ESE370: Circuit Level Modeling, Design, and Optimization for Digital Systems

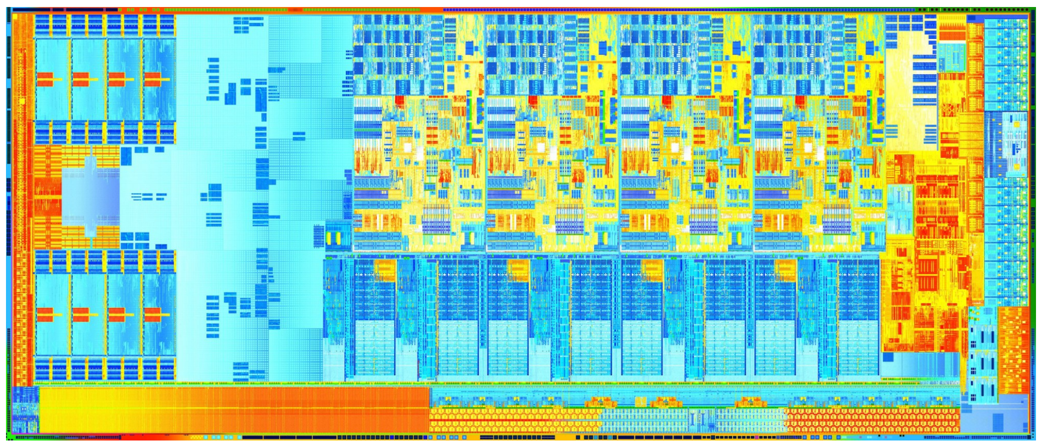
Lec 36: December 10,
Review

Objective

- At circuit level, how do we implement robust digital systems that are...

-High-speed
-Low-power
-Area-efficient

with given technology



3rd generation Intel Core

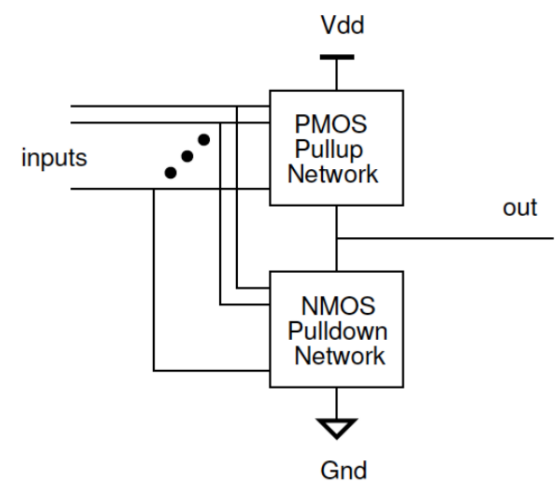
Learning

- How to **model** digital logic and reason about behaviors and performances
- How to **design** circuits and perform simulations for functional verification and performance analysis
- How to **optimize** circuits using various techniques

CMOS Logic

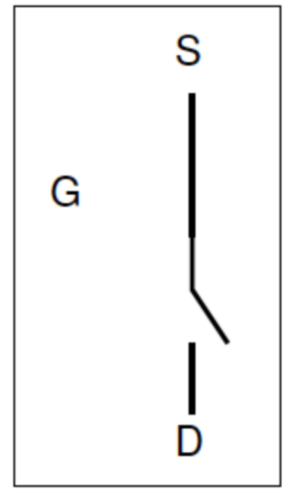
- Complementary metal-oxide semiconductor
- PMOS pullup network implements f
- NMOS pulldown network implements f/

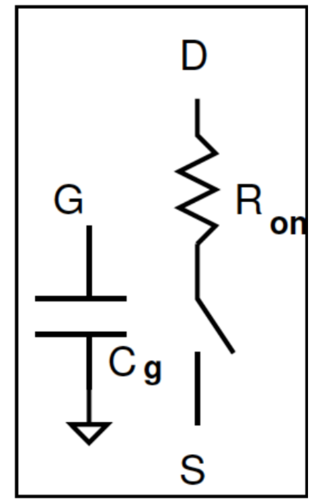
-Why not the other way around?



MOSFET Model

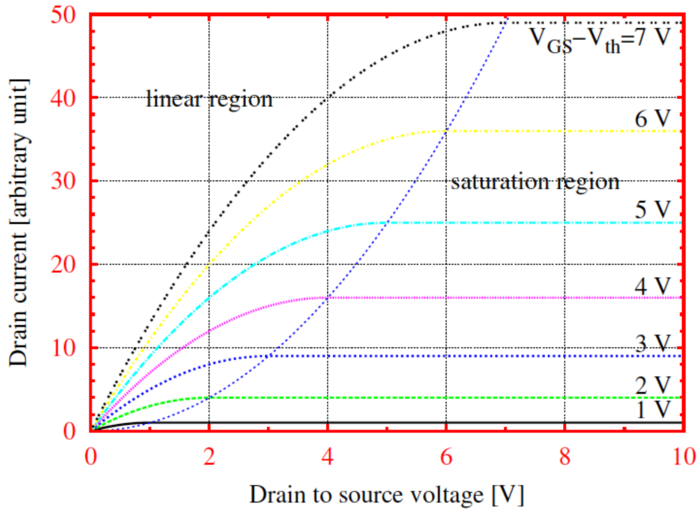
- Zeroth-order model:
  - Transistor as switch
- First order model:
  - Transistor as resistive driver (R_{on})
  - Transistor's gate as capacitive load (C_g)
- What do models tell?
  → Reason about logic
  → Reason about RC delay

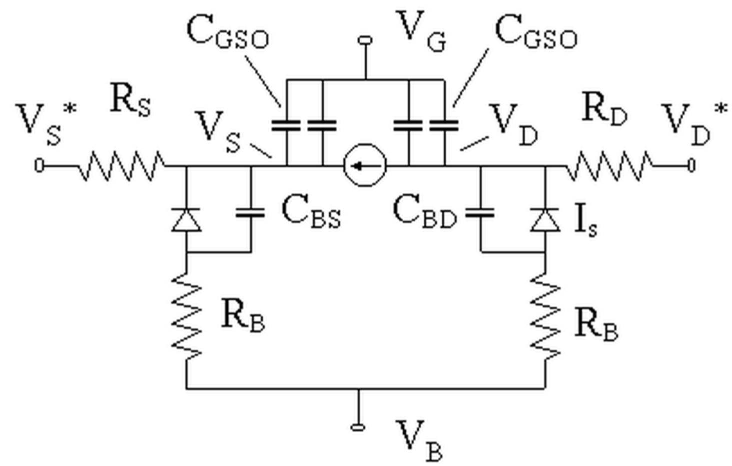




MOSFET Model

- Spice model of NMOS
- Parameters should look familiar





http://ecee.colorado.edu/~bart/book/book/chapter7/ch7_5.htm

Restoration

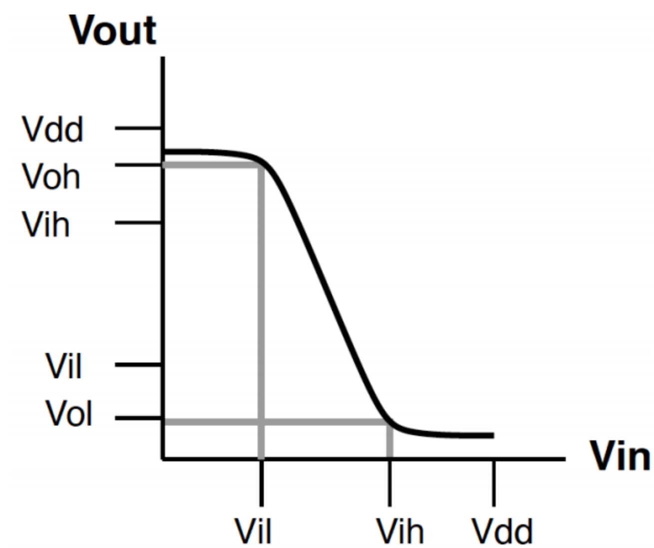
- Output not going to rail
- Noise problems

-Consequence?

→ Voltage seen at the input can be degraded

-What can we do?

→ Restoration with proper noise margins



Definition:

$NM_h = V_{oh} - V_{ih}$

: How much "high" output voltage can drop and still be recognized as "high"

$NM_l = V_{il} - V_{ol}$

: How much "low" output voltage can rise and still be recognized as "low"

Restoration

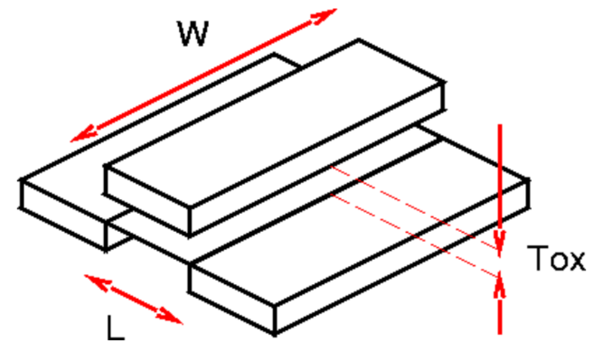
- Necessity observed throughout the course
  - Pass-transistor signal degradation (i.e. V_{dd} - V_{thn})
  - Ratioed-logic noise margin
  - Inductive noise
  - Crosstalk noise
  - Reflections

- Becomes more important as the circuit complexity increases
  - Want to maintain robust signal everywhere

MOS Transistor Operation

- Operating regions:

-Sub-threshold (cutoff), $V_{gs} < V_{th}$
-Resistive (linear), $V_{gs} > V_{th}$
-Saturation (active), $V_{ds} > V_{gs} - V_{th}$



- Strength scales by (W/L)
- Channel as varying resistance
- Short channel effects

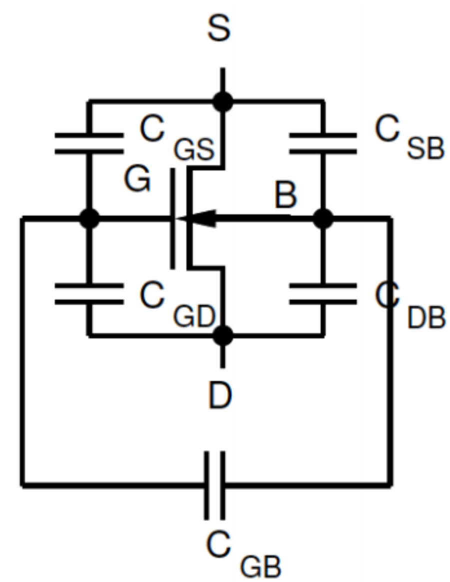
-Velocity saturation
-Drain-Induced Barrier Lowering (DIBL)
-Hot electron effect

MOS Transistor Capacitance

- Capacitances at each terminal of the transistor
- Capacitances vary with V_{gs} (operating region)

- Implications…?

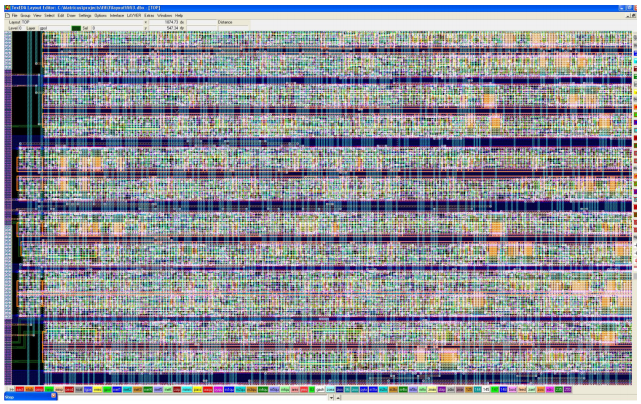
→ Overshooting

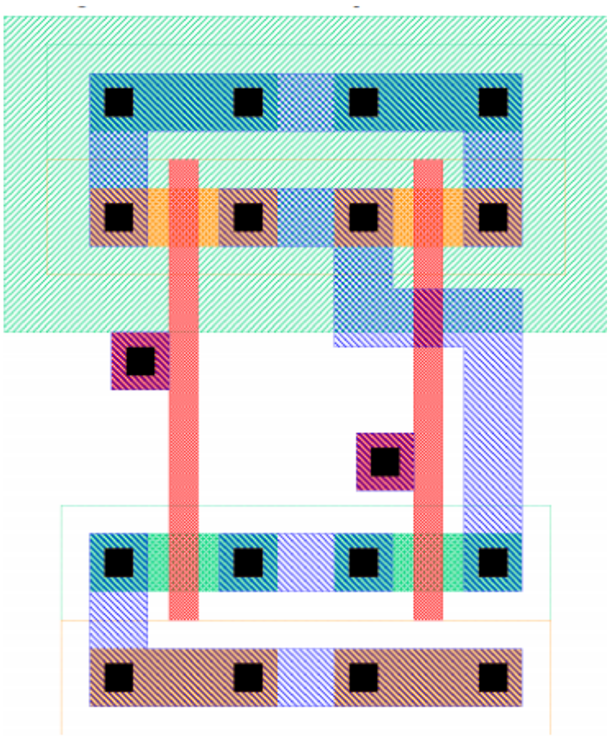


Layout

- Can identify what each part of layout is/does
- Design rules for fabrication
- Multiple metal layers for routing
- Bigger picture?

- Interconnect and parasitics
- Effects on performance



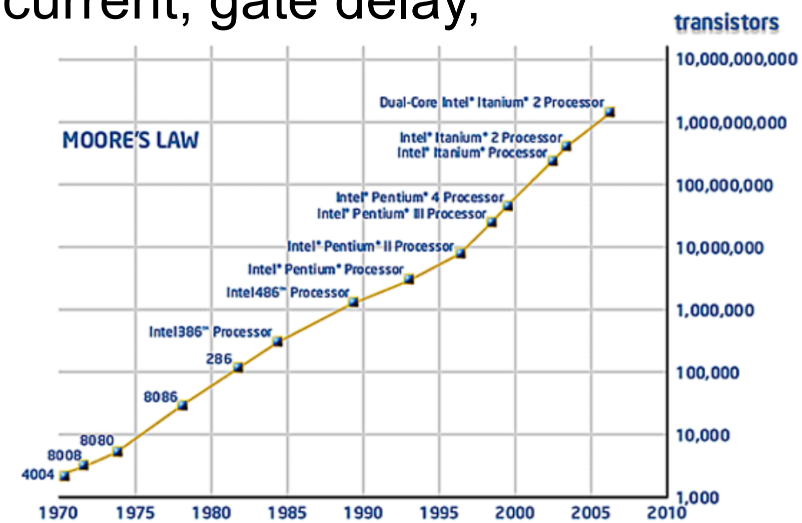


Scaling

- 32nm → 22nm → 14nm → 7nm
  - Final ITSR report 2015
- Observed scaling of:

Area, capacitance, resistance, threshold, current, gate delay, wire delay, and power

- Will Moore's law continue?
- Implications:
  - Material-science view
  - Power density limits
  - Other options for improvement...



τ Model

- $\tau = R_0C_0$ modeling for delay

$$C = \varepsilon_r \varepsilon_0 \frac{A}{d}$$

- Impact of transistor sizing (W and L) on R and C
- Fan-out, driving stages, and sizing

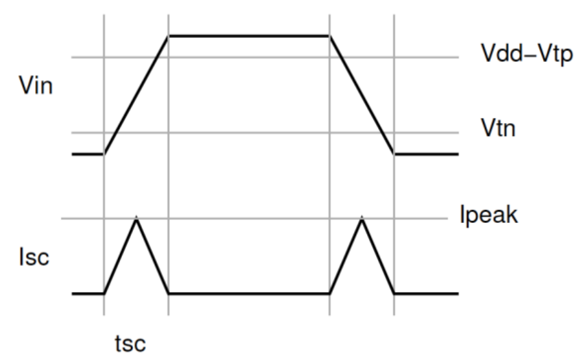
- i.e. Multiple inverter stages

$$R = \frac{\rho L}{A}$$

- Identify worst case delay scenarios for different gates
- Tradeoff between large gates vs small gates (# stages, fanin/fanout)

Energy and Power

- Static
  - Subthreshold leakage, gate-drain leakage
- Capacitive switching
  - Charge & discharge output load
- Short Circuit
  - When both N and P devices are on



Energy and Power

- Why Important…?

Capacitive Switching

$$E = V_{dd} \int I(t)dt$$

$$Q = CV = \int I(t)dt$$

$$E = CV_{dd}^2$$

Short Circuit

Q=I×t

Q=CV

$$E = V_{dd} \times \left(I_{peak} \times t_{sc} \times \left(\frac{1}{2} \right) \right)$$

$$E = V_{dd} \times Q_{sc}$$

$$E = C_{sc} V_{dd}^2$$

Static

$$I_{DS} = I_S \left(\frac{W}{L} \right) e^{\left(\frac{V_{GS}}{nkT/q} \right)} \left(1 - e^{-\left(\frac{V_{DS}}{kT/q} \right)} \right) \left(1 + \lambda V_{DS} \right)$$

Energy and Power Optimization

- Ignoring leakage,

- Energy proportional to V^2
- Delay proportional to 1/V
  - $\rightarrow E\tau$ - Energy & delay tradeoff

- V_{th} effect on
  - Speed
  - Switching energy
  - Leakage energy

- $P_{tot} = P_{static} + P_{sc} + P_{dyn}$
- $P_{dyn} + P_{sc} = a(½C_{load}+C_{sc})V^2f$
- $P_{tot} \approx a(½C_{load}+C_{sc})V^2f+VI'_s(W/L)e^{-Vt/(nkT/q)}$

- From project, logic family, logic optimization, sizing, …

Rich energy optimization space to explore

Ratioed Logic

- Build single pull-up (pull-down) control network
- Compared to CMOS,

Pros:

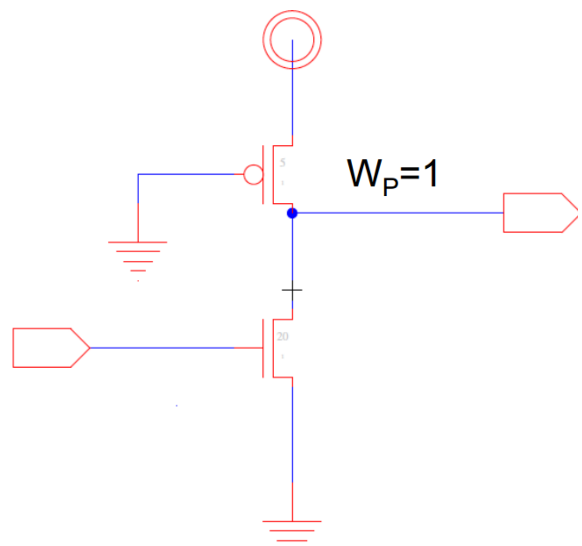
- Less transistor

→ Less area…?

→ Less capacitive load…?

Cons:

- Constant power dissipation
- Need careful sizing (noise margin)



Tradeoff between noise margin and area & capacitance

Pass Transistor Logic

- Simple switch-based logic
- Compared to CMOS,

Pros:

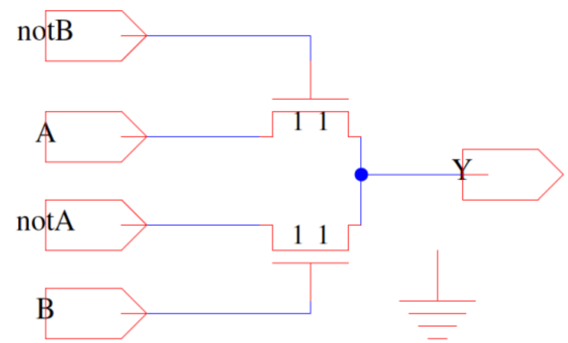
- Less transistor…?

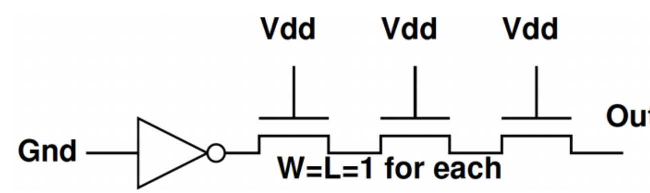
→ Less area…?

→ Less capacitive load…?

Cons:

- Needs restoration (buffering)
- Can be slow
- Limited voltage lowering for energy reduction





Pass transistor with restoration stages vs CMOS

Needs to take into account diffusion capacitance, γC_g→ Elmore delay

Elmore Delay

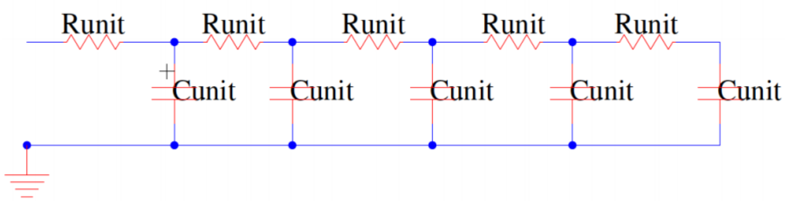
- Chain of pass-transistors modeled into RC chain

→ Cannot use simple τ model for delay (What if ɣ = 0?)

→ Use refined model for more accurate delay calculation

- Where else do we observe RC chain?

- Wire



$$Delay = \sum_{path} \left(R_i \times \sum_{i \xrightarrow{path} j} C_j \right)$$

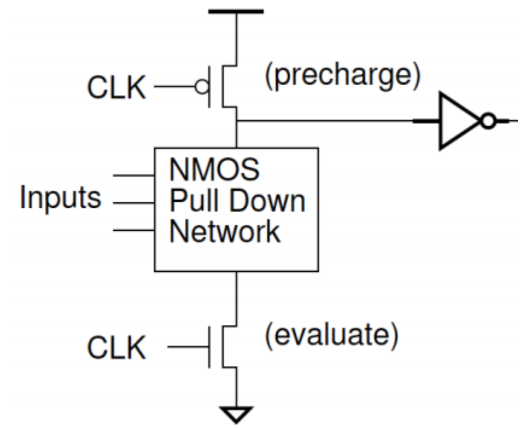
- Using elmore model,

Runit*(N*Cunit)+Runit((N-1)*Cunit
+Runit*(N-2)*Cunit+…+Runit*Cunit
=(Runit*Cunit)*(N+N-1+N-2+….1)
=Runit*Cunit*N^2/2

Synchronous Circuits & Clocking

- Reuse logic resources

→ Add state elements (latches, registers)



- Clocking discipline

- Setup and hold times
- clk → q delay for data output from state element

- Clocking can be used for dynamic logic family
  - Domino logic

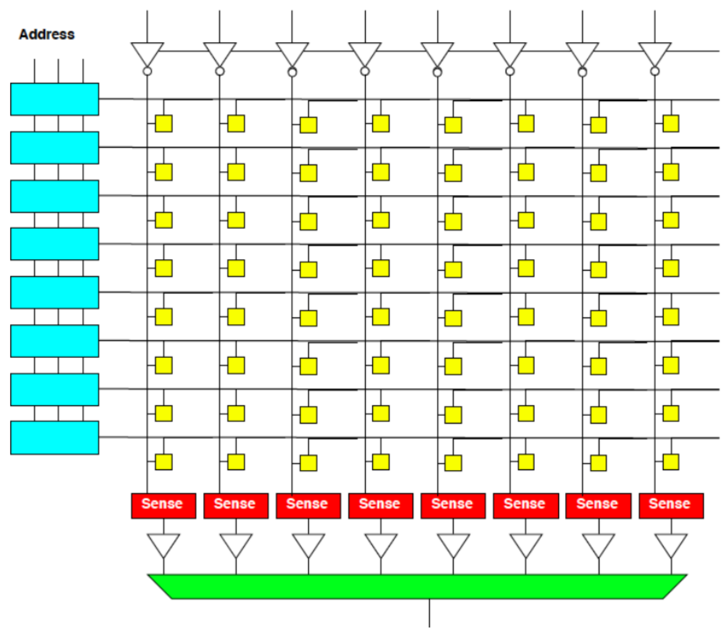
- Disable pull-up (pull-down) during evaluation; no static power
  - Needs precharge for disabling next-stage NMOS evaluation
  - Can involve extra clocking energy, precharge energy, complexity

Memory

- Memory bank organized for
  - Economic wire and area usage
  - Maximize storage density
  - Share peripheries

- Main components:

- SRAM Memory cell: cross-coupled inv.
- Write drivers (tristate buffers)
- Decoder (column/row)
- Precharge
- Sense amplifier



Memory

- What did we use to build?

  - CMOS
    - Memory cell cross-coupled inverter
    - Buffers
  - Pass transistor
    - Access transistor
    - Decoder
  - Ratioed Logic
    - NAND/NOR ROM
    - 6T SRAM sizing

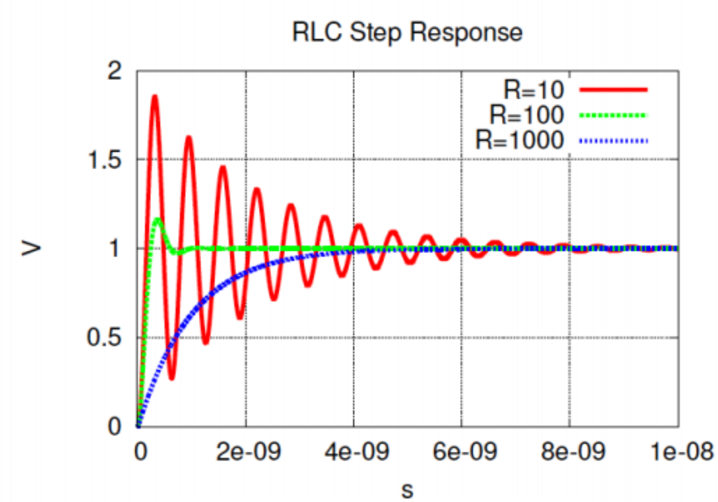
And more for energy optimization…

Memory

- Robustness
  - Charge-sharing effect and read-write upsets
  - → Need to carefully size cell
  - One solution was to use pre-charge of Vdd/2
  - → Prevent voltage swing and read-write upset
- Scaling
  - Deeper (more rows) memory will need strong driver and precharge
  - Wider (more columns) memory will need strong addressing
- Implications
  - Want to use high V_{th} from energy stand-point (sacrifice speed for energy)
  - Routing wires also scale with memory size
  - → Need to be concerned about parasitic capacitances, crosstalk, noise...

Inductive Noise

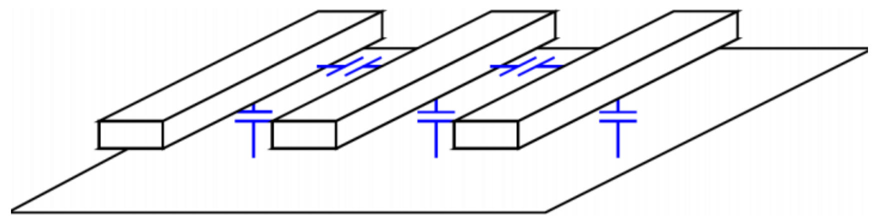
- Sources?
  - Wire (scales with length)
  - Bond & package pins
- Where?
  - Signal paths and power supplies
- Problems?
  - RLC response
  - Oscillation can dominate (HW8)
- Solutions?
  - Make wire short
  - Bypass capacitor (Lab 2)



$$R < \sqrt{\frac{4L}{C}}$$

Crosstalk

- Sources?
  - Wire (scales with size/spacing)
- Where?
  - Cables
  - PCB wires
- Problems?
  - Noise
  - Spurious transition
- Solutions?
  - Orthogonal routing
  - Increase pitch
  - Separate with ground/power shield wires (lab2)



Transmission Lines

- From LC lossless transmission line model
  - Signal propagates as wave down transmission line

- Behaviour at the end of the line is determined by termination type (short, open, or specific impedance)

- Where would the termination matching be important?
  - Vias
  - Branches
  - Cable-to-cable
  - Board-to-cable

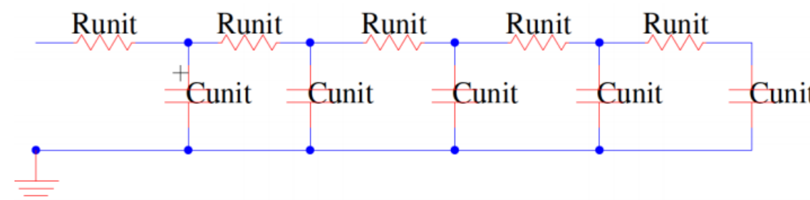
$$w = \frac{1}{\sqrt{LC}} = \frac{c_0}{\sqrt{\varepsilon_r \mu_r}}$$

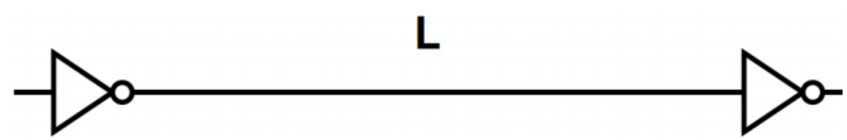
$$V_i\left(\frac{R - Z_0}{R + Z_0}\right) = V_r$$

$$V_i\left(\frac{R - Z_0}{R + Z_0} + 1\right) = V_t$$

Repeaters in Wiring

- Observed delay problems in RC chain
  - Delay will scale by L^2 (N^2)
  - Elmore delay
- How do we minimize delay?
  → Buffer the wires
- Parameters to consider for buffering
  - # of buffers
  - Length of segment
  - Size of buffers





Repeaters in Wiring

- Insights
  - Length of optimal segment is a function of technology (not a function of length of wire)
  - Same applies to the buffer sizing
  - Delay scales linearly in length with proper buffering

- Food for thought
  - Is buffering energy-efficient?

$$L_{seg}^{*} = \frac{L}{N} = \sqrt{2\left(\frac{R_{buf} \times \left(C_{self} + C_{load}\right)}{R_u \times C_u}\right)}$$

$$W = \sqrt{\frac{R_0 \times C_{wire}}{R_{wire} \times 2C_0}}$$

Admin

- Felicity review session TBD, See Piazza
- Final Exam Friday in Canvas, 12/17
- 12-2pm in Moore 212