ESE 570: Digital Integrated Circuits and VLSI Fundamentals

Lec 26: December 10, 2020

Review

1

Course Content

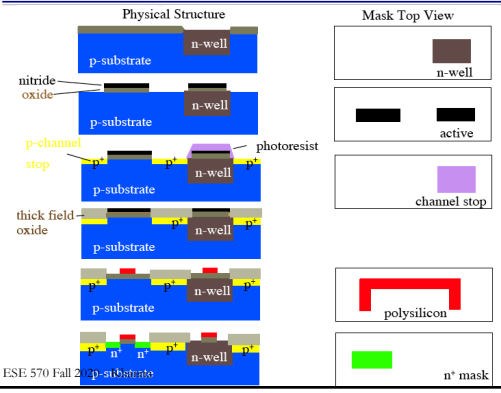
- Introduction
- Fabrication
- MOS Transistor Theory and Models
- MOS Models and IV characteristics
- Inverters: Static Characteristics and Performance
- Inverters: Dynamic Characteristics and Performance
- Combinational Logic Types (CMOS, Ratioed, Pass) and Performance
- Sequential Logic
- Dynamic Logic
- Memory Design
- Robust VLSI Design for Variation
- I/O Circuits and Inductive Noise
- CLK Generation
- Crosstalk
- Transmission Lines

2

MOS Devices and Characteristics

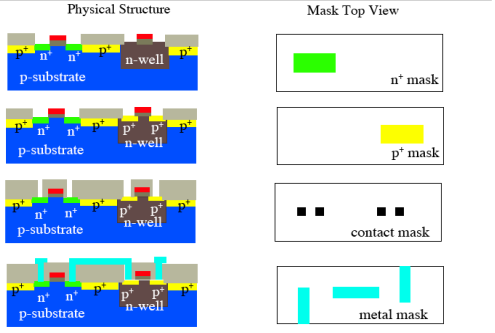
3

Typical N-Well CMOS Process



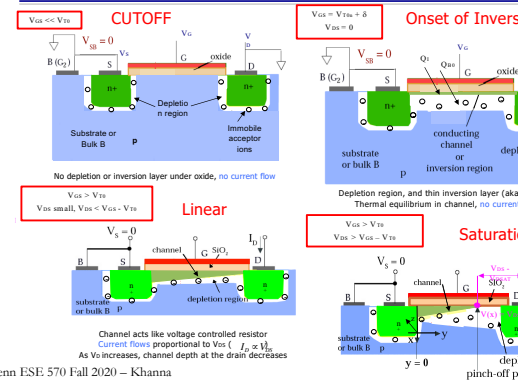
4

Typical N-Well CMOS Process



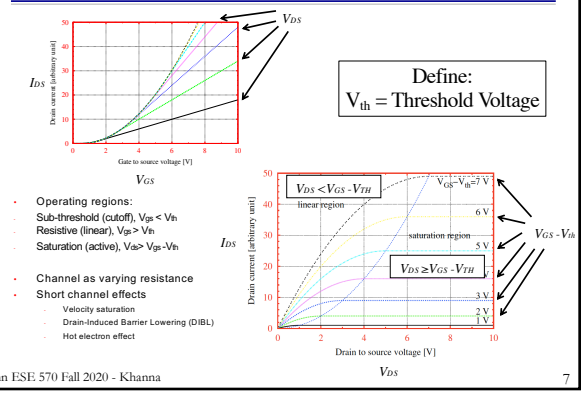
5

Regions of Operation



6

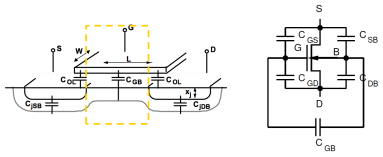
MOSFET – IV Characteristics



- Operating regions:
  - Sub-threshold (cutoff), $V_{gs} < V_{th}$
  - Resistive (linear), $V_{gs} > V_{th}$
  - Saturation (active), $V_{ds} > V_{gs} - V_{th}$
- Channel as varying resistance
- Short channel effects
  - Velocity saturation
  - Drain-Induced Barrier Lowering (DIBL)
  - Hot electron effect

7

MOSFET Parasitic Capacitance



- $C_{GS}=C_{GCS}+C_{GSO}$
- $C_{GD}=C_{GCD}+C_{GDO}$
- $C_{GB}=C_{GCB}$
- $C_{SB}=C_{diff}$
- $C_{DB}=C_{diff}$

- Any two conductors separated by an insulator form a parallel-plate capacitor
- Two types
  - Extrinsic – Outside the box (e.g. junction, overlap)
  - Intrinsic – Inside the box (e.g. gate-to-channel)

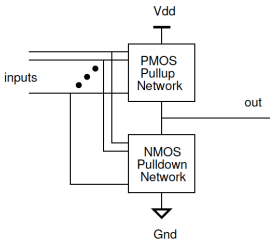
8

CMOS Logic

Penn

9

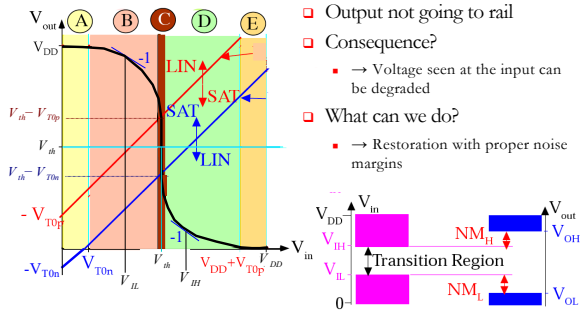
Static CMOS Gate Structure



- Drives rail-to-rail
  - Power rails are V_{dd} and Gnd
  - output is V_{dd} or Gnd
- Input connects to gates → load is capacitive
- Once output node is charged doesn't use energy (no static current)
- Output actively driven

10

Voltage Transfer Characteristic



- Output not going to rail
- Consequence?
  - → Voltage seen at the input can be degraded
- What can we do?
  - → Restoration with proper noise margins

11

Restoration

- Necessity observed throughout the course
  - Pass-transistor signal degradation (i.e. Vdd - Vthn)
  - Ratioed-logic noise margin
  - Inductive noise
  - Crosstalk noise
  - Reflections
- Becomes more important as the circuit complexity increases
  - Want to maintain robust signal everywhere

12

1st Order RC Delay Models

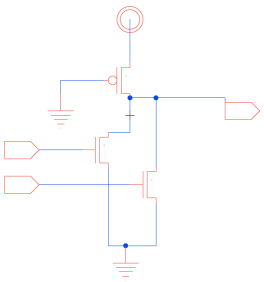
$\tau_{PHL} \approx 0.69 \cdot C_{load} \cdot R_n$ $\quad C_{load} \approx C_{dbn} + C_{dbp} + C_{int} + C_{gb}$

- Equivalent circuits used for MOS transistors
  - Ideal switch + "effective" ON resistance + load capacitance
    - Define unit resistance, R_u: "effective" ON resistance of transistor with min length and W=W_u (usually min width)
    - C_{gb} = C_g and C_{db} = C_{sb} = C_d for the unit n/pMOS transistors
  - Scale R and C with transistor sizing (W and L)
- Fan-out, driving stages, and sizing
- Identify worst case delay scenarios for different gates
- Tradeoff between large gates vs small gates (# stages, fanin/fanout)

13

Ratioed Logic

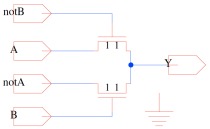
- Only build PDN (or PUN)
- Build NFET pulldown
  - Exploit high N mobility
- Pros:
  - Less transistor
    - Less area…?
    - Less capacitive load…?
- Cons:
  - Constant power dissipation
  - Need careful sizing (noise margin)
- Tradeoff between noise margin and area & capacitance

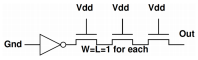


14

Pass Transistor Logic

- Simple switch-based logic
- Pros:
  - Less transistor…?
    - Less area…?
    - Less capacitive load…?
- Cons:
  - Needs restoration (buffering)
  - Can be slow
  - Limited voltage lowering for energy reduction
- Needs to take into account diffusion capacitance, ¥Cg⟶ Elmore delay





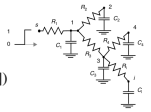
15

Elmore Delay: Distributed RC network

- The delay from source s to node i
  - N = number of nodes in circuit

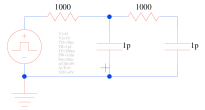
$$R_{ik} = \sum R_j \Rightarrow (R_j \in [path(s \rightarrow i) \cap path(s \rightarrow k)])$$

$$\tau_{Di} = \sum_{k=1}^{N} C_k R_{ik}$$



- Special Ladder Case:
- For each resistor C_i in path
  - Compute R_{ii} = sum of all Rs upstream of C_i

$$\tau_{DN} = \sum_{i=1}^{N} C_i \sum_{j=1}^{i} R_j = \sum_{i=1}^{N} C_i R_{ii}$$

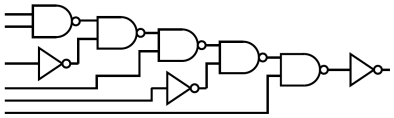


16

Design Abstraction and Performance

17

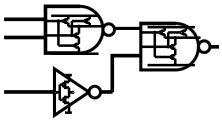
Digital Logic: Gate Level



- We care about design for performance
  - Functionality (e.g. F = A + B*C)
  - Speed
    - Each gate has a delay caused by the output resistance and capacitive load (which is the input capacitive load of the gate on the output)
    - Critical path defines delay
  - Power
    - Switching power (comprised of dynamic and short circuit power) and static power (I.e P_{tot}=P_{dyn}+P_{SC}+ P_{stat})
  - Area
    - For a gate the standard cell area, dependent on W and L of transistors

18

Digital Logic: Transistor Level



- We care about design for performance
  - Functionality (e.g. F = A + B*C)
    - Design for abstraction (VTC: switching voltage, high gain, noise margins)
  - Speed
    - Transistor sizing affects the output resistance and capacitance
  - Power
    - Switching power (comprised of dynamic and short circuit power) and static power (I.e $P_{tot}=P_{dyn}+P_{SC}+P_{stat}$). Transistor sizing affects drive current and impacts power consumption
  - Area
    - For a gate the standard cell area, dependent on W and L of transistors

19

Power Sources

- Static power
- Dynamic switching power
- Short circuit power

- $P_{tot} = P_{static} + P_{sw}$
- $P_{sw} = P_{dyn} + P_{sc} = a(C_{load}+C_{sc})V^2f$

- $P_{tot} \approx a(C_{load}+C_{sc})V^2f+VI'_s(W/L)e^{-Vt/(nkT/q)}$

20

Energy-Delay Tradeoff

- Tradeoff and V_{th} effect
  - Speed
  - Switching energy
  - Leakage energy
- Ignoring leakage,
  - Energy proportional to V^2
  - Delay proportional to 1/V
  - $E\tau$ - Energy & delay tradeoff
- From project, logic family, logic optimization, sizing, … Rich energy optimization space to explore

21

Design Space Dimensions

- Vdd
- Topology
  - Gate choice, logical optimization
  - Fanin, fanout, Serial vs. parallel
- Gate style / logic family
  - CMOS, Ratioed (N load, P load)
- Transistor Sizing
- Vth

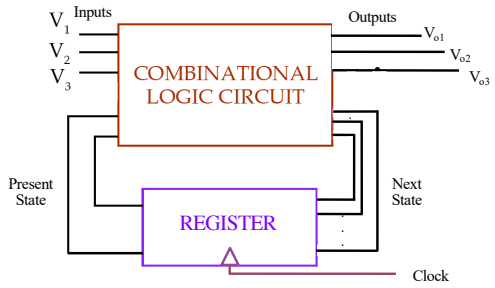
- The choices you make impact area, speed (delay), power

22

Sequential Logic

23

Sequential Circuit (or State Machine) Construct

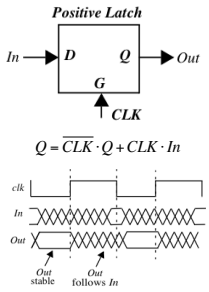


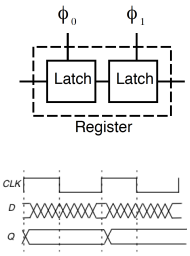
-> Register is used to Store Past Values of State(s) and Output(s)
-> Synchronous Sequential Circuit – clock, outputs change with clock event
-> Asynchronous Sequential Circuit – no clock, outputs change after inputs change

24

Memory Storage (Latches and Registers)

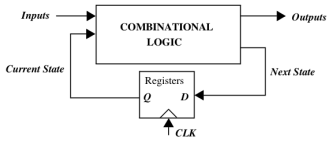
- *Level-sensitive device*
- *Edge-triggered storage element*





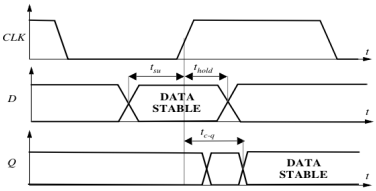
25

Latch Timing Issues



$$T \geq t_{c\text{-}q} + t_{plogic} + t_{su}$$

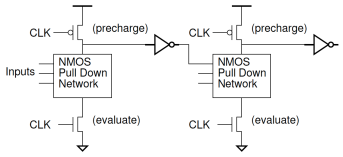
$$t_{cd\text{register}} + t_{cd\text{logic}} \geq t_{hold}$$



26

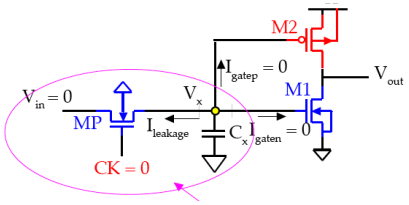
Domino Logic

- Single transition
  - Once transitioned, it is done → like domino falling
- All inputs at 0 during precharge
  - 'Outputs' pre-charged to 1 then inverted to 0
    - I.e. Inputs are pre-charge to 0
- Non-inverting gates



27

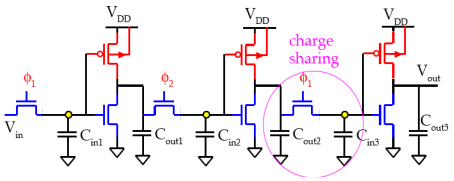
Charge Storage and Leakage



- Assume logic-high is stored onto V_x during active phase (CK=1)
- When CK=0, V_{in}→0

28

Charge Sharing



When $V_{out(i)}$ = 0V (or 5V) and $V_{in(i+1)}$ = 5V (or 0V) for i = 1,2 (stage)

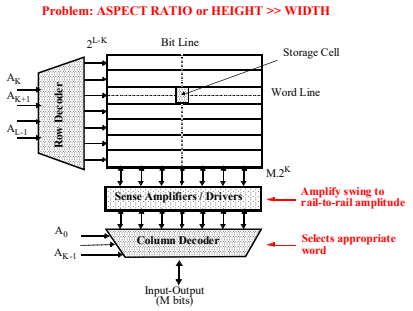
"Charge Sharing" is an issue when φ_1 and φ_2 switch

29

Memory

30

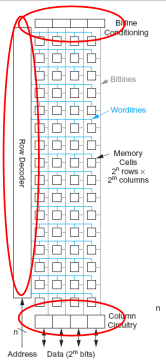
Array-Structured Memory Architecture



31

Array Architecture

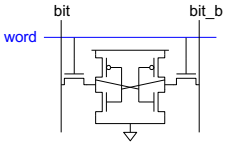
- Periphery Circuits
  - Row Decoders
    - Address bits for each word to reduce inputs
  - Bitline Conditioning
    - Precharging bitlines
    - Driving bitline capacitance
  - Sense Amplifiers
    - Amplify bitline voltage change to get rail-to-rail output and faster operation
  - Column Decoders



32

Read-Write Memory Cells

- Static (SRAM)
  - Data stored as long as supply is applied
  - Large (~6 transistors/cell)
  - Fast
  - Differential
- Dynamic (DRAM)
  - Periodic refresh required
  - Small (1-3 transistors/cell)
  - Slower
  - Single ended



33

VLSI Design for Manufacturing

34

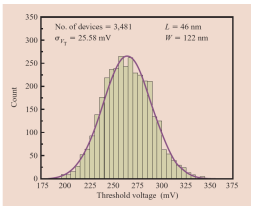
Structured Design Strategies

- Strategies common for complex hardware and software projects
  - Hierarchy: Subdivide the design in several levels of sub-modules
  - Modularity: Define sub-modules unambiguously and well defined interfaces
  - Regularity: Subdivide to max number of similar sub-modules at each level
  - Locality: Max local connections, keeping critical paths within module boundaries

35

Impact

- Changes parameters
  - W, L, t_{OX}, V_{th}



V_{th} Variability @ 65nm

- Higher V_{th}?
  - Not drive as strongly
  - $I_{d,vsat} \propto (V_{gs}-V_{th})$
  - Increased Delay
- Lower V_{th}?
  - Don't turn off as well → leak more

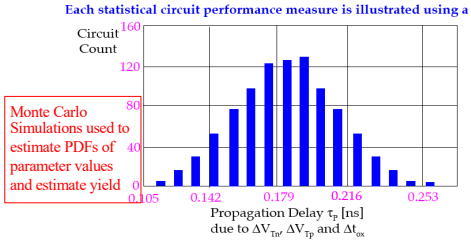
36

Parametric Yield

$\mathbf{r} = \mathbf{r}(\vec{\mathbf{x}}) = \mathbf{r}(\vec{\mathbf{d}} + \vec{\mathbf{s}})$ actual performance - random variable

$\mathbf{r}^0(\vec{\mathbf{d}}) = \mathbf{r}(\vec{\mathbf{d}} + \vec{\mathbf{s}}^0)$ nominal or designed performance - deterministic variable

Each statistical circuit performance measure is illustrated using a histogram

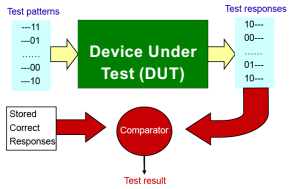


Monte Carlo Simulations used to estimate PDFs of parameter values and estimate yield

$$\text{parametric yield}\,(Y) = \frac{\text{total number of acceptable circuits}}{\text{total number of manufactured circuits}} \times 100\%$$

37

Testing Principle



- Observability: measure of the ease of observing a node by watching external output pins of the chip
- Controllability: measure of the ease of forcing a node to 0 or 1 by driving input pins of the chip

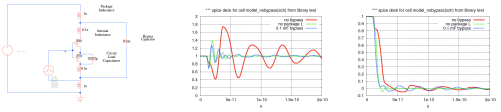
38

Wiring and Interconnect Considerations

39

Inductive Noise

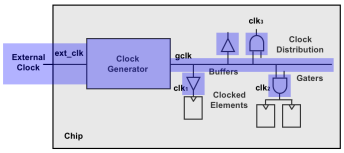
- Inductive noise
  - Originates in signal paths and supplies
  - Minimize wires when possible and add bypass capacitors



40

CLK Generation

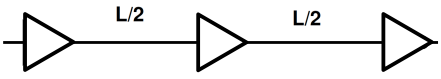
- CLK design and distribution is necessary for correct operation and timing
  - Tree, Grid, Mesh, H-Tree
  - CLK Gating



41

Optimally Buffered Wire Delay

- Wire delay linear once buffered optimally
  - Optimal buffering equalizes delays
    - Buffer delay, Delay on wire between buffers, Delay of wire driving buffer

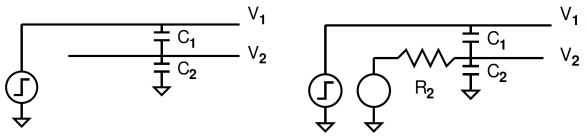


$$N = \sqrt{\frac{0.5 R_{wire} C_{wire}}{R_{buf}\left(C_{self} + C_{load}\right)}} \qquad W = \sqrt{\frac{R_{un} C_{wire}}{2 R_{wire} C_g}} = \sqrt{\frac{R_{un} C_{unit}}{2 R_{unit} C_g}}$$

42

Crosstalk

- Capacitive crosstalk
  - Diversion and recovery
  - Clocked and driven wires
    - Slow down transitions
  - Undriven wires voltage changed
  - Can cause spurious/false transitions



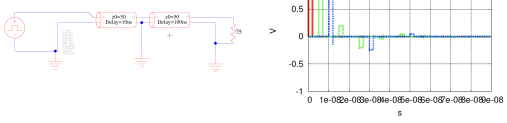
43

Transmission Lines

- Pulses travel as waves
- Transmission lines
  - high-speed
  - high throughput
  - long-distance signaling
- Termination

$$w = \frac{1}{\sqrt{LC}} = \frac{c_0}{\sqrt{\varepsilon_r \mu_r}} \qquad Z_0 = \sqrt{\frac{L}{C}}$$

$$V_r = V_i\left(\frac{R - Z_0}{R + Z_0}\right)$$



44

Admin

- Final Project
  - Design CLB
    - EC for best figure of merits
      - FOM = Area*avgEnergy/maxFrequency
  - Due 12/10 (last day of class)
    - Can turn in up to 12/17 with no penalty

45