LAYOUT DESIGN

LEC1: Intro

Physical (Back End) Design steps:

- Partitioning – Break the design into independent blocks also known as partitions (with as little interaction with each other)
  - Granularity (size or number of components (sub-blocks)) varies with abstraction level.
    - In digital circuits: Basic gates (with few MOSFETs) at the lowest level (Leaf cell) to IP blocks (with millions of MOSFETs) at a higher level of abstraction
- Floor planning – Decide the approximate coordinates of the blocks in the design
- Placement – Place the blocks and fix the coordinates of the blocks
- Power Planning – Power nets planned to reduce IR drops
- Clock tree synthesis (CTS) – Plan the clock distribution – Almost every block will have large number of clock points. Large fan-out – Clock buffers planned.
- Routing – Interconnect the blocks (cells) using wires (metal) to achieve the functionality with least wire delay
  - Global Routing – Input is floor plan. Decides loose plan for interconnects – estimate the routes and routing congestion
  - Detailed Routing – Actual routes of each interconnect with geometry, sizing and delay computation
- Physical Verification – verify whether the actual layout represents the desired RTL.

Digital ASIC Design methodology

Two types:

- Custom Design – Each cell is handcrafted (specifically designed)
- Semi-custom Design - Pre-designed functional blocks used

The granularity of the functional blocks may be fine or coarse

- Fine - Standard Cell Library – Collection of simple combinational and sequential elements with various views of each cell
  - Cell – may contain few MOSFETs
  - Views – Behavioural, circuit, Symbol, Layout, timing etc.
- Coarse – IP blocks
  - An IP block may contain millions of MOSFETs

Standard cell library is a collection of characterized functional basic cells

- Standard cells:
  - Simple and complex Combinational gates like NAND, NOR, AND, OR, AOI, OAI, XOR, XNOR with various Fan-ins and Drive strengths, Buffers, Clock buffers
  - Sequential elements like Flip-flops and Latches
  - Other cells like Tie-Hi, Tie-Low, Filler, IO Pads, Pad Filler etc.

Course granular cells are called Intellectual Properties (IP blocks)

- ALU, Integer and Floating point arithmetic Units, Register files, Memory blocks, IO controllers like USB, SPI, CAN, PCI etc.

Standard cell based Layout:

- **RTL is bound to Standard Cell Library**
- **Bound standard cells placed as rows – All rows have uniform width.**
- **Each row will have many standard cells abutting each other**
- **Rows may have space between them – called Channel – used for carrying interconnects (routing wires) – Older designs, when not enough metal layers were available. - Channel Routing**
- **If no space is left between rows. Rows abut each other – Routing strategy is Over the Cell routing. Alternate rows are flipped vertically to save space and reduce effective supply wire resistance.**

Routing:

- **Channel routing: Horizontal space (Channel) between rows and vertical space between cells in a row (feed-through) - used to route the interconnection of cells.**
- **Over the Cell routing: Metal layers other than ones used in the standard cell are used for routing.**
  - **Long length interconnects, clock and power distribution use top metal layers. Result in less interconnect resistance as top metal layers are thicker in comparison with lower metal layers**

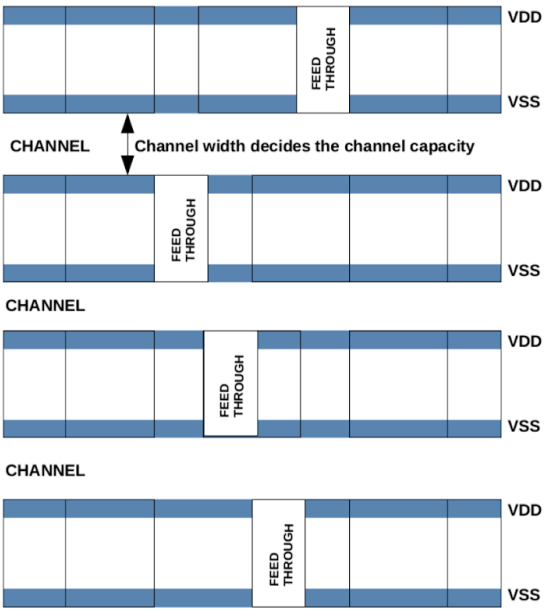
Over the cell routing additional notes:

1. Top layer metal thickness > bottom layer, thickness(doesn't refer to width here) is more implies, resistance is higher,hence, top layer metal offers least resistance. (hence used for power distribution.)

2. supposing say you wish to connect 2 gates, then poly to poly connection can be made with another poly itself. Although poly may have a larger resistance per square, but you save on contact resistance etc. so for short distances, you can use poly instead of metal for routing 2 transistor **gates.**

Channel routing additional notes: arrays of standard cells arranged as rows.

1. Inter row space - channel, intra-row gaps - feed through.
2. Feed through capacities don't generally require a large amount of space.
3. Odd and even metal lines are made of the same material, just present in different layers
4. Vias - by drilling a hole through the silicon dioxide lying between 2 consecutive metals,



Channel Routing

No inter-cell connections run over the cells.

Channels:

- Separation between rows of cells
- Used for routing interconnects along rows; both in x and y direction within the channel.
- Width decided by the number of wires, their widths and separation between the wires (pitch)

Feed-through:

- Separation between cells in a row
- Used for routing interconnects across rows; only run in vertical direction within the feed-through
- Width decided by the number of wires, their widths and separation between the wires (pitch)

Odd metal lines run horizontal

Even metal runs vertical

Different metals are connected using Vias

Abutted VDD, VSS lines and N-well

Over the Cell Routing

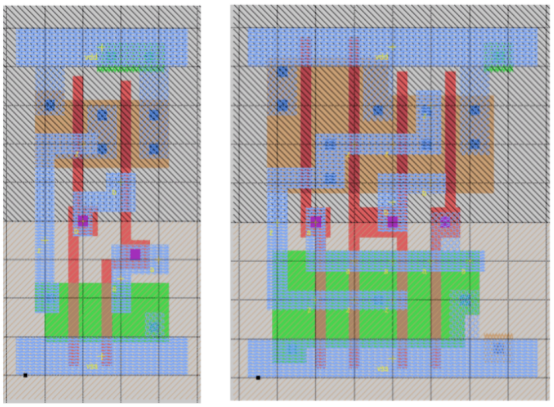
- No channels
- No feed-throughs
- Rows are flipped vertically and abut each other
- Only those metal layers that are not used in intra-cell routing are used for inter-cell routing
- Manhattan routing (only perpendicular horizontal and vertical lines
- Odd metal lines run horizontal
- Even metal runs vertical
- Different metals are connected using Vias

* Flipping allows Vss and Vdd of two different rows to be abutted -> larger Vss and Vdd lines
* Wider Vss range reduces overall resistance of Vss -> reduced drop across it
* N Well is used to place PMOS and white area - NMOS
* Different metal layers are used to avoid short circuit.

LEC2: Standard Cells

Examples of Standard cells

NAND2x1 NAND2x4



NAND2x1: 2 input NAND gate with a drive strength of 1

NAND2x4: 2 input NAND gate with a drive strength of 4

Drive strength: Indicative of capacity to drive fan-out of 'n' standard inverters

* all cells in standard lib have fixed height
* typically metal 1 is used for Vdd and Vss rail (blue)
* orange **horizontal** - p **diffusion**, green **horizontal** - n **diffusion**
* **poly** lines (red) which form the gate run **vertical**

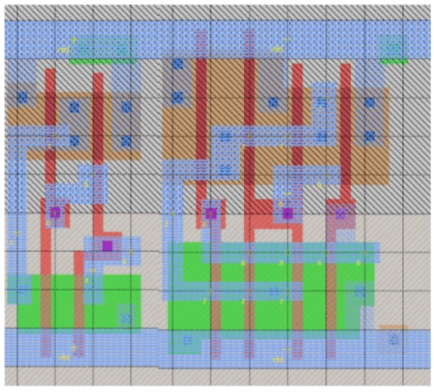
Features / Properties of Standard cells:

- **Fixed height, variable width – can create rows of standard cells**
- **VDD rail of fixed width at the top – runs horizontal – end to end**
  - **When cells are abutting horizontally, individual VDD rails touch each other and create a single VDD rail**
- **VSS rail of fixed width at the bottom – runs horizontal – end to end**
  - **When cells are abutting horizontally, individual VSS rails touch each other and create a single VSS rail**
- **N-well in the upper portion – runs end to end**
  - **When cells are abutting horizontally, individual N-wells touch each other and create a single N-well**
- **All P MOSFETs in the upper part of the cell**
- **All N MOSFETs in the lower part of the cell**

Features / Properties of Standard cells (continued):

- Single Line Diffusion technique is used
- N Diffusion runs horizontal as a line
- P Diffusion runs horizontal as a line
- To reduce the height, wide transistors are implemented as folded transistors
- Poly Lines representing the input Literals are running vertical
- P-substrate (P-well) contacts connected to the VSS rail
- N-well contacts connected to the VDD rail
- Only restricted metal layers (M1 or M1 and M2) used for interconnecting components within the cells (intra-cell routing), rest of the metal layers used for inter-cell routing
- Use array of Contacts and Vias as much as possible – improves reliability, yield as Contacts and vias are points of physical defects
- All objects are placed on Grids – Grid pitch is a (sub)multiple of λ – Auto routing becomes easier

Abutting cells to form a row of cells



Layout:

- Process of translation of a circuit schematic to the masks of all layers.
- Mask of a layer contains rectangular geometrical objects representing the shapes, dimensions and location of the material.

Design Rule Check:

Important from fabrication yield point of view, not from electrical behaviour or correctness point of view.

Two types – (i) Scalable (Lambda Rule) and (ii) Micron Rule

- Scalable (or λ Rule)
  - Conservative, dimensions defined in terms of λ (2λ = feature size) – all measurements are relative
  - Used in academic designs. Open source tools like MAGIC use these rules
  - The layout can be reused for a different process technology by linearly scaling the mask – portable across processes if layers match.
- μ Rule
  - Used in commercial designs. Commercial EDA tools use these Rules
  - All dimensions defined in terms of micrometers - actual
  - Very specific and tight. Saves space.
  - Not portable across process technology

Design Rules Define: (Refer to the example file: scmos.tech)

- Minimum dimension of an object in a given layer
  - eg. Diffusion must be at least 3λ
- Minimum separation between two objects in a given layer
  - eg. separation between two M1 should be at least 3λ
- Minimum separation between two different layers
  - eg. separation between Ndiff and PDiff should be at least 10λ
- Minimum overlaps between objects in different layers
  - eg. Overhang of Poly over Diffusion must be at least 2λ

Electrical Rule Check (ERC):

Important from electrical behaviour or correctness point of view.

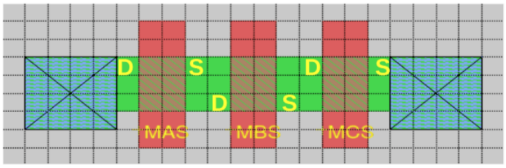
- No short circuit between VDD and VSS
- No open or unconnected nodes, hanging nodes
- Antenna effect (possibility of radiation, EMC / EMI compatibility)

LEC3: Layout Optimization

Creation of optimized layouts of MOSFETs in series, parallel, series parallel and multi-fingered MOSFET

Example of unoptimized layout of 3 MOSFETs in series

Example of optimized layout of 3 MOSFETs in series



Layout Optimization:

Reduces:

- Area required
  - Leakage
- Delay
  - Parasitic capacitance CDB, CSB of every MOSFET
  - On-Resistance of each of the MOSFET as the length of the Source and the Drain are reduces.
  - On-Resistance of each of the MOSFET as there are no contacts and wires connecting individual MOSFETs in series.

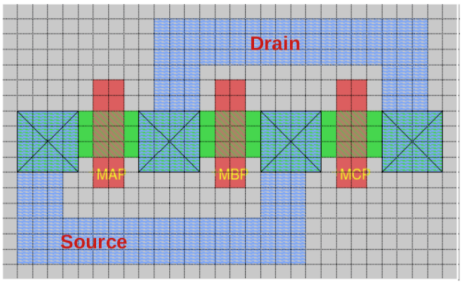
Improves:

- Speed: as delay is reduced
- More transistors can be packed, More functionality in the for the same area
- Reliability and yield: as the number of contacts are reduced

* MAS, MBS, MCS - A,B and C transistors in series(S) respectively
* Interconnect S and D with diffusion itself instead of using contacts and metal.
* Adv: reducing contact resistance and parasitic capacitance, leakage red, faster & compact.

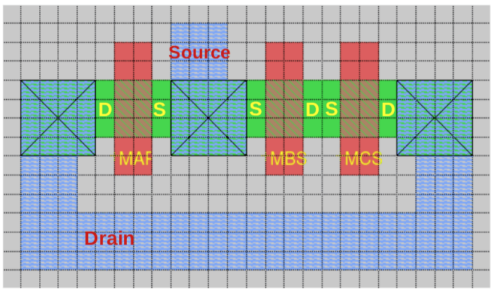
Off resistance remains same. **On resistance** (reduced) -
Originally = [contact res (from blue X) + diffusion res(green) + gate res + channel res + diffusion res(green) +contact res (from blue X)] x 3 + [metal resistance(light blue)] x 2
Now = [contact res]x2 + [diffusion res + gate and channel res]x3 -> reduced
Optimized both in terms of area and speed

Example of an optimized layout of 3 MOSFETs in parallel



Example of an optimized layout of MOSFETs in series – parallel combination.

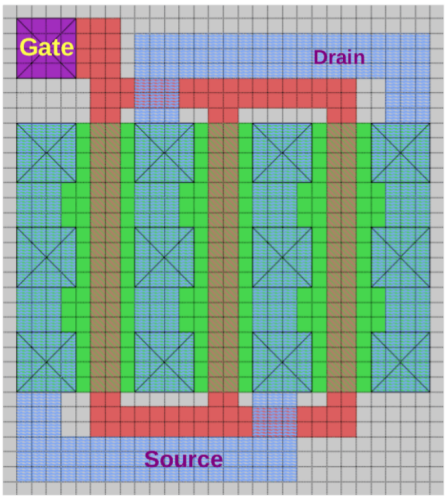
MAP is in parallel with a series combination of MBS and MCS



* Connect source of 3 transistors together. same goes for drain. Use single diffusion layer. Diffusion is set such that diff res is minimal.

Say, you want a large transistor of 30 micron (very large) height. Divide it into 3 transistor sof 10 micron width in parallel with a single gate contact (can have multiple poly also) ->

Example of a large transistor – A very wide MOSFET
Folded or Multi-fingered MOSFET



Folded, multi-fingered (3 fingers) MOSFET

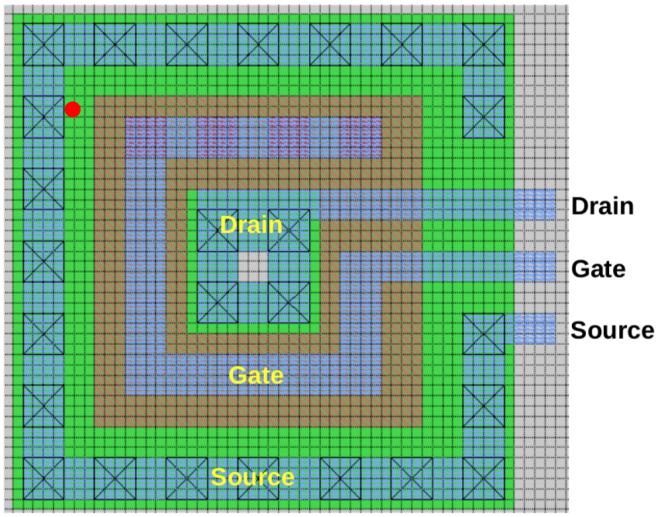
Use as many contacts as possible for Diffusion as well as Poly. Connect all contacts with appropriate metal, so that the Signal distribution and current density is uniform across the device

The figure shows only one Poly (Gate) contact

Why multiple contacts? DC current flows on the surface and not the core. Single contact - surface area would only be along the perimeter of a large rectangle.Multiple contacts (blue X) would result in larger surface area - sum of all the perimeters of box X. Contact res will also be reduced since the contacts are in parallel.

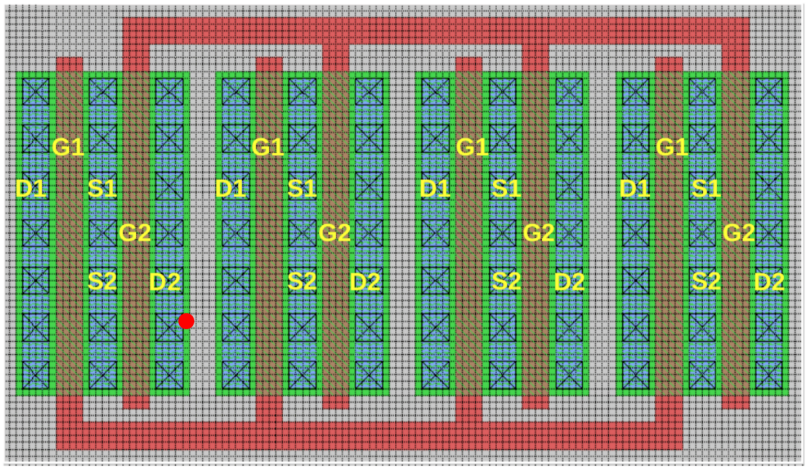
Doughnut shaped mosfet notes :
Very large transistor will also have high drain/ output capacitance. (in prev example, sum of 6 boxes area)
How to reduce? Instead of folding like in prev fig, fold it in doughnut shaped mosfet fashion. Now drain area - small, source may be large.

Doughnut shaped MOSFET



- Very wide transistor implemented
  - Width is the perimeter of the Poly Gate (Perimeter of the rectangular Poly)
  - Length of the Gate is the distance between Drain and the Source
- Very small Drain area
- Drain (Output node) Capacitance is very small

Multi-fingered (Four fingers) Matched NMOSFETs with Common Source (or Drain)



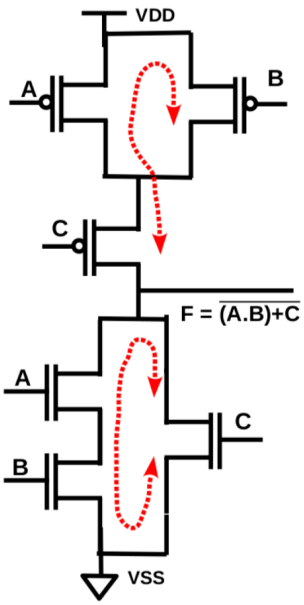
Only the source (or Drain) is common

Assumed that the nodes sharing the same label are connected with metal (M1)

Gates sharing the Same label are connected with Poly.

Performance will be matched.

LEC4: Single line diffusion + Euler's path



Euler's path:

- Each electrical node is considered to be a vertex/node in the graph
- Each MOSFETs Drain - Source (ie. the MOSFET or the input literal) is considered to be an edge in the graph;
- One graph for the PUN and one for the PDN

Procedure of obtaining Euler's Path:

- Start at any edge, traverse through the graph without without passing though the same edge twice. Note down the path.
- Collect all possible paths and list them
- Repeat the process for both the PUN and PDN graphs
- Identify common paths as a pair with one from PDN and other from PUN.
- These are the common Euler's Path.

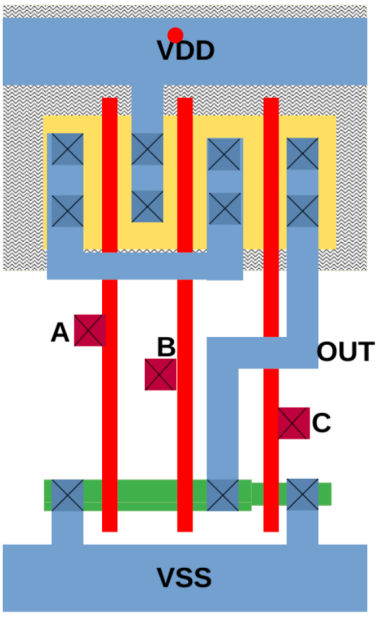
Example circuit $F = \overline{(A.B)+C}$:

Euler's Paths of PUN: ABC, BAC, CAB and CBA

Euler's Paths of PDN: ABC, BAC, CAB, CBA, BCA and ACB

Common Euler's Paths between PDN and PUN: ABC, BAC, CAB and CBA

Single Line Diffusion Layout of $F = \overline{(A.B)+C}$ based on Euler's Path

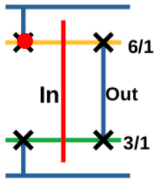


- This layout uses the order A-B-C; could use any of the ordered common path for PUN and PDN
- Note that there are no breaks in the Diffusions (Both P and N) – Single Line Diffusion
- Single vertical line of Poly for each input literal
- Note the width of the MOSFETs
  - NMOSFET corresponding to variable C

Note: in PDN - A and B switched in prev diagram to obtain this A-B-C version. A's source is connected to Vss. in Vdd - add NWell contact and similarly Vss also add contact to p.

Stick Diagram

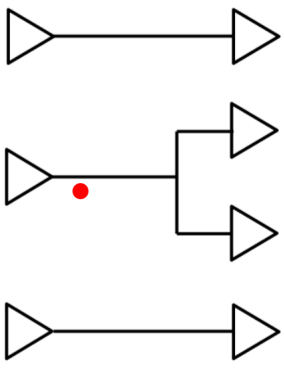
- A way of representing the layout.
- Each rectangular object in the layout is represented by a line (or stick) or a cross mark or simple box.
- Colour or line style or hash (stippled) patterns represent the material or the layer
- Unlike layout where the objects are drawn to scale, stick diagrams are not drawn to scale.
- Objects in a stick diagram depict the relative position of the objects and not to scale.
- Stick diagram may be labelled depicting the object meta-data like width, length, literal name of the MOSFET etc.



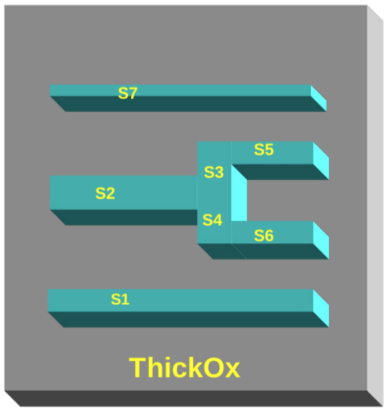
Simple stick diagram of a CMOS Inverter with labels "In", "Out" and dimensions of the NMOS and the PMOS.

THE WIRE

LEC 1:



The gates, their position and their interconnection, dimension etc. in the figure are just indicative and not specific.



- Interconnects can be partitioned into segments of identical widths
- S1 to S7 are segments of Metal (M1) of similar width
- Network of these segments can be formed
- Resistance and capacitance (to ground node) of each segment can be calculated
- Network of these resistances and capacitances can be formed
- Inductive effect is negligible
- Wires: capacitive effect dominates.
- Switch resistance and capacitance can also be included in this network

Impact of Interconnect parasitics:

- **Reduce reliability**
- **Affect performance and power consumption**
  - **Signal coupling – cross talk**
  - **Delay - Aggressor net on Victim net**

Types of parasitics and their effect (depends on the operational range of frequency)

- **Capacitive - Most dominant**
- **Resistive – Less dominant**
- **Inductive – Negligible**

Distribution of interconnects: Number of interconnects vs. their length

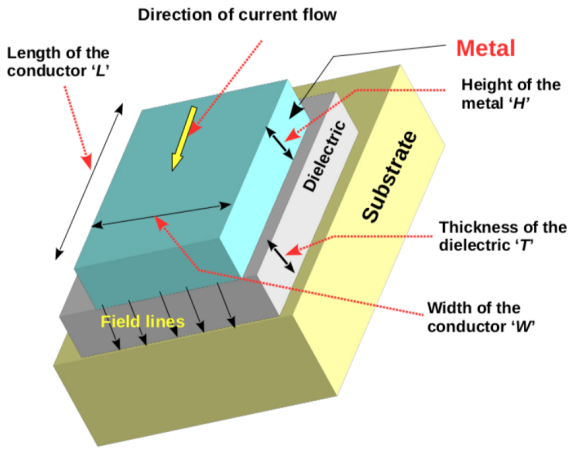
- **Very short – Huge number – Intra cell and inter cell between neighboring cells**
- **Short – Large number – Inter cell routes**
- **Very Long – Small number – Buses, Clock and Power distribution network – Global Interconnects**

Modeling of parasitics:

- **Parallel plate Capacitance between the neighboring wires and wires in the upper and lower metal layers can also be computed and included in the network**
- **Resistance and capacitance of the vias and contacts can also be modeled and included in the network**
- **Resistance and capacitance of the diffusion can also be modeled and included in the network**

* Aggressor is connected to a strong driver.

Parallel plate (Area) capacitance



Field lines shown are indicative, they exist everywhere below the metal conductor.

Parallel plate (Area) capacitance is given by:

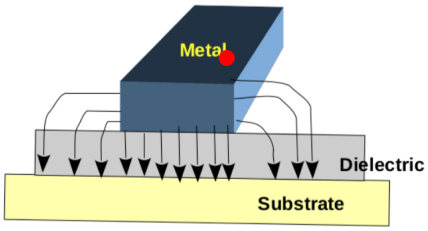
$C_{pp} = \epsilon WL / T$; Where

- **ϵ is the permittivity of the dielectric;**
- **W is the width of the conductor**
- **L is the length of the conductor**
- **T is the thickness of the dielectric**

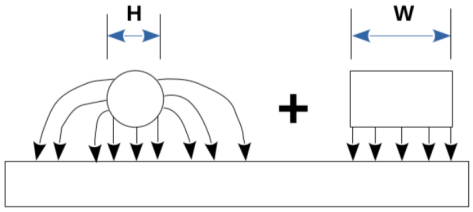
For computing parallel plate capacitance consider only the orthogonal field lines between two plates

* WL is the common overlap area between the 2 conductors.

Capacitance computation



Curved lines indicate the fringe field lines



Field lines that are not orthogonal to the metal and the substrate result in fringe capacitance.

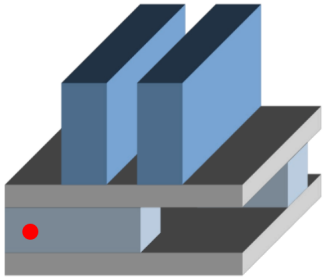
Total Capacitance = Sum of Capacitance due to Parallel plate and Fringe field.

$$C_{wire} = C_{pp} + C_{fringe}$$

$$C_{wire} = (\epsilon wL/T) + (2\pi\epsilon/\log(T/H))$$

Where

- Є is the permittivity of the dielectric;
- W is the width of the conductor
- H is the height (thickness) of the conductor
- w = W - H/2
- L is the length of the conductor
- T is the thickness of the dielectric



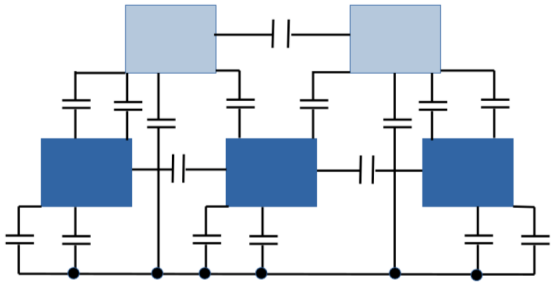
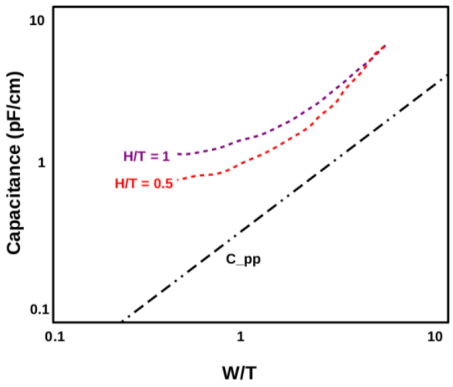


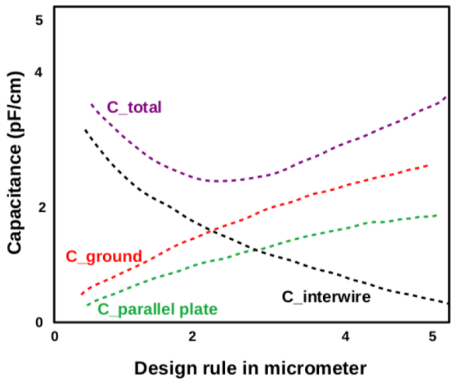
Illustration of parallel plate and fringe capacitance
Intra layer and interlayer capacitance

- Top metals have more cross sectional area as their height (thickness ie. top metals are taller) is more than the bottom metals and the chances are that top metals are also longer in lengths.
- If two top metals run parallel then the parallel plate capacitance (C_interwire) will be more between two wires since they are tall. Hence can result in more interaction between the signals on these wires resulting in interference (cross talk)

* Note: as seen in the figure, metals in alternate layers run in opposite ways (H & V)
* Fig 2: Between two rects of same color - interwire capacitance; between two diff colors - since there is no overlap, fringe capacitance; each of those colors will have capacitance to the ground; between dark blue and bottom wire - parallel plate + fringe capacitance.

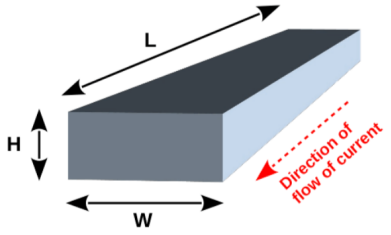


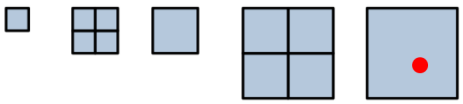
Capacitance of interconnect wire as a function of W/T (indicative only)



Interconnect Capacitance as a function of Design Rule (indicative only)

LEC2:





Resistances of all these squares are same.

- Keep the wire length as short as possible to reduce the resistance.
  - To reduce resistance, may increase the width but also results in increase in capacitance.

Wire Resistance:

- $R = \rho L / A$
  - where A is the cross sectional area of the wire; $A = HW$
- For a given technology process and a specific metal layer, ρ/H is a constant as the ρ of the material (metal is known) is a constant and H (the height of the conductor for a given metal layer) is also a constant.
- Let $\rho/H = R\square$; a constant for a given layer.
- R can be written as: $R = R\square\, L / W$
  - where $R\square$ is called as the Resistance per square; which is a constant for a given layer.
  - $R\square$ is called as the sheet resistance of that layer. The relation can be used for diffusion, Poly, or Metal layers.
  - The value of the sheet resistance can be found in the technology file. Eg. scmos.tech
- Resistance is proportional to length of the wire and inversely proportional to the width of the wire

Typical values of sheet resistance

Material	Sheet resistance Ω / □
N^- or P^- well diffusion	1000 to 1500
N^+, P^+ diffusion	50 to 150
N^+, P^+ diffusion with silicide	3 to 5
N^+, P^+ Polysilicon	150 to 200
N^+, P^+ Polysilicon with silicide	4 to 5
Aluminum	0.0 to 0.1

* Al will give smaller resistance for long distance compared to silicide. However, using Al -> will require additional metal contacts.

	Field	Active	Poly	Al 1	Al 2	Al 3	Al 4
Poly (area)	88						
Poly (fringe)	54						
Al 1 (area)	30	41	57				
Al 1 (fringe)	40	47	54				
Al 2 (area)	13	15	17	36			
Al 2 (fringe)	25	27	29	45			
Al 3 (area)	8.9	9.4	10	15	41		
Al 3 (fringe)	18	19	20	27	49		
Al 4 (area)	6.5	6.8	7	8.9	15	35	
Al 4 (fringe)	14	15	15	18	27	45	
Al 5 (area)	5.2	5.4	5.4	6.6	9.1	14	38
Al 5 (fringe)	12	12	12	14	19	27	52

Layer	Poly	Al 1	Al 2	Al 3	Al 4	Al 5
Capacitance	40	95	85	85	85	115

Typical Parasitic Capacitance

Table on the left:

Inter-layer Area and Fringe capacitance for a typical 250nm CMOS technology.

- **Rows represent top plate of the capacitor**
- **Columns represent the bottom plate of the capacitor**
- **Area capacitance is in aF/sq. Micron**
- **Fringe capacitance is in aF/micron**

Table on the left:

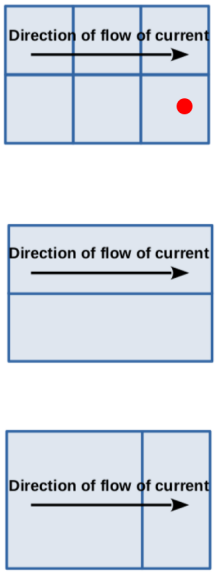
Intra-layer capacitance for a typical 250nm CMOS technology, with minimal spacing between wires .

Capacitance is in aF/micron

* area cap and fringe cap (due to edge) have **different** units
* aF => a : 10^-18 femto: 10^-15

* Al 5 (table 2) - 2 large wires in the upper layers (effective area is more due to height) -> larger intra-layer cap 

* separation between wires is also varying across Al1,2,3,4,5-> verify from tech file



Resistance calculation example:

Given: A strip of material dimension 3 x 2 units

Material can be Metal, Poly, Diffusion etc whose R□ is known.

Unit can be any, micron, nm etc. (as the resistance is expressed in per square)

Approach 1:

- **Consider the strip to be TWO rows of THREE squares each. These 3 squares can be considered to be in series, and two such strips to be in parallel (as shown in the figure in the middle).**
- **Resistance of each strip = 3 x *R*□**
- **Resistance of two such strips in parallel = *(*3 x *R*□ *) / 2***
- **Therefore effective resistance R = 1.5 x *R*□**

Approach 2:

- **Consider the given material to be consisting of TWO tiles as shown in the figure at the bottom. It can be considered to be two tiles in series, the first tile is a square and the second tile is half a square.**
- **Effective resistance *R* = *R*□ + 0.5*R*□ = 1.5*R*□**

Area capacitance calculation example:

Given: A strip of material dimension 3 x 2 microns

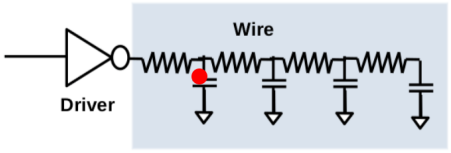
Material can be Metal, Poly, Diffusion etc whose *C* per sq. micron in farads is known.

- **Total area of the material = 3 x 2 = 6 sq. microns**
- **Total area capacitance is given by: Area in sq. microns x *C* per sq. micron**
- **Effective are capacitance = 6*C* farads**

The Distributed model of the wire

Distributed model:

- **Wires have resistive, capacitive and inductive parasitic components.**
  - **Distributed over the entire length; not lumped at a single point.**
- **Beyond High frequency range (> 30 MHz) the inductive components are dominant.**
- **At low and medium frequencies (< 3 MHz) capacitive components are dominant.**
- **Resistive component is to be accounted for at all frequency ranges.**
- **Should be described by partial differential equations**

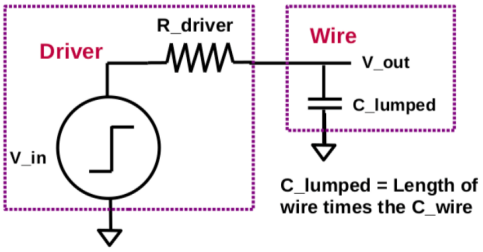


* complex in terms of computation (above)

The Lumped Model of the wire

Lumped model:

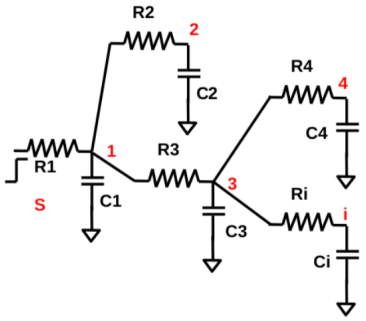
- **Takes into account only the component that is dominant over the operational range of frequency, and ignores the other parasitic components of that wire.**
  - **Lumps the dominant component**
- **At low and medium frequencies, capacitive component of the wire dominates as the wire resistance is very small compared to the driver resistance.**
- **Can be described by a ordinary differential equation**



- **A pessimistic model**
- **Inaccurate for long interconnects**

LEC 3: Elmore Delay

Lumped RC model: Elmore delay



All capacitances are modeled to be between the respective node and the reference (ground) node.

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

The dominant time constant $\tau_{Di} = \Sigma_{k=1}^{N} C_k R_{ik}$

Where

$R_{i1} = R1$

$R_{i2} = R1$

$R_{i3} = R1 + R3$

$R_{i4} = R1 + R3$

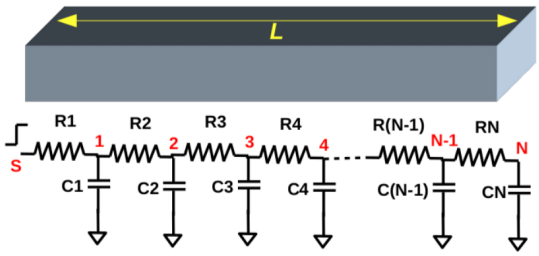
$R_{ii} = R1 + R3 + Ri$

Therefore

$\tau_{Di} = R1 \times C1 + R1 \times C2 + (R1+R3) \times C3 + (R1+R3) \times C4 + (R1+R3+Ri) \times Ci$

- ***S*** **is the driver node and all other nodes are fan-out nodes of the driver node – ie. driven nodes.**
- **τ_{Di} at each driven node will have a different value, indicating that the signal arrival time at these nodes are different.**

The Elmore Delay of a RC Chain (or a segment of a wire)



Let the wire be of uniform width and thickness with a certain *R*□.

Let the resistance of the wire be *r* per unit length.

Let the capacitance of the wire be *c* per unit length.

Let the length of the wire be *L*.

Consider the wire to be partitioned into a very large number of identical segments *N*.

- **Length of each segment is *L/N***
- **Resistance of each segment is *rL/N*.**
- **Capacitance of each segment is *cL/N*.**

If we substitute the values into the expression of Elmore model (There are no fan-outs in this case, it is a simple continuous wire and the expression is very simple);

we have τ_{Di} = *C1 R1 + C2 (R1 + R2) + C3 (R1 + R2 + R3) + ……. + Ci (R1 + R2 + …..+ Ri)*

Substituting the values of resistance and capacitance of each segment ,

τ_{DN} = $(L/N)^2$ (*rc + 2rc + 3rc+ ……+ Nrc*) = (rcL^2) *N* (*N*+1) / $2N^2$.

τ_{DN} = *RC* (*N* + 1) / 2*N*; where *R* = *rL* and *C* = *cL*.

For large *N* or infinitesimal length of segment, the dominant delay can be written as

τ_{DN} = RC / 2; Actual delay as per distributed model is HALF of the lumped delay model.

Lumped delay model is pessimistic. Inaccurate for long interconnects

τ_{DN} = rcL^2 / 2; Delay is a quadratic function of length

We have $\tau_{DN} = rcL^2 / 2$, Delay is a quadratic function of length

- To reduce delay, reduce the length of the wire

Backend designers always face the *Timing closure* problem.

Possible solution includes:

- Reduce the length of the longest path by rerouting the wire.
- May not be possible if the routing has already been done or no other route is possible etc.
- Break the path into smaller segments and introduce buffers between the segments.
- Introduce buffers wherever there is a large fan-out or a long wire

Example:

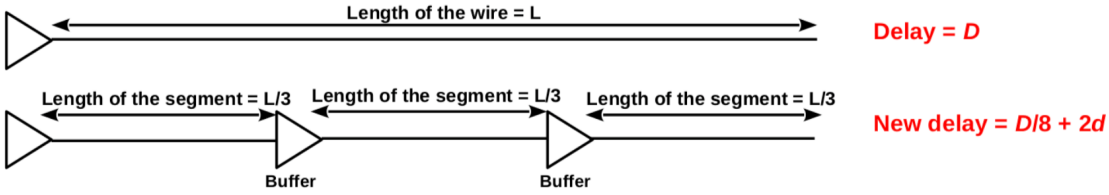
Consider a wire of length *L*, with a delay of *D*.

Let us partition it into *n* (3) segments and introduce *n*-1 (2) buffers. Let the buffer delay be *d*.

Total delay after partitioning the wire into n (3) segments and introducing n-1 (2) buffers is the sum of individual wire segment delay and the delay of the buffers

$= D(1/n)^2 + (n-1)d = D(1/3)^2 + 2d = D/8 + 2d$

With the knowledge of original delay *D*, desired delay, and the buffer delay d, choose *n* appropriately

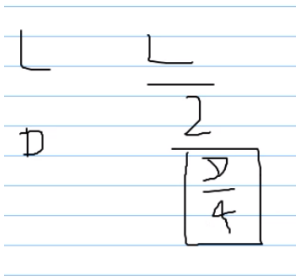


/* if a wire of length L gives us a delay of D, then if we make it into 3 wires of length L/3 then shouldn't we get a delay of D/3, i.e 3*(D/9), along with additional 2d delay from the buffers ?

But in class I think it was mentioned as D/8 + 2d

It's D/9 + 2d

The argument developed there was that the **delay is proportional to the square of the length**. (Since R and C both are reducing by a factor of L)
So if I reduce the length by half, delay reduces to quarter.



*/

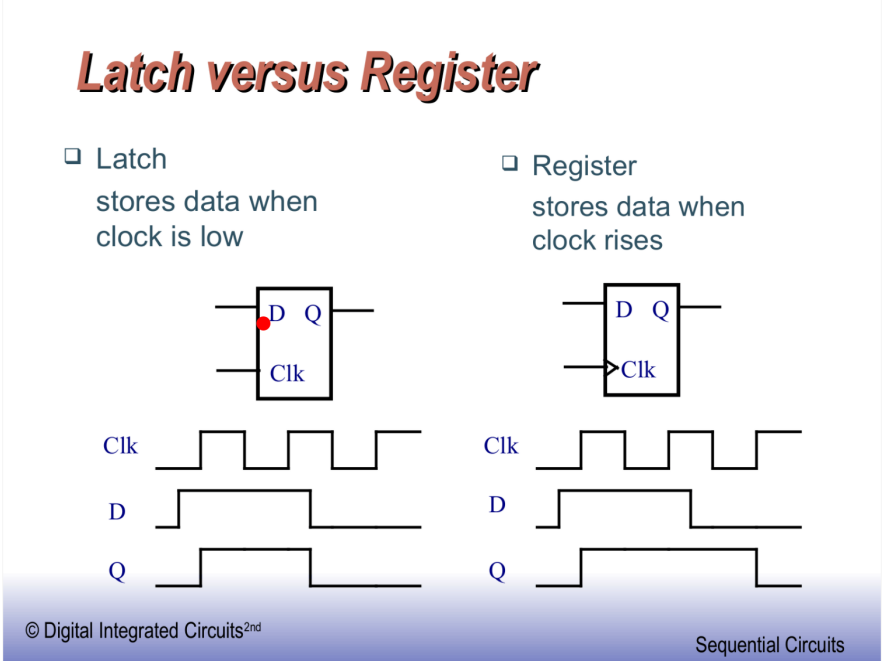
SEQUENTIAL CIRCUITS

LEC 1:

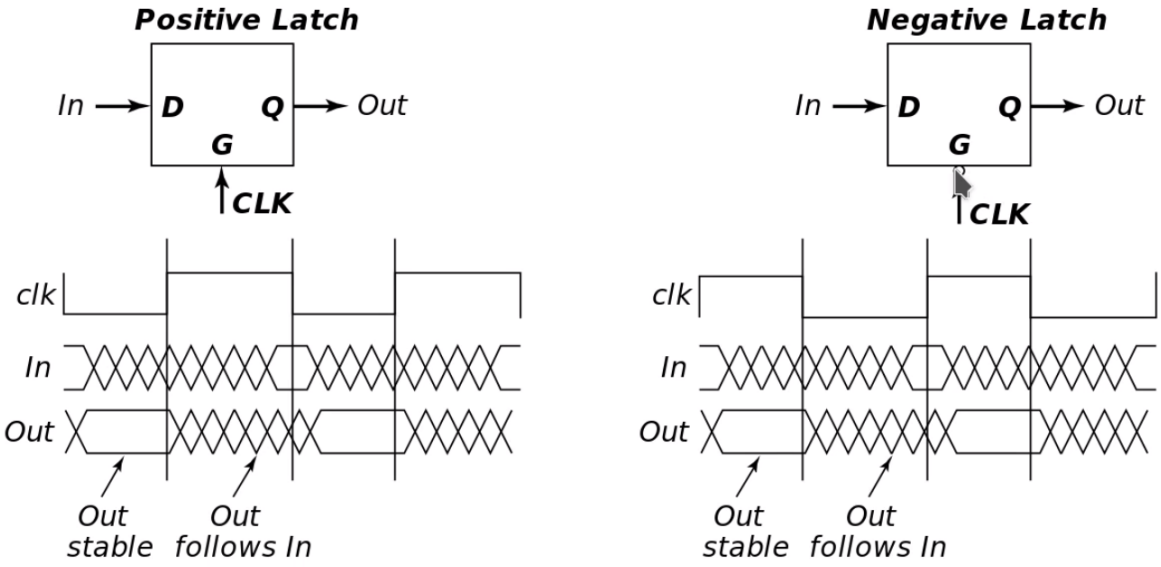
2 storage mechanisms - positive feedback; charge based

Latch - level sensitive

Register / Flip Flops - edge triggered



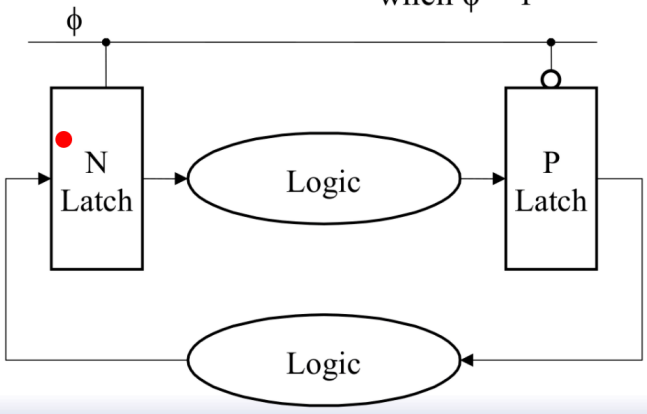
Latches



Latch-Based Design

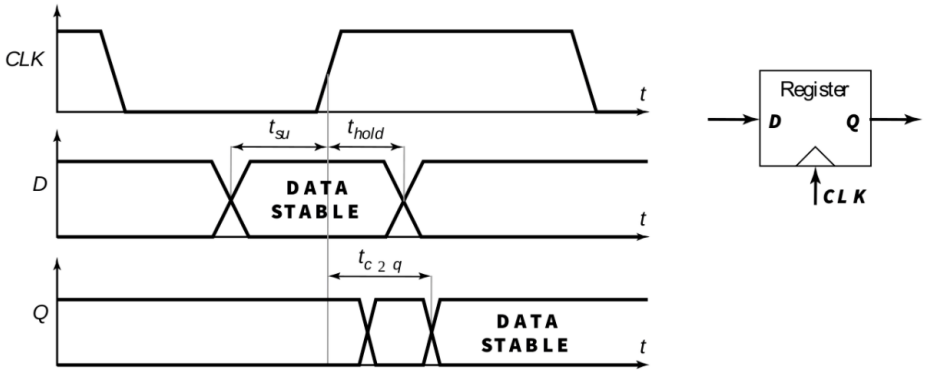
• N latch is transparent when $\phi = 0$

• P latch is transparent when $\phi = 1$



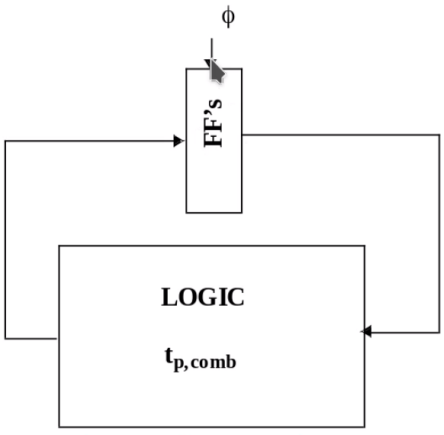
* transparent phase - D=Q

Timing Definitions



* **Tsetup** : flip flop requires the data to be stable for some time prior to the active edge of the clock, that duration is setup time.
* **Thold** : data should continue to remain stable further for some duration after active edge of clock
* If tsetup or thold is violated -> the data that is locked will not be locked -> metastable data
* Clock to Q delay: after the active edge, it takes some time(**Tc2q**) for the change in D to reflect in Q in a stable manner
* in transparent time, after how long D = Q in latch - **Td2q**

Maximum Clock Frequency

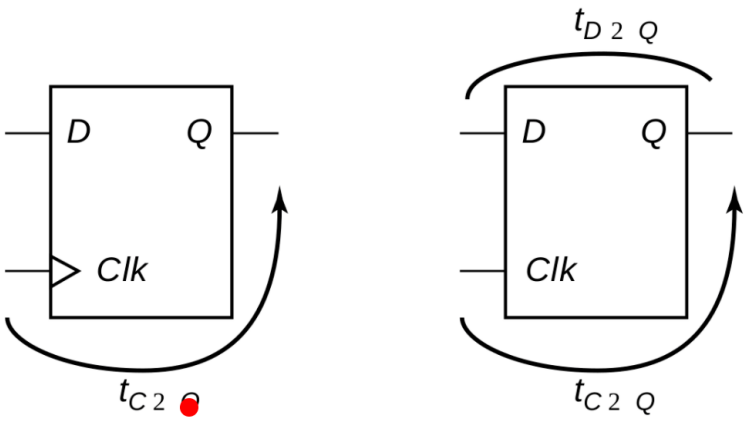


$t_{clk\text{-}Q} + t_{p,comb} + t_{setup} = T$

Also:
$t_{cdreg} + t_{cdlogic} > t_{hold}$

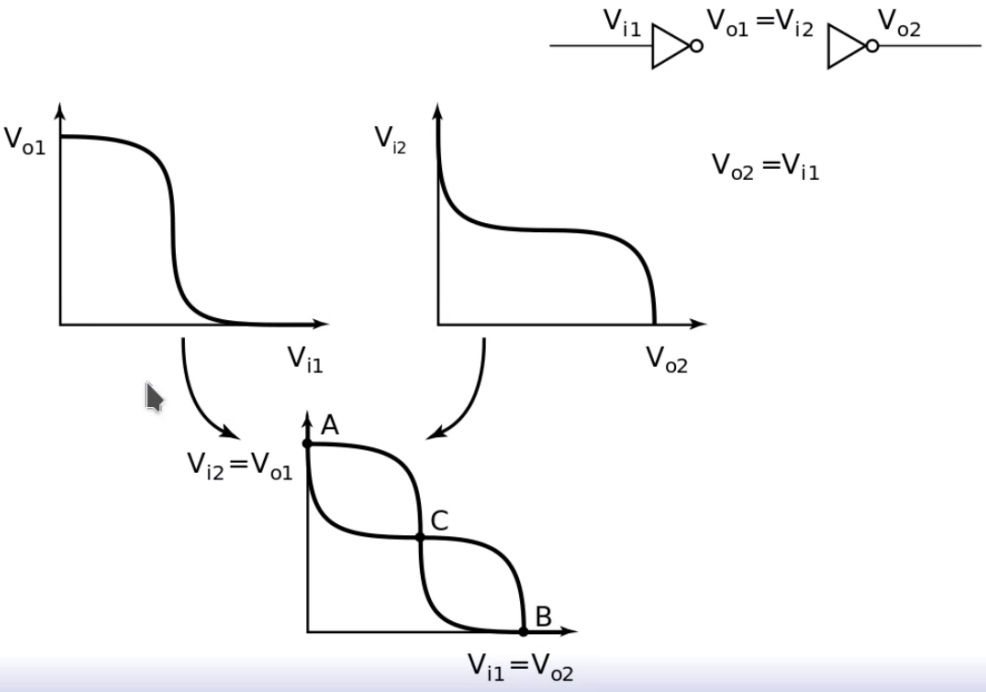
t_{cd}: contamination delay = minimum delay

Characterizing Timing



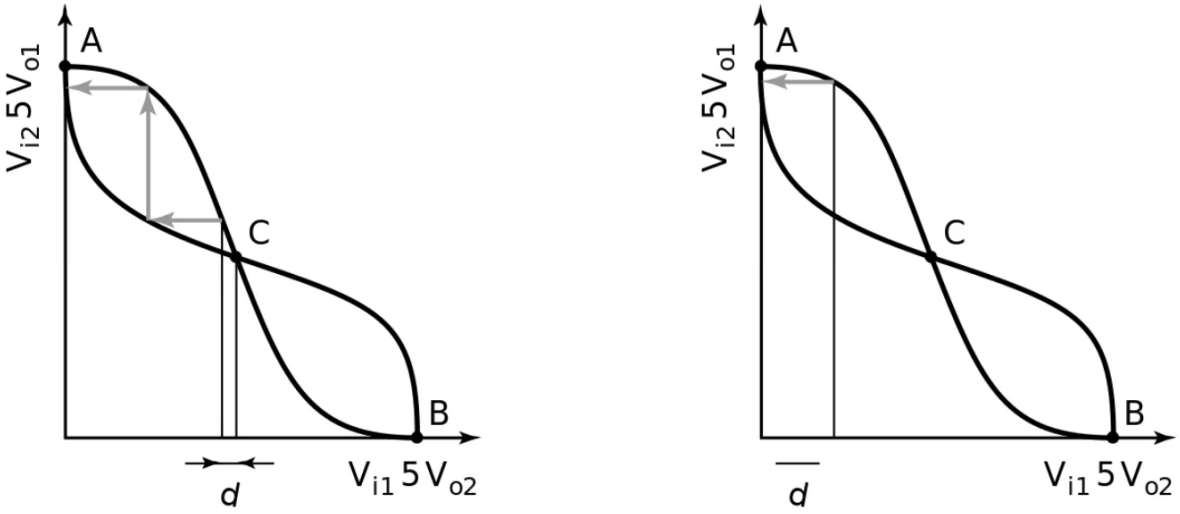
Register Latch

Positive Feedback: Bi-Stability



Latch=2 inverters connected back to back

Meta-Stability



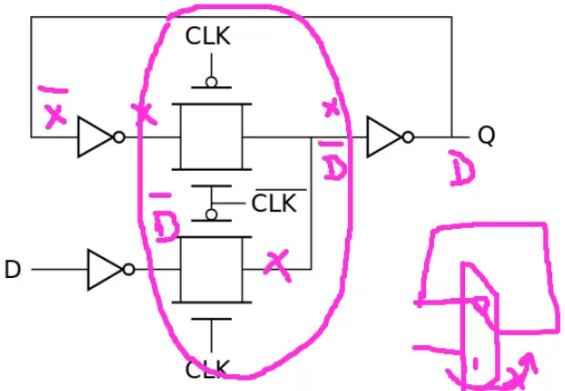
Gain should be larger than 1 in the transition region

Higher the gain, quicker the transition.

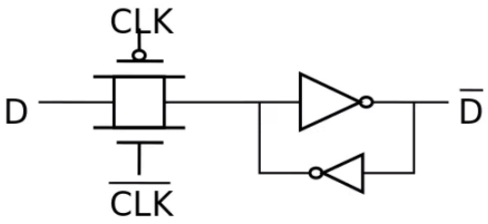
LEC2:

Writing into a Static Latch

Use the clock as a decoupling signal,
that distinguishes between the transparent and opaque states



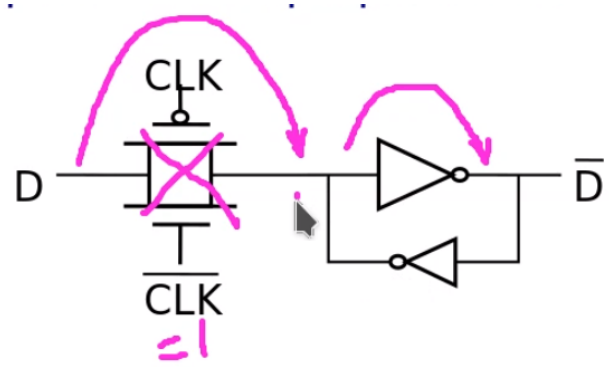
Converting into a MUX



Forcing the state
(can implement as NMOS-only)

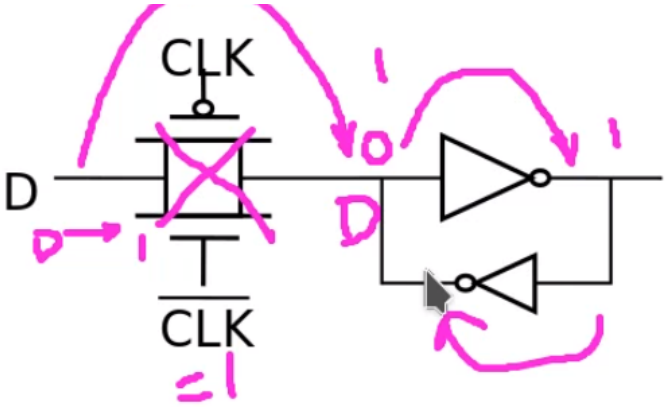
Figure (Left) :
Converting mux to latch - Pass transistor logic doesn't have any drive strength. That's why the output of the mux is fed to inverters.
Clock = 0: hold mode
Clock = 1: transparent mode

Figure (Right):



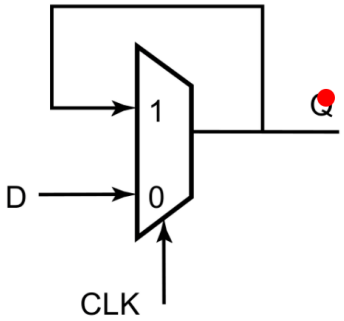
Clk = 1 : Cutoff + D reinforces itself

To avoid fight between 0 and 1 (driven from transistor logic) - the below inverter is weak and above is stronger



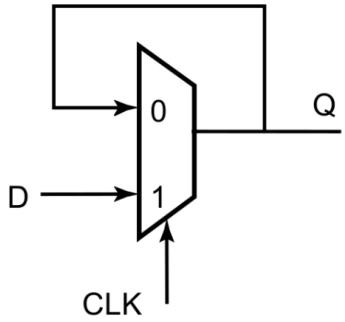
Mux-Based Latches

Negative latch
(transparent when CLK= 0)



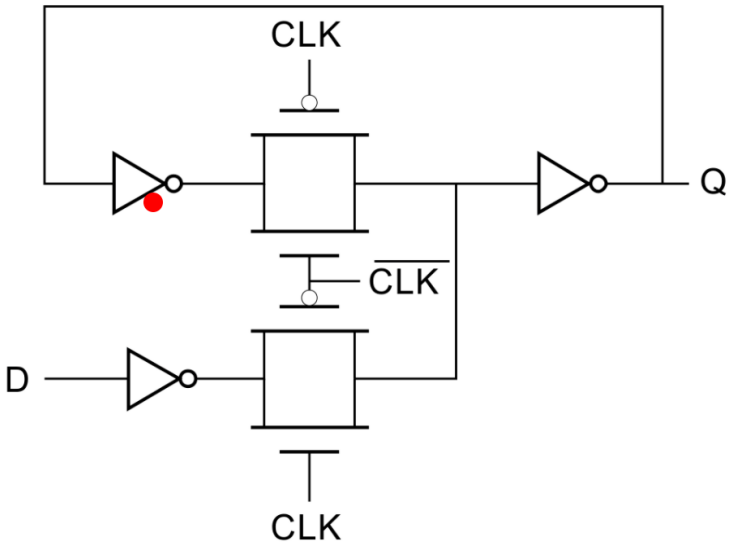
$$Q = \overline{Clk} \cdot Q + Clk \cdot In$$

Positive latch
(transparent when CLK= 1)



$$Q = Clk \cdot Q + \overline{Clk} \cdot In$$

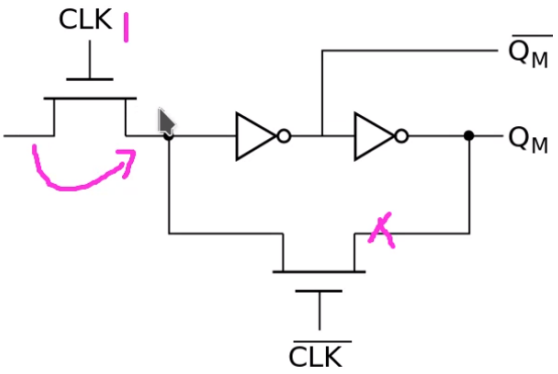
Mux-Based Latch

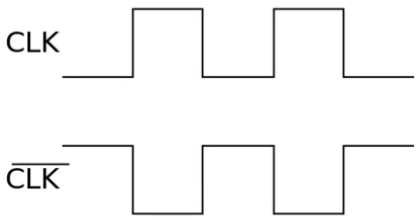


Inverters added to the design for drive strength.

Improved design with lesser number of transistors -

Mux-Based Latch

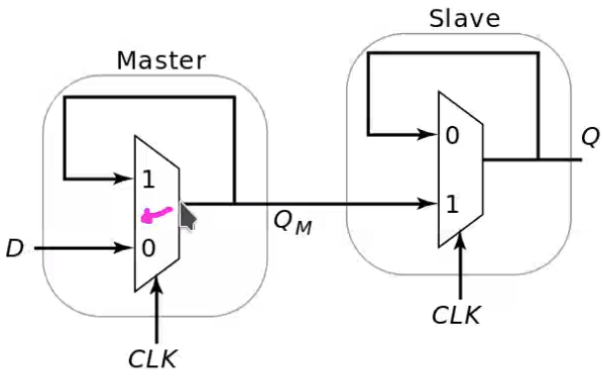


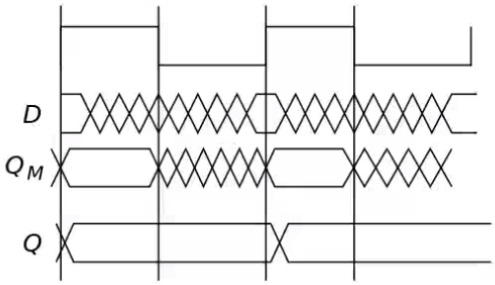


NMOS only

Non-overlapping clocks

Master-Slave (Edge-Triggered) Register





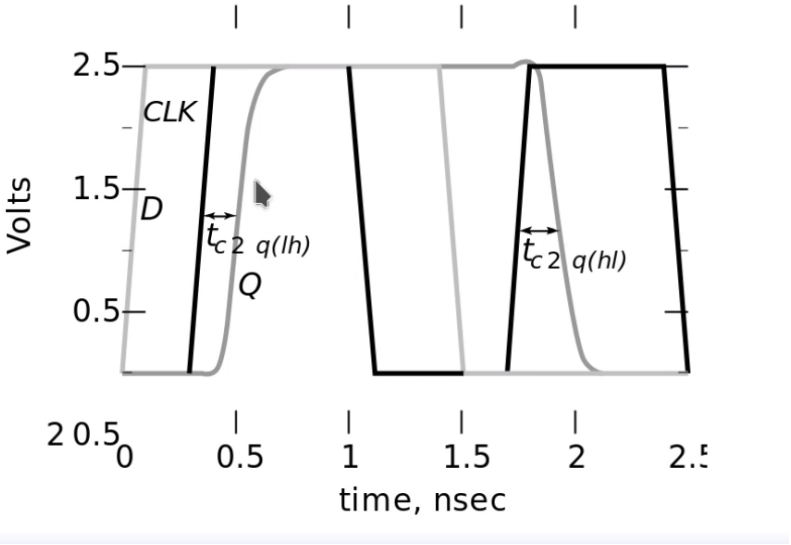
Two opposite latches trigger on edge
Also called master-slave latch pair

When M is hold mode, S in transparent and vice - versa.

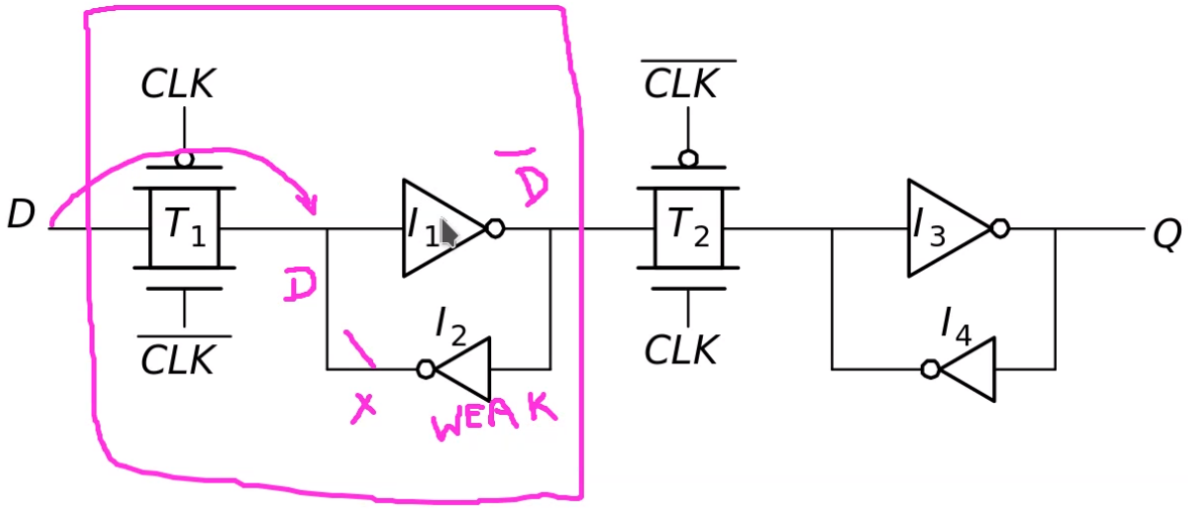
Transmission gate and inverter based M and S (Figure incomplete)

<check figure from moodle>

Clk-Q Delay

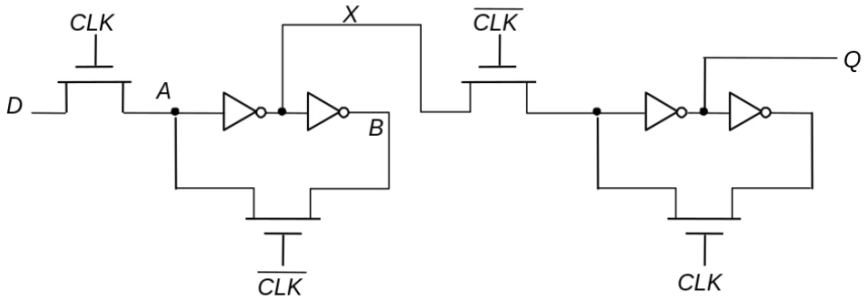


Reduced Clock Load Master-Slave Register

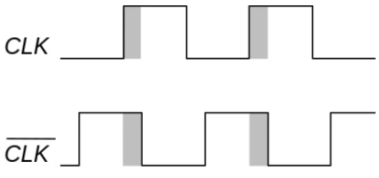


Weaker inverter occupies more area.

Avoiding Clock Overlap



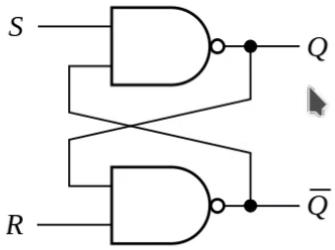
(a) Schematic diagram



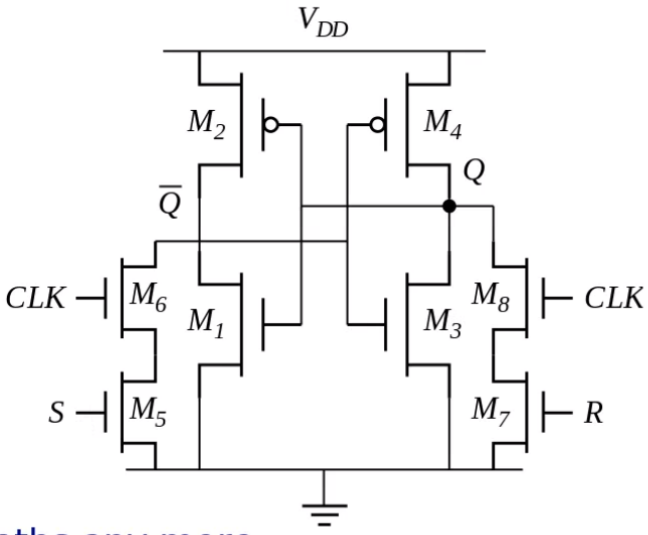
(b) Overlapping clock pairs

Cross-Coupled NAND

Cross-coupled NANDs



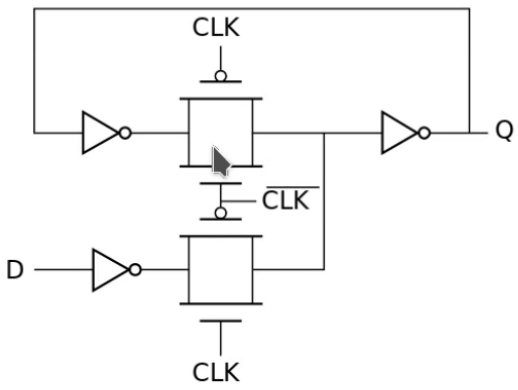
Added clock



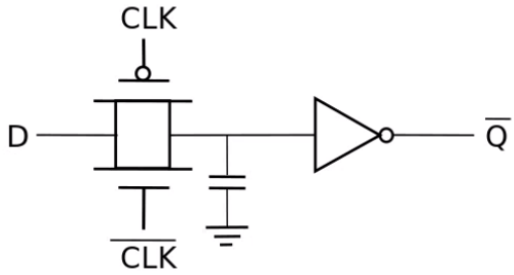
This is not used in datapaths any more,
but is a basic building memory cell

Storage Mechanisms

Static



Dynamic (charge-based)



Making a Dynamic Latch Pseudo-Static