Circuit Techniques for Leakage Power Reduction

Power Dissipation in CMOS

- Dynamic P.D. $\mathbf{P_{sw} = C_L V^2_{DD} f_{CLK} \alpha}$
- Static P.D. : $\mathbf{P_{static} = V_{DD} I_{off}}$
- $I_{sub} = \mu_0 C_{ox}(W/L)v_T^2 \, exp\{(V_{GS} - V_{TH} + nV_{DS})/\eta v_T\}$

Power Dissipation in CMOS

- If a high α is assumed, then lower V_{DD} and V_{TH} would be chosen to reduce the dynamic power dissipation and to provide acceptable noise margins, respectively
- However, if α is smaller and V_{DD} is kept same, dynamic dissipation will decrease and static dissipation will remain the same.

Power Dissipation in CMOS

- Since $\mathbf{P_{sw} \propto V^2_{DD}}$ $\Rightarrow$ Reduce V_{DD}
- If V_{DD} is reduced and V_{TH} is retained the same, the noise margin and speed will reduce
- To improve noise margins, V_{TH} must also be scaled down

Power Dissipation in CMOS

- Subthreshold leakage current increases exponentially when V_{TH} is reduced
- The resultant higher static dissipation may then offset the reduction in the dynamic portion of the dissipation
- Hence the devices need to be designed to have threshold voltages that maximize the net reduction in the dissipation

Power-Delay product – dependence on Vdd

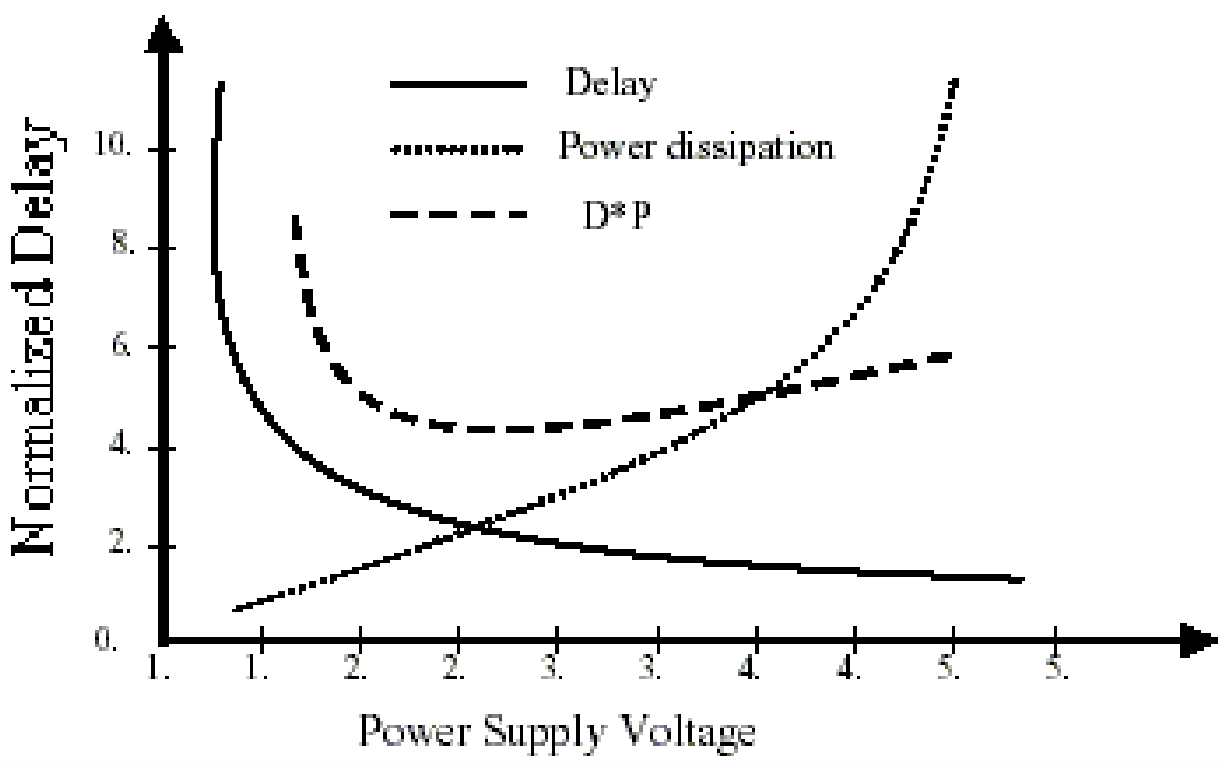
- As V_{DD} is reduced, delay increases monotonically
- To compensate for this increased delay, increase W/L ratio of transistor
- Because of interconnect capacitance, the power-delay product initially decreases and then increases when the W/L ratio is increased and V_{DD} is reduced so as to keep the delay constant
- Exploiting parallelism and pipelining :

  As V_{DD} is reduced, the degree of parallelism or the number of stages of pipelining is increased to compensate for the increased delay
- Latency and control circuitry overhead increases, and these consume power

Power-Delay product – dependence on Vdd

$$T_{PHL} = \frac{C_{load}}{K_n(V_{DD}-V_{tn})}\left[\frac{2V_t}{V_{DD}-V_{tn}} + ln\left(\frac{4(V_{DD}-V_{tn})}{V_{DD}} - 1\right)\right] \quad ...(4)$$

$$T_{PLH} = \frac{C_{load}}{K_p(V_{DD}-|V_{tp}|)}\left[\frac{2|V_{tp}}{V_{DD}-|V_{tp}|} + ln\left(\frac{4(V_{DD}-|V_{tp}|)}{V_D} - 1\right)\right] \quad ...(5)$$



Levels of Power Optimization:

- Layout Level:
  - Reduce node capacitance
- Circuit Level:
  - Static / Dynamic Style
  - Pass Transistor / Normal CMOS
  - Synchronous / Asynchronous
- Logic Level:
  - Use of automatic tools to locally transform the circuit and select realizations for its pieces from a pre-characterized library
- Higher Level: Various structural choices
  - Ripple carry / carry-look-ahead / carry select Adder
- System Level : Power modes

1. Standby Leakage control using transistor stacks (Self reverse bias):

Subthreshold leakage current flowing through a stack of series connected transistors reduces when more than one transistor of the stack is turned off

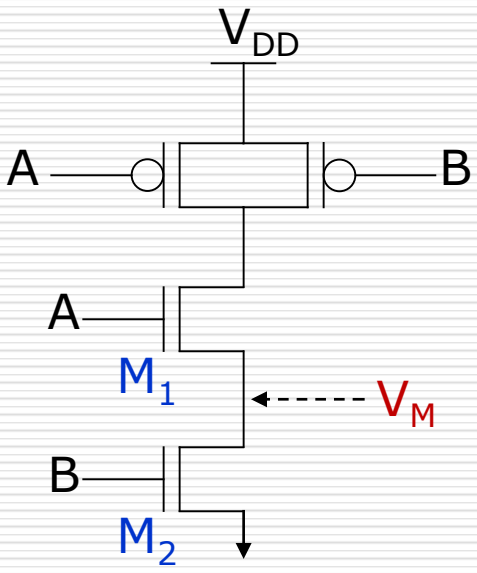
This effect is known as the “stacking effect”

Two types:

a)Natural Stacking

a)Artificial Stacking

Natural stacking



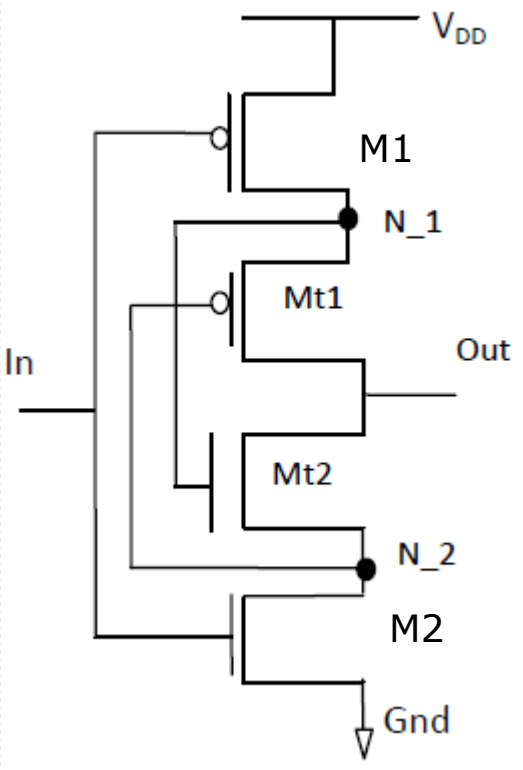
In the 2-input NAND gate shown, when both M_1 and M_2 are turned off, the voltage at V_M is more positive due to a small drain current

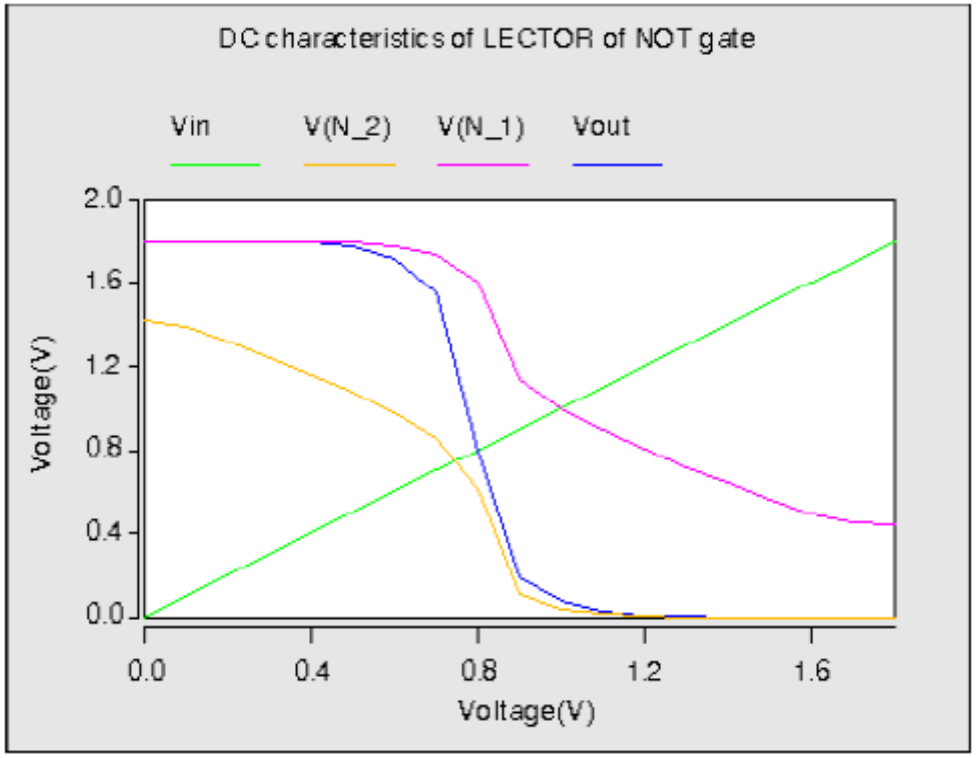
Positive potential at the intermediate node has the following effects:

- As V_M is positive, gate-to-source voltage of M_1 (V_{gs1}) becomes negative and so the sub-threshold current reduces greatly
- Due to $V_M > 0$, V_{sb1} of M_1 increases resulting in increased V_{TH} (due to body effect) and thus reducing sub-threshold leakage
- As $V_M > 0$, V_{ds1} of M_1 is lesser, resulting in increasing V_{TH} (due to less DIBL effect), thus reducing sub-threshold leakage

$$I_{sub} = \mu_0 C_{ox}(W/L)v_T^2 \, exp\{(V_{GS} - V_{TH} + nV_{DS})/\eta v_T\}$$

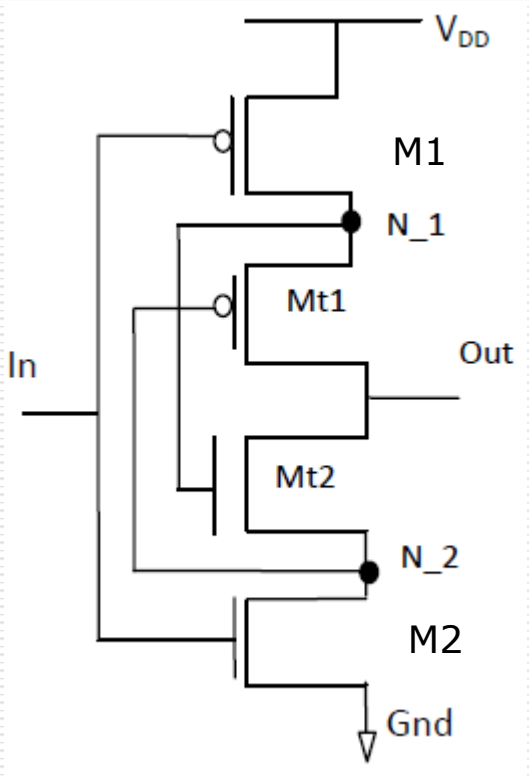
Artificial stacking - LECTOR Inverter

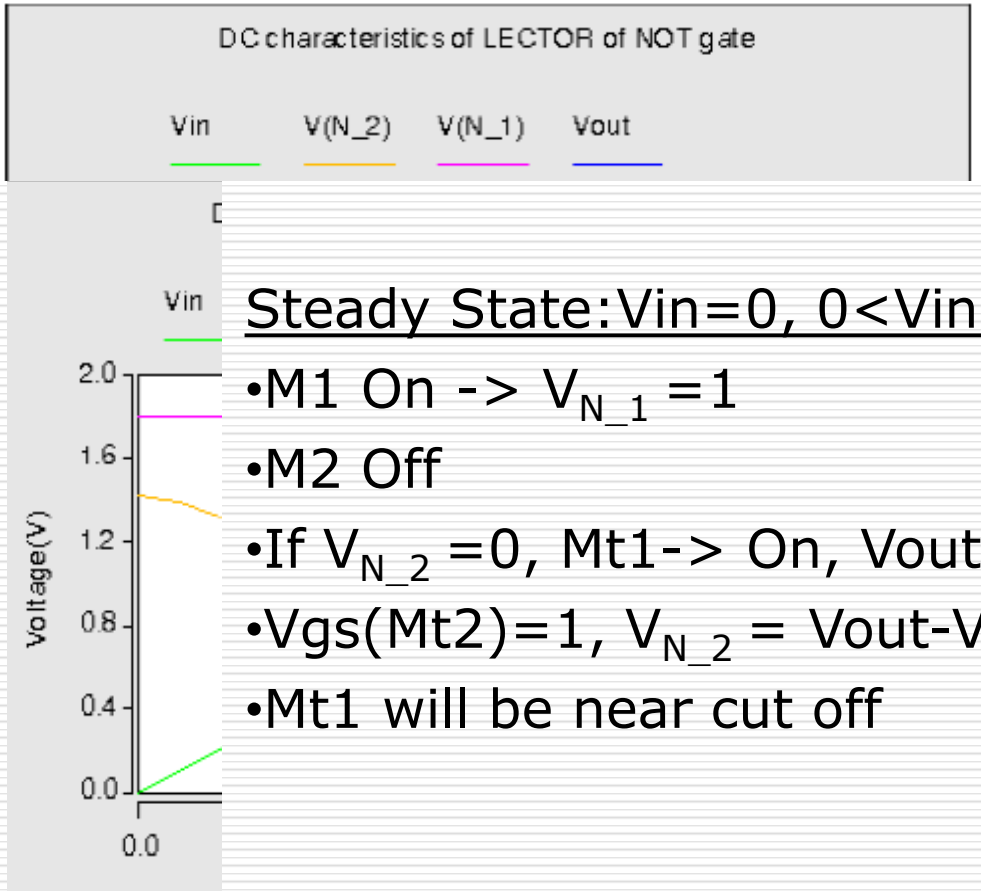




- The Leakage Control Transistors Mt1 and Mt2 are inserted between Nodes N_1 and N_2 and they act as self-controlled stacked transistors
- One of Mt1 and Mt2 will be near cut off region

Artificial stacking - LECTOR Inverter

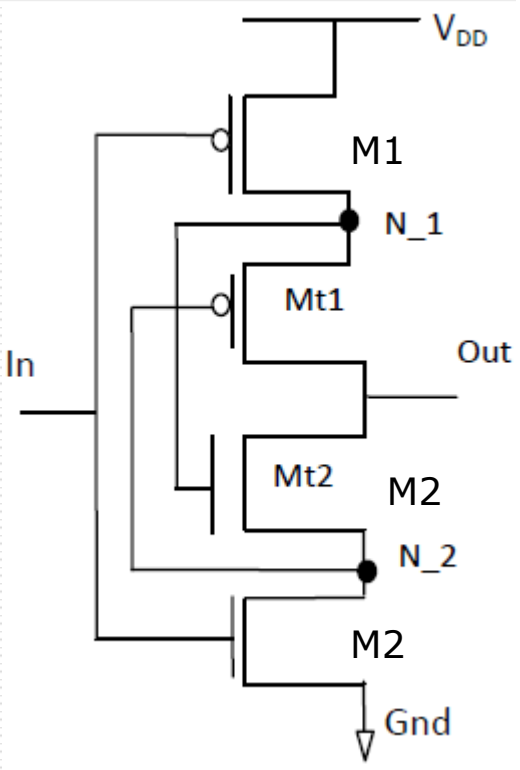


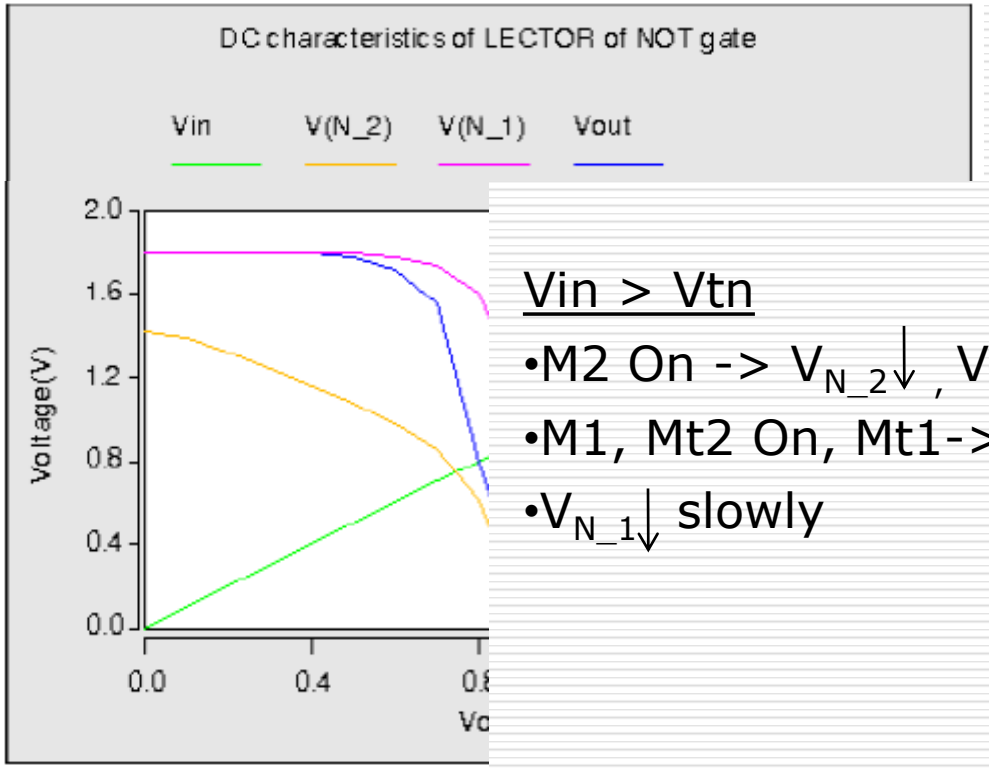


Steady State:Vin=0, 0<Vin<Vtn

- M1 On -> V_{N_1} =1
- M2 Off
- If V_{N_2} =0, Mt1-> On, Vout ->1
- Vgs(Mt2)=1, V_{N_2} = Vout-Vt
- Mt1 will be near cut off

Artificial stacking - LECTOR Inverter

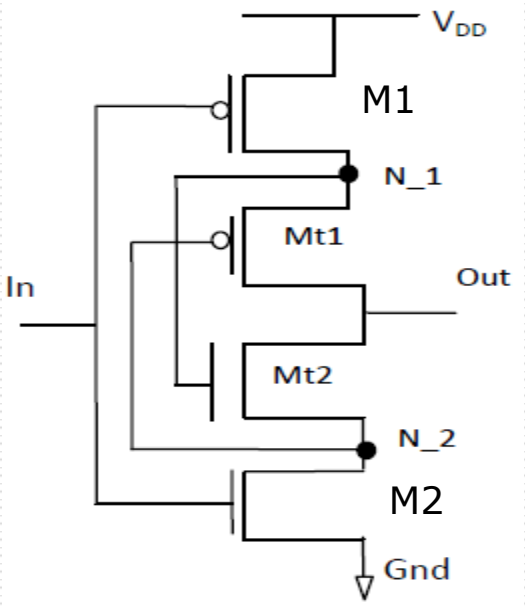


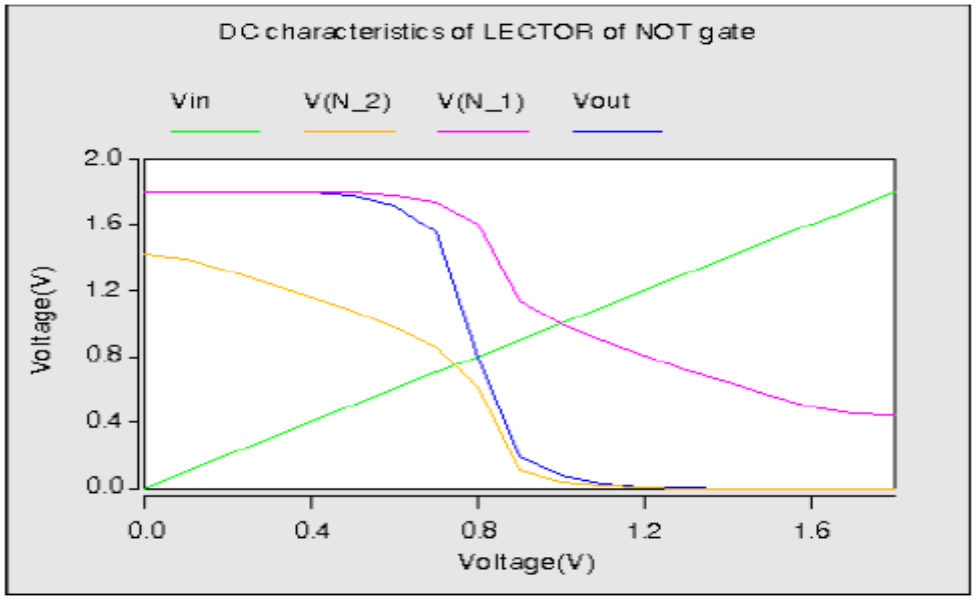


Vin > Vtn

- M2 On -> $V_{N\_2}\downarrow$, Vout$\downarrow$
- M1, Mt2 On, Mt1->on
- $V_{N\_1}\downarrow$ slowly

Artificial stacking - LECTOR Inverter

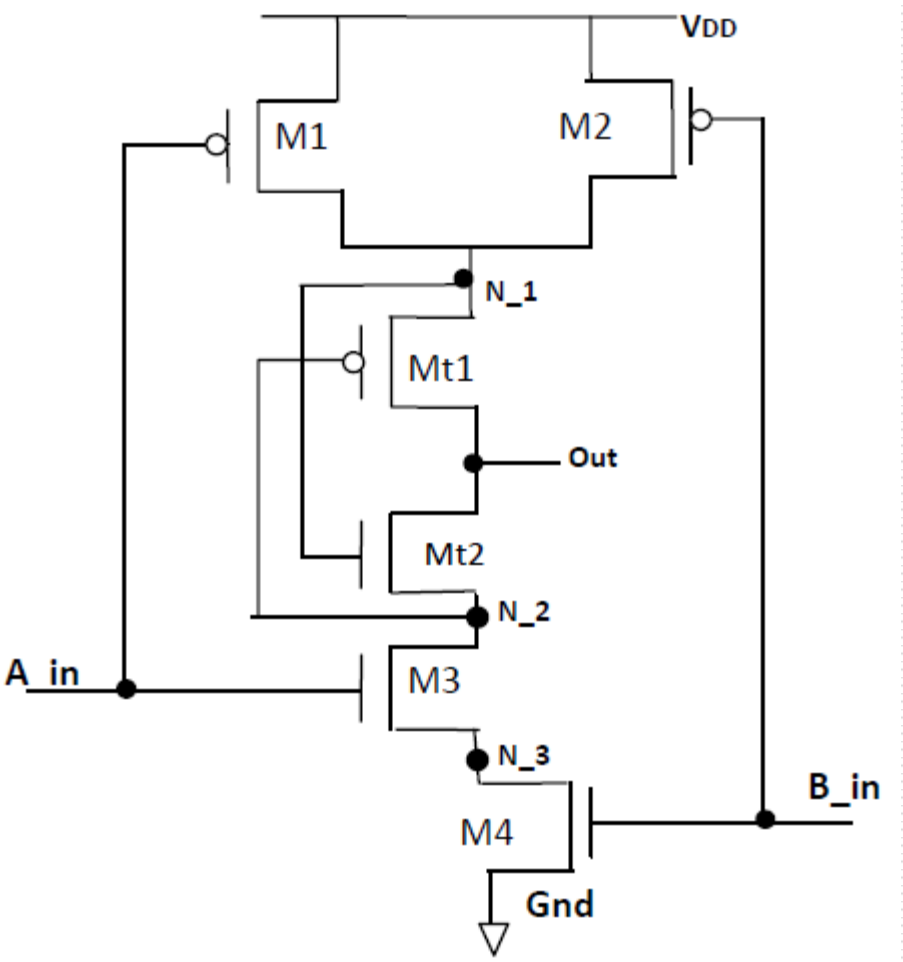


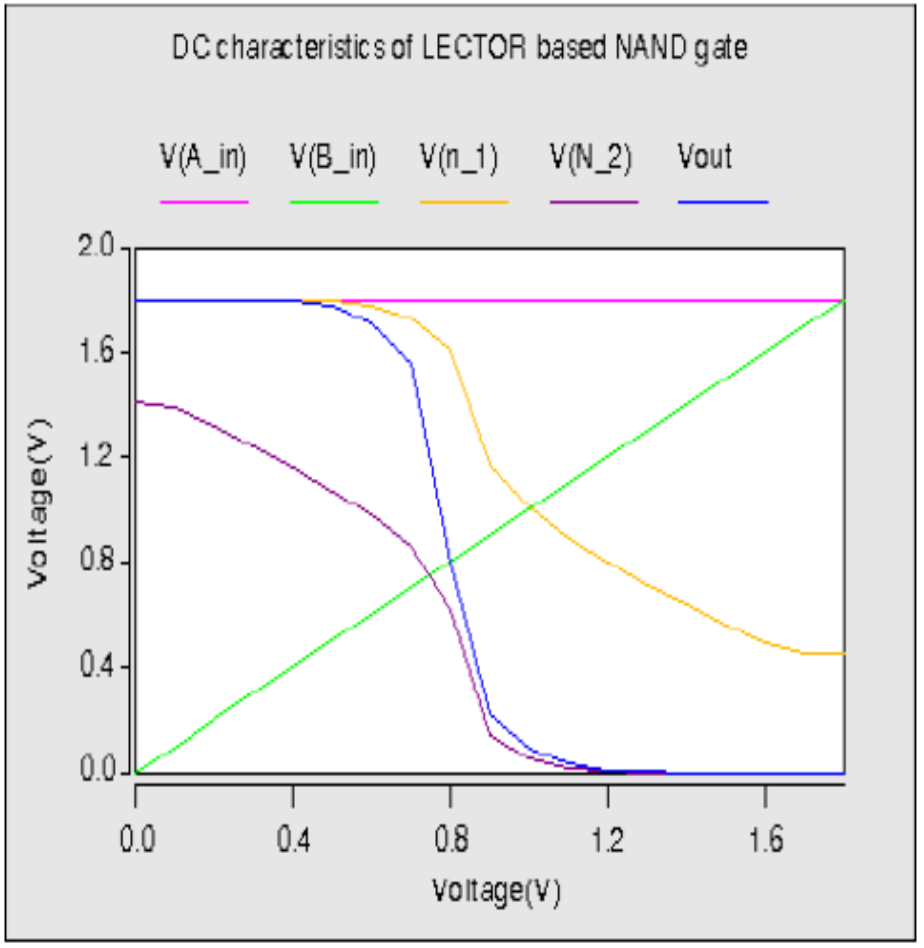


Vin > Vdd - |Vtp|

- M1 Off, M2 On
- V_{N_2} , Vout =0
- V_{N_1} =|Vtp|,
- Mt2 will be near cut off

Two input LCT NAND gate





2. Multiple V_{TH} Techniques:

- Multiple-threshold CMOS circuit has both high and low threshold transistors in a single chip

- The high threshold transistors can suppress the sub-threshold leakage current, while the low threshold transistors are used to achieve high performance

$$V_T = V_{T0} + \gamma[(2\Phi_b + |V_{SB}|)^{1/2} - (2\Phi_b)^{1/2}]$$

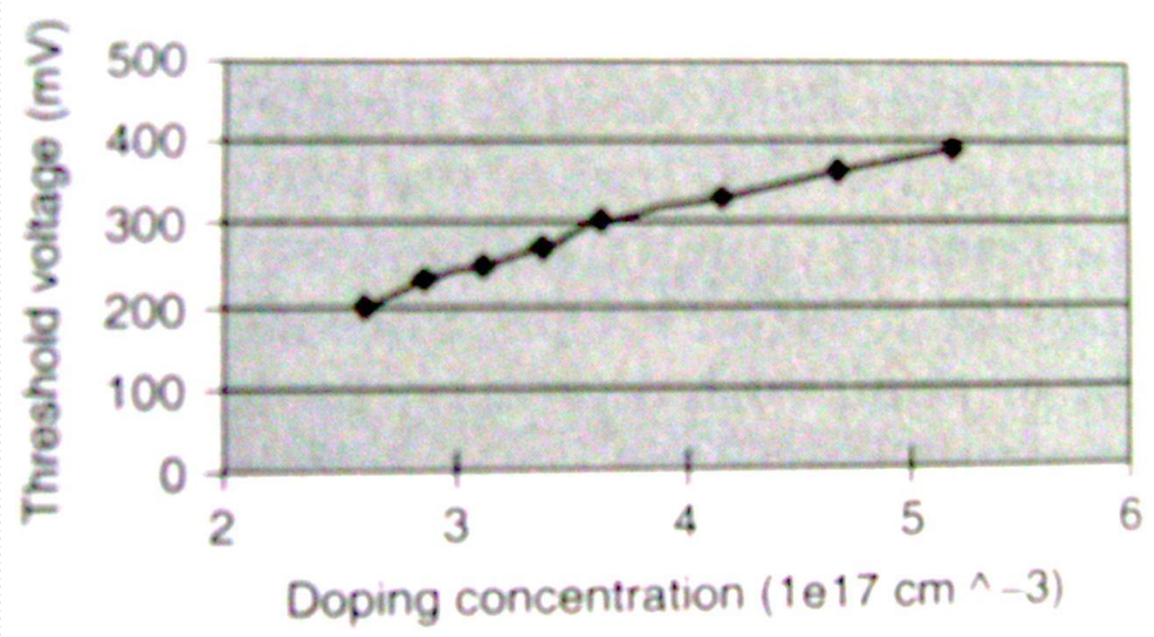
Where, $\gamma = (t_{ox}/\varepsilon_{ox})(2q\,\varepsilon_{Si}N_A)^{1/2}$ and $\Phi_b = kT/q\, ln(N_A/N_i)$;

N_i – carrier concentration in Intrinsic silicon.

Multiple threshold voltages can be achieved by:

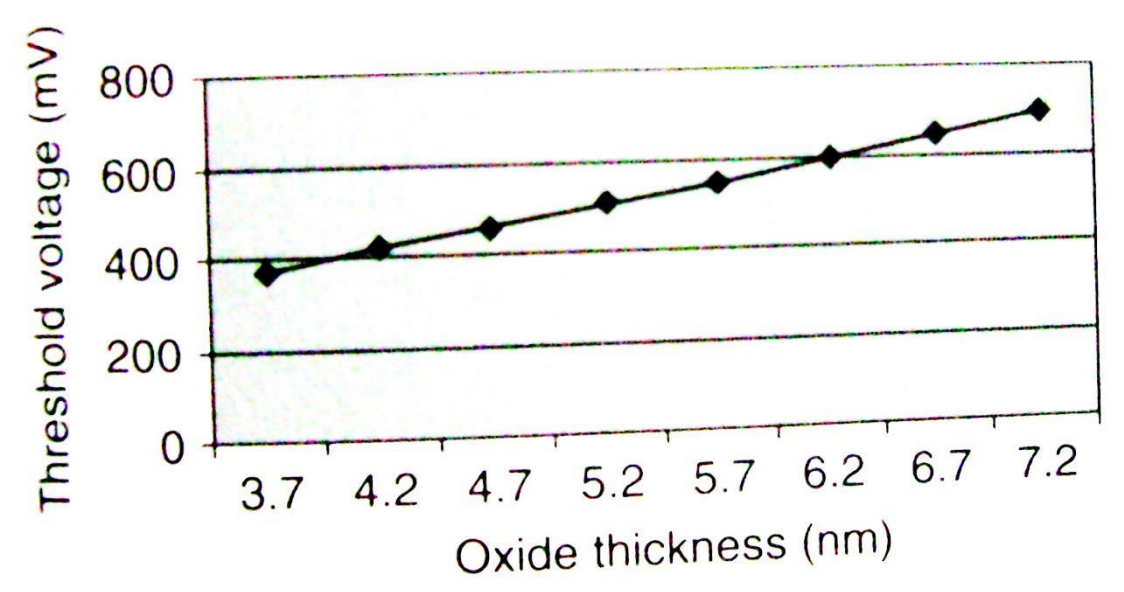
- Multiple Channel Dopings
- Multiple Oxide CMOS (MOXCMOS) Circuits
- Multiple Channel Lengths
- Multiple Body Biases
- Multithreshold-voltage CMOS (MTCMOS)
- Dual Threshold CMOS
- Variable Threshold CMOS (VTMOS)
- Dynamic Threshold CMOS (DTMOS)
- Double Gate Dynamic Threshold SOI CMOS (DGDT-MOS)

a. Multiple Channel Dopings



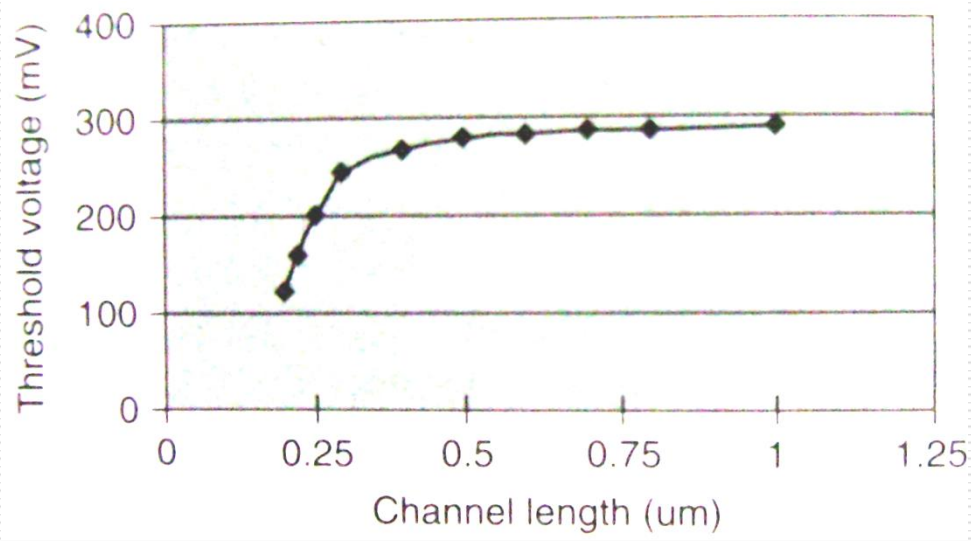
- Multiple threshold voltages can be achieved by adjusting the channel doping densities
- Additional masks are required
- Commonly used technique but difficult to achieve dual threshold voltages when the threshold voltages are very close to each other

b. Multiple Oxide CMOS (MOXCMOS) Circuit



- Gate oxide thickness (t_{ox}) can be used to modify V_{TH}
- Dual V_{TH} can be achieved by depositing two different oxide thicknesses. Higher the t_{ox} higher is V_{TH}
- In addition to the reduction in subthreshold leakage, higher oxide thickness also reduces gate oxide tunneling and dynamic power dissipation (lower C_L)

c. Multiple Channel Lengths



- For short channel devices, the threshold voltage decreases with the decrease in channel length (V_{TH} roll-off). So, different threshold voltages can be achieved by using different channel lengths
- For transistors with feature sizes close to 0.1μm, halo doping techniques have to be used to suppress short channel effects

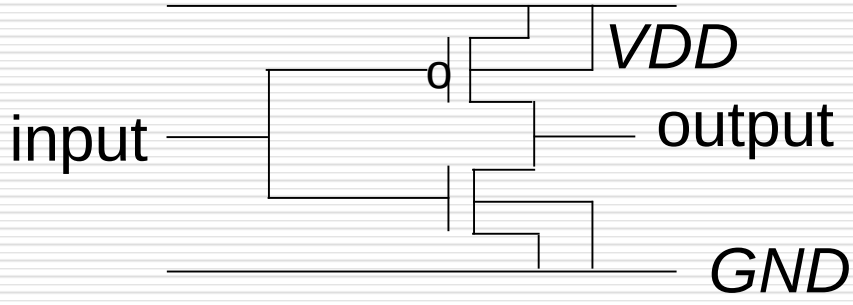
- This causes the V_{TH} roll-off to be very sharp and hence, it is non trivial to control V_{TH} near the minimum feature size.

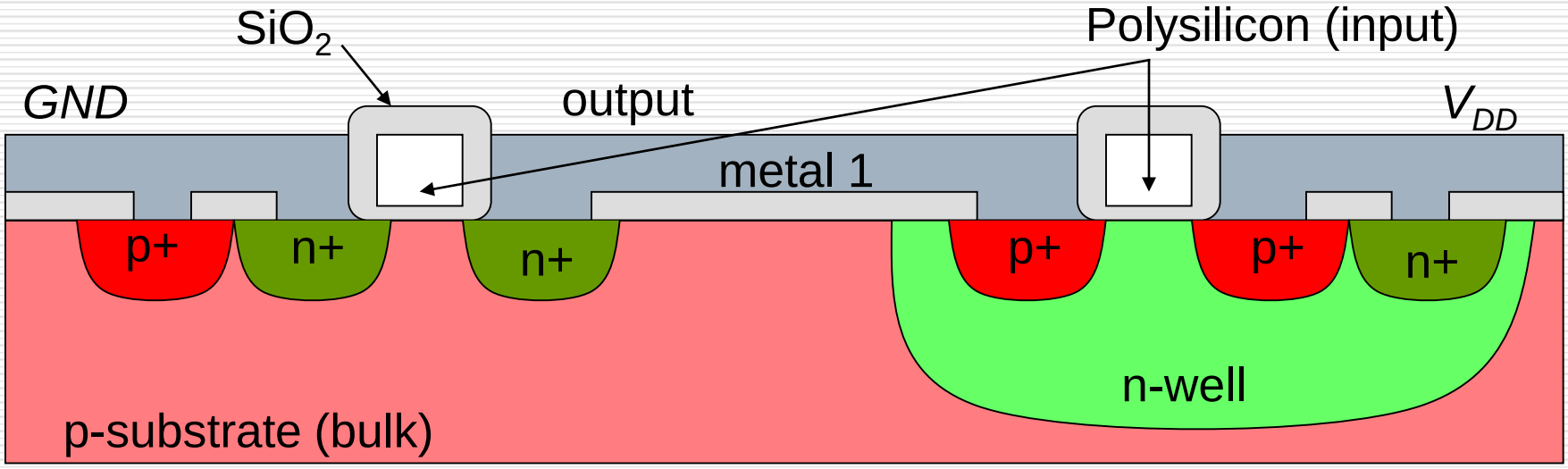
- The longer channel lengths for the high V_{TH} transistors will increase the gate capacitance, which in turn increase dynamic dissipation.

d. Multiple Body Biases

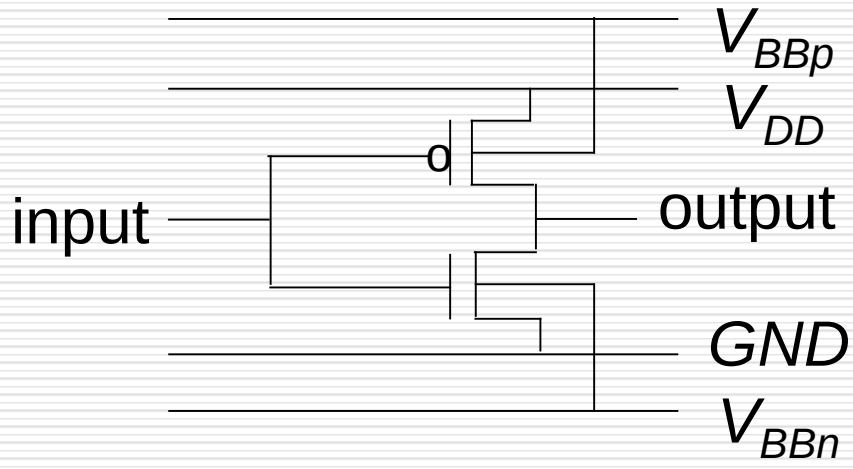
- Body biasing is changed to modify V_{TH}
- The transistors should have separate well
- Easier in case of SOI devices, since they are isolated naturally.

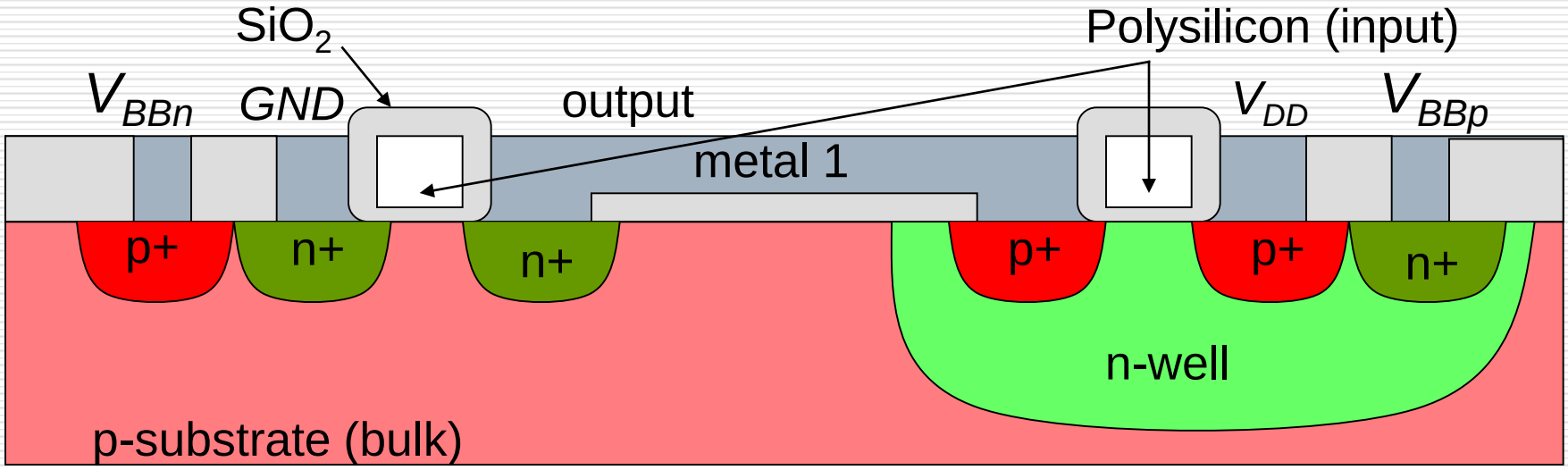
Normal CMOS Inverter



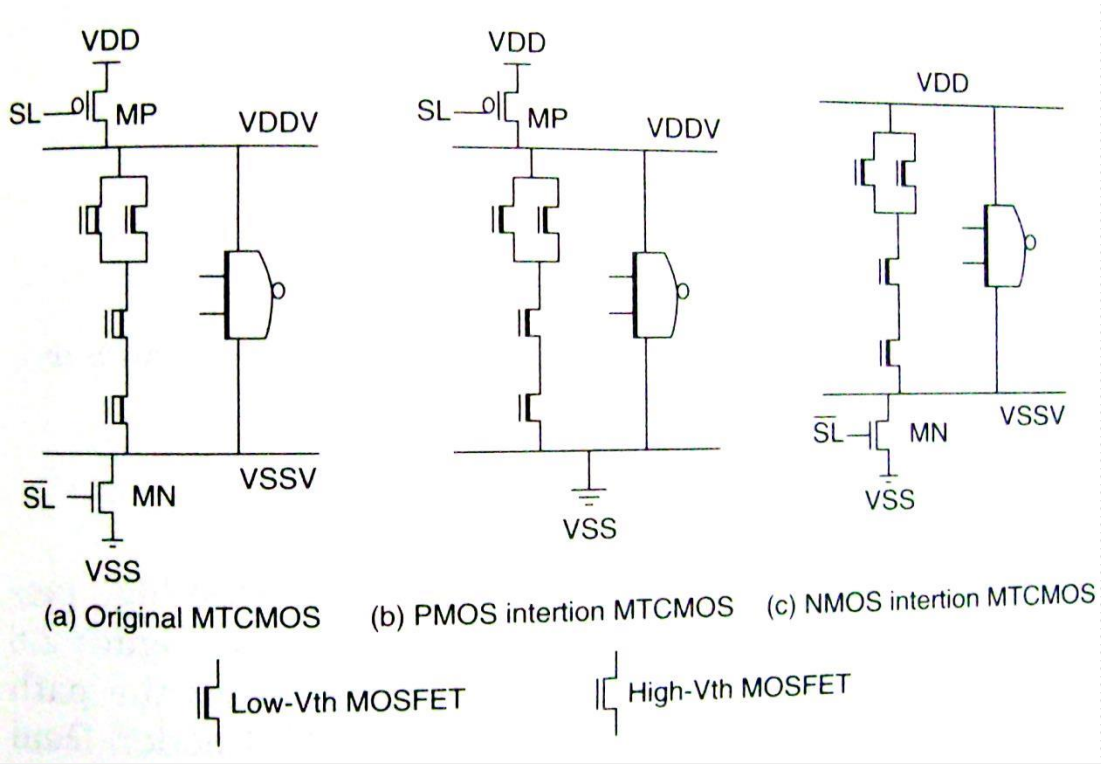


Leakage Reduction by Body Bias





e. Multithreshold-voltage CMOS (MTCMOS)



- High threshold devices are inserted in series into low-V_{TH} circuitry
- Uses a sleep control scheme
- In the active mode, SL is set low and sleep control high-V_{TH} transistors (MP and MN) are turned on

- Since their on-resistances are small, the virtual supply voltages (V_{DDV} and V_{SSV}) almost function as real power lines
- In the stand-by mode, SL is set high, MN and MP are turned off and the leakage current is low
- The NMOS insertion scheme is preferable, since the NMOS on-resistance is smaller at the same width and it can be sized smaller than corresponding PMOS

Super cut-off CMOS (SCCMOS):

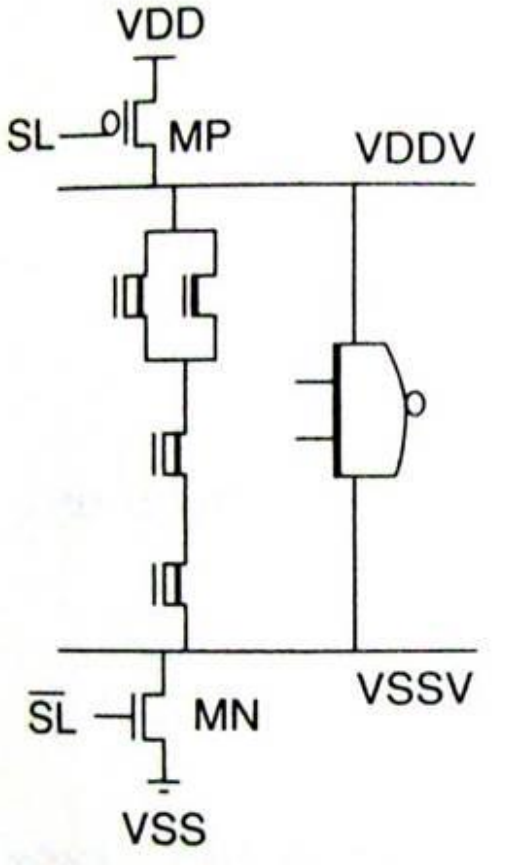
- MTCMOS can only reduce the standby leakage power and the large inserted MOSFETs can increase the area and delay

- If the data retention is required in the standby mode, an extra high-V_{TH} memory circuit is needed to maintain the data

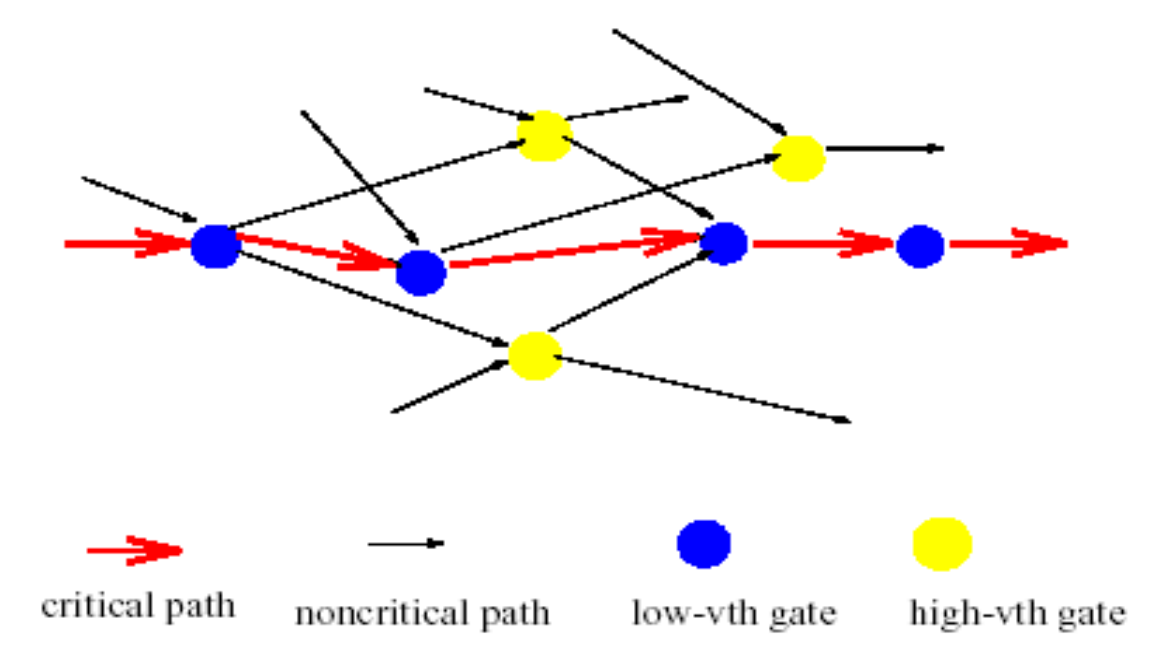
- Instead of using high-V_{TH} sleep control transistors as in MTCMOS, SCCMOS circuit uses low-V_{TH} transistors with an inserted gate bias generator

Super cut-off CMOS (SCCMOS):

- For the PMOS(NMOS), in the active mode, the gate is applied 0V(V_{DD}) and the virtual V_{DD}(V_{SS}) line is connected to V_{DD}(V_{SS})
- In the standby mode, the gate is applied V_{DD}+0.4V(V_{SS}-0.4V) to fully cut-off the leakage current
- SCCMOS circuits can work at lower supply voltages



f. Dual Threshold CMOS



- A design generally can have many paths from inputs to outputs and not all are critical paths
- High Vth transistors are used in noncritical paths so as to reduce leakage current, while the performance is maintained by the use of low-V_{TH} transistors in the critical paths

- But, some gates on non-critical paths may also be assigned low V_{TH} transistors to prevent those paths from becoming critical
- No additional leakage control transistors are required and both high performance and low power can be achieved simultaneously
- Dual V_{TH} CMOS has the same critical delay as the single low V_{TH} CMOS circuit, but the transistors in noncritical paths can be assigned high V_{TH} to reduce leakage power
- Dual threshold technique is good for leakage power reduction without delay and area overhead

A Sample Calculation:

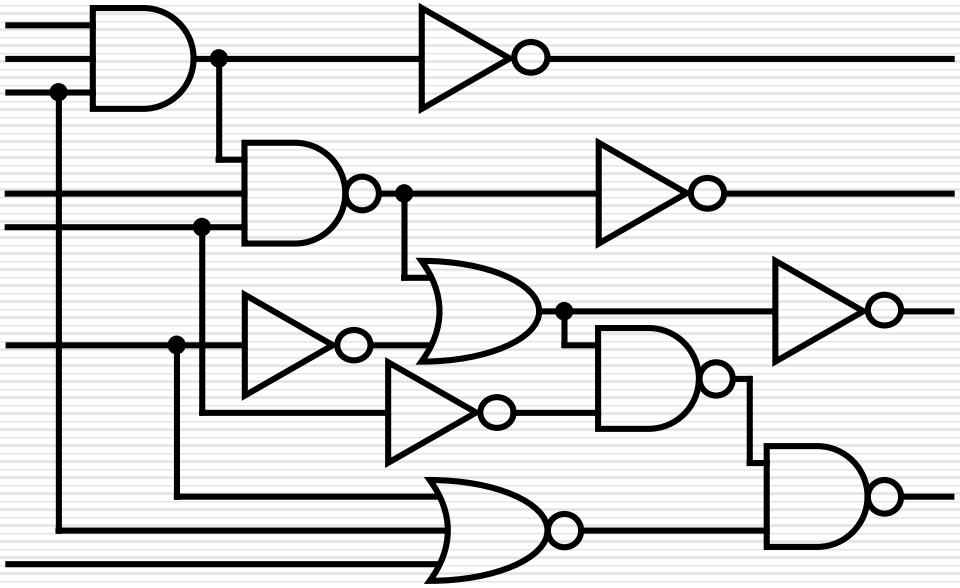
- V_{DD} = 1.2V, 100nm CMOS process
- Transistor width, W = 0.5μm
- OFF device ($V_{gs} < V_{TH}$) leakage
  - I_0 = 20nA/μm, for low threshold transistor
  - I_0 = 3nA/μm, for high threshold transistor
- Say, a chip has 100M transistors
- If all low-Vth transistors are used,
  - Leakage power $=(100\times10^6/2)(0.5\times20\times10^{-9}\text{A})(1.2\text{V})$
    $= \mathbf{600mW}$
- If all high-Vth transistors are used,
  - Leakage power $= (100\times10^6/2)(0.5\times3\times10^{-9}\text{A})(1.2\text{V})$
    $= \mathbf{90mW}$

For a Dual-Threshold Chip:

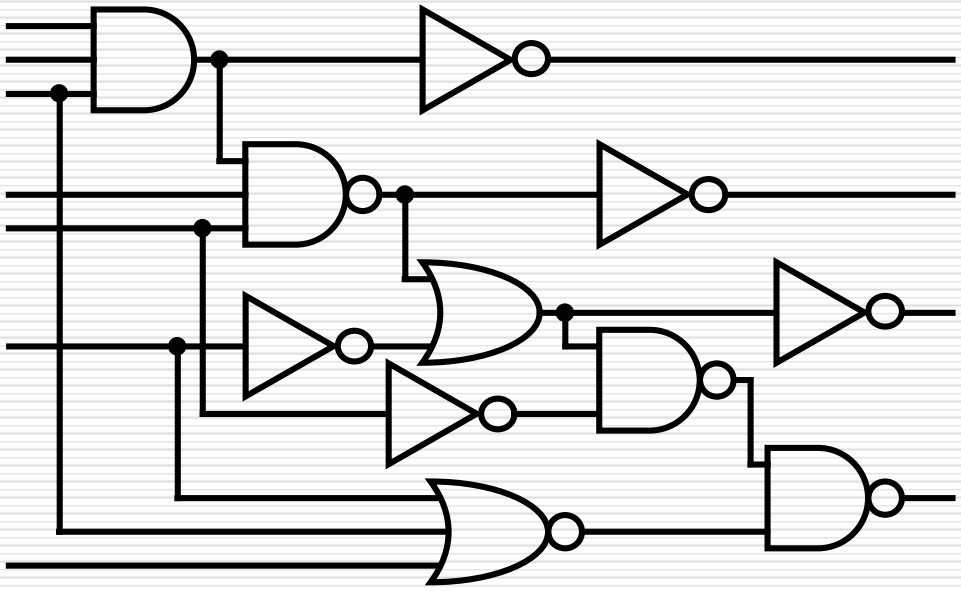
- Low-threshold only for 20% transistors on critical path.
- Leakage power = 600×0.2 + 90×0.8
  = 120 + 72
  = **192 mW**

Problem: Leakage Reduction

Following circuit is designed in 65nm CMOS technology using low threshold transistors. Each gate has a delay of 5ps and a leakage current of 10nA. Given that a gate with high threshold transistors has a delay of 12ps and leakage of 1nA, optimally design the circuit with dual-threshold gates to minimize the leakage current without increasing the critical path delay. What is the percentage reduction in leakage power? What will the leakage power reduction be if a 30% increase in the critical path delay is allowed?



Solution 1: No Delay Increase



Three critical paths are :

from the first, second and third inputs to the last output

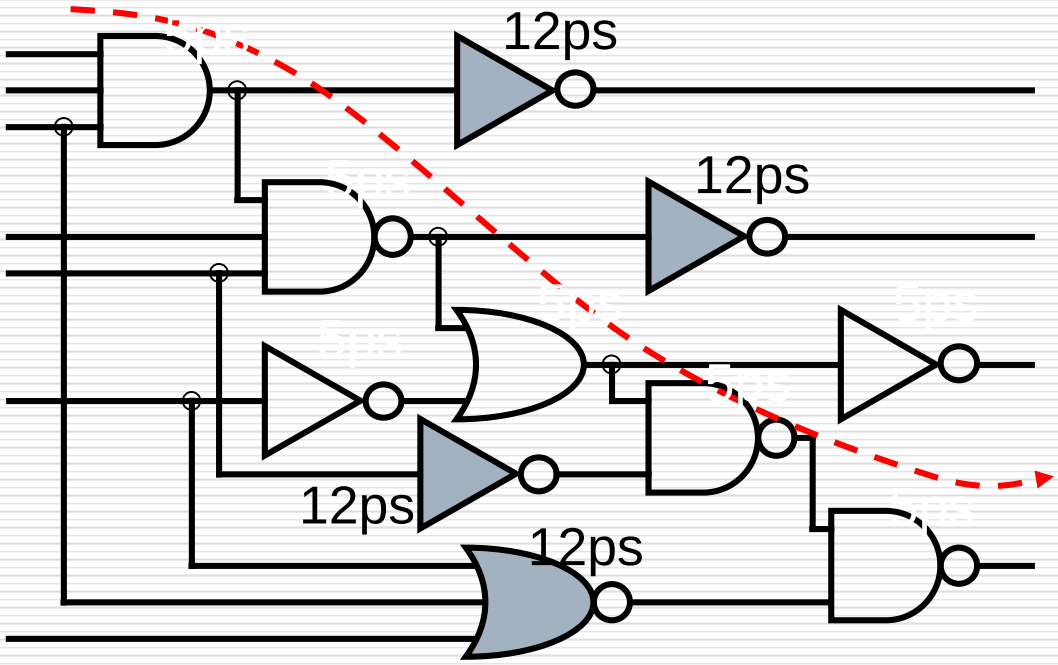
Each path has five gates and delay of 25ps. Hence the critical path delay = 25ps

None of the five gates on the critical path can be assigned high threshold ones

Also, the two inverters that are on four-gate long paths cannot be assigned high threshold ones, because then the delay of those paths will become 27ps

Solution 1: No Delay Increase

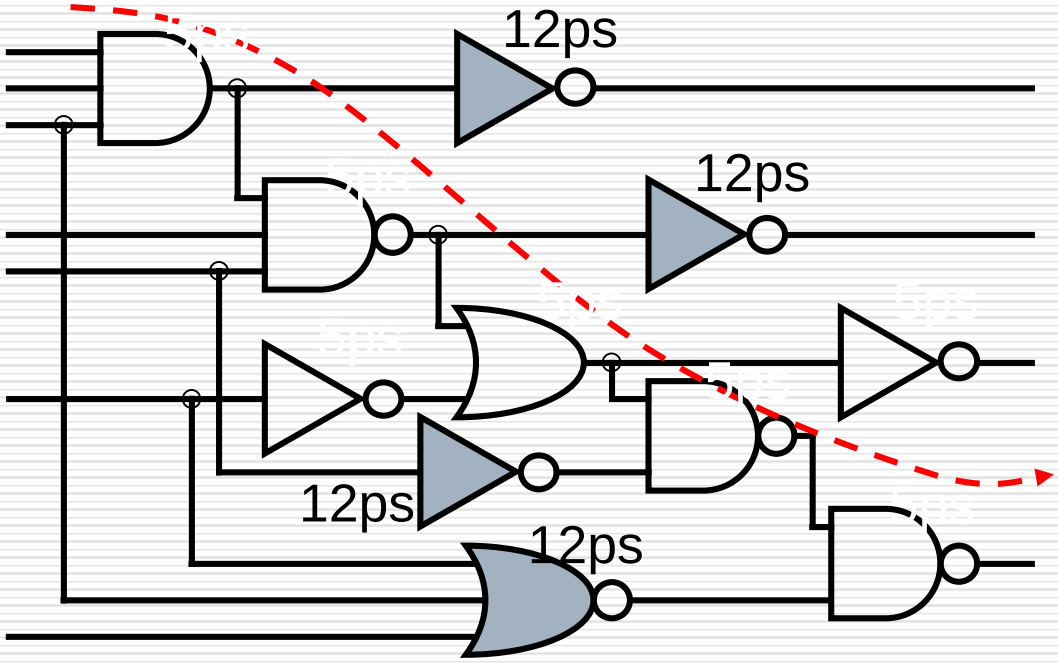
Critical path delay = 25ps



The remaining three inverters and the NOR gate can be assigned high threshold. These are shaded in the circuit.

Solution 1: No Delay Increase

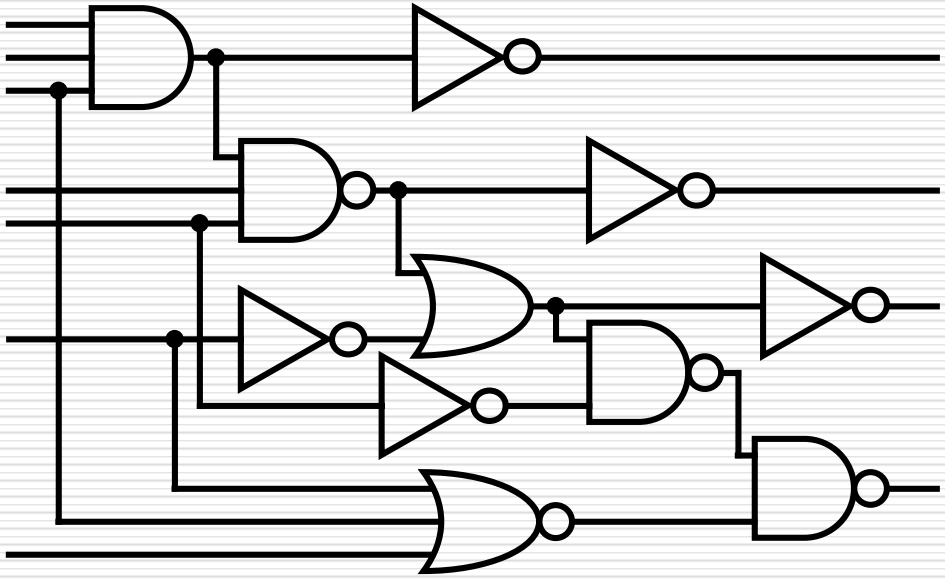
Critical path delay = 25ps



The reduction in leakage power is :

= [11×10 – (4×1+7×10)]/(11×10) = **32.73%**

Solution 2: 30% Delay Increase



The allowed new delay : 25 + 30%(25) = 32.5 ps

We can replace as many low threshold gates with high threshold ones to reduce leakage as long as the modified critical path delay requirement is met

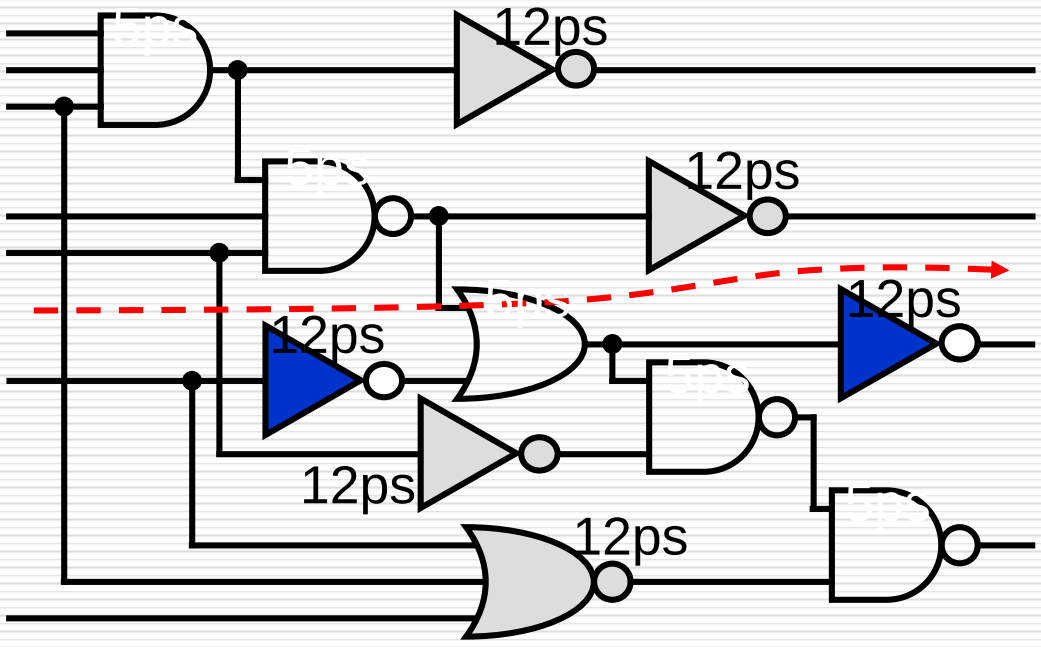
There are 2, 3, 4 and 5 gate paths

2 gate paths can have both high threshold gates
3-gate paths can have 2 high threshold gates
4 and 5 gate paths can have only one high threshold gate

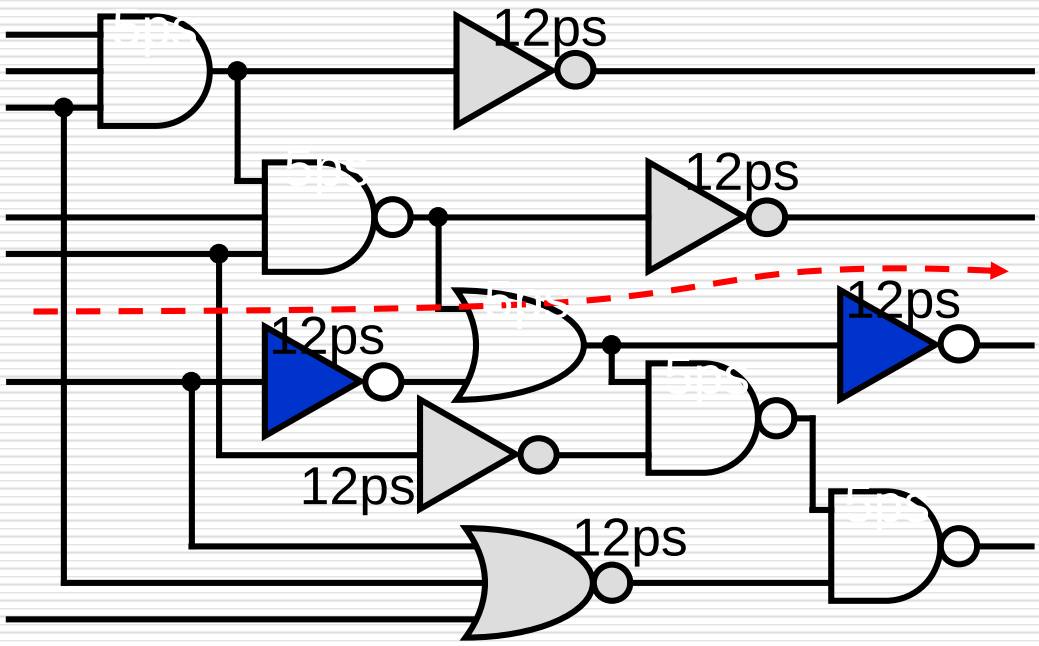
Solution 2: 30% Delay Increase

Several solutions are possible.

One solution is shown below where six high threshold gates are shown with shading and the critical path is shown by a dashed red line arrow.



Solution 2: 30% Delay Increase



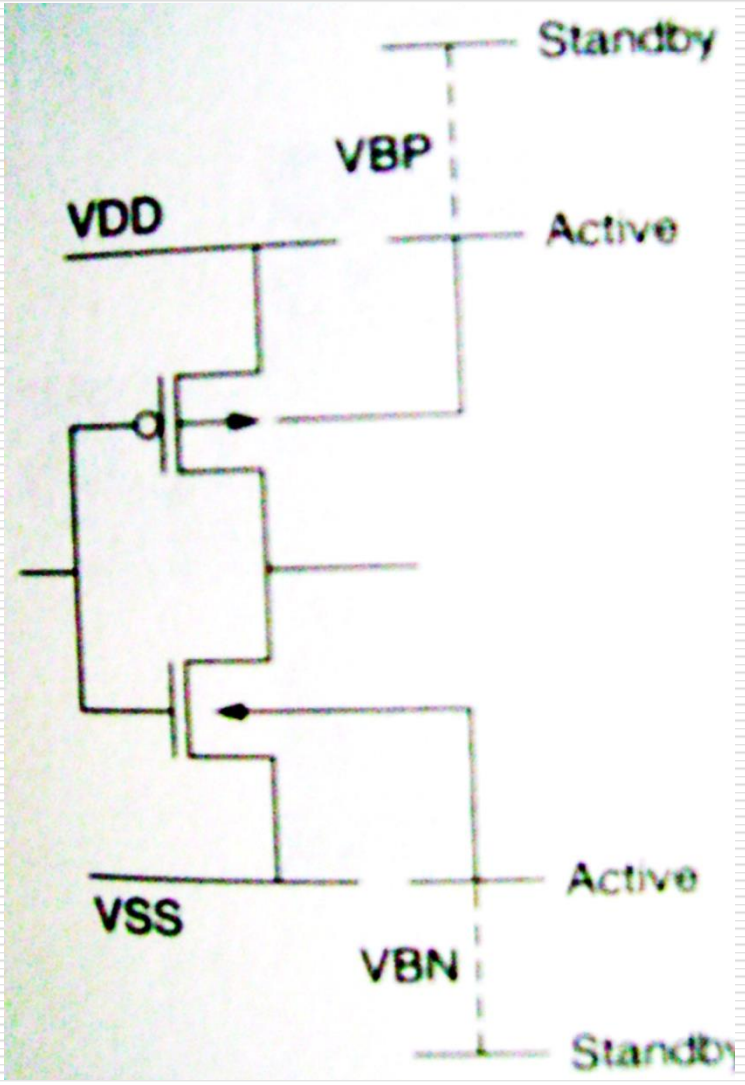
The reduction in leakage power is :

= [11×10 – (6×1+5×10)]/(11×10) = **49.09%**

The new critical path delay = 29ps

g. Variable Threshold CMOS (VTMOS)

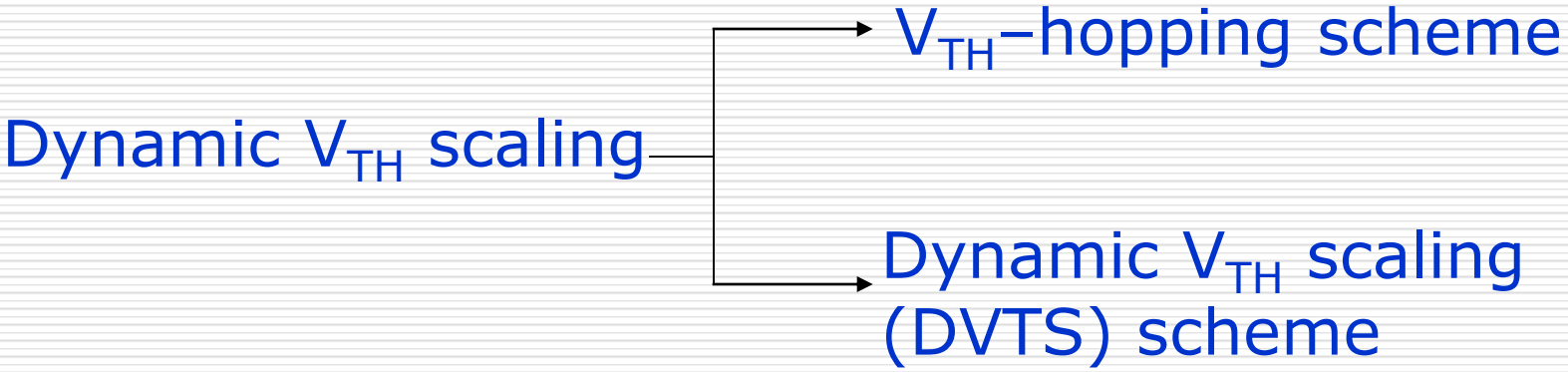
- This is a body-biasing design technique.
- A substrate bias circuit is used to control the body bias.
- In active mode, nearly zero body bias is applied.
- In standby mode, a deeper reverse body bias is applied to increase Vt.
- In active mode, a slightly forward substrate bias can be used to reduce Vt and increase the circuit speed



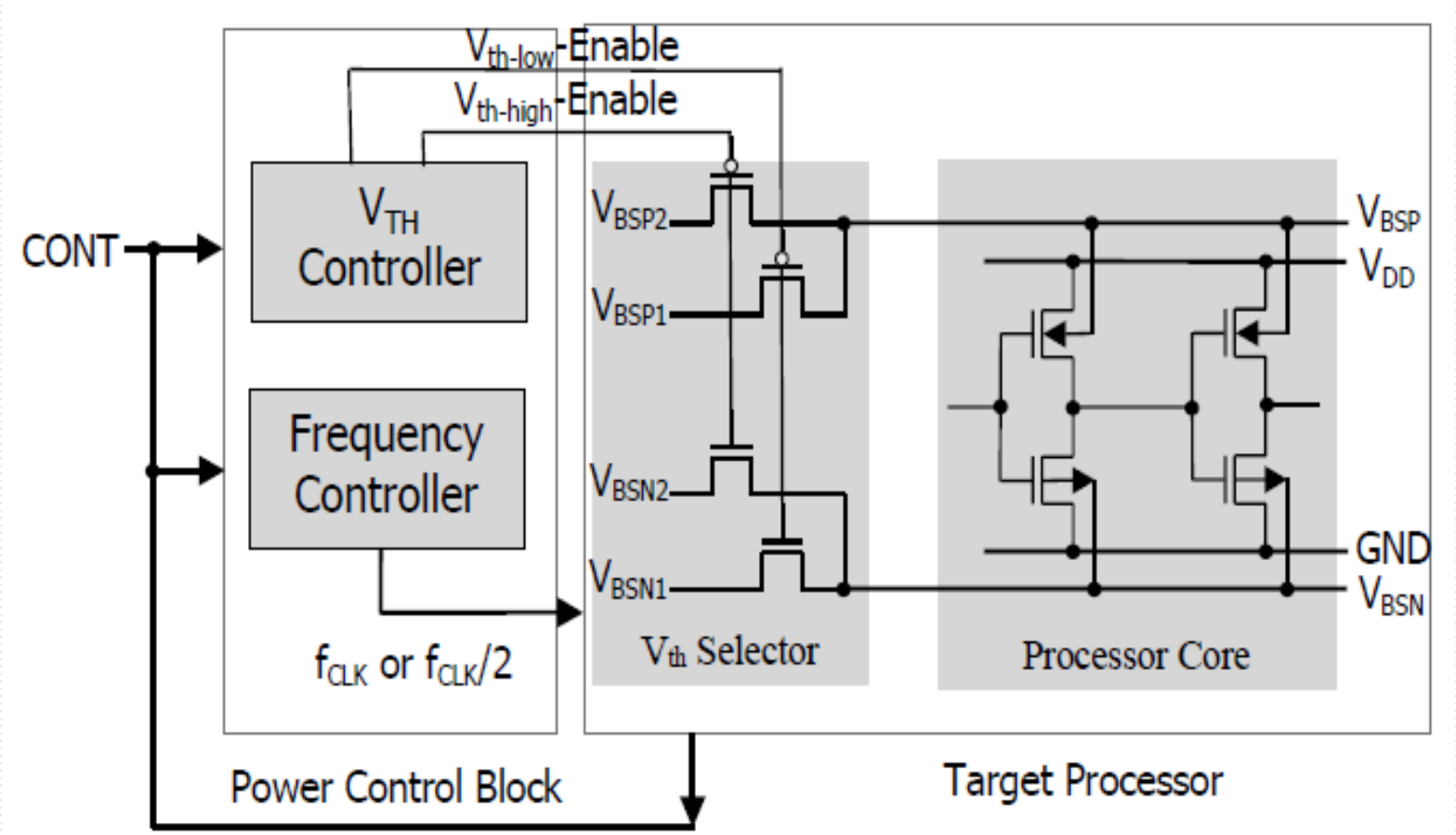
- Requires routing the body grid that adds to the overall chip area.
- It has been shown that in 0.35μm technology reverse body biasing lowers IC leakage by 3 orders of magnitude.
- The effectiveness of reverse body bias to lower I_{OFF} decreases as technology scales.

Dynamic V_{TH} Techniques:

- Dynamic threshold voltage scaling is a technique for active leakage power reduction.
- This scheme utilizes dynamic adjustment of V_{TH} through back-gate bias control depending on the workload of a system.
- When the workload decreases, less power is consumed by increasing V_{TH}.

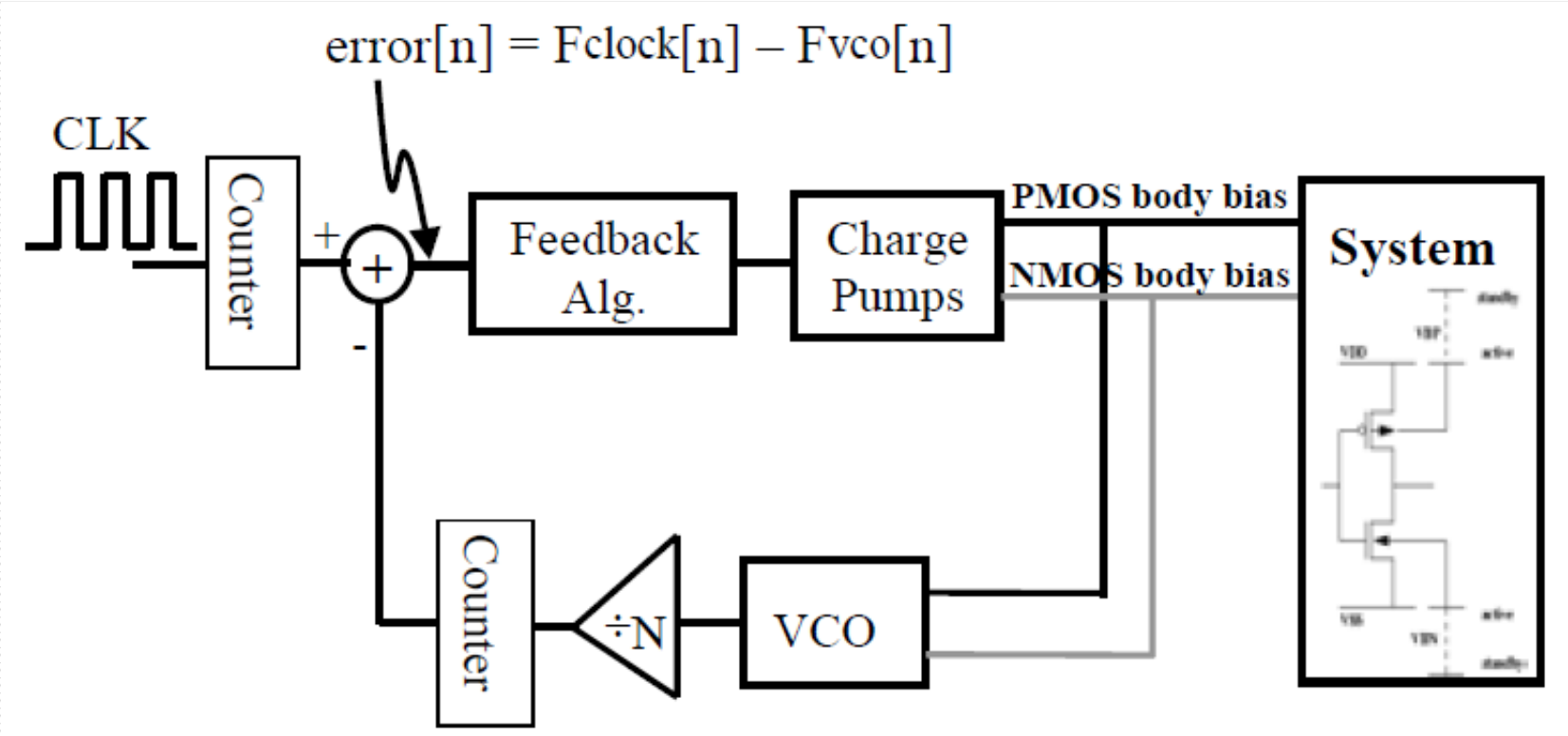


V_{TH}–hopping scheme



Schematic diagram of V_{TH}–hopping

Dynamic V_{TH} scaling (DVTS) scheme



Schematic of the DVTS hardware

Problem 1:

- A 32 bit off-chip bus operating at 1V and 2GHz clock rate is driving a capacitance of 2pF/bit. Each bit is estimated to have a toggling probability of 0.5 at each clock cycle. What is the power dissipation in operating the bus?

Solution:

- Total capacitance, C = 32×2 = 64pF
- Power dissipation, P = $\alpha\ C_L\ V_{DD}^2\ f_{CLK}$
- = $0.5\times(64\times10^{-12})\times1^2\times2\times10^9$ W
- = **64 mW**

Problem 2:

- The chip size of a CPU is 1cm×1cm with clock frequency of 500MHz operating at 1.2V. The length of the clock routing is estimated to be twice the circumference of the chip. Assume that the clock signal is routed on a metal layer with width of 1μm and the parasitic capacitance of the metal layer is 1fF/μm^2. What is the power dissipation of the clock signal?

Solution:

- Total capacitance, C = 4cm×2×1u×1fF/u.u = 80pF
- For clock signal, $\alpha = 1$
- Power dissipation, P
  - $= \alpha\ C_L\ VDD^2\ f_{CLK}$
  - = 1×(80pF)×1.2×1.2×(500Meg)
  - = 57.6mW

Problem 3:

- State and prove a theorem specifying the condition for zero short-circuit power consumption in a CMOS gate.

Solution:

- Theorem – A CMOS gate consumes no short-circuit power when $V_{DD} \leq V_{tn} + |V_{tp}|$, i.e., supply voltage is lower than the sum of the threshold voltage magnitudes for the n and p channel MOSFETs.

- Proof: The short-circuit conduction requires that a pull-up path through pMOS devices and a pull-down path through nMOS devices should be simultaneously on. If the common gate voltage for both devices is V_{in}, where $0 \leq V_{in} \leq V_{DD}$, then a necessary condition for short-circuit conduction is:

$$V_{tn} \leq V_{in} \leq V_{DD} - |V_{tp}|$$

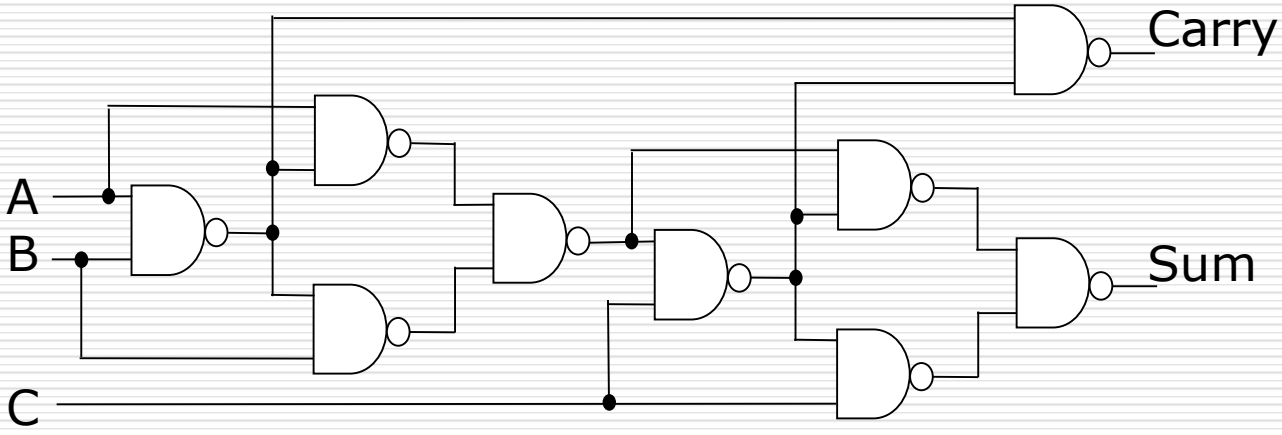
- In order to make this condition impossible, we must ensure that the upper bound on V_{in} does not exceed the lower bound. Thus,

$$V_{DD} - |V_{tp}| \leq V_{tn}$$

Therefore, $V_{DD} \leq V_{tn} + |V_{tp}|$

Problem 5:

The following circuit is implemented in 65 nm CMOS technology, which suffers from high leakage. For high speed, only low threshold transistors have been used. Each gate has a delay of 5ps and a leakage current of 10nA.

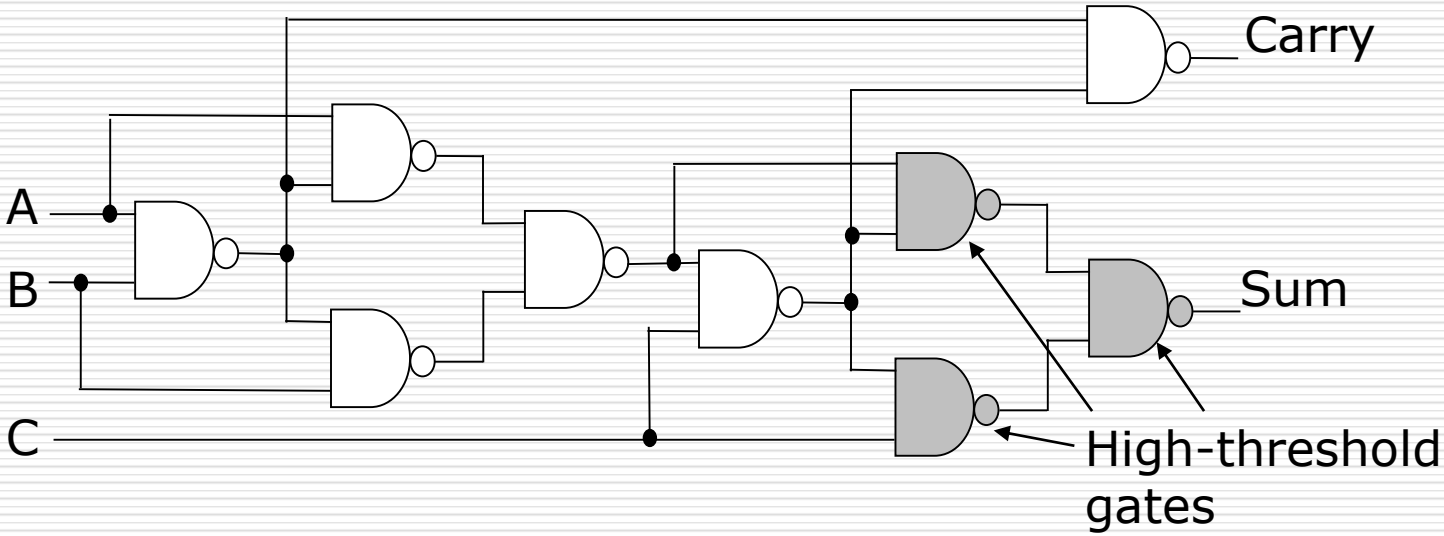


(a) Given that a gate with high threshold transistors has a delay of 12ps and leakage of 1nA, optimally assign thresholds to gates to minimize the leakage current without increasing the critical path delay for Carry output. Assume that delay of Sum output is not critical. What is the percentage reduction in leakage power?

(b) Is it possible to further reduce leakage by redesigning the circuit? If yes, show how much reduction you can obtain.

(c) Are the dual threshold designs better or worse than the all low threshold design for glitch power?

(a) Given that the critical path delay of the carry output must not increase, only three gates (shown shaded in the following diagram) can have high threshold.



Therefore,

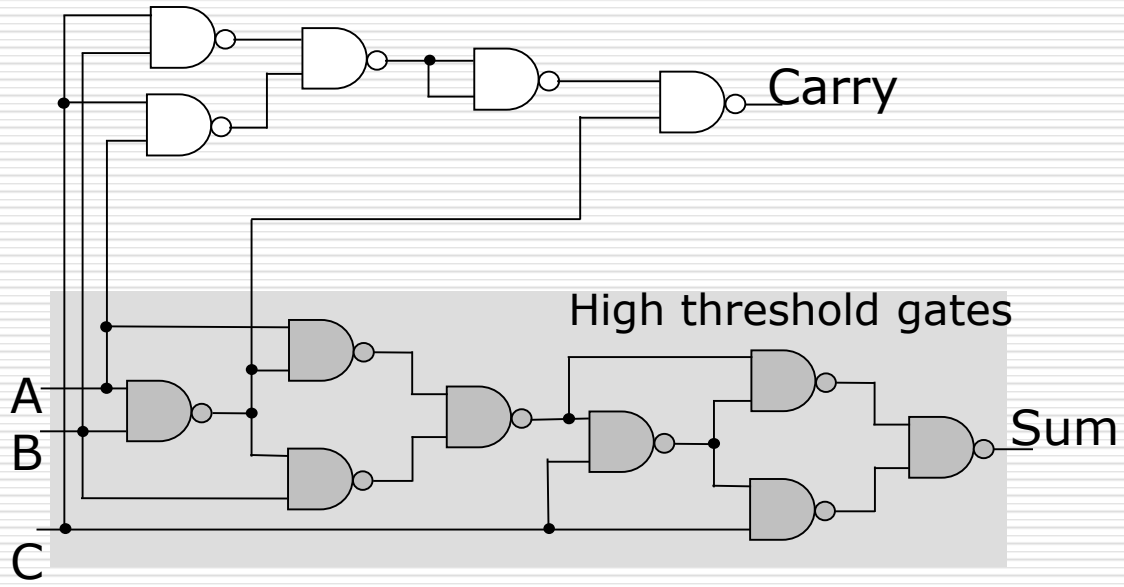
Leakage reduction = 100 [9×10 – (6×10 + 3×1)] / (9×10)

= **30%**

(b) The carry signal can be resynthesized from its truth table. Thus,

$$\text{Carry} = AB + BC + CA = [\,(AB)'\{(BC)'\,(CA)'\}]'$$

The following circuit implements this using only two-input NAND gates:



Leakage reduction = 100 [9×10 – (5×10 + 8×1)] / (9×10)

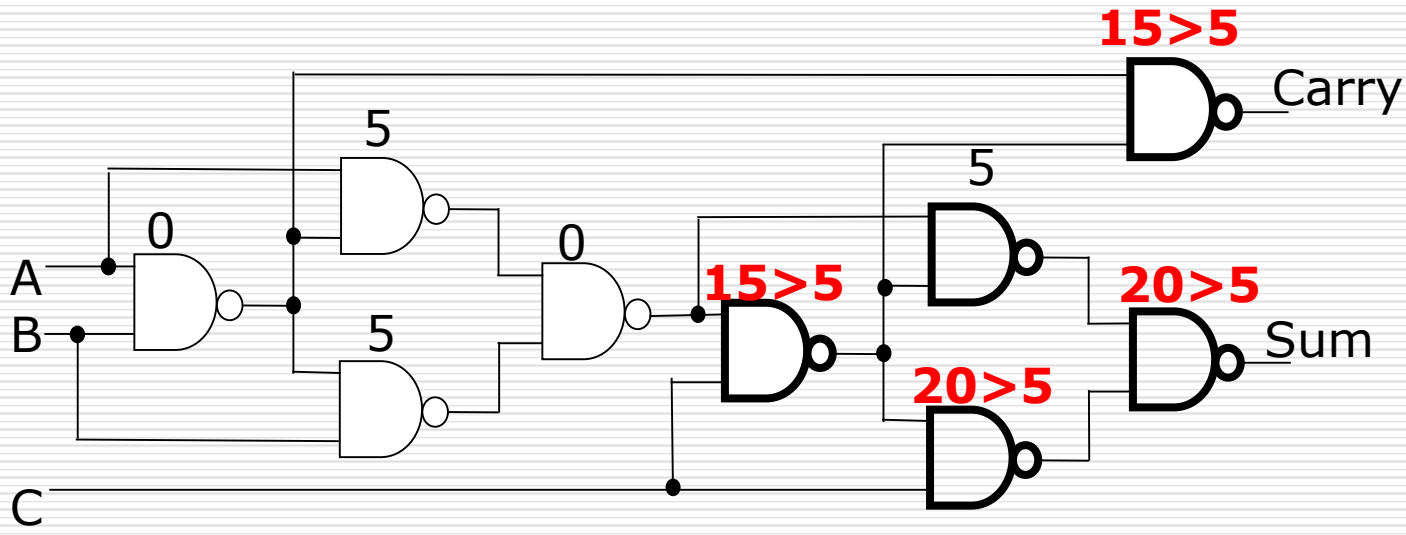
= **35.6%**

Although a slightly higher leakage reduction is achieved, this circuit will have more dynamic power dissipation because it contains 44.4% more gates. This design can be much improved if a three-input NAND gate with similar delay (5ps) is available.

(c) To analyze the glitch power, we examine the glitch suppression condition, i.e.,

differential path delay at a gate < inertial delay of the gate, for all gates.

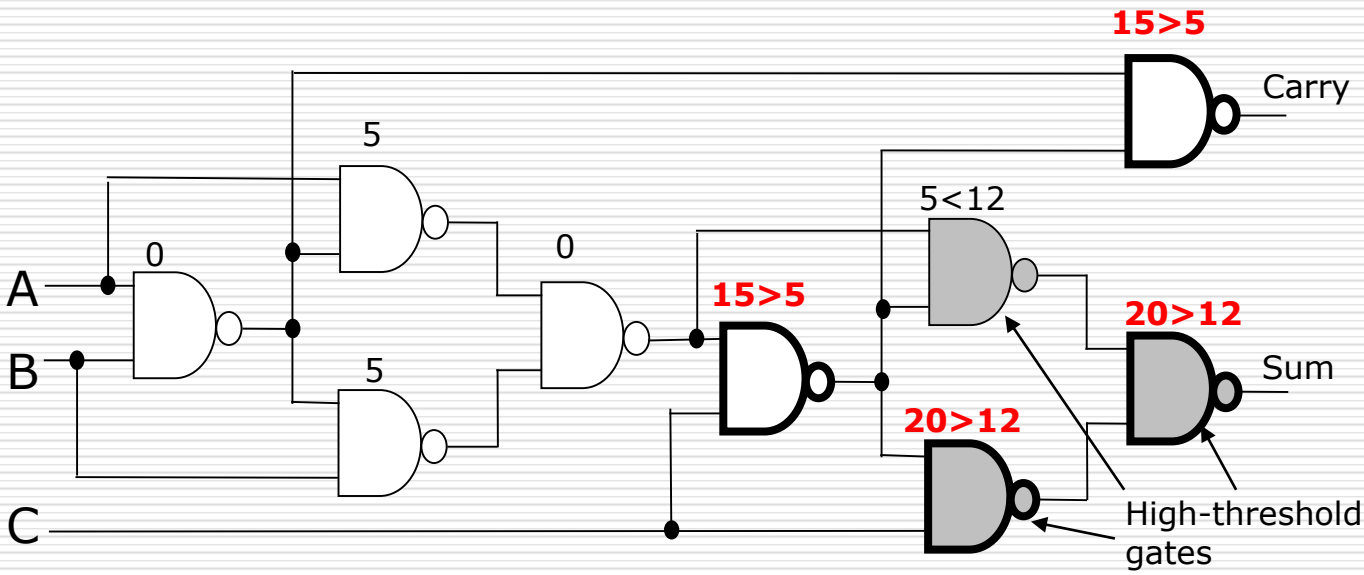
In the original circuit, all gates have 5ps delay. The maximum differential delay is shown below



All low-threshold gates

Gates that can potentially produce glitches are shown in bold

Dual-threshold circuit



The number of gates with potential glitches is reduced by one.
In general, we can assume that the dual threshold circuit, because of larger inertial delays, will produce fewer glitches.

Problem 6:

A CMOS processor has a rated supply voltage 1.5V and clock frequency 2GHz. Its average power consumption is 100W, which consists of 75W dynamic power and 25W static power. Assuming that the delay of a gate in the technology is proportional to $V_{DD}/(V_{DD} - V_{th})$, where threshold voltage V_{th} = 0.5V. A low energy mode uses a lower supply voltage and a reduced frequency clock.

Determine the voltage and clock frequency that will minimize the average energy consumption per cycle.

Compare the power consumption and energy per cycle for the rated and low energy modes.

Solution:

We use the given data to establish the following for a supply voltage V:

(a) Dynamic power, $P_{dyn} = C V^2 f$

where C is the total average capacitance switched per cycle.

At the rated voltage and frequency, we have

$P_{dyn} = C (1.5)^2\ 2\ G = 75$ or $C = 16.67$ nF

Therefore, $P_{dyn} = 16.67 V^2 f$

(b) Dynamic energy per cycle, $E_{dyn} = P_{dyn} / f = 16.67 V^2$

(c) Clock frequency, f, is inversely proportional to gate delay:

$f = k(V - V_t)/V$, where k is a constant of proportionality.

Using rated voltage and frequency, we get $k = 3$ GHz.

Therefore,

$f = 3\ (V - V_t)/V$ GHz

(d) Static power is proportional to supply voltage:

$P_{stat} = k' V$, where k' is a constant of proportionality.

Using the rated voltage data, we get $k' = 25/1.5 = 16.67$,

Therefore, $P_{stat} = 16.67\ V$

Static energy per cycle is obtained by multiplying P_{stat} with clock period (1/f).

Thus,

$E_{stat} = (16.67/3)\ V^2/(V - V_t) = 5.56\ V^2/(V - 0.5)$

(e) Total energy per cycle is given by,

$$E_{total} = 16.67\ V^2 + 5.56\ V^2/(V - 0.5)$$

To minimize this, we set its derivative to 0 and obtain the following quadratic equation:

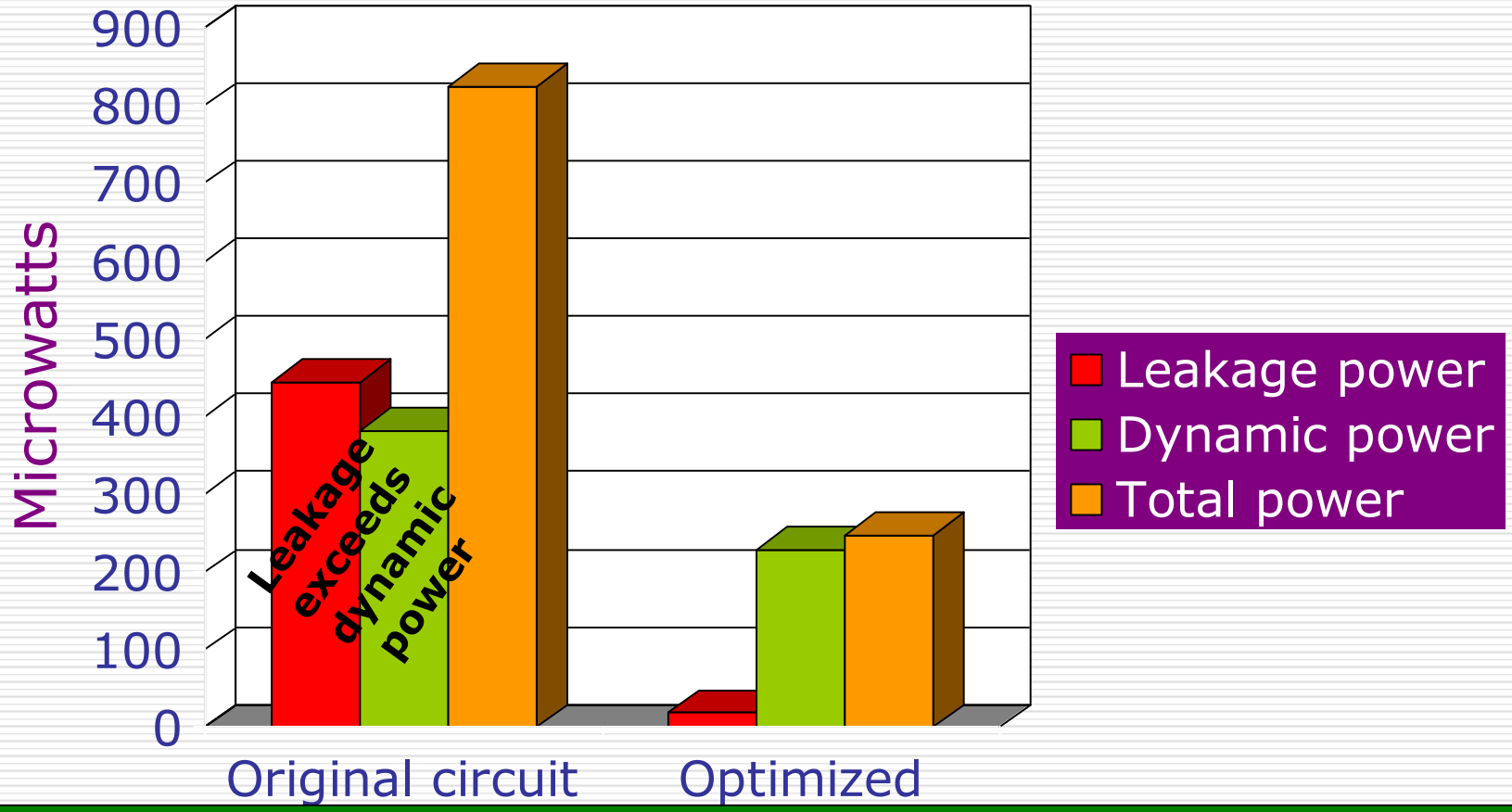
$$6\ V^2 - 5\ V + 0.5 = 0$$

There are two roots, V = 0.72 and 0.1167. We discard the second root as it is lower than the threshold voltage of 0.5 V.

The following table compares the two modes of operation. Voltage, clock frequency, power and energy per cycle are highlighted.

	Rated mode	Low energy mode	Reduction (%)
Voltage	**1.5 V**	**0.72 V**	**52%**
Clock frequency	**2 GHz**	**917 MHz**	**54%**
Dynamic energy/cycle	37.5 nJ	8.64 nJ	77%
Static energy/cycle	12.5 nJ	13.1 nJ	– 4.8%
Total energy/cycle	**50.0 nJ**	**21.74 nJ**	**56.5%**
Dynamic power	75.0 W	7.92 W	89.4%
Static power	25.0 W	12.0 W	52.0%
Total power	**100.0 W**	**19.92 W**	**80.1%**

Leakage & Dynamic Power Optimization 70nm CMOS c7552 Benchmark Circuit @ 90°C



Y. Lu and V. D. Agrawal, "CMOS Leakage and Glitch Minimization for Power-Performance Tradeoff," *Journal of Low Power Electronics (JOLPE)*, vol. 2, no. 3, pp. 378-387, December 2006.

Summary

- Leakage power is a significant fraction of the total power in nanometer CMOS devices.
- Leakage power increases with temperature; can be as much as dynamic power.
- Dual threshold design can reduce leakage.
  - Reference: Y. Lu and V. D. Agrawal, "CMOS Leakage and Glitch Minimization for Power-Performance Tradeoff," *J. Low Power Electronics,* Vol. 2, No. 3, pp. 378-387, December 2006.
  - Access other paper at http://www.eng.auburn.edu/~vagrawal/TALKS/talks.html