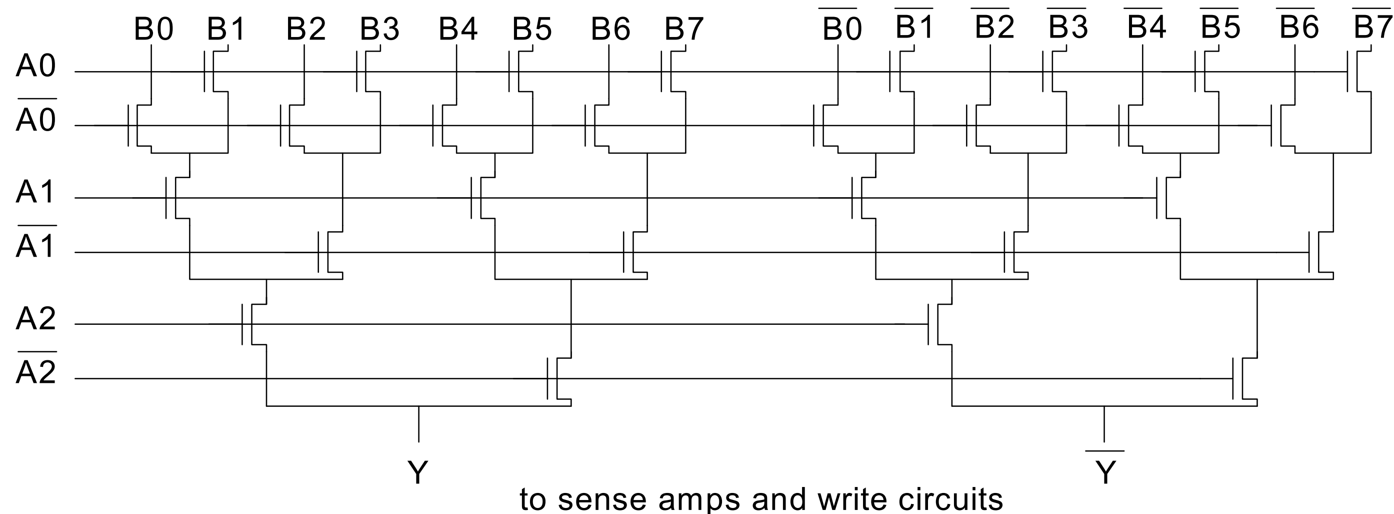
Pass Transistor Based Column Decoder



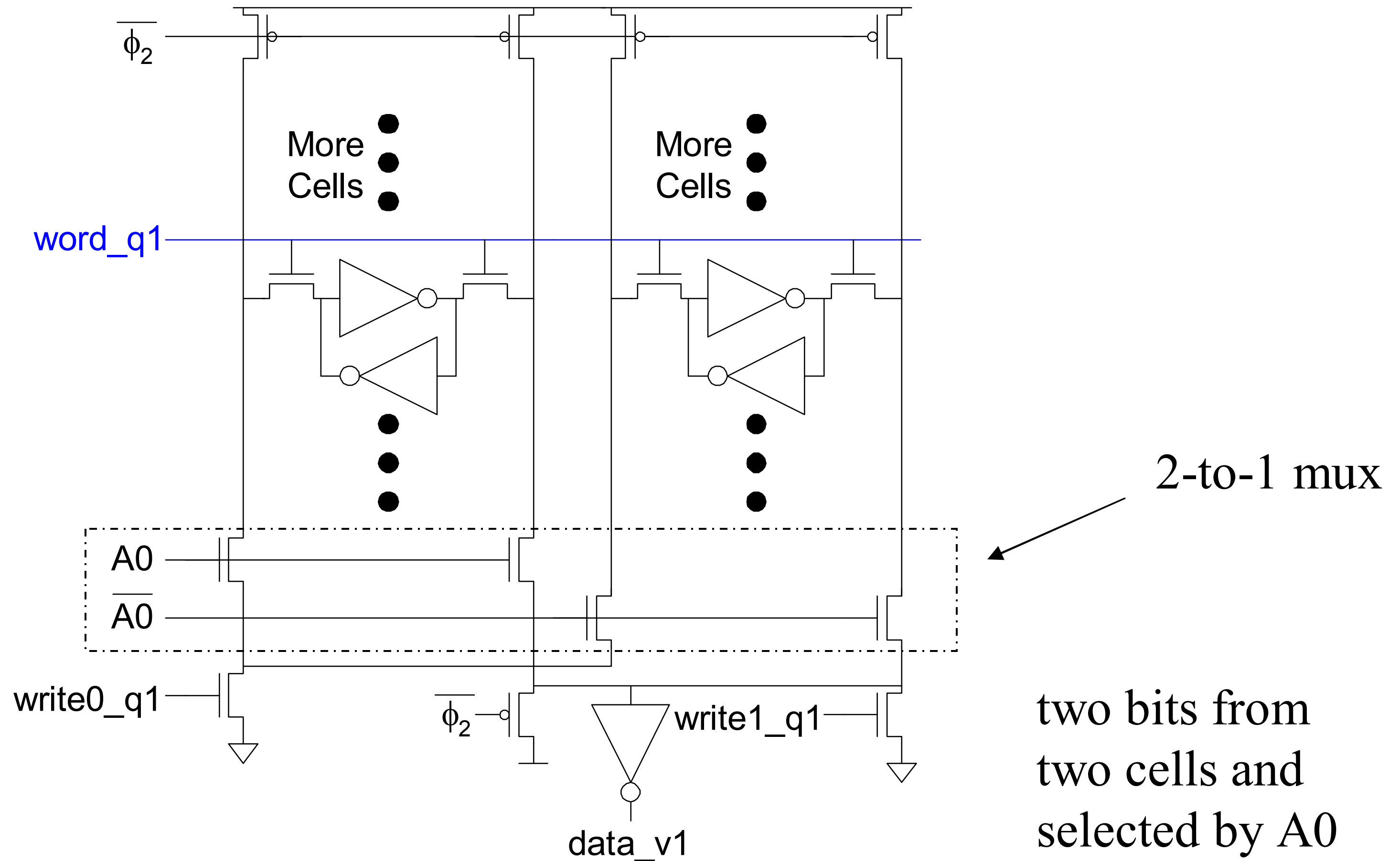
- Advantage: speed since there is only one extra transistor in the signal path
- Disadvantage: large transistor count

Tree Decoder Mux

- Column MUX can use pass transistors
  - Use nMOS only, precharge outputs
- One design is to use k series transistors for 2^k:1 mux
  - No external decoder logic needed



Ex: 2-way Muxed SRAM

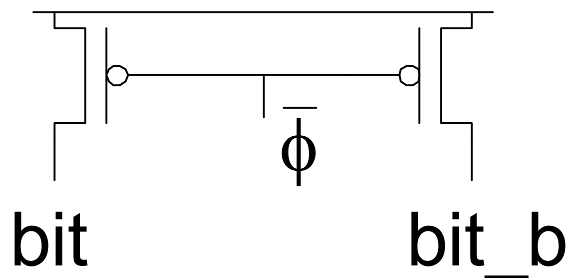


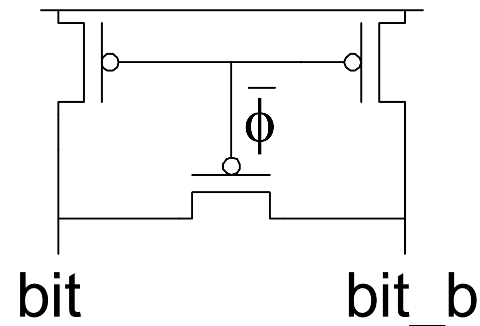
2-to-1 mux

two bits from two cells and selected by A0

Bitline Conditioning

- Precharge bitlines high before reads





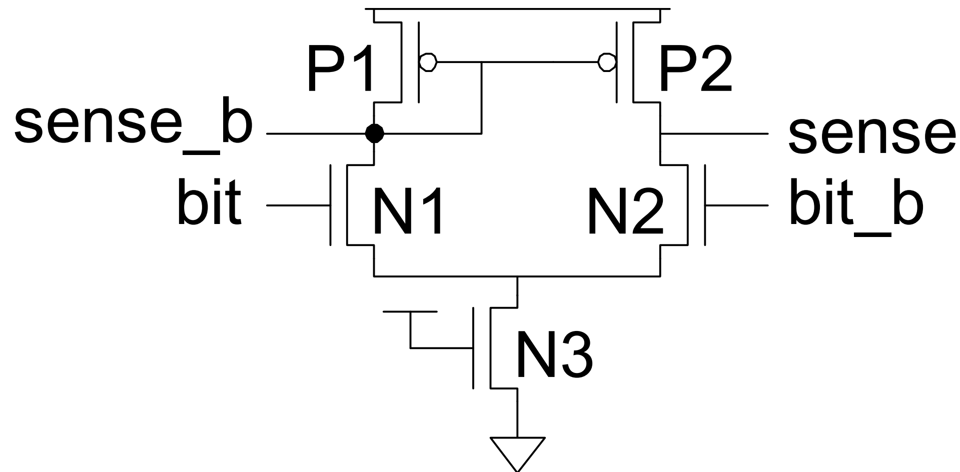
- Equalize bitlines to minimize voltage difference when using sense amplifiers

Sense Amplifiers

- Bitlines have many cells attached
  - Ex: 32-kbit SRAM has 256 rows x 128 cols
  - 128 cells on each bitline
- $t_{pd} \propto (C/I)\,\Delta V$
  - Even with shared diffusion contacts, 64C of diffusion capacitance (big C)
  - Discharged slowly through small transistors (small I)
- *Sense amplifiers* are triggered on small voltage swing (reduce ΔV)

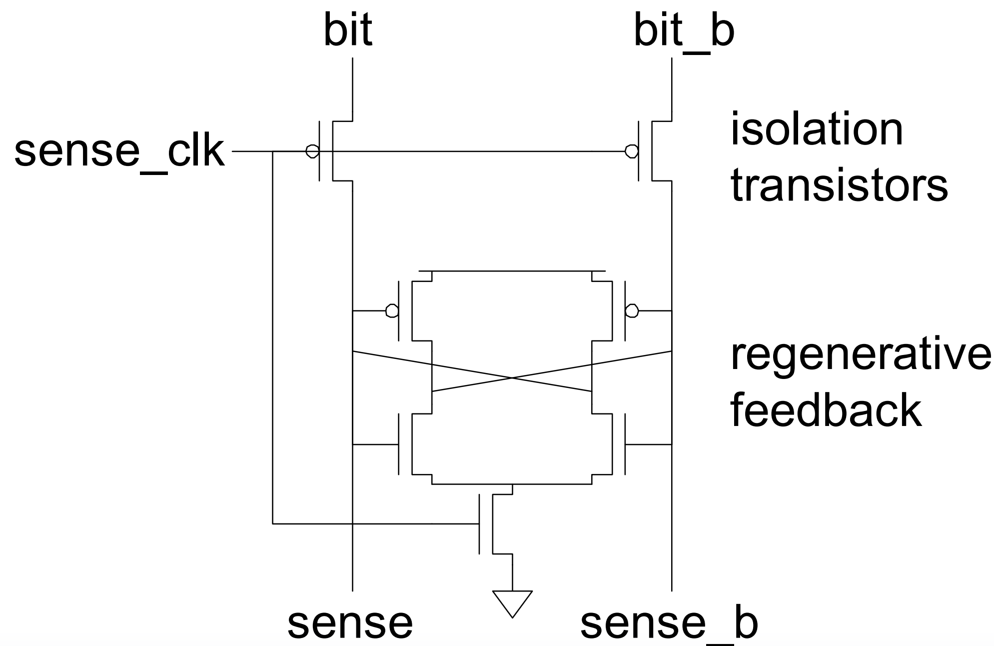
Differential Pair Amp

- Differential pair requires no clock
- But always dissipates static power

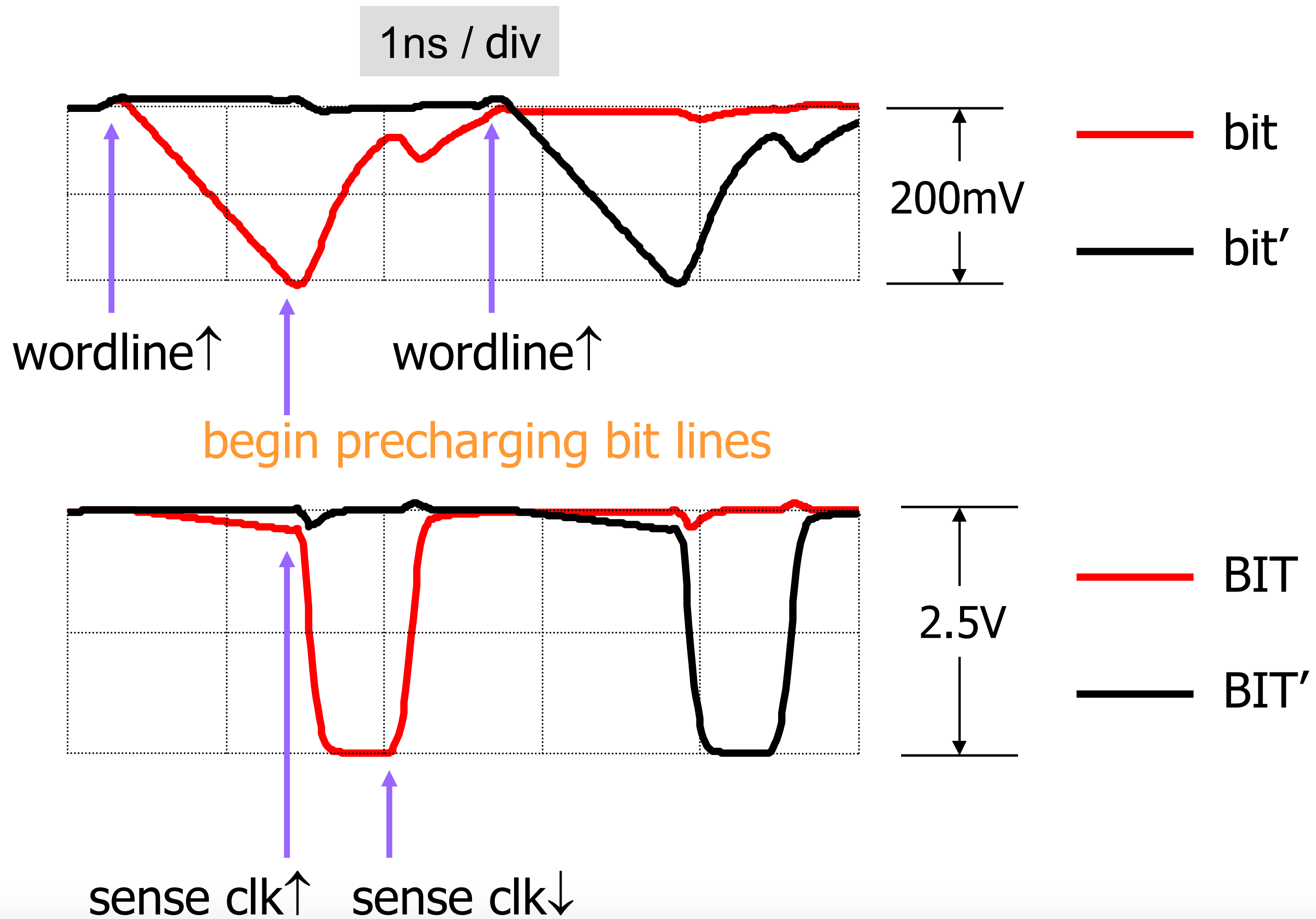


Clocked Sense Amp

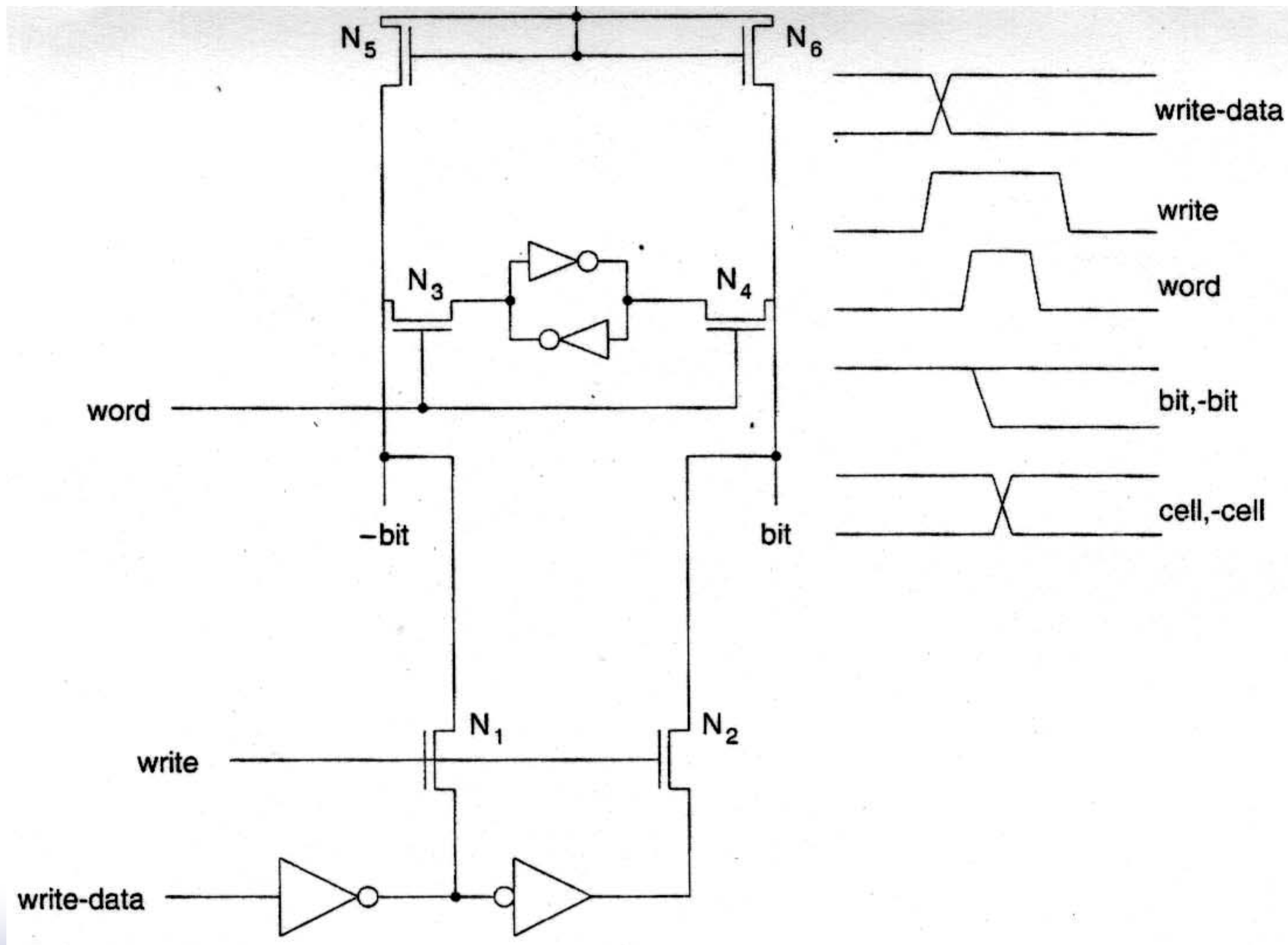
- Clocked sense amp saves power
- Requires sense_clk after enough bitline swing
- Isolation transistors cut off large bitline capacitance



Sense Amp Waveforms

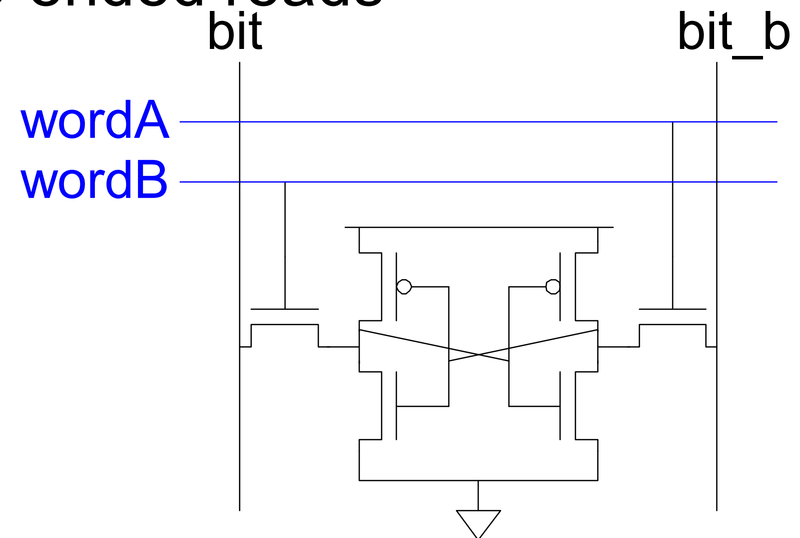


Write Driver Circuits



Dual-Ported SRAM

- Simple dual-ported SRAM
  - Two independent single-ended reads
  - Or one differential write



wordA reads bit_b (complementary)

wordB reads bit (true)

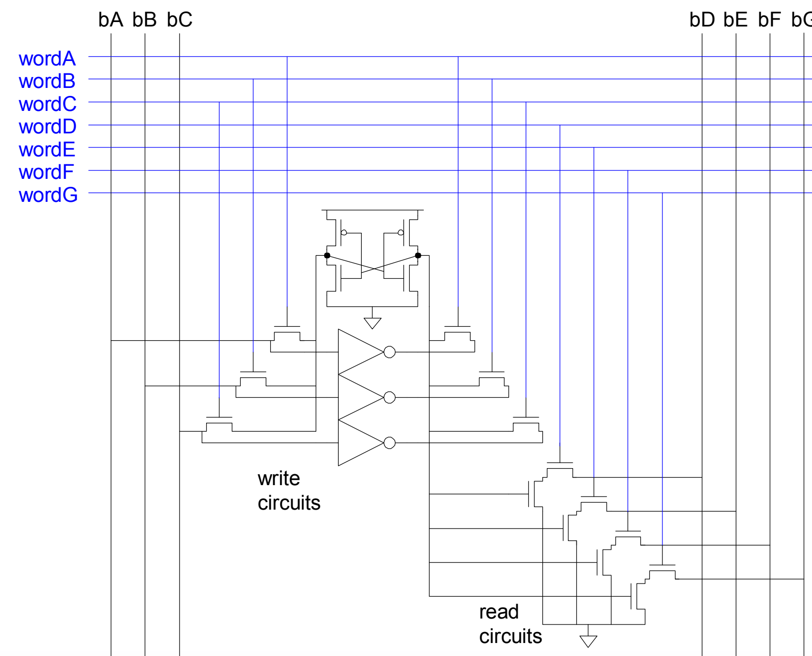
- Do two reads and one write by time multiplexing
  - Read during ph1, write during ph2

Multiple Ports

- We have considered single-ported SRAM
  - One read or one write on each cycle
- *Multiported* SRAM are needed for register files
- Examples:
  - Multicycle MIPS must read two sources or write a result on some cycles
  - Pipelined MIPS must read two sources and write a third result each cycle
  - Superscalar MIPS must read and write many sources and results each cycle

Multi-Ported SRAM

- Adding more access transistors hurts read stability
- Multiported SRAM isolates reads from state node
- Single-ended design minimizes number of bitlines



Logical effort of RAMs

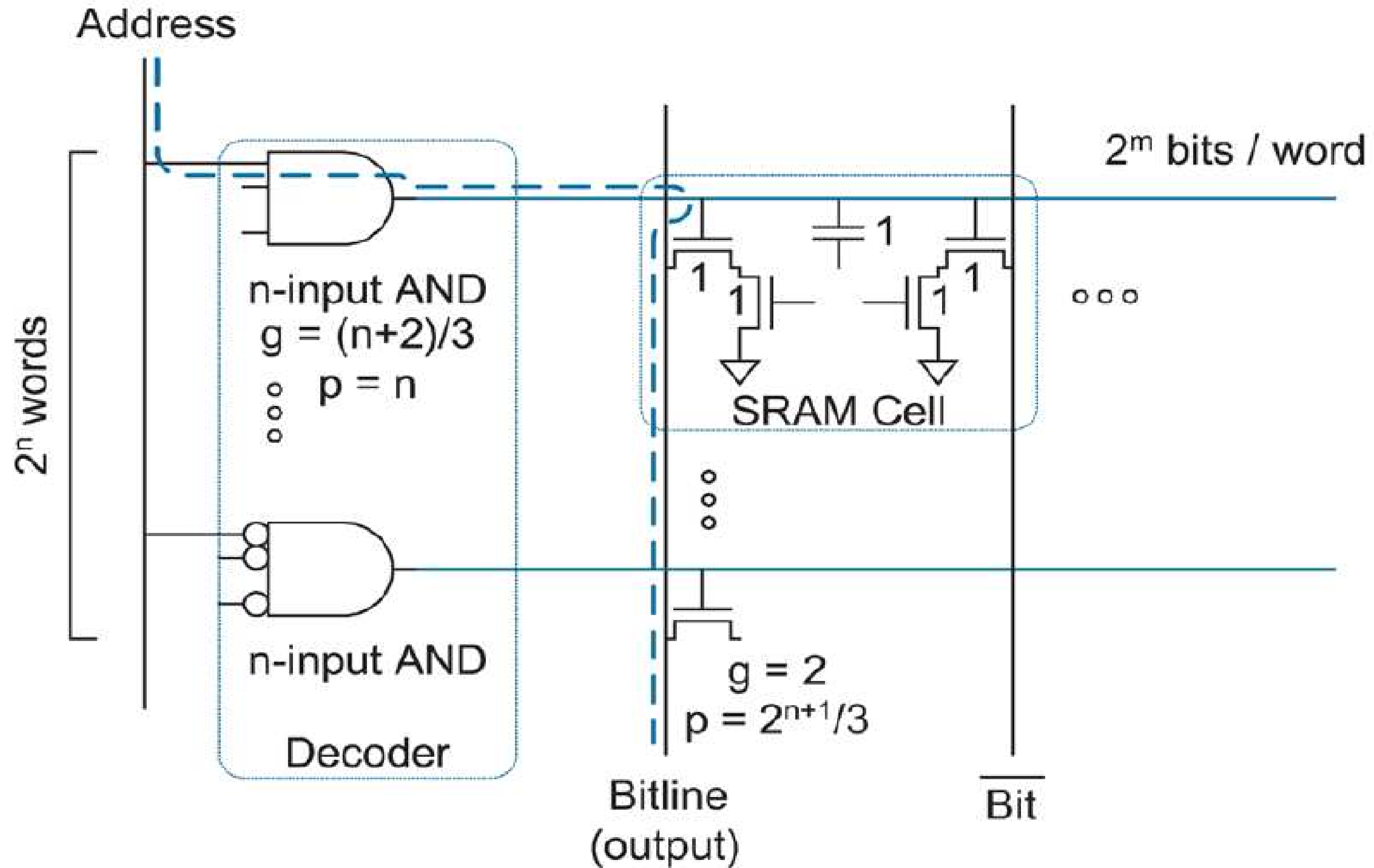
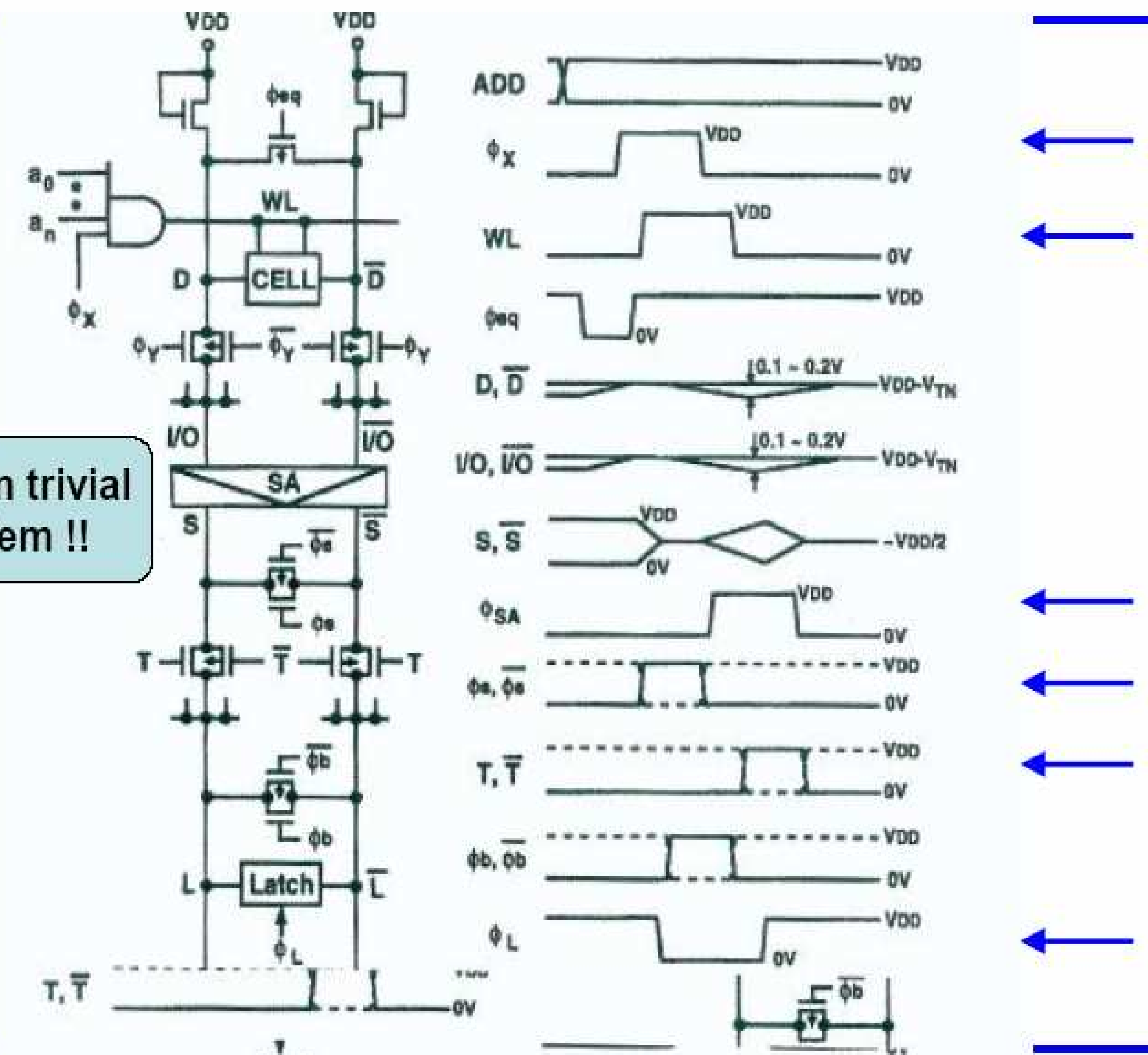


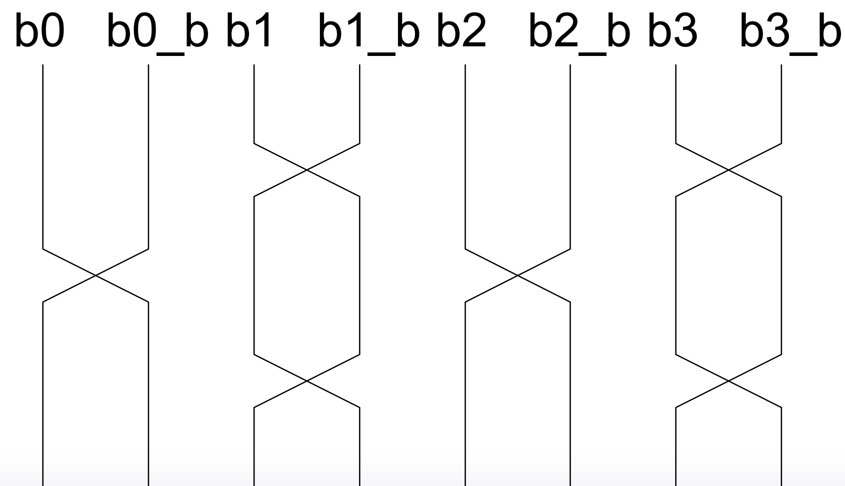
FIG 11.25 Critical path for read of small SRAM



Timing is non trivial design problem !!

Twisted Bitlines

- Sense amplifiers also amplify noise
  - Coupling noise is severe in modern processes
  - Try to couple equally onto bit and bit_b
  - Done by *twisting* bitlines

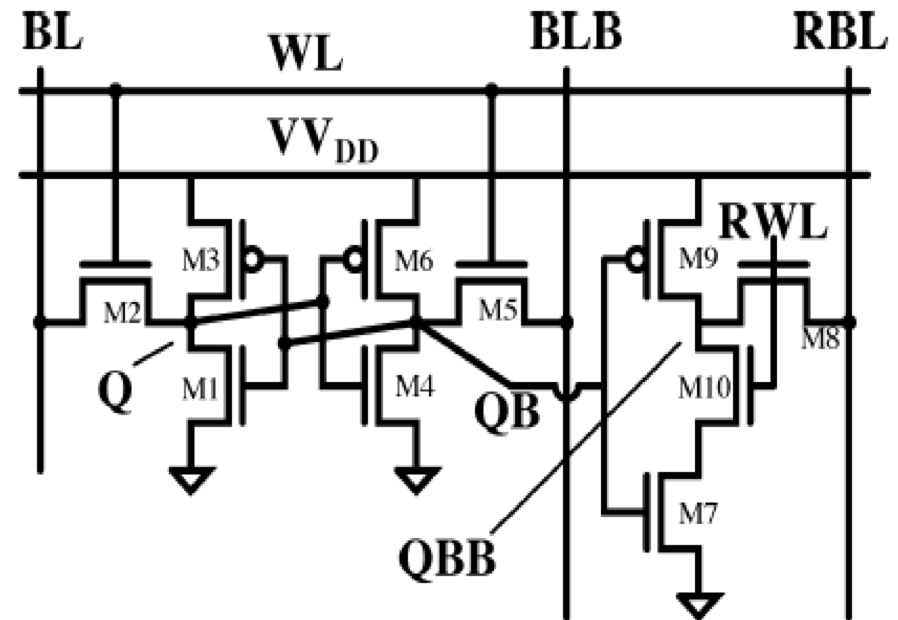


Alternative SRAM Cells

- Low Voltage/High Leakage/Process Variations crowd the operating margins of conventional SRAM
- Alternative Sense Amplifiers, column and row arrangements, adaptive timing, smaller hierarchy, redundant and spare rows/columns have all been addressed in the literature with some success.
- Some problems come from the cell design itself– modifying the cell can break conflicting demands for optimization

10T

- Features
  - BL Leakage reduction
- Approaches
  - Separated Read port
  - Stacked effect by M10
- Performance
  - 400mV@475kHz, 3.28uW
  - 320mV W/O Read error@27℃
  - 380mV W/O Write error@27℃
  - Vmin=300mV@1% bit errors
  - 256 bits/BL

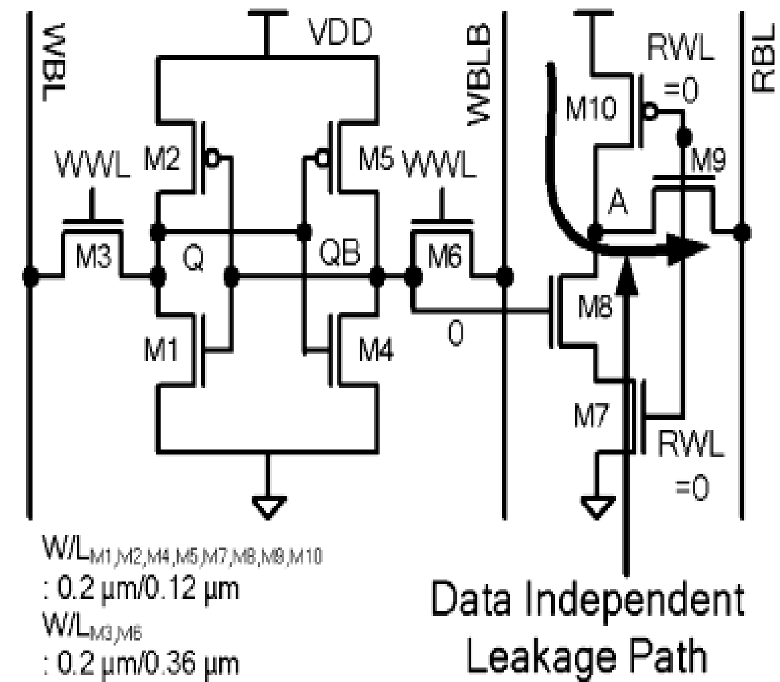


A 256-kb 65-nm Sub-threshold SRAM Design for Ultra-Low-Voltage Operation

B. Calhoun & A. Chandrakasan, JSSC, 2007

10T

- Features
  - BL leakage reduction of data
- Approaches
  - Virtual GND Replica
  - Reverse Short Channel Effect
  - BL Writeback
- Performance
  - 0.2V@100kHz, 2uW
  - 1024 bits/BL
  - 130nm process technology

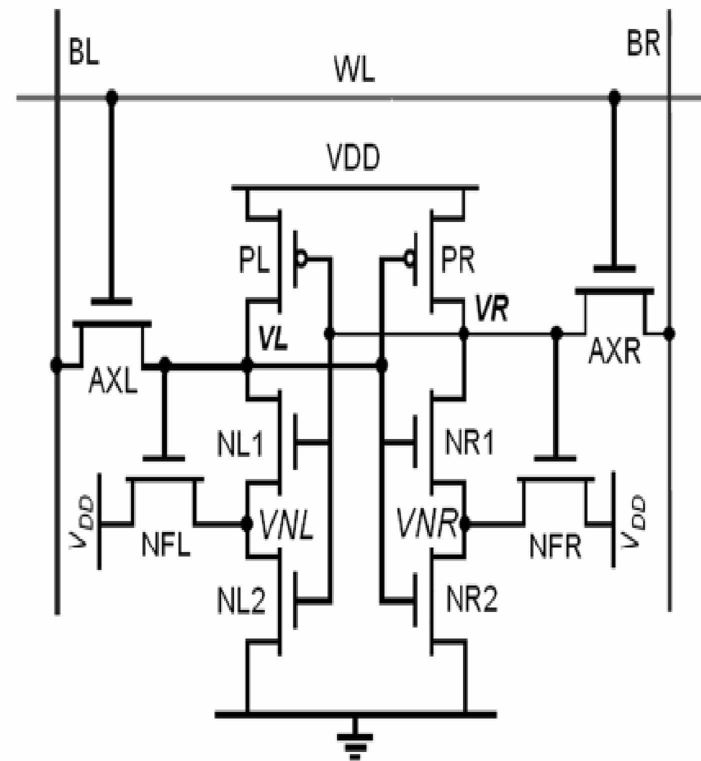


A High-Density Subthreshold SRAM with Data-Independent Bitline Leakage and Virtual Ground Replica Scheme
Chris Kim, ISSCC, 2007

STMicro/Intel/UCSD/THNU

10T

- Features
  - ST cell array can work @160mV
  - 2.1x larger than 6T cell
- Approaches
  - Schmitt Trigger based cell
  - Good stability @ LowVDD
  - Good scalability
- Performance
  - Read SNM ↑ 1.56x @VDD=0.4V
  - More power saving
  - Leakage power ↓ 18%
  - Dynamic power ↓ 50%
  - Hold SNM @150mV is 2.3x of 6T
  - 130nm process

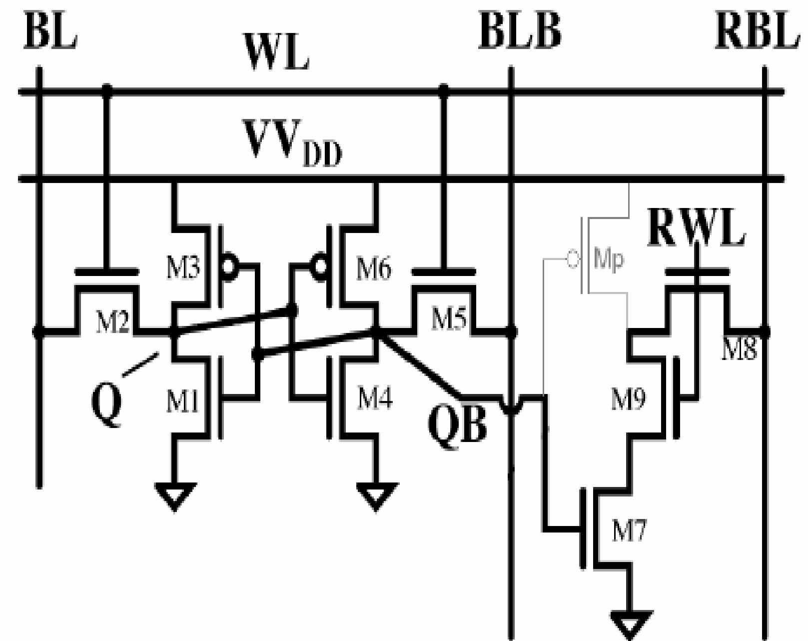


A 160mV Robust Schmitt Trigger Based Subthreshold SRAM

K. Roy, JSSC, 2007

9T

- Features
  - Modifying from 10Tcell
  - 17% more area than 6T cell
  - 16.5% less area than 10T cell
- Approaches
  - More leakage saving than 8T cell
  - Separated read port
- Performance
  - 128 bits/BL @350mV ,100MHz
  - Hold SNM=117mV @300mV
  - Stand-by power: 6uW
  - 65nm process



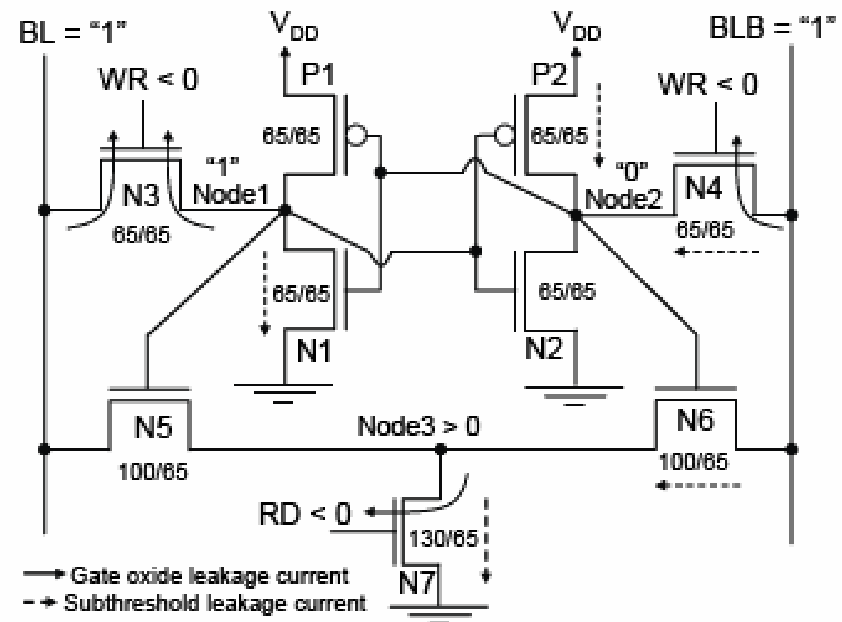
A 100MHz to 1GHz, 0.35V to 1.5V Supply 256x64 SRAM Block using Symmetrized 9T SRAM cell with controlled Read

S. A. Verkila,et al, Conference on VLSI Design, 2008

9T

- Features
  - Read stability enhancement
  - Leakage power reduction
- Approaches
  - Separated read port
  - Min. sizing of N3, N4 and negative

    Vg7, and larger Node3 durir stand-by mode for leakage reduction
- Performance
  - 2x R-SNM cf. 6T
  - 22.9% leakage power reduction
  - 65 nm PTM

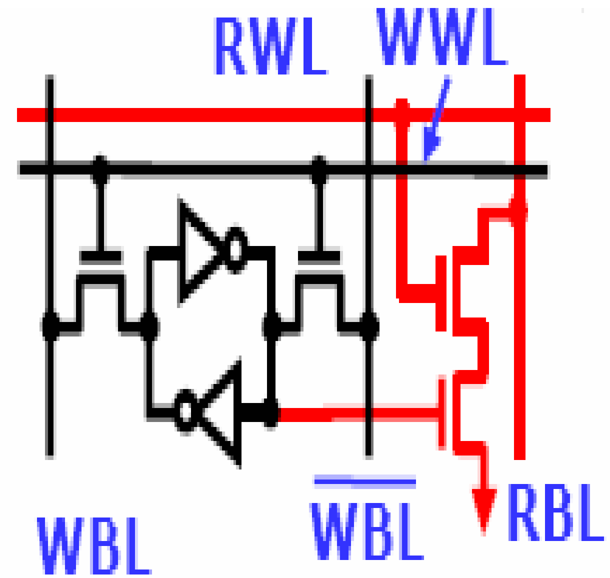


High Read Stability and Low Leakage Cache Memory Cell

Z. Liu and V. Kursun, IEEE Conference, 2007

8T

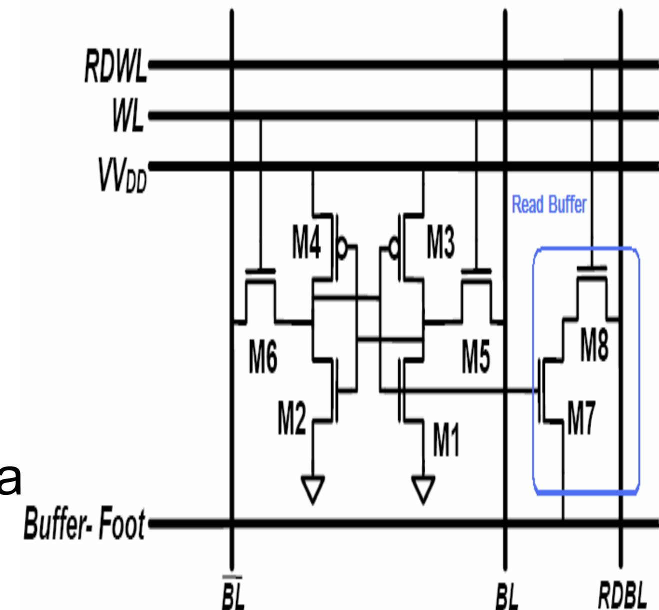
- Features
  - No read disturb
  - About 30% area penalty
- Approaches
  - Separate Read & Write WL
  - Separated read port
- Performance
  - Larger SNM than 6T
  - Better scalability than 6T



Stable SRM Cell Design for the 32nm Node and Beyond
Leland Chang et. al,
Symp. on VLSI 2005

8T

- Features
  - No read disturb
  - Low VDD(350mV)
  - Low subthreshold(Sub. Vt) leakage
- Approaches
  - Separate Read &Write WL
  - Separated read port
  - Foot-drivers reduce the sub.Vt leaka
- Performance
  - 65nm process ,128 cells/row
  - Operating @ 25KHz
  - 2.2uW leakage power

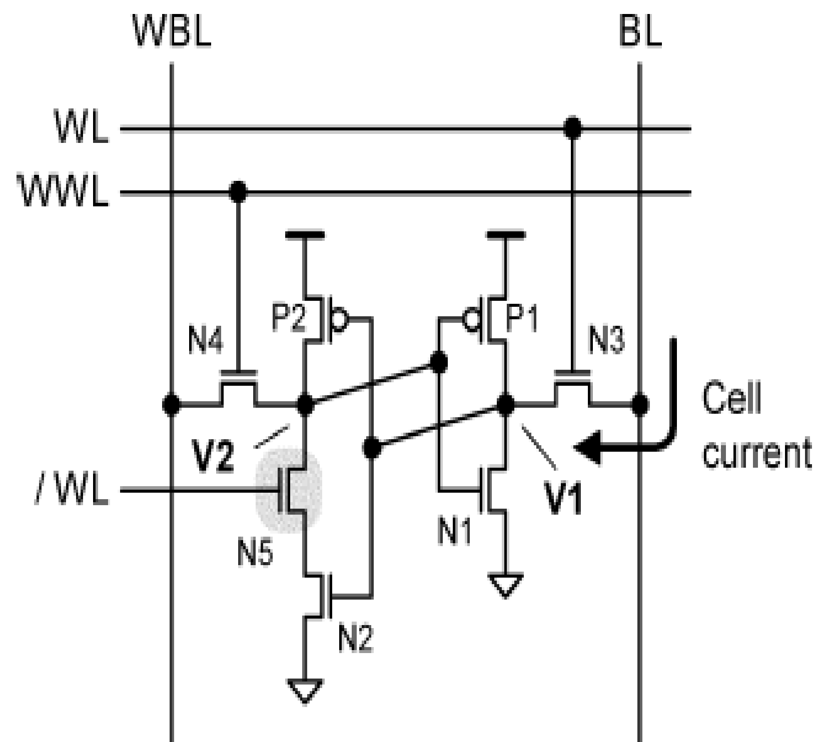


A 256kb 65nm 8T Subhreshold SRAM Employing Sense-Amplifier Redundancy

N. Verma ,and A. P. Chandrakasan, JSSC,2008

7T

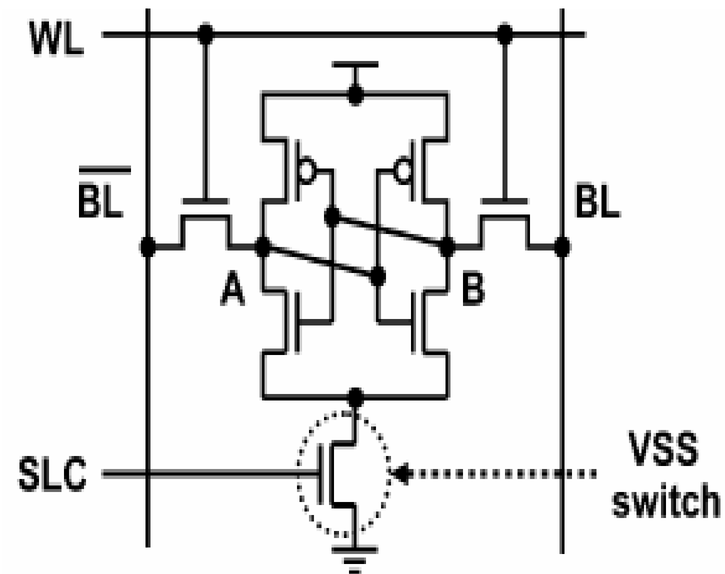
- Features
  - 23% smaller than Conv. 6T bitcell
  - Low VDD(440mV)
  - Not suit for low speed demanc
- Approaches
  - Separate Read &Write WL
  - Seperate Read &Write BL
  - Data protection nMOS:N5
- Performance
  - 20ns access time@0.5V
  - 90nm process



A Read-Static-Noise-Margin-Free SRAM Cell forLow-VDD and High-Speed Applications NEC, JSSC, 2006

7T

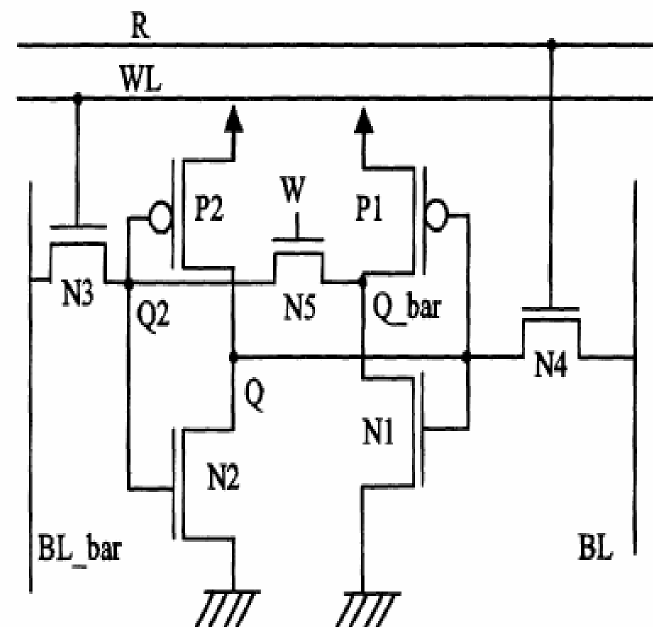
- Features
  - 90% power saving
- Approaches
  - BL swing: VDD/6
- Performance
  - 0.35um proces
  - *Leakage* not controlled well



90% Write Power-Saving SRAM Using Sense-Amplifying Memory Cell
T. Sakurai et. al., JSSC, 2004

7T

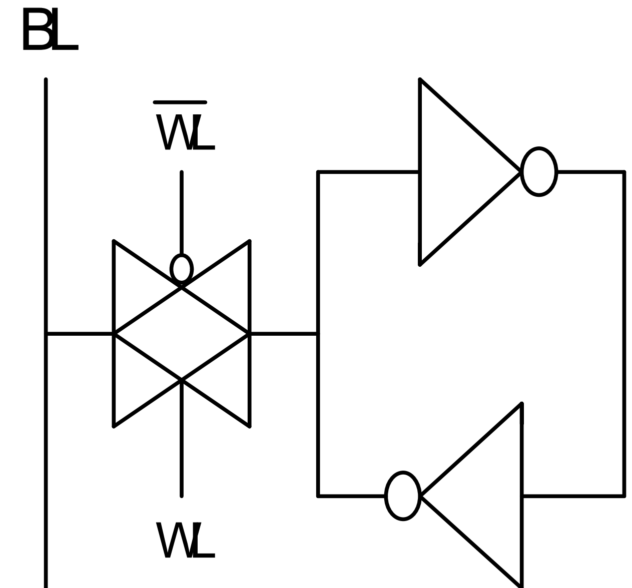
- Features
  - Low write power
  - SNM is effected by “Read pattern” (Read 0-N2,P2,N4 & Read 1-N1,P1,N3,N5)
  - 17.5% larger than 6T
- Approaches
  - Reducing write power by cut off the (feedback) connection to BL
- Performance
  - 0.18um proces
  - 49% write power saving



Novel 7T SRAM Cell For Low Power Cache Design
R. Aly, M. Faisal and A. Bayoumi
IEEE SoC Conf., 2005

6T

- Features
  - Single-ended
  - Low VDD
- Approaches
  - Adjustable header/footer (virVDD, virGND)
- Performance
  - VDD range: 1.2V~193mV
  - Vmin=170mV with 2% redundancy



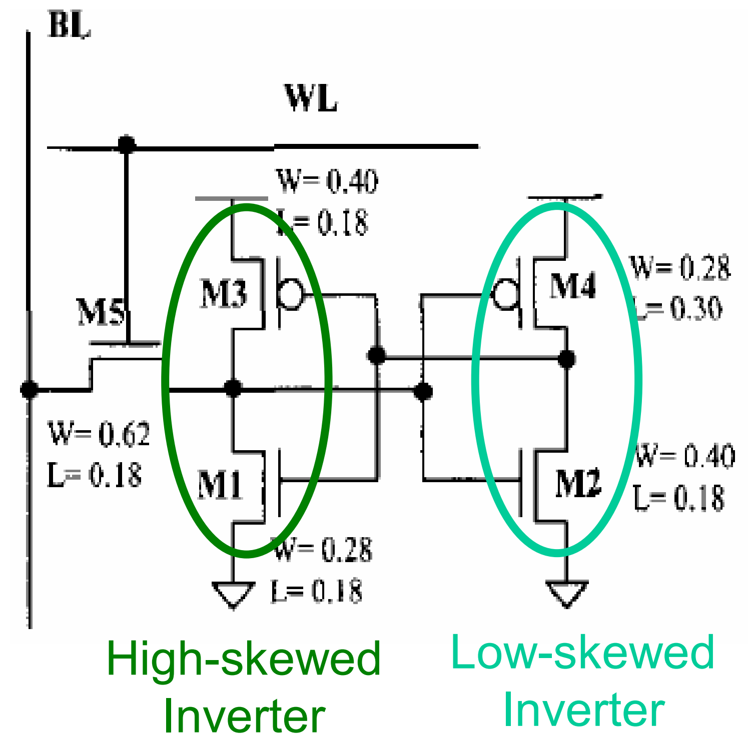
A Sub-200mV 6T SRAM in 0.13 μ m CMOS

ISSCC, 2007

STMicro/Intel/UCSD/THNU

5T

- Features
  - Single-ended
  - Single BL, Single WL
  - Area 23% smaller than 6T
- Approaches
  - BL precharge to Vpc=600mV
  - Asymmetric cell sizing
  - Differential SA is used for Read
- Performance
  - 75% BL leakage reduction cf. 6T
  - SNM is *50% lower* than the 6T's
  - 0.18um process



A High Density, Low Leakage, 5T SRAM for Embeded Caches

I. Carlson et.al., ESSCIRC, 2004

Example Electrical Design: UCSD 32nm prototype

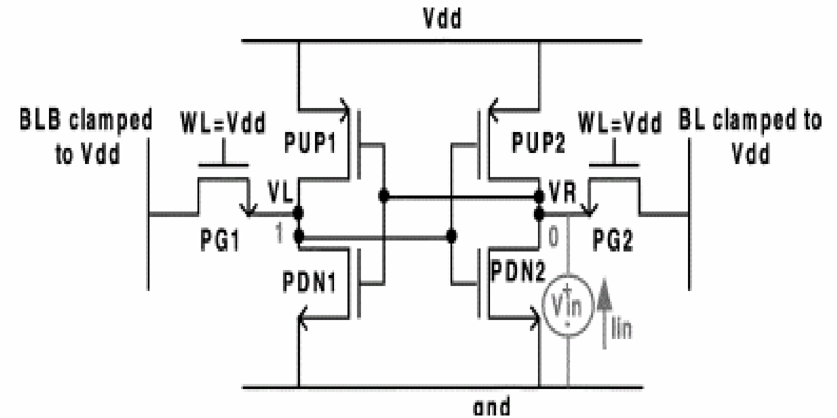
- Butterfly (read stability)
- N-curves (read and write stability)
- I_{read} (read stability and access time)
- VDD_{HOLD} (data retention)
- $I_{leakage}$ (power and data retention)

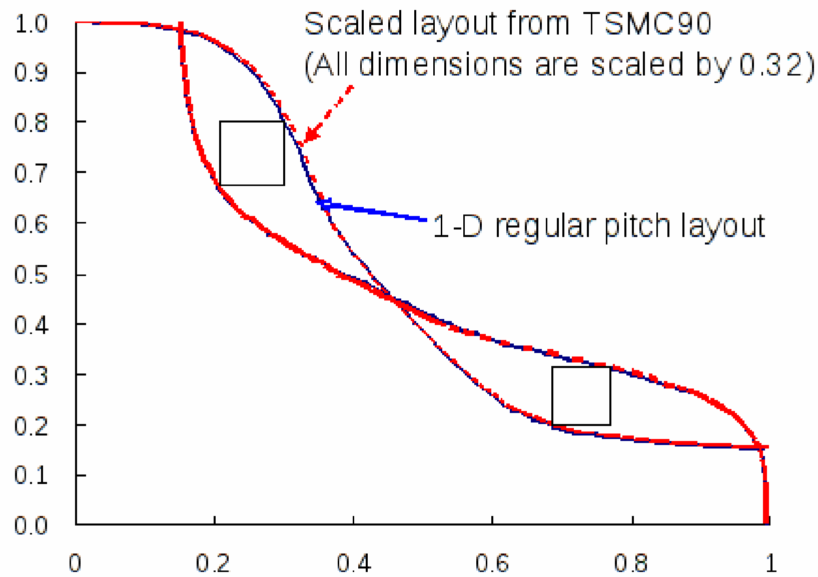
- SPICE Model:
  - 32nm HKMG (high-K/metal-gate) from PTM
- Reference Design
  - Scaled bitcell from TSMC 90nm bitcell

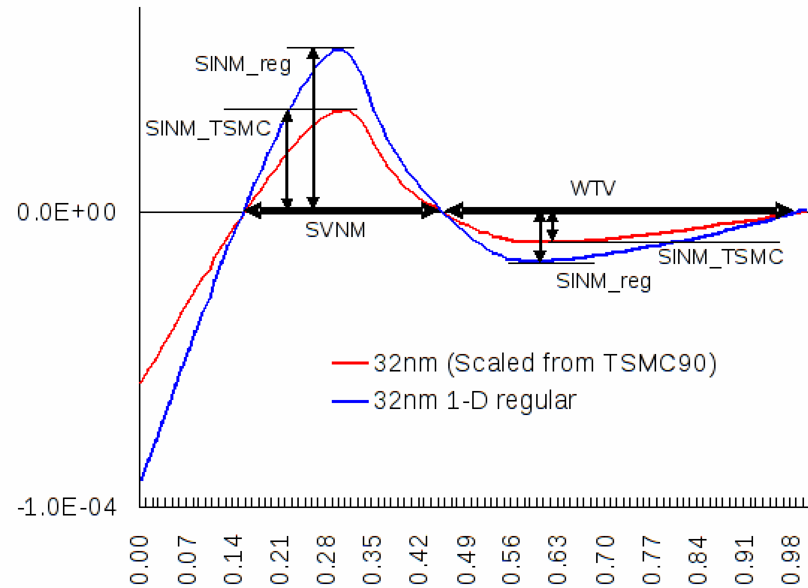
	TSMC 90nm		32nm scaled from TSMC 90nm (REFERENCE)		32nm proposed (for 30x12, 25x12)	
	L (nm)	W (nm)	L (nm)	W (nm)	L (nm)	W (nm)
Pull-up	100	100	32	32	32	44
Pull-down	100	175	32	56	32	88
Pass-Gate	115	120	37	38	32	44

Butterfly and N-Curves

- Measure method
  - Increase VR and measure VL
  - Increase VL and measure VR
  - Make voltage transfer curve in VR and VL axes → Butterfly
  - Measure I_{in} → N-curve



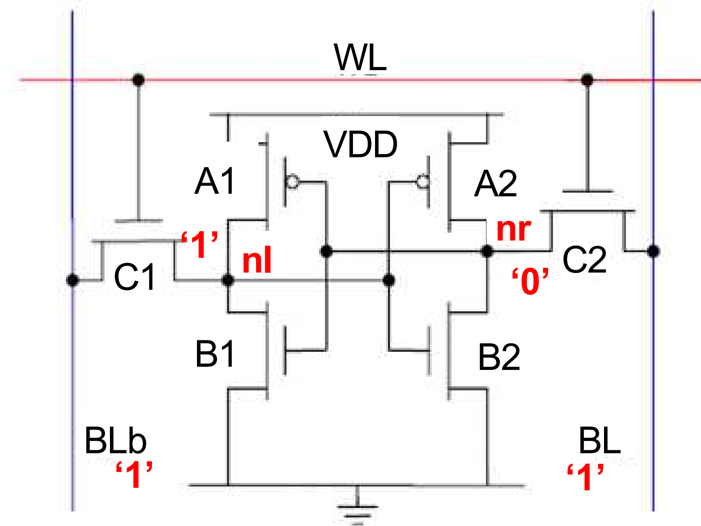




I_{read}, $I_{leakage}$ and VDD_{HOLD}

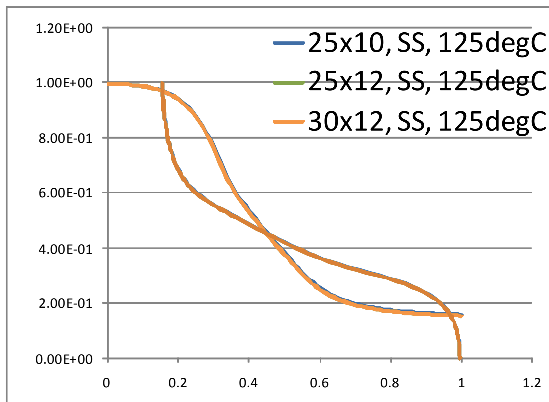
- I_{read}
  - Measure bitline current when WL switches to high
- $I_{LEAKAGE}$
  - Measure VDD (or VSS) current when WL=0
- VDD_{HOLD}
  - Decreasing VDD voltage, while WL=0
  - Measure minimum VDD voltage when | V(nl) - V(nr) | = 'sensing margin'

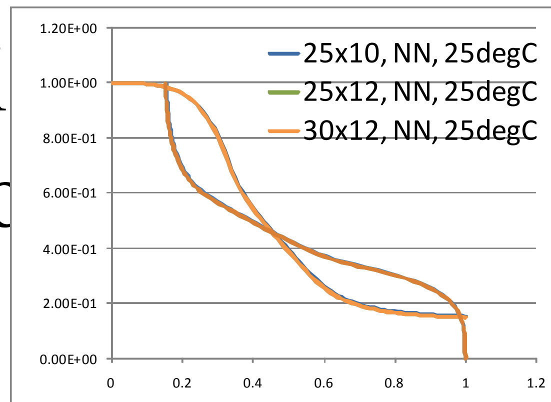
(100mV is assumed)

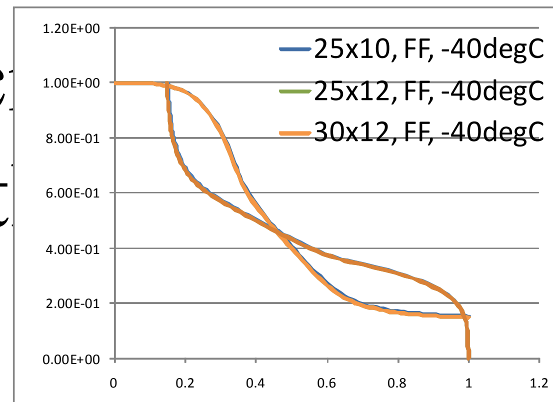


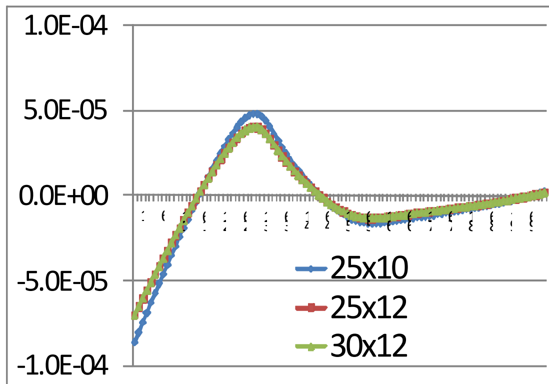
	REFERENCE	32nm proposed (for 30x12 and 25x12)
I_{read}	41.2 uA	66.7 uA
$I_{leakage}$	85.4 nA	142.7 nA
VDD_{HOLD}	110 mV	118 mV

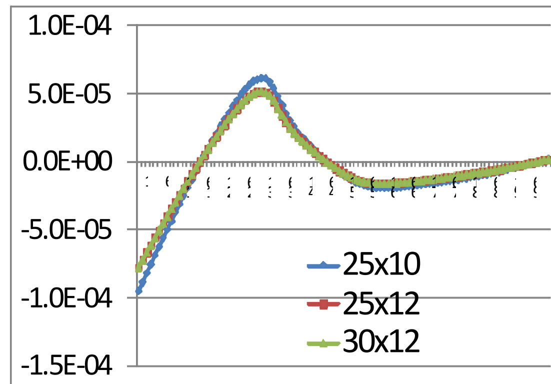
Corner Simulation: Butterfly and N-Curve

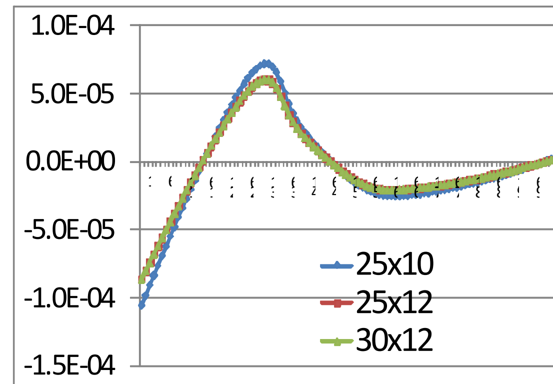






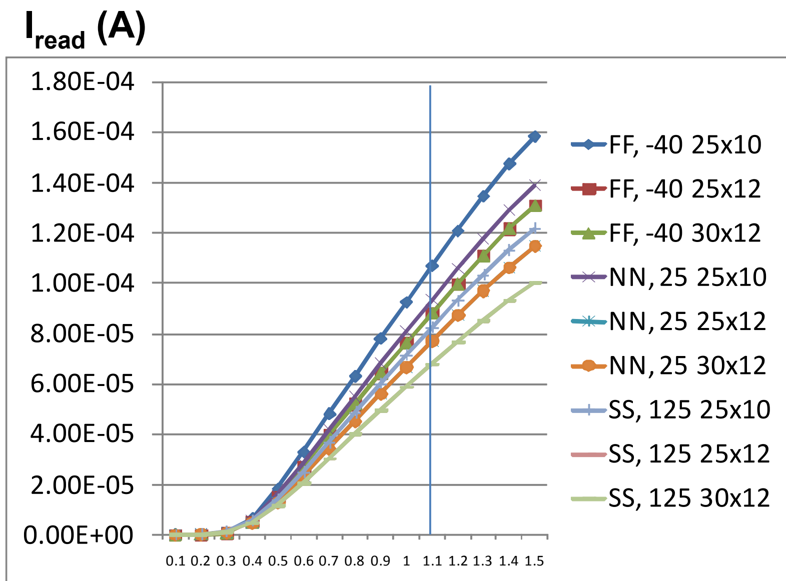


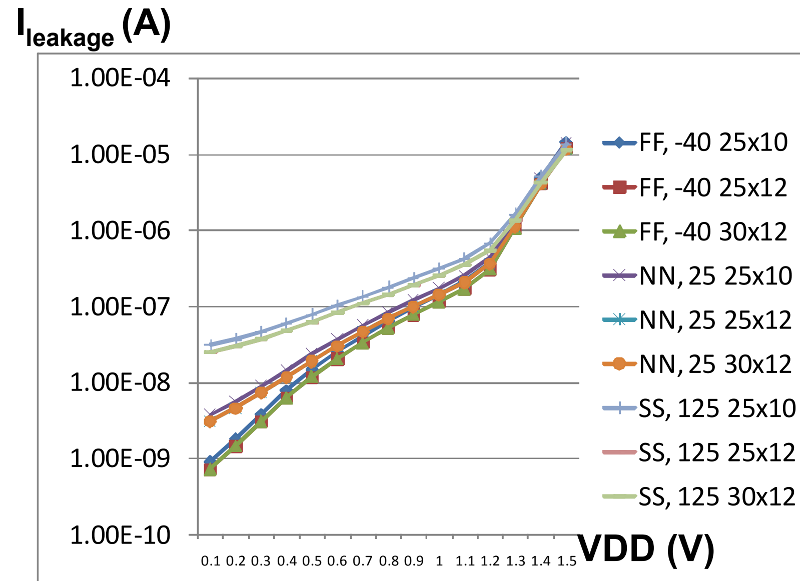


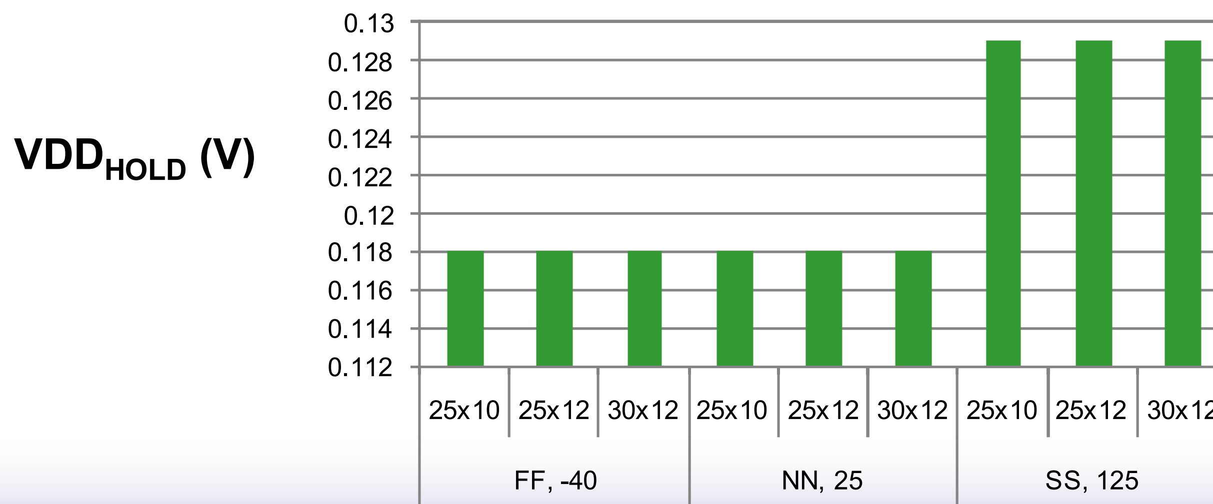


(SS, 125degC, 1.0V) **(NN, 25degC, 1.0V)** **(FF, -40degC, 1.0V)**

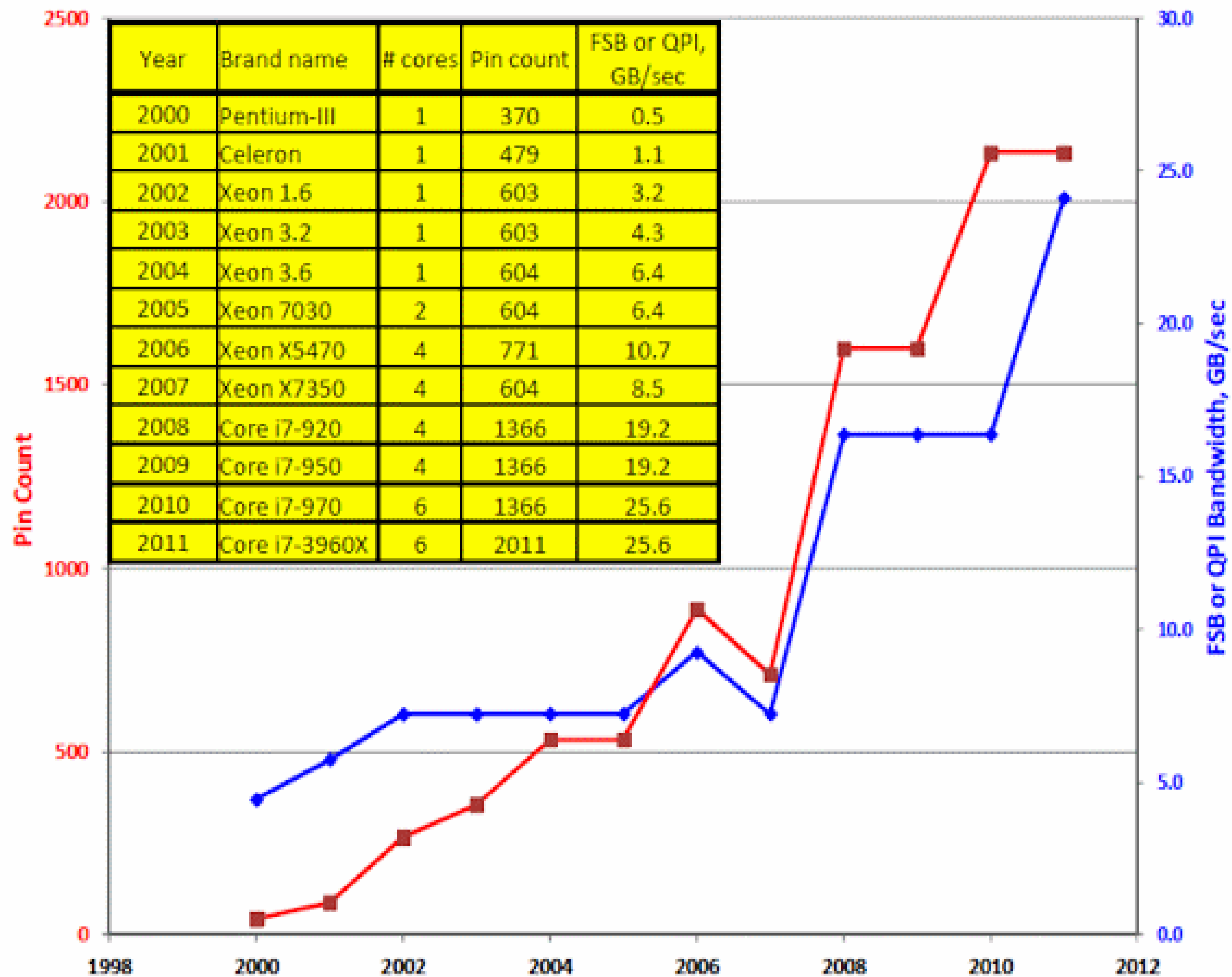
Corner Simulation: I_{read} , $I_{leakage}$ and VDD_{HOLD}







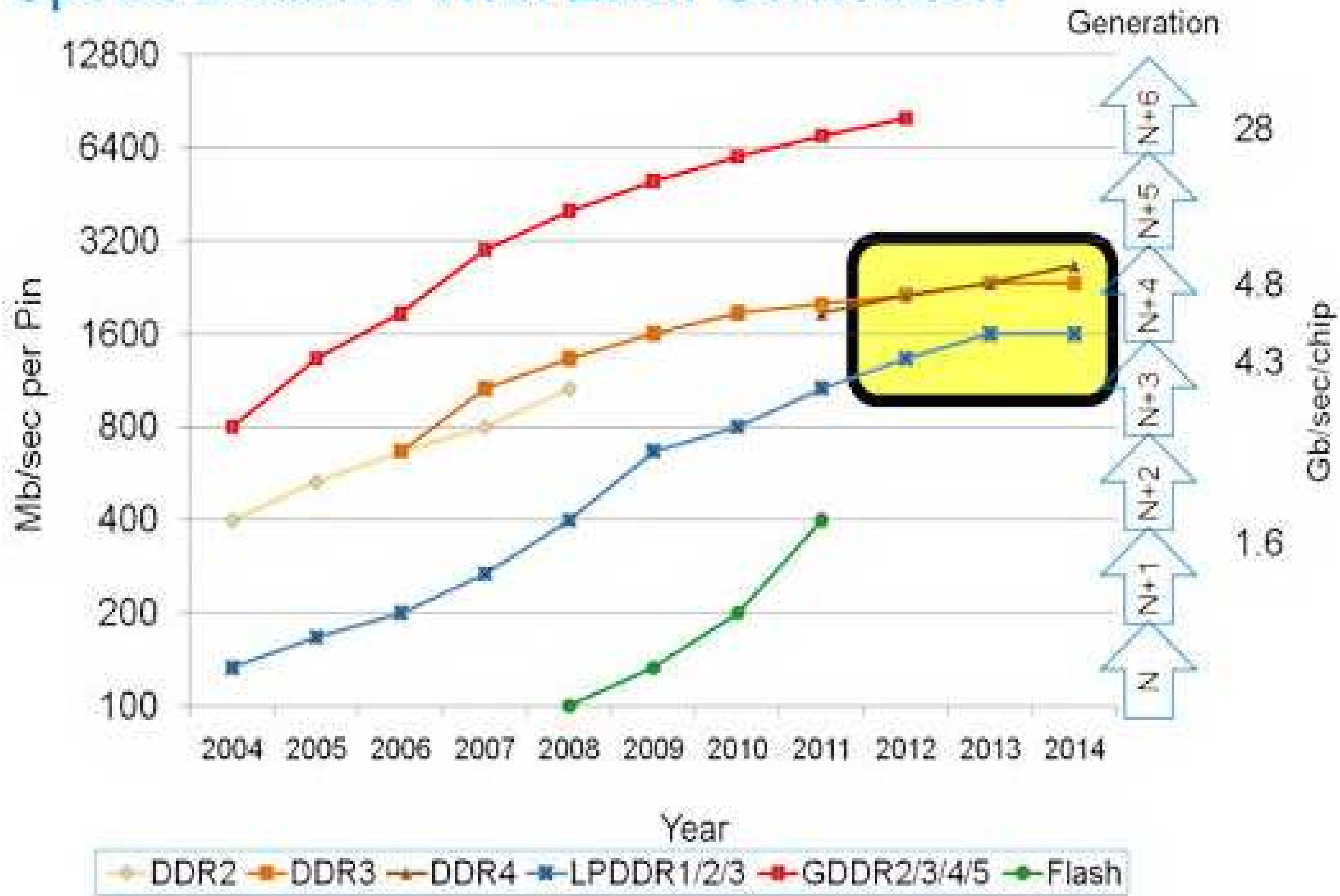
Processor Memory Bandwidth/Pins



Year	Brand name	# cores	Pin count	FSB or QPI, GB/sec
2000	Pentium-III	1	370	0.5
2001	Celeron	1	479	1.1
2002	Xeon 1.6	1	603	3.2
2003	Xeon 3.2	1	603	4.3
2004	Xeon 3.6	1	604	6.4
2005	Xeon 7030	2	604	6.4
2006	Xeon X5470	4	771	10.7
2007	Xeon X7350	4	604	8.5
2008	Core i7-920	4	1366	19.2
2009	Core i7-950	4	1366	19.2
2010	Core i7-970	6	1366	25.6
2011	Core i7-3960X	6	2011	25.6

GPU Bandwidth/Core (local and PCIe)

Year	GPU	Shader Procs (# Cores)	PCIe Speed	GDDR Bandwidth
2003	GeForce PCX4300	2	40	5.33
2004	GeForce PCX5300	5	40	5.33
2005	GeForce 6500	7	40	5.33
2006	GeForce 7900GT	64	40	42.2
2007	GeForce 8800GT	112	80	57.6
2008	GeForce 9800GT	112	80	57.6
2009	GeForce GTX280	240	80	141.7
2010	GeForce GTX480	480	80	177.4
2011	GeForce GTX580	512	80	192.4


Speed Doubles With Each Generation
Generation
12800
6400
3200
1600
800
400
200
100
Mb/sec per Pin
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
Year
N+6
N+5
N+4
N+3
N+2
N+1
N
28
4.8
4.3
1.6
Gb/sec/chip
DDR2
DDR3
DDR4
LPDDR1/2/3
GDDR2/3/4/5
Flash

Why 3D DRAM?