ECE 425

Spring 2026

Review Notes for ECE 425 (Intro to VLSI System Design)

These notes aren’t fully comprehensive.

Aniketh Tarikonda (aniketh8@illinois.edu)

Contents

1 Intro

- the semiconductor market is growing, and thus we need VLSI people (literally entirety of lecture 1)

- (Brief) History of Computers:
  - Until the 20th century, we had mechanical computers, abacuses, etc.
  - Vacuum tubes are invented, and can implement boolean logic. First computer is built with vacuum tubes.
  - Vacuum tubes are replaced by transistors
  - Transistors decreased in size, complexity of systems increase
  - ICs invented, can print transistors through lithography. Early LSI era.
  - Technology improves → higher resolution → hundreds of billions of transistors on chips. VLSI era.

- Modern chips use CMOS (complementary nMOS and pMOS networks) to implement digital logic.

Obligatory Moore's Law and Dennard Scaling Mention rahh I can never escape it

- Dynamic MOSFET Power Consumption: $P_{\text{dyn}} = nCV^2 f_{\text{clk}}$

- If dimensions of the transistor scale by $\sim 0.7x$, area scales by roughly a half.
  - Capacitance (C) and Voltage (V) scale linearly wrt. dimensions, f_{clk} scales by $\frac{1}{0.7}$
  - $2x$ transistors in the same area
  - End result is that scaling dimensions has *no* effect on P_{dyn}
  - This observation was the basis for **Dennard Scaling**

- Dennard Scaling ended around 2005/2006

- Economies of Scale, increasing R&D costs for fabrication lead to companies outsourcing fabrication to certain specialized companies (e.g. TSMC, GlobalFoundries, etc.)
  - Rise of EDA industry and global standards for semiconductors (e.g. GDSII)
  - Tools, libraries, PDKs, etc.

- Semiconductors have short market windows, short product life cycles, stiff competition
  - Certain chips need to be low cost, some need to have really good power efficiency

- Modern ASIC/Chip Design Workflow: Design → Architecture → RTL → Gate-Level Netlist → Physical Design (floorplanning, layout, pnr) → fab does their thing, we do Post-Silicon Validation

pagebreak()

2 Midterm 1

2.1 Intro to MOS Transistors

- We build chips out of silicon because its a semiconductor, has four valence electrons and can form nice crystal lattices (covalent bonds) , has nice thermal properties, and is relatively abundant.
- We can dope it with an element that has 3/5 valence electrons so that we introduce holes and electrons which travel around the lattice (doping)
  - Doping is generally done via diffusion: exposing silicon to superheated phosphorus/boron gas.
- Electrons drifting → Current, improves conductivity
- **n-type semiconductors have extra electrons**
- **p-type semiconductors have extra holes (lack of electrons)**

PN Junctions

- a P-N junction forms a diode
- Initially the electrons in the n-type fill the holes near the junction, forming a **depletion layer**
- When we put a higher electric potential on the anode (p-type side), current will flow (assuming its greater than the threshold voltage)
  - This is called **forward biasing**
- Alternatively, we can **reverse bias** the PN junction, which causes the depletion layer to grow and current to stop flowing.

nMOS transistors (invert p and n for pMOS)

- depletion layer forms around n-wells
- only have current when p-substrate has a higher potential
- four terminal device: gate, source, drain, body
- When gate voltage increases beyond a threshold:
  - An inversion region forms under the gate with electrons as charge carriers
  - Creates an n-channel: current flows from source to drain.
- We want to keep n-well at a higher potential than the p-type substrate (otherwise we forward bias it)
  - Body connection for nMOS is to GND (reverse biased)
- Holes move slower than electrons[1] by a factor of $2 - 3$x, which is why pMOS transistors are usually sized $\sim 2 - 3$x larger than nMOS.
- NMOS passes logical 0 well, passes a degraded 1. PMOS passes logical 1 well, but degraded 0.

[1] Technically holes are just the lack of electrons. In any case, this is because holes “travel” in the valence band, whereas electrons travel in the conduction band.

- **CMOS** - combination of PMOS and NMOS
  - pull-up network (PUN) of pMOS, pull-down network (PDN) of NMOS
  - if PUN and PDN are both on, we have a short circuit.
  - if PUN and PDN are both off, the output is floating (high-Z)

- PUN is the logical complement of PDN

- **Demorgan's Law**
  - $(A' + B') = (AB)'$
  - $(A'B') = (A + B)'$

2.2 Intro to Layout

Layout Design Rules

- this is very idealized because we're using a relatively ancient process node (FreePDK45nm)
  - modern fabrication processes are *significantly* more complex

2.2.1 Lambda (λ) Design Rules

- λ coresponds with half of the minimum feature size
- feature size is the minimum transistor channel length, or the minimum width of the polysilicon wire
- allows easy scaling for different (old) processes.
- not applicable to modern (sub 90nm) processes

Rules:

1. metal and diffusion have minimum width and spacing of 4λ
2. contacts are $2\lambda \times 2\lambda$, surrounded by λb on layers above and below
3. polysilicon width is 2λ
4. polysilicon and contacts have spacing of 3λ from others
5. polysilicon overlaps by 2λ where it is desired, spacing of λ away from areas where no transistor is desired
6. n-well surrounds pMOS by 6λ, avoids nMOS by 6λ

2.2.2 Guides for Optimized CMOS Layouts

- optimize boolean expression before drawing stick diagram, layout
- horizontal V_{DD} rail on top, GND rail on bottom
  - p-diffusions close to V_{DD}
  - n-diffusions close to GND
- minimize metal lengths
- polysilicon lines are high-Ω, and should generally run vertically. Avoid turns
- **merge diffusions**
  - e.g. NAND gate, drain and source of PUN can be merged. the drains of PDN can also be merged.
  - large savings on area, routing, performance

- size transistors appropriately such that equivalent resistances remain somewhat minimized

Gate Layout with Euler Paths

- draw schematic
- find Euler path (doesn't have to end at the starting point)
- ensure label/ordering is the same for PUN/PDN
- If you do this correctly, you can create designs with nice, straight polysilicon.
- can be not-so-trivial at times, this is an NP-hard problem
- Avoid multiple metal layers, if possible (leave room for routing)
- Don't forget metal-poly, metal-metal, metal-diffusion contacts

2.2.3 Common Combinational and Sequential Circuit Elements

- Most are self-explanatory, really basic
  - AOI22: "and - or - invert" ($Y =\sim ((AB) \mid (CD))$)
    - AOI21 just passes C (no 2nd NAND gate)
  - OAI22: "or - and - invert" ($Y =\sim ((A \mid B) \,\&\, (C \mid D))$)
    - OAI21 just passes C (no 2nd NOR gate)

Non-restoring Transmission Gate

- nMOS and pMOS in parallel
- called "non-restoring" because output voltage isn't being driven by V_{DD} or GND
  - signal slowly gets degraded as you put many non-restoring gates in series

Tri-States

- a transmission gate is one way to build a (non-restoring) tri-state, when $\text{EN} = 0$, the output is high-Z
- A Restoring Tri-State Inverter uses two pMOS and two nMOS in series, outputs are directly driven by V_{DD} and GND

Multiplexers

- can be built via NAND/NOR/AOI22 gates, but not very optimal
- can be built with two transmission gates (non-restoring)
- can be build with a pair of tri-state inverters
- Larger muxes (e.g. 4-1, 8-1) can be build hierarchically using 2-1 muxes, or flattened (4 or 8 tristates)
- With multiplexers, inverters, and tri-states, you can build sequential elements such as D-latches.

- By placing two D-Latches in series with an inverter between their CLK inputs, we create a DFF (posedge-triggered FF)
- Back-to-Back DFFs can malfunction due to clock skew, race conditions
  - Thus, we can insert buffers/gates to add some combinational delay between the DFFs[2]

2.3 MOS Transistor Theory

- Naming: "Source" of a MOSFET is the source of majority charge carriers (electrons in nMOS, holes in pMOS)
- Three operation modes: Accumulation, Depletion, inversion
  - Accumulation: $V_g - V_b < 0$
  - Depletion: $0 < V_g - V_b < V_t$
  - Inversion: $V_g - V_b > V_t$
- Three Regions in the Shockley (wild ass wiki article btw) Model: Cutoff, Linear, Saturation
- When $V_{\text{gs}} < V_t$, the device is in cutoff (no channel forms)
- Channel forms when V_g exceeds channel voltage V_c by at least V_t
  - $V_c \approx V_s$ near source, $V_{\text{gs}} = V_g - V_s > V_t$, and so the n-channel forms
  - Near the drain, $V_c \approx V_d$, and so $V_{\text{gd}} = V_{\text{gs}} - V_{\text{ds}} < V_t$, and so the n-channel is pinched off
- In saturation $V_{\text{ds}} > V_{\text{ov}}$, current will still flow, but theres no dependence on V_{ds}[3]

Deriving the Ideal MOSFET Equation(s)

- Overview: MOS structure is effectively a parallel plate capacitor (important thing to keep in mind for later discussions on timing/delay)
- How can we derive the amount of charge in the channel, and how long each charge takes to cross it?
- Capacitor Eqn: $Q_{\text{charge}} = C_g V$ where C_g is the gate capacitance which we can derive using the parallel plate capacitor eqn:
- $C_g = \varepsilon_{\text{ox}} \frac{WL}{t_{\text{ox}}} = C_{\text{ox}} WL.$
  - ε_{ox} is the permittivity of the gate oxide, t_{ox} is the gate thickness, WL is the area (width times length)
- Gate voltage $V = \left(V_g - V_c\right) - V_t$.
  - We can use the approximation $V_c \approx \frac{V_s + V_d}{2}$ to simplify this to $V = \left(V_{\text{gs}} - \frac{V_{\text{ds}}}{2}\right) - V_t$
- With the gate voltage/capacitance, we can approximate the charge that forms under the gate. However, $I = \frac{\Delta Q}{\Delta t}$, so we also need to derive this Δt.

[2] Don't overdo this or we end up with setup time failures

[3] V_{ov} is the overdrive voltage, defined as $V_{\text{gs}} - V_t$ in nMOS (or $V_{\text{sg}} - |V_t|$ in pMOS)

- Electric Field (lateral) is $E = \frac{V_{\text{ds}}}{L}$ where L is channel length. We also know $v_{\text{electron}} = \mu E$ where μ is electron mobility.
  - time required for an electron to cross the channel is $t = \frac{L}{v_{\text{electron}}}$
  - $\Delta t = \frac{L}{v} = \frac{L^2}{\mu V_{\text{ds}}}$
- With these two components, we can now derive the MOSFET IV-equation for an nMOS in the linear region (Shockley model)
  - $$I_{\text{ds}} = \mu C_{\text{ox}} \left(\frac{W}{L}\right) \left(V_{\text{ov}} V_{\text{ds}} - \frac{1}{2} V_{\text{ds}}^2\right) \tag{1}$$
- To calculate the current at saturation, we plug in $V_{\text{ds}} = V_{\text{ov}}$, at which point we get:
  - $I_{\text{ds}} = \frac{1}{2} \mu C_{\text{ox}} \left(\frac{W}{L}\right) \left(V_{\text{ov}}^2\right)$
- How do we increase I_{ds}?
  - Increase amount of charge in the channel
    - Increase t_{ox} (not in your control)
    - Increase L (bad idea, charges have to cross a longer distance)
    - Increase transistor width W (this is a good idea)
    - Increase gate voltage (also a good idea, **although there are caveats**)
  - Decrease time to cross channel
    - Decrease L (this is intrinsically limited by your minimum feature size)
    - Increase V_d / V_{ds}

2.3.1 Non-Ideal Effects

- There are two E-fields in a MOSFET
  - a vertical one $E_{\text{vert}} = \frac{V_{\text{gs}}}{t_{\text{ox}}}$
  - a lateral one $E_{\text{lat}} = \frac{V_{\text{ds}}}{L}$
- Electrons are attracted upwards due to E_{vert}, directed them right towards the gate oxide
  - They collide/scatter off of the substrate-oxide interface, degrading velocity (lowering mobility)
  - Increase total drain-to-source path
- To account for E_{vert}, we introduce $\mu_{\text{eff}} < \mu_e$
  - μ_{eff} is determined by best-fitting experimental data.
  - $E_c = 2 \frac{v_{\text{sat}}}{\mu_{\text{eff}}}$

$$v = \begin{cases} \frac{\mu_{\text{eff}} E}{\left(1 + \frac{E}{E_c}\right)} & E < E_c \\ 1 & E \geq E_c \end{cases} \tag{2}$$

Channel Length Modulation

- A reversed-biased PN-junction forms a depletion region

  - Width of depletion region L_d grows when when we further reverse-bias it

- Thus, the channel length is actually smaller when we take this into account:
  - $L_{\text{eff}} = L - L_d$, shorter L_{eff} means higher current
  - Increasing V_{ds} past saturation grows the depletion region, thus increasing L_d, thus resulting in a shorter L_{eff}, which results in a higher current.

- To account for this, we add a term to the saturation equation to linearly approximate this contribution:

$$I_{\text{ds}} = \frac{1}{2}\mu C_{\text{ox}}\left(\frac{W}{L}\right)(V_{\text{ov}})^2(1 + \lambda V_{\text{ds}}) \tag{3}$$

Threshold Voltage Effects

- Depends on body voltage, drain voltage, channel length
- intrinsically depends on material, doping concentration, oxide geometry, temperature, etc.
- Body Effect: $V_t = V_{\text{t0}} + \gamma\left(\sqrt{\varphi_s + V_{\text{sb}}} - \sqrt{\varphi_s}\right)$
  - γ is the body coefficient, φ_s is the fermi Potential.
- Drain-Induced Barrier Lowering (DIBL)
  - high V_d creates a large depletion region
  - depletion region spills over to the channel
  - easier for gate to create inversion layer in the channel
  - Another linear approximation: $V_{\text{t'}} = V_t - \eta V_{\text{ds}}$ (threshold voltage decreases)

Leakage Currents

- **Subthreshold Leakage**: Current flowing from source to drain when in cutoff
  - Caused by weak inversion layers, short-channel effects, thermal agitation
  - The primary leakage concern for lowly students in ECE425

$$I_{\text{ds}} = I_{\text{ds0}} \exp\left[\frac{V_{gs} - V_{t0} + \eta V_{ds} - k_\gamma V_{sb}}{n v_T}\right]\left(1 - \exp\left[\frac{-V_{ds}}{v_T}\right]\right) \tag{4}$$

$$I_{\text{ds0}} = \beta v_T^2 \exp[1.8] \tag{5}$$

- Gate Leakage: current flowing between gate and body
  - Charges can quantum tunnel through the gate oxide

- Junction Leakage: current flowing between source and body, drain and body
  - Diode Leakage: fast/heated electrons can cross the depletion region
  - Band-To-Band Tunnel (BTBT): electrons can tunnel through the PN-junction.
  - gate-induced drain leakage (GIDL): high V_d, low V_g causes pronounced BTBTn beneath gate overlap

Temperature Effects

- Higher temperature causes reduced mobility, lower V_t
- Electrons now have more energy and are more likely to tunnel
- Exacerbates most of the other non-ideal effects
- This is why you should keep your chips relatively cool (or at least be smart with packaging)

Parameter Variation

- We can't expect transistors to behave ideally
- "Fast Assumption" : L_{eff} is short, V_t is low, t_{ox} is thin.
- "Slow Assumption" : L_{eff} is long, V_t is high, t_{ox} is thick.
- "Typical Assumption" is in the middle of these two.
- We need to ensure our design works for all of these assumptions
- **Process Corners** are a way of graphically representing this parameter variation