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

Lec 7: February 14, 2024

Layout and Area, MOS Scaling

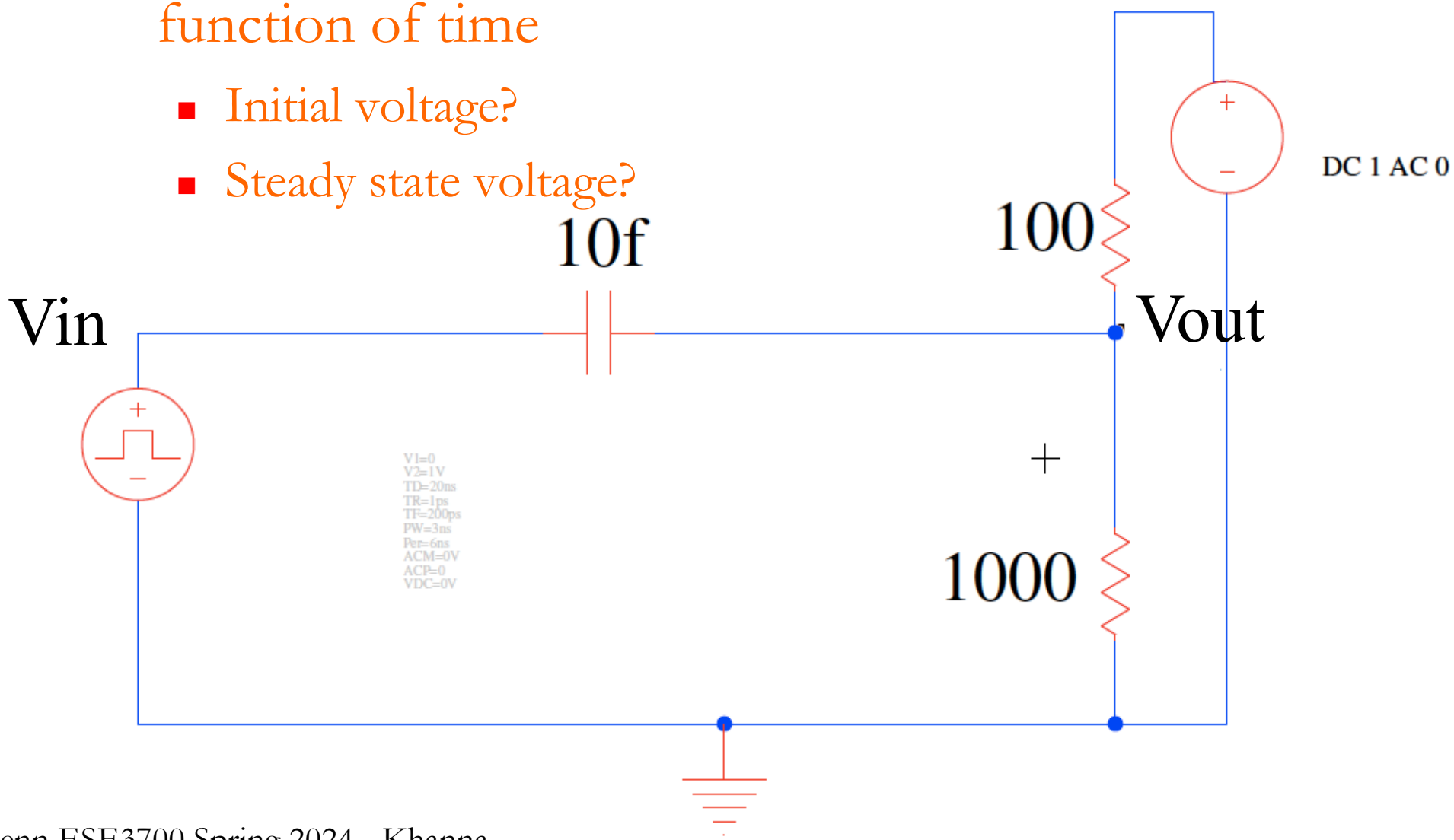
One Parasitic Capacitance Implication

Feedback Capacitance C_{gd}

Penn

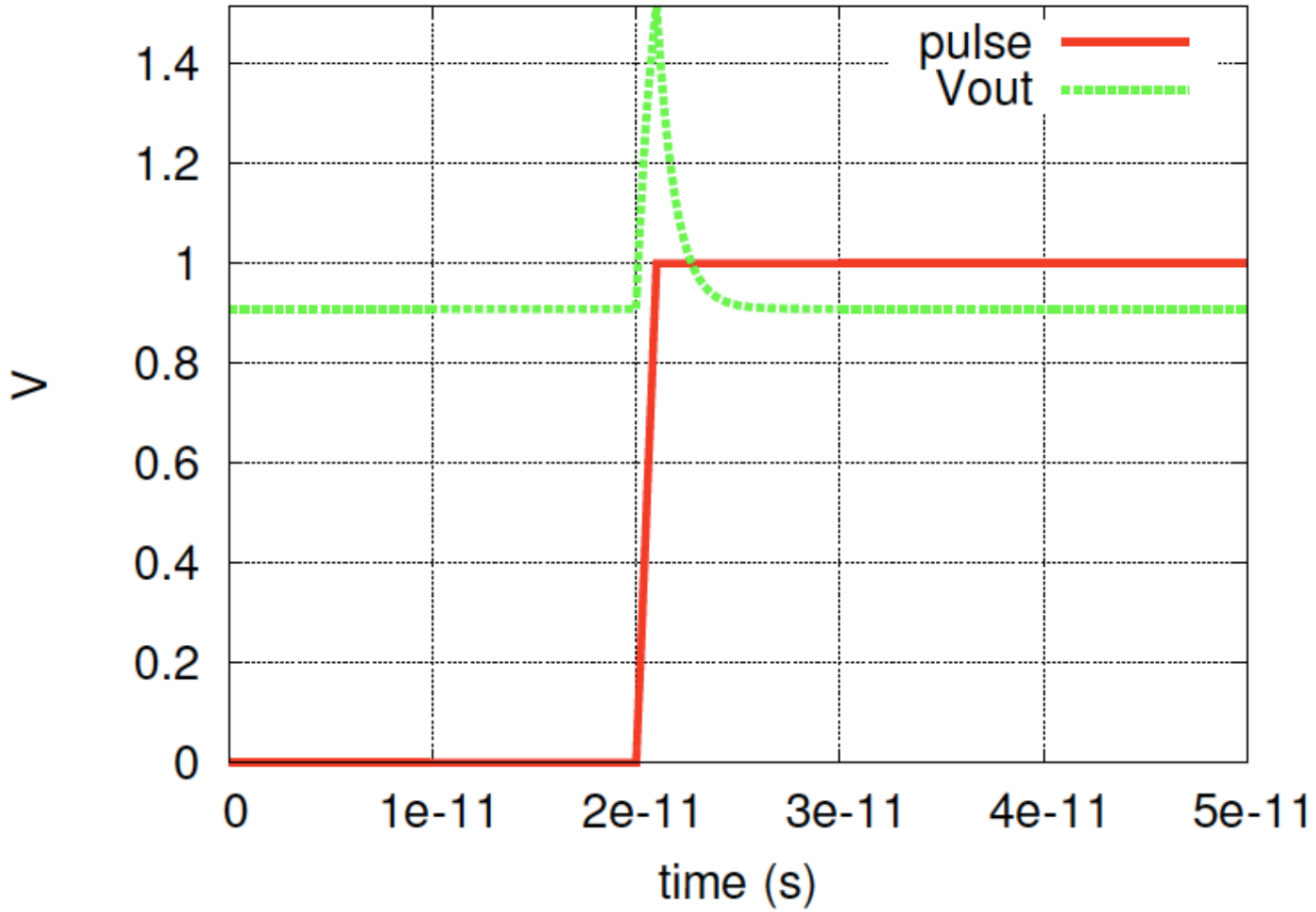
Step Response? (Preclass 1)

- Vin steps from 0 to 1, what does Vout look like as a function of time
  - Initial voltage?
  - Steady state voltage?



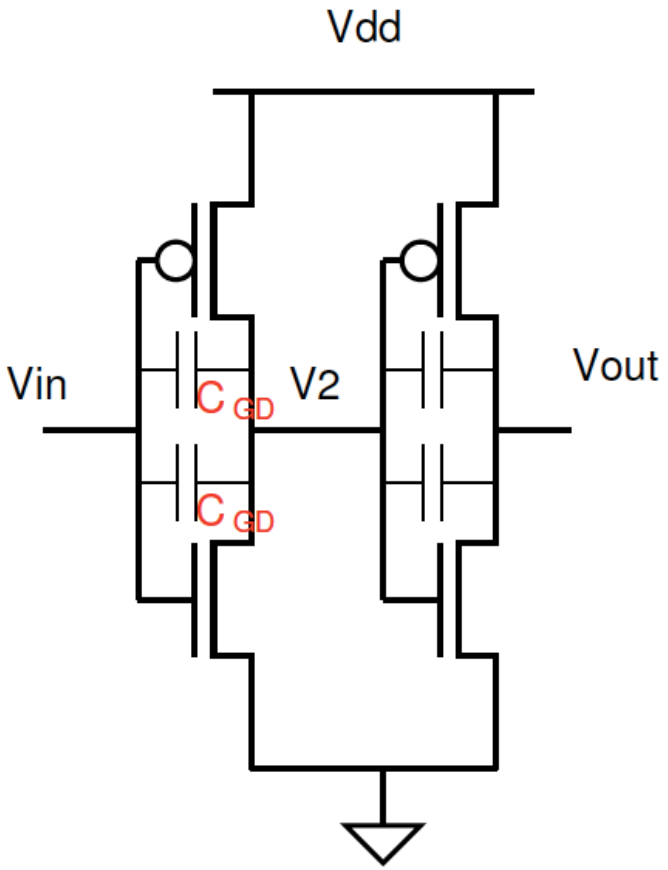
Step Response

Voltage peaking!

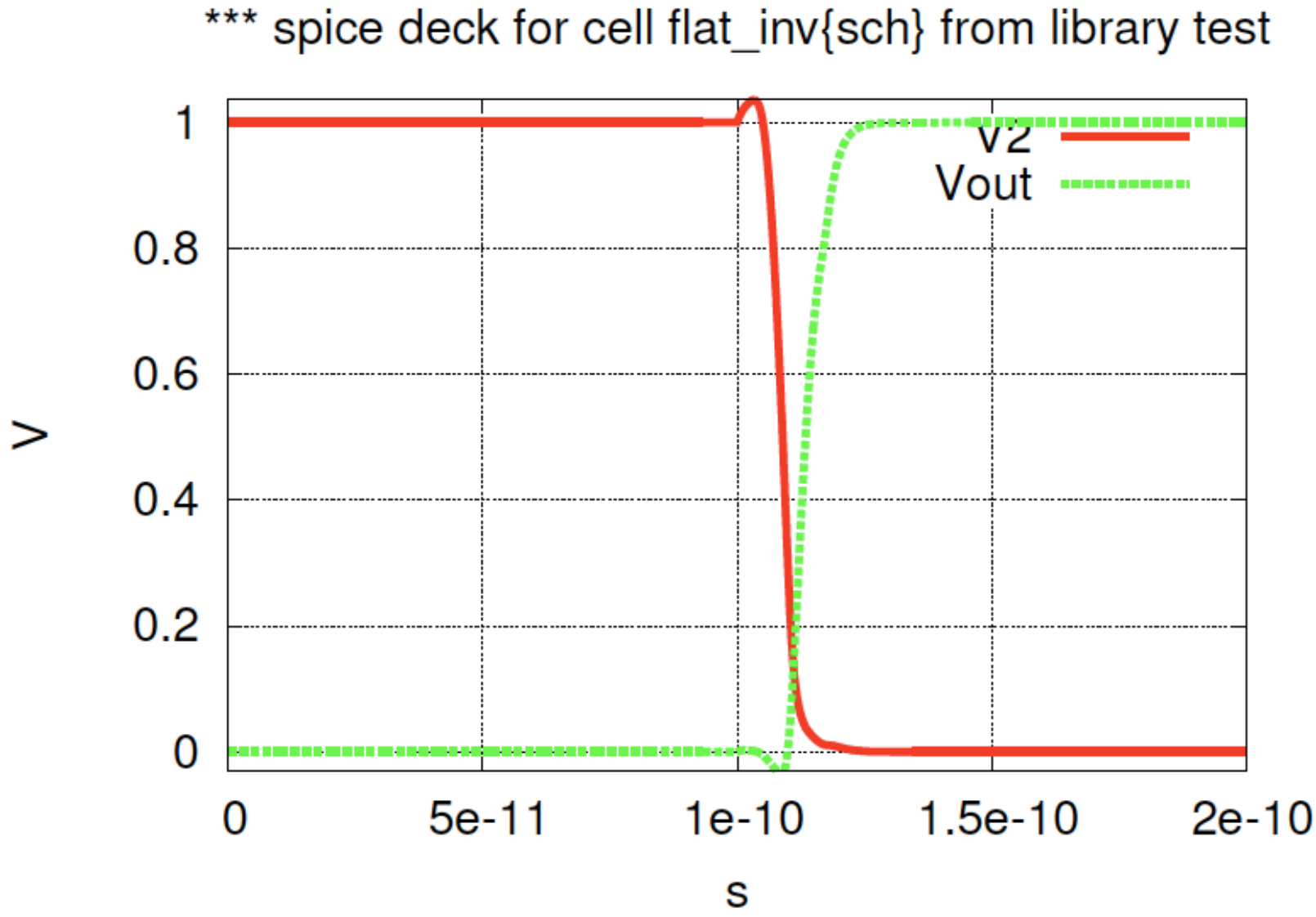


Impact of C_{GD}

- What does C_{GD} do to the switching response here?
  - V_2
  - V_{out}



Impact of C_{GD}

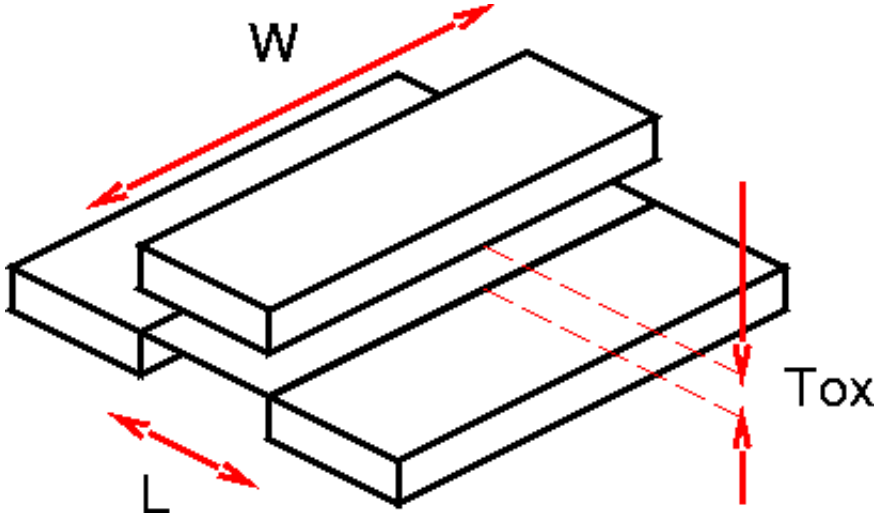


Today

- Layout
  - Transistors
  - Gates
- Design rules
- Standard cells
- VLSI Scaling Trends/Disciplines

Transistor

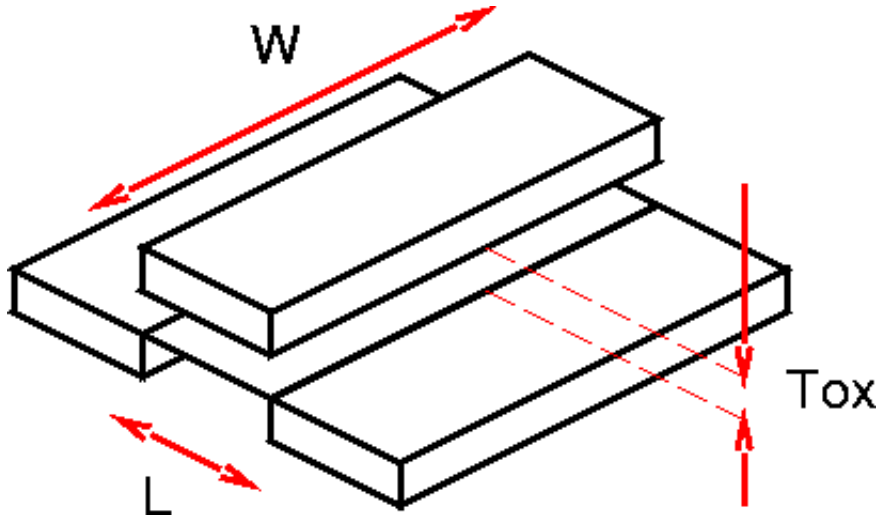
Side view



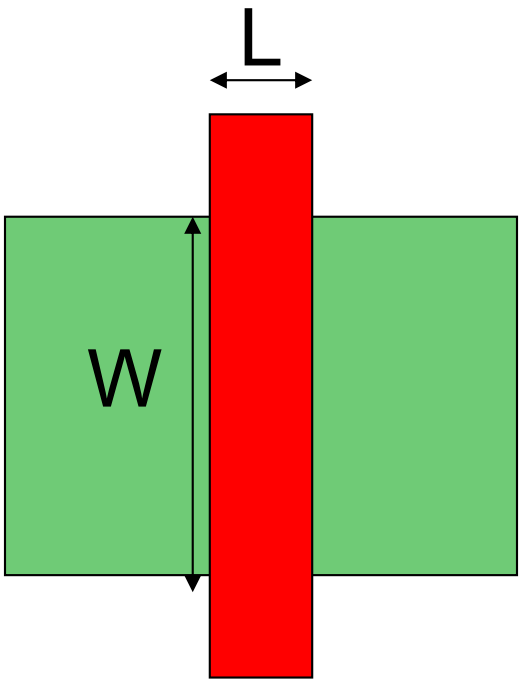
Perspective view

Layout

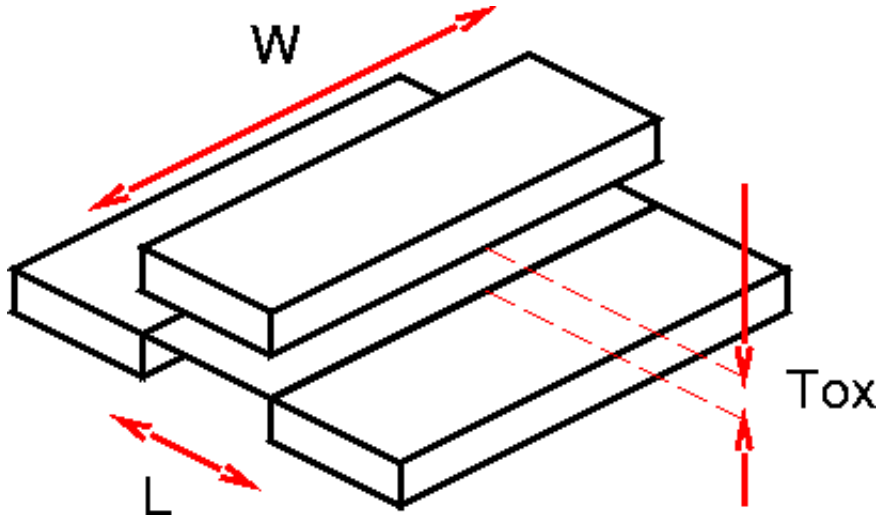
- Sizing & positioning of transistors
- Designer controls W, L
- t_{ox} fixed for process
  - Sometimes thick/thin oxide "flavors"



NMOS Geometry



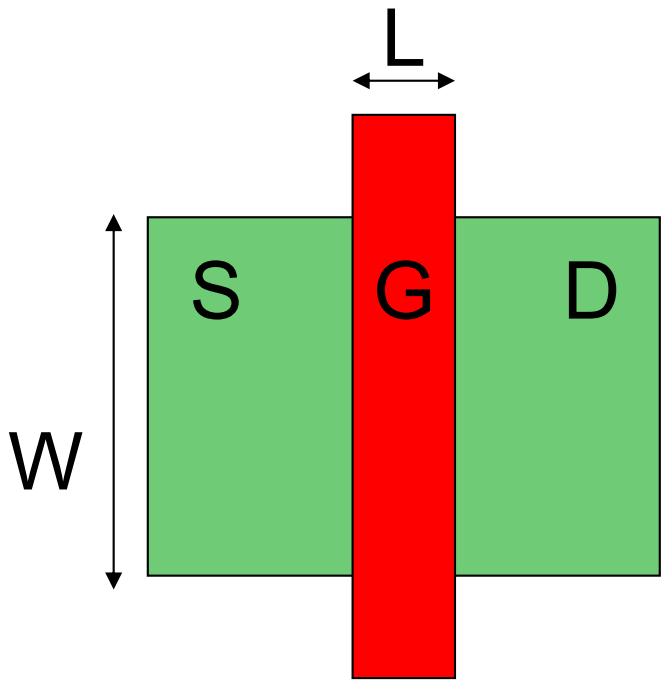
Top view



Perspective view

NMOS Geometry

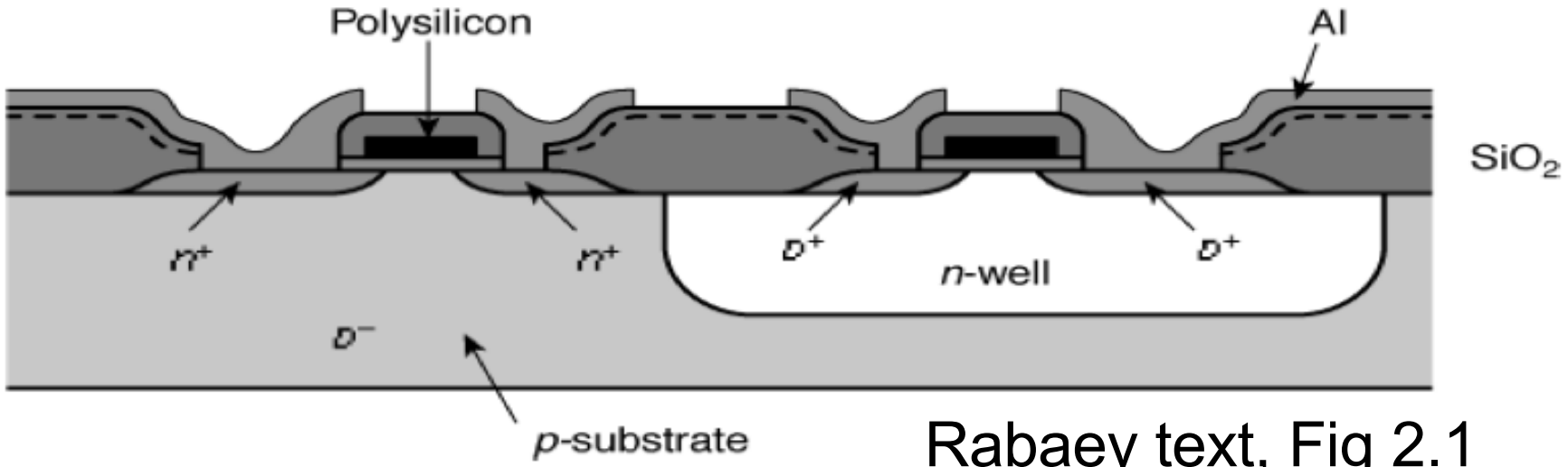
- Color scheme
  - Red: gate (polysilicon material)
  - Green: source and drain areas (n type diffusion)



Top view

NMOS vs PMOS

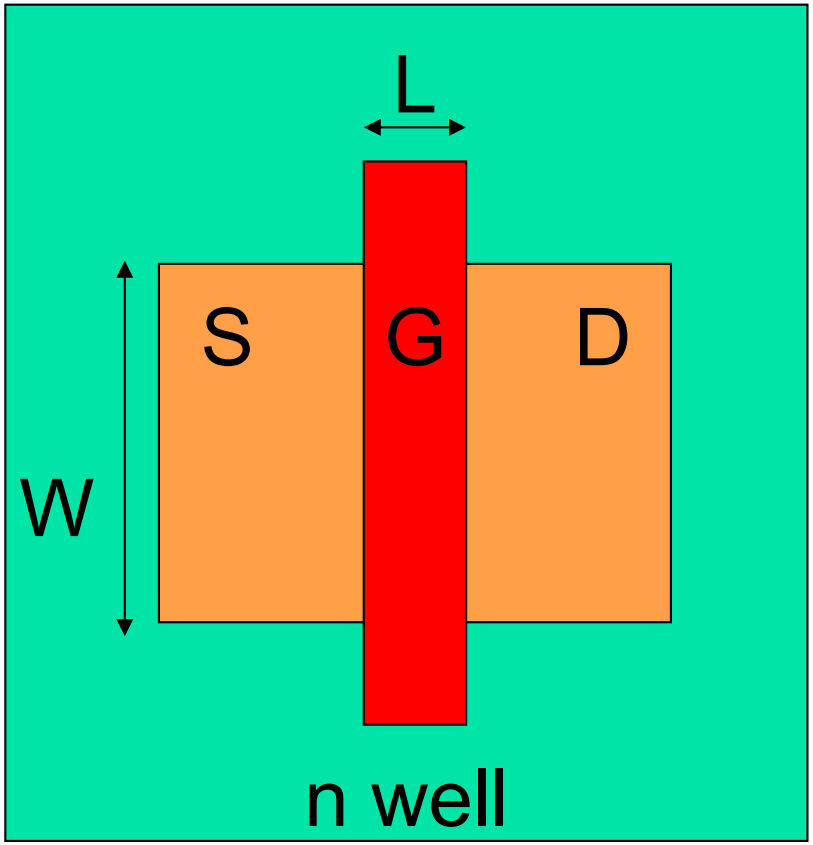
- NMOS built on p substrate
- PMOS built on n substrate
  - Needs an N-well



Rabaey text, Fig 2.1

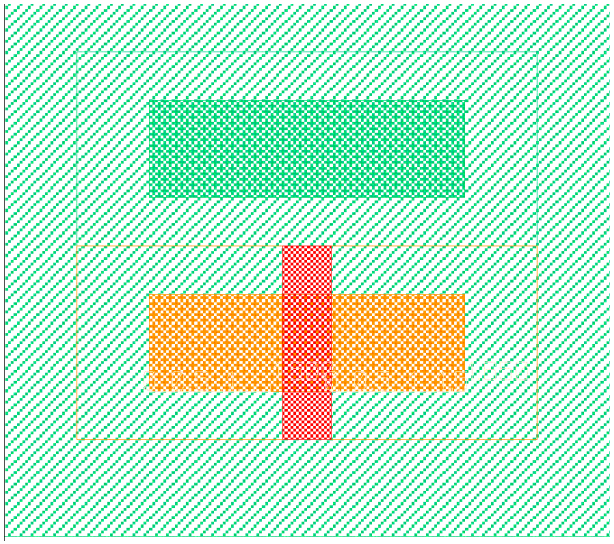
PMOS Geometry

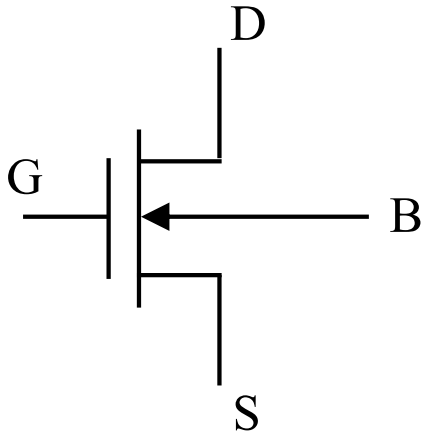
- Color scheme
  - Red: gate
  - Orange: source and drain areas (p type)
  - Green: n well
- NMOS built on p wafer
  - Must add n well material to build PMOS

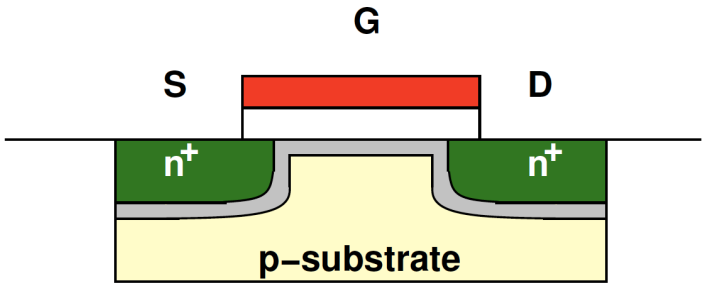


Body Contact

- "Fourth terminal"
- Needed to set voltage around device
  - PMOS: $V_b = V_{dd}$
  - NMOS: V_b = GND
- At right: PMOS (orange) with bulk contact (dark green)

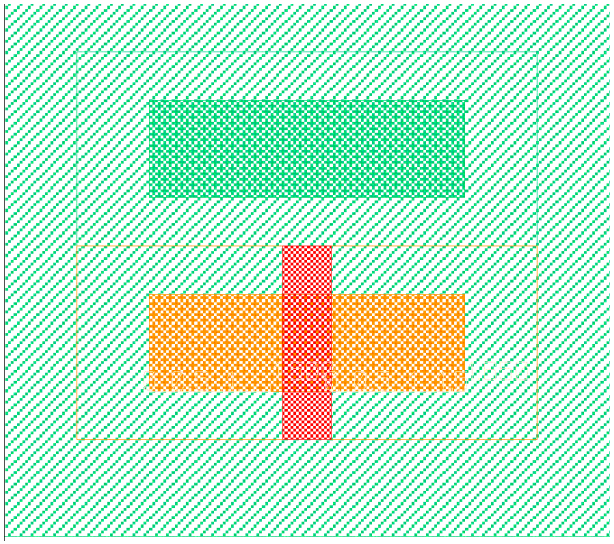


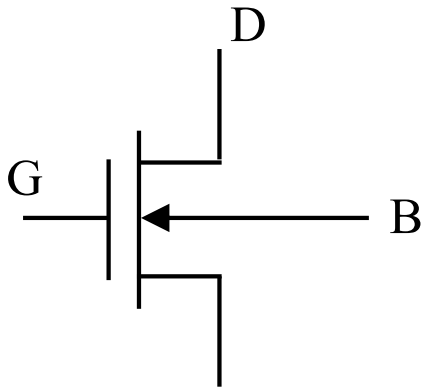


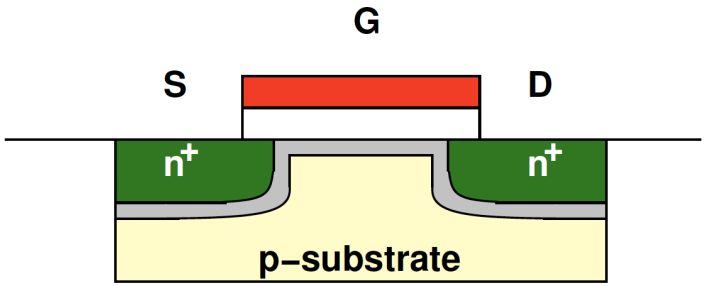


Body Contact

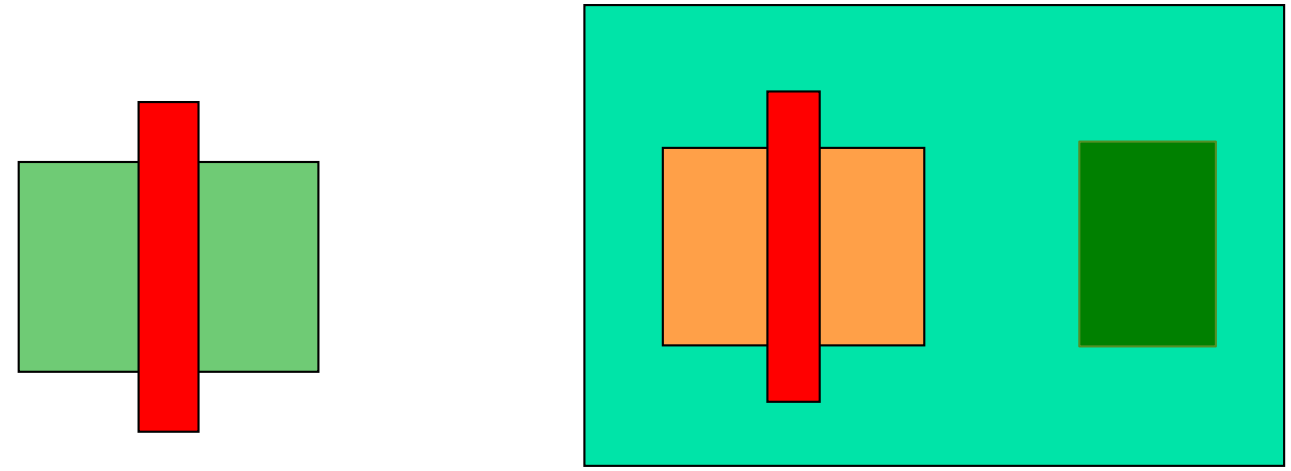
- Needed to set voltage around device
  - PMOS: $V_b = V_{dd}$
  - NMOS: $V_b = GND$
- What happens if NMOS body contact is V_{dd}?
  - Polarity of field wrong
  - Increase Vth (need higher voltage to invert the channel)

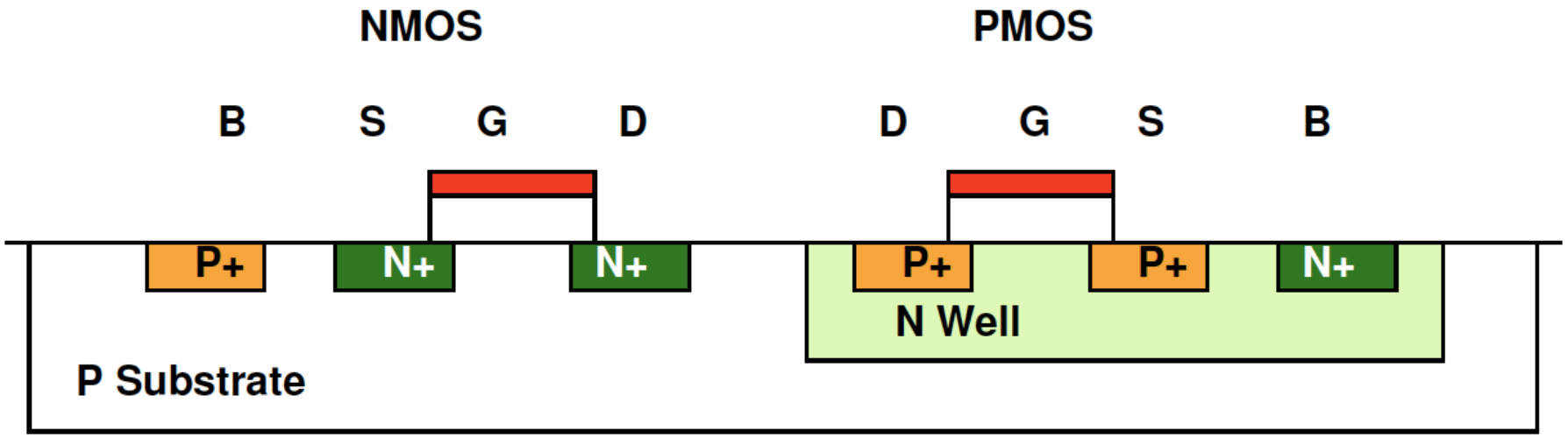






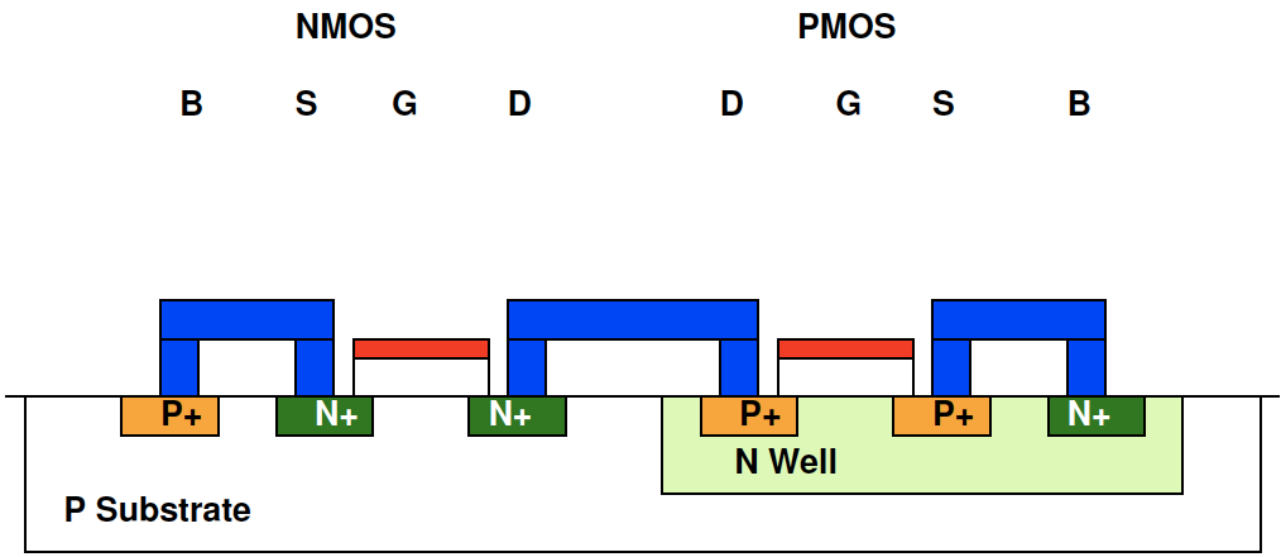
Transistor Geometry





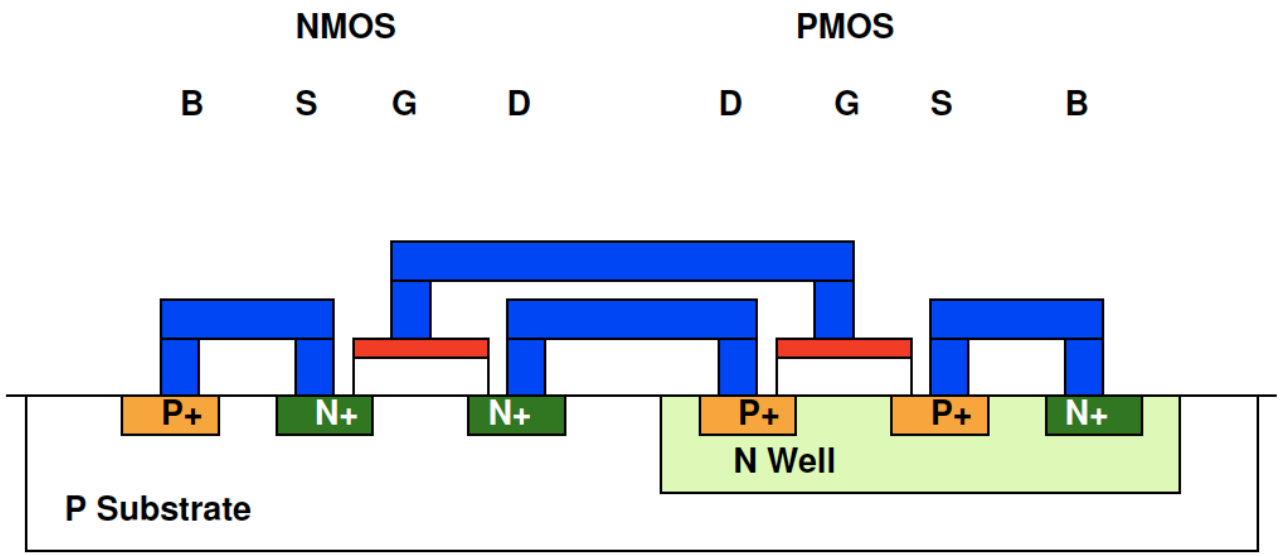
Interconnect

- Connect transistors
  - Different layers of metal
    - "Contact" - metal to transistor
    - "Via" - metal to metal

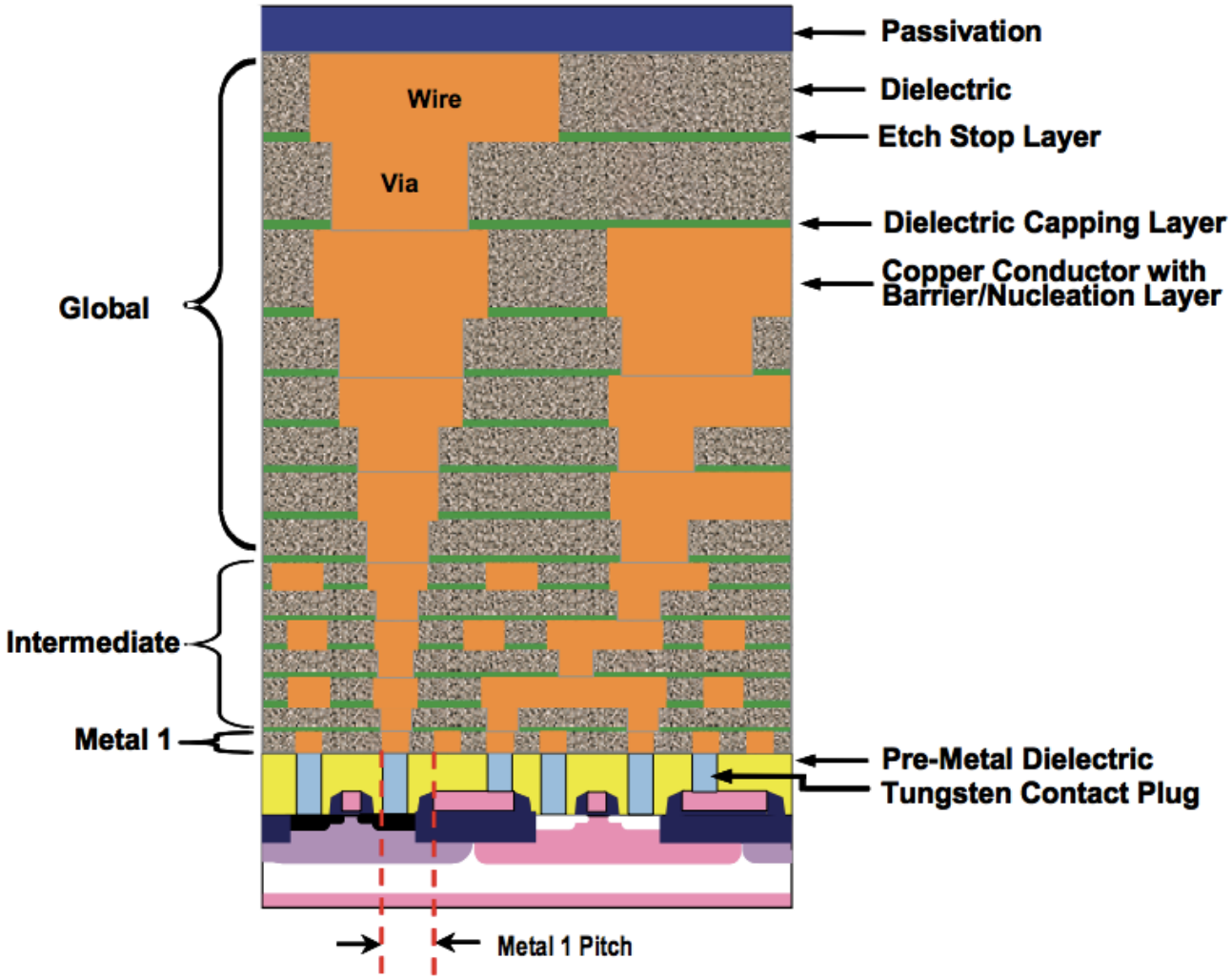


Interconnect

- Connect transistors
  - Different layers of metal
    - "Contact" - metal to transistor
    - "Via" - metal to metal



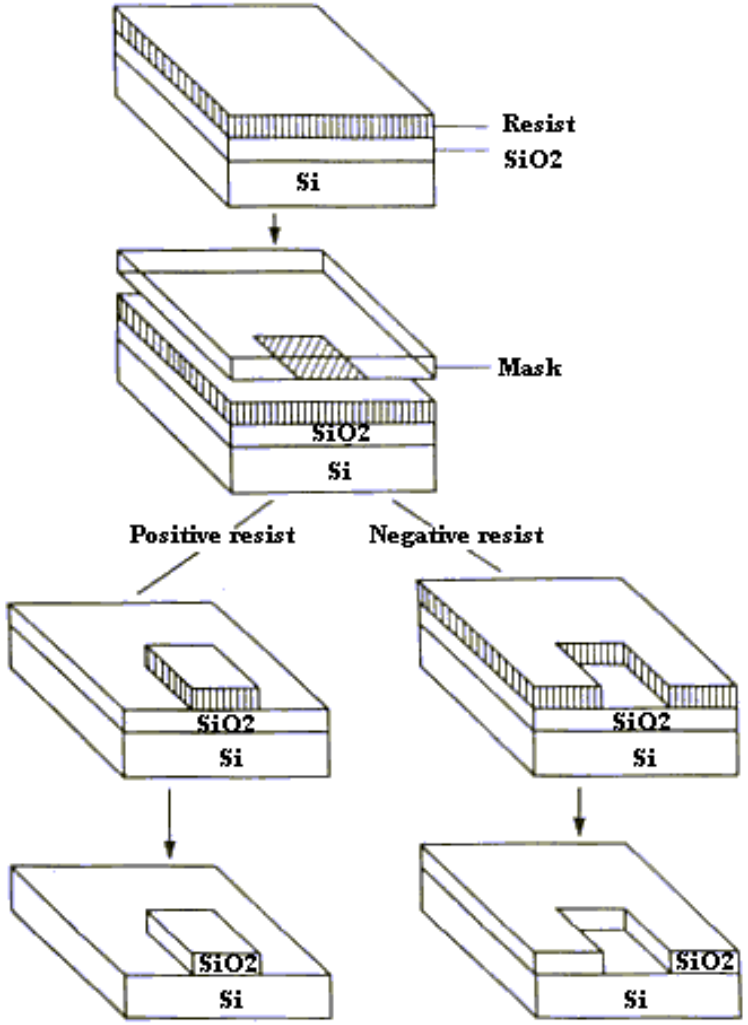
Interconnect Cross Section



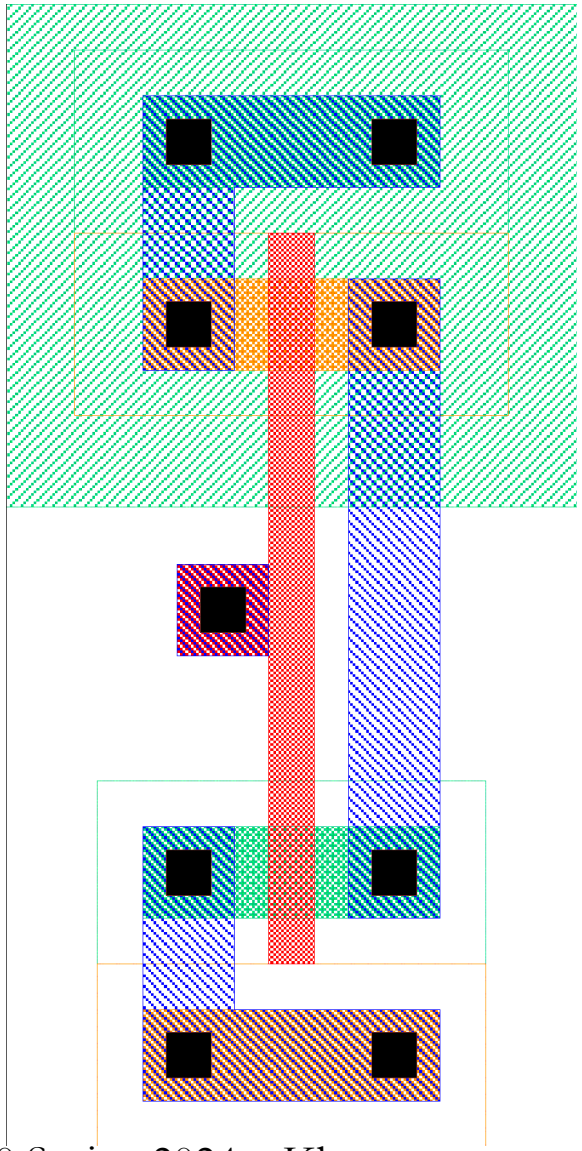
ITRS 2007

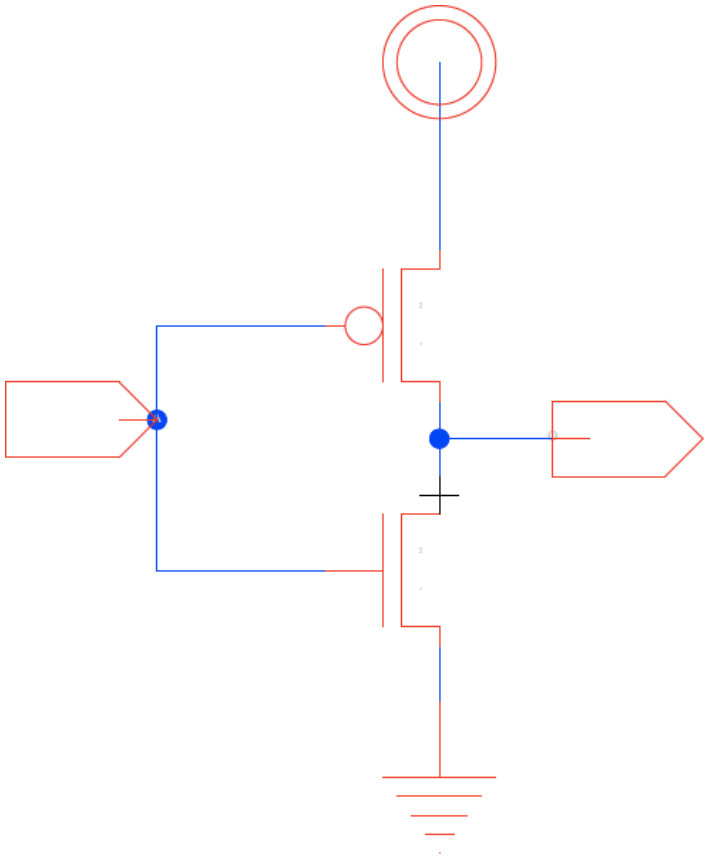
Masks

- Define areas want to see in layer
  - Think of "stencil" for material deposition
- Use photoresist (PR) to form the "stencil"
  - Grow PR over entire wafer
  - Expose PR through mask
  - PR dissolves in exposed areas
  - Material is deposited/etched
    - Only "sticks" in area w/ dissolved PR

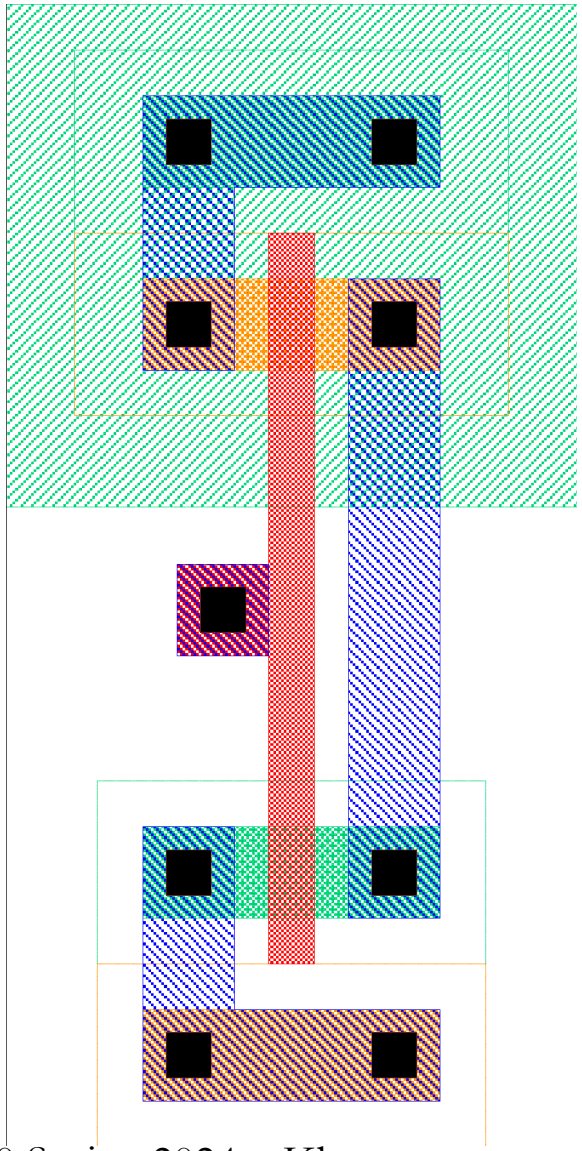


Reverse Engineer Inverter Layout (Preclass 2)



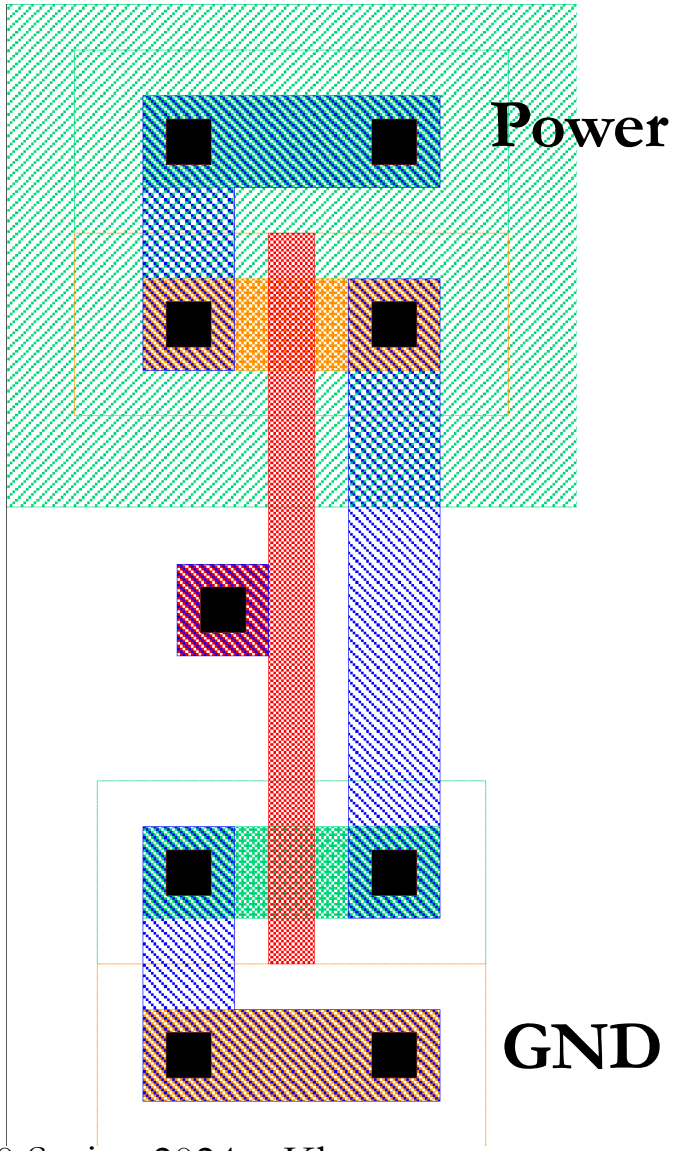


Layout Revisited

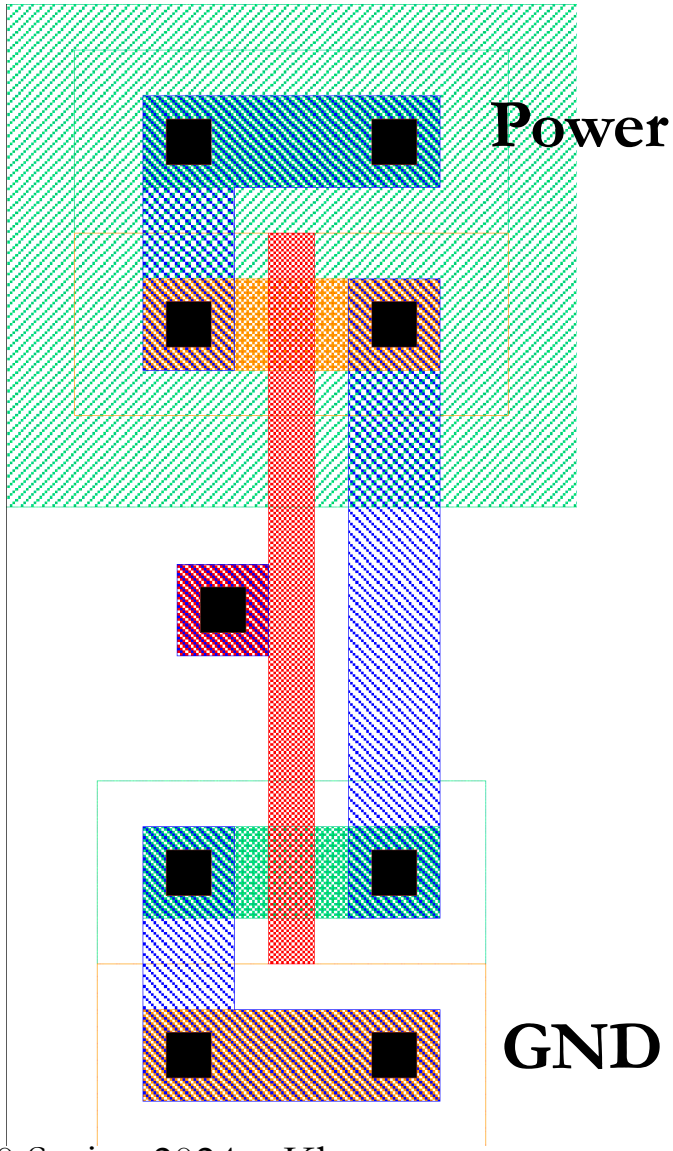


- How to "decode" circuit from layout?

Reverse Engineer Inverter Layout

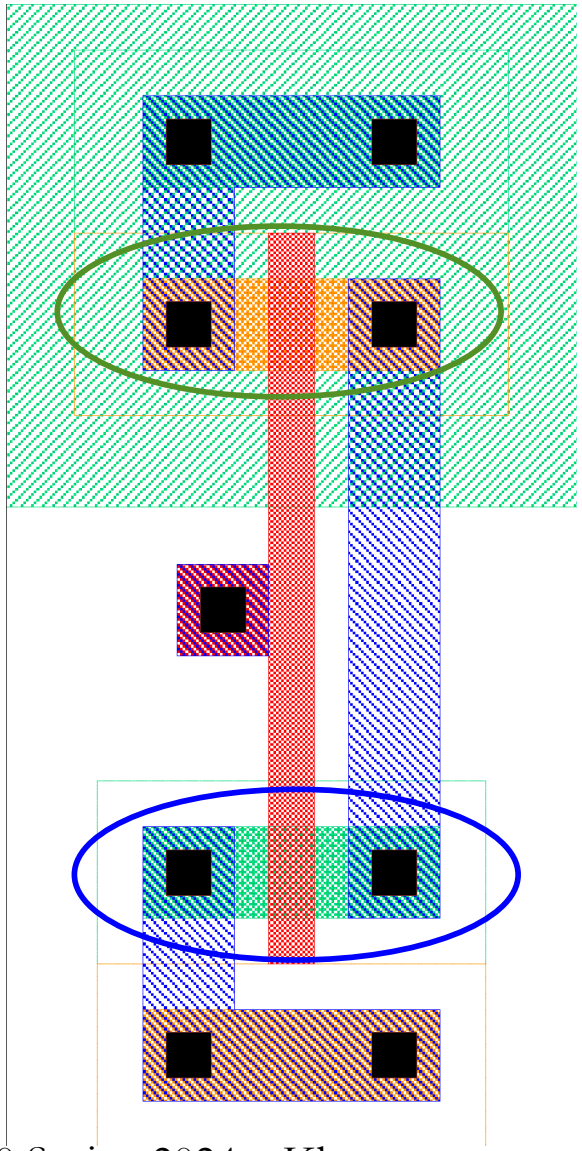


Reverse Engineer Inverter Layout

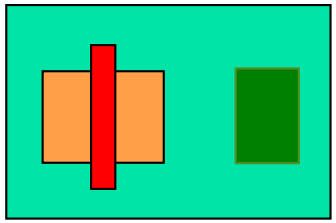


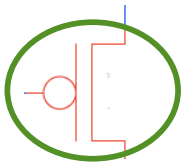
- Where is PMOS transistor?
- NMOS?

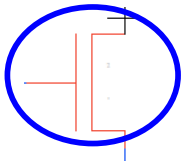
Layout to Circuit

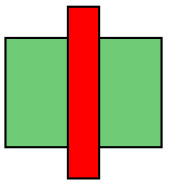


- 1. Identify transistors

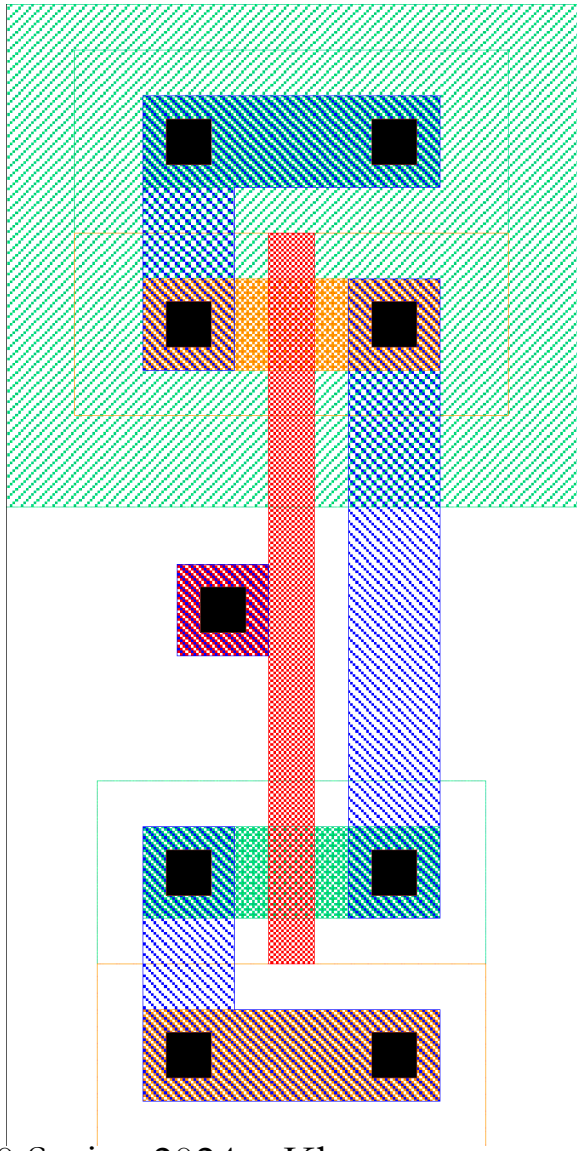






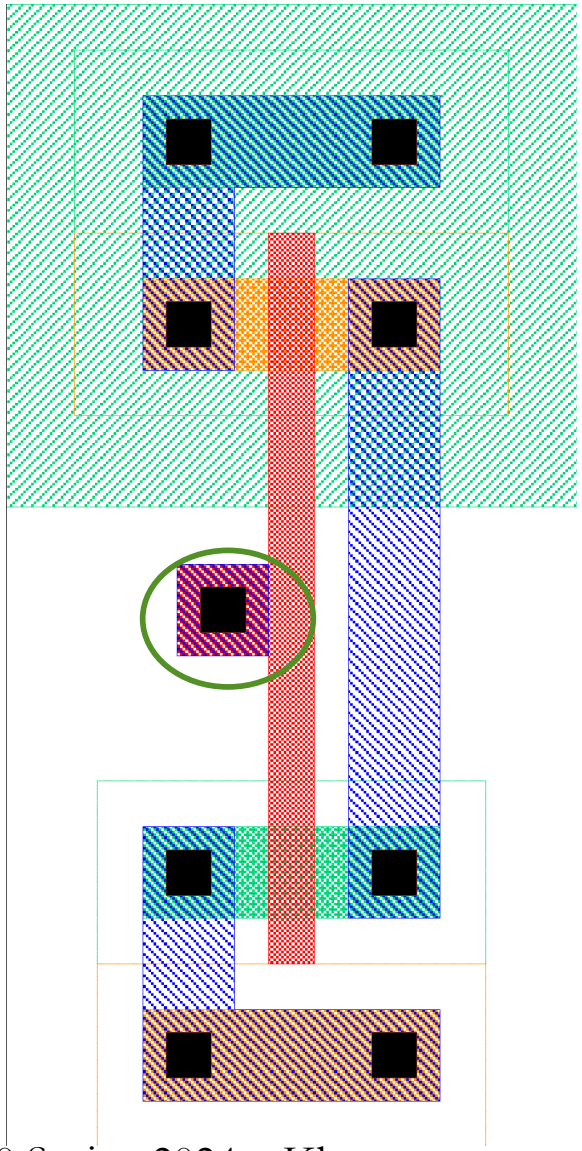


Inverter Layout



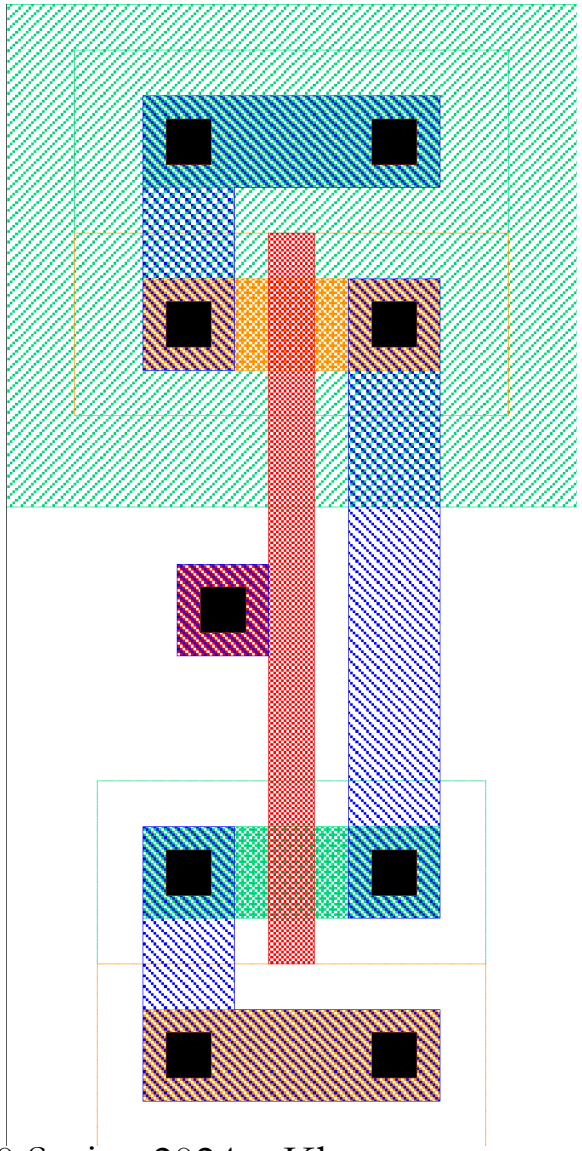
- Where is Input?

Inverter Layout



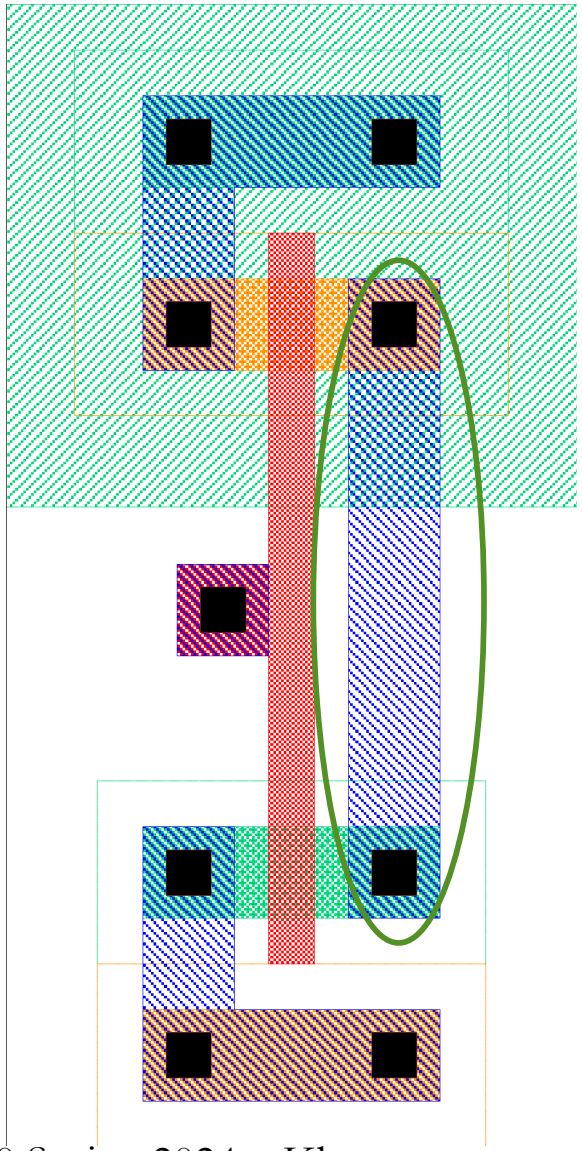
- Where is Input?

Inverter Layout



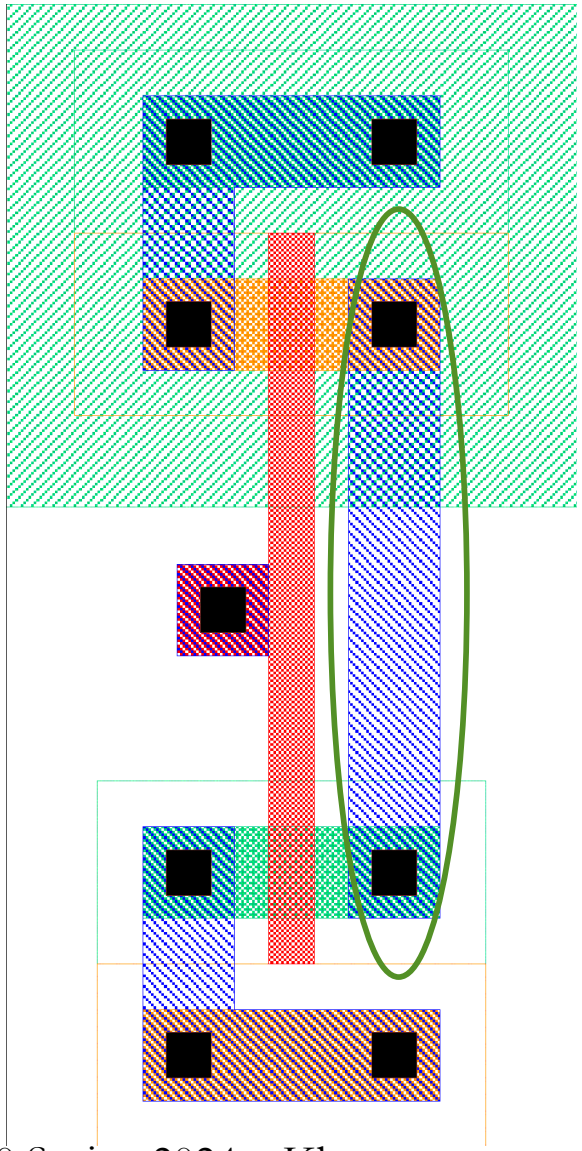
- Where is Output?

Inverter Layout

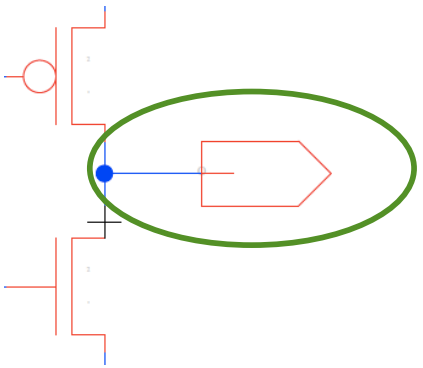


- Where is Output?

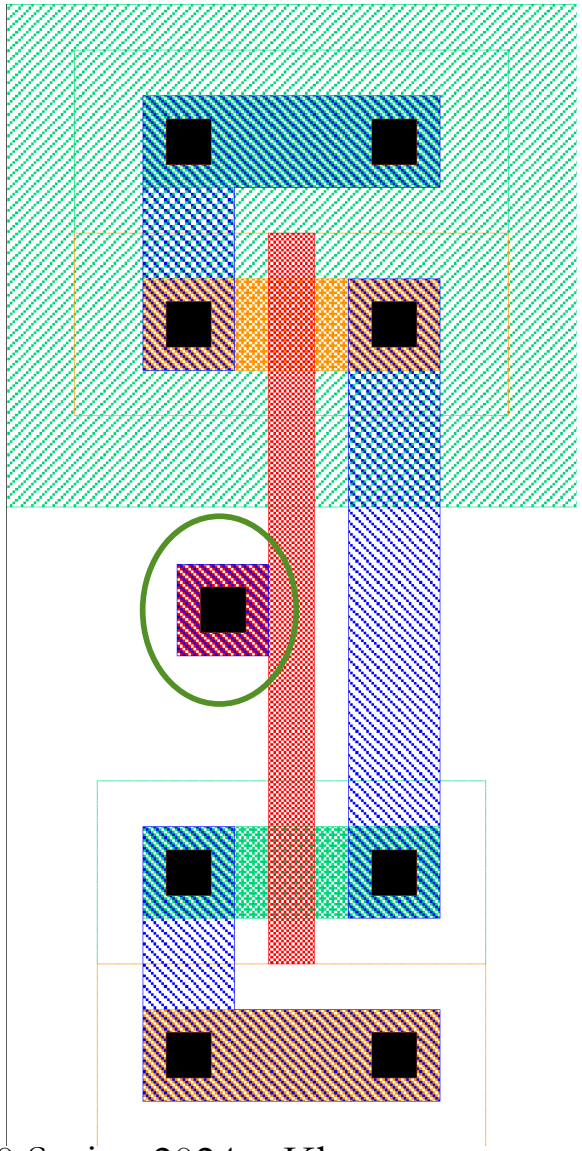
Layout to Circuit



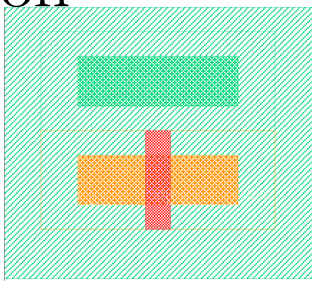
- 2. Add connections
  - Drain connection

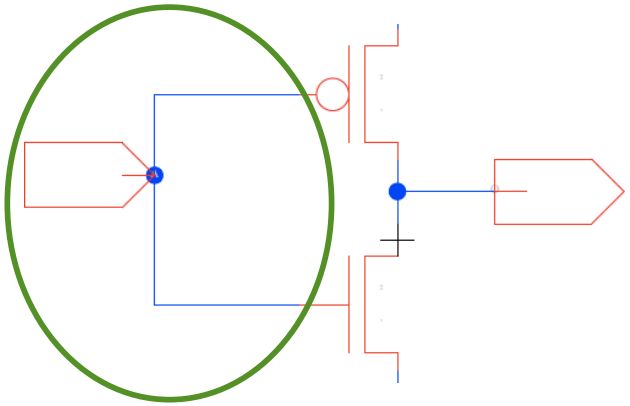


Layout to Circuit

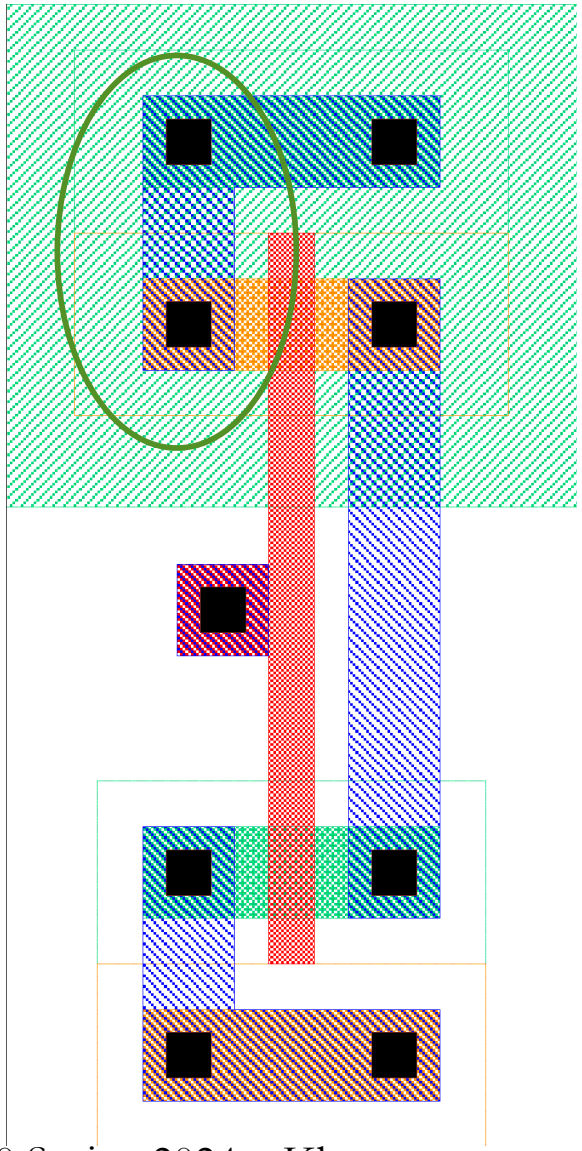


- 2. Add connections
  - Gate connection

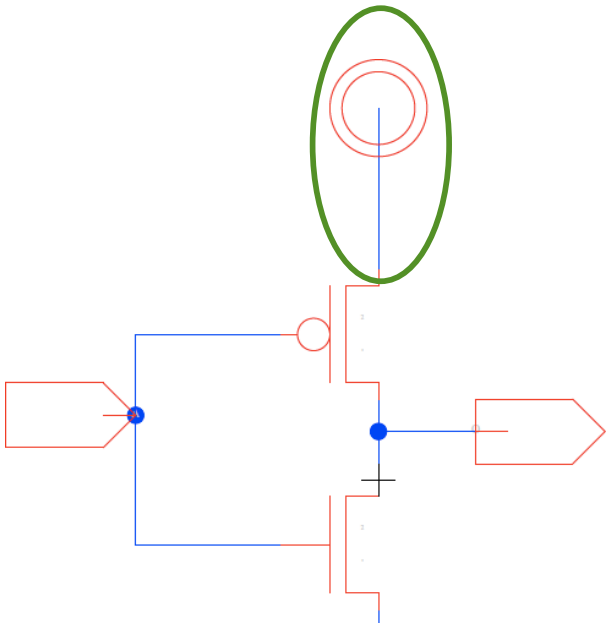




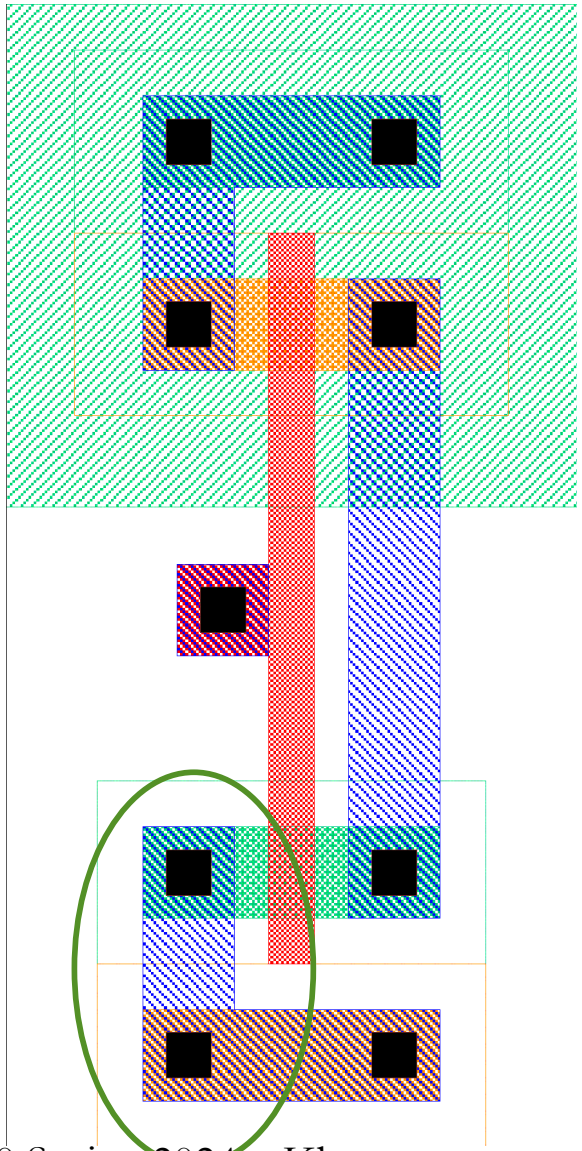
Layout to Circuit



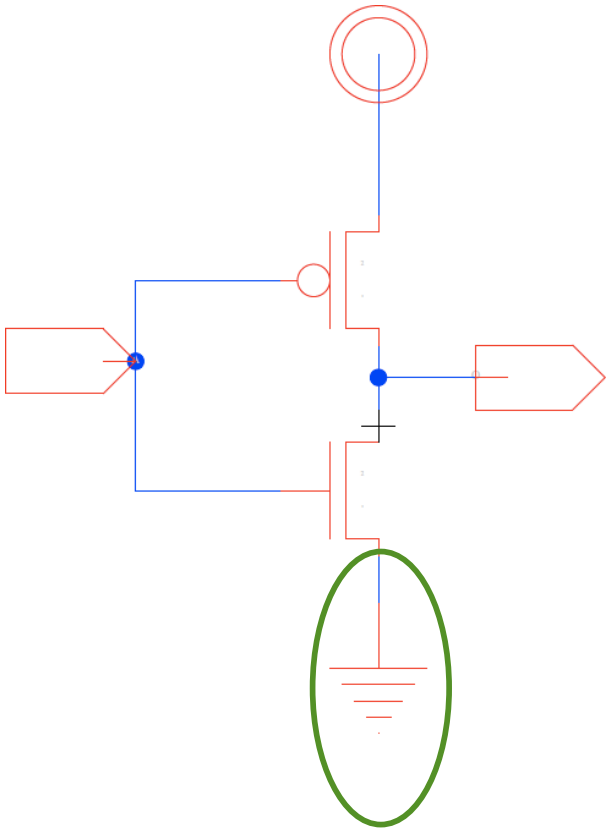
- 2. Add connections
  - pMOS-source to VDD



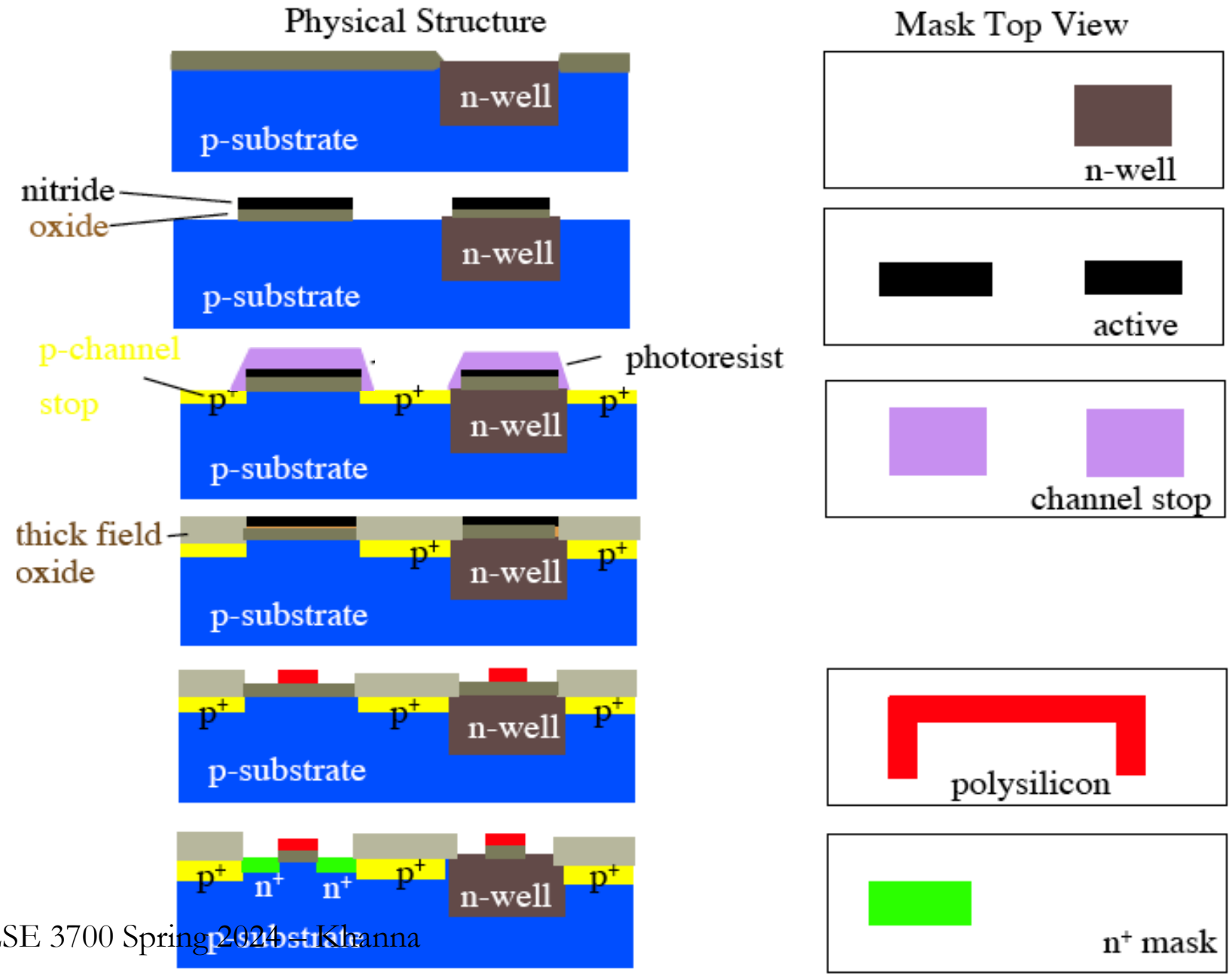
Layout to Circuit



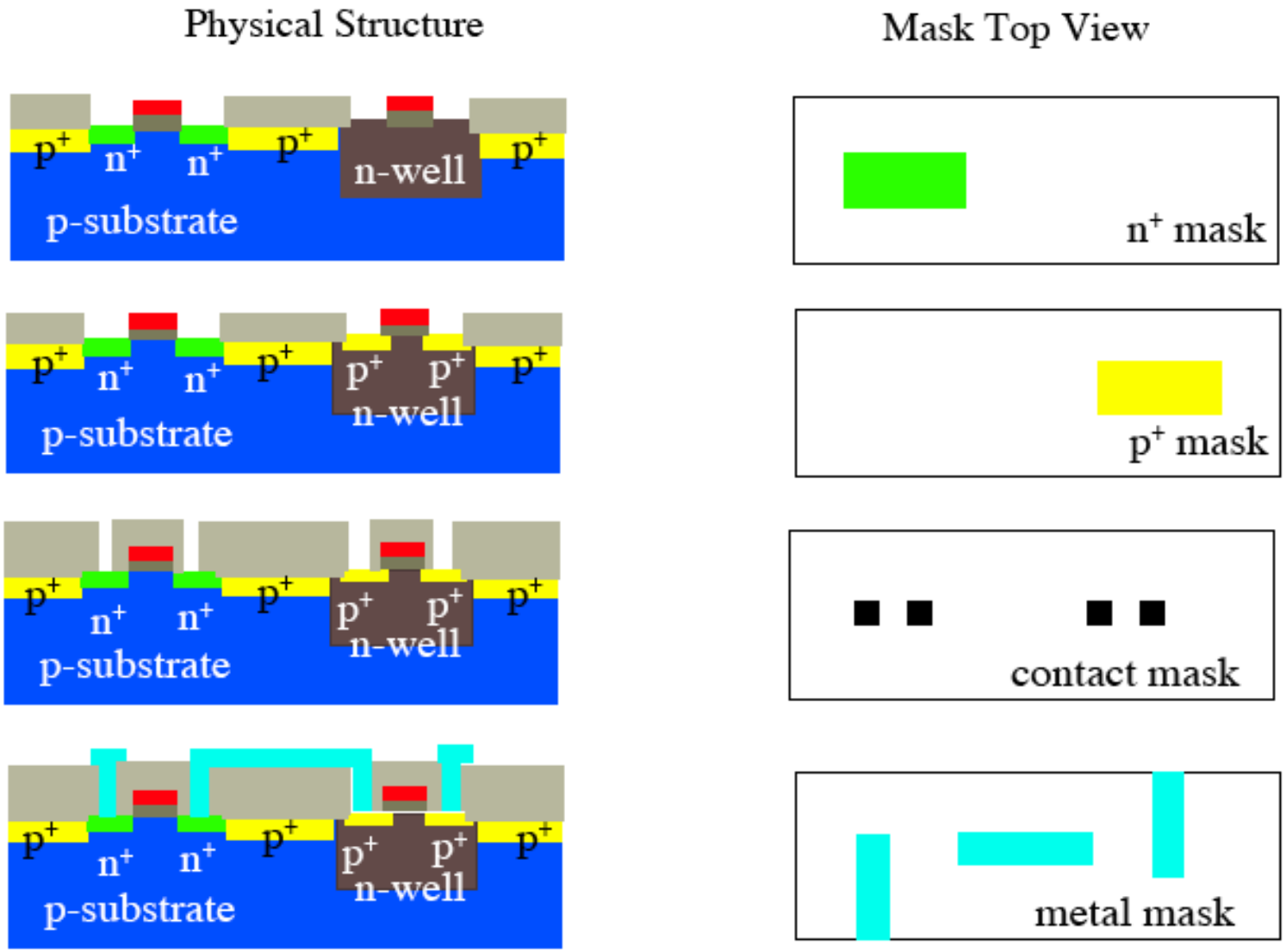
- 2. Add connections
  - nMOS source to GND



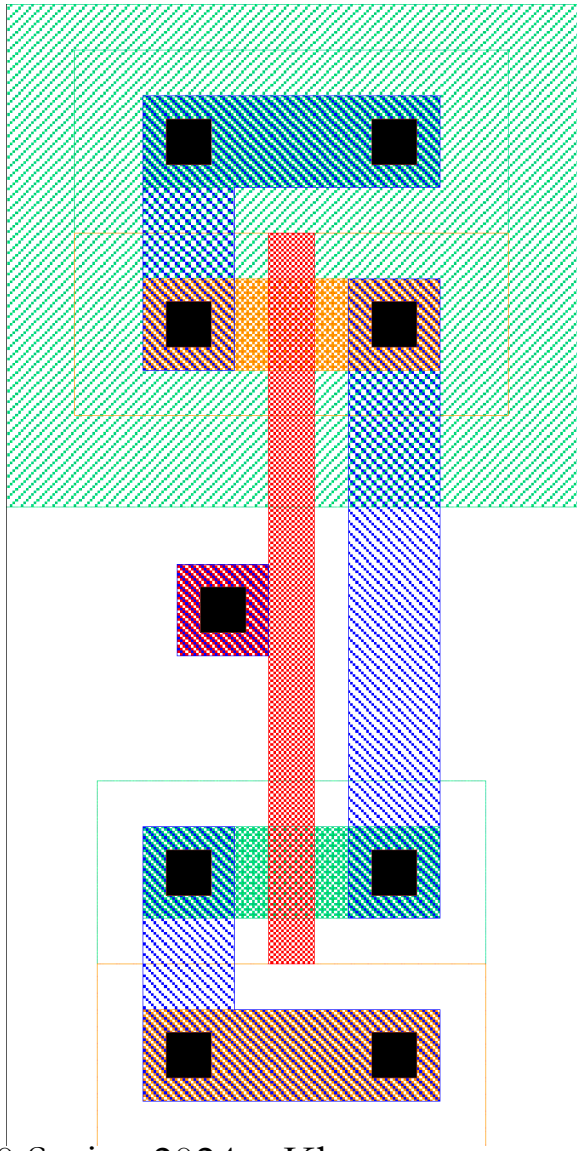
Typical N-Well CMOS Process



Typical N-Well CMOS Process



Design Rules



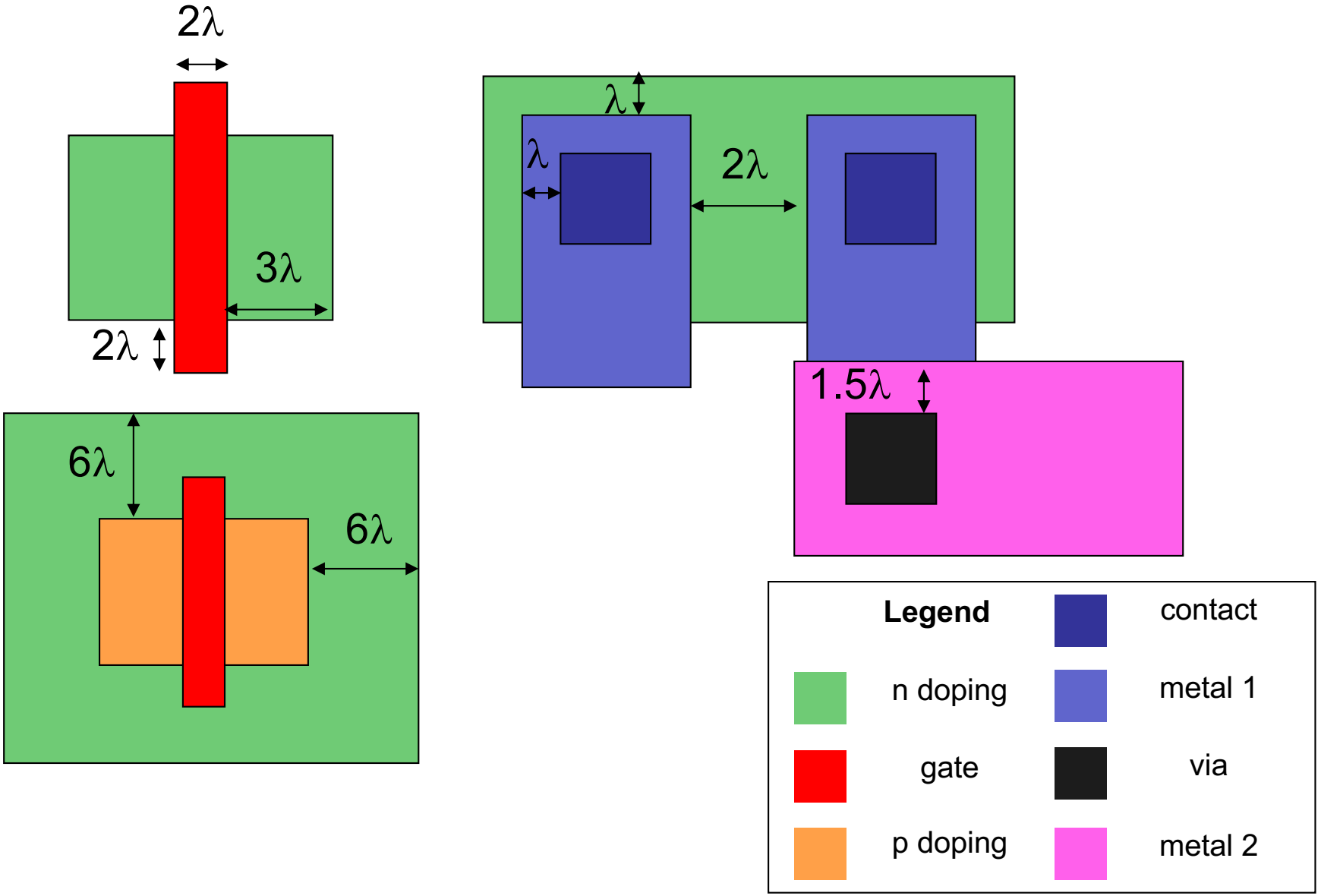
- Why not adjacent transistors?
  - Plenty of empty space
  - If area is money, pack in as much as possible
    - Shortens connections
- Recall: processing is imprecise
  - Margin of error for process variation

Design Rules

- Contract between process engineer & designer
  - Minimum width/spacing
  - Can be (often are) process specific

- Lambda rules: scalable design rules
  - In terms of $\lambda = 0.5\ L_{min}$ (L_{drawn})
  - Can migrate designs from similar process with lambda factor

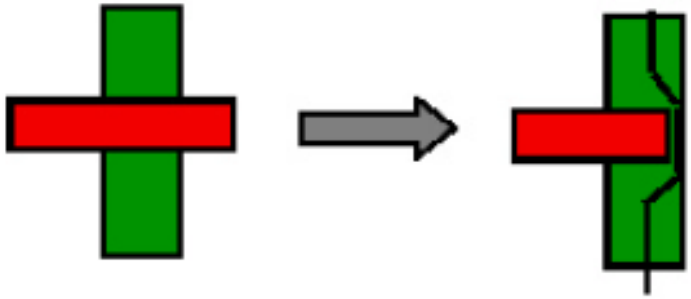
Design Rules: Some Examples



Potential Consequences of Design Rule Violations

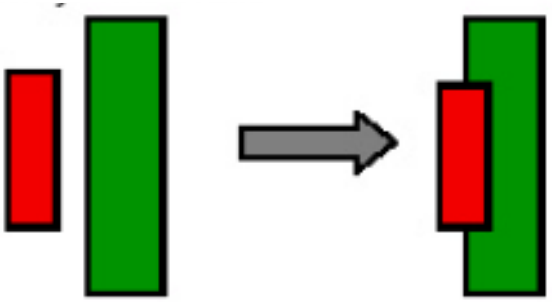
- Inter-Layer Design Rule Origins

Intended Transistor



Catastrophic Error – Unintended misalignment cause Source-Drain short circuit

Intended Unrelated Poly & Diffusion



Catastrophic Error – Unintended overlap cause fabrication of a parasitic Transistor

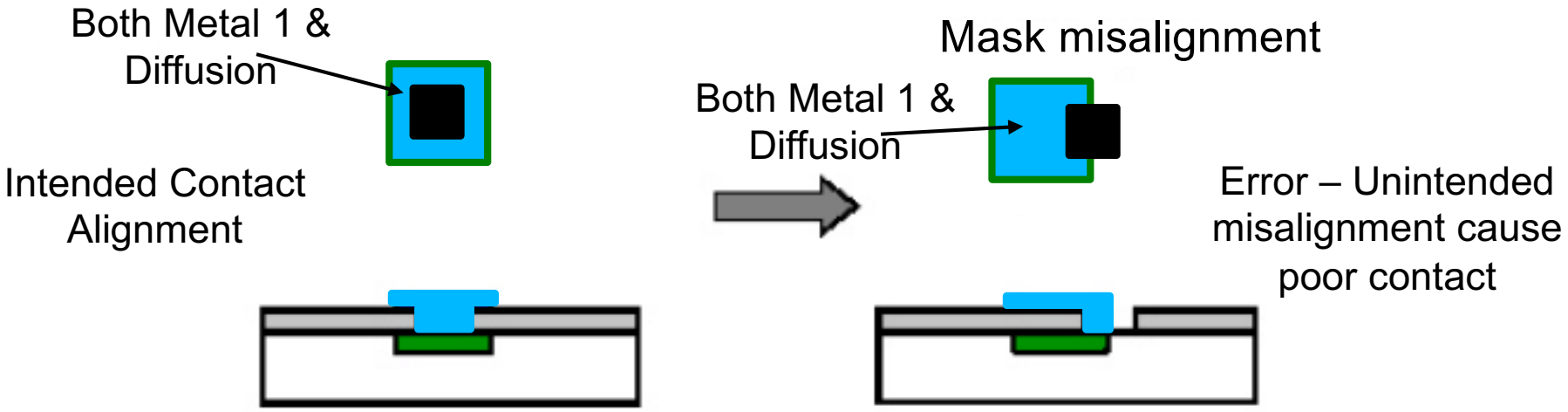
Potential Consequences of Design Rule Violations

- Inter-Layer Design Rule Origins

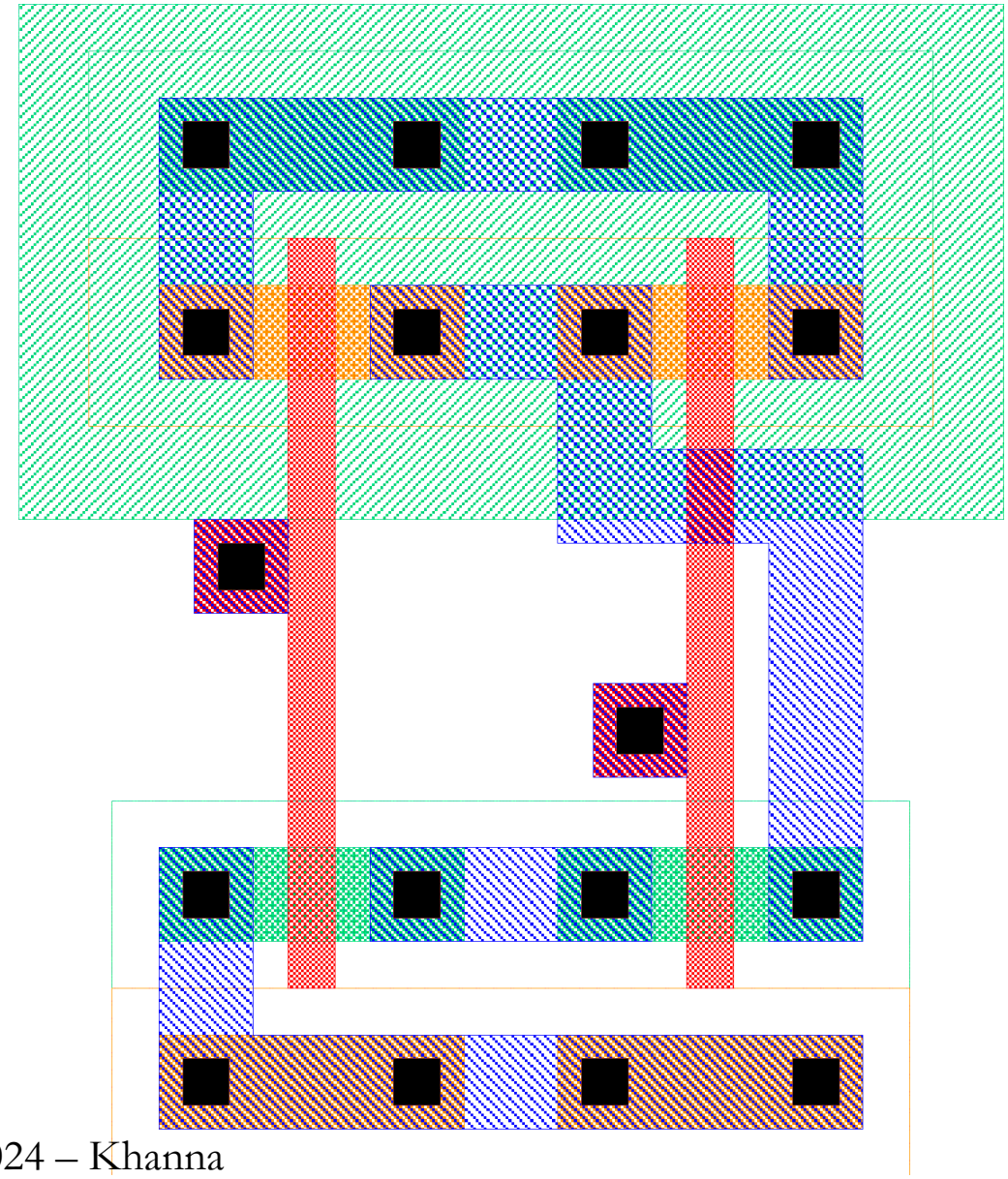
Contact and Via Masks

M1 contact to n-diffusion
M1 contact to p-diffusion
M1 contact to poly
-> Contact Mask

Mn contact to Mn-1 for n = 2, 3,..
-> Via Mask

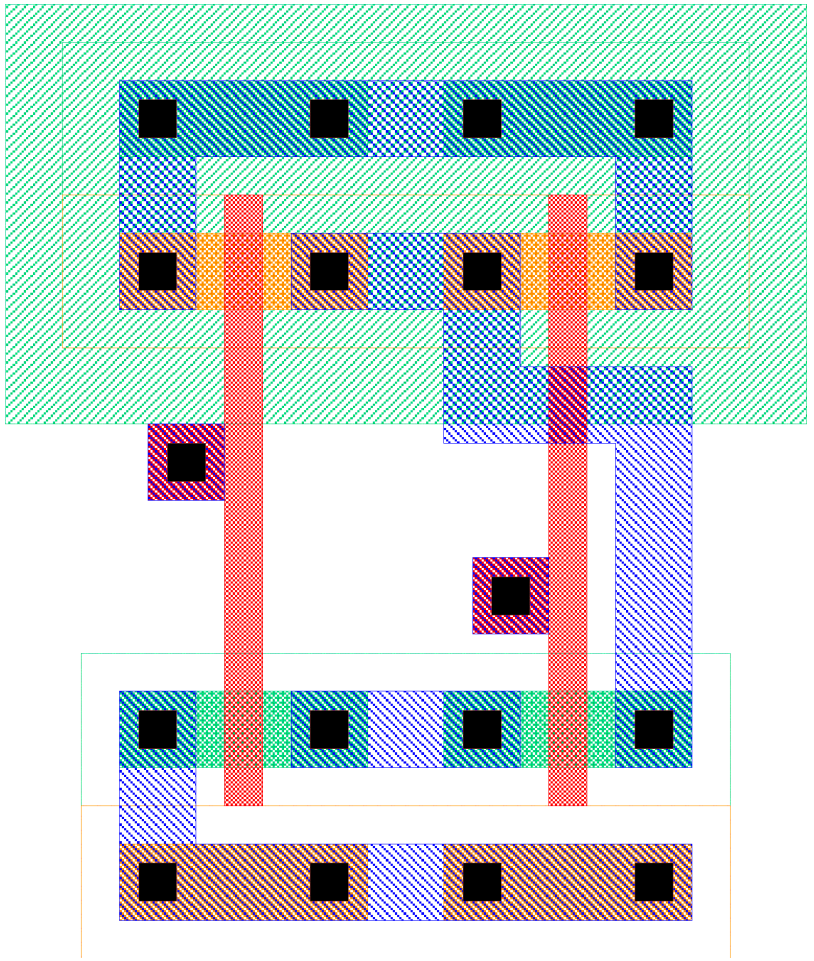


Layout #2 (preclass 3)



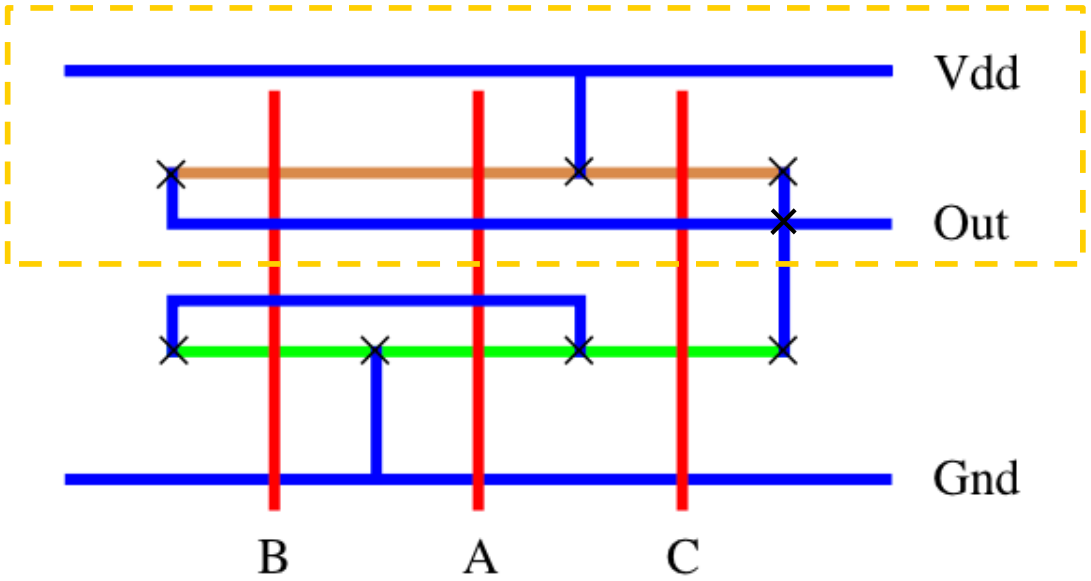
Layout #2 (preclass 3)

- How many transistors?
  - PMOS?
  - NMOS?
- How connected?
  - PMOS, NMOS?
- Inputs connected how?
- Outputs?
- What is it?



Symbolic Layout (Preclass 4)

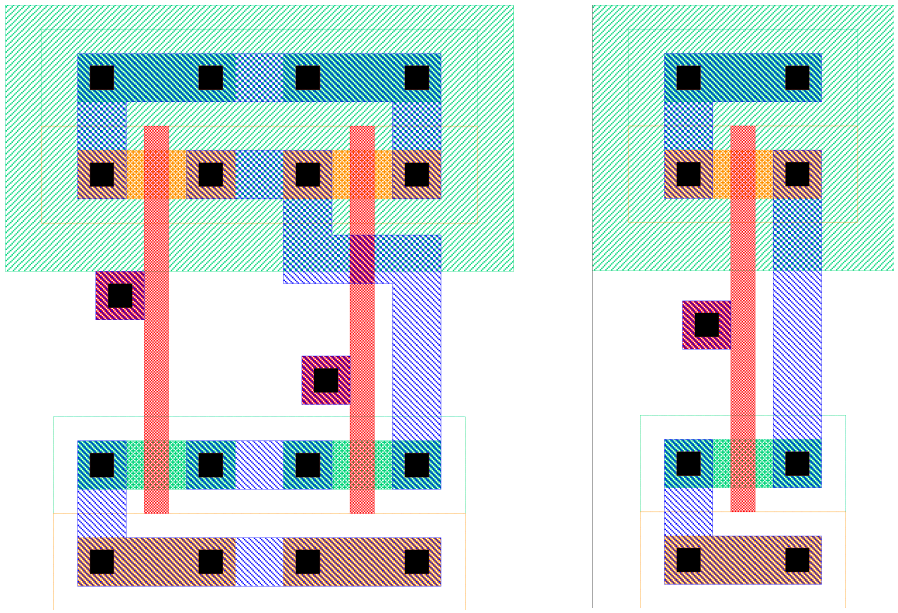
- Stick diagrams capture spatial relationships, but abstract away design rules



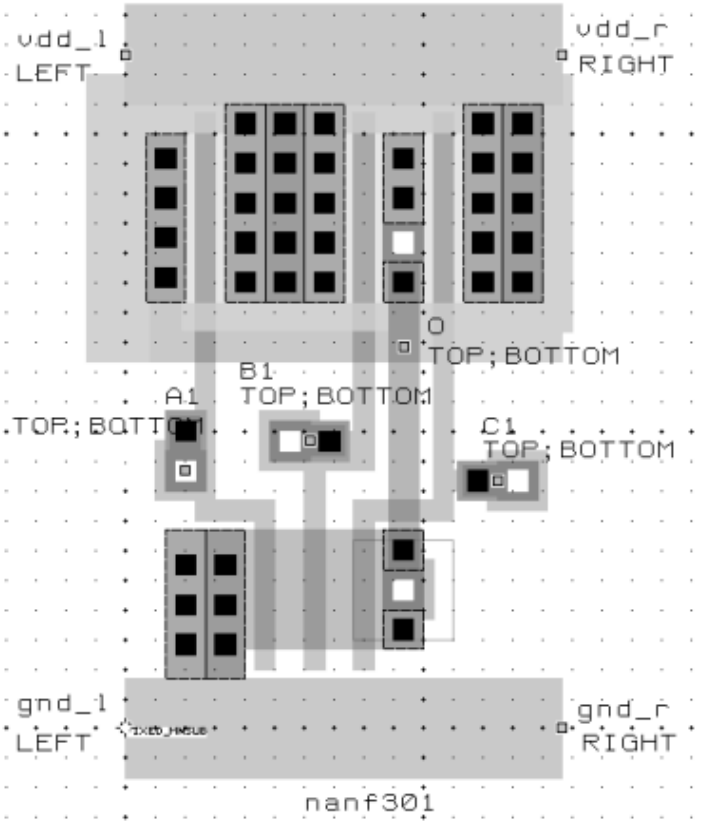
- What is the gate function?
  - How many NMOS? PMOS? D/S connections?
  - Draw schematic

Standard Cells

- Lay out gates so that heights match
  - Rows of adjacent cells
  - Standardized sizing of gate heights
- Motivation: automated place and route
  - EDA tools convert HDL to layout



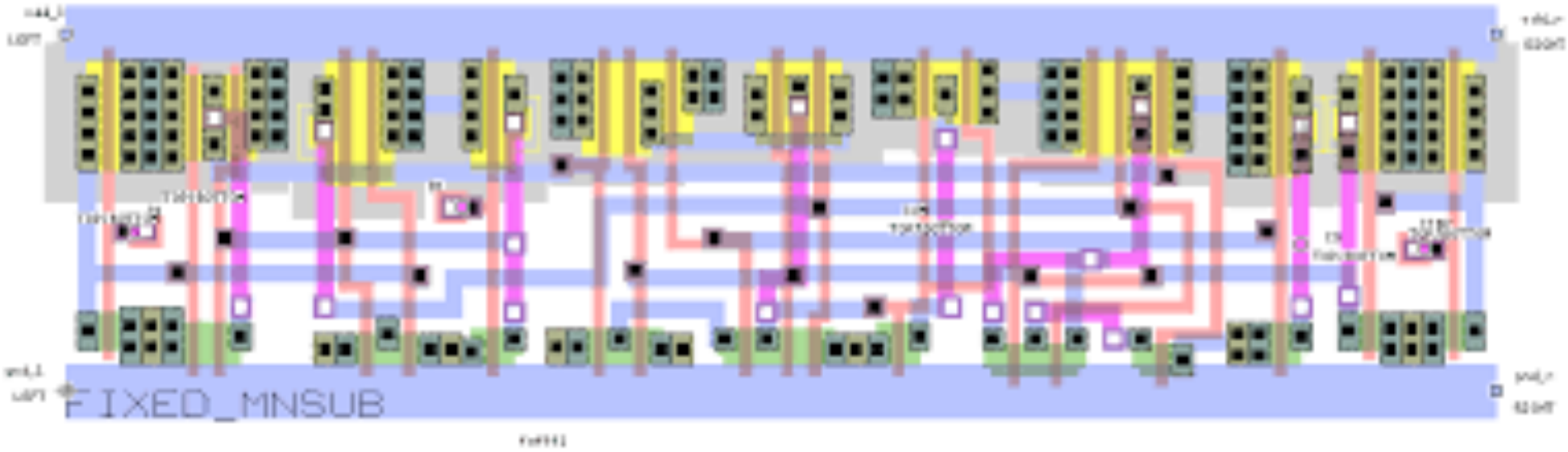
Standard Cells



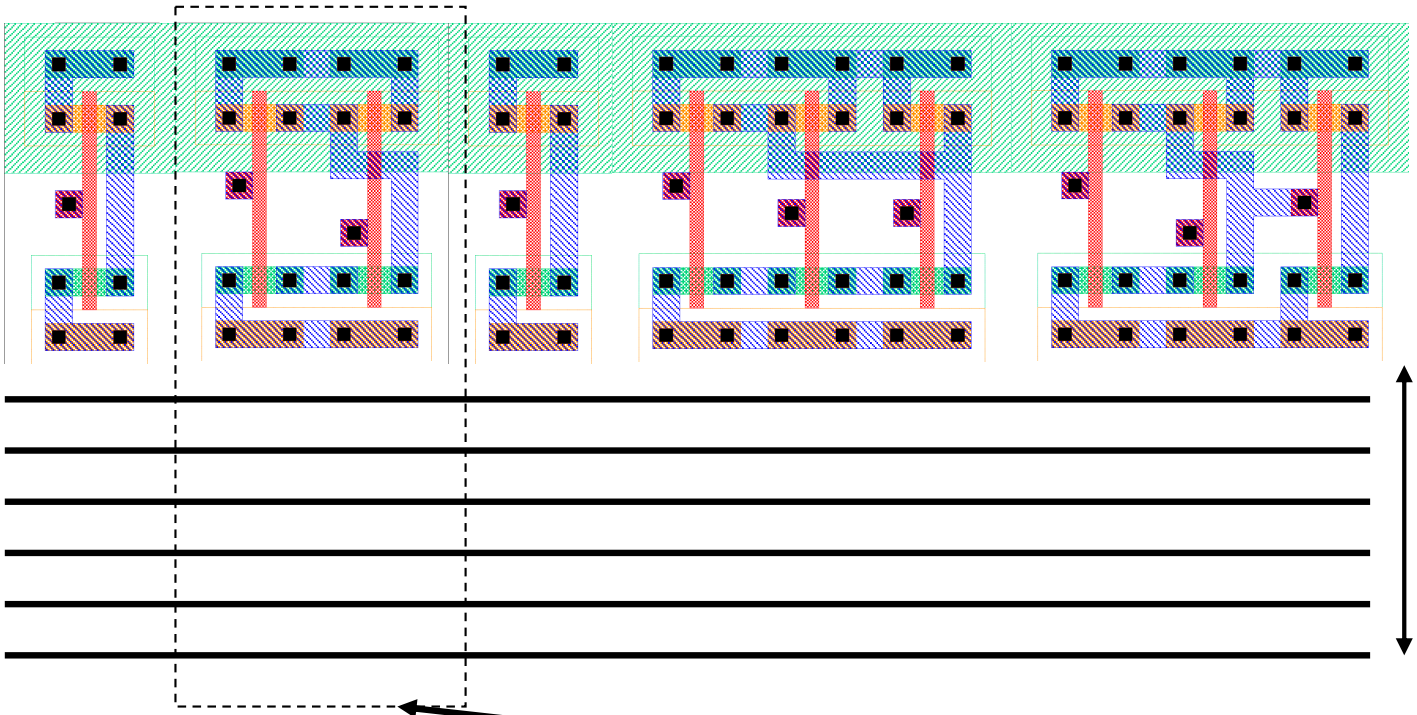
Fanout 4x	0.5 μm	1.0 μm	2.0 μm
A1_tphl	0.595	0.711	0.919
A1_tplh	0.692	0.933	1.360
B1_tphl	0.591	0.739	1.006
B1_tplh	0.620	0.825	1.1.81
C1_tphl	0.574	0.740	1.029
C1_tplh	0.554	0.728	1.026

3-input NAND cell
(from Mississippi State Library)
characterized for fanout of 4 and
for three different technologies

Standard Cell Layout Example



Standard Cell Area

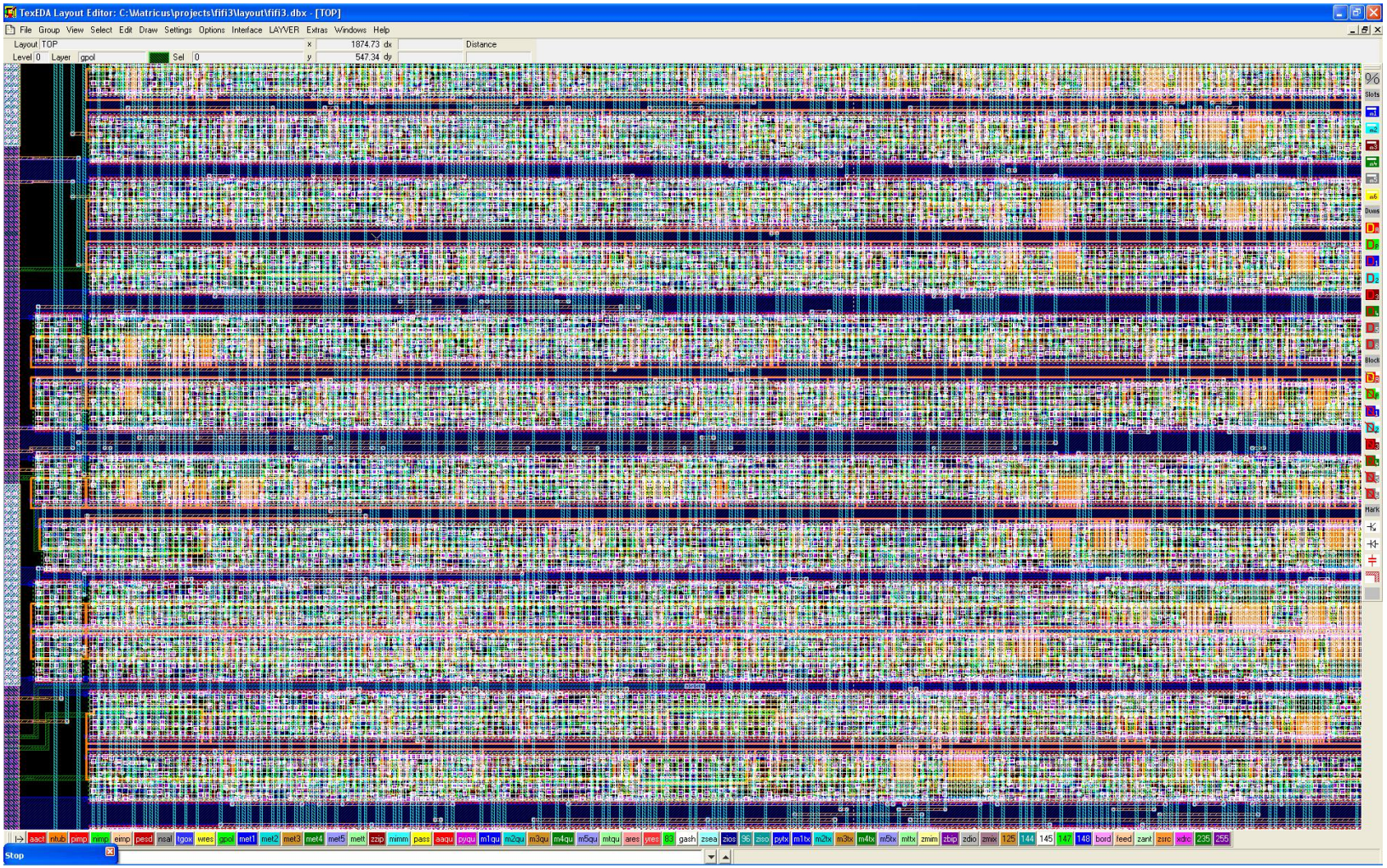


All cells uniform height

Width of channel determined by routing

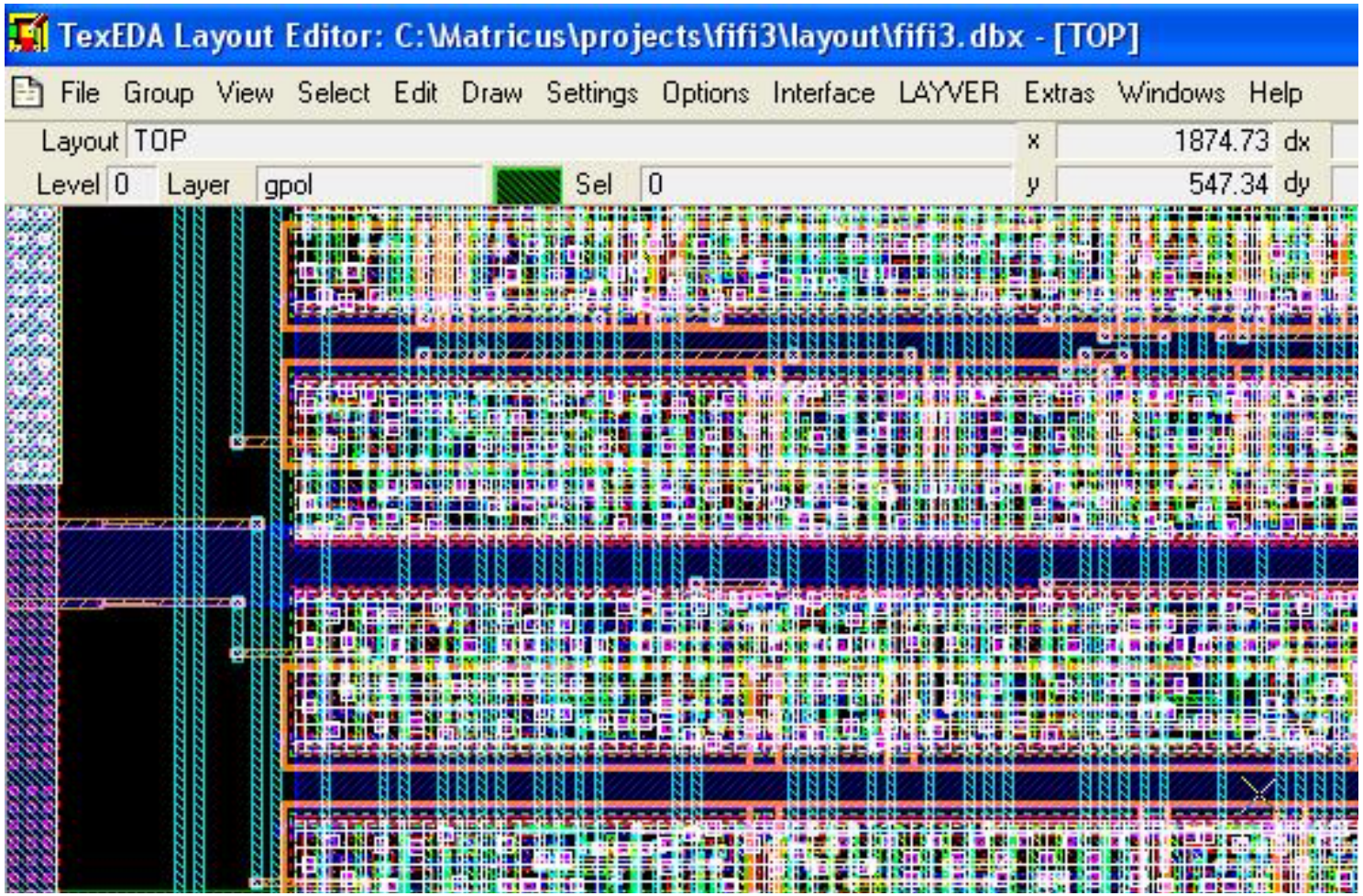
Cell area

Standard Cell Layout Example



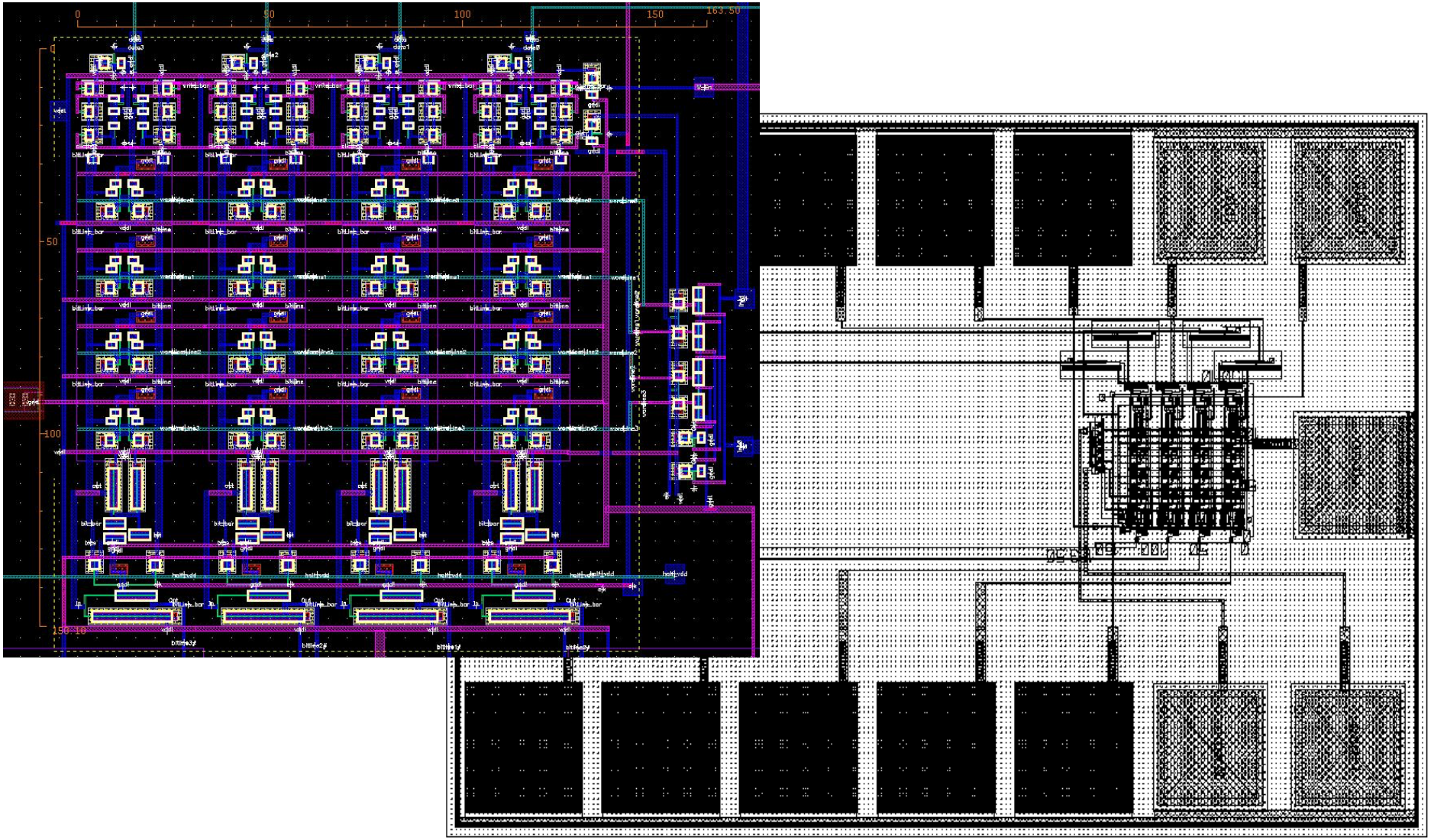
http://www.laytools.com/images/StandardCells.jpg

Standard Cell Layout Example



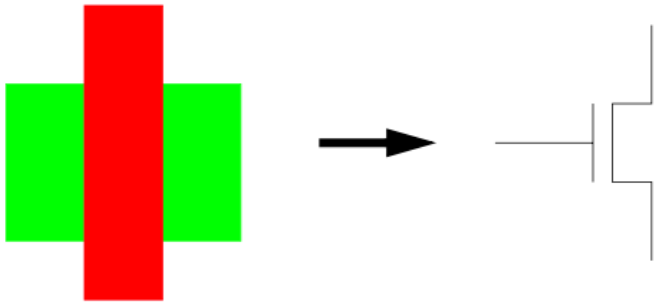
http://www.laytools.com/images/StandardCells.jpg

4x4 6T SRAM Memory



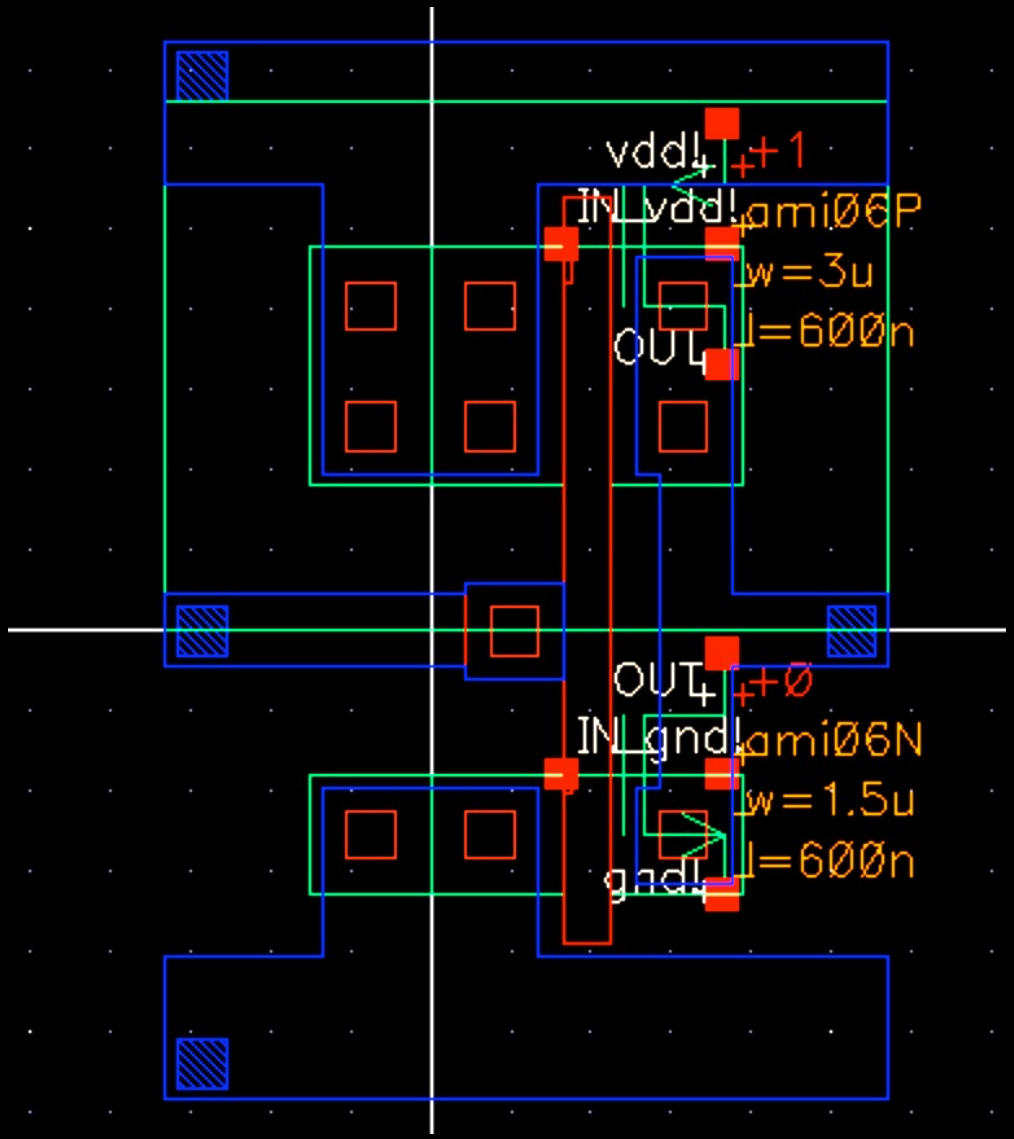
Circuit Extraction

- Circuit extraction extracts a schematic representation of a layout, including transistors, wires, and possibly wire and device resistance and capacitance.

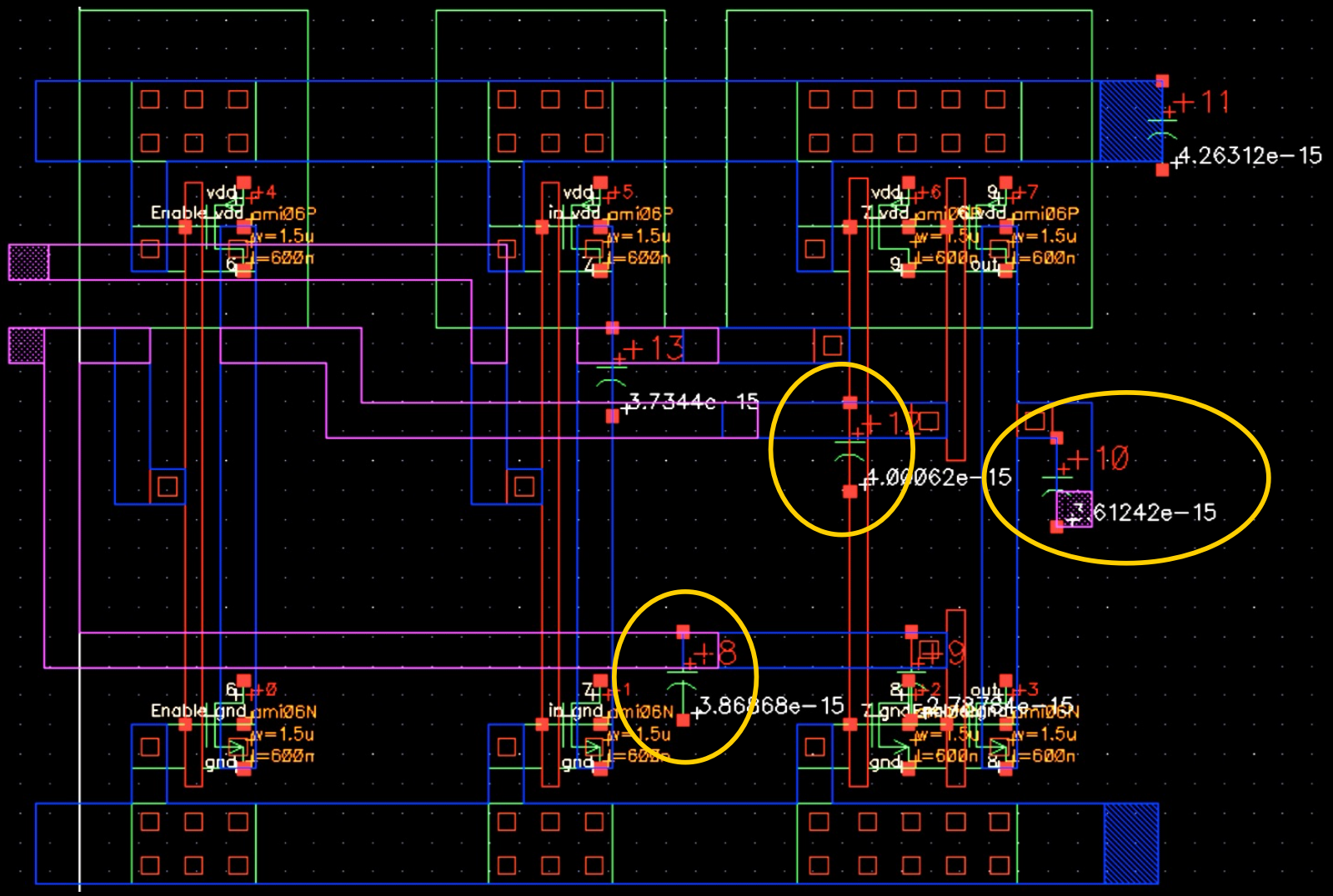


- Circuit extraction is used for LVS, and for spice simulation of layouts

Circuit Extraction



Circuit Extraction

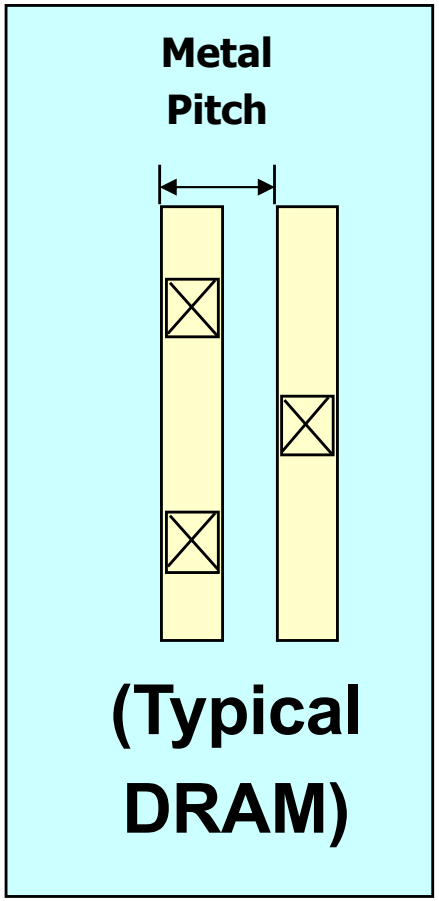


Scaling

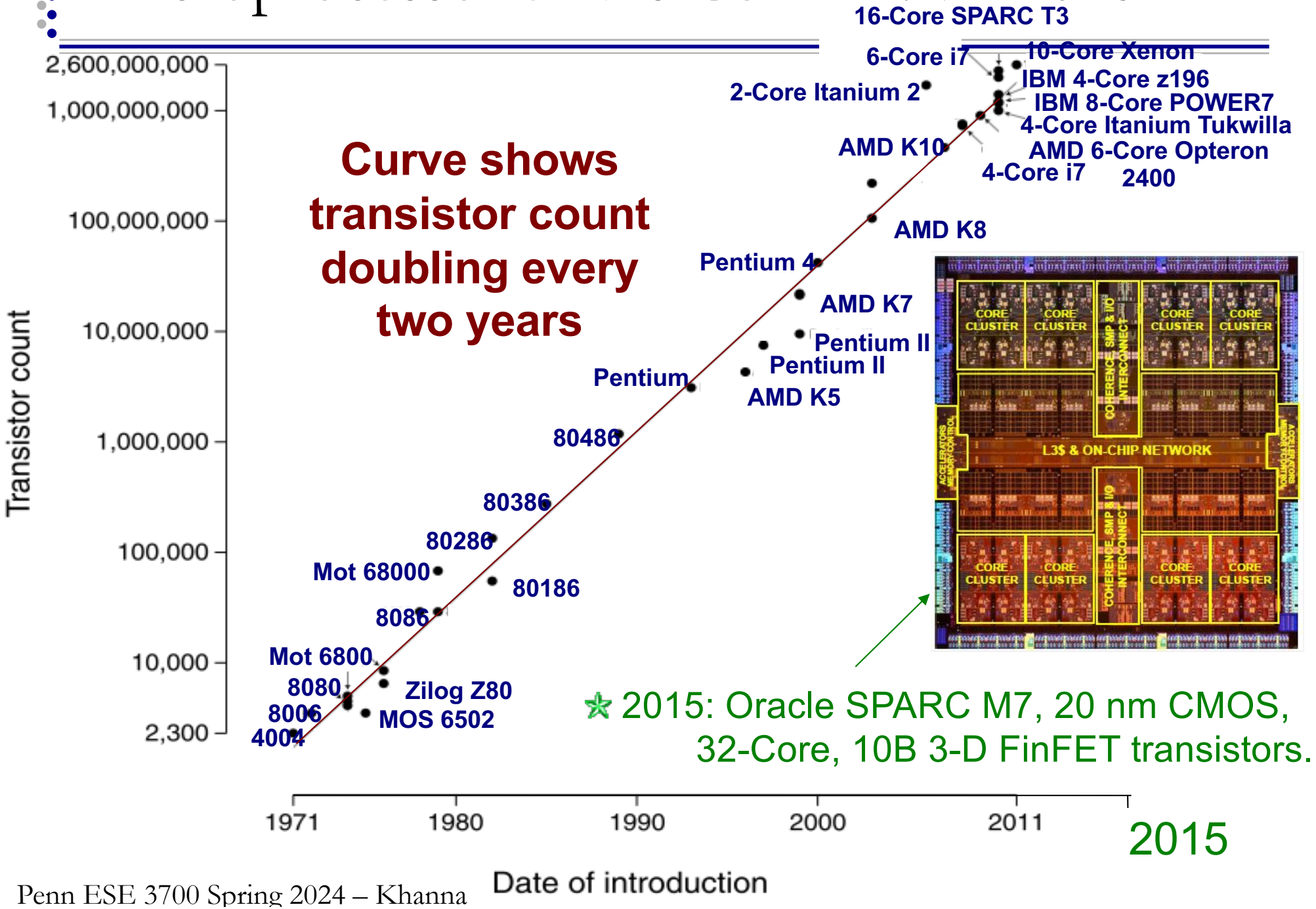
Scaling Technology

- **Premise:** features scale "uniformly"
  - everything gets better in a predictable manner

- **Parameters:**
  - λ (lambda) -- Mead and Conway (L=2λ)
  - F -- Half pitch – ITRS (F=2λ=L)
  - S – scale factor – Rabaey
    - F'=F/S
      - S>1

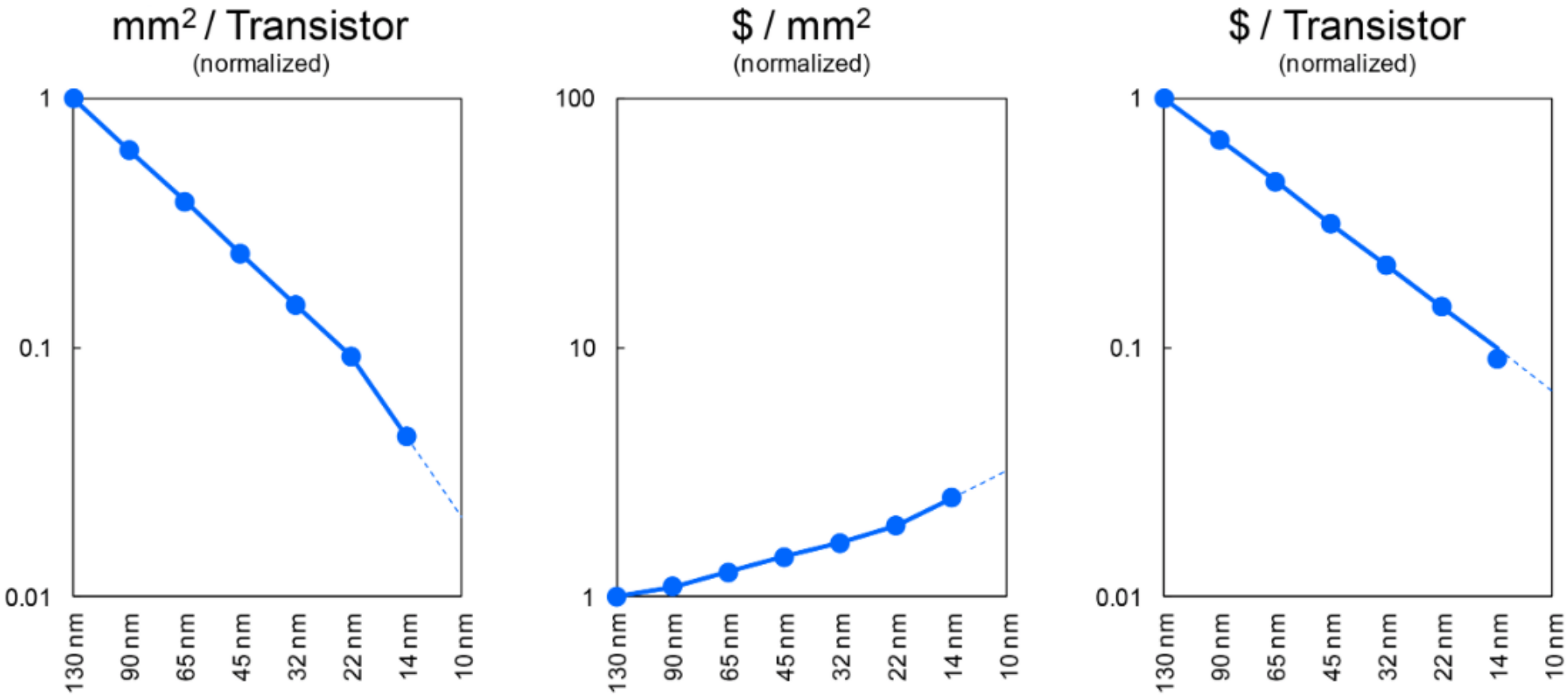


Microprocessor Trans Count 1971-2015



★ 2015: Oracle SPARC M7, 20 nm CMOS, 32-Core, 10B 3-D FinFET transistors.

Intel Cost Scaling

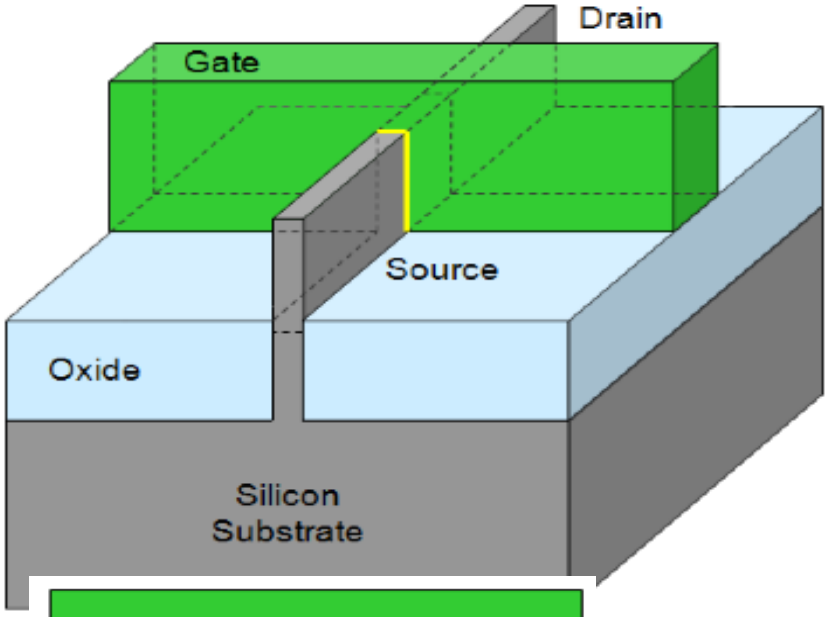


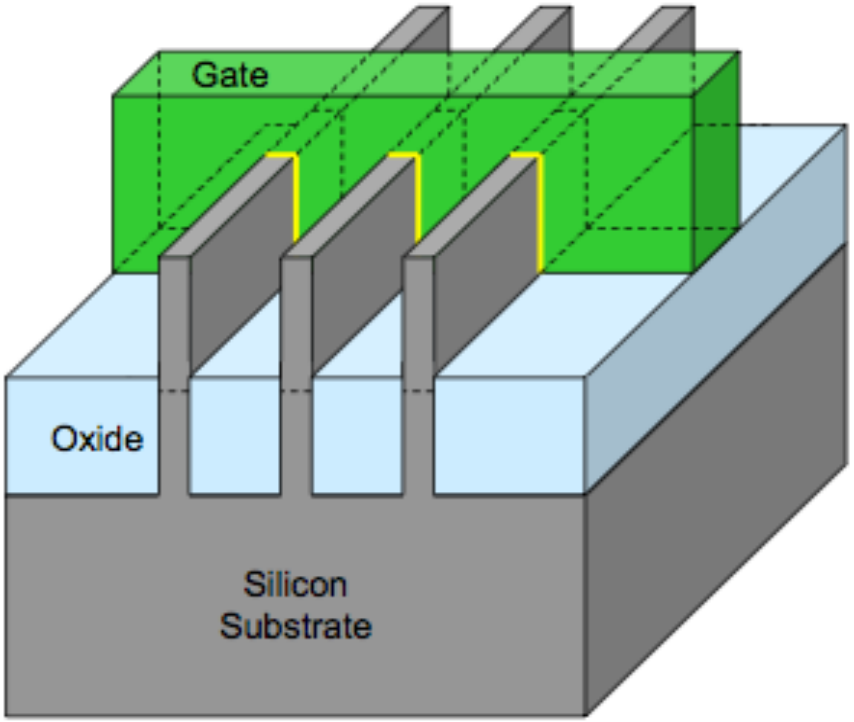
http://www.anandtech.com/show/8367/intels-14nm-technology-in-detail

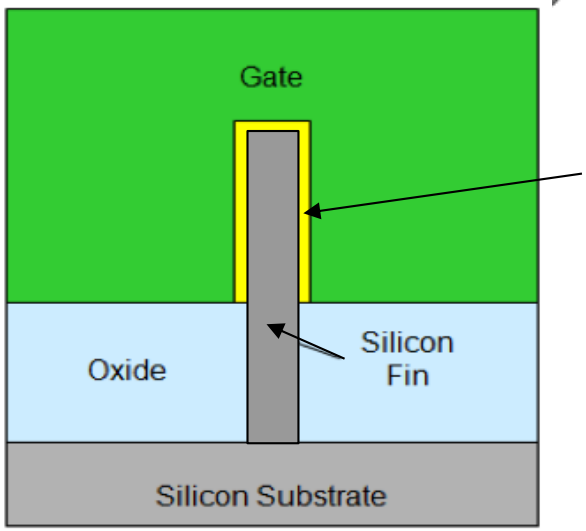
More Moore → Scaling

- Geometrical Scaling
  - continued shrinking of horizontal and vertical physical feature sizes
- Design Equivalent Scaling
  - design technologies that enable high performance, low power, high reliability, low cost, and high design productivity even if neither geometrical nor equivalent scaling can be used

22nm 3D FinFET Transistor







High-k gate dielectric

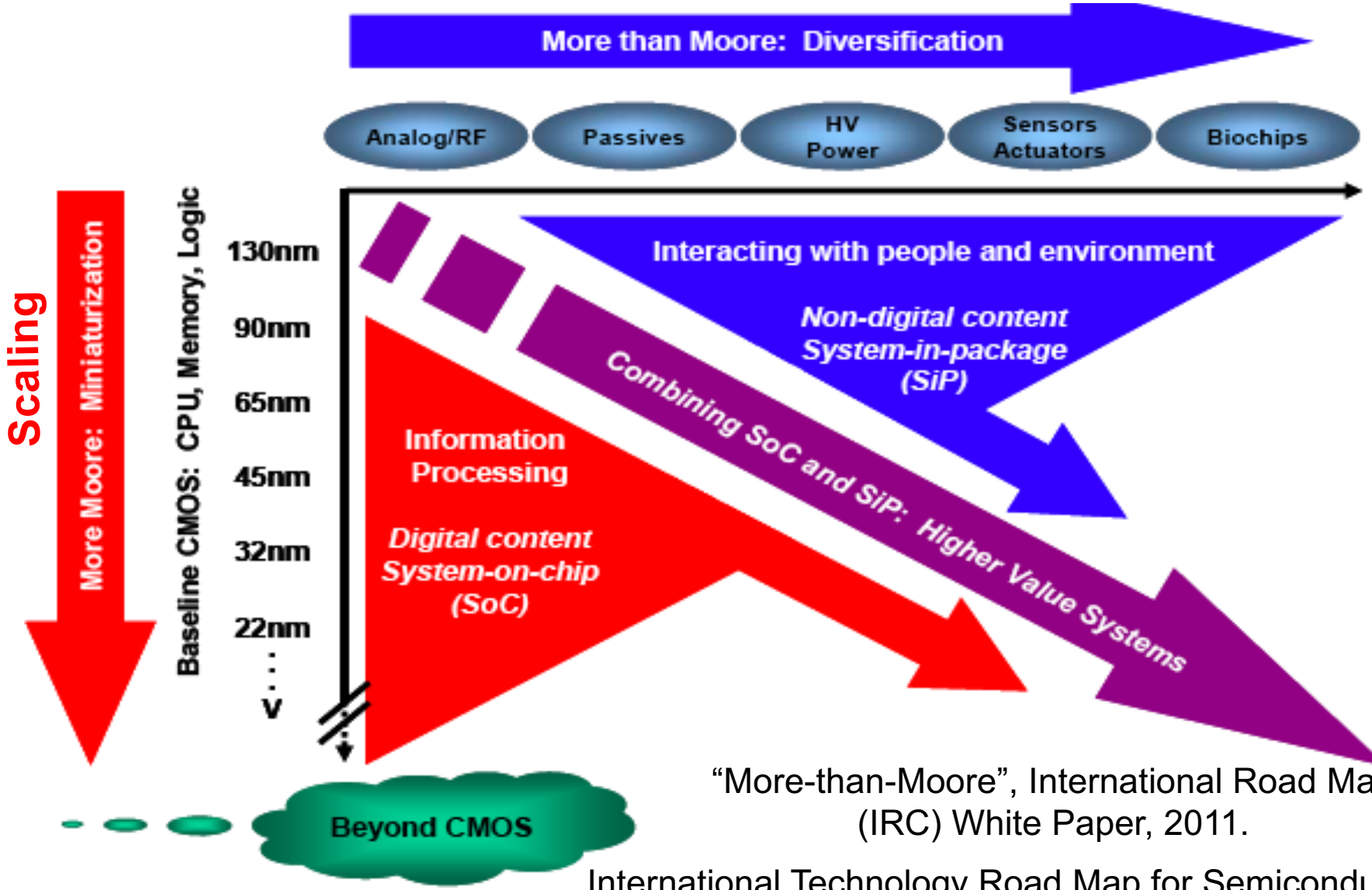
Tri-Gate transistors with multiple fins connected together increases total drive strength for higher performance

http://download.intel.com/newsroom/kits/22nm/pdfs/22nm-Details_Presentation.pdf

ITRS Roadmap

- International Technology Roadmap for Semiconductors
  - Try to predict where industry going
- ITRS 2.0 started in 2015 with new focus
  - System Integration, Heterogeneous Integration, Heterogeneous Components, Outside System Connectivity, More Moore, Beyond CMOS and Factory Integration.
- http://www.itrs2.net/

More-than-Moore



"More-than-Moore", International Road Map (IRC) White Paper, 2011.

International Technology Road Map for Semiconductors

Question (Preclass 5)

- Scaling from 32nm → 22nm, what is 1/S?
  - Scaling minimum gate length
  - And pitch distance

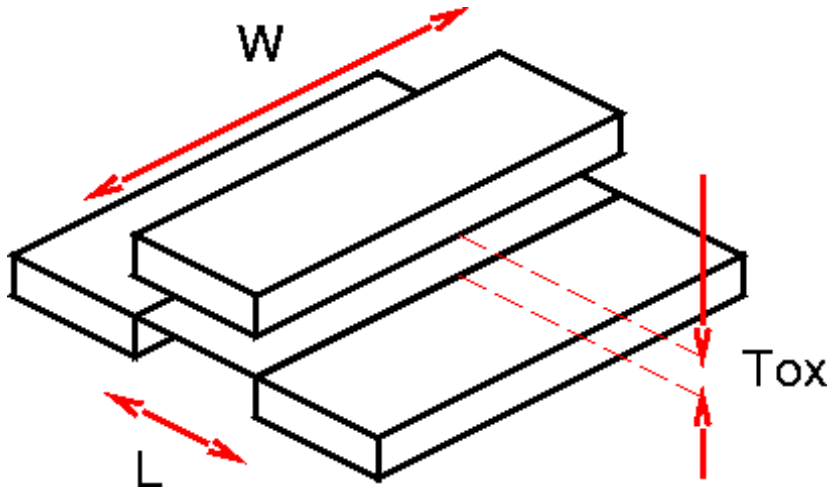
MOS Transistor Scaling - (1974 to present)

1/S=0.7
per technology node
[0.5x per 2 nodes]

Source: 2001 ITRS - Exec. Summary, ORTC Figure, Andrew Kahng

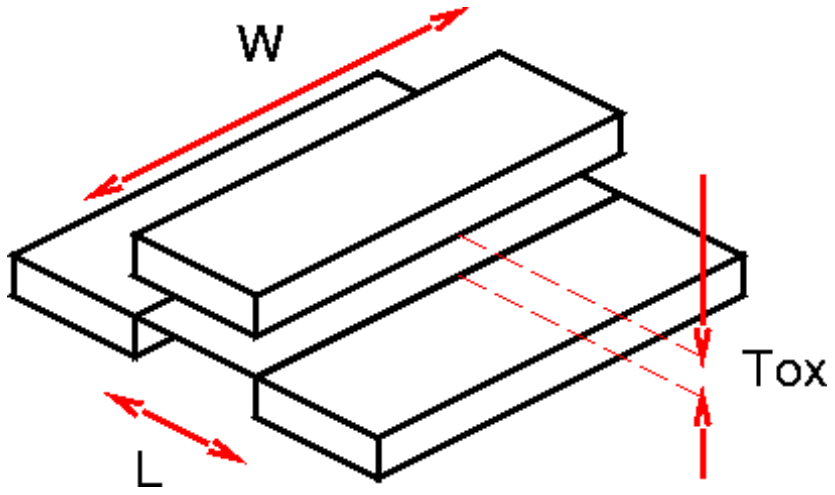
Scaling

- Channel Length (L)
- Channel Width (W)
- Oxide Thickness (t_{ox})
- Doping (N_a)
- Voltage (V_{DD},V_t,)



Full Scaling (Ideal Scaling)

- Channel Length (L) 1/S
- Channel Width (W) 1/S
- Oxide Thickness (t_{ox}) 1/S
- Doping (N_a) S
- Voltage (V_{DD},V_t,) 1/S



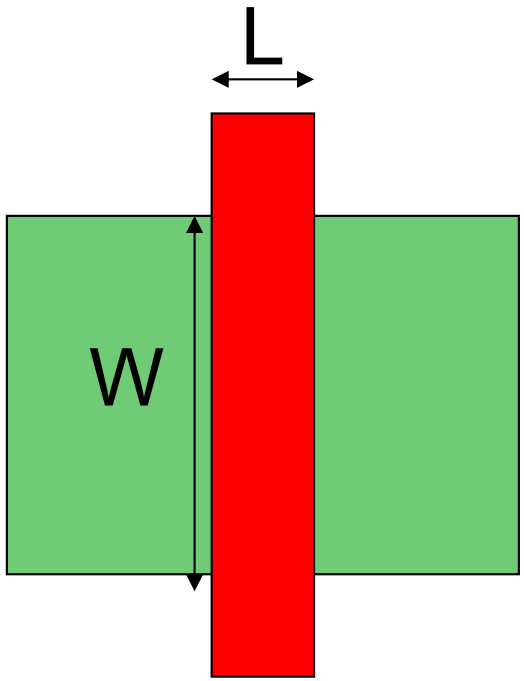
Effects on Physical Properties and Specs?

- Area
- Capacitance
  - C_{ox} and C_{gate}
- Resistance
- Current (I_d)
- Gate Delay (τ_{gd})
- Wire Delay (τ_{wire})
- Power
  - Same frequency
  - Scaled frequency
- Power Density
  - Same frequency
  - Scaled frequency

Preclass 6

Area

- $\lambda' \rightarrow \lambda/S$
- Area impact?
- $A = L \times W$
-

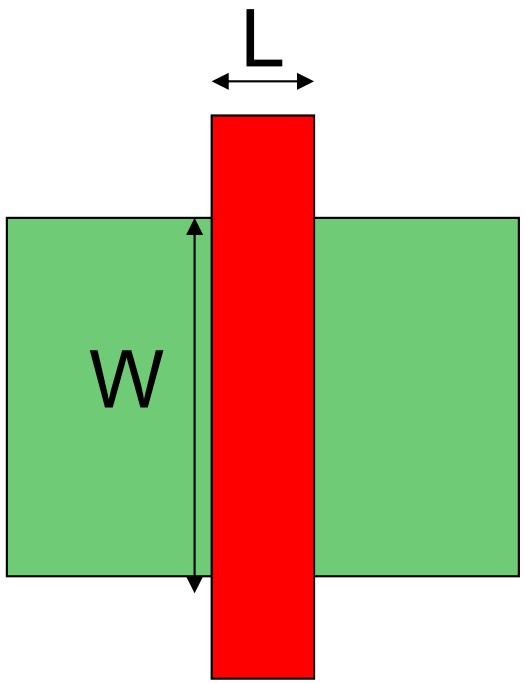


Area

- $\lambda' \rightarrow \lambda/S$
- Area impact?
- $A = L \times W$
- $A' \rightarrow A/S^2$

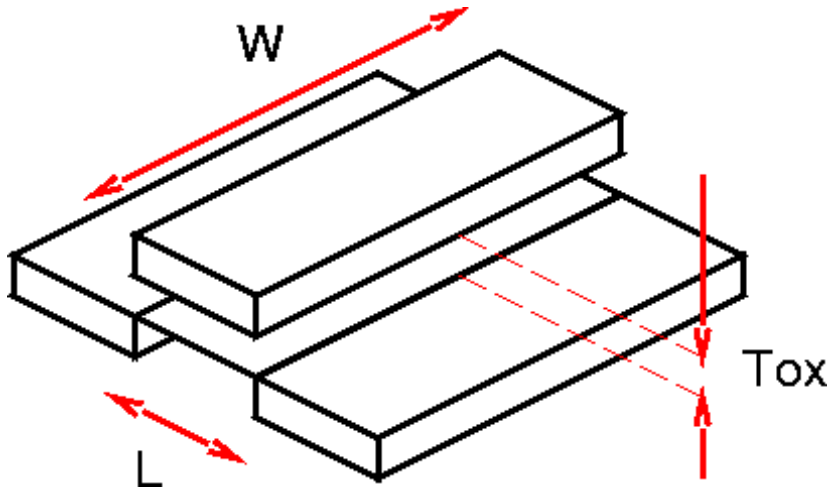
- 32nm → 22nm
- 50% area
- 2 × transistor capacity for same area

$1/S=0.7$



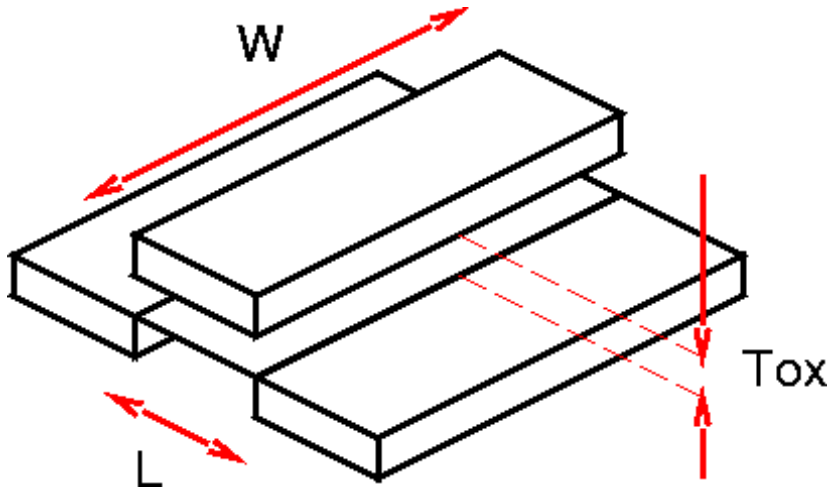
Capacitance

- Capacitance per unit area scaling?
  - $C_{ox} = \varepsilon_{SiO_2}/t_{ox}$
  - $t'_{ox} \rightarrow t_{ox}/S$



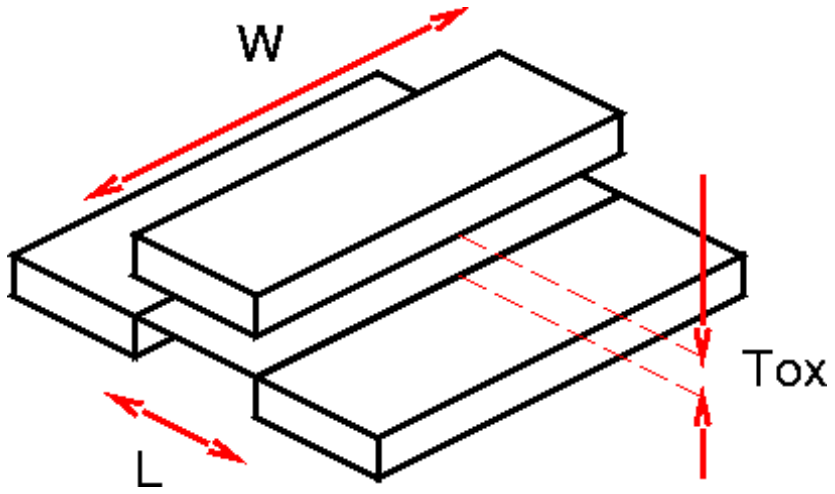
Capacitance

- Capacitance per unit area scaling?
  - $C_{ox} = \varepsilon_{SiO_2}/t_{ox}$
  - $t'_{ox} \rightarrow t_{ox}/S$
  - $C'_{ox} \rightarrow C_{ox} \times S$



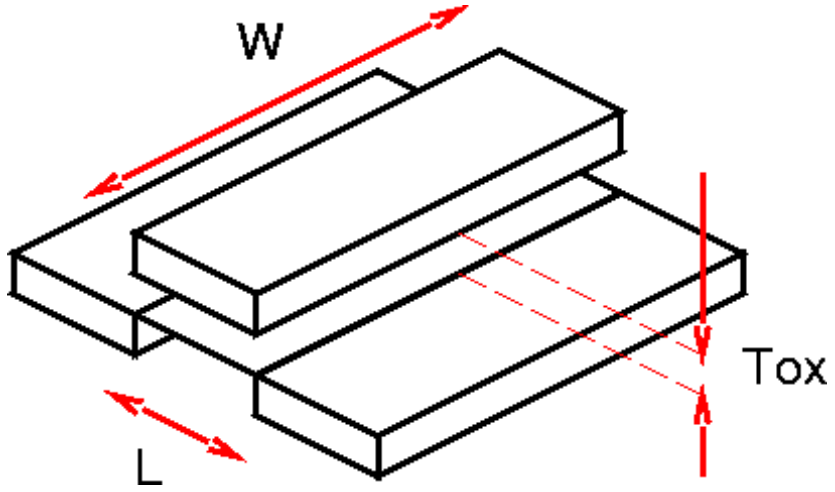
Capacitance

- Gate Capacitance scaling?
  - $C_{gate} = A \times C_{ox}$
  - $A' \rightarrow A/S^2$
  - $C'_{ox} \rightarrow C_{ox} \times S$



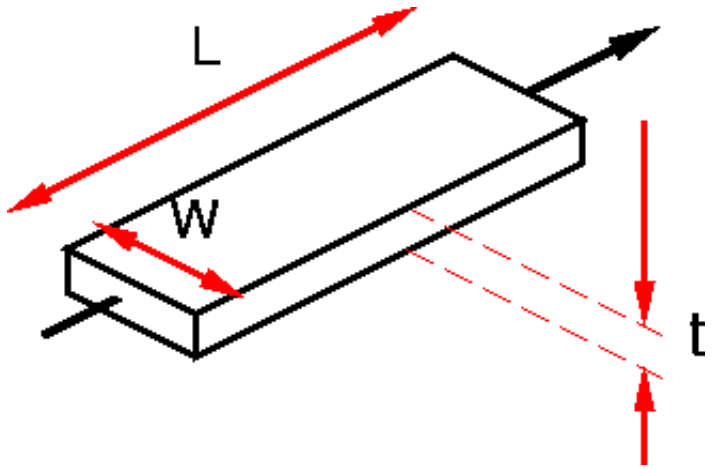
Capacitance

- Gate Capacitance scaling?
  - $C_{gate} = A \times C_{ox}$
  - $A' \rightarrow A/S^2$
  - $C'_{ox} \rightarrow C_{ox} \times S$
  - $C'_{gate} \rightarrow C_{gate}/S$



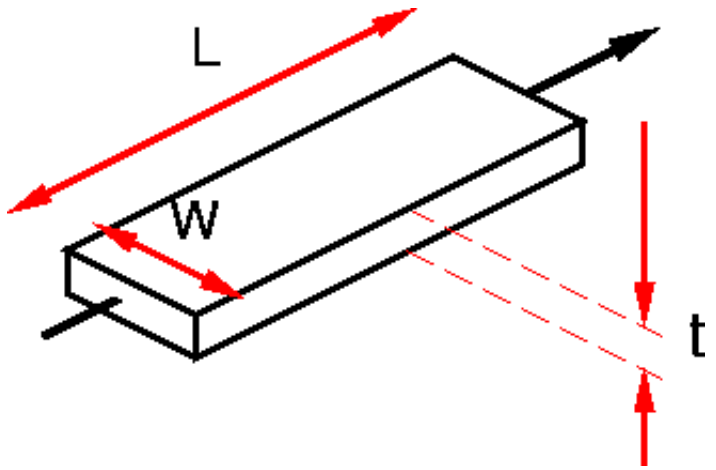
Wire Resistance

- Resistance scaling?
- $R=\rho L/(W*t)$
  - L, t remain similar (not scaled)
- W→ W/S



Wire Resistance

- Resistance scaling?
- $R=\rho L/(W*t)$
  - L, t remain similar (not scaled)
- W → W/S
- R' → R×S



Current

- Which Voltages matter here? (V_{gs},V_{ds},V_{th}…)
- Transistor charging looks like voltage-controlled current source
- Saturation Current scaling?

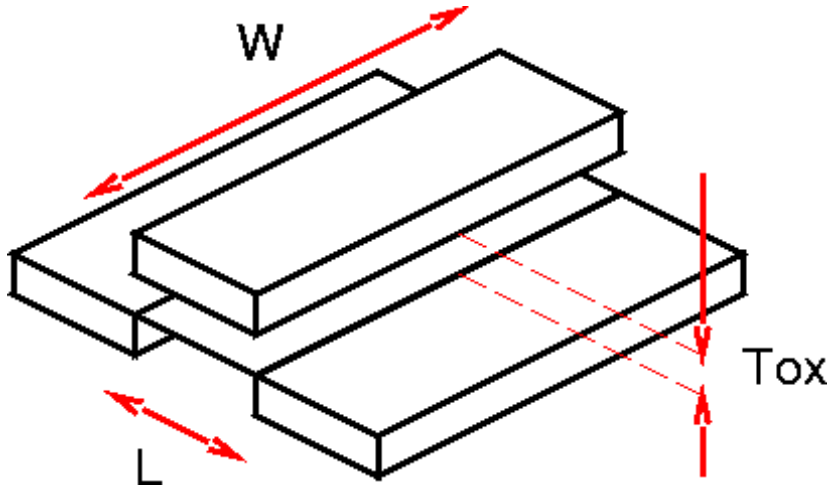
$$I_d=(\mu C_{OX}/2)(W/L)(V_{gs}-V_{TH})^2$$

V_{gs},V_{TH}: V'→ V/S

W'→ W/S

L'→ L/S

C'_{ox} → C_{ox}×S



Current

- Which Voltages matters here? (V_{gs},V_{ds},V_{th}…)
- Transistor charging looks like voltage-controlled current source
- Saturation Current scaling?

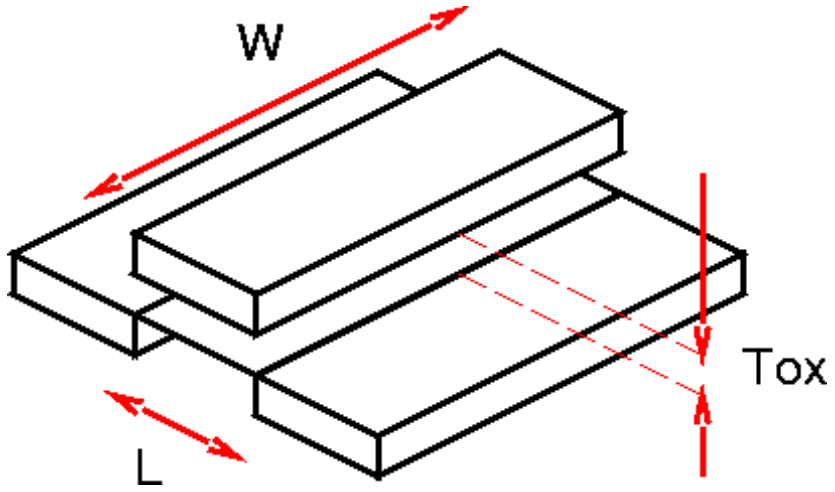
$$I_d=(\mu C_{OX}/2)(W/L)(V_{gs}-V_{TH})^2$$

V_{gs},V_{TH}: V'→ V/S

W'→ W/S

L'→ L/S

C'_{ox} → C_{ox}×S



$$I'_d=(\mu C_{OX}S/2)((W/S)/(L/S))(V_{gs}/S-V_{TH}/S)^2$$

Current

- Which Voltages matters here? (V_{gs},V_{ds},V_{th}…)
- Transistor charging looks like voltage-controlled current source
- Saturation Current scaling?

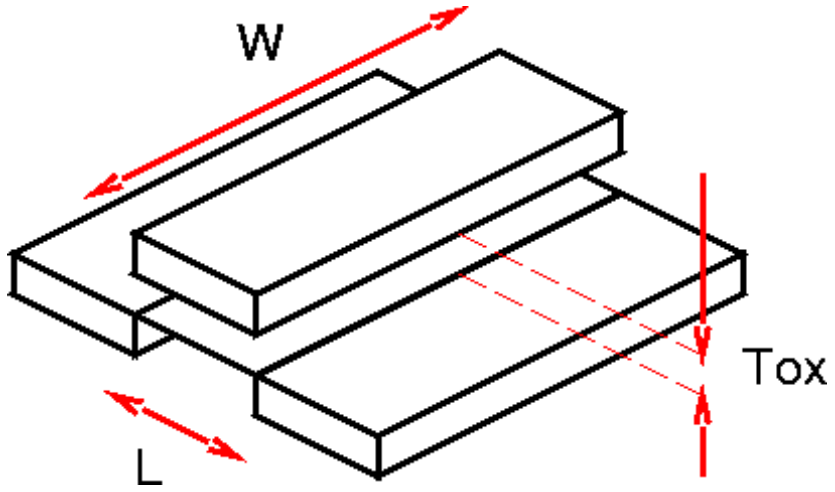
$$I_d=(\mu C_{OX}/2)(W/L)(V_{gs}-V_{TH})^2$$

V_{gs},V_{TH}: V'→ V/S

W'→ W/S

L'→ L/S

C'_{ox} → C_{ox}×S



I'_d→I_d/S

Current

- Velocity Saturation Current scaling?

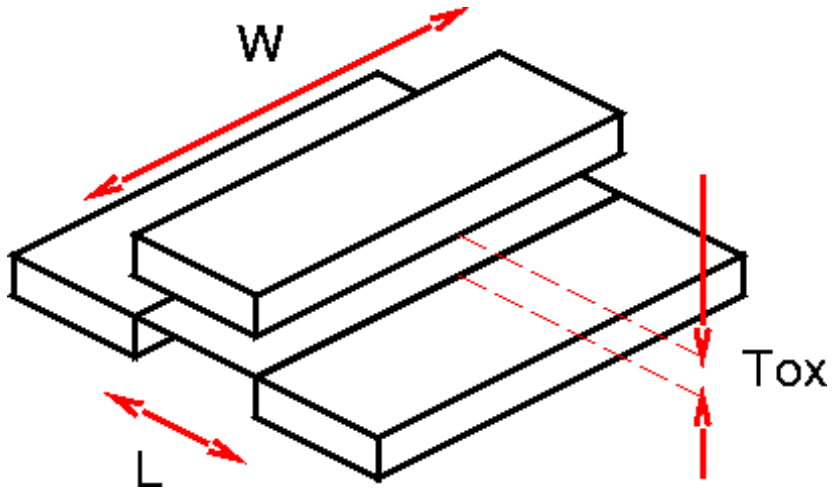
$V_{gs}, V_{TH}: V' \rightarrow V/S$

$L' \rightarrow L/S$

$W' \rightarrow W/S$

$C'_{ox} \rightarrow C_{ox}S$

$$I_{DS} \approx \nu_{sat} C_{OX} W \left(V_{GS} - V_{TH} - \frac{V_{DSAT}}{2} \right)$$



Current

- Velocity Saturation Current scaling?

$V_{gs}, V_{TH}: V' \rightarrow V/S$

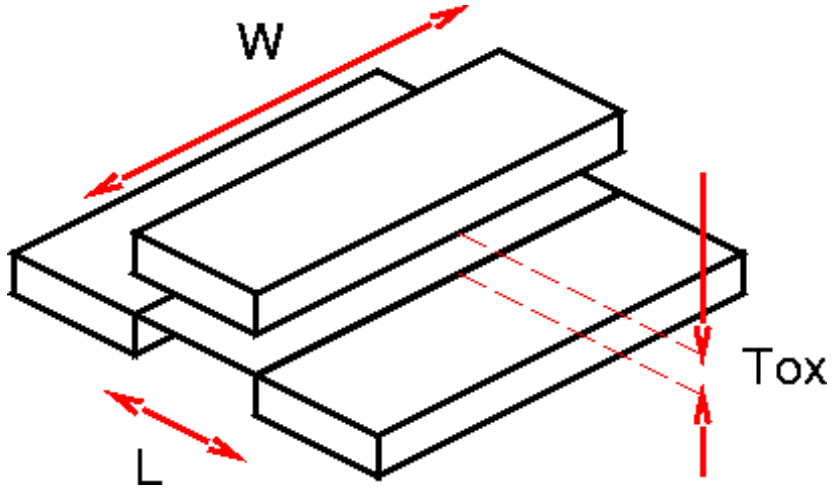
$L' \rightarrow L/S$

$W' \rightarrow W/S$

$C'_{ox} \rightarrow C_{ox}S$

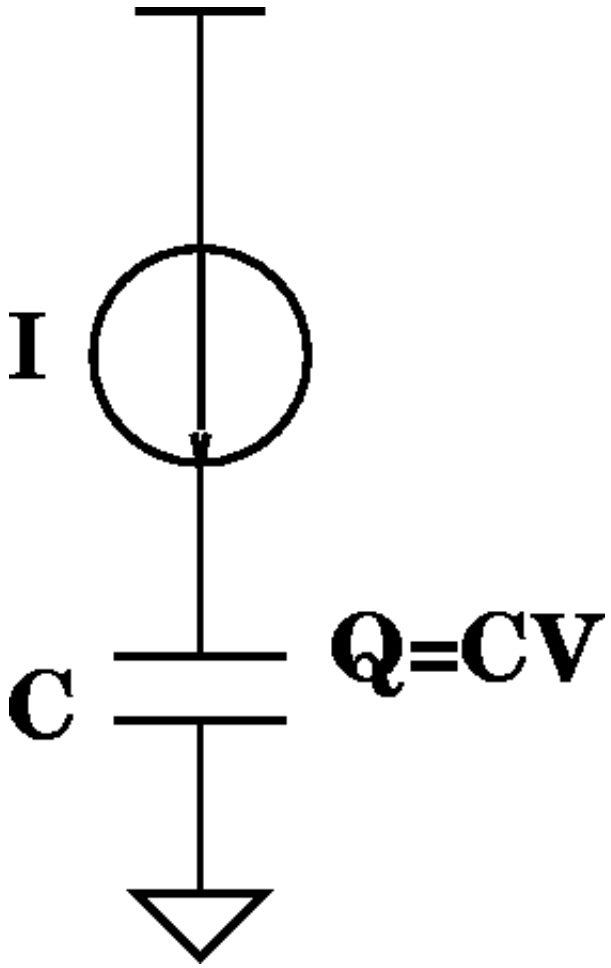
$I'_d \rightarrow I_d/S$

$$I_{DS} \approx \nu_{sat} C_{OX} W \left(V_{GS} - V_{TH} - \frac{V_{DSAT}}{2} \right)$$



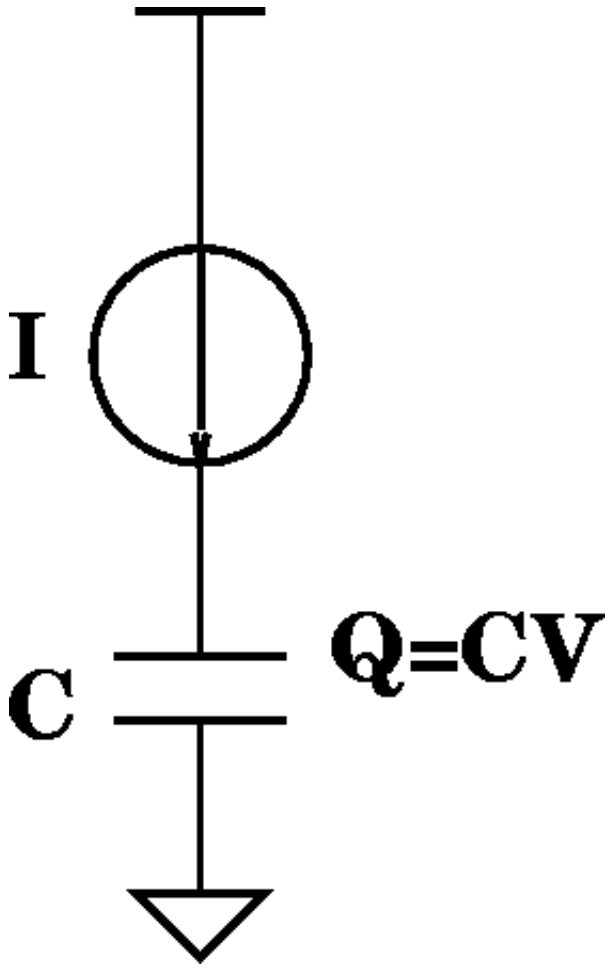
Gate Delay

- Gate Delay scaling?
- $\tau_{gd}=Q/I=(CV)/I$
- V'→ V/S
- $I'_d \rightarrow I_d/S$
- $C_g' \rightarrow C_g/S$



Gate Delay

- Gate Delay scaling?
- $\tau_{gd}=Q/I=(CV)/I$
- V'→ V/S
- $I'_d \rightarrow I_d/S$
- $C_g' \rightarrow C_g/S$
- $\tau'_{gd} \rightarrow \tau_{gd}/S$



Wire Delay

- Wire delay scaling?
- $\tau_{wire} = R \times C$
- R' → R×S
- C' → C/S

  - Again assuming (logical) wire lengths remain constant

Wire Delay

- Wire delay scaling?
- $\tau_{wire} = R \times C$
- R' → R×S
- C' → C/S
- $\tau'_{wire} \rightarrow \tau_{wire}$

  - Again assuming (logical) wire lengths remain constant

Power Dissipation (Dynamic)

- Capacitive (Dis)charging scaling?
  - $P=(1/2)CV^2f$
  - V'→ V/S
  - C' → C/S

Power Dissipation (Dynamic)

- Capacitive (Dis)charging scaling?
  - $P=(1/2)CV^2f$
  - $V' \rightarrow V/S$
  - $C' \rightarrow C/S$
  - $P' \rightarrow P/S^3$

Power Dissipation (Dynamic)

- Capacitive (Dis)charging scaling?
  - $P=(1/2)CV^2f$
  - $V' \rightarrow V/S$
  - $C' \rightarrow C/S$
  - $P' \rightarrow P/S^3$
- Increase Frequency?
  - $\tau_{gd} \rightarrow \tau_{gd}/S$
  - So: $f \rightarrow f \times S$

Power Dissipation (Dynamic)

- Capacitive (Dis)charging scaling?
  - $P=(1/2)CV^2f$
  - $V' \rightarrow V/S$
  - $C' \rightarrow C/S$
  - $P' \rightarrow P/S^3$
- Increase Frequency?
  - $\tau_{gd} \rightarrow \tau_{gd}/S$
  - So: $f \rightarrow f \times S$
  - $P \rightarrow P/S^2$

Effects?

Area	$1/S^2$
Capacitance (C_{ox}, C_g)	S, 1/S
Resistance	S
Threshold (V_{th})	1/S
Current (I_d)	1/S
Gate Delay (τ_{gd})	1/S
Wire Delay (τ_{wire})	1
Power	$1/S^3$, $1/S^2$

1/S=0.7

Power Density

- P' ➔ P/S^2 (increased frequency)
- P' ➔ P/S^3 (same frequency)
- A' ➔ A/S^2

- Power Density: P/A two cases?

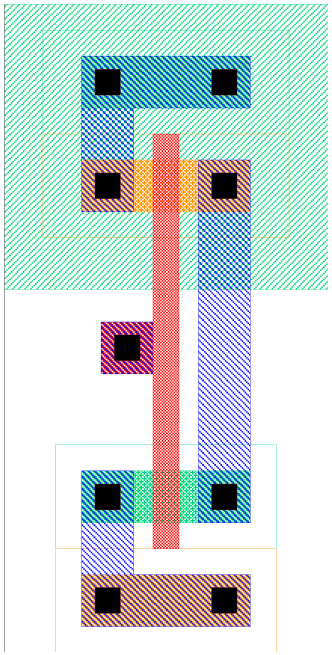
Power Density

- P' ➔ P/S^2 (increased frequency)
- P' ➔ P/S^3 (same frequency)
- A' ➔ A/S^2

- Power Density: P/A two cases?
  - P/A ➔ P/A increase freq.
  - P/A ➔ (P/A)/S same freq.

Big Idea

- Layouts are physical realization of circuit
  - Geometry tradeoff
    - Can decrease spacing at the cost of yield
    - Design rules
- Can go from circuit to stick diagram/layout or stick diagram/layout to circuit by inspection
- Moderately predictable VLSI Scaling
  - unprecedented capacities/capability growth for engineered systems
  - …but hits physical limit



Admin

- HW3 out now – due 2/16 (Friday)
  - Don't forget the demo/video of SPICE workflow
    - Get TA checkoff or submit video in Canvas
- Midterm 1 Wednesday 2/21 (next week)
  - 1.5 hrs during class in Moore 216
  - Midterm 1 TA review session, Saturday 2-4pm
    - location TBD. Keep an eye on Ed
  - Covers lecture 1-6
  - Old exams posted on previous years websites
    - Note old exams were for 2hrs
- HW 4 posted on 2/21

Midterm 1 - Coverage

- Lec 1 - 6
  - Identify CMOS/non-CMOS
  - Identify CMOS function
  - Any logic function → CMOS gate
  - Noise Margins / Restoration
  - Circuit first order switching rise/fall times
    - Output equivalent resistance
    - Load capacitance
  - MOS Model
    - Identify transistor region of operation
    - Analysis with transistor IV models
    - MOS capacitance models

Acknowledgement

- Prof. André DeHon (University of Pennsylvania)
- Prof. Jing Li (University of Pennsylvania)