LECTOR: A Technique for Leakage Reduction in CMOS Circuits

Narender Hanchate, *Student Member, IEEE,* and Nagarajan Ranganathan, *Fellow, IEEE*

***Abstract*—In CMOS circuits, the reduction of the threshold voltage due to voltage scaling leads to increase in subthreshold leakage current and hence static power dissipation. We propose a novel technique called LECTOR for designing CMOS gates which significantly cuts down the leakage current without increasing the dynamic power dissipation. In the proposed technique, we introduce two leakage control transistors (a p-type and a n-type) within the logic gate for which the gate terminal of each leakage control transistor (LCT) is controlled by the source of the other. In this arrangement, one of the LCTs is always "*near its cutoff voltage*" for any input combination. This increases the resistance of the path from V_{dd} to ground, leading to significant decrease in leakage currents. The gate-level netlist of the given circuit is first converted into a static CMOS complex gate implementation and then LCTs are introduced to obtain a leakage-controlled circuit. The significant feature of LECTOR is that it works effectively in both active and idle states of the circuit, resulting in better leakage reduction compared to other techniques. Further, the proposed technique overcomes the limitations posed by other existing methods for leakage reduction. Experimental results indicate an average leakage reduction of 79.4% for MCNC'91 benchmark circuits.**

***Index Terms*—Deep submicron, leakage power, power optimization, transistor stacking.**

I. Introduction

POWER dissipation is an important consideration in the design of CMOS VLSI circuits. High power consumption leads to reduction in the battery life in the case of battery-powered applications and affects reliability, packaging, and cooling costs. The main sources for power dissipation are: 1) capacitive power dissipation due to the charging and discharging of the load capacitance; 2) short-circuit currents due to the existence of a conducting path between the voltage supply and ground for the brief period during which a logic gate makes a transition; and 3) leakage current. The leakage current consists of reverse-bias diode currents and subthreshold currents. The former is due to the stored charge between the drain and bulk of active transistors while the latter is due to the carrier diffusion between the source and drain of the OFF transistors.

The short-circuit power dissipation can be reduced to 10% of total power dissipation by designing the circuit to have equal input and output rise/fall edge times [1]. The power dissipation resulting from switching activity is the dominant component for technology processes with feature size larger than 1 μm. With technology processes maturing toward the deep-submicron regime, the feature sizes of the transistors are becoming smaller, thereby reducing the load capacitances. The reduction in feature size also forces a reduction in the supply voltage. The voltage scaling technique takes benefit of the quadratic dependence of switching power on supply voltage for dynamic power savings. However, this technique pays a penalty for the operation of the circuit by increasing the delay drastically as supply voltage approaches the threshold voltage V_t of the devices [2]. In order to facilitate voltage scaling without affecting the performance, threshold voltage has to be reduced. In general, the ratio between the supply voltage and the threshold voltage should be at least 5, so that the performance of CMOS circuits is not affected [3]. This also leads to better noise margins and helps to avoid the hot-carrier effects in short-channel devices [4].

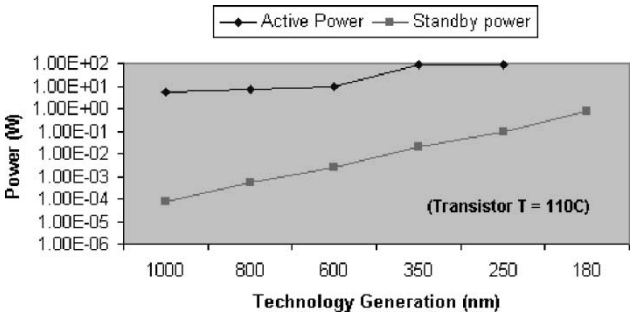

Fig. 1. Voltage supply and threshold voltage scaling trend.

Scaling down of threshold voltage V_t results in exponential increase of the subthreshold leakage current [5]. The supply voltage and threshold voltage scaling trends for Intel's microprocessor process technologies are discussed in [6]. It can be seen from Fig. 1 that the leakage power is only 0.01% of the active power for 1-μm technology, while it is 10% of the active power for 0.1-μm technology. There is a fivefold increase in leakage power as the technology process advances to a new generation. Projecting these trends, it can be seen that the leakage power dissipation will equal the active power dissipation within a few generations. Hence, efficient leakage power reduction methods are very critical for the deep-submicron and nanometer circuits.

In this paper, we describe a new leakage power reduction technique called LECTOR (LEakage Control TransistOR) for designing CMOS circuits. The rest of the paper is organized as follows. Section II describes briefly the prior works on leakage power reduction and their limitations. Section III introduces the transistor models used for estimating the leakage power. Our design strategy and an approach for minimizing the area overhead

Manuscript received December 11, 2002; revised August 20, 2003. This paper was recommended by Associate Editor Dr. R. Secareanu.

The authors are with the Nanomaterials and Nanomanufacturing Research Center, Department of Computer Science and Engineering, University of South Florida, Tampa, FL 33620 USA (e-mail: narender@csee.usf.edu; ranganat@csee.usf.edu).

Digital Object Identifier 10.1109/TVLSI.2003.821547

are described in Sections IV and V, respectively. Results are presented in Section VI, followed by conclusions in Section VII.

II. Related Work

Numerous methods for leakage power control have been reported in the literature. The work in [7] makes use of the dependence of the leakage current on the input vector to the gate. With additional control logic, the circuit is put into a low-leakage standby state when it is idle and restored to the original state when reactivated. Reactivation state forces the need to remember the original state information before going to low-leakage standby state. This requires special latches, thereby increasing the area of the circuit by about five times in the worst case [8]. Also, the amount of time for which the unit remains in idle state should be long enough so that the dynamic power consumed in forcing the circuit to low-leakage state and the leakage power dissipated in the standby state together is less than the leakage power without the technique.

Another technique for leakage power control is power gating, which turns off the devices by cutting off their supply voltage [9], [10]. This technique makes use of a bulky NMOS and/or PMOS device (sleep transistor) in the path between the supply voltage and ground. The sleep transistor is turned on when the circuit is active and turned off when the circuit is in idle state with the help of sleep signal. This creates virtual power and ground rails in the circuit. Hence, there is a significant detrimental effect on the switching speed when the circuit is active. The identification of the idle regions of the circuit and the generation of the sleep signal need additional hardware capable of predicting the circuit states accurately. This additional hardware consumes power throughout the circuit operation even when the circuit is in an idle state to continuously monitor the circuit state and control the sleep transistors.

The use of multiple threshold voltage CMOS (MTCMOS) technology for leakage control is described in [11] and [12]. The transistors of the gates are at low threshold voltage and the ground is connected to the gate through a high-threshold voltage NMOS gating transistor. The logical function of a gating transistor is similar to that of a sleep transistor. The existence of reverse conduction paths tend to reduce the noise margin or in the worst case may result in complete failure of the gate [4]. Moreover, there is a performance penalty since high-threshold transistors appear in series with all the switching current paths. A variation of MTCMOS technique is the Dual V_t technique, which uses transistors with two different threshold voltages. Low-threshold transistors are used for the gates on the critical path and high-threshold transistors are used for those not in the critical path [4], [13], [14]. In both MTCMOS and Dual V_t methods, additional mask layers for each value of threshold voltage are required for fabricating the transistors selectively according to their assigned threshold voltage values. This makes the fabrication process complex.

In addition to these limitations, the techniques discussed above suffer from turning-on latency, that is, when the idle subsections of the circuit are reactivated, they cannot be used immediately because some time is needed before the subcircuit returns to its normal operating condition. The latency for power gating is typically a few cycles, and for Dual V_t technology, is much higher [15]. Also, these techniques are not effective in controlling the leakage power when the circuit is in active state.

In [16], the authors use the concept of forced stacks for leakage control. Forced stacking introduces an additional transistor for every input of the gate in both N- and P-networks. This ensures that two transistors are OFF instead of one for every OFF-input of the gate and hence makes a significant savings on the leakage current. However, the loading requirements for each input introduced by the forced stacking reduces the drive current of the gate significantly. This results in a detrimental impact on the speed of the circuit.

In [10], a blend of sleep transistors and the stacking effects are used to reduce leakage power. This method identifies a circuit input vector for which the leakage current of the circuit is the lowest possible. The sleep signal controlled transistors are inserted away from the critical path where only one transistor is OFF when low-leakage input vector is applied to the circuit. Hence, this technique is input-vector dependent. Moreover, as this technique uses sleep transistors, it needs additional hardware for controlling them. This additional hardware consumes power in both idle and active states of the circuit.

In this work, we develop a new technique for leakage control in CMOS circuits. The proposed technique avoids the problems associated with the techniques discussed above.

III. Preliminaries

In this section, we briefly describe the models used in this work for estimating power dissipation for short-channel MOSFETs. The leakage current calculation is not straightforward due to the highly nonlinear behavior of the drain current of the device with respect to source/drain voltages. We have used the Berkeley Short-Channel IGFET (BSIM) Predictive Technology Model to estimate the leakage power dissipation, as it fits well with HSPICE simulations [5]. In the BSIM model, the threshold voltage V_t is expressed as

$$V_t = V_{FB} + \phi_s + k_1\sqrt{\phi_s} - k_2\phi_s - \eta V_{dd} \tag{1}$$

where V_{FB} is the flatband voltage, ϕ_s is twice the Fermi potential, k_1 and k_2 represent the nonuniform doping effect, and η models the drain-induced barrier lowering (DIBL) effect, an undesirable punch-through current flowing between the source and drain below the surface of the channel. The leakage current for NMOS transistors operating in weak-inversion region (i.e., $V_{gs} = 0$) is given by

$$I_s = I_0 \exp\left(\frac{(V_{gs} - V_t)}{nV_T}\right)\left(1 - \exp\left(-\frac{V_{ds}}{V_T}\right)\right) \tag{2}$$

where V_T is the thermal voltage and is given by q/kT, n is the subthreshold slope coefficient, and $I_0 = \mu_0 C_{\text{ox}}(W_{\text{eff}}/L_{\text{eff}})V_t^2 e^{1.8}$. Equation (2) gives a simple method for estimating the leakage current in a single NMOS transistor. A similar expression for the leakage current in a single PMOS transistor can be obtained.

In most CMOS logic design styles, the gates consist of series-parallel networks of PMOS and NMOS transistors. The