UNIT-II

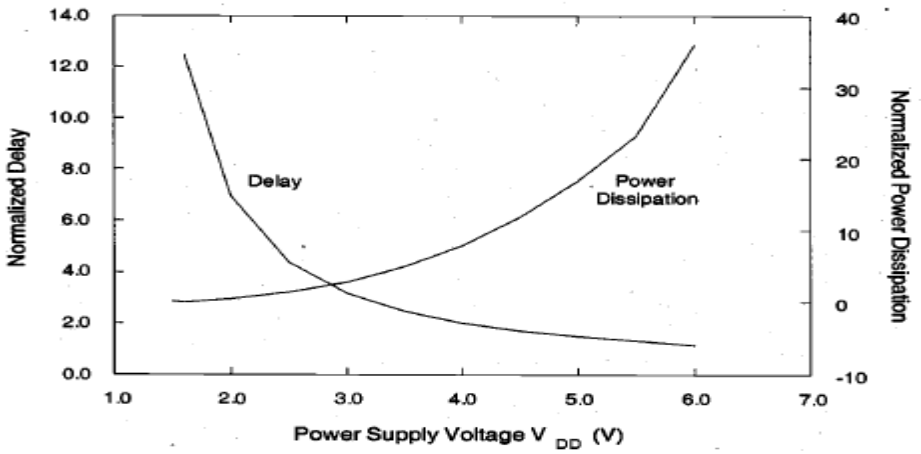
LOW POWER VLSI DESIGN APPROACHES

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 VDD 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 3.3 V, the issue of time-domain performance should also be addressed carefully. Reduction of the power supply voltage with a corresponding scaling of threshold voltages, in order to compensate for the speed degradation. *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,

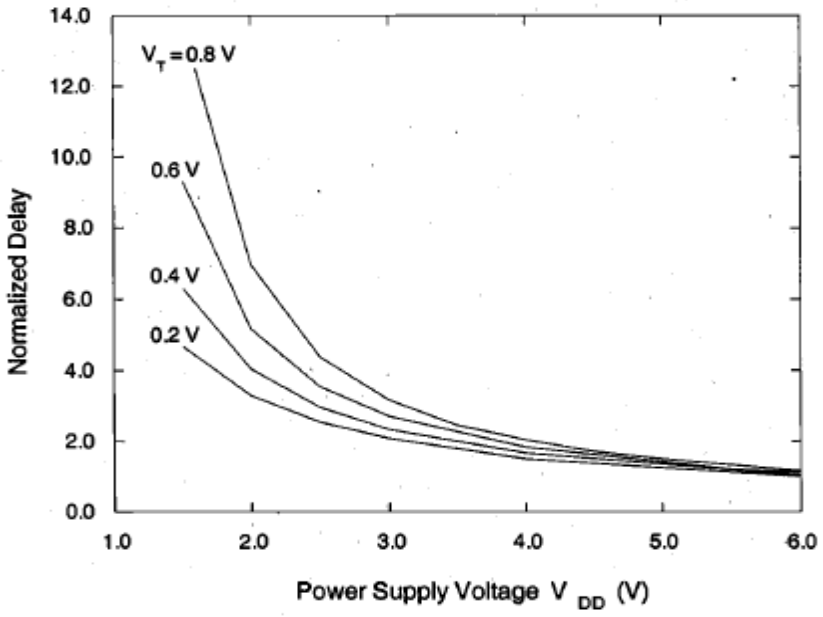
$$\tau_{PHL} = \frac{C_{load}}{k_n\left(V_{DD} - V_{T,n}\right)}\left[\frac{2V_{T,n}}{V_{DD} - V_{T,n}} + ln\left(\frac{4\left(V_{DD} - V_{T,n}\right)}{V_{DD}} - 1\right)\right]$$

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



the dependence of circuit speed on the power supply voltage may also influence the relationship between the dynamic power dissipation and the supply voltage. The above equation suggests a quadratic improvement (reduction) of power consumption as the power supply voltage is reduced. However, this interpretation assumes that the switching frequency (i.e., the number of switching events per unit time) 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 (i.e., the operating frequency) will drop as the propagation delay becomes larger with the reduction of the power supply voltage. The net result is that the dependence of switching power dissipation on the power supply voltage becomes stronger than a simple quadratic relationship, shown in Figure: It is important to note that the voltage scaling is distinctly different from *constant-field scaling* , where the power supply voltage as well as the critical device dimensions (channel length, gate oxide thickness) and doping densities are scaled by the same factor. 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 expressions show that the negative effect of reducing the power supply voltage upon delay can be compensated for, if the threshold voltage of the transistors *(VT)* is scaled down accordingly. However, this approach is limited because the threshold voltage may not be scaled to the same extent as the supply voltage. When scaled linearly, reduced threshold voltages allow the circuit to produce the same speed-performance at a lower ***VDD.*** Figure shows the variation of the propagation delay of a CMOS inverter as a function of the power supply voltage, and for different threshold voltage values.

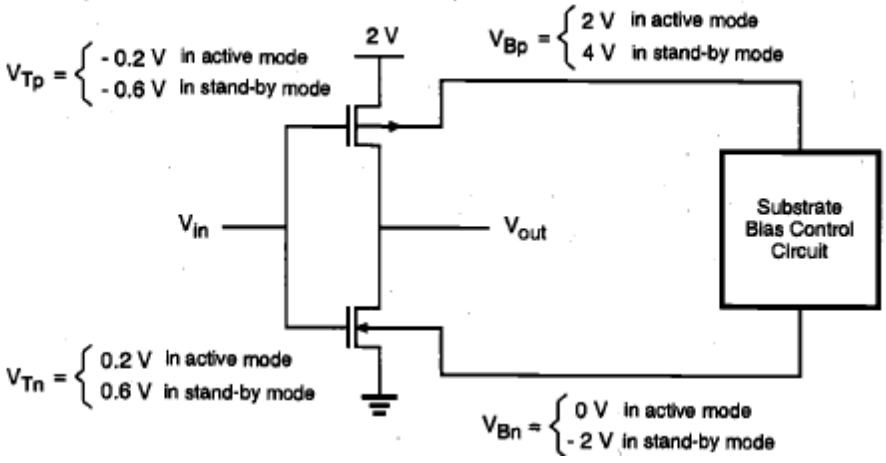


The reduction of threshold voltage from 0.8 V to 0.2 V can improve the delay at *VDD*= 2 V by a factor of 2. The positive influence of threshold voltage reduction upon propagation delay is

specially pronounced at low power supply voltages, for *VDD* < 2 V. It should be noted, however, that using low- V_T transistors raises significant concerns about noise margins and sub-threshold conduction. Smaller threshold voltages lead to smaller noise margins for the CMOS logic gates. The sub-threshold conduction current also sets a severe limitation against reducing the threshold voltage. For threshold voltages smaller than 0.2 V, leakage due to sub-threshold conduction in stand-by, i.e., when the gate is not switching, may become a very significant component of the overall power consumption. 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 (such as leakage and high stand-by power dissipation) 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 *(VDD)* and a low threshold voltage (VT) 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 output is not switching. One possible way to overcome this problem is to *adjust* the threshold voltages of the transistors in order to avoid leakage in the stand-by mode, by changing the substrate bias.
The threshold voltage VT of an MOS transistor is a function of its source-to-substrate voltage *VSB,* 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 onnected to *VDD.* This ensures

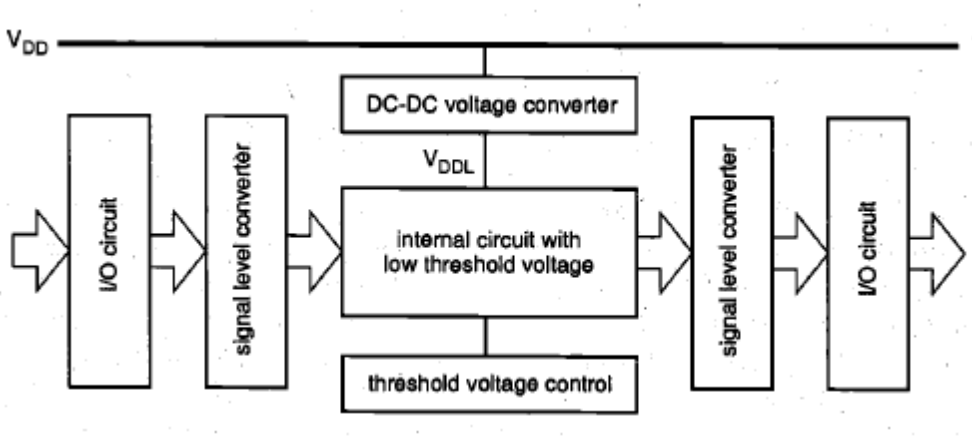


that the source and drain diffusion regions always remain reverse-biased with respect to the substrate, and that the threshold voltages of the transistors are not significantly influenced by the body (back gate-bias) effect. In VTCMOS circuit technique, on the other hand, the transistors are designed 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, as shown in Fig.

When the inverter circuit is operating in its *active* mode, the substrate bias voltage of the nMOS transistor is *VOn* = 0 and the the substrate bias voltage of the pMOS transistor is ***VBP*** = *VDD.* Thus, the inverter transistors do not experience any back gate-bias effect. The circuit operates with low *VDD and* low ***VT,*** benefiting from both low power dissipation (due to low *VDD)* and high switching speed (due to low ***VT).*** When the inverter circuit is in the *stand-by* mode, however, the substrate bias control circuit 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 *VTl* and ***VT,*** 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 Scheme (SATS).

The variable-threshold CMOS circuit design techniques are very effective for reducing the sub-threshold leakage currents and for controlling threshold voltage values in low *VDD* - **low** VT applications. However, this technique usually requires twin-well or triple-well CMOS technology in order to apply different substrate bias voltages to different parts of the chip. Also, separate power pins may be required if the substrate bias voltage levels are not generated on-chip. The additional area occupied by the substrate bias control circuitry is usually negligible compared to the overall chip area.



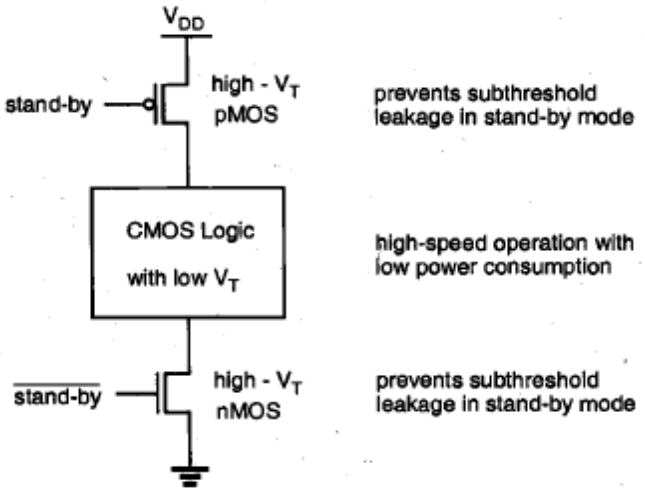
Block diagram of a typical low-power chip.

The internal supply voltage is generated on-chip, by a DC-DC converter circuit.

circuits of the chip usually operate with a higher external supply voltage, in order to increase the noise margins and to enable communication with the peripheral devices. An on-chip DC-DC voltage converter generates the low internal supply voltage ***VDDL,*** which is used by the internal circuitry. Two signal swing converters (level converters) are used to reduce the voltage swing of the incoming input signals, and to increase the voltage swing of the outgoing output signals, respectively. The internal low-voltage circuitry can be designed using VTCMOS techniques, where the threshold voltage control unit adjust the substrate bias in order to suppress leakage currents.

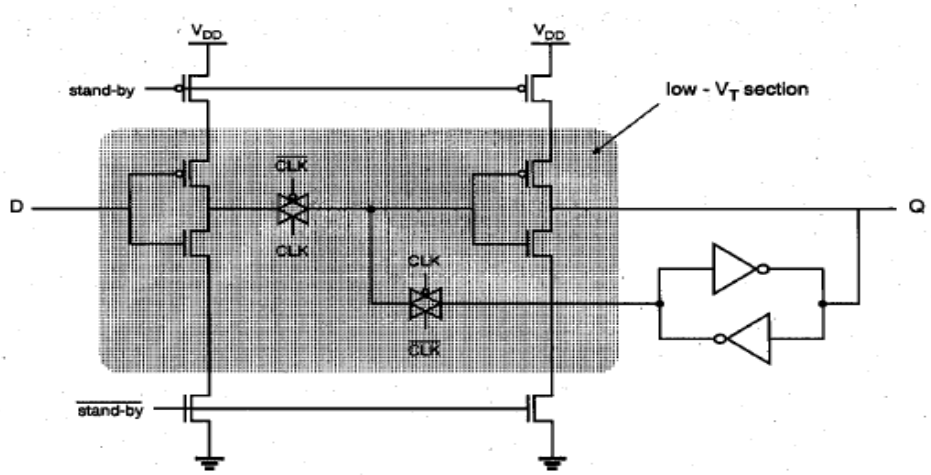
Multiple-Threshold CMOS (MTCMOS) Circuits

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 (both n-MOS and p-MOS) with two different threshold voltages in the circuit. Here, low-VT transistors are typically used to design the logic gates where switching speed is essential, whereas high- VT transistors are used to effectively isolate the logic gates in stand-by and to prevent leakage dissipation. The generic circuit structure of the MTCMOS logic gate is shown



Generic structure of a multiple-threshold CMOS (MTCMOS) logic gate.

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 circuit is driven into stand-by mode, on the other hand, the high-VT 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. Figure shows a simple D-latch circuit designed with the MTCMOS technique. The critical signal propagation path from the input to the output consists exclusively of low- V_T transistors, while a cross-coupled inverter pair consisting of high-V_T transistors is used for preserving the data in the stand-by mode.

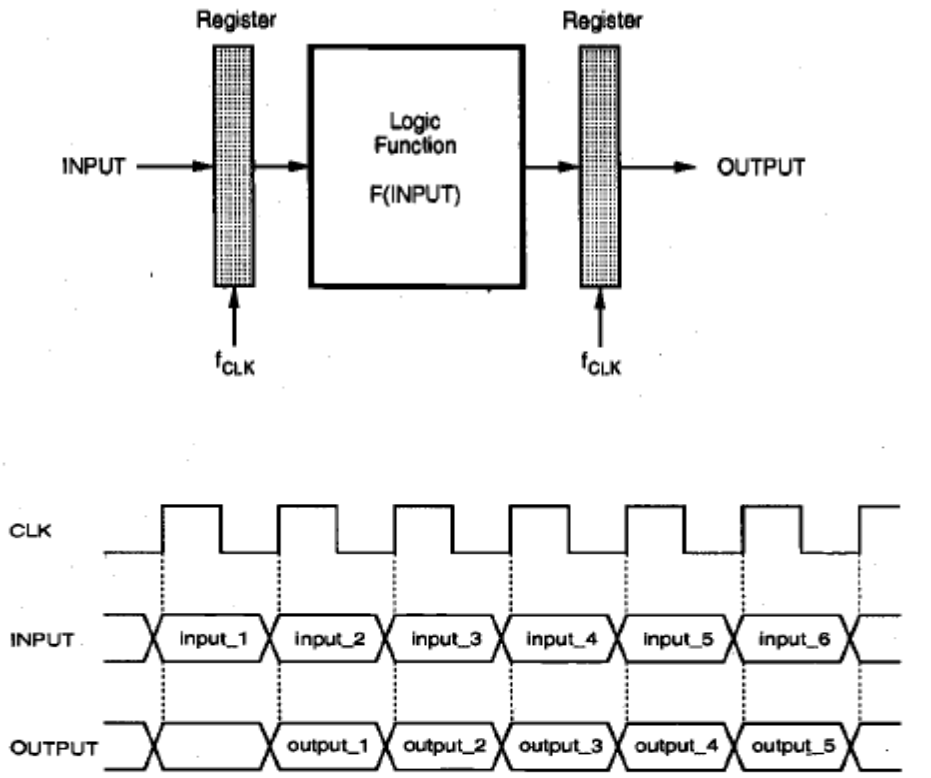


Low-power/low-voltage D-latch circuit designed with MTCMOS technique.

The MTCMOS technique is conceptually easier to apply and to use compared to the VTCMOS technique, which usually requires a sophisticated substrate bias control mechanism. It does not require a twin-well or triple-well CMOS process; the only significant process-related overhead of MTCMOS circuits is the fabrication of MOS transistors with different threshold voltages on the same chip. One of the disadvantages of the MTCMOS circuit technique is the presence of series-connected stand-by transistors, which increase the overall circuit area and also add extra parasitic capacitance. While the VTCMOS and MTCMOS circuit techniques can be very effective in designing low-power/low-voltage logic gates, they may not be used as a universal solution to low-power CMOS logic design. In certain types of applications where variable threshold voltages and multiple threshold voltages are infeasible due to technological limitations, system-level architectural measures such as pipelining and hardware replication techniques offer feasible alternatives for maintaining the system performance (throughput) despite voltage scaling.

Pipelining Approach

First, consider the single functional block shown in Fig. which implements a logic function **F(INPUT)** of the input vector, **INPUT.** Both the input and the output vectors are sampled through register arrays, driven by a clock signal CLK. Assume that the critical path in this logic block (at a power supply voltage of *VDD)* allows a maximum sampling frequency off CLK; in other words, the maximum input-to-output propagation delay *pmax* of this logic block is equal to or less than *TCLK = lfCLK.* Figure shows a simplified timing diagram of the circuit. A new input vector is latched into the input register array at each clock cycle, and the output data becomes valid with a latency of one cycle.



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

Let *C total* be the total capacitance switched every clock cycle. Here, *C total,* consists of (i)the capacitance switched in the input register array, (ii) the capacitance switched to implement the logic function, and (iii) the capacitance switched in the output register array. Then, the dynamic power consumption of this structure can be found as

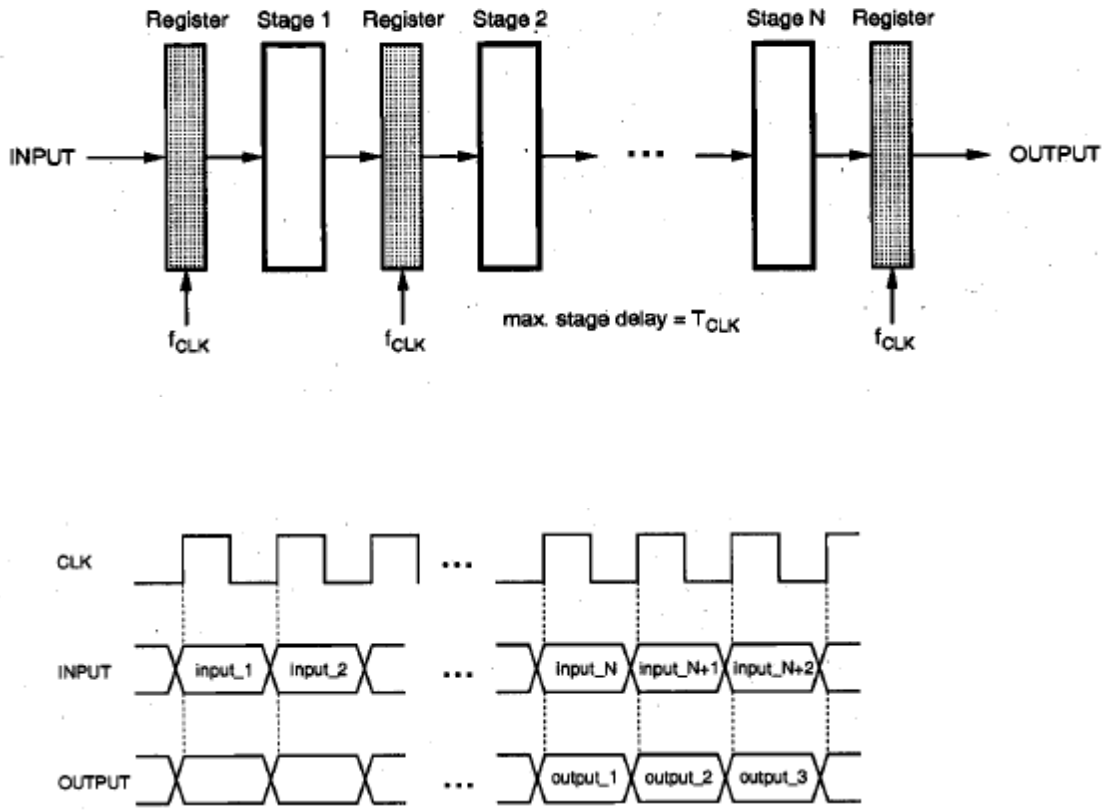
$$P_{reference} = C_{total} \cdot {V_{DD}}^2 \cdot f_{CLK}$$

The logic function F(INPUT) has been partitioned into N successive stages, and a total of (N- 1) register arrays have been introduced, in addition to the original input and output registers, to create the pipeline. All registers are clocked at the original sample rate, *fCLK.* If all stages of the partitioned function have approximately equal delays of

$$\tau_P(pipeline\_stage) = \frac{\tau_{P,max}(input - to - output)}{N} = T_{CLK}$$

Then the logic blocks between 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 *VDD fnew'* to effectively slow down the circuit by a factor N



N-stage pipeline structure realizing the same logic function as shown in Fig. The maximum pipeline stage delay is equal to the clock period, and the latency is N clock cycles.

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

$$P_{pipeline} = \left[C_{total} + (N-1)C_{reg}\right] \cdot V_{DD,new}{}^2 \cdot f_{CLK}$$

where Creg represents the capacitance switched by each pipeline register. Then, the power

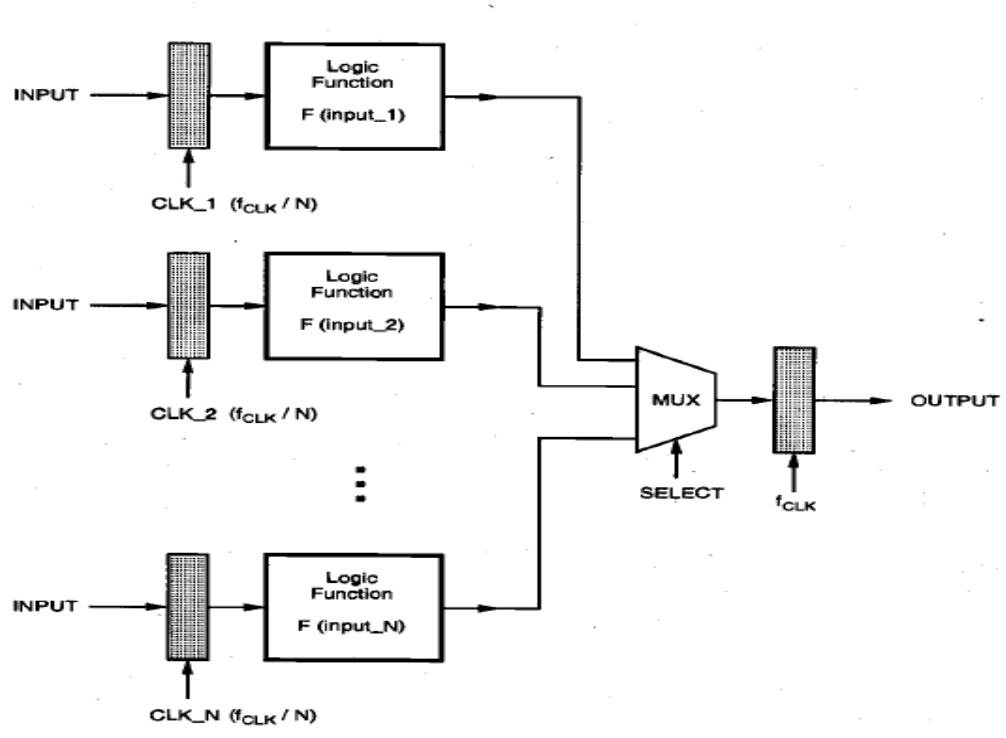
reduction factor achieved in a N-stage pipeline structure is

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

As an example, consider replacing a single-stage logic block *(VDD =5 V' fCLK* = 20 MHz) with a four-stage pipeline structure, running at the same clock frequency. This means that the propagation delay of each pipeline stage can be increased by a factor of 4 without sacrificing the data throughput. Assuming that the magnitude of the threshold voltage of all transistors is 0.8 V, the target speed reduction can be achieved by reducing the power supply voltage from *5* V to approximately 2V With *(Creg/Ctotal) =0.1,* the overall power reduction factor is as 0.2. This means that replacing the original single-stage logic block with a four-stage pipeline running at the same clock frequency and reducing the power supply voltage from *5* V to 2 V will provide a switching power saving of about 80%, while maintaining the same throughput as before. The architectural modification described here has a relatively small area overhead. A total of (N-1) register arrays have to be added to convert the original single-stage structure into a pipeline. While trading off area for lower power, this approach also increases the latency from one to N clock cycles. Yet in many applications such as signal processing and data encoding, latency is not a very significant concern.

Parallel Processing Approach (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(INPUT) in parallel, as shown in Fig. 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



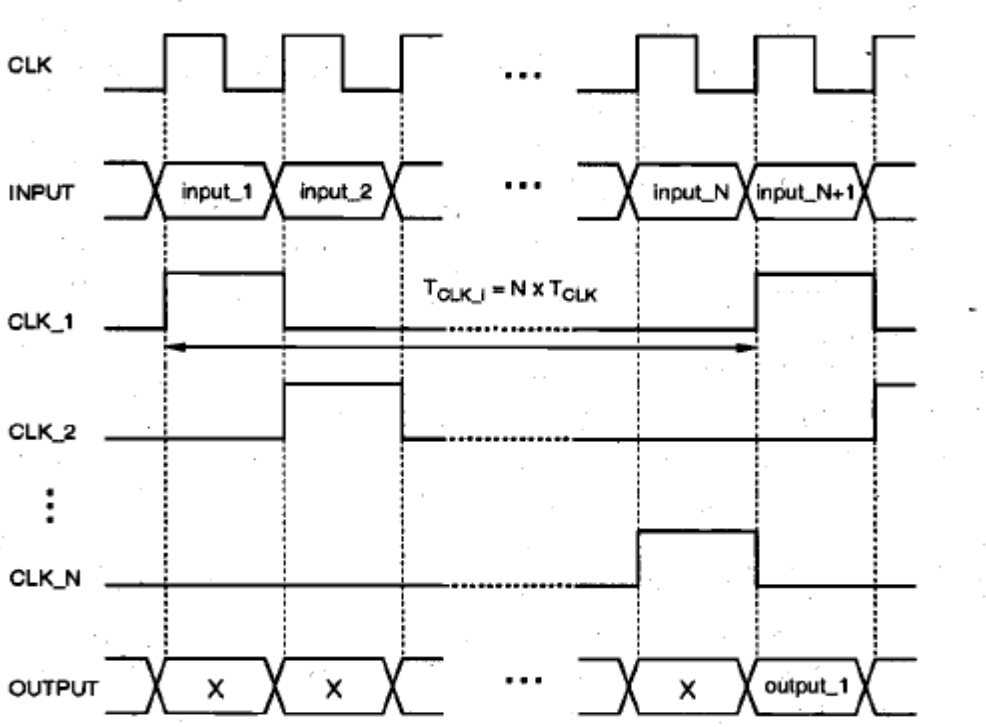
N-blockparallel structure realizing the same logic function as in Fig..
The input registers are clocked at a lower frequency of *(fcL* / N).

processing blocks. Gated clock signals, each with a clock period of (N *TCLK),* are used to load each register every N clock cycles. This means that the clock signals to each input register are skewed by *TCLK,* such that each one of the N consecutive input vectors is loaded into a different input register. Since each input register is clocked at a lower frequency of *(VCLK* I N), the time allowed to compute the function for each input vector is increased by a factor of N. This implies that the power supply voltage can be reduced until the critical path delay equals the, new clock period of (N *TCLK).* The outputs of the N processing blocks are multiplexed and sent to an output register which operates at a clock frequency of ***CLK'*** ensuring the same data throughput rate as before. The timing diagram of this parallel arrangement is given in Fig.

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 *VDD newl* to effectively slow down the circuit. The new supply voltage can be found, as in the pipelined case. The total dynamic power dissipation of the parallel structure (neglecting the dissipation of the multiplexer) is found as the sum of the power dissipated by the input registers and the logic blocks operating at a clock frequency of *(fCLK* N), and the output register operating at a clock frequency *offCLK.*

$$P_{parallel} = N \cdot C_{total} \cdot V_{DD,new}{}^2 \cdot \frac{f_{CLK}}{N} + C_{reg} \cdot V_{DD,new}{}^2 \cdot f_{CLK}$$

$$= \left(1 + \frac{C_{reg}}{C_{total}}\right) \cdot C_{total} \cdot V_{DD,new}{}^2 \cdot f_{CLK}$$



Simplified timing diagram of the N-block parallel structure

Note that there is also an additional overhead which consists of the input 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

$$\frac{P_{parallel}}{P_{reference}} = \frac{V_{DD,new}^{\ 2}}{V_{DD}^{\ 2}} \cdot \left(1 + \frac{C_{reg}}{C_{total}}\right)$$

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

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

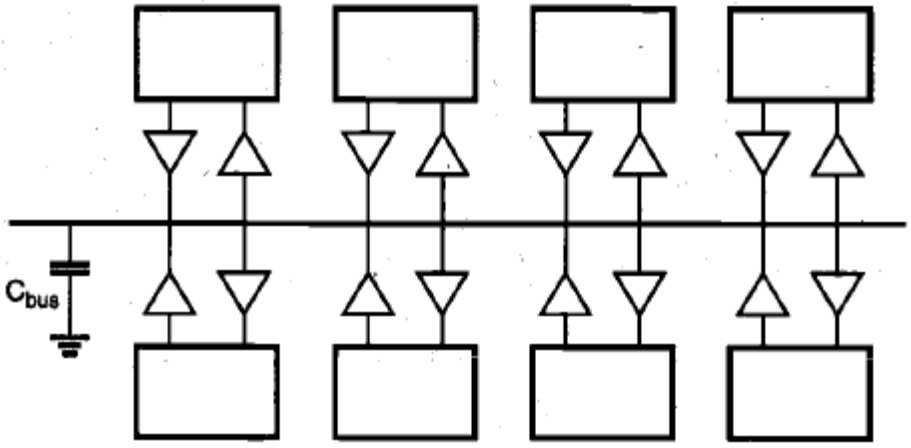
Two obvious consequences of this approach are the increased area and the increased latency. A total of N identical processing blocks must be used to slow down the operation (clocking) speed by a factor of N. In fact, the silicon area will grow even faster than the number of processors because of signal routing and the overhead circuitry. The timing diagram in Fig shows that the parallel implementation has a latency of N clock cycles, as in the N-stage pipelined implementation. Considering its smaller area overhead, however, the pipelined approach offers a more efficient alternative for reducing the power dissipation while maintaining the throughput.

Reduction of Switched Capacitance

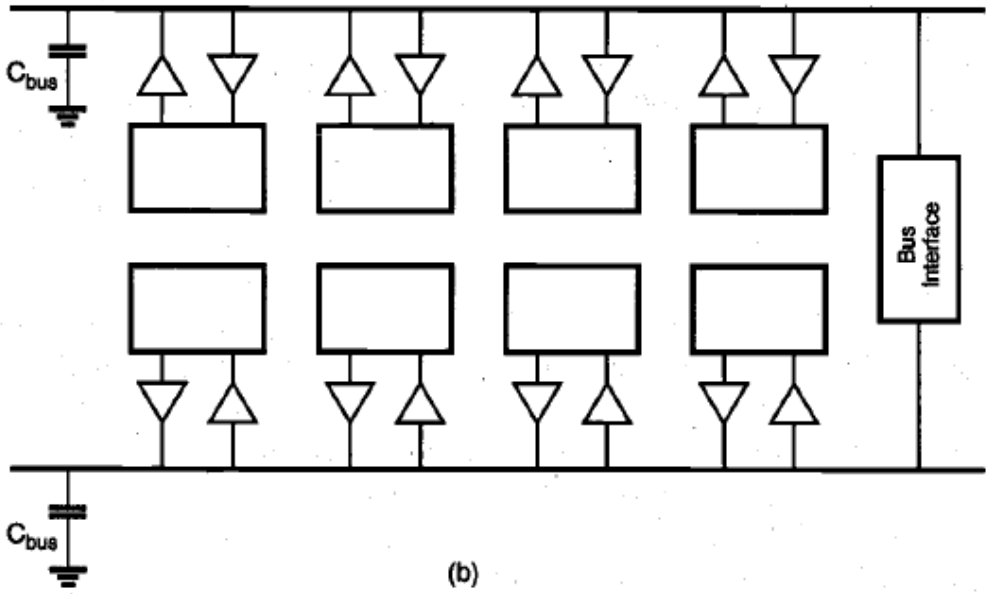
It was already established in the previous sections that the amount of switched capacitance plays a significant role in the dynamic power dissipation of the circuit. Hence, reduction of this parasitic capacitance is a major goal for low-power design of digital integrated circuits. In this Section, we will consider various techniques at the system level, circuit level and physical design (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. A simple example is the use of a global bus structure for data transmission between a large number of operational modules . If a single shared bus is connected to all modules as in fig. this structure results in a large bus capacitance due to (i) the large number of drivers and receivers sharing the same transmission medium, and (ii) the parasitic capacitance of the long bus line. Obviously, driving the large bus capacitance will require a significant amount of power consumption during each bus access. Alternatively, the global bus structure can be partitioned into a number of smaller dedicated local buses to handle the data transmission between neighboring modules, as shown in Fig. In this case, the switched capacitance during each bus access is significantly reduced, although multiple buses may increase the overall routing area on the chip.



(a)



(b)

(a)Using a single global bus structure for connecting a large number of modules on chip results in large bus capacitance and large dynamic power dissipation.
(b) Using smaller local buses reduces the amount of switched capacitance, at the expense of additional chip area.

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 (pass-transistor logic) 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.

The threshold-voltage drop through n-MOS transistors while transmitting a logic " 1 " makes swing restoration necessary in order to avoid static currents in subsequent inverter stages or logic gates (cf. Chapter 9). In order to provide acceptable output driving capabilities, inverters are usually attached to pass-gate outputs, which increases the overall area, time delay and the switching power dissipation of the logic gate. Because pass-transistor structures typically require complementary control signals, dual-rail logic is used to provide all signals in complementary form. As a consequence, two complementary n-MOS pass-transistor networks are necessary in addition to swing restoration and output buffering circuitry, effectively diminishing the inherent advantages of pass transistor logic over conventional CMOS logic. Thus, the use of pass-transistor logic gates to achieve low power dissipation must be carefully considered, and the choice of logic design style must ultimately be based on a detailed comparison of all design aspects such as silicon area, overall delay as well as switching power dissipation.

Mask-Level Measures

The amount of parasitic capacitance that is switched (i.e. charged up or charged down) 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 (especially the drain and source regions) 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 between dynamic performance and power dissipation should be carefully considered in critical circuits. Especially in circuits driving a large *extrinsic* capacitive loads, e.g., large fan-out or routing capacitances, the transistors must be designed with larger dimensions. Yet in many other cases where the load capacitance of a gate. is mainly *intrinsic,* the transistor sizes can be kept at a minimum. Note that most standard cell libraries are designed with larger transistors in order to accommodate a wide range of capacitive loads and performance requirements. Consequently, a standard-cell based design may have considerable overhead in terms of switched capacitance in each cell.