International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 3, March - 2013

"Stack Keeper Approach: A Novel Technique To Reduce Leakage Power In Cmos Vlsi Design"

Amreen parveen[1] , Amrita Shreevastava,[2] Mala Kushwah,[3]

Abstract—

Leakage power consumption plays a significant role in current CMOS technology. International Technology Roadmap for semiconductors reports that leakage power consumption dominates the total chip power consumption as technology advances to nanoscale. Most of the battery operated applications such as cell phones, Laptops etc requires a longer battery life, which can be made possible by controlling leakage current flowing through the CMOS gate. This paper presents leakage current mechanisms and different leakage reduction techniques to reduce leakage power consumption. We propose a novel leakage reduction technique named "Galeorstack" which can achieve better leakage reduction by maintaining exact logic state than the other techniques discussed in this paper. The proposed technique has been verified and compared with the other techniques for NOR and EXOR logic circuits and implemented using standard cells of 90nm CMOS process from CADENCE TOOLS. GaleorStack technique would be the best choice to the designer for the low leakage and less delay while achieving exact logic state

INTRODUCTION

The gate leakage currents with technology scaling, leakage power is increasingly significant in CMOS circuits as the technology scales down. The leakage power is as much as 50% of the total power in the 90nm technology and is becoming dominant in more advanced CMOS technologies with smaller feature sizes. Also, the leakage in active mode is significantly larger due to the higher die temperature in active mode. Although many leakage reduction techniques have been proposed, most of them can only reduce the circuit leakage power in standbymode. In this paper, we present a novel active leakage power reduction technique using dynamic power cutoff, called the dynamic power cutoff technique (DPCT). To reduce the active leakage power, we target the idle part of the circuit when it is in active mode. First, the switching window for each gate, during which a gate makes its transitions, is identified by static timing analysis. Then, the circuit is optimally partitioned into different groups based on the minimal switching window (MSW) of each gate. Finally, power cutoff transistors are inserted into each group to control the power connections of that group. The power of each gate is only turned on during a small timing

Window within each clock cycle, which results in significant active leakage power savings. Standby leakage can also be reduced by turning off the power connections of all gates all of the time once the circuit is idle. This technique also reduces dynamic power and short-circuit powe by reducing the circuit glitches.

In this work, we present a novel active leakage power reduction technique using dynamic power cutoff, called the dynamic power cutoff technique (DPCT). We propose a new minimal switching window (MSW) for CMOS gates to identify when the gate is active, which is equal to the worst-case delay of the gate. We propose a heuristic partitioning algorithm based on dynamic program to partition the circuit into groups based on the MSW of each gate so that the cost of adding extra power cutoff controls will be minimized without sacrificing much of the leakage power savings. We propose a six-step approach to implement DPCT. We also present the procedures to do power grid analysis and process variation analysis on DPCT.

IJERT

I. Power Dissipation in CMOS Circuits

There are three sources of power dissipation in CMOS digital circuits: dynamic power, short circuit power, and leakage power. Formerly, the dynamic power was dominant and the other two parts were negligible. But leakage power is becoming more and more significant as the CMOS technology goes into the deep submicron scale. Now, all three are important and leakage power is beginning to dominate.

Dynamic Power

Dynamic power is the power required to charge and discharge the load capacitances whentransistors switch. Suppose that we have a CMOS inverter with load capacitance *C*, which is shown in Figure 2.1. This charging and discharging process repeats $T f_{sw}$ times over an interval of *T*, where f_{sw} is the frequency of the input signal.

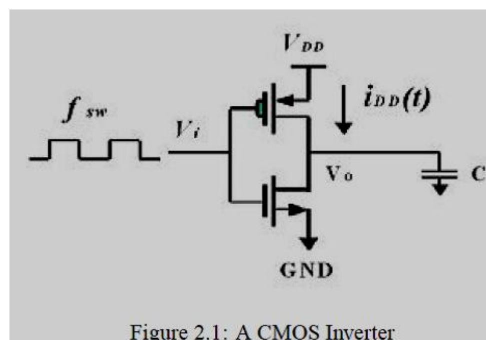

Figure 2.1: A CMOS Inverter

ISSN: 2278-0181
Vol. 2 Issue 3, March - 2013

So, the dynamic power can be calculated by the following formula:

$$P_{dynamic} = \frac{1}{T}\int_0^T i_{DD}(t)V_{DD}\,dt = \frac{V_{DD}}{T}\int_0^T i_{DD}(t)\,dt = \frac{V_{DD}}{T}(Tf_{sw}CV_{DD}) = CV_{DD}^2 f_{sw} \qquad (2.1)$$

The dynamic power can also be formalized as:

$$P_{dynamic} = \alpha C_L V_{DD}^2 f$$

where f is the clock frequency and α is the node transition activity factor.

Short-Circuit Power

When transistors switch, both *n*MOS and *p*MOS networks may be momentarily on at once. This leads to a blip of short circuit current. The short circuit power is given by:

$$P_{short-circuit} = I_{mean}V_{DD}$$

where I_{mean} is average short-circuit current. For a symmetric inverter shown in Figure 2.1,

$$I_{mean} = \frac{\beta}{12}(V_{DD} - 2V_t)^3\frac{t_{rf}}{t_p}$$

where V_{DD} is the power supply voltage, $V_t = V_{tn} = -V_{tp}$ is the threshold of the MOSFETs, $\beta = \beta_n = \beta_p$ is the β of the MOSFETs, $t_r = t_f = t_{rf}$ are the rising and falling times of the input pulse, and t_p is the period of the input pulse.

Leakage Power

Leakage power, also called static power, is due to the off-state current of a transistor when it is off. Suppose that there are *N* transistors in a circuit, and I_{offi} is the off-state current of the *i*th transistor. Then, the total leakage power of the circuit can be expressed in the following formula:

$$P_{leakage} = V_{DD}\sum_{i=1}^{N} I_{offi}$$

There are mainly six short-channel leakage mechanisms as illustrated in Figure 2.2 [34]. *I*1 is the reverse-bias *pn* junction leakage; I_2 is the sub-threshold leakage; I_3 is the oxide tunnelling current; I_4 is the gate current due to hot-carrier injection; I_5 is the *gate-induced drain leakage* (GIDL); and I_6 is the channel punchthrough current. Currents I_2, I_5, and I_6 are off-state leakage mechanisms, while I_1 and I_3 occur in both ON and OFF states. I_4 can occur in the off state, but more typically occurs when the transistor bias states are in transition.

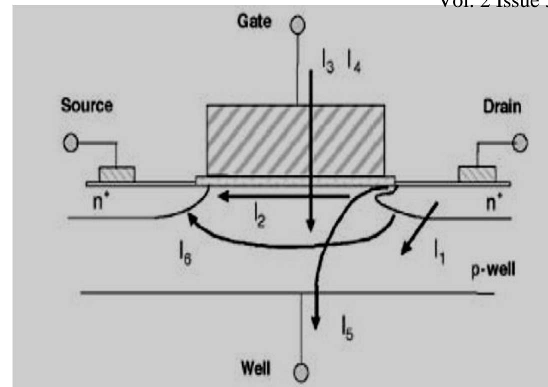

Figure 2.2: Leakage Current Mechanisms of Deep-Submicron Transistors

II. PN - Junction Reversed-Bias Current

Drain and source to well junctions are typically reverse biased, causing *pn* junction leakage current. The *pn* junction reverse-bias leakage is a function of junction area and doping concentration. If both *n* and *p* regions are heavily doped (this is the case for advanced MOSFETs using heavily doped shallow junctions and halo doping for better short channel effects (SCEs)), *band-to-band tunneling* (BTBT) dominates the *pn* junction leakage. The tunneling current density is given by:-

$$J_{b2b} = A\frac{EV_{app}}{\sqrt{E_g}}exp\left(-B\frac{E_g^{\frac{3}{2}}}{E}\right),\ A = \frac{\sqrt{2m^*}q^3}{4\pi^3h^2},\ and\ B = \frac{\sqrt{2m^*}}{3qh}$$

where m^* is effective mass of the electron; E_g is the energy band gap; V_{app} is the applied reverse bias; E is the electric field at the junction; q is the electronic charge; and h is Planck's constant.

III. Subthreshold Leakage Current

Subthreshold or weak inversion conduction current between source and drain in a MOS transistor occurs when the gate voltage is below V_{th}. It typically dominates modern device off-state leakage. The weak inversion current can be expressed based on the following :-

$$I_{ds} = I_{ds0}e^{\frac{V_{gs}-V_{th}}{nv_T}}\left(1 - e^{\frac{-V_{ds}}{v_T}}\right)$$

$$I_{ds0} = \beta v_T^2 e^{1.8}$$

IJERT

where V_{th} is the threshold voltage; V_{gs} is gate-source voltage; V_{ds} is drain-source voltage; v_T is the thermal voltage; I_{ds0} is the current at the threshold and is dependent on process and device geometry; the $e_{1.8}$ term was found empirically; and n is a process-dependent term affected by the depletion region characteristics and is typically in the range of 1.4-1.5 for CMOS processes. The inverse of the slope of the log10 I_{ds} versus V_{gs} characteristic is called the *subthreshold swing* (St). Subthreshold slope indicates how effectively the transistor can be turned off (rate of decrease of I_{off}) when V_{gs} is decreased below V_{th}. S_t is given by Equation 2.8, where Cdm is the capacitance of the depletion layer, and Cox is the gate oxide capacitance.

$$S_t = 2.3\frac{kT}{q}(1+\frac{C_{dm}}{C_{ox}})$$

Many factors affect the subthreshold current, such as temperature, body effect, DIBL (drain induced barrier lowering), the narrow-width effect, the effect of channel length, and V_{th} rolloff.

IV. Leakage Power Reduction Techniques

The reduction in leakage current has to be achieved using both process and circuit-level techniques. At the process level, leakage reduction can be achieved by controlling the dimensions (length, oxide thickness, junction depth, etc.) and doping profiles in transistors. At the circuit level, threshold voltage and leakage current of transistors can be effectively controlled by controlling the voltages of different device terminals [drain, source, gate, and body (substrate)].

Device-Level Leakage Reduction Techniques

Well engineering is always used to improve short-channel characteristics. By changing the doping profile in the channel region, the distribution of the electric field and potential contours can be changed. The goal is to optimize the channel profiles to minimize the OFF-state leakage while maximizing the linear and saturated drive currents. Supersteep retrograde wells and halo implants have been used as a means to scale the channel length and increase the transistor drive current without causing an increase in the OFF-state leakage current.

Retrograde Doping

Retrograde channel doping is a vertically nonuniform, low-high channel doping. It is used to improve the *short channel effects* (SCEs) and to increase surface channel mobility by creating a low surface channel concentration followed by a highly doped subsurface region. The low surface concentration increases surface channel mobility by minimizing channel impurity scattering while the highly doped subsurface region acts as a barrier against punchthrough. Figure 2.6 shows a schematic band-bending diagram at the threshold condition of an extreme retrograde profile with an undoped surface layer of thickness. For the same gate depletion width, the surface electric field and the total depletion charge of an extreme retrograde channel is one-half that of a uniformly doped channel. This reduces the threshold voltage and improves mobility.

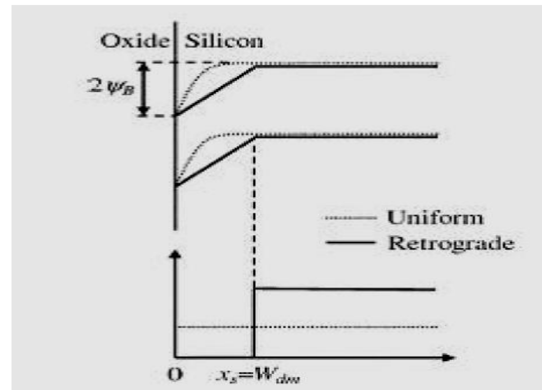

Figure 2.6: Band Diagrams (Shown on Top) at the Threshold Condition for a Uniformly Doped and an Extreme Retrograde-Doped Channel (Doping Profiles Shown at Bottom) .

Halo Doping

Halo doping or nonuniform channel profile in a lateral direction was introduced below the 0.25μm technology node to provide another way to control the dependence of threshold voltage on channel length. For n-channel MOSFETs, more highly p-type doped regions are introduced near the two ends of the channel as shown in Figure 2.7. Under the edges of the gate, in the vicinity of what will eventually become the end of the channel, point defects are injected during sidewall oxidation. These point defects gather doping impurities from the substrate, thereby increasing the doping concentration near the source and drain ends of the channel.

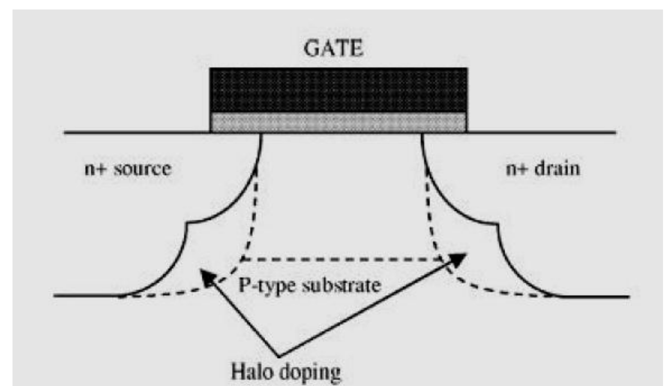

Figure 2.7: Halo or Nonuniform Channel Doping

A more highly doped p-type substrate near the edges of the channel reduces the charge-sharing effects from the source and drain fields, thus reducing the width of the depletion region in the drain-substrate and source-substrate regions.

IJERT

Reduction of charge-sharing effects reduces the threshold voltage degradation due to channel length reduction. Thus, threshold voltage dependence on channel length becomes more flat and the off-current becomes less sensitive to channel length variation. The reduction in drain and source junction depletion region widths also reduces the barrier lowering in the channel, thus reducing DIBL. Since the channel edges are more heavily doped and junction depletion widths are smaller, the distance between source and drain depletion regions is larger. This reduces the punchthrough possibility .

V.Circuit-Level Leakage Reduction Techniques

In this section, we will review six major circuit design techniques for leakage reduction in digital circuits: transistor stacking, input vector control, multiple V_{th}, supply voltage scaling (multiple and dynamic V_{DD}), power cut-off, and dynamic power-gating using the Shannon Expansion.

Transistor Stacking

Subthershold leakage current flowing through a stack of series-connected transistors reduces when more than one transistor in the stack is turned off. This effect is known as the stacking effect, which is shown in Figure 2.8. The technique of inserting an extra series connected transistor in the pulldown path of a gate and turning it off in the standby-mode of operation is known as forced stacking .

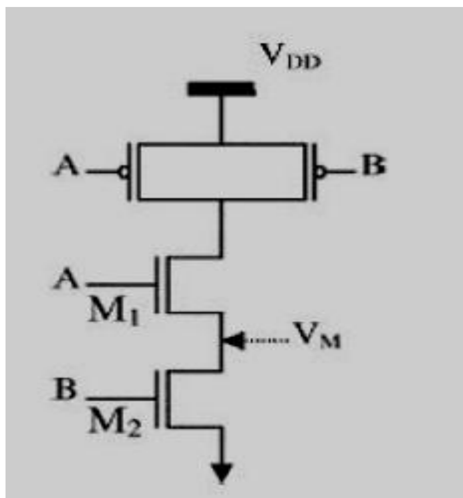

Figure 2.8: Stacking Effect in Two-Input NAND Gate

The extra transistor is turned on during the regular mode of operation and turned off during the idle mode of operation. When the extra transistor is turned off, the intermediate source voltage increases, which results in a decrease in the subthreshold current through the top transistor. Hence, the total subthreshold leakage through a two-transistor stack is reduced. Forced stacking only works for stand by leakage power reduction. Another way of using the stacking effect for leakage reduction is to replace a single transistor with two transistors of the same size. This is equivalent to replacing a low threshold transistor with a high threshold transistor in the dual-threshold transistor technique. Static timing analysis is needed to identify those gates on non-critical paths for possible insertion of stacking transistors. Similar algorithms as high threshold transistor insertion in the dual-threshold transistor technique can be used.for the detail introduction of the dual-threshold transistor technique.

Leakage Reduction by Input Vector Control

Due to the stacking effect, the subthreshold leakage through a logic gate depends on the applied input vector. This makes the total leakage current of a circuit dependent on the states of the primary inputs. It has been shown that the leakage current ratio between different input combinations can be as high as 10. The goal can then be expressed as finding the input pattern that maximizes the number of disabled (off) transistors in all stacks across the circuit. One possible way is to perform an exhaustive circuit-level simulation for all input patterns to find the pattern with the minimum leakage current. However, this approach is not practical for large circuits. Z. Chen *et al.* proposed a genetic algorithm to locate the vector that results in the near minimal leakage current. J. Halter and F. Najm used probabilistic methods to reduce the number of simulations necessary to find a solution with a desired accuracy. SAT-based formulation were also proposed for finding the minimum leakage vector at the circuit inputs.

Leakage Reduction by Multiple Threshold Voltage Designs

One way of decreasing the leakage current is to increase the threshold voltages of transistors. Multiple-threshold CMOS technologies, which provide both high- and low-threshold transistors in a single chip, can be used to deal with the leakage problem. The high-threshold transistors can suppress the subthreshold leakage current, while the low-threshold transistors are used to achieve high performance. Several multiple-threshold circuit design techniques have been developed recently, including multi-threshold CMOS, dual-threshold CMOS, variable threshold CMOS, and dynamic threshold CMOS.

Multi-Threshold Voltage CMOS.

Multi-threshold voltage CMOS (MTCMOS) reduces the leakage by inserting high-threshold devices in series with low-threshold circuitry. Figure 2.9 shows the schematic of an MTCMOS circuit. In the active mode, the sleep control transistors (MP and MN) are turned on.Since their on-resistances are small, the virtual supply voltages (*VDDV* and *VSSV*) almost function as real power lines. In the standby

IJERT

mode, MN and MP are turned off, and the leakage current is low. In fact, only one type of high transistor is enough for leakage control. Figures 2.9 (b) and (c) show the *p*MOS insertion and *n*MOS insertion schemes, respectively. The *n*MOS insertion scheme is preferable, since the *n*MOS on-resistance is smaller at the same width; therefore, it can be sized smaller than the corresponding *p*MOS. This technique is only effective for standby leakage power reduction.

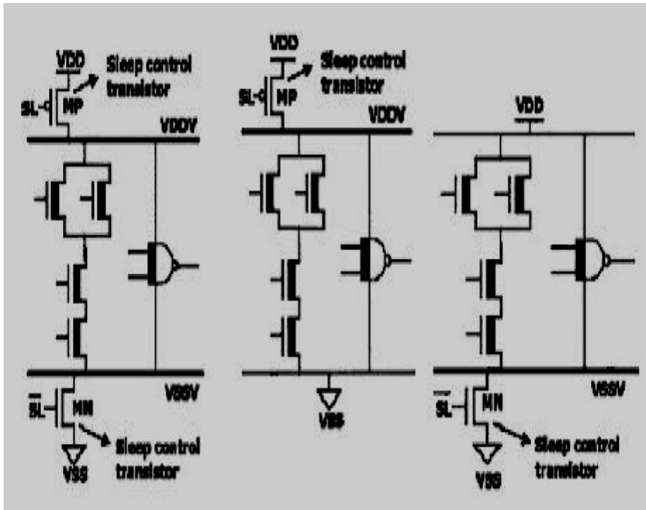

Figure 2.9: (a) Original MTCMOS (b) *p*MOS Insertion MTCMOS (c) *n*MOS Insertion MTCMOS

Dual-Threshold CMOS.

Another approach of MTCMOS is to use high-threshold voltage devices on noncritical paths to reduce the leakage power while using low-threshold devices on critical paths so that the circuit performance is maintained. This technique has been called dual-threshold CMOS. It is an integer linear program to choose an optimal assignment of dual-*Vth* for all of the transistors or gates in the circuit. Various heuristic algorithms are proposed to solve this problem for big circuits Dual-threshold CMOS is a very effective approach for leakage reduction in both active mode and standby mode. More than 80% of leakage power savings have been reported. Comparied with other leakage reduction techniques, it requires very little 18 modification of the circuit design. It can also be combined with transistor sizing and multiple *VDD* to get more leakage power savings. Y. Lu *et al.* combine dual-*Vth* assignment with path balancing using integer linear programing to reduce both leakage and dynamic glitch power simultaneously. Thus, dual-threshold CMOS is widely used in modern CMOS fabrication lines.

Variable Threshold CMOS

Variable threshold CMOS (VTMOS) is a technique, which uses the body bias voltage to change the threshold of CMOS transistors. It has been reported that reverse body biasing lowers integrated circuit leakage by three orders of magnitude in a 0.35μm technology. However, it was also shown that the effectiveness of reverse body bias in lowering leakage decreases as technology scales. This technology also requires routing the body grid, which will add to the overall chip area.

Dynamic Threshold CMOS.

In dynamic threshold *CMOS* (DTMOS), the threshold voltage is altered dynamically to suit the operating state of the circuit. It can be achieved by tying the gate and body together. DTMOS can be developed in bulk technologies by using triple wells. Doping engineering is needed to reduce the parasitic components. The supply voltage of DTMOS is limited by the diode built-in potential in bulk silicon technology. The *pn* diode between source and body should be reverse biased. Hence, this technique is only suitable for ultra-low voltage (0.6V and below) circuits in bulk CMOS. Another way for dynamic threshold design is to control the body bias voltage dynamically through a bias-control circuit depending on the workload of the system. When the workload becomes less, the bias control circuit will change the body bias to increase the threshold to reduce the power.

IJERT

VI. Leakage Reduction by Power Cut-Off

Instead of using low*VDD* for active mode and high *VDD* for standby mode, the power supply can be cut-off during the standby state and resumed during the active mode. This is called power cut-off technology. Two different power cut-off CMOStechnologies have been proposed: *super cut-off CMOS* (SCCMOS)and *zigzag super cut-off CMOS* (ZSCCMOS).

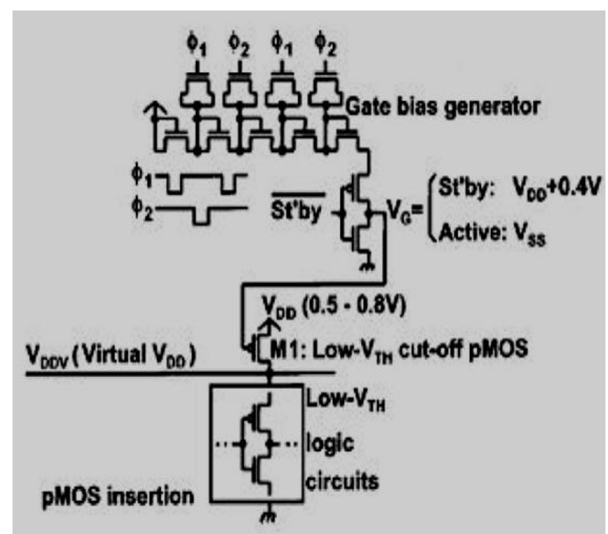

Figure 2.10: Concept of SCCMOS

The SCCMOS scheme was proposed and demonstrated to achieve high speed and low standby current with sub-1V supply voltages. In Figure 2.10, the low-V_{th} cut-off *p*MOS, M1, whose *Vth* is 0.1-0.2V, is inserted in

series to the logic circuits consisting of low-*Vth* MOSFETs. The gate voltage of M1, *VG*, is grounded in an active mode to turn M1 on. When the logic circuits enter standby operation, *VG* is overdriven to *VDD*+0.4V to completely cut off the leakage current. A problem associated with this scheme is that data can get lost during the long sleep period due to the leakage current. SCCMOS also suffers from a long wake-up time and a high current peak at the sleep-to-active transition. This is due to the virtual *VDD* node being discharged (charged) during the sleep period and being charged (discharged) when returning to active mode. A *zigzag super cut-off CMOS* (ZSCCMOS) method was then proposed to improve the operating speed by eliminating the series-connected switches while achieving the relaxation of the high-voltage stress at the cut-off switch. Tschanz *et al.* incorporated the power cut-off technology with the clock-gating scheme for leakage power reduction in a microprocessor. The gated-clock signal is used to synchronize the power cut-off controls of the respective circuit blocks, so that not only dynamic power but also leakage power can be reduced when the circuit block is in standby mode.

Dynamic Power Gating Using the Shannon Expansion

Bhunia *et al.* proposed an active leakage reduction technique using supply gating. They use the Shannon expansion to identify the idle part of the circuit and dynamically apply supply gating to those idle parts so that active leakage power is saved. Based on the Shannon expansion, each function $f(x1, x2, ..., xn)$ can be expanded into two parts based on variable xi

$$f(x_1, x_2, \ldots, x_n) = x_i CF_1 + x_i' CF_2$$
$$CF_1 = f(x_1, x_2, \ldots, x_i = 1, \ldots, x_n);$$
$$CF_2 = f(x_1, x_2, \ldots, x_i = 0, \ldots, x_n);$$

Based on the above expansion, $f(x1, x2, ..., xn)$ can be implemented in a circuit shown in Figure 2.11. For such an implementation, xi acts as a power gating signal for the circuit and only half the circuit is active at any time. In a big combinational circuit, all of the Boolean functions could be expanded by applying the Shannon expansion recursively and implemented into a similar circuit architecture. Thus, only a partial circuit will be active at any time and active leakage power is saved. With this technique, 15% to 88% total power reduction in MCNC benchmarks are reported.

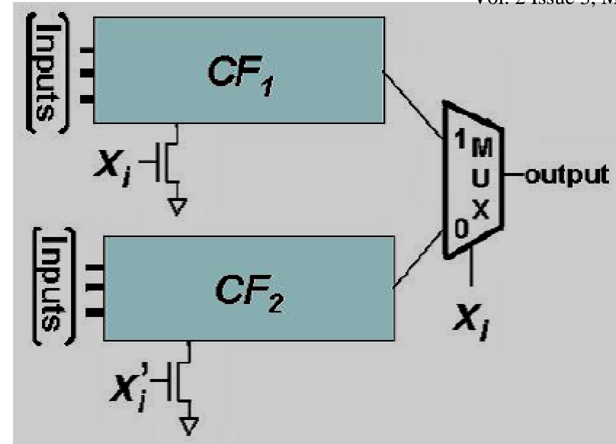

Figure 2.11: Dynamic Supply Gating Using the Shannon Expansion

VII. Conclusion

All of the techniques described above can be used to reduce the leakage power of a circuit in standby mode. However, even when the circuit is active, it still consumes a significant 21 amount of leakage in deep-submicron CMOS technologies. In fact, the leakage power in the active mode is significantly larger due to higher die temperature in active mode. Among the technologies we described above, dual-threshold CMOS, DTMOS, and all of the device-level techniques are effective for active leakage reduction. Most of the other schemes only work for standby leakage reduction. However, the dual-threshold technique does not reduce the leakage on critical paths. Thus, it does not help much for timing-optimized circuits, whose paths are usually well balanced. The DTMOS is usually achieved by tying the gate and body together. So, the supply voltage of DTMOS is limited by the diode built-in potential in bulk silicon technology. Hence, this technique is only suitable for ultra-low voltage (0.6V and below) circuits in bulk CMOS. Thus, more effective active leakage reduction techniques are still very desirable. Here we are proposing a new leakage reduction technology, called the *dynamic power cutoff technique*, which reduces both active leakage and standby leakage.

IJERT