

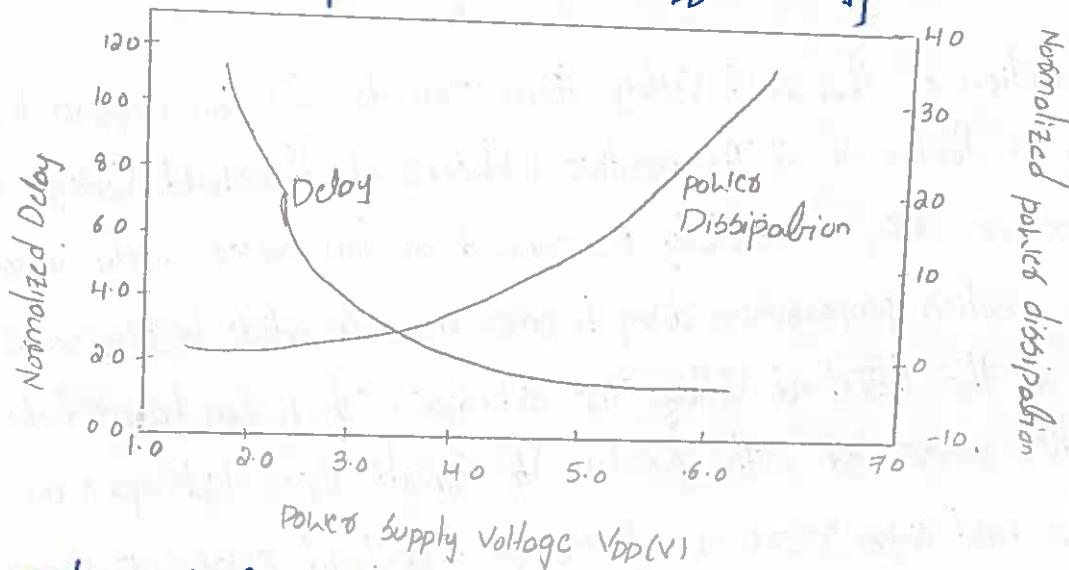
## Low power Design through Voltage Scaling:

The Switching power dissipation in CMOS digital integrated Circuits is a strong function of the power supply voltage. Therefore, reduction of V<sub>DD</sub> emerges as a very effective means of limiting the power consumption. Given a certain technology, the circuit designer may utilize on-chip DC-DC converters and/or separate power pins to achieve this goal. The savings in power dissipation comes at a significant cost in terms of increased circuit delay.

When considering drastic reduction of the power supply voltage below the new standard of 3.3V, the issue of time domain performance should also be addressed carefully. Influence of Voltage Scaling on Power and Delay. Although the reduction of power supply voltage significantly reduces the dynamic power dissipation, the inevitable design trade off is the increase of delay. This can be seen easily by examining the following propagation delay expressions for the CMOS inverter circuit.

$$\tau_{PHL} = \frac{C_{load}}{k_n(V_{DD}-V_{Thn})} \left[ \frac{2V_{Thn}}{V_{DD}-V_{Thn}} + \ln \left[ \frac{4(V_{DD}-V_{Thn})}{V_{DD}} - 1 \right] \right]$$

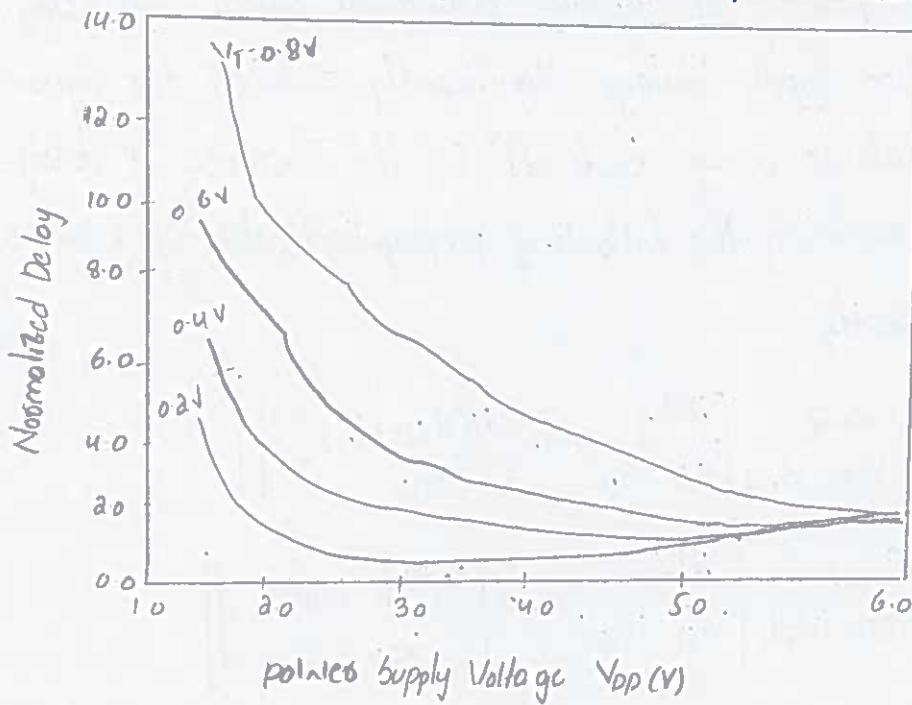
$$\tau_{PLH} = \frac{C_{load}}{k_p(V_{DD}-V_{Thp})} \left[ \frac{2V_{Thp}}{V_{DD}-V_{Thp}} + \ln \left[ \frac{4(V_{DD}-V_{Thp})}{V_{DD}} - 1 \right] \right]$$



The dependence of Circuit Speed on the power supply voltage may also influence the relationship b/w the dynamic power dissipation and the supply voltage. The above graph suggests a quadratic improvement of power consumption as the power

Supply Voltage is reduced. However, this expectation assumes that the switching frequency remains constant. If the circuit is always operated at the maximum frequency allowed by its propagation delay, the number of switching events per unit time will drop as the propagation delay becomes larger with the reduction of the power supply voltage. Here, we examine the effects of reducing the power supply voltage for a given technology, hence, key device parameters and the load capacitances are assumed to be constant.

The propagation delay equations show that the negative effect of reducing the power supply voltage upon delay can be compensated for, if the threshold voltage of the transistors is scaled down accordingly. When scaled linearly, reduced threshold voltages allow the circuit to produce the same speed performance at a lower  $V_{DD}$ .

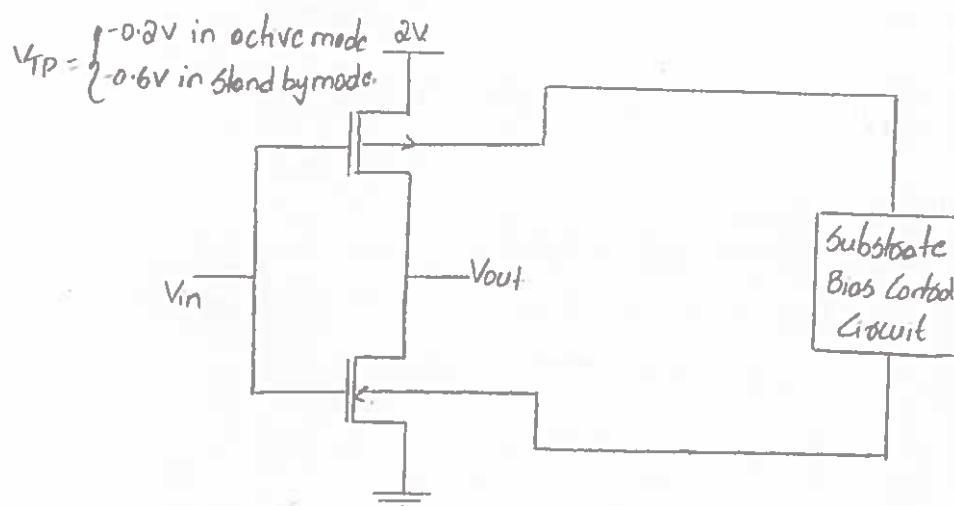


The reduction of threshold voltage from  $0.8V$  to  $0.2V$  can improve the delay at  $V_{DD} = 2V$  by a factor of 2. The positive influence of threshold voltage reduction upon propagation delay is specially pronounced at low power supply voltages, for  $V_{DD} < 2V$ . In addition, propagation delay becomes more sensitive to process related fluctuations of the threshold voltage. The techniques which can be used to overcome the difficulties associated with the low  $V_T$  circuits. These techniques are called Variable threshold CMOS (VTCMOS) and Multiple threshold CMOS (MTCMOS).

## Variable Threshold CMOS (VTCMOS) Circuits:

Using a low supply voltage and a low threshold voltage in CMOS logic Circuits is an efficient method for reducing the overall power dissipation, while maintaining high speed performance. Yet designing a CMOS logic gate entirely with low VT transistors will inevitably lead to increased sub threshold leakage, and consequently, to higher stand by power dissipation when the op is not switching.

The threshold voltage  $V_T$  of an MOS transistor is a function of its source to substrate voltage  $V_{SB}$ . In conventional CMOS logic Circuits, the substrate terminals of all nMOS transistors are connected to ground potential, while the substrate terminals of all PMOS transistors are connected to  $V_{DD}$ .



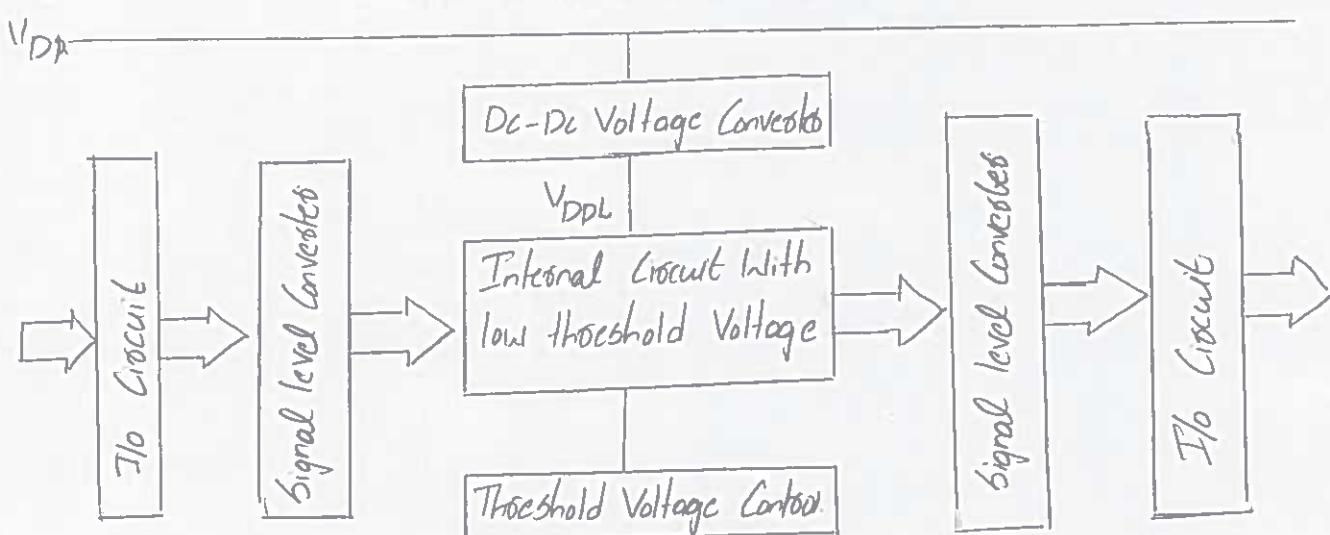
This ensures that the source and drain diffusion regions always remain reverse biased with respect to the substrate, and that the threshold voltages of the transistors are not significantly influenced by the body effect. In VTCMOS circuit technique, on the other hand, the transistors are designed inherently with a low threshold voltage, and the substrate bias voltages of nMOS and PMOS transistors are generated by a variable substrate bias control circuit.

When the inverter circuit is operating in its active mode, the substrate bias voltage of the nMOS transistor is  $V_{SB} = 0$  and the substrate bias voltage of the PMOS transistor is  $V_{SB} = V_{DD}$ . Thus, the inverter transistors do not experience any back gate bias effect. The ckt operates with low  $V_{DD}$  and low  $V_T$ , benefiting from both low power dissipation & high switching speed.

When the inverter circuit is in the stand by mode, however, the substrate bias control ckt generates a lower substrate bias voltage for the nMOS transistor and a higher substrate bias voltage for the PMOS transistor. As a result, the

magnitudes of the threshold voltages  $V_{T1}$  and  $V_T$ , both increase in the stand by mode, due to the back gate bias effect. Since the sub threshold leakage current drops exponentially with increasing threshold voltage, the leakage power dissipation in the stand by mode can be significantly reduced with this technique.

The VTCMOS technique can also be used to automatically control the threshold voltages of the transistors in order to reduce leakage currents, and to compensate for process related fluctuations of the threshold voltages. This approach is also called the self Adjusting Threshold Voltage (V<sub>T</sub>) scheme.



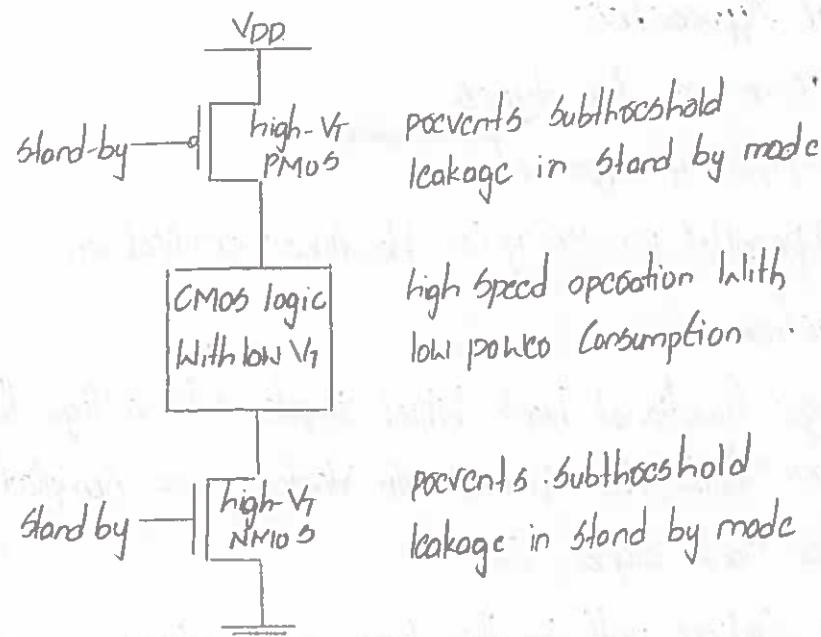
Block Diagram of a typical low power chip

#### Multiple Threshold CMOS (MTCMOS) Circuits:

Another technique which can be applied for reducing leakage currents in low voltage circuits in the stand by mode is based on using two different types of transistors with two different threshold voltages in the circuit. (HOLI) Here, low  $V_T$  transistors are typically used to design the logic gates where switching speed is essential, high  $V_T$  transistors are used to effectively isolate the logic gates in stand by mode and to prevent leakage dissipation.

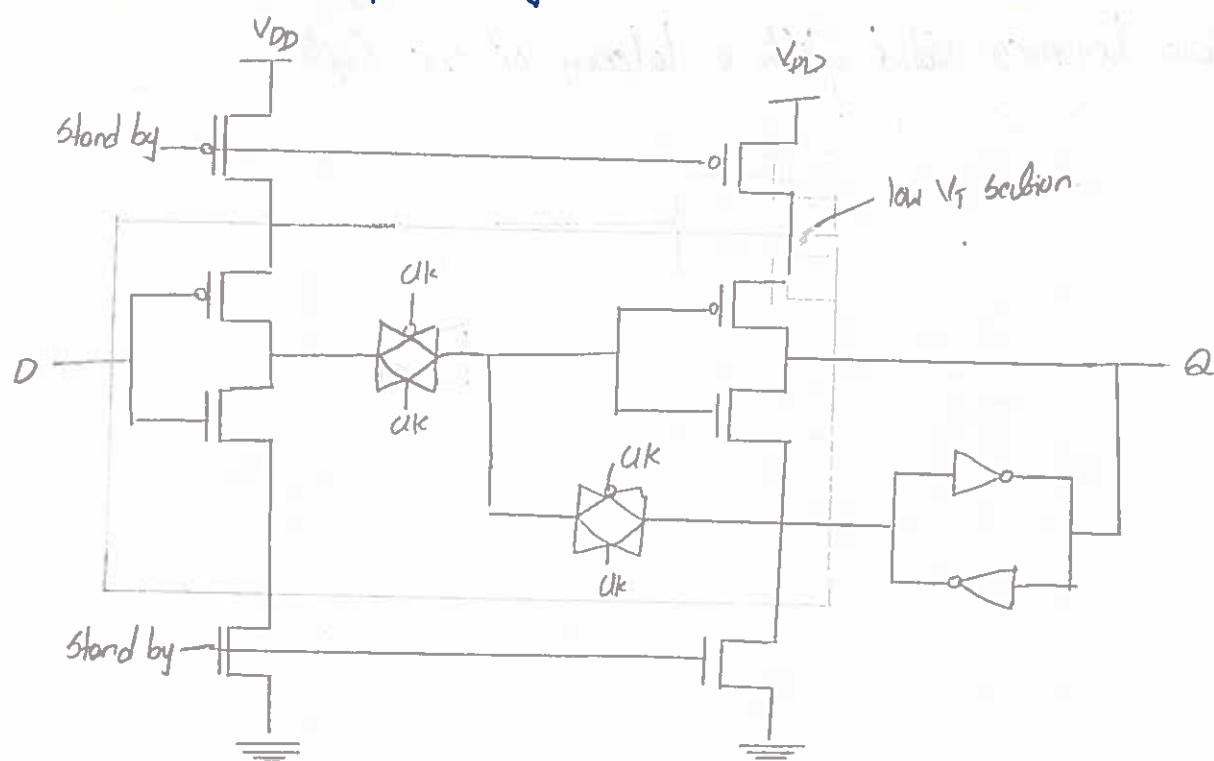
In the active mode, the high  $V_T$  transistors are turned on and the logic gates consisting of low- $V_T$  transistors can operate with low switching power dissipation and small propagation delay. When the ckt is driven into

Stand by mode, on the other hand, the high  $V_T$  transistors are turned off and the conduction paths for any sub threshold leakage currents that may originate from the internal low  $V_T$  circuitry are effectively cut off.



General Structure of a Multiple Threshold CMOS (MTCMOS) logic gate

The Critical Signal propagation path from the  $D$  to  $Q$  consists exclusively of low  $V_T$  transistors. While a Cross Coupled inverter pair consisting of high  $V_T$  transistors is used for precharging the data in the stand by mode.



Low power/low Voltage D-latch Circuit designed with MTCMOS technique

The MTCMOS technique is conceptually easier to apply and to use compared to the VT莫斯 technique, which usually requires a sophisticated substrate bias to control mechanism. It does not require a twin well or triple well CMOS process.

one of the disadvantages of the NMOS Ckt technique is the presence of series connected stand by transistors, which increase the overall circuit area and also add extra parasitic capacitance.

### Architectural level Approaches:

There are two types

1) Pipelining Approach

2) parallel processing (or) Hardware replication

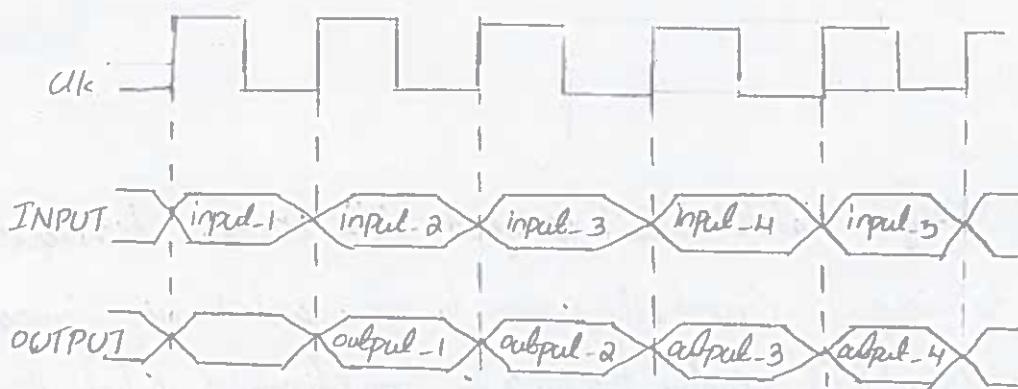
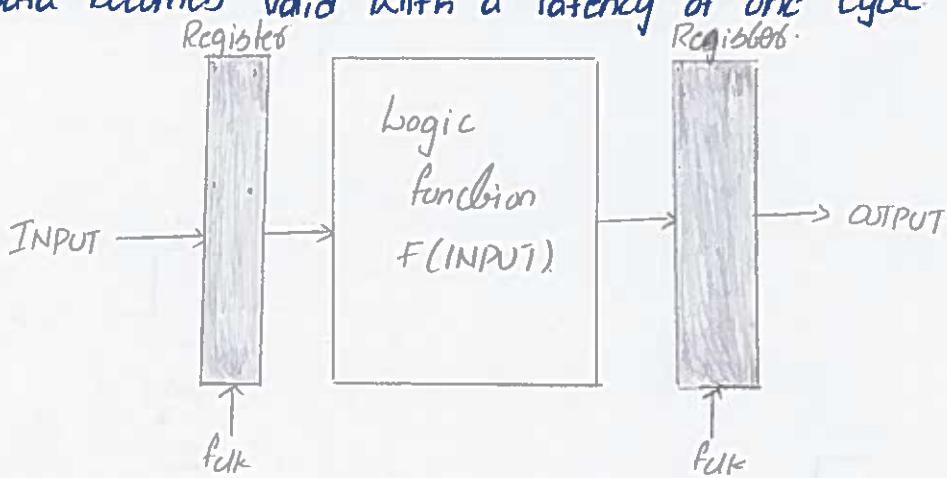
### Pipelining Approach:

\* Consider the single functional block which implements a logic function  $F(\text{INPUT})$  of the i/p Vector, INPUT. Both the i/p and o/p Vectors are sampled through registers arrays, driven by a clock signal CLK.

\* Assume that the Critical path in this logic block allows a maximum sampling frequency of CLK.

\* In other words, the maximum i/p to o/p propagation delay  $T_{max}$  of this logic block is equal to or less than  $T_{CLK} = 1/f_{CLK}$ .

\* A new i/p Vector is latched into the i/p registers array at each clock cycle, and the o/p data becomes valid with a latency of one cycle.



Single stage implementation of a logic function & its simplified timing diagram.

Let  $C_{total}$  be the total Capacitance switched every clock cycle. Hence,  $C_{total}$  consists of

- The Capacitance switched in the i/p registers array.
- The Capacitance switched to implement the logic function
- The Capacitance switched in the o/p registers array.

Then, the dynamic power consumption of this structure can be found as

$$P_{dynamic} = C_{total} \cdot V_{DD}^2 \cdot f_{CLK}$$

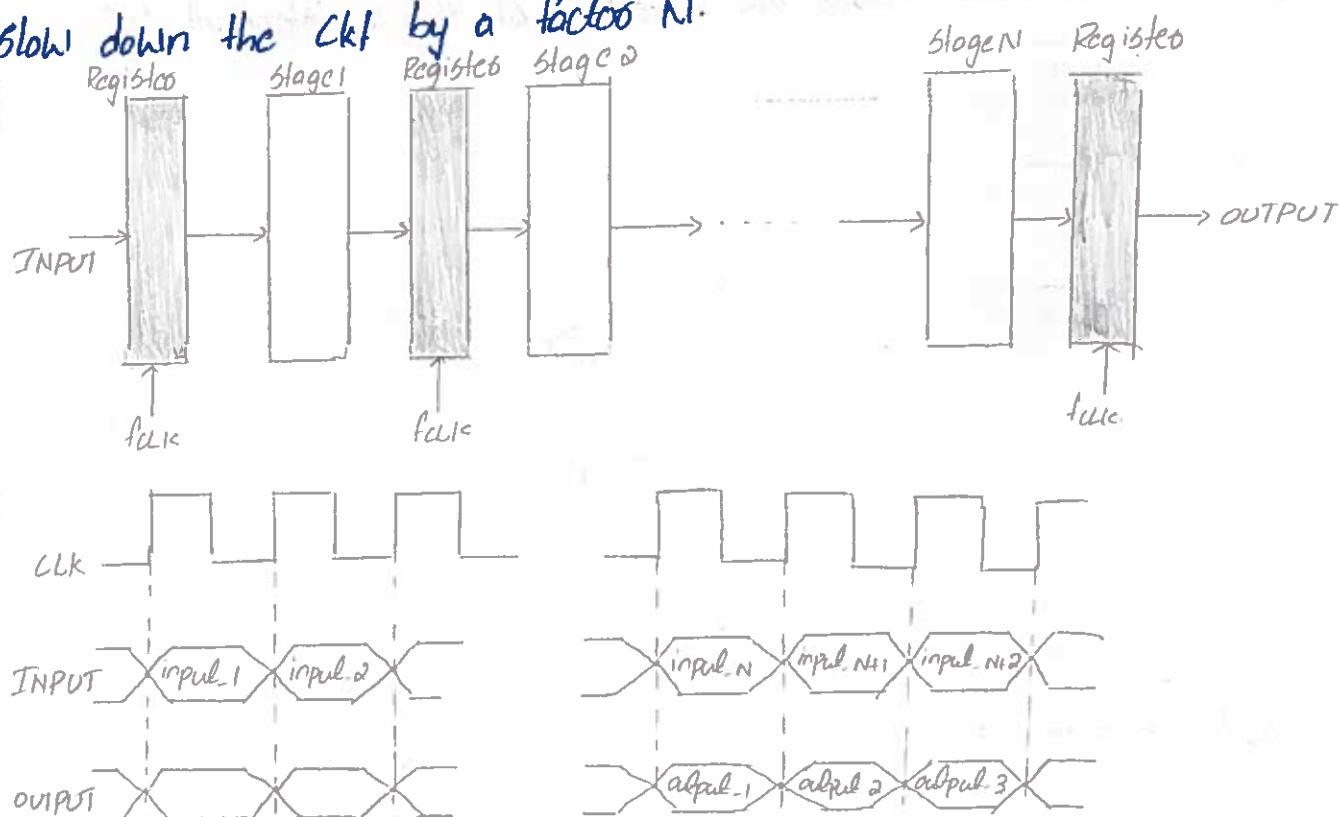
The logic function  $F(\text{INPUT})$  has been partitioned into  $N$  successive stages and a total of  $(N-1)$  registers arrays have been introduced, in addition to the original i/p and o/p registers, to locate the pipeline. All registers are clocked at the original sample rate,  $f_{CLK}$ . If all stages of the partitioned function have approximately equal delays of

$$T_p(\text{pipeline-stage}) = \frac{T_p, \text{max(input-to-output)}}{N}$$

$$= T_{CLK}$$

Then the logic blocks b/w two successive registers can operate  $N$  times slower while maintaining the same functional throughput as before. This implies that the power supply voltage can be reduced to a value of  $V_{DD/N}$  to effectively

Slow down the Ckt by a factor  $N$ .



$N$  Stage pipeline structure realizing the same logic function.

The dynamic power consumption of the N-stage pipelined structure with a lower supply voltage and with the same functional throughput as the single stage structure can be approximated by

$$P_{\text{Pipeline}} = [C_{\text{total}} + (N-1)C_{\text{eq}}] \cdot V_{\text{DD,new}}^2 \cdot f_{\text{CLK}}$$

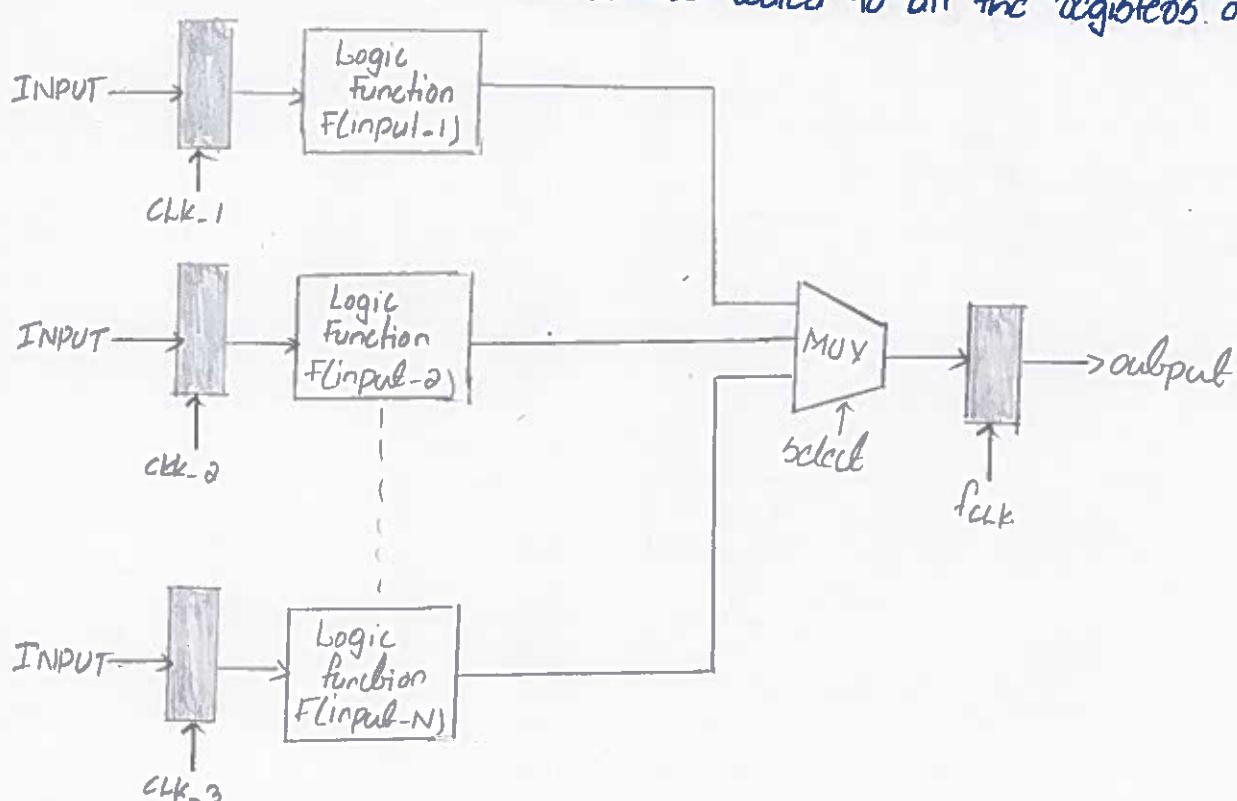
Where  $C_{\text{eq}}$  represents the capacitance switched by each pipeline register. Then, the power reduction factor achieved in a N stage pipeline structure is

$$\frac{P_{\text{Pipeline}}}{P_{\text{Performance}}} = \frac{[C_{\text{total}} + (N-1)C_{\text{eq}}]V_{\text{DD,new}}^2 f_{\text{CLK}}}{C_{\text{total}} \cdot V_{\text{DD}}^2 \cdot f_{\text{CLK}}} \\ = \left[ 1 + \frac{C_{\text{eq}}(N-1)}{C_{\text{total}}} \right] \frac{V_{\text{DD,new}}^2}{V_{\text{DD}}^2}$$

### Parallel Processing Approach or Hardware Replication.

Another method for trading off area for lower power dissipation is to use parallelism or hardware replication. This approach could be useful especially when the logic function to be implemented is not suitable for pipelining.

- \* Consider N identical processing elements, each implementing the logic function  $F(\text{INPUT})$  in parallel.
- \* Assume that the consecutive input vectors arrive at the same rate as in the single stage case. The input vectors are routed to all the registers of the N.

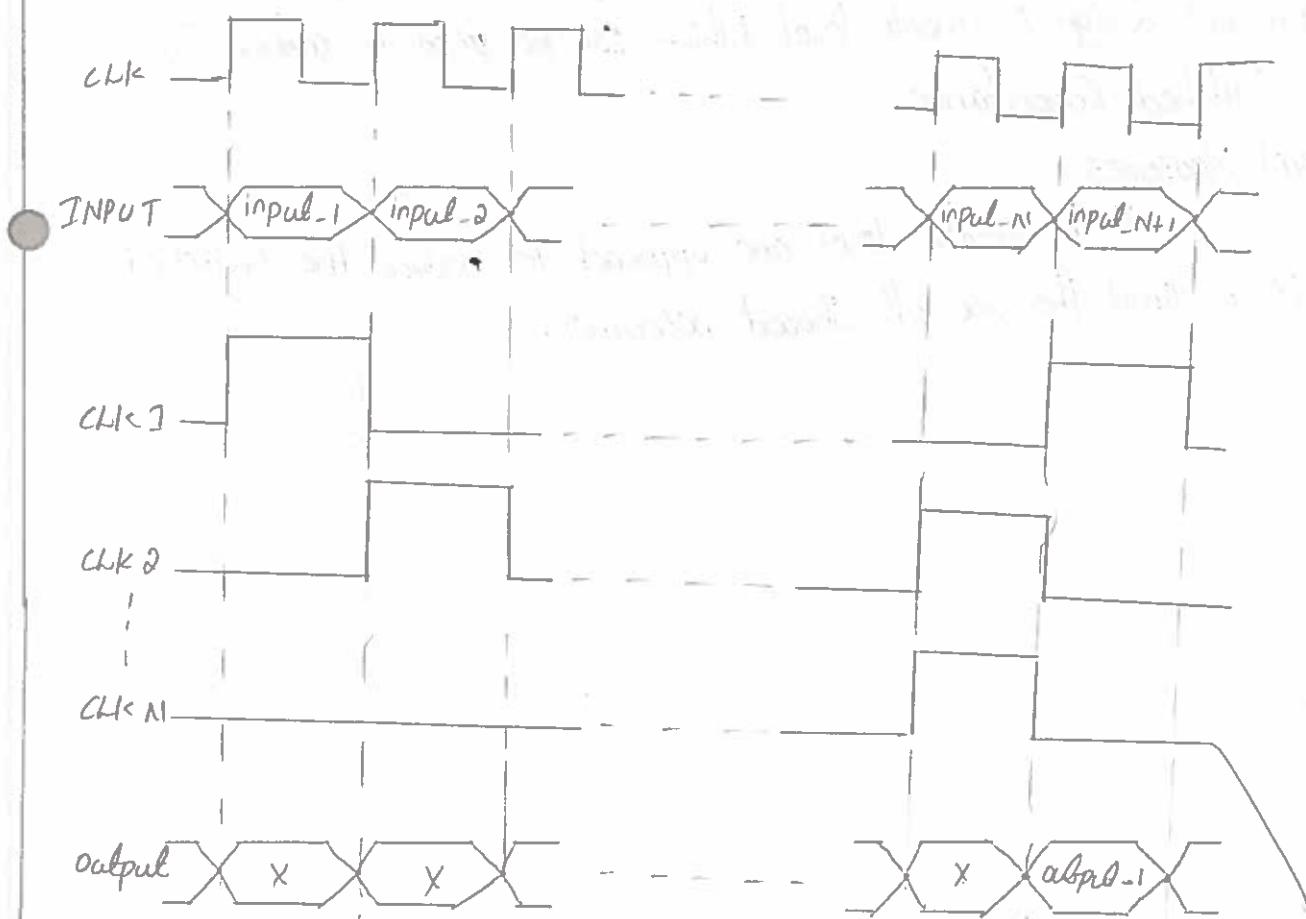


N-block parallel structure realizing the same logic function

- \* Grated clock signals, each with a clock period of, are used to load each register every N clock cycles.
- \* This means that the clock signals to each input registers are skewed by  $T_{CLK}$ , such that each one of the N consecutive ip Vectors is loaded into a different ip registers.
- \* The outputs of the N processing blocks are multiplexed and sent to an output registers which operates at a clock frequency of  $clk$  ensuring the same data throughput rate as before.
- \* Since the time allowed to Compute the function for each input Vector is increased by a factor of N, the power supply Voltage can be reduced to a value of  $V_{DD}$  new to effectively slow down the Circuit.
- \* The total dynamic power dissipation of the parallel structure is found as the sum of the powers dissipated by the ip registers and the logic blocks operating at a clock frequency, and the op registers operating at a clock frequency off  $clk$ .

$$P_{parallel} = N \cdot C_{total} \cdot V_{DDnew}^2 \cdot \frac{f_{CLK}}{N} + C_{op} \cdot V_{DDnew}^2 \cdot f_{CLK}$$

$$= \left[ 1 + \frac{C_{op}}{C_{total}} \right] C_{total} \cdot V_{DDnew}^2 + f_{CLK}.$$



Simplified timing diagram of the N-block parallel structure.

Note that  $\theta_{\text{soc}}$  is also an additional overhead which consists of the i/p routing capacitance, all of which are increasing functions of  $N$ . If this overhead is neglected, the amount of power reduction achievable in a  $N$ -block parallel implementation is

$$\frac{P_{\text{parallel}}}{P_{\text{reference}}} = \frac{V_{DD\text{new}}^2}{V_{DD}^2} \left[ 1 + \frac{C_{\text{req}}}{C_{\text{total}}} \right]$$

The lower bound of switching power reduction realizable with architecture driven Voltage Scaling is found, assuming zero threshold voltage, as

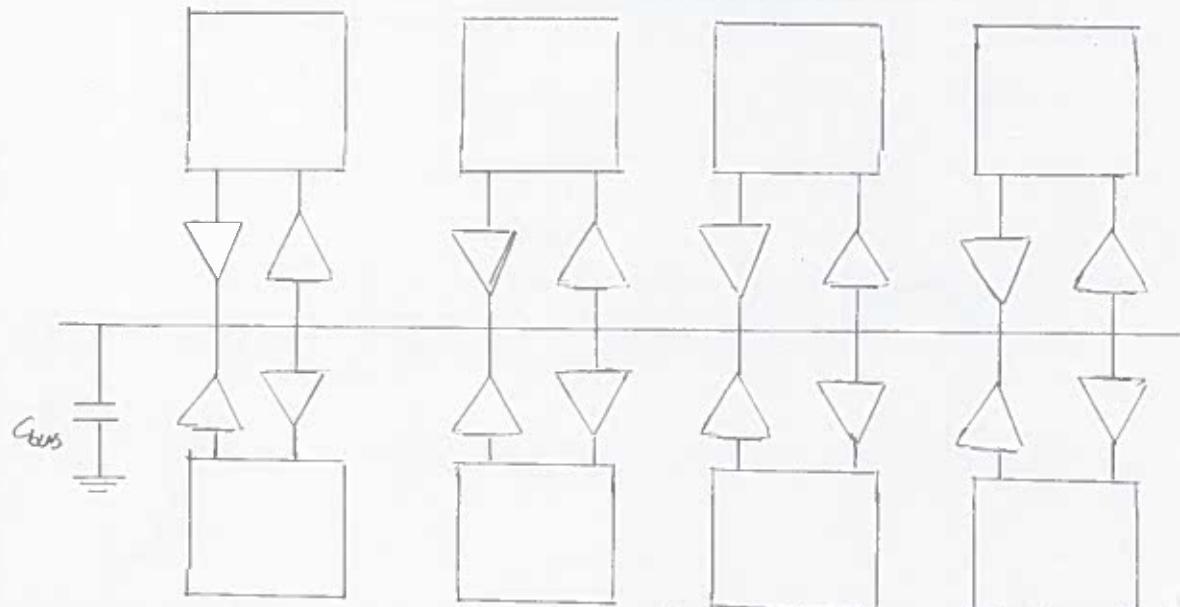
$$\frac{P_{\text{parallel}}}{P_{\text{reference}}} \geq \frac{1}{N^2}$$

### Reduction of Switched Capacitance:

- \* It was already established in the previous sections that the amount of switched capacitance plays a significant role in the dynamic power dissipation of the circuit.
- \* Hence, reduction of this parasitic capacitance is a major goal for low power design of digital integrated circuits.
- \* In this section, we will consider various techniques at the system level, circuit level and physical design (or) mask level which can be used to reduce the amount of switched capacitance.

### System-Level Measures:

At the system level, one approach to reduce the switched capacitance is to limit the use of shared resources.



Using a single global bus structure for connecting a large number of modules on chip results in large bus capacitance & large dynamic power dissipation.

## Circuit-Level Measures:

The type of logic style used to implement a digital Circuit also affects the output load Capacitance of the Circuit. The Capacitance is a function of the number of transistors that are required to implement a given function.

For example, one approach to reduce the load Capacitance is to use transfer gates instead of conventional CMOS logic gates to implement logic functions. pass gate logic design is attractive since fewer transistors are required for certain functions such as XOR and XNOR.

Therefore, this design style has emerged as a promising alternative to conventional CMOS, for low power design. Still a number of important issues must be considered for pass gate logic.

## Mask level Measures:

The amount of parasitic Capacitance that is switched during operation can be also reduced at the physical design level or mask level. The parasitic gate and diffusion Capacitances of MOS transistors in the Circuit typically constitute a significant amount of the total Capacitance in a Combinational logic Circuit.

Hence, a simple mask-level measure to reduce power dissipation is keeping the transistors at minimum dimensions whenever possible and feasible, thereby minimizing the parasitic Capacitances. Designing a logic gate with minimum size transistors certainly affects the dynamic performance of the Circuit, and this trade off b/w dynamic performance and power dissipation should be carefully considered in Critical Circuits.

You in the office & I am now at home and have had  
An attack of a cold. It has got to the point where I can't  
even eat or drink. I have been trying to get some sleep  
but I just can't seem to get it. I am really tired and  
my head hurts like crazy. I think I might have a

flu. I have a sore throat and my nose is all clogged up.  
I think I should go to the doctor and get some medicine.

I am not sure if I should go to work tomorrow. I am still  
feeling pretty bad and I don't know if I will be able to  
get through the day without feeling worse.

I am going to try and get some rest tonight. I am not  
sure if I will be able to sleep though. I am still feeling  
kind of tired and my head is still hurting.

I am going to try and get some rest tonight. I am not  
sure if I will be able to sleep though. I am still feeling