L3. Digital Noise

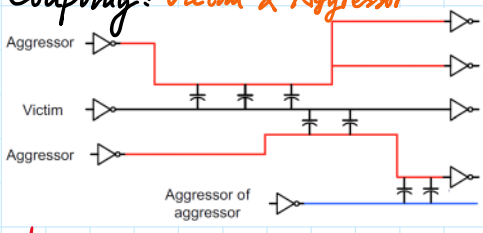
2023年1月3日 19:38

Noise Type:

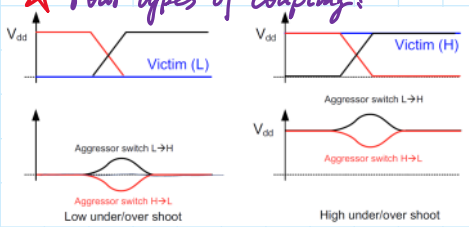
① Digital: Deterministic, Repeatable (usually large)

② Analog: Random, Small.

Coupling: Victim & Aggressor

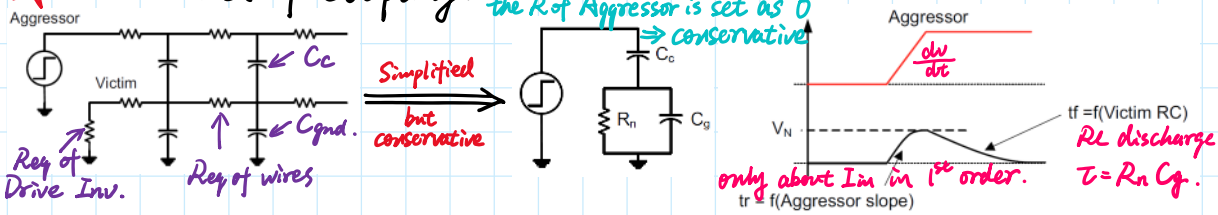


Four types of Coupling:

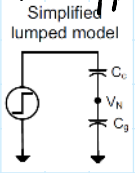


RL Model of Coupling:

the R of Aggressor is set as 0 ⇒ conservative



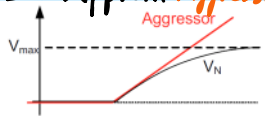
1st Approx: Ignore Rn ⇒ conservative

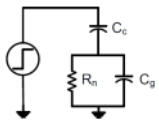


$\Delta V_N = \Delta V_a \cdot \frac{C_c}{C_g + C_i}$ charge sharing.

Noise Amp. Aggressor Amp.

2nd Approx: Aggressor Never Sat.





$V_{N,Sat.} = R_N \cdot C_i \frac{dV_a}{dt}$

always have a current until C_g is charged to $V_N = I_{in}/R \Rightarrow I_{in} = I_R$, No more ΔV at C_g.

Coupling Parameter:

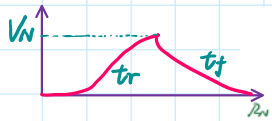
Narrow Pulse ⇒ better

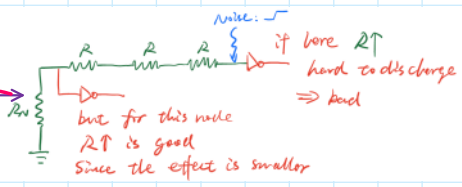
	V_N	t_r	t_f
Victim driver ↑	↓	same	↓
Aggressor ↓	↓	↑	same
C_g ↑	↓	same	↑
R_w ↓	Depends on topology		
C_c ↓	↓	same	↓

↑ Best Solution

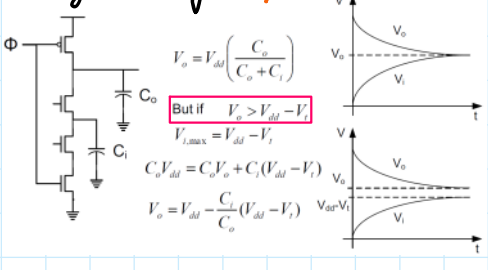
close to Noise → Small; Far → Large.

- Victim driver stronger makes it a strong aggressor
- Rise time depends on aggressor slope
- Fall time depends on victim's RC

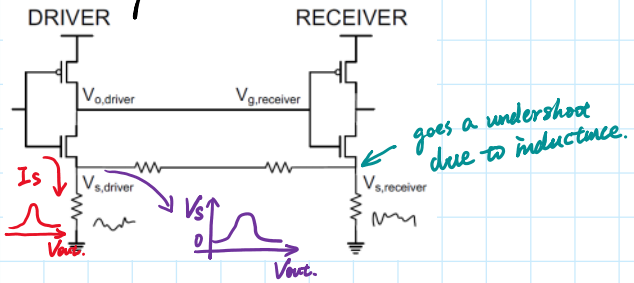




Charge Sharing: C/Ctotal or related to Vth.

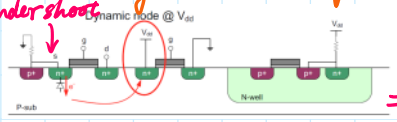


IR-Drop:



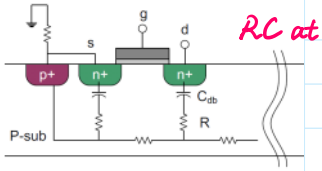
Substrate Coupling:

① Minority Carrier injection



Carriers are injected by pn junction
May hit the dynamic net.
⇒ Protect by guard ring.

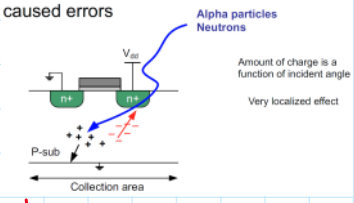
② Ground Bounce:

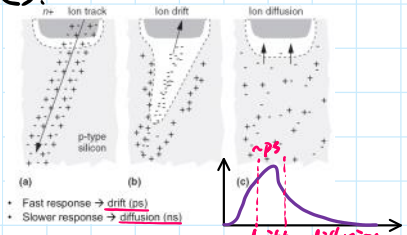


RC at the sub.

Soft Error Rate (SER):

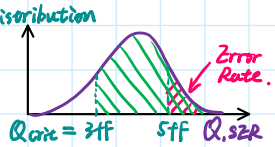
Non-deterministic generated radiation caused errors





Critical Parameter

1. Critical charge: Min charge stored on node
2. Collection Area: Area sensitive to SER



FIT: Number of Failures in 10^9 hours

Particals:

- Alpha Particles:
  - Lead in chips and old packaging, materials:
    - Flux: 0.01 events/cm²h
  - Fix with cleaner processes (Clean lead, new materials)
    - Flux: 20x reduction with new materials
  - Background radiation:
    - Flux: 0.005 events/cm²h
    - Can be shielded (coating of the package with special materials)
  - Charge = 10-20 fC

- Neutrons:
  - Background cosmic radiation:
    - Flux: 0.005 events/cm²h
  - Charge: 100-200 fC ↖ Larger charge.
  - Cannot be shielded

More critical.

SER for Different Circuit.

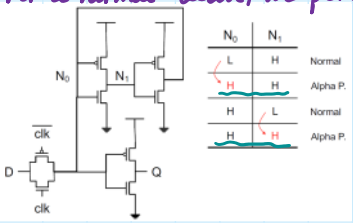
- With scaling less charge is stored in memory cells → More sensitive to SER
- More cells/unit area → SER collection are is decreasing
- DRAM scaling → 2 effects cancel
- SRAM scaling -> holding steady
- But: more multi-failure -> interleaving

A small tech node:
Less Qcrit, but smaller Area.

SRAM has error control code, usually detect 2-bit, fix 1-bit.
To avoid 2-bit flip, the bits of same word are placed separately.
e.g. W_1B_0 W_2B_0 W_3B_0 W_1B_1

Hardened Latch: Avoid SER

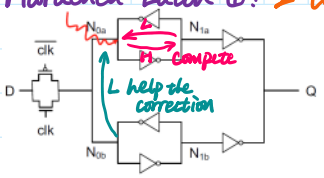
For a normal latch, the pervious state before soft error is unknow.



both High for N_1, N_0 — Soft error.
but 2 possiable pervious states
⇒ Cannot Recover.

Hardened Latch ①: 2 Latches, Data backup

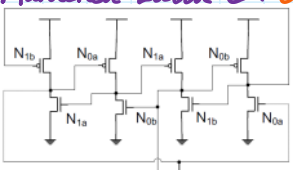
Hardened Latch ①: 2 Latches, Data backup

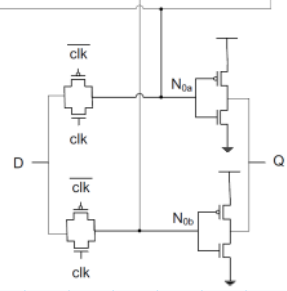


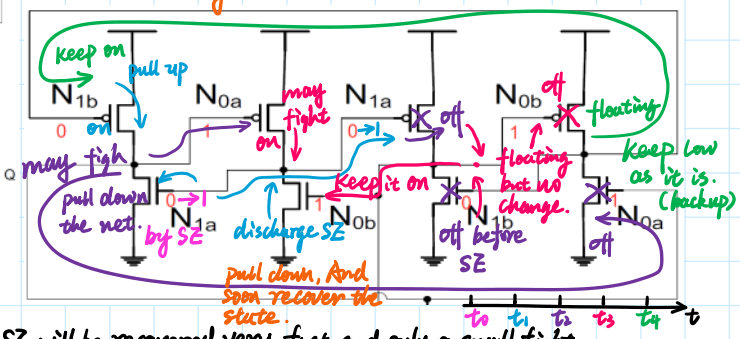
	N_{0a}	N_{1a}	N_{0b}	N_{1b}
Normal	L	H	L	H
	H	L	H	L
Alpha P.	L (Error)	L	H	L (Backup Data)

Hardened Latch ②: Cause floating nets for a single change

if the other side keeps the same, it will be soonly corrected since the floating net cannot make any action to the other side.







For small charges: the SE will be recovered very fast and only a small fight

But leakage may dominate if charge is too much — may fail.

★ Noise Analysis — Global View

★ Latch Noise: Data Noise

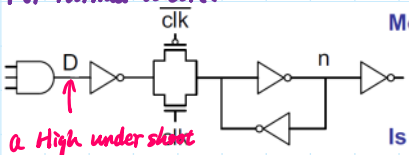
The clock noise is not common: Strong driver

Data noise is more common:

⇒ will be latched and affect the circuit.

at Setup-Hold Window

For normal latch:

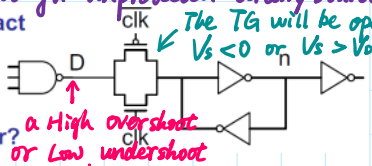


a High undershoot or Low overshoot will affect data at falling edge.

More compact

Is this better?

But for unprotected drain/source:



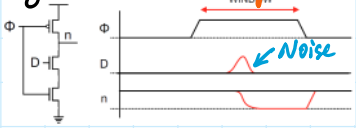
The TG will be open if $V_S < 0$ or $V_S > V_{DD}$

a High overshoot or Low undershoot can affect data at clk=Low)

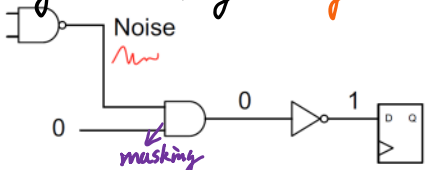
⇒ More sensitive to noise in a longer window.

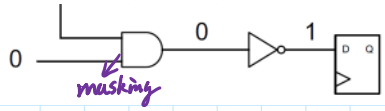
★ Latches' input must be protected.

★ Dynamic Node: Larger Sensitive Window

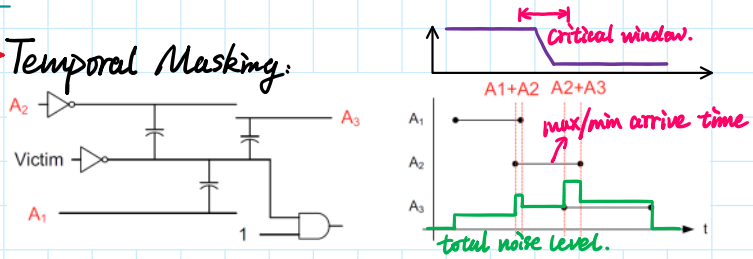


★ Logical Masking: No big deal if it can be masked at clk edge.





Temporal Masking:

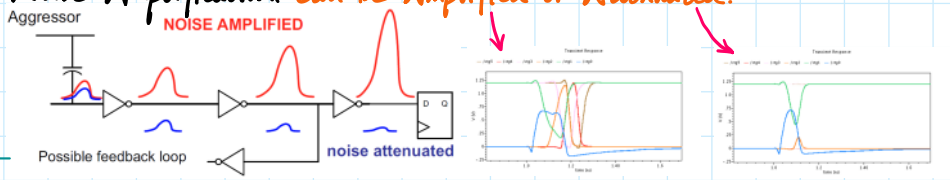


The total noise of one net is the sum of aggressors — may not at the same time

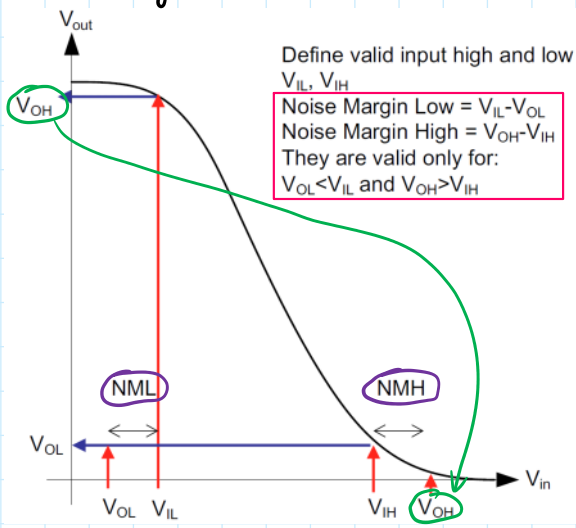
If the next stage is a FF, only need to worry about the noise at rising edge. (falling of Master)

Noise Margin — Local Analysis

Noise Amplification: Can be Amplified or Attenuated.



Noise Margin: a Local constrain to avoid amplification



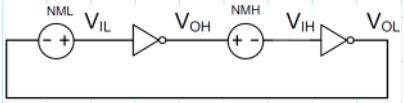
Define valid input high and low voltages V_{IL}, V_{IH}

Noise Margin Low = V_{IL}-V_{OL}

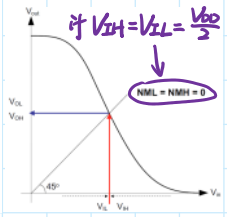
Noise Margin High = V_{OH}-V_{IH}

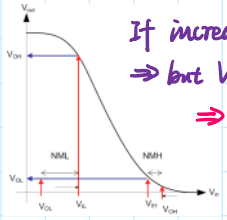
They are valid only for:

V_{OL}<V_{IL} and V_{OH}>V_{IH}



* V_{IL} & V_{IH}: Worest Condition allowed for input (set by human)
* V_{OL} & V_{OH}: Worest Condition of output (due to V_{IH} & V_{IL})
* Noise Margin: Margin left for noise under worest case.
  ⇒ V_{IL} & V_{IH} − V_{OL} & V_{OH}

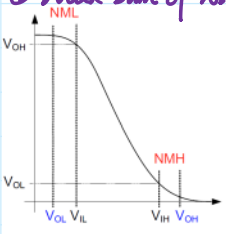




If increase V_{IL} ⇒ NML↑
⇒ but V_{OH}↓ ⇒ NMH↓
⇒ Trade off in N.M.

Optimal Noise Margin

① Max Sum of NML + NMH: V_{IL}/V_{IH} at Slope of −1 of VTC



$$NML = V_{IL} - V_{OL} \qquad NMH = V_{OH} - V_{IH}$$

$$\max \sum NML, NMH = V_{IL} - V_{OL} + V_{OH} - V_{IH}$$

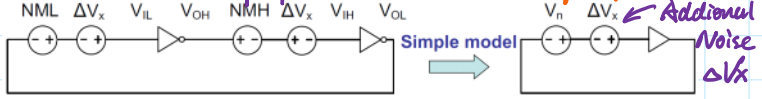
$$V_{OUT} = H(V_{IN})$$

$$F = \sum NML + NMH = V_{IL} - H(V_{IH}) + H(V_{IL}) - V_{IH}$$

$$\frac{\partial F}{\partial V_{IL}} = 0 = \frac{\partial [V_{IL} + H(V_{IL})]}{\partial V_{IL}} \qquad H'(V_{IL}) = -1$$

$$\frac{\partial F}{\partial V_{IH}} = 0 = \frac{\partial [-H(V_{IH}) - V_{IH}]}{\partial V_{IH}} \qquad H'(V_{IH}) = -1$$

② Avoid Noise Amplification: V_{IL}/V_{IH} at Slope of $\frac{1}{2}$ of VTC



k is the gain of the inverter

$$\Delta V_o = k\Delta V_I$$

$$\Delta V_I = \Delta V_n + \Delta V_n [(\alpha V_x + k\alpha V_x) k + \alpha V_x] k + \alpha V_x \cdots$$

for $k<1$, $\Delta V_o = k \cdot \left(\frac{1-k^n}{1-k}\right)$

- Unity gain k=1 oscillate Not safe.

$$\frac{\partial V_o}{\partial V} = \infty$$

k is the gain of the inverter

$$\Delta V_o = k\Delta V_I$$
$$\Delta V_I = \Delta V_x + \Delta V_o$$
$$\Delta V_o = k(\Delta V_x + \Delta V_o)$$
$$\Delta V_o(1-k) = k\Delta V_x$$
$$\Delta V_o = \Delta V_x \frac{k}{1-k}$$

$$\boxed{\frac{\partial V_o}{\partial V_x} = \frac{k}{1-k}}$$

- Unity gain k=1 oscillate Not safe.

for $k<1$, $\Delta V_o = k\cdot(\frac{1-k^n}{1-k})$

$[\Delta V_x + k\Delta V_x)k + \Delta V_x]\cdot k + \Delta V_x \cdots$

$\Rightarrow \Delta V_x(1+k+k^2+k^3\cdots)\cdot k$

$\frac{\partial V_o}{\partial V_x} = \infty$

- For $\frac{\partial V_o}{\partial V_x} = 1$ k=1/2 unity noise gain

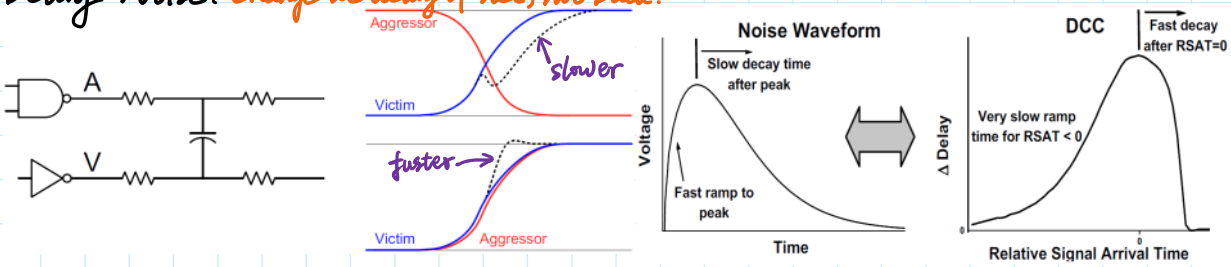
★ For safty use $k=-\frac{1}{2}$ point (actually close to k=-1 since transition is sharp)

★ Estimation in N.M. Analysis:

*Optimistic: 1. only VTC of INV (ignore other gates like NAND/NOR)
2. Assume nominal gates (No PVT)

*Pesmistic: 1. It is DC Analysis, actual noise has limited width
2. No masking
3. Not every gate is in the limit ($V_{IN} = V_{IL}/V_{IH}$) — N.M. is the worest case

★ Delay Noise: Change the delay of net, not Data.



★ Effective Caps of Coupling: $C_c \Rightarrow C_{g1}$ & C_{g2}.

★ If $= I_1 = I_2$

① Opsite Switch in same dV/dt.: Diff. Signal.

$V_1(t) = V_{DD} - V_2(t)$

$I_f = C_c \frac{\partial[V_1(t)-V_2(t)]}{\partial t} = 2C_c \frac{\partial V_1(t)}{\partial t}$

$\Rightarrow I_1 = C_{g1}\frac{\partial V_1(t)}{\partial t} = -I_2 = -C_{g2}\frac{\partial V_2(t)}{\partial t}$

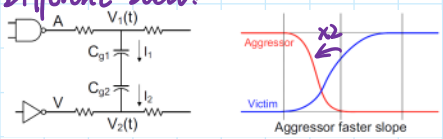
$\Rightarrow C_1 = C_2 = 2C_c$

② Same switch with same slew:

$I_f = 0$

$\Rightarrow C_1 = C_2 = 0$

③ Different slew:



$V_1(t) = V_{DD} - 2V_2(t)$

$I_f = -3C_c \frac{\partial V_2(t)}{\partial t}$

$\Rightarrow I_1 = C_{g1}\cdot\frac{\partial V_1(t)}{\partial t} = -2C_{g1}\frac{\partial V_2(t)}{\partial V_2(t)}$

$I_2 = -C_{g2}\cdot\frac{\partial V_2(t)}{\partial t}$

$\Rightarrow C_{g1} = \frac{3}{2}C_c$, $C_{g2} = 3C_c$.

★ Noise Avoidance:

- Size up victim driver :
  - Makes it a stronger aggressor
  - Only effective if wire is not too long

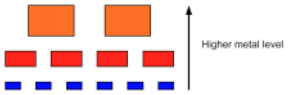
- Spacing
  - Coupling capacitance is reduced
  - Total capacitance is reduced

- Size up victim driver :
  - Makes it a stronger aggressor
  - Only effective if wire is not too long
  - Delay impact (increases input load)

- Widening wires:
  - Reduces resistance → Good for long wire with strong driver
  - Increases grounded capacitance
  - Can be good (or bad) for delay

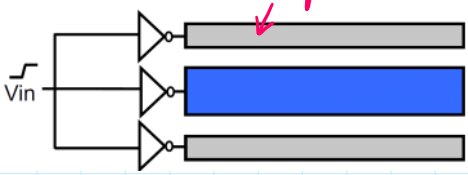
poor wire poor driver

- High level metal → Wider and more spacing



- Bus encoding
  - Minimize the number of transitions in a bus
  - Address/Data buses have different behavior
- Minimize capacitive coupling:
  - Switch the shields with the victim
  - Speed up delay

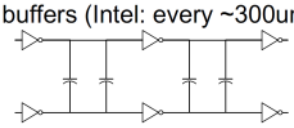
sacrifical wire.



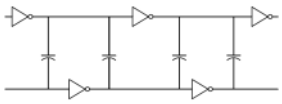
- Spacing
  - Coupling capacitance is reduced
  - Total capacitance is reduced
  - Good for delay
  - Area penalty

- Shielding (always for clk)
  - Coupling capacitance is eliminated
  - Total capacitance is increased

- Insert buffers (Intel: every ~300um)



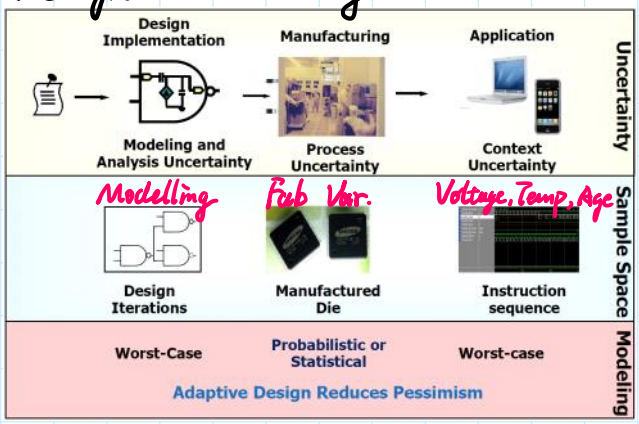
- Staggered buffers
  - Injects both positive/negative noise → cancel



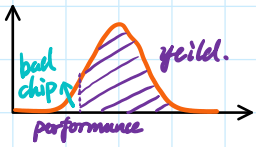
L4. Process Variations & DFM

2023年2月4日 16:58

Design Uncertainty:



It is a statistical process
Usually in random distribution



Modeling Errors: Model may not be accurate for some case (e.g. sub-Vth)

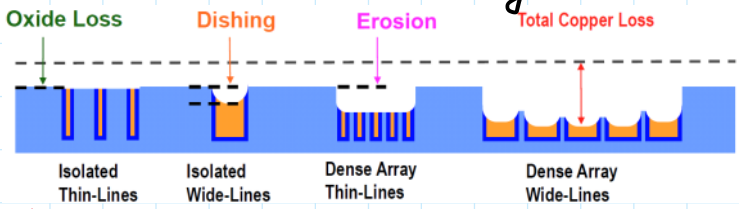
Manufacturing Variation

e.g. Tox: $\frac{\sigma}{\mu}$ = 3%; Leff: $\frac{\sigma}{\mu}$ = 5~7%; Vth: $\frac{\sigma}{\mu}$ = 5%

⇒ Different Process has different effect to variation.

One physical para. can cause multi device para. variation.

Chemical Mechanical Polishing: Metal thickness

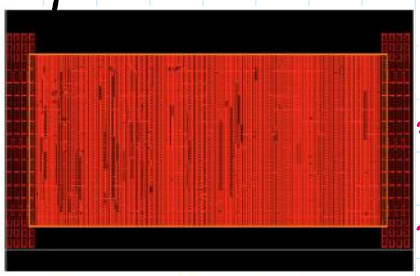


Different Metal Density has different Polishing depth (SiO_2 is harder than metal).

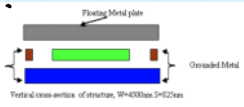
Tmetal ↓: C↓ but R↑

Use Metal Fill to have uniform metal distribution:

Drawn routes in processor block



Pre-tape out inserted fill

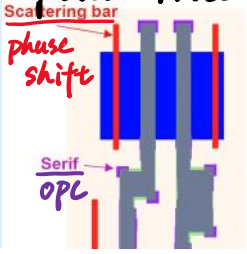


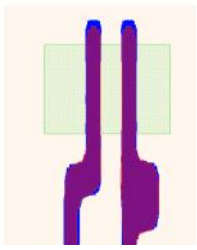
*If M_{fill} is grounded:
Large C_g but small C_c.
*If M_{fill} is floating:
Large C_c

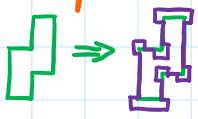
Lithography Related Variation: With resolution enhanced tech (RET)

Min Feature Size: $K_1 \cdot \lambda / NA$

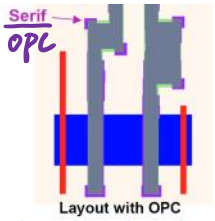
Optical Process Correction: Add pattern to compensate

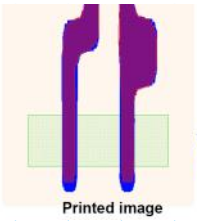




e.g. 

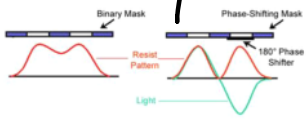
Since litho may not good at the edges of layout.





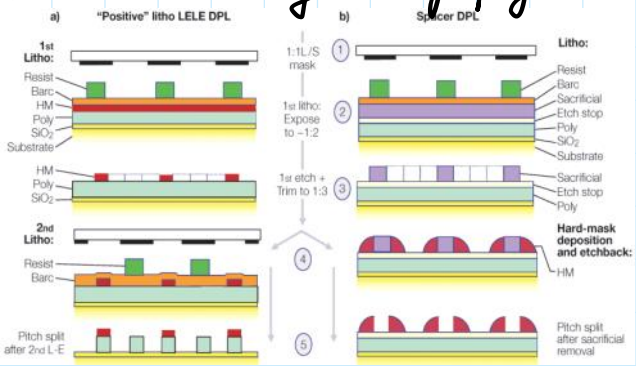
at the edges of layout.

Phase Shifted Masks



Enhance Resolution

Double Parttoning Lithography: closed gate may have different variation

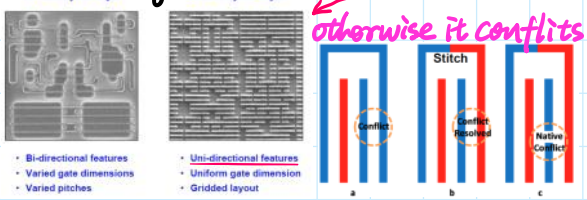


⇐ Two ways of Double Patterning.

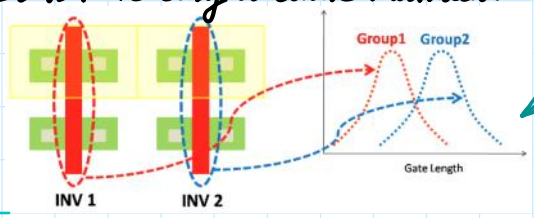
It gives very good accuracy

but it layout should in one direction

otherwise it conflits



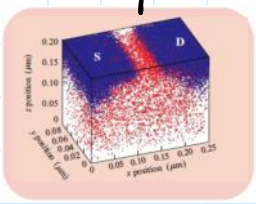
It is more easy to cause mismatch — Especially for SRAM

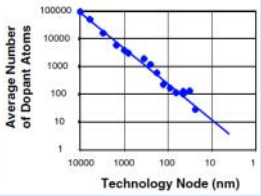


Different Process

Different Distribution

Ion Implantation — Random Dopants.



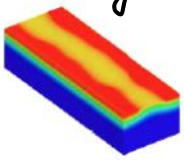


Number of Dopants is small in advanced node

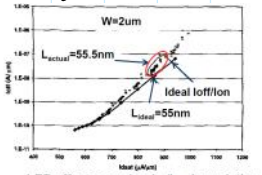
Maybe only ~10 dopants

⇒ More easy to have variation

Line Edge Roughness: Radom Distribution of Photon



Potential distribution in channel with LER (Simulation)



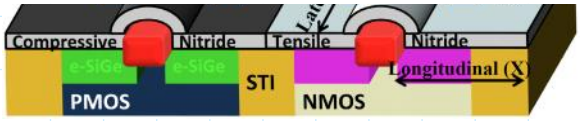
LER effects causing I_{off}/I_{on} degradation
Source: Wu et al., SISPAD 2002

⇒ changed Ion/Ioff and on-off ratio if Leff changed

Mobility Enhancing — Stress

	NMOS	PMOS
X	Tensile	Compressive
Y	Tensile	Tensile
Z	Compressive	Tensile

X	Tensile	Compressive
Y	Tensile	Tensile
Z	Compressive	Tensile

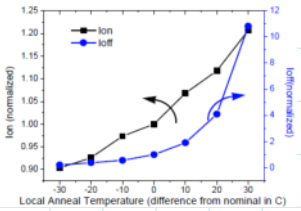


The stress changes with the Layout of drain & drain contact

RTA Variation:

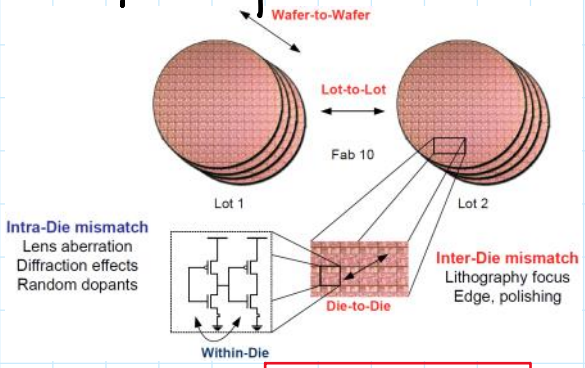
To avoid the gate metal damage, use rapid thermal annealing (RTA)

⇒ Time is too short, there may be Temp. Variation specifically.



⇒ changed Vth & Rext and Ion/Ioff.

Classifiction of PV:



1. Lot-to-Lot
2. Wafer-to-Wafer
3. Die-to-Die
4. Within-Die

Process Variation

- Systematic Error
  - RTA
  - Stress
  - OPC
  (repeatable)
- Random Variation
  - Die-to-Die
  - Within Die
  - WiD error also DTD
    - Spacial Correlated
      - Gate Length
      - Dose of Light
      - Etch Rate
    - Independent
      - RDF
      - LER

There will be a batch Calibration that test some chips and calibrate the whole batch ⇒ Less Die-to-Die variation is easy for calibration

Simulation of PV:

E.g. for gate length:

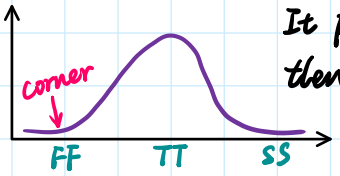
$$L_{eff} = L_{nom} + \Delta L_{system} + \Delta L_{Die\text{-}to\text{-}Die} + \Delta L_{random}$$

($\Delta L_{system} + \Delta L_{Die\text{-}to\text{-}Die}$: shared; ΔL_{random}: Mismatch)

Corner Simulation: only captured Die-to-Die variation

Shared Mismatch

★ Corner Simulation: only captured Die-to-Die Variation



It pick a point in the plot and then apply to all gates

⇒ No Mismatch.

⇒ Underestimation

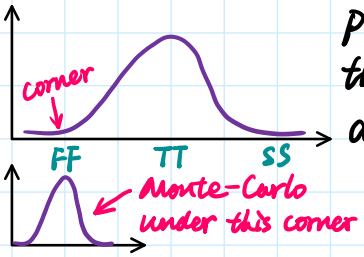
★ Global Monte-Carlo:

Randomly pick points on the distribution

⇒ It includes Die-to-Die variation into Local simulation

⇒ Overestimation

★ Local Monte-Carlo: Most Accurate

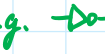


pick a corner first — Die-to-Die

then monte-carlo for distribution around this corner — Mismatch

★ PV and Circuit Delay:

★ Gates in Series: td is the Sum.

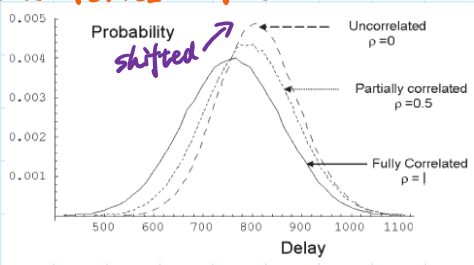
e.g. 

If uncorrelated paths: Mean adds, Std = $\sqrt{\sigma_1^2+\sigma_2^2\cdots} = \sqrt{n}\sigma$

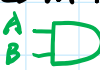
★ $\left(\frac{\sigma}{\mu}\right)_{path} = \frac{1}{\sqrt{n}} \cdot \left(\frac{\sigma}{\mu}\right)_{gate}$

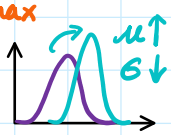
If correlated:

★ $\left(\frac{\sigma}{\mu}\right)_{path} = \sqrt{\frac{1+\rho(n-1)}{n}} \left(\frac{\sigma}{\mu}\right)_{gate}$



★ Gates in Parallel: $td = \{t_{path}\}_{max}$

e.g. 



If uncorrelated paths: Mean ↑; Std ↓

If correlated paths and $\rho=1$: Mean & Std keeps the same

★ Operating Context Variation

1. Supply Voltage: IR Drop, $L\,di/dt$ Drop

- Operating Context Variation
  1. Supply Voltage: IR Drop, $L\,di/dt$ Drop
  2. Temp. Hot Spot
  3. Relibility: NBTI/PBTI, TDDB, HCI (hot currier degradation)
  4. Electromigration

L5. Adaptive Design

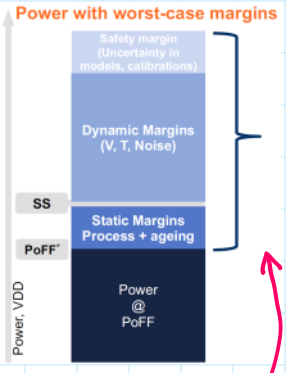
2023年2月18日 23:40

Motivation:

Variations in Circuits. Time & Spacial

SPATIAL REACH	STATIC: EXTREMELY SLOW	DYNAMIC: SLOW-CHANGING	DYNAMIC: FAST-CHANGING
GLOBAL	Inter-die process variations Life-time degradation (NBTI, TDDB)	Package/Die VDD fluctuations Ambient temperature variations	PLL jitter IR Drop
LOCAL	Intra-die process variations	Temperature hot-spots	IR Drop Coupling noise (capacitive and Ldi/dt) Local Clock-jitter (IR drop in clock-tree)

Slow TEMPORAL RATE OF CHANGE Fast



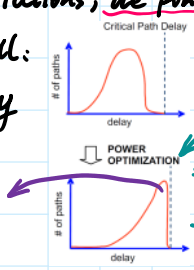
* To provide margin to all of the variations, the power is high

And for circuits that optimazed well:

⇒ A lot of circuits has similar delay

⇒ the No. of critical path is high

- More sensitive to variations



No. of critical path is high

⇒ More chance to fail

⇒ leave Large margin to keep yeild

Adaptive Method:

Adaptive design is used to reduce the margin need by variation

It has several methods:

"Always correct" approaches ⇒ with Margin

- Predict point of first failure and add small margin to this point
- Look-up table based techniques
- "Canary" circuits
- In-situ delay detection

"let fail and correct" approaches ⇒ No margin

- Self-calibrating techniques - Circuits tune until point of failure
- Require recovery mechanisms
- Eliminates margins due to global and local variations

Always Correct Approaches.

Look-up Table Based DVFS:

Two ways: Design Based & Post Silicon (use tester)

Set some voltage & Frequency Combinations

Voltage	Frequency	Power
1.65V	800MHz	900mW
1.3V	600MHz	450mW
0.75V	200MHz	50mW

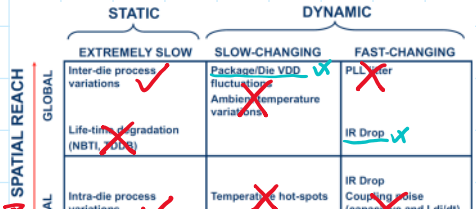
Rise voltage first ⇒ rise freq.

Drop freq. first ⇒ Drop voltage

} ⇒ keep circuit op. safe.

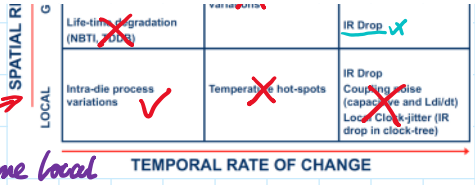
May use a glitchless Mux for PLL re-clocking

* For Design Based LUT, solves no variations ⇒ Not Adaptive.



variations ⇒ Not Adaptive.

* For Post-Silicon Based LUT:



Mostly static & global, but still some local

*Advantages

- Very easy to design and deploy
- Exploits low CPU utilization epochs through DVFS for better energy efficiency

*Disadvantages

- Design time tables do not eliminate margins – only gives frequency / energy trade-off (i.e, dynamic but not adaptive)
- Post Silicon tables: only addresses static variations (global/local process)
- Requires worst-case safety margins for all other variations
- Tester time calibration introduces significant inefficiency because of the different frequency points

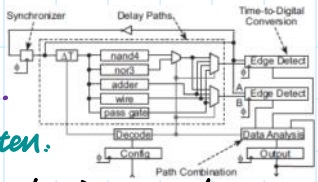
Canary Circuit: Mirror the critical path

Test the delay of canary circuit ⇒ Adjust V_{DD} to fit.

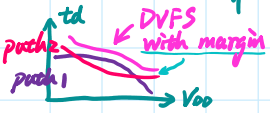
Canary Blocks:

(1) use inverters: Ring Osc. Easy but not very accurate
⇒ May have different critical path — Different Delay Pattern

(2) Critical Path Moniter:

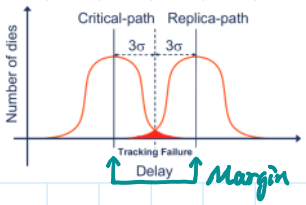


combination of gates ⇒ More accurate.
Diff V_{DD} can lead to different delay partten.
e.g.

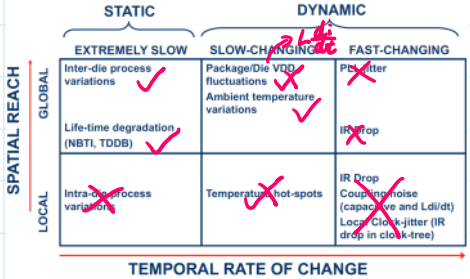


A small mistrack gives small margin
⇒ Higher Effeciency

Local Variations:



The delay caused by local variation can not be tracked
Even if the canary circuit has margin, it still chance to fail
⇒ May need localized ⇒ Put close critical path



Mainly track global & slow variation

*Advantages

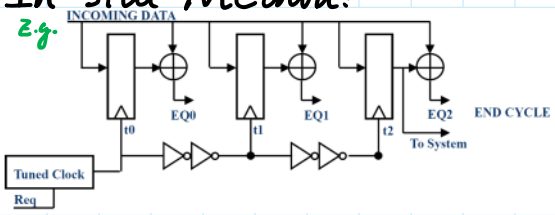
- Relatively easier to design and deploy
- No changes to processor design
- No need for tester time

*Disadvantages

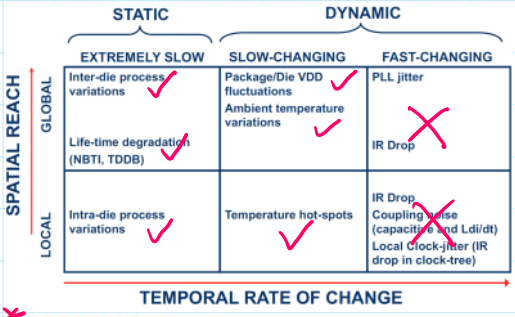
- Adapts to more variations: slow and static, global variations
- Does not address: fast variations, local variations
- In some cases canary circuits can be tuned to reduce margins further but with more testing cost

☆In-situ Method:

e.g.



Detect Data by different FF:
⇒ Can see it is fast/slow/just in time.



Need at least 1 clk to response
⇒ Not work for very fast change.

*Advantages

- Captures Local and Global variations – monitors the actual circuit delay and not a copy
- Eliminates margin for all slow moving variations

*Disadvantages

- More invasive in circuit design – need to change the circuits themselves internally.
- Some delay overhead from monitors
- Requires periodic halting of processor to run worst-case vectors

⇒ only works if it is worst-case vector

☆Let Fail & Correct:

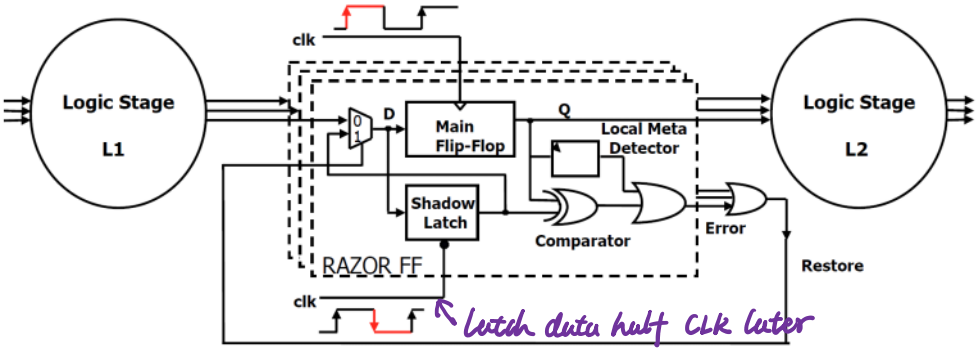
The circuit will fail & be corrected with a certain recovery cycle

Let Fail & Correct:

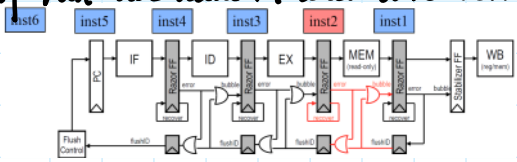
The circuit will fail & be corrected with a certain error rate

⇒ It reduce the margin to zero

Razor I:

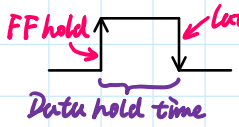


Compare Main FF with Latch: if Fail use data in Latch to recover.



Problem:

(1) Hold & Max Detection Time: Data need to be held half CLK

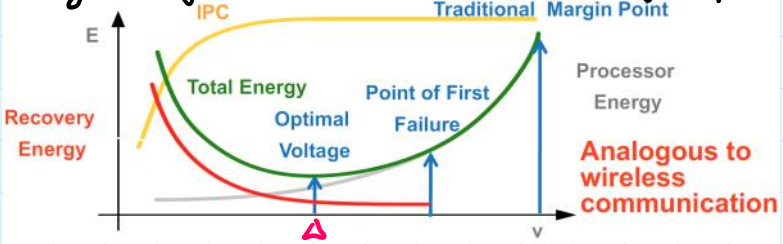


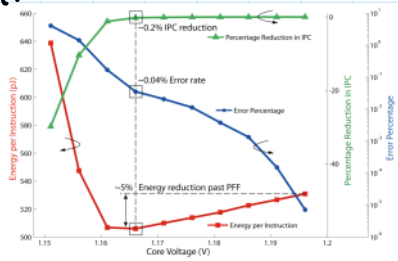
it is also the max detection time.

otherwise Latch will hold the next data.

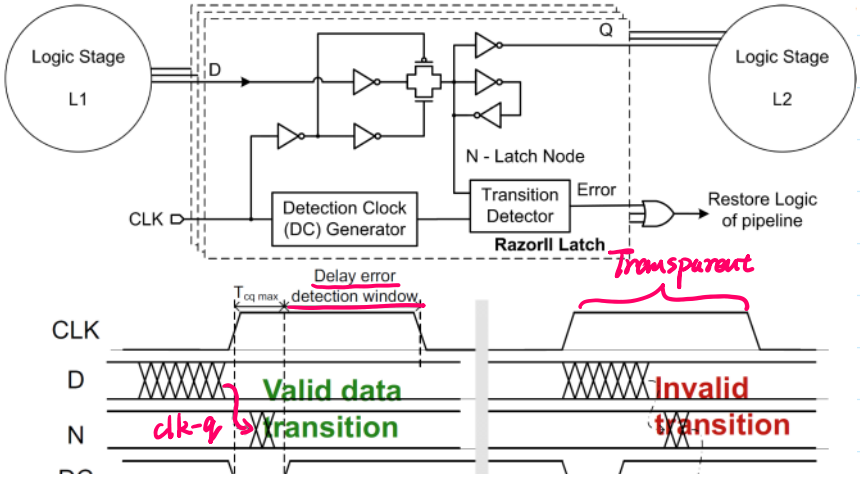
(2) Very Large OR gate for error detection: slow & large.

By choosing a suitable error rate, the efficiency is optimal.





Razor II:



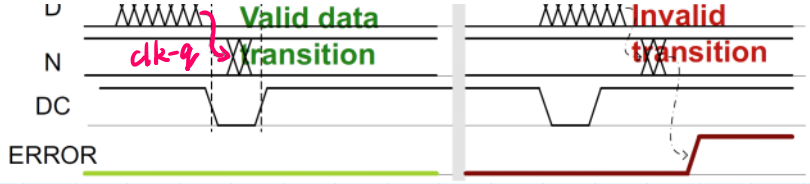
(1) Give a pulse at clk edge

Block the transition detector

(2) Detector transition in rest of the times ⇒ Yes, means fail

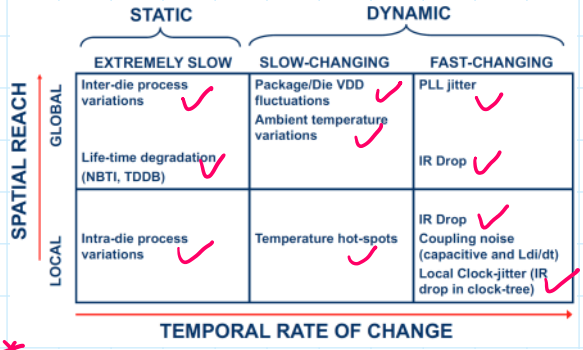
(Latch transparent)

(3) Flag and Refetch Instruction



(3) Flag and Refetch Instruction

✱Conclution:



if delay within detection range.
⇒ Track all variations.

*Advantages

- Eliminates all margins
- Also captures instruction dependent delay variations

*Disadvantages

- Requires architectural changes in the processor design for error correction mechanism
- Hold time constraint increases overhead of approach with larger timing speculation window

L6. Low Power

2023年2月7日 19:26

★ Review of Dynamic Power:

see ZECS427.

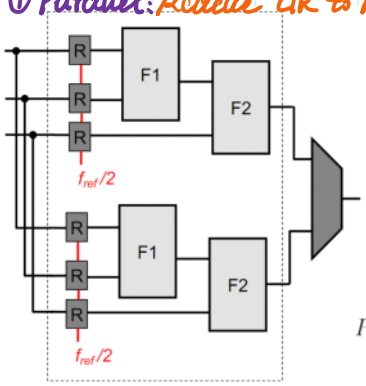
- Reduce C:
  - Sizing, P/N ratio, shorter routes
  - New circuit structures
- Reduce α:
  - Clock gating
  - Bus encoding → already discussed for a diff. sig. $C_g = 2C$.
- Scaling V_{dd}: Very effective
  - Parallel/Pipeline tradeoff
  - Low swing signaling/clocking
  - Dynamic Voltage Scaling (DVS)

★ Multiple V_{DD}:

Since power: $P = \alpha C V_{DD}^2 f$, Reduce V_{DD} can reduce $V_{DD}^2 f$ — 3rd order.

★ Parallel & Pipelined Datapath:

① Parallel: Reduce clk to half



$$f_{par} = f_{ref}/2$$
$$C_{par} = (2 + ov_{par}) \cdot C_{ref}$$
$$V_{dd,par} = \epsilon_{par} \cdot V_{dd,ref}$$

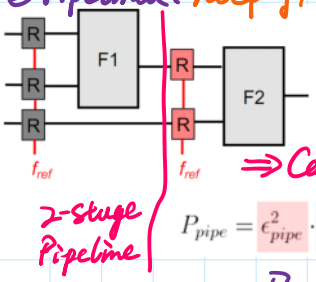
V_{DD} can drop to ~half.

Almost 1

$$P_{par} = \epsilon_{par}^2 \cdot \left(\frac{2 + ov_{par}}{2}\right) \cdot P_{ref}$$

⊖ 1. Large Area
2. Large delay / Data hazard.

② Pipelined: keep f, reduce V_{DD}



$$f_{pipe} = f_{ref}$$
$$C_{pipe} = (1 + ov_{pipe}) \cdot C_{ref}$$
$$V_{dd,pipe} = \epsilon_{pipe} \cdot V_{dd,ref}$$

⇒ Can Reduce V_{DD} since the load is lighter.

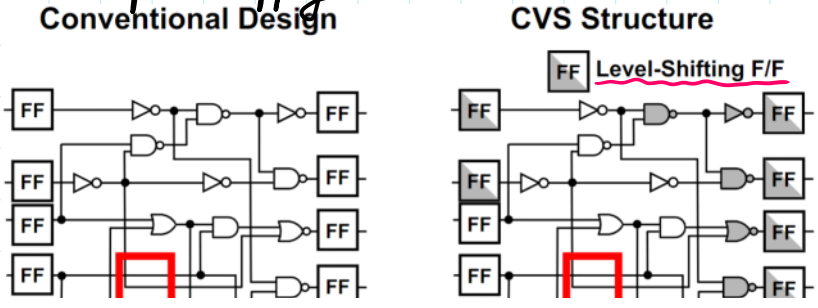
$$P_{pipe} = \epsilon_{pipe}^2 \cdot (1 + ov_{pipe}) \cdot P_{ref}$$

⊖ Large delay / Data hazard.

	Parallel	Pipelined
Area:	↑	—
Delay:	↑	↑
Leakage:	↑	—

⇒ Pipeline is a better choice

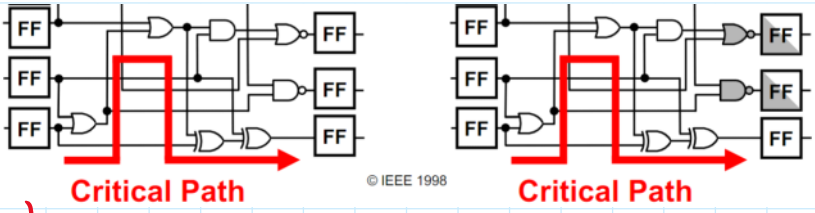
★ Multiple Supply in one Block: Low V_{DD} in non-critical path.



① Level Shifted by FF

② Only go from $V_{DDH} \rightarrow V_{DDL}$ inside the block (No LS)

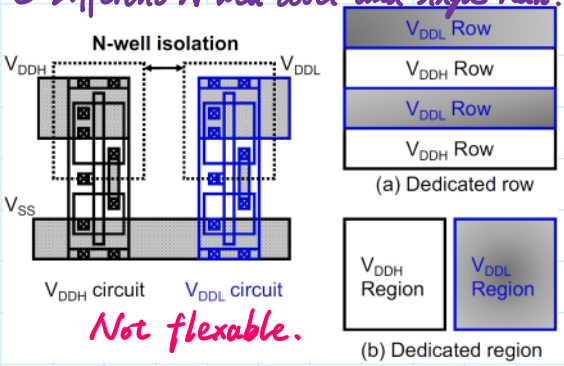
③ Keep Critical Path in V_{DDH}



③ Keep Critical Path in V_{DDH}

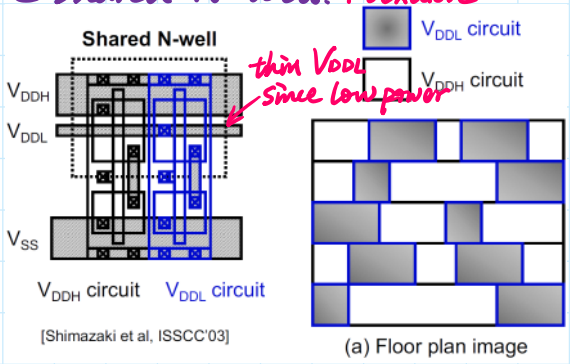
☆ Standard Cell in multi-V_{DD}:

① Different N-well level and single Rail:

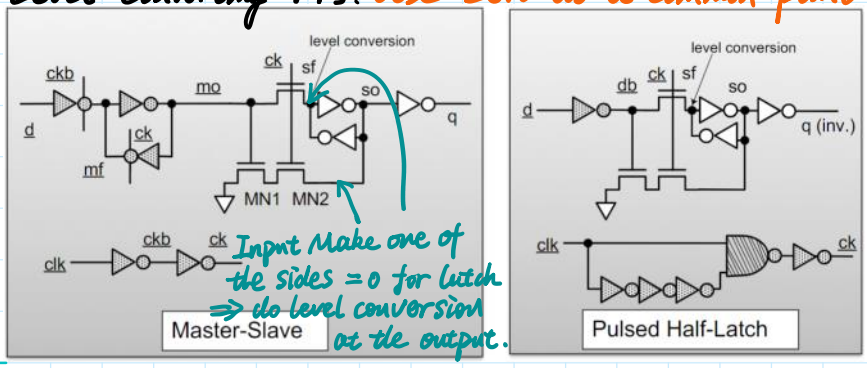


Not flexable.

② Shared N-well: Flexable



☆ Level Converting FFs: Use zero as a common point to convert levels.



☆ Dynamic Power Reduction: DVFS

Reduce the power ∝ duty³ by DVFS.

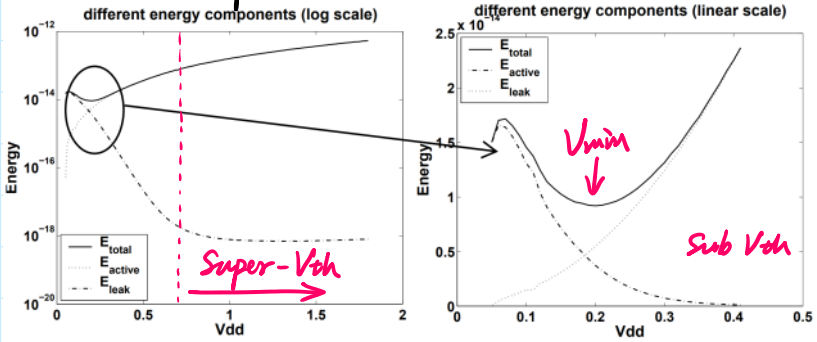
☆ Find the min power point for scaling: Bo Zhai, Blaauw, DAC, 2004

① $V_{DD} > V_{th}$: super-threshold region: Energy can be saved by smaller V_{DD}

1. Active Energy drops ∝ V_{DD}^2↓ $E = aCV_{DD}^2$
2. Delay rises ∝ V_{DD}↓
3. I_{leak} keeps the same.

} Static Energy keeps the same

$E = V_{DD} \cdot I_{leak} \cdot Time$



② Sub-V_{th} region: Has a V_{min} for min Energy

1. Active Energy drops ∝ V_{DD}^2↓

② Sub-Vth region: Has a Vmin for min Energy

1. Active Energy drops $\propto V_{DD}^2 \downarrow$
2. Delay rises $\propto \exp(V_{DD})\downarrow$
3. Ileak keeps the same.

} Static Energy rises $\propto \exp(V_{DD})\downarrow$

e.g. For 0.18μm tech: $V_{min} = [1.587\ln(\eta \cdot n_{eff}) - 2.355] \cdot m \cdot \frac{kT}{q}$ ← Not related to Vth

n: Stage Number
α: Active Factor

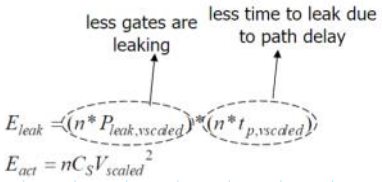
* There is a $n_{eff} \equiv \frac{n}{\alpha}$;

It means the ratio between leaking and active gates.

$n_{eff} < 10$, No Vmin — too many switching gates, the Energy can be reduced by keep reducing V_{DD}.

★ DVFS works well for Small n_{eff}:
— less stages & large α

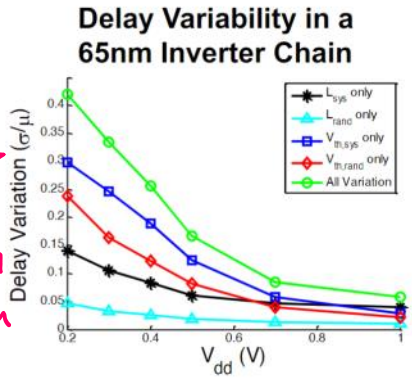
- Factors affecting V_{min}: α, n, T, S_S
  when $\alpha \uparrow, n \downarrow, T \downarrow, S_S \uparrow (m \downarrow)$, $V_{min} \downarrow$
  $V_{min} = [1.587\ln(\eta \cdot n_{eff}) - 2.355] \cdot m \cdot \frac{kT}{q}$
- Vmin dependency on α and n:



★ Problems in Low Voltage Design:

- Sensitivities → exponential
  - Supply voltage
  - Threshold voltage
  - Temperature
- Complicates SRAM design
  - Alternative cell design again helpful
- Makes timing difficult
  - Short paths, long paths
  - Increased clock skew
  - Margins grow

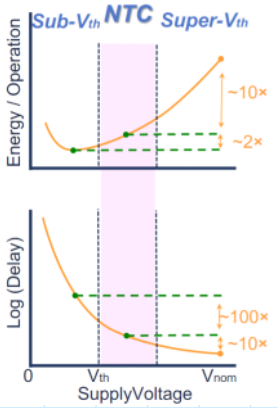
Low V_{DD} ⇒ Large variation Hard for circuit Design



⇒ Can use Near Threshold Computing (NTC)

* Save a Large amount of energy

* But only add a small delay.



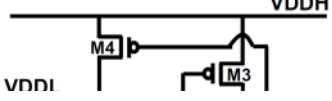
★ Level Converters:

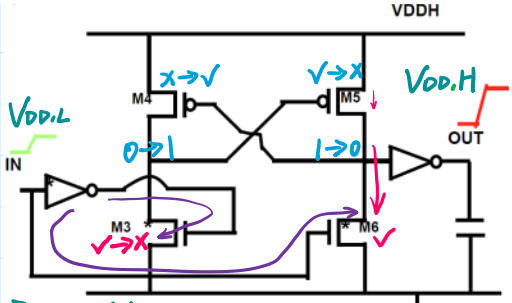
★ DCVS & Pass Gate: Use Different Current level.

① DCVS:

Io ↑

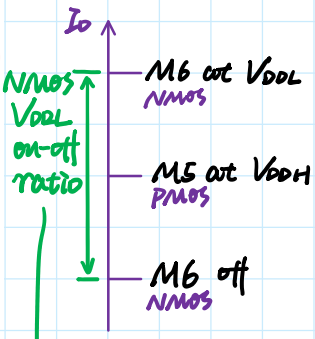
② PG: Similar to DCVS

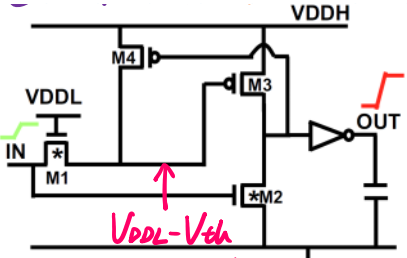




Even M6 may not fully on at $V_{DD,L}$ by have a weak PMOS and feedback it will finally stablelized.

High power but great connection



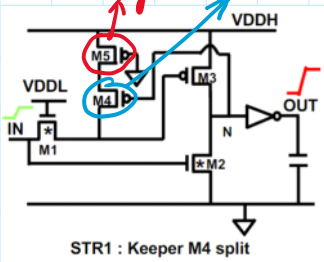


weaken M3 when M2 is on.

If the on-off ratio is not large (i.e. V_{DDL} is too small) The design margin for PVT variation is also small.

③ Improved PG:

Large L + Small L = Large L

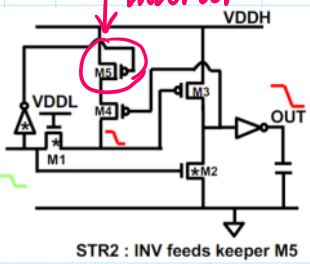


STR1 : Keeper M4 split

Since this PMOS should be weak ⇒ Large L give large load to output.

⇒ use M4 & M5 in series to reduce load.

Similar but use inverter

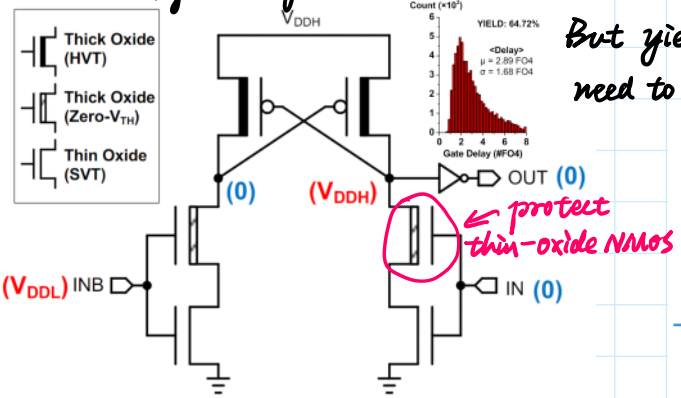


STR2 : INV feeds keeper M5

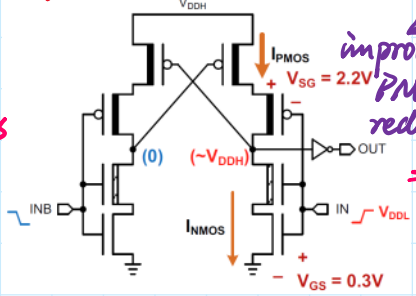
☆ High Voltage Convertor

Since for a very low V_{DDL} (say 300mV) the pervious converter may not be able to work

⇒ Use High Voltage Convertor: Thick oxide device is for high voltage.



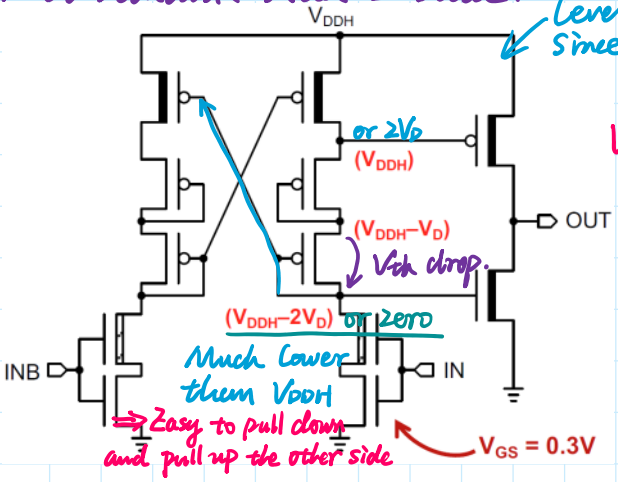
But yield is low since NMOS driven by V_{DDL} need to fight with PMOS driven V_{DDH}



improved by add a PMOS but only reduce a V_{DDL}

⇒ Not good.

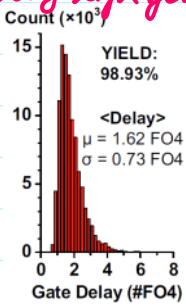
Final Solution: Add 2 Diode:



Level shift since it is not full swing

⇒ Easy to pull down and pull up the other side

Very high yeild



⊖ High Leakage

⊖ Only work for large V_{DDH} otherwise $V_{DDH} - 2V_D$ too small

L7. Energy Recovery (Adiabatic)

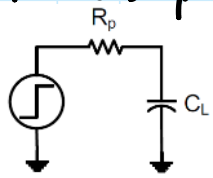
2023年2月18日 23:40

★ Energy in Digital Circuit:

Once there is I & V acrossing resistor (Transistor)

⇒ There is power consumption.

For a step to RC: $E = \frac{1}{2}CV_{step}^2$



So, if make it to very small steps:

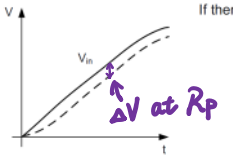
$$\star\ Z = \left[\frac{1}{2}C\left(\frac{V}{n}\right)^2\right]\cdot n = \frac{1}{2}C\frac{V^2}{n}$$

For $n \to \infty$ (Ramp)

$P \to 0$

Since the ΔV between $R_p \to 0$

In practice: Use slow slew



$E = I^2RT$

If there is a small change in voltage over time:

$$I = \frac{CV}{T}$$

$$E = \left(\frac{CV}{T}\right)^2 RT$$

$$= \left(\frac{RC}{T}\right) CV_{dd}^{\ 2}$$

$T \to \infty \qquad E = 0$

If $T \gg RC$:

$\Rightarrow Z \to 0$

E.g. $RC \sim 100ps \Rightarrow T = 1ns$

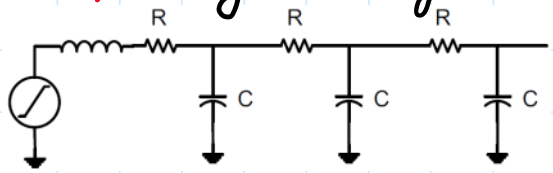
$f_{clk} = 1GHz$ – Not too slow

★ Zero Power Clock:

After making RC loss to be small, to achieve a zero clock power

⇒ Need to reuse the energy stored in Caps

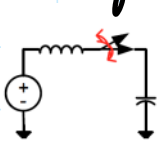
★ Slowly Oscillating LC Networks.

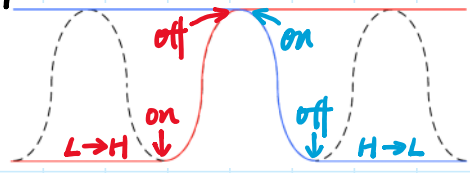


if $T \gg RC$,

$Z \to 0$

For Logic operation: use switch to sample/hold the clock





ideally no power loss for LC.

only power loss at R_P.

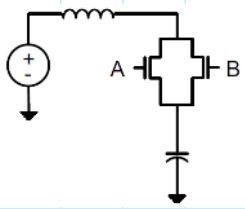
Tridtional Clock loss all energy in cap.

★ Adiabatic Logic:

★ Pass-gate Logic:

Adiabatic Logic:

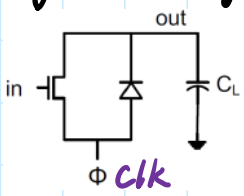
Pass-gate Logic:

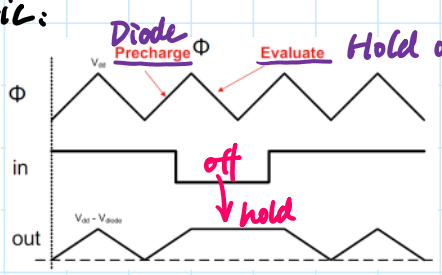


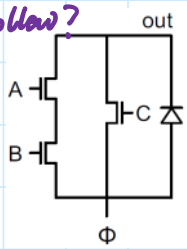
⊕ 1. Low Energy; 2. Auto Clock Gating

⊖ 1. Must switch at right time
2. t_d = 1 clock phase
3. Swing (Vth drop)
4. Cap is changing (f_c changing)
5. Need high Q for low loss

Dynamic Logic:

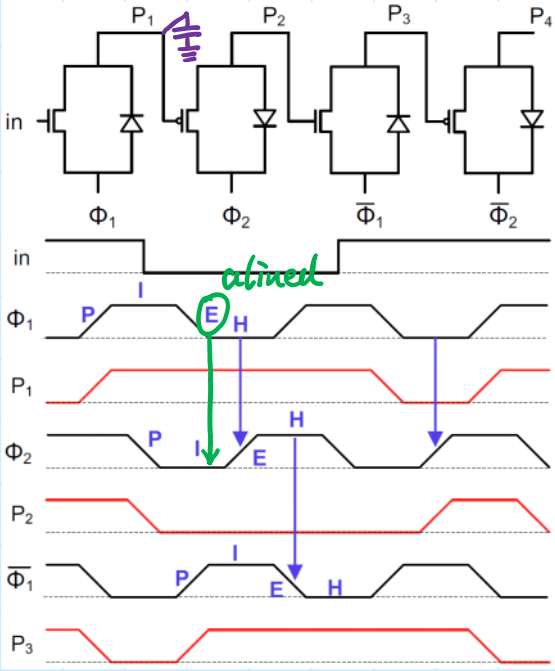




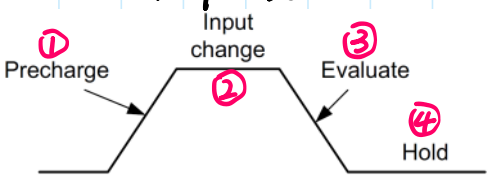


$Out = \overline{(A \cdot B) + C}$

*In cascade:



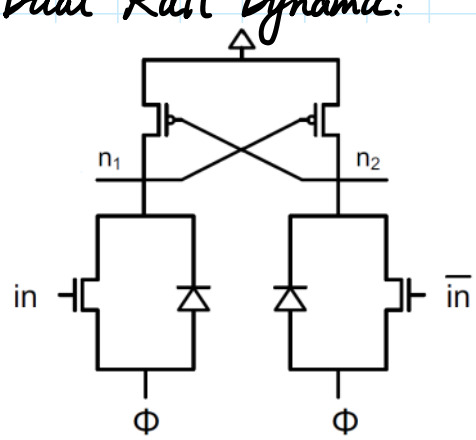
Need 4-phase Clock: $\Delta\phi = 90°$

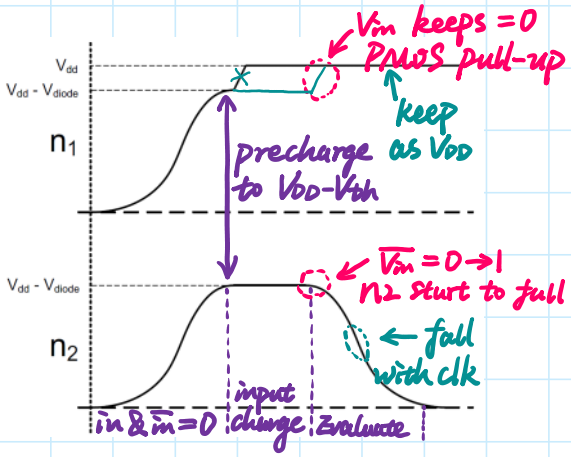


*Problems:

1. Diode give a V_D drop
2. "Free Pipelining"
   ⇒ but no arbitry connection.
3. Cap is changing (f_c changing)
   ⇒ More loss if $f_{clk} \neq f_c$.

Dual Rail Dynamic:

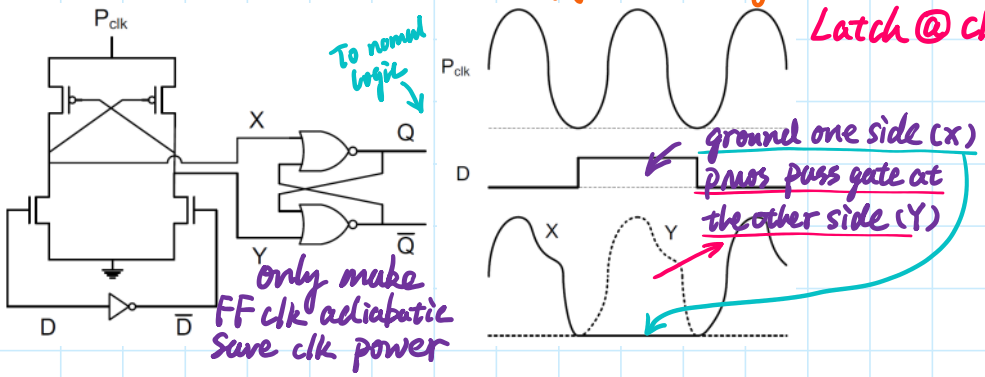




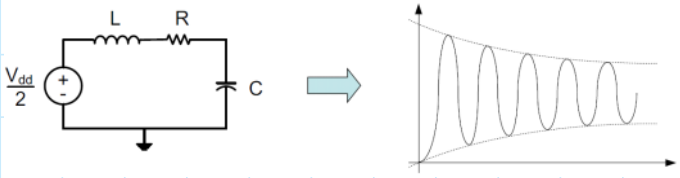
⊕ 1. Equal load Cap (No f_c change): see half of C every switch
2. Full-rail realized by cross-coupled PMOS
⊖ Too many transistors (6T for a Inv)

★ Adiabatic FF: PMOS Energy Recovering FF (pTERF)

Latch @ clk=Low

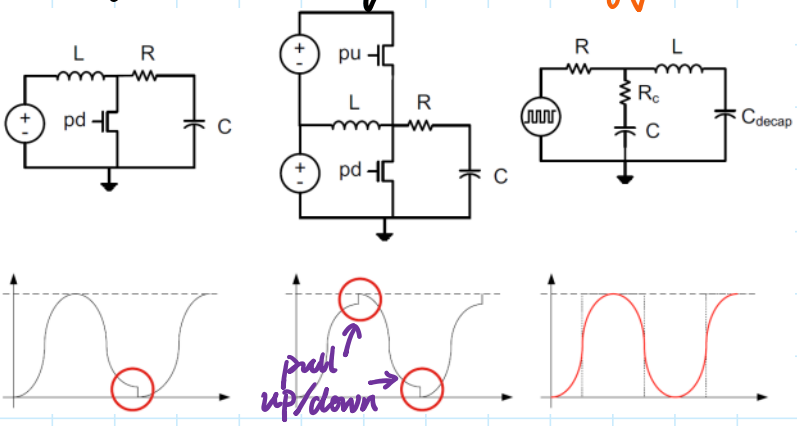


★ Clock Generation:



$W_0 = \frac{1}{\sqrt{LC}}$
⇒ But have loss

⇒ Injection Recovery: Add Energy to LC at clock = Low & High



L8. Leakage

2023年3月21日 0:24

Leakage Mechanisms.

Subthreshold Leakage. Covered in EECS 427. L9.

Isub formula & DIBL:

$$I_{DS} = 2n\mu C_{ox}\frac{W}{L}\left(\frac{kT}{q}\right)^2 e^{\frac{V_{GS}-V_{TH}}{nkT/q}}\left(1-e^{\frac{-V_{DS}}{kT/q}}\right) = I_S e^{\frac{V_{GS}-V_{TH}}{nkT/q}}\left(1-e^{\frac{-V_{DS}}{kT/q}}\right)$$ — w/o DIBL

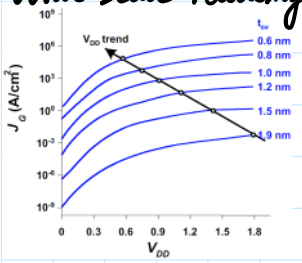
$$P_{leak} = (I_0\frac{W}{W_0}10^{\frac{-V_{TH}}{S}})(V_{DD}10^{\frac{\lambda_d V_{DD}}{S}})$$ — w/ DIBL

⇒ Sub-th leakage power is a strong function of V_{DD}.

Gate-Induced Drain Leakage: GIDL

*PMOS has small gate leakage — hole tunneling is less.

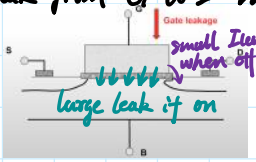
With scale reducing: t_{ox}↓ ⇒ More gate leakage.



Also, V_{DD}↓ with scaling. ⇒ help to reduce a little.

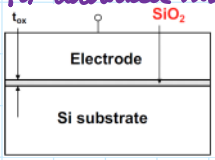
At the channel is off, only a little Ileak from G to S via overlap.

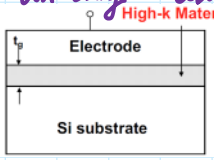
⇒ can be ignored when MOS is off.



High-k Gate Oxide:

For advanced node, t_{ox} only ~ atoms ⇒ large leakage ⇒ use thick oxide but high-k

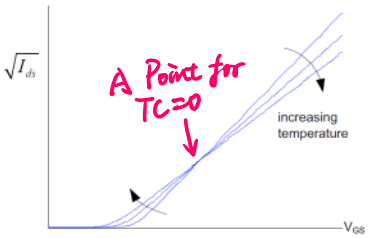


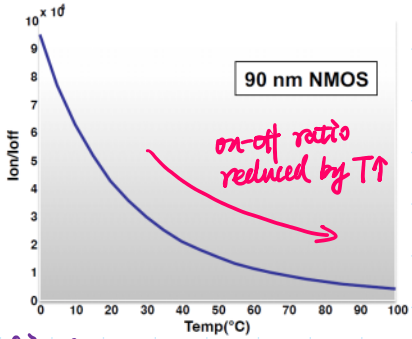


	High-k vs. SiO_2	Benefits
Gate capacitance	60% greater	Faster transistors
Gate dielectric leakage	> 100% reduction	Lower power

Temperature Effect:

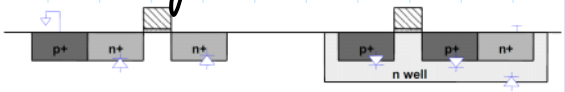
- Increasing temperature
  - Reduces mobility
  - Reduces V_{TH}
- I_{ON} decreases with temperature μ↓
- I_{OFF} increases with temperature Vth↓





Since different region, different effects give different affections.

Other Leakage: Diode



- Electron-hole pair generation in depletion region of reverse-biased diodes
- Diffusion of minority carriers through junction
- For sub-50nm technologies with highly-doped pn junctions, tunneling through narrow depletion region becomes an issue

band-to-band Tunneling (BTBT)

Strong function of Temp.

HT ⇒ large leakage

- For sub-50nm technologies with highly-doped pn junctions, tunneling through narrow depletion region becomes an issue

Strong function of temp.
HT ⇒ large leakage

Leakage Reduction:

Reduce Vth: ×10 leakage ⇌ 20%~30% delay

	Low-Vt; 0.9V	High-Vt; 0.9V	Low-Vt; 1.8V	High-Vt; 1.8V
Leakage (norm)	1	0.06	1	0.07
Delay (norm)	1	1.30	1	1.20

10× 10×
30% 20%
30% > 20%.

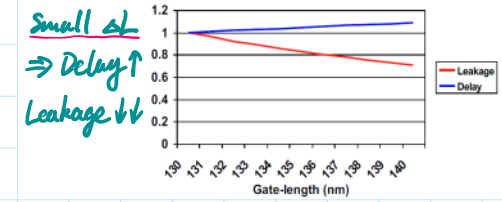
$$Id \propto \frac{Vth}{(V_{DD}-V_{th})}$$

But for low V_{DD}, the advantage from HVT is less since delay increase more.

Lower Temperature: Cooling

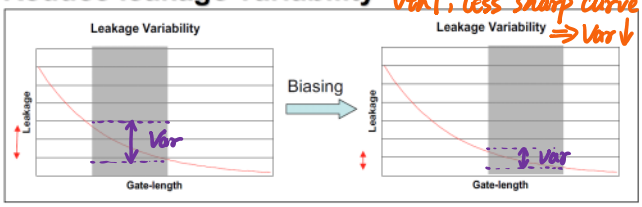
Increase Gate Length: L ↑, Vth ↑

- Reducing leakage due to V_{th} roll-off



- Reduce leakage variability

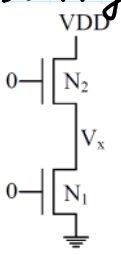
Vth↑, less sharp curve ⇒ Var↓



Decrease V_{DD}: DIBL $I_L \propto \exp(V_{DD})$

Low-Power Techniques:

Stacking Transistors:



→ DIBL by Vx

Same Isub for N1 & N2

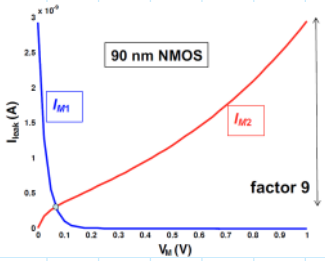
$$I_{sub} = \underbrace{I_{off}10^{\frac{\eta(V_x-V_{DD})}{S}}}_{N1} = \underbrace{I_{off}10^{\frac{-V_x+\eta((V_{DD}-V_x)-V_{DD})-k_\gamma V_x}{S}}}_{N2}$$

DIBL Body effect.
Vgs

$$\text{★ } V_x = \frac{\gamma V_{DD}}{1+2\gamma+k_\gamma} \approx 100mV$$

$$I_{sub} = I_{off}10^{\frac{-\eta V_{DD}\left(\frac{1+\eta+k_\gamma}{1+2\eta+k_\gamma}\right)}{S}} \approx I_{off}10^{\frac{-\eta V_{DD}}{S}}$$

⇒ Isub reduced 10x by Stacking a transistor.

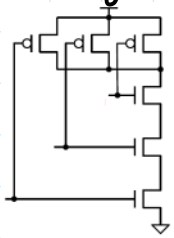


Leakage Reduction	
2 NMOS	9x
3 NMOS	17x
4 NMOS	24x
2 PMOS	8x
3 PMOS	12x
4 PMOS	16x

2x 1.5x

→ but the advantages from stacking is not significant when n>2.

Stacking Order Dependence of Gate Leakage:



State	I_{sub}	I_{gate}	I_{total}
000	0.382	0.000	0.382
001	0.709	6.339	7.048
010	0.709	1.275	1.275
011	5.626	12.677	18.303
100	0.676	0.000	0.676
101	3.804	6.339	10.143
110	3.804	0.000	3.804
111	28.273	19.015	47.288

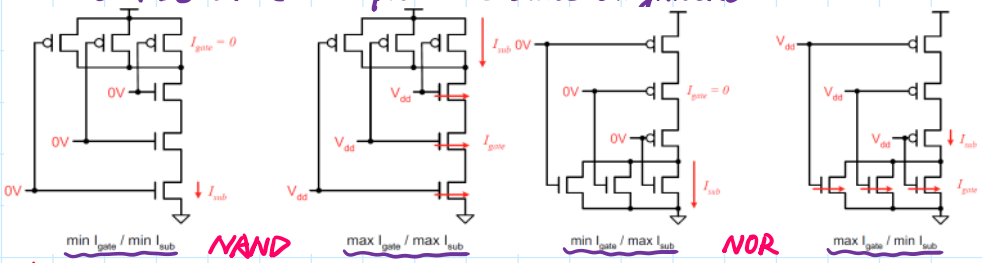
Gate leakage is significant when channel is on.
e.g. Igate reduces with no. of on transistors.

→ $I_{gate} \approx 0$ when off or $V_D = V_S$.

110	3.804	0.000	3.804
111	28.273	19.015	47.288

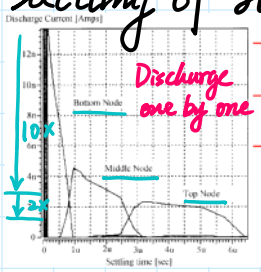
$I_{gate} \approx 0$ when on or $V_{DD} - V_{th}$.

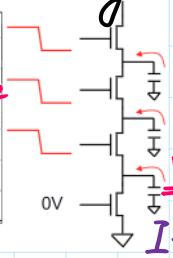
This can be used to improve the state assignment.



min I_{gate} / min I_{sub} NAND max I_{gate} / max I_{sub} min I_{gate} / max I_{sub} NOR max I_{gate} / min I_{sub}

★ Settling of Stacking:



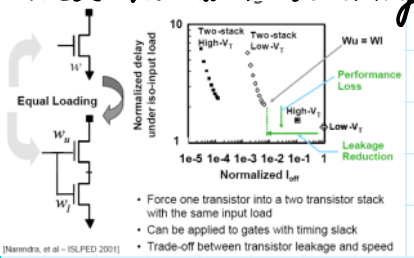


Due to the setting time, the advantage of short on/off is less.

$V_{DD} - V_{th}$ ⇒ ~80mV

$I = I_{leak, w/o\ stacking}$

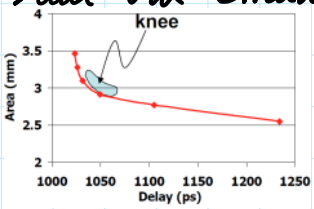
★ Forced Transistor Stacking:



use LVT but stacking

⇒ Some td↑ ⇌ Ileak↓↓

★ Dual Vth Circuit:



After the knee, by the same td reduction the Area & Ileak increase exponentially.

⇒ 1. Optimize to knee; 2. change some gate with HVT.

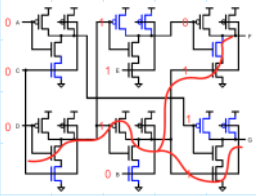
Not significantly rise td, but reduce leakage.

Mainly apply for non-critical path. And it is the only way reducing active leakage. others only for sleep leakage.

⊕ Simple to implement ⊖ only 2x reduction

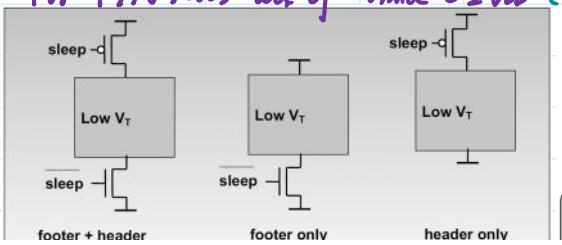
★ Combination of Vth & State Assignment:

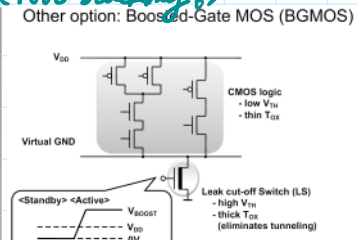
⇒ Assign HVT for leaking state.

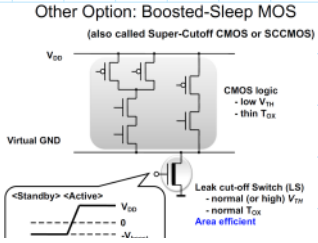


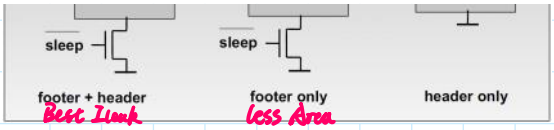
★ Power Gating: MTCMOS.

* For P/N MOS all of $V_{mid} \approx \frac{1}{2} V_{DD}$ (Not stacking!)

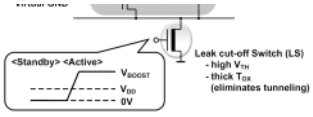


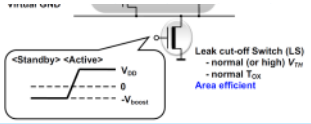




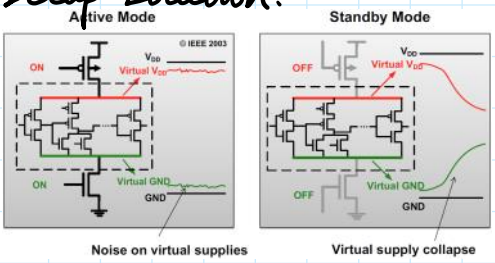


Best Ileak　less Area

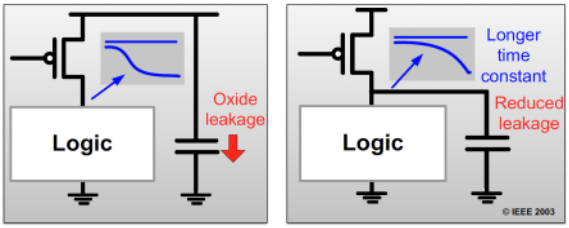




★Decap Location:

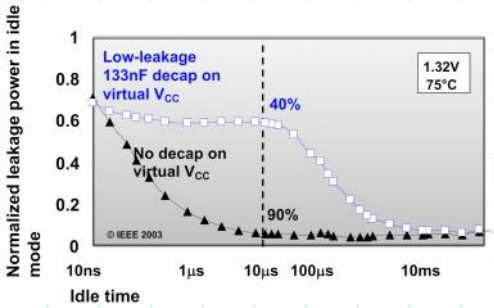


* The power gating creats virtual supply
* The header/footer can be sized as ~10% total
  ⇒ Not all the gates switching in same time



	on supply rails	on virtual rails
Performance		✓
Convergence time	✓	
Oxide leakage savings		✓

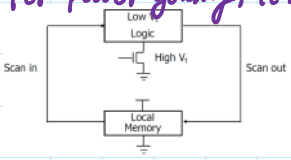
Also, for Cap on virtual rail:
the rush open may induce a supply noise
⇒ Affect other blocks. (turn-on gradully).



★ The location of Decap depends on application
⇒ On virtual rail is not effective for fast on/off.

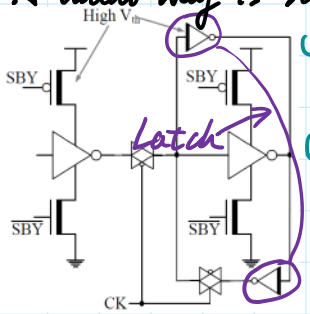
★Preserving State:

For power gating, it requires retaining the state at sleep mode.



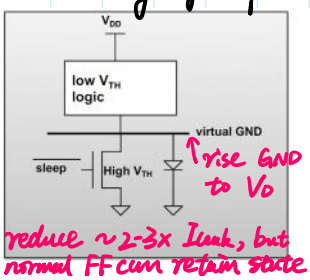
One easy way is scan out all the state
⇒ But takes a long time.

* A usual way is state retaining FFs:

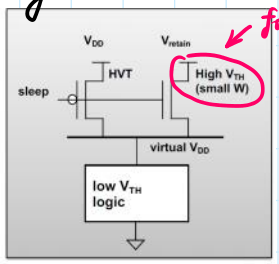


① When Sleep:
Use HVT Latch to reduce Isub
② When work:
Just a normal Latch.
But it requires both footer & header
⇒ otherwise no leakage reduction

* other ways for preventing state:

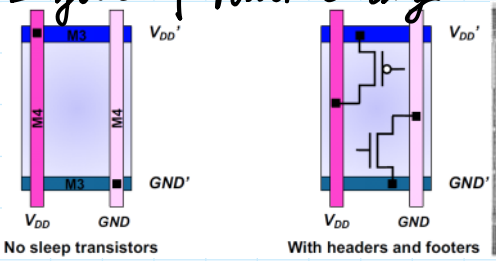


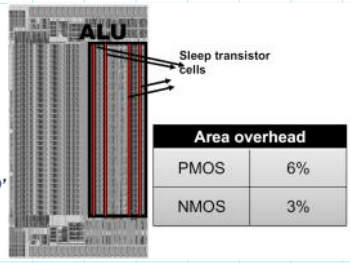
↑rise GND to V_D
reduce ~2-3x Ileak, but normal FF can retain state



for retain state

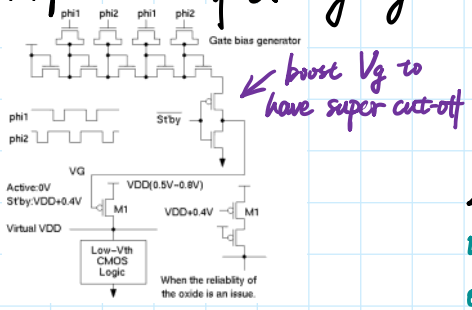
Layout of Power Gating

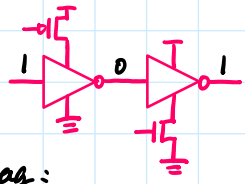




Area overhead	
PMOS	6%
NMOS	3%

Super Cut-off & Zigzag MTCMOS

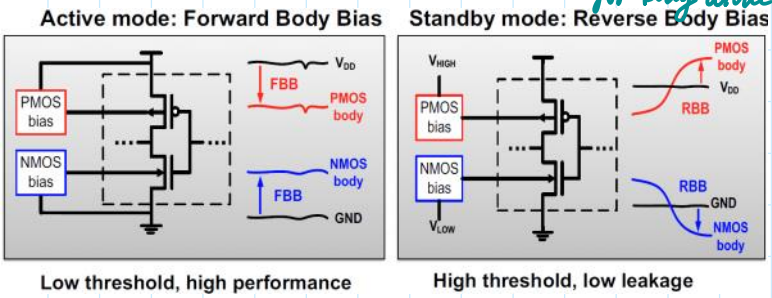


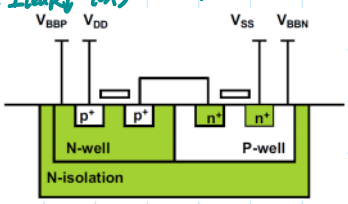


Zigzag:
use state assignment of logic only foot/head the needed one.

Dynamic Body Biasing:

No delay reduction, but need triple-well, Limited Vth adjustment, need charge body. (power↑)
(~0.1V, since K≈0.1 and Vo=0.7 for body diode, Ileak↓~10x)

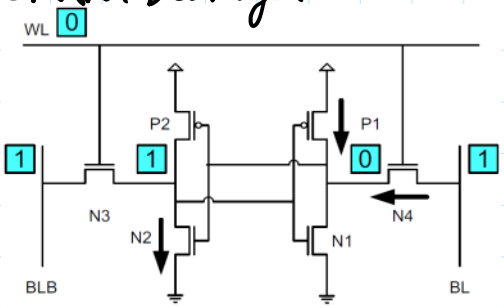




Conclusion:

	Transistor Stacking	Power Gating	Dynamic Body Biasing	Dual Vt
Pros	▪ Conventional technology ▪ No performance impact	▪ Conventional technology ▪ Conceptually simple ▪ Most effective	▪ Reuse of standard designs ▪ No performance impact	▪Works at run time ▪ No / minor performance loss ▪ No layout change ▪ extremely common
Cons	▪ Limited impact ▪ Special registers	▪ Performance impact of serial transistor ▪ Changes in design flow	▪ Triple well ▪ Slow activation ▪ Does not fare well with technology scaling	▪ Limited leakage reduction ▪ More mask layer(s)
Potential Savings	▪ 1 – 2x Very Fast	▪ 2 – 1000x Fast	▪ 2 – 20x slow	▪ 2 – 5x also works at run time

SRAM Leakage:



1. Make V_{DD} Low / V_{SS} High for latch. (Drowsy SRAM)
2. Floating BL $\Rightarrow \frac{V_{DD}}{2}$ by leakage $\Rightarrow$ Access Transistor $I_{leak}\downarrow$
3. Body Biasing
4. Underdrive WL. (~-100mV)

L9. Power Supply

2023年3月21日 0:24

☆ Power Supply Model:

For power supply of the chip:

*Important to keep it low impedence for all frequency

⇒ low R & L; more parallel Cap.

e.g. Power = P. Max ripple is r·Vdd

⇒ $Z < r\cdot(Vdd)^2/P$. For P=100W, r=0.1, $Z < 1m\Omega$

☆ Voltage Drop: IR & $L\frac{di}{dt}$

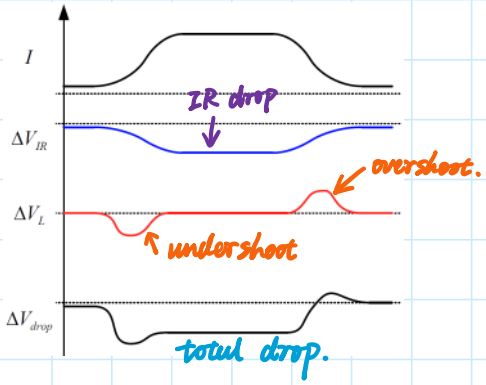
*IR drop:

$\Delta V = IR$; $P = IV$ ⇒ if P↑ ⇒ V↓ ⇒ I↑

⇒ Make ΔV larger

*$L\frac{dI}{dt}$ drop.

if freq ↑; dI↑, dt↓ ⇒ $L\frac{dI}{dt}$↑↑↑



☆ Power Supply Noise:

① Sustained drop: $td \propto \frac{CV}{I} \propto \frac{1}{(Vdd-Vss-Vt)}$

For ΔV at Vdd ⇒ Circuit delay td↑

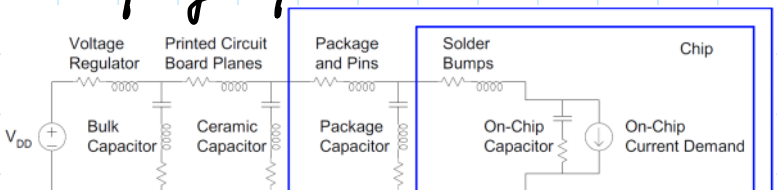
② AC Noise: Functional / Delay

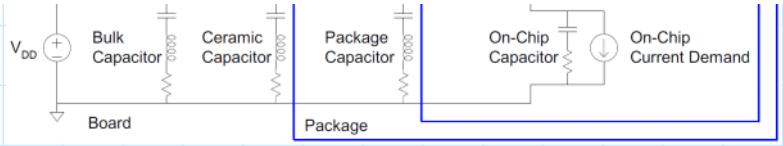
may have ringing.

③ Over/under shoot: Exacerbated TDDB

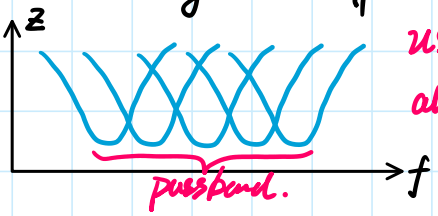
Also limited $V_{DD,max}$ to be small (margin for over/under shoot)

☆ Decoupling Caps:

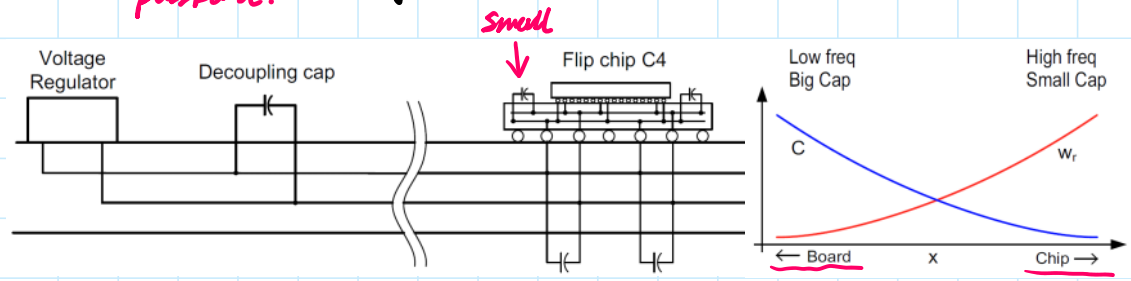




There are different decaps ⇒ Large C comes with Large R_{ESR} & L_{ESL}.
Thus, they have different ω_0.



use different decap to make the Z_{in} low at all the frequency of interests.

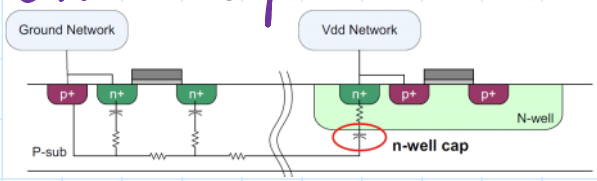


* 1. usually the cap close to the chip is small but high frequency
* 2. The decaps usually placed at where self-inductance is dominant.
i.e. $L_{decap} \gg L_{wire}$, otherwise it has no help to the circuit.

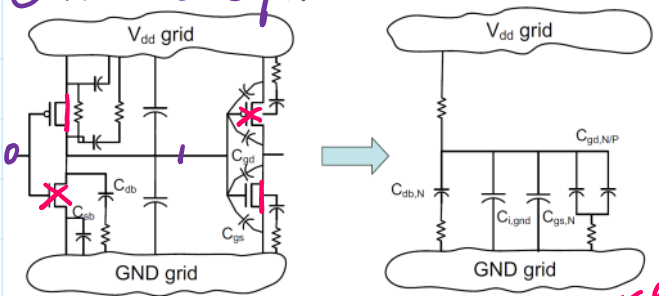
Decap in Circuit.

Implicit:

① N-well Cap:



② Parastic Caps.



Some of the cap between V_{DD} & V_{SS} can be considered as decap.

$$C_{eff} = \frac{1}{2}\left(C_{db,N} + C_{gs,N} + C_{db,P} + C_{gs,P}\right) + \frac{1}{2}\left(C_{i,gnd} + C_{i,Vdd}\right) + C_{gd,N} + C_{gd,P}$$

always there.

half chance for $V_{in} = V_{SS}/V_{DD}$

For the cap in circuit: non-switching caps are decaps.

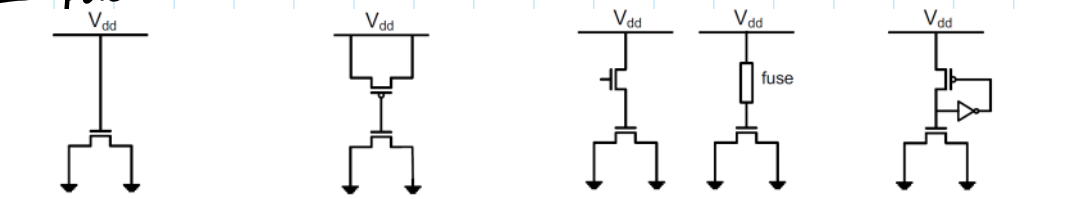
$P = f C_{switch} V_{dd}^2 \Rightarrow C_{switch} = P/f V_{dd}^2$

Due to the activity rate: $C_{switch} = s C_{total}$.

Due to the activity rate: $C_{switch} = SC_{total}$.

$$C_{decap} = (1-S)C_{total} = \frac{1-S}{S}\frac{P}{fV_{dd}^2}$$

Explicit:



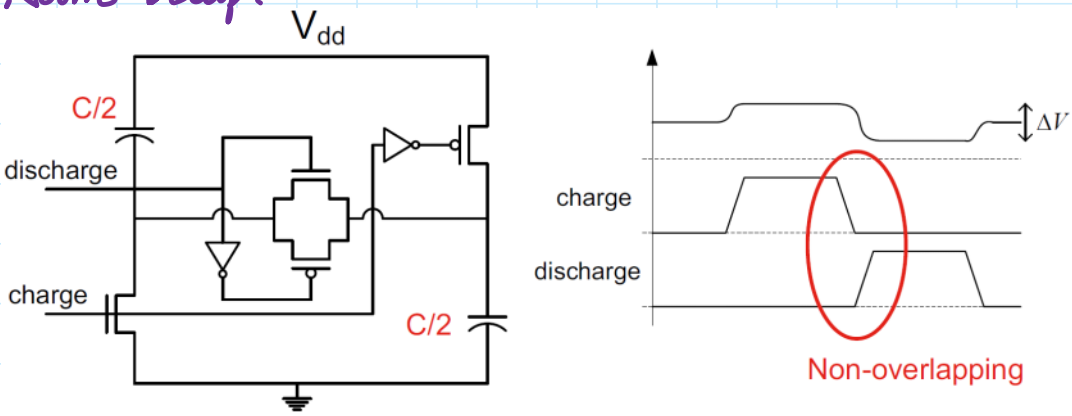
- Reliability: Thin oxide gate can short. Yield problem
- High gate leakage

- Short only if double failure
- Gate voltage is half, less sensitive to gate failure
- Less cap

- Turns off the cap in case of oxide failure
- Less gate leakage
- Unusable if the transistor fails

- Variation of previous

Active Decap:

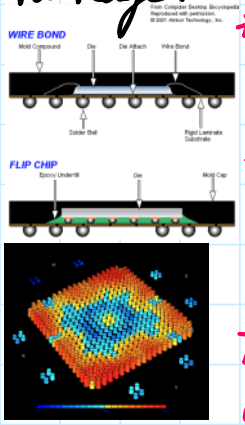


Charge in parallel; discharge in series: like charge pump.

($\frac{1}{2}V_{dd} \to V_{dd}$) ($2V_{dd} \to V_{dd}$) $\Rightarrow$ Provide more charge ($Q=CV$)

Normal Cap: $Q = \Delta VC$ $\Rightarrow$ Active: $Q = \frac{1}{2}CV_{DD} + \frac{C}{2}\Delta V$

Package:



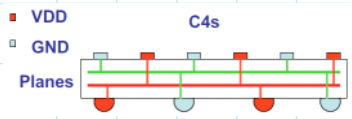
*Wire bound:

Cheap, but large L, small No. of pin (~nH) (~200)

*Flip Chip:

Less L, expensive, more pins. (~0.1nH) (~1000)

There is package induced variation due to the pin location.



Heat Dissipation:

- ΔT: temperature rise on chip

heat sink usually.

Heat Dissipation:

$\Delta T = \theta_{ja} \cdot P$

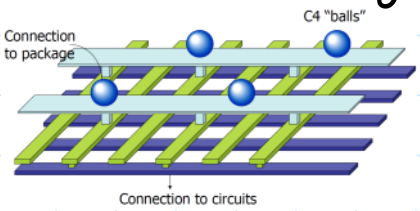
- ΔT: temperature rise on chip
- θ_{ja}: thermal resistance of chip junction to ambient — can in series/parallel.
- P: power dissipation on chip
- Ohm's Law for heat

heat sink usually.

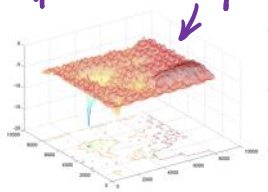
e.g. $\Delta T = 45°C$, $\theta_{pack} = 1°C/W$, $\theta_{chip} = 0.5°C/W$

$\Rightarrow P = 45/(1+0.5) = 30W$.

Chip Power Delivery:



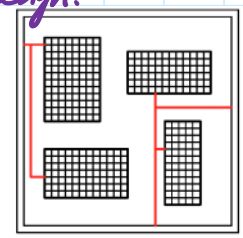
Different Vdd drop



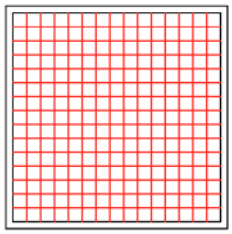
Can use top layer metal for power

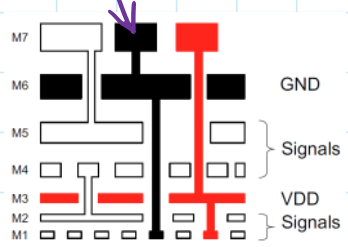
Power grid design:

Tree
- Local grids
- Simple
- Metal utilization is efficient
- Single wire failure

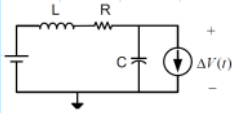


Grid
- High end processor
- Lots of metal
- Lower resistance
- Lines up with C4
- Less J, better electromigration
- Spatial variation is lower
- Highly redundant



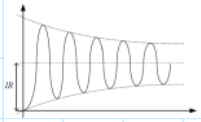


Power Grid Model:



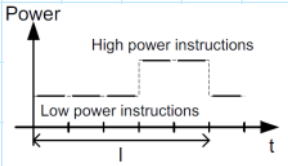
For the change of I: $\Delta V(t) = IR + I\sqrt{\frac{L}{C}} e^{\frac{-Rt}{2L}} \cdot \cos(\omega_r t - \theta)$

$\Rightarrow$ IR drop + resonate LC



$\Delta V_L \propto I\sqrt{\frac{L}{C}}$ $\Rightarrow$ decap reduce $\propto \sqrt{C}$; Id reduce $\propto$ Id.

R gives IR drop, but helps dump the resonace.



$\omega_r \neq \frac{k\omega_{clk}}{l}$ k→harmonic, l→loop

Loop of on/off. instruction.

Also, the LC may resonate.

For pervious, $W_0 > W_{clk}$, the harmonic may induce resonance

Now, $W_{clk} > W_0$; harmonic is not a problem but the on-off of the blocks.

Solution:

Reduce R: Use more metal. Tree → Grid → Plane

$\Delta V_{IR} \downarrow$ but less damping $Q \uparrow$

Reduce L: Thin package, bondwire, flip-chip. More pads

$\Delta V_L \downarrow$ $\omega_r \uparrow$ $Q \downarrow$ difficult to control

Increase C: Decoupling capacitance, but only $\sqrt{C}$

$\Delta V_{IR} \downarrow$ $\Delta V_L \downarrow$ $Q \downarrow$ Area $\uparrow$

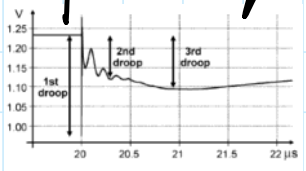
Decrease I: Turn modules on slowly, larger reset latency

$\Delta I \downarrow$ $\Delta V_{IR} \downarrow$ $\Delta V_L \downarrow$

Decrease f: turn modules on slowly, larger reset latency

$\Delta I \downarrow \quad \Delta V_{IR} \downarrow \quad \Delta V_L \downarrow$

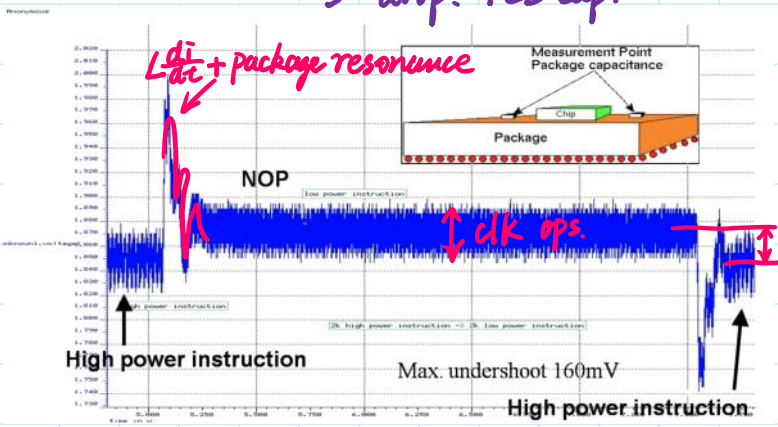
☆ Different Frequency Drops:



1st droop: decap (on-chip)

2nd droop: package cap.

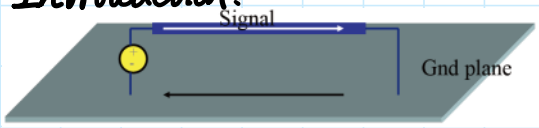
3rd droop: PCB cap.



L10. Inductance

2023年3月27日 0:40

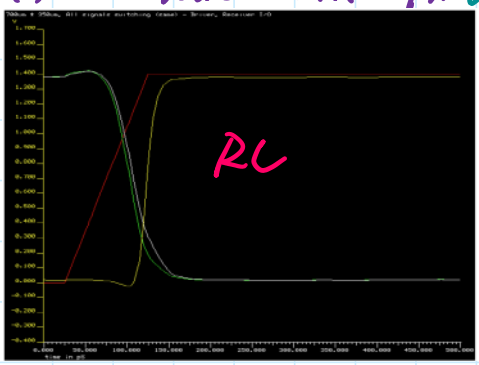
Introduction:

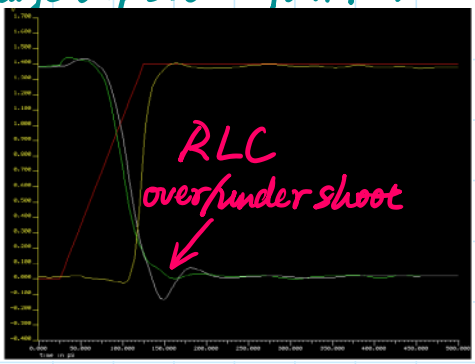


for inductance:

(1) Self: L & Mutual: M

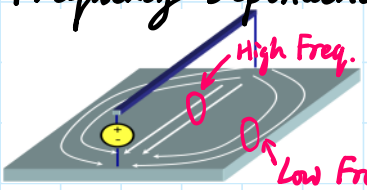
(2) Must have return loops: Large loop size ⇒ flux ↑ ⇒ L ↑.





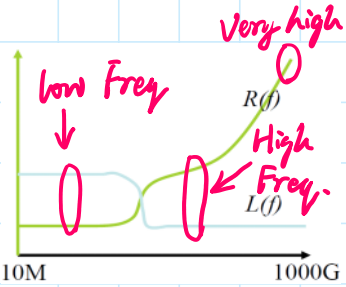
⇒ Important for high speed IC chip.

Frequency Dependent:



For normal case: $Z = R + j\omega L$

but in chip with high freq, the L also change with f.



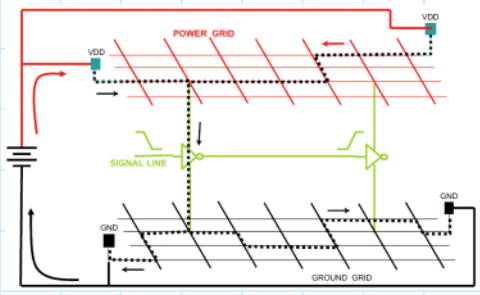
(1) Low Freq: The current is wide spread: L ↑, R ↓

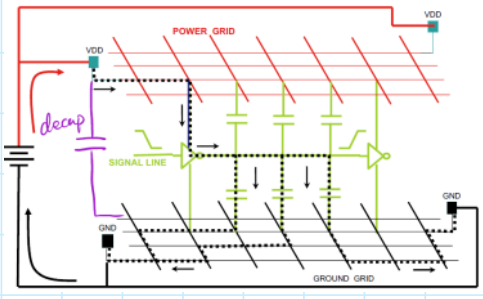
(2) High Freq: The current is limited in a narrow path under the wire: L ↓ R ↑

(3) Very High Freq: The skin effect makes R even larger.

← current only flow at at the surface.

Current in the Circuit:



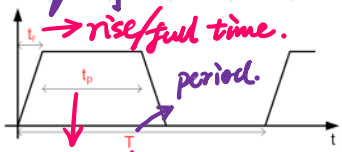


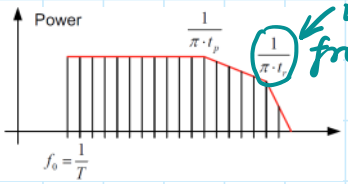
Flow through power grid, also with decaps.

⇒ More decaps helps the current response.

Inductance in Circuit:

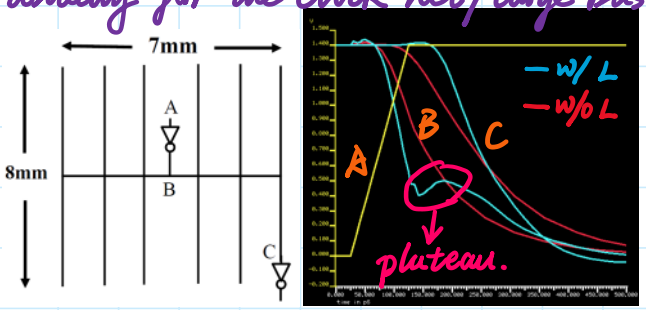
Freq. of interests:





Switching with RLC:

usually for the clock net/large bus:



With L: at C

just switch but more delay

⊖ More noise

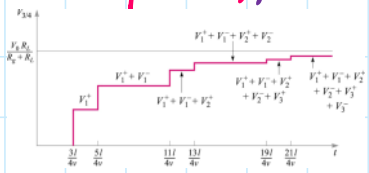
⊕ Less Ishort, fast switch.

For the circuit with L modelled: pluteau.

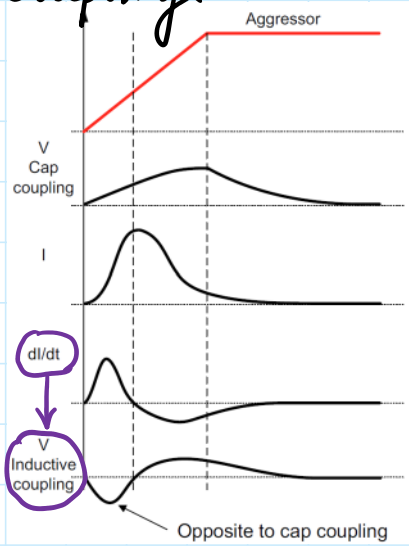
⇒ Current in L with step response will rise gradually.

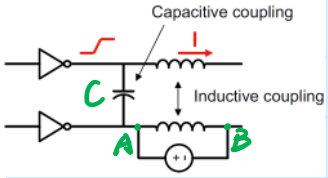
like open first (fast switch to a pluteau); then drop (I↑)

Also can be explained by DC response of a Tx line:



Coupling:

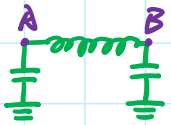




Two parallel wire may have inductive coupling:

This coupling only provide ΔV,

but the reference Vo should be found.

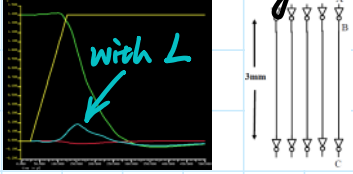
e.g. 

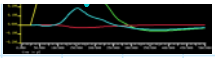
$C_1 = C_2 \Rightarrow V_A = -\frac{1}{2}\Delta V;\ V_B = \frac{1}{2}\Delta V$

$C_1 = C,\ C_2 = 0 \Rightarrow V_A = -\Delta V;\ V_B$ not change.

only C_1 provide current.

For bus, there may be crosstalks

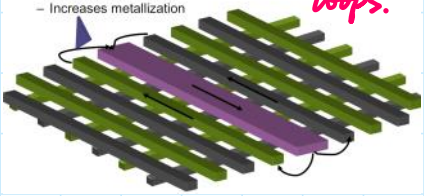




Inductance Avoidance.

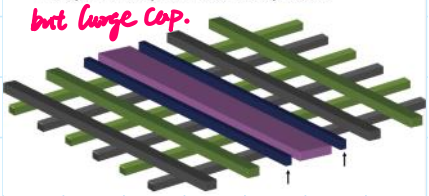
① Denser power/ground grid
- Close-by return paths
- Increases metallization

small return loops.

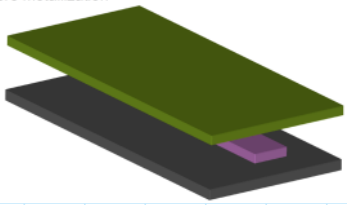


② Sandwich signal line between power/ground shields
- Keep the return paths as close and possible

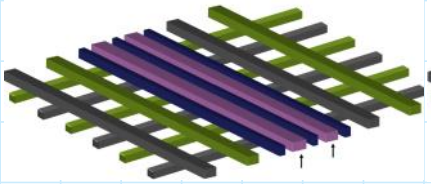
but large cap.



③ Reduction of inductance at low and high frequency
- More metallization

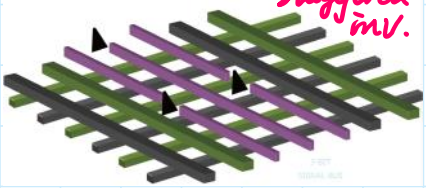


④ Split wide conductor into multiple wires with shields in between. [Massoud '98]

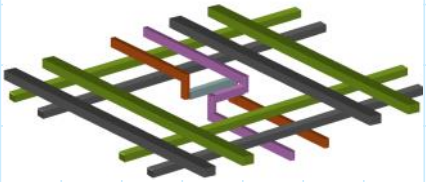


⑤ Reduce the overlap between adjacent wires
- Capacitive and inductive coupling reduced

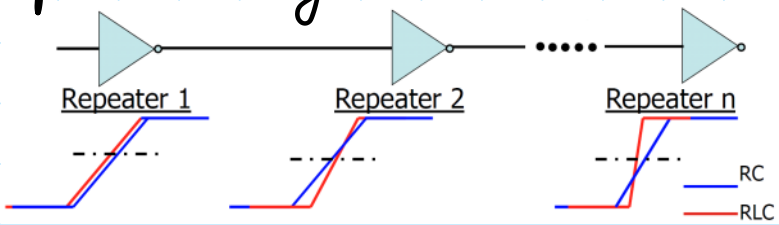
Staggered inv.



⑥ Nets routed to create opposite loops [Zhong '00]
- Magnetic flux from adjacent nets cancels out



Optimal Delay:

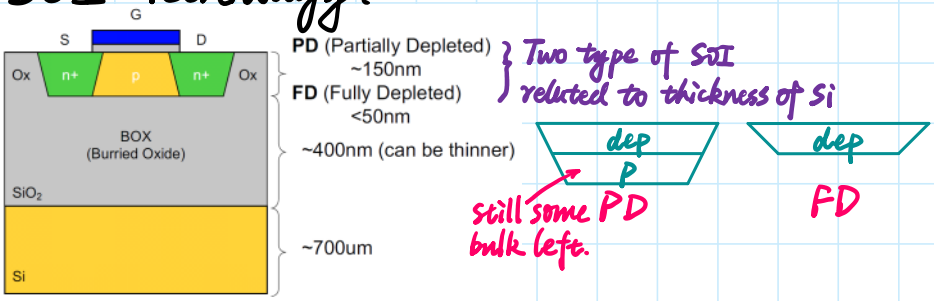


The inductance can be opt. to achieve a low delay. by fast slew

L11. SOI

2023年4月5日 20:15

☆SOI Technology:



Two type of SOI related to thickness of Si

still some PD bulk left.

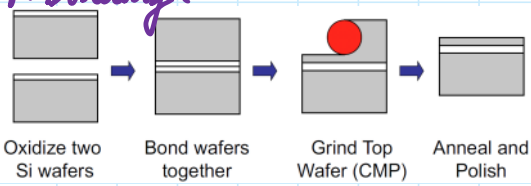
FD

*Main Features:

1. No junction leakage, less drain/source Cap – fast & Low power
2. Radiation hardening (SER)
3. PD: Floating Body: Hysteresis, parastic BJT, leaky.
   FD: No Body: like a bulk device.
4. Thermal Isolation: SiO_2 has 2 order lower thermal conductivity.
   ⇒ Not significant globally, but can cause local hot point.

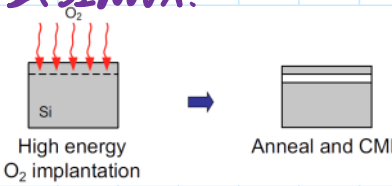
☆Fabrication:

1. Bonding:



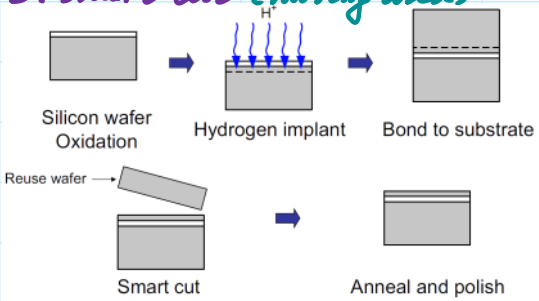
Expensive since need two wafers.

2. SIMOX:



O_2 implantation cause damage of surface.

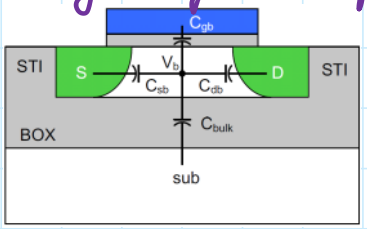
3. Smart Cut (mostly used)



☆PD SOI Device

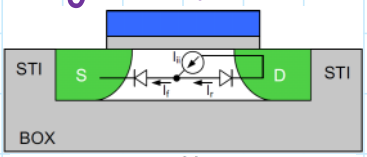
PD SOI Device

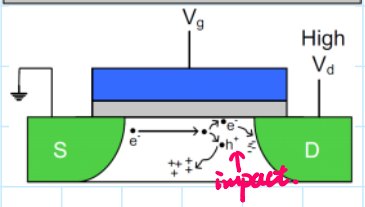
Body Charge – AC Caps: For fast changes (switching)



1. Change V_b
2. C_{gb} only there when no channel.
3. C_{sb} is voltage dependent.

Body charge – DC: slow change (leakage)

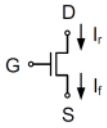




1. SB forward biased current (~nA, in nS)
2. DB reverse biased current (~pA, in uS)
3. Tunneling current of gate (in uS)
4. Impact Ionization: Rise the body voltage.
5. Thermal generation/recombination

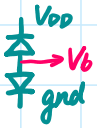
In actual case: AC+DC

Body Voltage in DC:



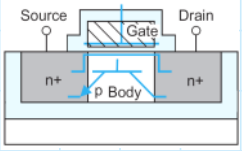
	G	S	D	V_b	V_t
①	0	0	0	0	V_{t0}
②	0	0	V_{dd}	0.35V	Reduced ← reduced by $V_b>0$
③	0	V_{dd}	V_{dd}	V_{dd}	small ← very small V_{ch}

For ②: There is If & Ir. If = Ir, make $V_b \approx 0.35V$

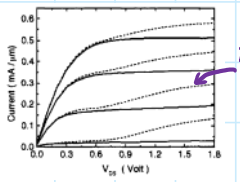


if $V_G \neq 0$, there is impact ionization $\Rightarrow V_b\uparrow \Rightarrow V_{th}\downarrow$

Parasitic BJT: at where Vb is high, but Vs drop to low



if there is a I_b (Body → Source)
it will induce a large i_c (Drain → Source)

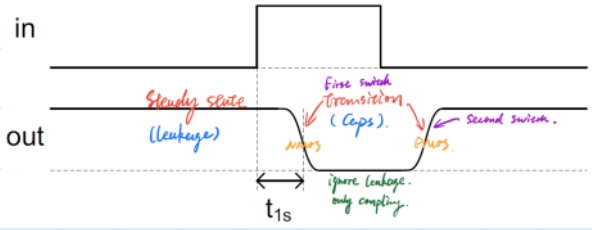


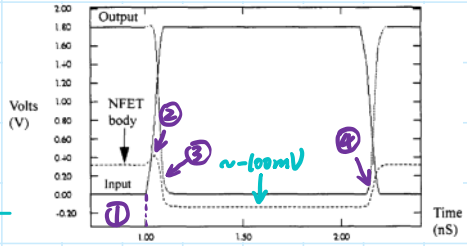
By having implant ionization & BJT
Current is higher

Switch Order: AC+DC

① NMOS First Switch:

① NMOS First Switch:





① Before switch:

Leakage (DC) set the $V_b = 0.35V$

② Switching: Cgb coupling, V_b↑

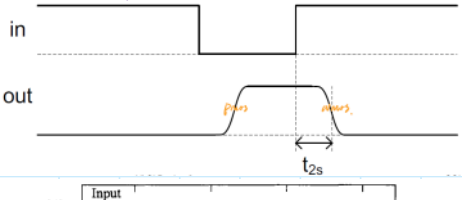
But Cgb disappear when NMOS is on.

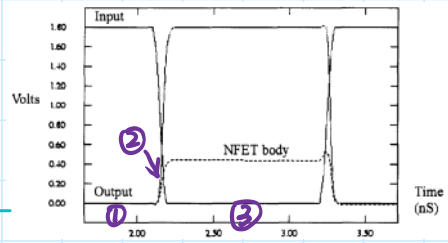
③ Switching (Vout start drop): V_b↓ (coupling)

④ PMOS switching: only Cdb coupling (+~450mV)

No Cgb coupling since channel is on at begining.

② NMOS Second Switch: PMOS switch output from L→H first





① Before switch:

Leakage (DC) set the $V_b = 0V$

② Switching (Vout start drop): V_b↑ (coupling)

Small Cgb coupling since channel is on at begining.

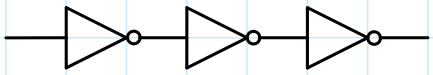
③ After 1st switch: V_b rises some amount

④ NMOS Switching: with Cgb & Cdb coupling

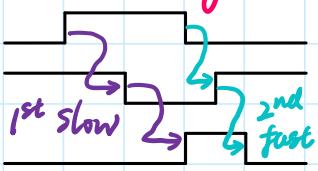
V_b = 450mV, ⇒ V_{th}↓ ⇒ fast switching.

★ The first switching is slower than switch at second (Same for PMOS)

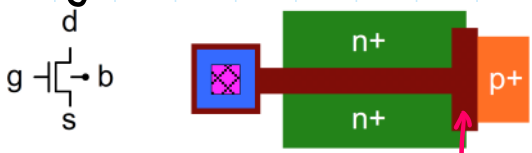
It induces a effect called Pulse narrowing



⇒ Pulse may disappear after a long propagation



★ Body Contact: Add a p^+ to the well.



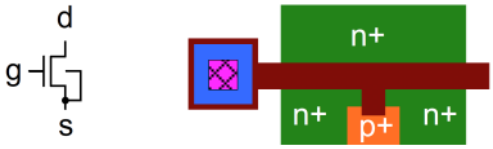
Tie body to GND

– *Slower than bulk

– But no delay variation

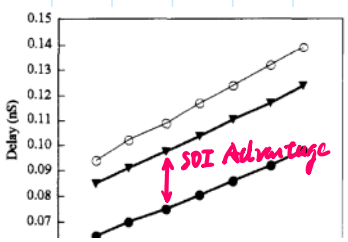
– No bipolar

isolate doping.

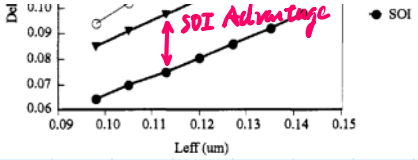


Source tie

- Increase gate cap
- Increase layout area
- Some delay variation
- No bipolar
- No body effect, V_{sb}=0



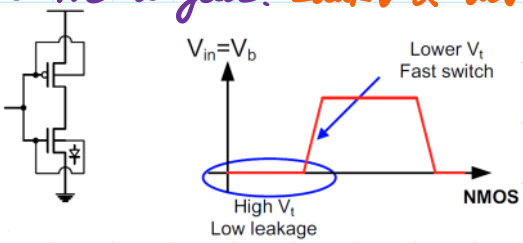
Slower than normal bulk device.



☆ Smart Body Contact:

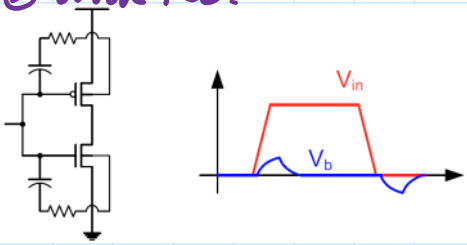
usually at the last buffer ↙ to drive heavy load.

① Tie to gate: I_{leak}↓ & t_d↓



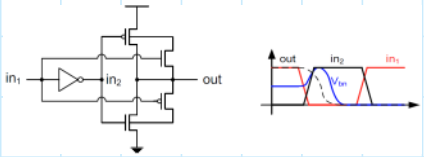
But: 1. Add gate cap.
2. Miller effect
3. RC delay of back gate
4. limited Vdd < 0.6V (forward of Diode)

② With RC:



No Vdd limitation
But hard to make

③ Smart version:



No leakage reduction
Better as a driver
For Buffers.

☆ FD SOI: No body effect.

- Thin film thickness < 50nm
- Complete depletion when channel forms
- No bipolar
- Much less hysteresis
- Better subthreshold swing ~70mV/dec
- Lower V_{th} for same I_{off}
- But: S/D resistance
  - V_{th} sensitive to Si thickness
  - High V_{th} more difficult to fabricate

e.g. UTBB-SOI from ST.

- STMicroelectronics 28nm
- Ultra-thin body/box (7nm body, 25nm box)
- Advantages
  - strong back bias effect
  - undoped channel so low RDF
- Disadvantages
  - hard to manufacture
  - Ss penalized with thin box

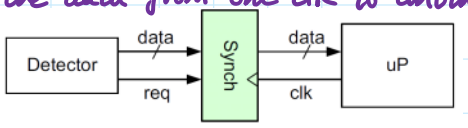
L12. Synchronization & Metastability

2023年4月6日 16:48

Synchronization:

In a system, there may be multiple clks.

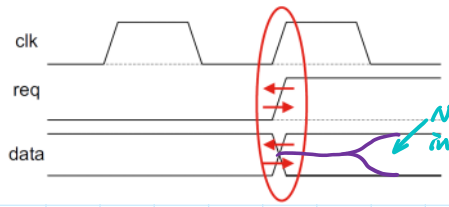
the data from one clk to another requires synchronization.



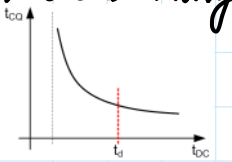
if the data/request change at the edge of clk

⇒ Induce a indeterment state.

(Due to the metastability of the FF/latch)

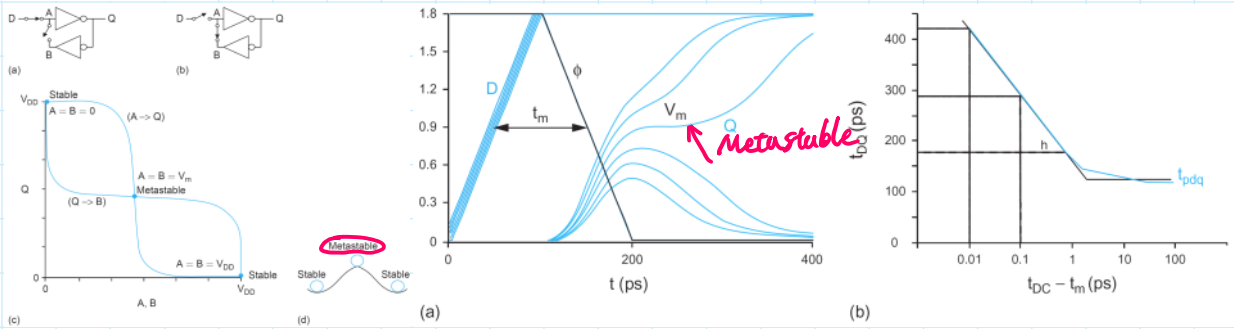


Metastability:

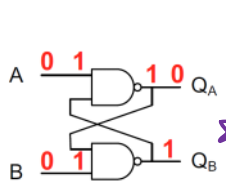


For FF/Latch, there is a setup time.

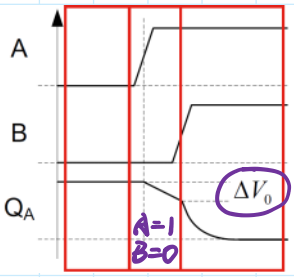
⇒ td ↑ when input is close to clk edge.



Use a crosscoupled nand as a example:



A	B	Q_A	Q_B
0	0	1	1
1	0	0	1
1	1	0	1



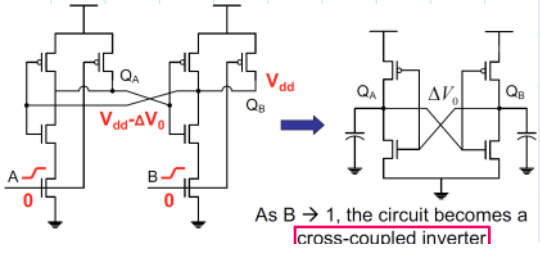
*At A=1, B=0. There is a ΔV drop by input: input data at edge.

Since $I = C\frac{dV}{dt} \Rightarrow \Delta V = \frac{I_D}{C_L}\cdot\Delta t$.

By set $t_u = I_D/C_L$ called Apperture Window.

⇒ $\Delta V_0 = \Delta t / t_u$ — the initial voltage difference.

*At A=1, B=1, The circuit latches: evaluating the input data.

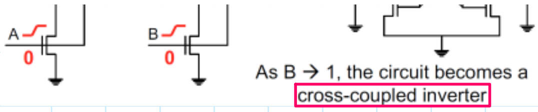


when the latch evaluating data:

$\Delta I = g_m \Delta V \Rightarrow d\Delta V = \frac{\Delta I dt}{C_L} = \frac{g_m \Delta V dt}{C_L}$

$\Rightarrow \int \frac{d\Delta V}{\Delta V} = \int \frac{\Delta I}{C_L}\cdot dt$ $\tau = \frac{C_L}{g_m}$ (speed of circuit)

$\Rightarrow \Delta V(t) = \Delta V_0 e^{(\frac{g_m}{t_d}\cdot t)} = \Delta V_0 e^{(\frac{t}{\tau})}$



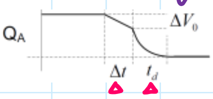
$\Rightarrow \int \Delta V = \int C_L\, dt$ ... $\tau = g_m$ (speed of circuit)

$\Rightarrow$ ★ $\Delta V(t) = \Delta V_0 e^{(\frac{g_m}{C_L} \cdot t)} = \Delta V_0 e^{(\frac{t}{\tau})}$

Equation to find the developed voltage in time t.

Define a resolving time t_d for $\Delta V(t_d) = 1$ — the unit voltage that means fully developed.

$\Rightarrow$ ★ $t_d = -\tau_r \cdot \ln(\Delta V_0) = -\tau_r \cdot \ln(\Delta t / t_a)$ $\Rightarrow$ fast circuit, less t_d

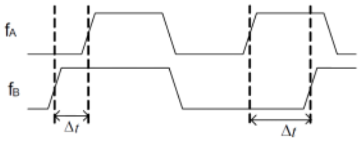


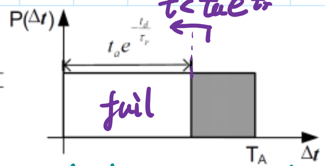
if $\Delta V_0 = 0$ or $\Delta t = 0 \Rightarrow t_d \to \infty$: The data is defined by noise

★ Probability of failure:

Define $P(\Delta V(t_d) < 1)$ as failure rate.

for $t_d < t_a e^{\frac{-t_d}{\tau_r}}$, the sync fails.





★ $P(\Delta V(t_d) < 1) = P(\Delta t < t_a e^{\frac{-t_d}{\tau_r}}) = t_a e^{\frac{-t_d}{\tau_r}} / T_A = f_A t_a e^{\frac{-t_d}{\tau_r}}$

The Δt is randomly given at the range of faster clk (clock A). The area of T_A that less than $t_a e^{\frac{-t_d}{\tau_r}}$ means fail.

The fail rate depends on the faster clock.

The sync event happens every clock of the slower clock (clock B)

$\Rightarrow$ ★ Frequency of Failure $= f_A \cdot f_B \cdot t_a e^{\frac{-t_d}{\tau_r}}$

★ Mean Time Between Failure (MTBF): $e^{\frac{t_d}{\tau_r}} / (f_A \cdot f_B \cdot t_a)$

e.g.

f_A = event frequency on line A = 50MHz
f_B = event frequency on line B = 300MHz
t_a = apperture of flop = 100ps
t_d = Wait time until decision = 10ns
τ_r = regeneration constant of flop = 200ps

Use higher frequency

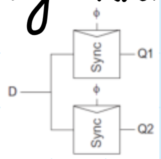
$P_{failure} = t_a f_B e^{-\frac{t_d}{\tau_r}} = 100ps \cdot 300MHz \cdot e^{-\frac{10ns}{200ps}}$

$= 5.786 x 10^{-24}$

$MTBF = 1/f_{failure} = 1/P_{failure} f_A = 1/5.786 x 10^{-21} \cdot 50MHz = 3.45 x 10^{15}$ sec

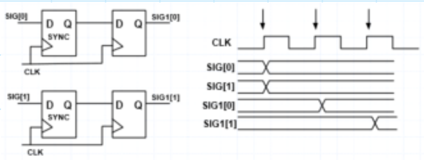
$= 10^8$ years $\Rightarrow$ Not too good (x bus no. x chip no.)

★ Synchronization Pitfall:



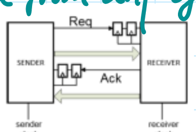
1. Not sync one signal in two places.

$\Rightarrow$ May have different result.



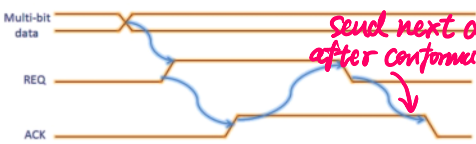
For bus, only use one synchronizer

$\Rightarrow$ if use for every bit, can get data from diff cycle.



★ Correct Way: Handshaking

- Sender outputs data and THEN asserts REQ
- Receiver latches data and THEN asserts ACK
- Sender deasserts REQ, will not reassert it until ACK deasserts
- Receiver sees REQ deasserted, deasserts ACK when ready to continue



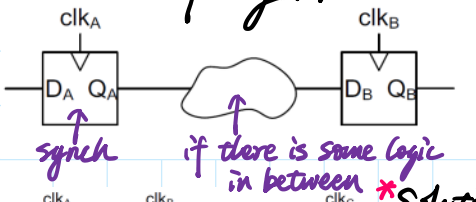
★ Fix Failure:

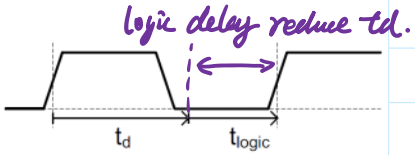
- Smaller cap $t_a = \frac{C_L}{I_d}$
  - Shield load cap at output of flops
- Wait longer
  - Application-specific, latency ok?
- Lower frequency
- Faster gates
  - Technology scaling helps but freq ↑ too
- Add completion logic
- Add more sampling FFs
  - Multi-stage synchronizers

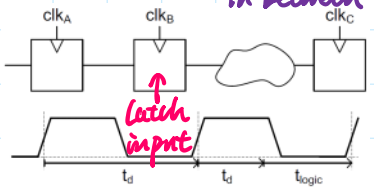
– Wait longer
  • Application-specific, latency ok?
– Lower frequency

– Add completion logic
– Add more sampling FFs
  • Multi-stage synchronizers

More Sampling FF:

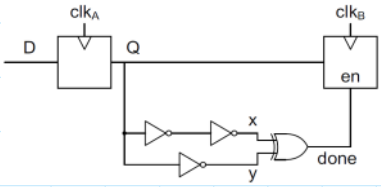


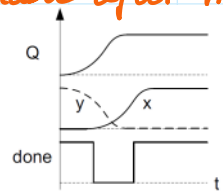




*Solution: use FFB to sample synch.
⇒ No logic in between ⇒ Mux td.

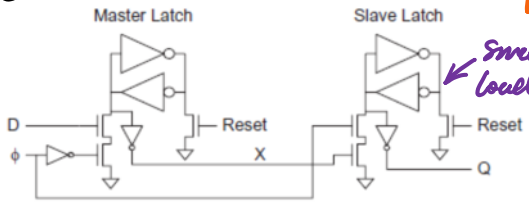
Completion Detector: Enable after transition





But logic itself can be metastable
⇒ Metastability cannot be permanently fixed.

Jumb Latch: Small load Cap. fast.



Small load Caps
But need reset.