Leakage Control for Deep-Submicron Circuits

Kaushik Roy, Hamid Mahmoodi-Meimand, and Saibal Mukhopadhyay

School of Electrical and Computer Engineering, Purdue University,
West Lafayette, IN, USA
{kaushik, mahmoodi, sm}@ecn.purdue.edu

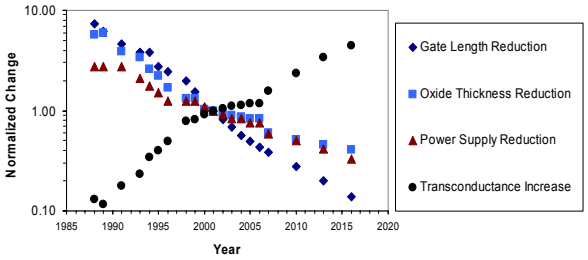
Abstract

High leakage current in deep sub-micron regimes is becoming a significant contributor to power dissipation of CMOS circuits as threshold voltage, channel length, and gate oxide thickness are reduced. Consequently, leakage control and reduction are very important, especially for low power applications. 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 profile 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)).

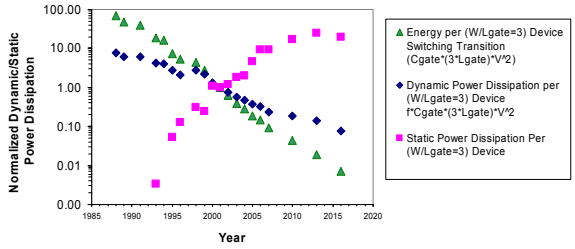
1. Introduction

To achieve higher density and performance and lower power consumption, CMOS devices have been scaled for more than 30 years. Transistor delay times have decreased by more than 30% per technology generation resulting in doubling of microprocessor performance every two years. Supply voltage (V_{DD}) has been scaled down in order to keep the power consumption under control. Hence, the transistor threshold voltage (V_{th}) has to be commensurately scaled to maintain a high drive current and achieve the performance improvement. However, the threshold voltage scaling results in the substantial increase of the subthreshold leakage current [1]. Fig. 1 shows projections for transistor physical dimensions, supply voltage, and device power consumption according to the International Technology Roadmap for Semiconductors (ITRS) [2]. All the parameters are normalized to their values in the year 2001. As shown in Fig. 1(b), due to the substantial increase in the leakage current, the static power consumption is expected to exceed the switching component of the power consumption unless effective measures are taken to reduce the leakage power.

For a CMOS circuit, the total power dissipation includes dynamic and static components during the active mode of operation. In the standby mode, the power dissipation is due to the standby leakage current. Dynamic power dissipation consists of two components. One is the switching power due to charging and discharging of load capacitance. The other is short circuit power due to the non-zero rise and fall time of input waveforms. The static power of a CMOS circuit is determined by the leakage current through each transistor. The dynamic (switching) power (P_D) and leakage power (P_{LEAK}) are expressed as:



(a)



(b)

Fig. 1: ITRS projections for transistor scaling trends and power consumption (a) physical dimensions and supply voltage (b) device power consumption [2].

$$P_D = \alpha f C V_{dd}^2 \quad (1)$$

$$P_{LEAK} = I_{LEAK} \cdot V_{dd} \quad (2)$$

where α is the switching activity; f is the operation frequency; C is the load capacitance; V_{dd} is the supply voltage; and I_{LEAK} is the cumulative leakage current due to all the components of the leakage current described in the previous section. Due to all the leakage mechanisms described in the previous section, leakage current (power) increases dramatically in the scaled devices. Particularly, with reduction of threshold voltage (to achieve high performance), leakage power becomes a dominant component of the total power

consumption in both active and standby modes of operation. Hence, in order to suppress the power consumption in low-voltage circuits, it is necessary to reduce the leakage power in both the active and standby modes of operation. 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 profile 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)).

This paper is organized as follows. In section 2, different leakage current components and mechanisms in deep sub-micron transistors are explored which is essential to guide solutions for reducing power and leakage per transistor. Device options for leakage reduction, which are based on channel engineering, are explained in the first part of section 3. The second part of section 3 explores different circuit techniques for leakage control in logic and memory. Finally, the conclusion of the paper appears in section 4.

2. Transistor Leakage Mechanisms

Six short channel leakage mechanisms are illustrated in Fig. 2. I_1 is the reverse bias *pn* junction leakage; I_2 is the subthreshold leakage; I_3 is the oxide tunneling 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, 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 during the transistor bias states in transition.

2.1. *pn* Junction Reverse Bias Current (I_1)

Drain and source to well junctions are typically reverse-biased causing *pn* junction leakage current. A reverse bias *pn* junction leakage (I_1) has two main components: One is minority carrier diffusion/drift near the edge of the depletion region and the other is due to electron-hole pair generation in the depletion region of the reverse biased junction. For an MOS transistor, additional leakage can occur between the drain and well junction from gated diode device action (overlap of the gate to the drain-well *pn* junctions) or carrier generation in drain to well depletion regions with influence of the gate on these current components. *pn* junction reverse bias leakage (I_{REV}) 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), Band-To-Band Tunneling (BTBT) dominates the *pn* junction leakage [3].

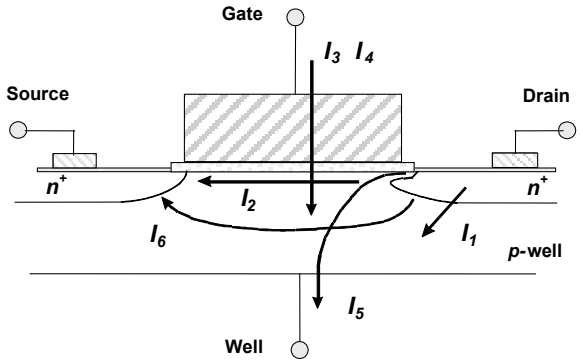


Fig. 2: Summary of leakage current mechanisms of deep submicron transistors.

2.2. Subthreshold Leakage (I_2)

Subthreshold or weak inversion conduction current between source and drain in an MOS transistor occurs when gate voltage is below V_{th} [4]. Weak inversion typically dominates modern device off-state leakage due to the low V_{th}. The weak inversion current can be expressed based on the following equation [4]:

$$I_{ds} = \mu_0 C_{ox} \frac{W}{L}(m-1)(v_T)^2 \times e^{\frac{(V_g - V_{th})}{m v_T}} \times \left(1 - e^{\frac{-v_{DS}}{v_T}}\right) \quad (3)$$

where

$$m = 1 + \frac{C_{dm}}{C_{ox}} = 1 + \frac{\frac{\varepsilon_{si}}{W_{dm}}}{\frac{\varepsilon_{ox}}{t_{ox}}} = 1 + \frac{3 t_{ox}}{W_{dm}} \quad (4)$$

V_{th} is the threshold voltage, and $v_T=KT/q$ is the thermal voltage. C_{ox} is the gate oxide capacitance; μ_0 is the zero bias mobility; and m is the subthreshold swing coefficient (also called body effect coefficient). W_{dm} is the maximum depletion layer width, and t_{ox} is the gate oxide thickness. C_{dm} is the capacitance of the depletion layer, and C_{ox} is the capacitance of the insulator layer.

In long channel devices, the subthreshold current is independent of the drain voltage for V_{DS} larger than few v_T. On the other hand, the dependence on the gate voltage and threshold voltage is exponential.

In long-channel devices, the source and drain are separated far enough that their depletion regions have no effect on the potential or field pattern in most part of the device. Hence, for such devices, the threshold voltage is virtually independent of the channel length and drain bias. In a short channel device, however, the source and drain depletion width in the vertical direction and the source-drain potential have a strong effect on the band bending over a significant portion of the device. Therefore, the threshold voltage and consequently the subthreshold current of short channel devices vary with the drain bias. This effect is referred to as Drain-Induced Barrier Lowering (DIBL). Fig. 3

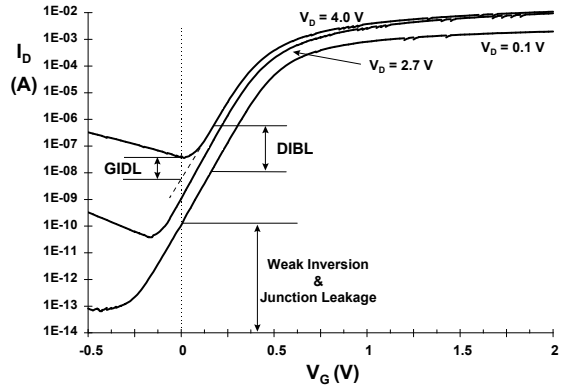


Fig. 3: n-channel I_D vs. V_G showing DIBL, GIDL, weak inversion, and pn junction reverse bias leakage components [7].

illustrates the DIBL effect as it moves the I_D-V_G curve up and to the left as the drain voltage increases.

2.3. Tunneling Into and Through Gate Oxide (I_3)

Reduction of gate oxide thickness results in increase in the field across the oxide. The high electric field coupled with low oxide thickness results in tunneling of electrons from substrate to gate and also from gate to substrate through the gate oxide, resulting in the gate oxide tunneling current.

The mechanism of tunneling between substrate and gate poly-silicon can be primarily divided into two parts, namely, (I) Fowler-Nordheim (FN) tunneling and (II) direct tunneling. In the case of FN tunneling, electrons tunnel through a triangular potential barrier, whereas in the case of direct tunneling, electrons tunnel through a trapezoidal potential barrier.

2.4. Injection of Hot Carriers from Substrate to Gate Oxide (I_4)

In a short channel transistor, due to high electric field near the Si/SiO_2 interface electrons or holes can gain sufficient energy from the electric field to cross the interface potential barrier and enter into the oxide layer. This effect is known as hot carrier injection. The injection from Si to SiO_2 is more likely for electrons than holes as electrons have a lower effective mass than that of holes and the barrier height for holes (4.5 eV) is more than that for electrons (3.1 eV) [6].

2.5. Gate Induced Drain Leakage (I_5)

Gate Induced Drain Leakage (GIDL) is due to high field effect in the drain junction of an MOS transistor. When the gate is biased to form an accumulation layer at the silicon surface, the silicon surface under the gate has almost same potential as the p-type substrate. Due to presence of accumulated holes at the surface, the surface behaves like a p-region more heavily doped than the substrate. This causes the depletion layer at the surface to be much narrower than elsewhere. The narrowing of depletion layer at or near the surface causes field crowding or an increase in the local electric field, and thereby enhancing the high field effects near that region. When the negative gate bias is large (i.e. gate at zero or negative and drain at V_{DD}), the n+ drain region under the gate can be depleted and even inverted. This causes more field crowding and the peak field increase, resulting in a dramatic increase of high field effects such as avalanche multiplication and band-to-band tunneling. As a result of all these effects, minority carriers are emitted in the drain region underneath the gate. Since the substrate is at a lower potential for minority carriers, the minority carriers that have been accumulated or formed at the drain depletion region underneath the gate are swept laterally to the substrate, completing a path for the GIDL [7]. Thinner oxide thickness and higher V_{DD} (higher potential between gate and drain) enhance the electric field and therefore increase GIDL.

2.6. Punchthrough (I_6)

In short channel devices, due to the proximity of the drain and the source, the depletion regions at the drain-substrate and source-substrate junctions extend into the channel. As the channel length is reduced, if the doping is kept constant, the separation between the depletion region boundaries decreases. An increase in the reverse bias across the junctions (with increase in V_{ds}) also pushes the junctions nearer to each other. When the combination of channel length and reverse bias leads to the merging of the depletion regions, punchthrough is said to have occurred. In sub-micron MOSFETs a V_{th}-adjust implant is used to have a higher doping at the surface than that in the bulk. This causes a greater expansion of the depletion region below the surface (due to smaller doping there) as compared to the surface. Thus the punchthrough occurs below the surface [8]. An increase in the drain voltage beyond the value required to establish the punchthrough, lowers the potential barrier for the majority carriers in the source. Thus, more of these carriers cross the energy barrier and enter into the substrate, and the drain collects some of them. The net effect is an increase in the subthreshold current. Furthermore, punchthrough degrades the subthreshold slope.

3. Leakage Reduction Techniques

In this section, we first consider major process techniques and then consider several circuit techniques for leakage control and reduction. Though most of the process and circuit techniques described here, are used to control the subthreshold leakage, some of them can be used to control other leakage components too.

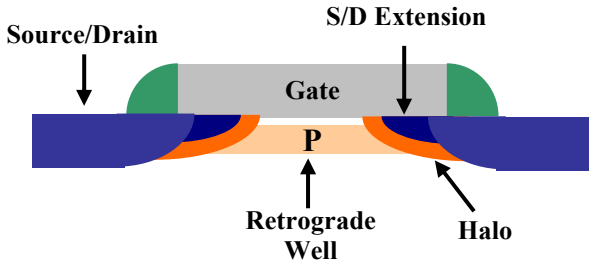


Fig. 4: Graphical representation of different aspects of well engineering [12].

Reducing all the components of leakage by both process and circuit level techniques is of major interest.

3.1. Channel Engineering for Leakage Reduction

In addition to gate oxide thickness and junction scaling, another technique to improve short channel characteristics is well engineering. 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 profile to minimize the off-state leakage while maximizing the linear and saturated drive currents. Super Steep Retrograde Wells (SSRW) 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. Fig. 4 is a schematic representation of the transistor regions that are affected by the different types of well engineering [9]. Retrograde well engineering changes the 1-dimentional characteristics of the well profile by creating a retrograde profile toward the Si/SiO_2 surface. The halo profile creates a localized 2-dimentional dopant distribution near the S/D extension regions. The use of these two techniques to increase the device performance, while keeping leakage to a tolerable limit, is discussed in the following sections.

3.1.1. Retrograde Doping

In order to maintain acceptable off-state leakage with continually decreasing channel lengths, both the oxide thickness and the gate-controlled depletion width in silicon must be reduced in proportion to the channel length to offset the degradation in short channel effects for extremely small devices. This requires an increase in the channel doping concentration. This leads to a higher threshold voltage for a uniformly doped channel. However, if the threshold voltage is not scaled, the device performance for low supply voltages will degrade due to the large reduction in gate drive. To reduce the gate-controlled depletion width while fulfilling the V_{th} reduction trend, retrograde doping can be used.

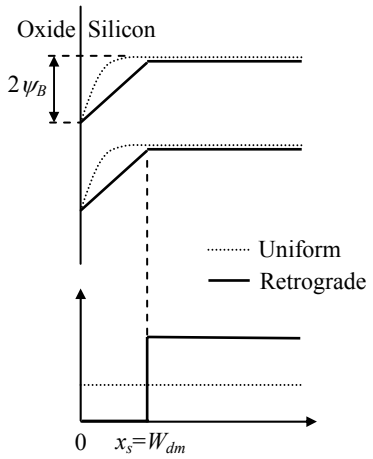


Fig. 5: Band digrams (shown on top) at the threshold condition for a uniformly doped and an extreme retrograde

Retrograde channel doping is a vertically nonuniform, low-high channel doping. It is used to improve the short channel effects 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.

Fig. 5 shows a schematic band-bending diagram at the threshold condition of an extreme retrograde profile with an undoped surface layer of thickness x_s. For the same gate depletion width (W_{dm}), 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.

3.1.2. Halo Doping

Halo doping or non-uniform channel profile in lateral direction was introduced below 0.25 μm 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 Fig. 4. 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 end of the channel. 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. As the channel length is reduced,

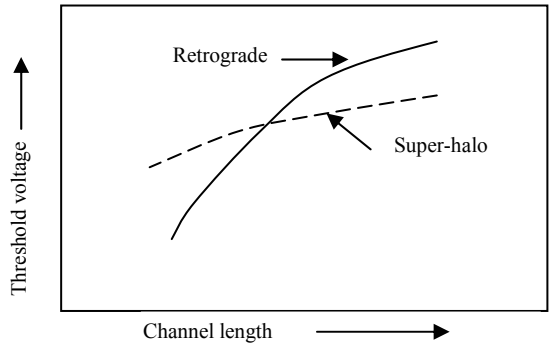


Fig. 6: Short channel threshold-voltage roll-off for retrograde and super-halo (vertical and lateral non-uniform dopings).

these highly doped regions consume a larger fraction of the total channel. 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 as shown in Fig. 6. Hence, the off-current becomes less sensitive to channel length variation. The reduction in drain and source junction depletion region width 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. The higher doping near the channel edges causes larger band-to-band tunneling, and higher GIDL. The band-to-band tunneling currents in the high-field region near the drain ultimately limit the halo doping level.

3.2. Circuit Techniques for Leakage Reduction

In this section, four major circuit design techniques namely, transistor stacking, multiple V_{th}, dynamic V_{th}, and supply voltage scaling (multiple and dynamic V_{DD}) for leakage reduction in digital circuits (logic and memory) are described.

3.2.1. Standby Leakage Control Using Transistor Stacks (Self Reverse Bias)

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". The stacking effect is best understood by considering a two input NAND gate as shown in Fig. 7. When both M_1 and M_2 are turned off, the voltage at the intermediate node (V_M) is positive due to small drain current. Positive potential at the intermediate node has three effects:

1) Due to the positive source potential V_M, gate to source voltage of M1 (V_{gs1}) becomes negative, and hence the subthreshold current reduces substantially.

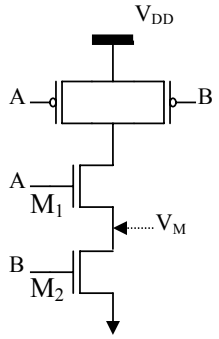


Fig. 7: Stacking effect in 2-input NAND gate.

2) Due to $V_M > 0$, body to source potential (V_{bs1}) of M_1 becomes negative resulting in an increase in the threshold voltage (larger body effect) of M_1, and thus reducing the subthreshold leakage.
3) Due to $V_M > 0$, the drain to source potential (V_{ds1}) of M_1 decreases resulting in an increase in the threshold voltage (less DIBL) of M_1, and thus reducing the subthreshold leakage.

The leakage of a two-transistor stack is an order of magnitude less than the leakage in a single transistor. An analysis of the subthreshold leakage through a stack of n- transistor is shown in [10].

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. The most straightforward way to find a low leakage input vector is to enumerate all combinations of primary inputs. For a circuit with n primary inputs, there are 2^n combinations for input states. Due to the exponential complexity with respect to the number of primary inputs, such an exhaustive method is limited to circuits with a small number of primary inputs. For large circuits, a random search based technique can be used to find the best input combinations. This method involves generating a large number of primary inputs, evaluating the leakage of each input, and keeping track of the best vector giving the minimal leakage current. A more efficient way is to employ the genetic algorithm to exploit historical information to speculate on new search points with expected improved performance to find a near optimal solution [10]. The reduction of standby leakage power by application of input vector is a very effective way of controlling the subthreshold leakage in the stand-by mode of operation of a circuit. In [11], a stack transistor insertion technique is given. For the gates with high subthreshold leakage in non-critical paths a leakage control transistor (low V_{th}) is inserted in series and is turned off during the stand by mode. The technique can effectively reduce the leakage current using single threshold voltage.

3.2.2. Multiple V_{th} Designs

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. Multiple threshold voltages can be achieved by the following methods:

- **Multiple Channel Doping**
  Multiple threshold voltages can be achieved by adjusting the channel doping densities. For this approach, two additional masks are required. This technique is commonly used to modify the threshold voltages. However, the threshold voltage can vary due to the non-uniform distribution of the doping density, making it difficult to achieve dual threshold voltages when the threshold voltages are very close to each other.
- **Multiple Oxide CMOS (M_{OX}CMOS)**
  Gate oxide thickness can be used to modify the threshold voltage of a transistor. Dual V_{th} can be achieved by depositing two different oxide thicknesses. For transistors in non-critical path, having a higher oxide thickness results in a high threshold voltage, and hence low subthreshold leakage. On the other hand, lower oxide thickness, and hence lower threshold voltage, in critical paths maintains the performance. Higher oxide thickness not only reduces the subthreshold leakage, it also reduces the followings:
  1) Gate oxide tunneling, since the oxide tunneling current exponentially decreases with an increase in the oxide thickness.
  2) Dynamic power consumption, since higher oxide thickness reduces the gate capacitance, which is beneficial for reduction of the dynamic power [12].

  For deep sub-micron devices, increasing the gate oxide thickness has an adverse effect of increasing short channel effect. In order to reduce the short channel effect, the aspect ratio of the device (AR) must be kept large enough. Aspect ratio of a device indicates the short channel immunity of the transistor – the larger the ratio is, the less the short channel effects are. Hence, increased oxide thickness of a transistor should be associated with channel length increase in order to prevent severe short channel effects. An advance process technology is required for fabricating M_{OX}CMOS. An algorithm for M_{OX}CMOS design is given in [12].
- **Multiple Channel Length**
  For short channel transistors, the threshold voltage decreases with the decrease in channel length (V_{th} roll-off). Hence, different threshold voltages can be achieved by using different channel lengths. Multiple channel length design uses the conventional CMOS technology. However, for the transistors with feature sizes close to 0.1μm, halo techniques have to be used to suppress the short channel effect. This causes the V_{th} roll-off to be very sharp, and hence, it is non-trivial to control the threshold voltage near the minimum feature size for such technologies. Longer channel lengths for high threshold transistors increase the gate capacitance, which has negative effect on performance and power.
- **Multiple Body Bias**
  For bulk silicon devices, the body voltage can be changed to modify the threshold voltage. If separate body biases are applied to different NMOS transistors, the transistors cannot share the same well, and therefore, triple well technologies are required. However, it is easier to change the body bias of partially-depleted Silicon On Insulator (SOI) devices, since the SOI devices are isolated naturally. For example, consider the double gate fully-depleted SOI, whose front gate and back gate surface potentials are strongly coupled to each other. The threshold voltage can be adjusted by biasing the back gate voltage.

Based on the above multiple threshold technologies, several multiple-threshold circuit design techniques have been developed recently, as explained in the following sections.

3.2.2.1. Multi-Threshold-Voltage CMOS (MTCMOS)

Multi-Threshold-Voltage CMOS (MTCMOS) reduces the leakage by inserting high threshold devices in series to low-V_{th} circuitry [13]. Fig. 8(a) shows the schematic of an MTCMOS circuit. A sleep control scheme is introduced for efficient power management. 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 (VDDV and VSSV) almost function as real power lines. In the standby mode, SL is set high, MN and MP are turned off, and the leakage current is low.

In fact, only one type of high V_{th} transistor is enough for leakage control. Fig. 8 (b) and (c) show the PMOS insertion and NMOS insertion schemes, respectively. The NMOS insertion scheme is preferable, since the NMOS on-resistance is smaller at the same width, and therefore, it can be sized smaller than corresponding PMOS. MTCMOS can be easily implemented based on existing circuits. A 1-Volt DSP chip for mobile phone applications has been developed recently [14]. However, MTCMOS can only reduce the standby leakage power, and the large inserted MOSFETs can increase the area and delay. Moreover, if data retention

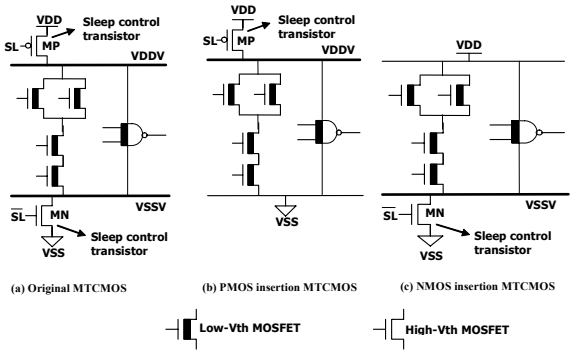


Fig. 8: Schematic of MTCMOS circuit [21].

is required in the standby mode, an extra high-V_{th} memory circuit is needed to maintain the data [15]. Instead of using high V_{th} sleep control transistors as MTCMOS, super cut-off CMOS (SCCMOS) technique uses low-V_{th} transistors with an inserted gate bias generator [16]. For the PMOS (NMOS) insertion, the gate is applied to 0V (VDD) in the active mode, and the virtual VDD (VSS) line is connected to supply VDD (VSS). In the standby mode, the gate is applied to VDD+ΔV (VSS-ΔV) to fully cut off the leakage current. Compared to MTCMOS, SCCMOS circuits can work at lower supply voltages.

3.2.2.2. Dual Threshold CMOS

For a logic circuit, a higher threshold voltage can be assigned to some transistors in non-critical paths so as to reduce the leakage current, while the performance is maintained due to the use of low threshold transistors in the critical path(s) [17]. Therefore, no additional leakage control transistors are required, and both high performance and low power can be achieved simultaneously. Fig. 9 illustrates the basic idea of a dual-V_{th} circuit. Fig. 10 shows the path distribution of dual V_{th} and single V_{th} CMOS for a 32-bit adder. Dual V_{th} CMOS has the same critical delay as the single low V_{th} CMOS circuit, but the transistors in non-critical paths can be assigned high V_{th} to reduce leakage power. Dual threshold technique is good for leakage power reduction during both standby and active modes without delay and area overhead.

3.2.2.3. Variable Threshold CMOS (VTMOS)

Variable threshold CMOS (VTMOS) is a body-biasing design technique [18]. Fig. 11 shows the VTMOS scheme. In order to achieve different threshold voltages, a self-substrate bias circuit is used to control the body bias. In the active mode, a nearly zero body bias is applied. While in the standby mode, a deeper reverse body bias is applied to increase the threshold voltage and cut off the leakage current. This scheme has been used in a two dimensional discrete cosine

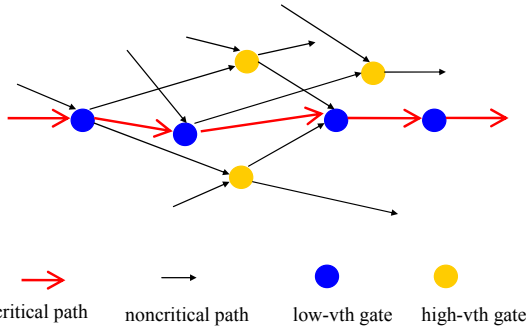


Fig. 9: Dual Vth CMOS circuit.

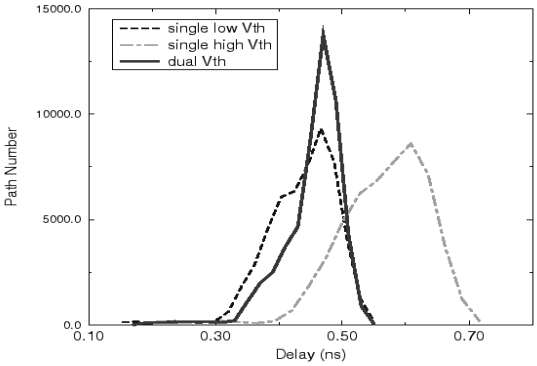


Fig. 10: Path distribution of dual V_{th} and single V_{th} CMOS.

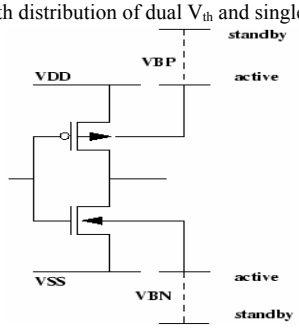


Fig. 11: Variable Threshold CMOS (VTCMOS).

transform core processor [18]. Furthermore, in the active mode, a slightly forward substrate bias can be used to increase the circuit speed while reducing short channel effects [19]. Providing the body potential requires routing the body grid that adds to the overall chip area. Keshavarzi *et. al.* reported that reverse body biasing lowers IC leakage by three orders of magnitude in a 0.35 µm technology [20]. However, more recent data showed that the effectiveness of reverse body bias to lower I_{OFF} decreases as technology scales [20].

3.2.2.4. Dynamic Threshold CMOS (DTMOS)

For dynamic threshold CMOS, the threshold voltage is altered dynamically to suit the operating state of the circuit. A high threshold voltage in the standby mode gives low leakage current, while a low threshold

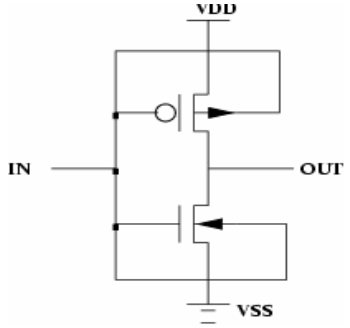


Fig. 12: Schematic of DTMOS inverter.

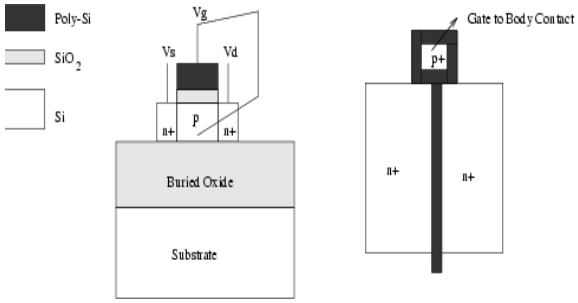


Fig. 13: SOI DTMOS structure and layout.

voltage allows for higher current drives in the active mode of operation. Dynamic threshold CMOS can be achieved by tying the gate and body together (DTMOS) [21]. Fig. 12 shows the schematic of a DTMOS inverter. DTMOS can be developed in bulk technologies by using triple wells. "Doping engineering" is needed to reduce the parasitic components [22]. Stronger advantage of DTMOS can be seen in partially depleted Silicon-on-Insulator (SOI) devices. Fig. 13 shows the SOI DTMOS structure and layout. In [22], excellent DC inverter characteristics down to 0.2V and good ring oscillator performance down to 0.3V are achieved using this method. 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.

3.2.2.5. Double Gate Dynamic Threshold SOI

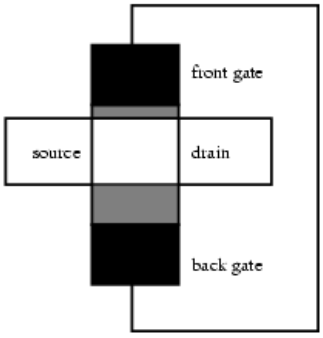


Fig. 14: DGDT SOI MOSFET structure.

CMOS (DGDT-MOS)

Double gate dynamic threshold voltage (DGDT) SOI MOSFET [23] combines the advantages of DTMOS and double gate fully depleted (FD) SOI MOSFETs without any limitation on the supply voltage. Fig. 14 shows the structure of a DGDT SOI MOSFET. DGDT SOI MOSFET is an asymmetrical double gate SOI MOSFET. Back gate oxide is thick enough to make the threshold voltage of the back gate larger than the supply voltage. Since the front gate and back gate surface potentials are strongly coupled to each other, the front gate threshold voltage changes dynamically with the back gate voltage. Results show that DGDT SOI MOSFETs have nearly ideal symmetric subthreshold characteristics. Compared to symmetric double gate SOI CMOS, the power delay product of DGDT SOI CMOS is smaller.

3.2.3. Dynamic V_{th} Designs

Dynamic threshold voltage scaling is a technique for active leakage power reduction. This scheme utilizes dynamic adjustment of frequency through back-gate bias control depending on the workload of a system. When the workload decreases, less power is consumed by increasing V_{th}. Two varieties of dynamic V_{th} scaling have been proposed as described below.

3.2.3.1. V_{th}-Hopping Scheme

Fig. 15 shows the schematic diagram of the V_{th}-hopping scheme [24]. Using the control signal (CONT), which is obtained from software, the power control block generates select signals, $V_{th\text{-}low}$-Enable and $V_{th\text{-}high}$-Enable, which in turn control the substrate bias for the circuit. When the controller asserts $V_{th\text{-}low}$-Enable, V_{th} in the target processor reduces to $V_{th\text{-}low}$. On the other hand, when the controller asserts $V_{th\text{-}high}$-Enable, the target processor V_{th} becomes $V_{th\text{-}high}$. CONT is controlled by software through a software feedback loop scheme. CONT also controls the operation frequency of the target processor. When the controller asserts $V_{th\text{-}low}$-Enable, the frequency controller generates f_{CLK}, and when the controller asserts $V_{th\text{-}high}$-

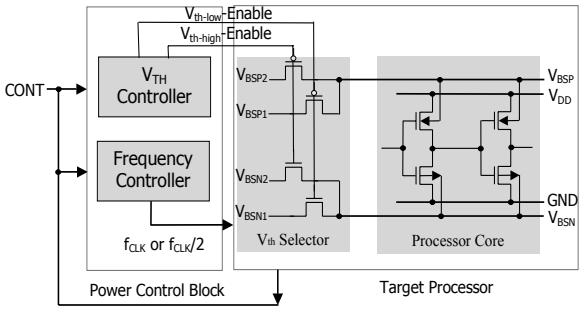


Fig. 15: Schematic diagram of V_{th}-hopping [34].

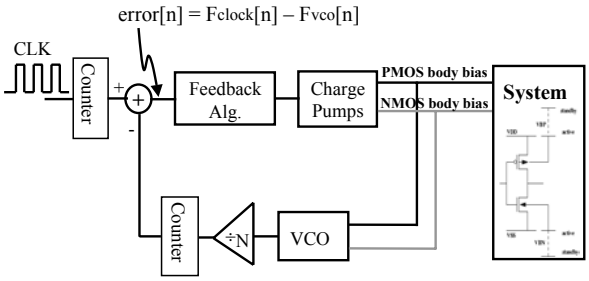


Fig. 16: Schematic of DVTS hardware [35].

Enable, the frequency controller generates $f_{CLK}/2$ (say).

3.2.3.2. Dynamic V_{th} Scaling (DVTS) Scheme

A block diagram of the DVTS scheme and its feedback loop is presented in Fig. 16 [25]. A clock speed scheduler, which is embedded in the operating system, determines the (reference) clock frequency at run-time. The DVTS controller adjusts the PMOS and NMOS body bias so that the oscillator frequency of the VCO tracks the given reference clock frequency. The error signal, which is the difference between the reference clock frequency and the oscillator frequency, is fed into the feedback controller. The continuous feedback loop also compensates for variation in temperature and supply voltage.

3.2.4. Supply Voltage Scaling

Supply voltage scaling was originally developed for switching power reduction. It is an effective method for switching power reduction because of the quadratic dependence of the switching power on the supply voltage. Supply voltage scaling also helps reduce leakage power, since the subthreshold leakage due to DIBL decreases as the supply voltage is scaled down [26]. For a 1.2V, 0.13μm technology, it is shown that the supply voltage scaling has significant impacts on subthreshold leakage and gate leakage (reductions in the orders of V^3 and V^4 , respectively) [27].

To achieve low-power benefits without compromising performance, two ways of lowering supply voltage can be employed: static and dynamic supply scaling. In static supply scaling, multiple supply voltages are used as shown in Fig. 17. Critical and noncritical paths or units of the design are clustered and powered by higher and lower supply voltages, respectively [28]. Since the speed requirements of the noncritical units are lower than the critical ones, supply voltage of noncritical units can be lowered without degrading system performance. Whenever an output from a low V_{DD} unit has to drive an input of a high V_{DD} unit, a level conversion is needed at the interface. The secondary voltages may be generated off-chip or regulated on-die from the core supply. Dynamic supply scaling overrides the cost of

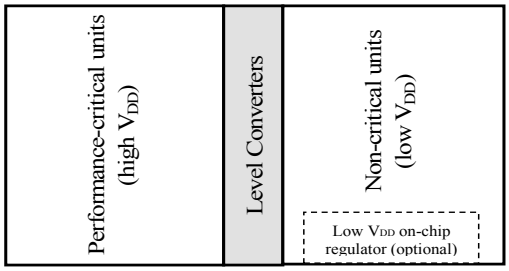


Fig. 17: Two-level multiple supply voltage scheme [41].

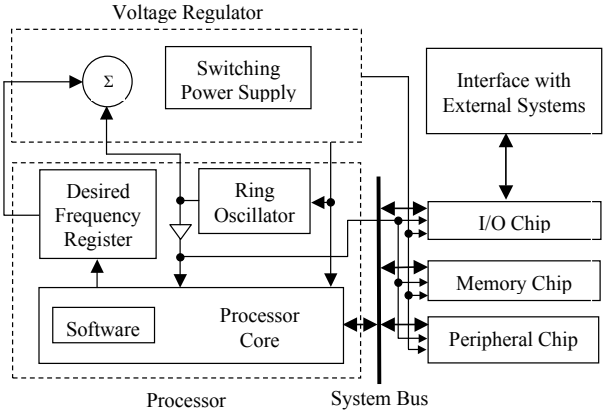


Fig. 18: Dynamic Voltage Scaling (DVS) architecture [42].

using two supply voltages, by adapting the single supply voltage to performance demand. The highest supply voltage delivers the highest performance at the fastest designed frequency of operation. When performance demand is low, supply voltage and clock frequency is lowered, delivering reduced performance but with substantial power reduction. There are three key components for implementing Dynamic Voltage Scaling (DVS) in a general-purpose microprocessor: an operating system that can intelligently determine the processor speed, a regulation loop that can generate the minimum voltage required for the desired speed, and a microprocessor that can operate over a wide voltage range. Fig. 18 shows a DVS system architecture [30]. Control of the processor speed must be under software control, as the hardware alone may not distinguish whether the currently executing instruction is part of a compute-intensive task or a non-speed-critical task. Supply voltage is controlled by hard-wired frequency-voltage feedback loop, using a ring oscillator as a replica of critical path. All chips operate at the same clock frequency and same supply voltage, which are generated from the ring oscillator and the regulator.

3.2.5. Leakage Reduction Methods for Cache Memory

State-of-the-art microprocessor designs devote a large fraction of the chip area to memory structures, e.g. multiple levels of instruction and data caches,

translation look-aside buffers, and prediction tables. For instance, 30% of Alpha 21264 and 60% of Strong ARM processors are devoted to cache and memory structures [31]. Caches account for a large (if not dominant) component of leakage energy dissipation in recent designs, and will continue to do so in the future. Recent energy estimates for 0.13µ process technology indicate that leakage energy accounts for 30% of L1 cache energy and as much as 80% of L2 cache energy. To address the above-mentioned problem, several techniques have been proposed in the literature, as explained in the following sections.

3.2.5.1. Data Retention Gated-Ground Cache (DRG-Cache)

Data Retention Gated-Ground Cache (DRG-Cache) puts the unused portions of the memory core to low leakage mode to reduce power. The key idea is to introduce an extra NMOS transistor (Fig. 19) in the leakage path from the supply voltage to the ground of the Static Random Access Memory (SRAM) cells; the extra transistor is turned 'on' in the used and turned 'off' in the unused sections, essentially "gating" the supply voltage of the cells. Fig. 19 shows the anatomy of the DRG-Cache. Gated-ground achieves significantly lower leakage because of the two off transistors connected in series, reducing the leakage current by orders of magnitude; this effect is due to the self reverse-biasing of the stacked transistors, which is called the stacking effect, as described earlier.

Similar to conventional gating techniques, the gated-ground transistor can be shared among multiple SRAM cells from one or more cache blocks. This amortizes the overhead of the extra transistor. Because the size of the gated-ground transistor plays a major role in the data retention capability and stability of the DRG-Cache, and also affects the power and performance savings, the gated-ground transistor must be carefully sized (Fig. 19) with respect to the SRAM cell transistors. While the gated-ground transistor must be made large enough to sink the current flowing through the SRAM cells during a read/write operation in the active mode and to enhance the data retention capability of cache in the standby mode, a large gated-ground transistor may reduce the stacking effect, and thereby diminishing the energy savings. Moreover, large transistors also increase the area overhead due to gating. In DRG-Cache the gated-ground transistor is shared by a row of SRAM cells. The gated-ground transistor is controlled by the row decoder logic of the conventional SRAM. The cells are turned 'on' only when the row is being read from or when data is written into the row. However, this requires the row decoder to drive a larger gate capacitance associated with the gated-ground transistor unlike conventional caches. To maintain the performance proper sizing of the decoder is required.

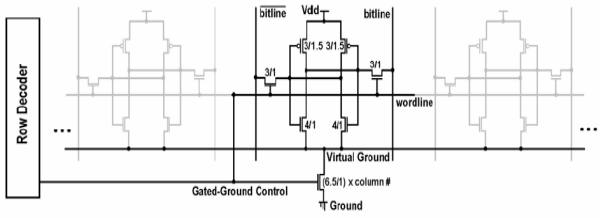


Fig. 19: Anatomy of Data Retention Gated-Ground Cache (DRG-Cache) [44].

Conventional SRAM stores the data as long as power supply is 'on'. This is because the cell storage nodes, which are at '0' and '1', are firmly strapped to the power rails through conducting devices (by a pulldown NFET in one inverter and a pull-up PFET in the other inverter). When the gated-ground transistor is ON, the DRG cache behaves exactly like a conventional SRAM in terms of data storage. Turning 'off' the gated-ground cuts-off the leakage path to the ground. However, it also cuts-off the opportunity to firmly strap nodes, which are at '0', to the ground. This makes it easier for a noise source to write a '1' to that node. Turning 'on' the gated-ground transistor restores the '0' data. Simulation results show that data is not lost even if the gated-ground transistor is turned off for indefinite time [32].

3.2.5.2. Drowsy-Cache

Significant leakage reduction can also be achieved by putting the cache into a low power drowsy mode [33]. In the drowsy mode, the information in the cache line is preserved. However the line has to be reinstated to a high power mode before its contents can be accessed. One technique for implementing a drowsy cache is to switch between two different supply voltages in each cache line [33]. Due to short channel effect in deep sub-micron devices, subthreshold leakage current reduces significantly with voltage scaling. The combined effect of reduced leakage and supply voltage gives large reduction in the leakage power.

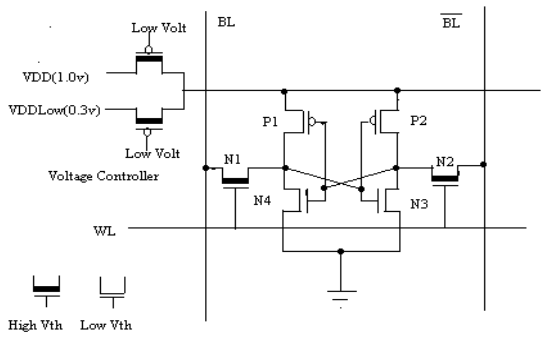


Fig. 20: Schematic of drowsy memory circuit [45].

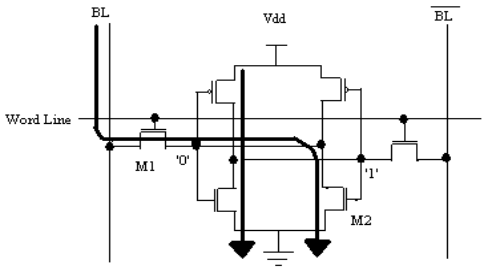


Fig. 21: The two dominant leakage paths (V_{dd} to ground and bitline to ground) for a 6-transistor SRAM cell. Leakage through these two paths consist a high percentage of the total leakage [47].

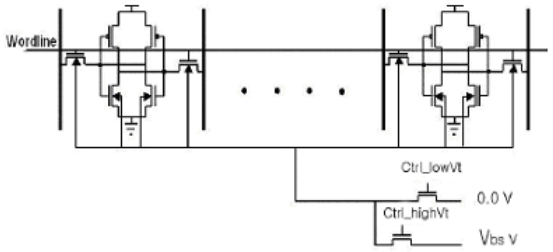


Fig. 22: Schematic of a dynamic V_{th} SRAM set [46].

Fig. 20 illustrates the circuit schematic of a SRAM cell connected to the voltage controller. One PMOS pass gate switch supplies the normal supply voltage (VDD) (in the active mode), and the other supplies the low supply voltage (VDDLow) (in the stand-by mode) for the drowsy cache line. Each pass gate is a high-V_{th} device to prevent leakage current from the normal supply to the low supply through the two PMOS pass gate transistors. A separate voltage controller is needed for each cache line. By scaling the voltage of the cells to approximately 1.5 times of V_{th}, the state of the memory cell can be maintained. For a typical 70nm process, the drowsy voltage is about 0.3V [33]. Since the capacitance of the power rail is very low, the transition time between the high and low power state is low. High-V_{th} devices are used as the pass transistors that connect the internal inverters of the memory cell to the read/write lines (N1 and N2). This reduces the leakage through the pass transistors, since the read/write lines are maintained in high power mode.

3.2.5.3. Dynamic Threshold Voltage (V_{th}) SRAM

Dynamic V_{th} SRAM (DTSRAM) architecture can be used to reduce leakage energy dissipation in memory structures. Using body biasing, the subthreshold leakage can be reduced without sacrificing data stability [34]. In a time-based dynamic V_{th} scheme, high V_{th} is assigned to the cache lines which are not accessed for a certain time period (30 μs ~ 100 μs) and a low V_{th} is assigned to the cache lines which are in frequent use to maintain high performance [35]. Fig. 21 depicts the two dominant leakage paths for a conventional 6-transistor SRAM cell, the V_{dd} to ground and the bitline to ground leakage paths. These two leakage paths make up a high percentage of the total leakage [35]. Fig. 22 shows the schematic of a DTSRAM cache line. The NMOS substrate can be switched to 0V for high performance. When the cache line is not in use, the substrate can be switched to a negative voltage (V_{bs}) to reduce the leakage. Since the transition energy required for a single substrate bias transition is much more than the leakage energy saved during one clock cycle, V_{th} transition cannot be made every clock cycle [35]. Moreover, the performance loss due to negative body bias (i.e. high V_{th}) is considerable. To overcome these difficulties, properties of temporal and spatial locality of cache access can be used. In [35], a time based scheme is described, which instead of turning a cache line to high V_{th} state right after its access, leaves the cache line in low V_{th} for a certain time period (30 μs ~ 100 μs). This ensures that the upcoming accesses within this time period will not impose any energy or delay penalties. Moreover, using the spatial locality of program reference, instead of only turning 'on' the accessed cache line, a portion of the cache containing the accessed cache line is turned 'on'. Consequently, subsequent accesses occur in the turned 'on' portion of the cache. A capacitor-discharging scheme is described in [35] to implement the body-bias control circuit.

4. Conclusion

With the continuous scaling of CMOS devices, leakage current is becoming a major contributor to the total power consumption. In current deep submicron devices with low threshold voltages, subthreshold and gate leakage have become dominant sources of leakage and are expected to increase with the technology scaling. To manage the increasing leakage in deep-submicron CMOS circuits, solutions for leakage reduction have to be sought both at the process technology and circuit levels. At the process technology level, well engineering techniques by retrograde and halo doping are used to reduce leakage and improve short channel characteristics. At the circuit level, transistor stacking, multiple V_{th}, dynamic V_{th}, multiple V_{dd}, and dynamic V_{dd} techniques can effectively reduce the leakage current in high performance logic and memory designs.

Acknowledgment:

This work was supported in part by Semiconductor Research Corporation, DARPA, Intel, and IBM.

References

[1] V. De and S. Borkar, "Technology and Design Challenges for Low Power and High Performance," in Proceedings of

International Symposium on Low Power Electronics and Design, pp. 163–168, August 1999.
[2] 2001 International Technology Roadmap for Semiconductors, http://public.itrs.net/
[3] Y. Taur and T. H. Ning, Fundamentals of Modern VLSI Devices, New York, USA: Cambridge University Press, 1998, ch. 2, pp. 94-95.
[4] Y. Taur and T. H. Ning, Fundamentals of Modern VLSI Devices, New York, USA: Cambridge University Press, 1998, ch. 3, pp. 120-128.
[5] A. Keshavarzi, K. Roy, and C. F. Hawkins, "Intrinsic Leakage In Low Power Deep Submicron CMOS Ics," in Proceedings of International Test Conference, pp. 146–155, 1997.
[6] Y. Taur and T. H. Ning, Fundamentals of Modern VLSI Devices, New York, USA: Cambridge University Press, 1998, ch.2, pp. 97-99.
[7] K. Roy and S. C. Prasad, Low-Power CMOS VLSI Circuit Design, New York, USA: Wiley Interscience Publications, 2000, ch. 2, pp. 28-29.
[8] K. Roy and S. C. Prasad, Low-Power CMOS VLSI Circuit Design, New York, USA: Wiley Interscience Publications, 2000, ch. 2, pp. 27-28.
[9] S. Thompson, *et. al.*, "MOS Scaling: Transistor Challenges for the 21st Century," Intel Technology Journal Q3'98.
[10] Z. Chen, *et. al.*, "Estimation of Standby Leakage Power in CMOS Circuits Considering Accurate Modeling of Transistor Stacks," in Proceedings of International Symposium on Low Power Electronics and Design, pp. 239-244, 1998.
[11] M. C. Johnson, D. Somasekhar, and K. Roy, "Leakage Control with Efficient Use of Transistor Stacks in Single Threshold CMOS", Design Automation Conf., pp.442-445, 1999.
[12] N. Sirisantana, L. Wei, and K. Roy, "High-Performance Low-Power CMOS Circuits Using Multiple Channel Length And Multiple Oxide Thickness," in Proceedings of International Conference on Computer Design, pp. 227 –232, 2000.
[13] S. Mutoh, *et. al.*, "1-V Power Supply High-Speed Digital Circuit Technology with Multi-threshold Voltage CMOS," IEEE Journal of Solid-State Circuits, vol.30, pp. 847-854, August 1995.
[14] S. Mutoh, S.Shigematsu, Y. Matsuya, H. Fukuda, and J. Yamada, "A 1-V Multi-threshold Voltage CMOS DSP with an Efficient Power Management for Mobile Phone Application," IEEE Inter. Solid-State Circuits Conference, pp. 168-169, 1996.
[15] S. Shigematsu, *et. al.*, "A 1-V High Speed MTCMOS Circuit Scheme For Power-Down Applications," IEEE Journal of Solid-State Circuits, vol.32, pp. 861-869, June 1997.
[16] H. Kawaguchi, K. Nose, and T. Sakurai, "A CMOS Scheme for 0.5V Supply Voltage with Pico-Ampere Stanby Current," IEEE International Solid-State Circuits Conference, pp. 192-193, 1998.
[17] L. Wei, *et. al.*, "Design and Optimization of Dual Threshold Circuits for Low Voltage Low Power Applications", IEEE Transactions on VLSI Systems, pp. 16-24, March 1999.
[18] T. Kuroda, *et. al.*, "A 0.9V 150MHz 10mW 4mm 2-D Discrete Cosine Transform Core Processor with Variable-Threshold-Voltage Scheme," in Digest of Technical Papers of IEEE International Solid-State Circuits Conference, pp. 166-167, 1996.
[19] Y. Oowaki, et. al., "A Sub-0.1um Circuit Design with Substrate-Over-Biasing," in Digest of Technical Papers of IEEE International Solid-State Circuits Conference, pp. 88-89, 1998.
[20] A. Keshavarzi, *et. al.*, "Effectiveness of Reverse Body Bias for Low Power CMOS Circuits," in Proceedings of 8th NASA Symposium on VLSI Design, pp. 2.3.1-2.3.9, 1999.
[21] F. Assaderaghi, *et. al.*, "A Dynamic Threshold Voltage MOSFET(DTMOS) for Ultra-Low Voltage Operation", IEEE International Electron Devices Meeting, pp. 809-812, 1994.
[22] C. Wann, *et. al.*, "Channel Profile Optimization And Device Design For Low-Power High-Performance Dynamic-Threshold MOSFET," in Digest of Technical Papers of IEEE International Electron Devices Meeting, pp. 113-116, 1996.
[23] L. Wei, Z. Chen, and K. Roy, "Double Gate Dynamic Threshold Voltage (DGDT) SOI MOSFETs for Low Power High Performance Designs," in Proceedings of IEEE International SOI Conference, pp. 82-83, 1997.
[24] K. Nose, *et. al.*, "VTH-Hopping Scheme to Reduce Subthreshold Leakage for Low-Power Processors," IEEE Journal of Solid-State Circuits, vol. 37, pp. 413-419, March 2002.
[25] C.H. Kim and K. Roy, "Dynamic V_{TH} Scaling Scheme For Active Leakage Power Reduction," Conference on Design, Automation and Test in Europe, pp. 163 –167, 2002.
[26] A.J. Bhavnagarwala, *et. al.*, "A minimum total power methodology for projecting limits on CMOS GSI," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 8, pp. 235-251, June 2000.
[27] S.Tyagi, et. al, "A 130 nm Generation Logic Technology Featuring 70 nm Transistors, Dual Vt Transistors and 6 Layers of Cu Interconnects," in Digest of Technical Papers of International Electron Devices Meeting, pp. 567 -570, 2000.
[28] M. Takahashi, et. al., "A 60-mw MPEG4 Video Codec Using Clustered Voltage Scaling With Variable Supply-Voltage Scheme," IEEE Journal of Solid-State Circuits, vol. 33, pp. 1772 –1780, November 1998.
[29] R. K. Krishnamurthy, A. Alvandpour, V. De, and S. Borkar, "High-performance and Low-power Challenges for Sub-70nm Microprocessor Circuits," in Proceedings of IEEE Custom Integrated Circuits Conf, pp. 125-128, 2002.
[30] S. Lee and T. Sakurai, "Run-Time Voltage Hopping For Low-Power Real-Time Systems," in Proceedings of IEEE/ACM Design Automation Conference, pp. 806–809, 2000.
[31] S. Manne, *et. al.*, "Pipeline Gating: Speculation Control for Energy Reduction" in Proceeding of the International Symposium on Computer Architecture, pp. 32–141, 1998.
[32] A. Agarwal, H. Li, and K. Roy, "DRG-Cache: A Data Retention Gated-Ground Cache for Low Power," in Proceedings of Design Automation Conference, pp.473-478, 2002.
[33] K. Flautner, *et. al.*, "Drowsy Caches: Simple Techniques for Reducing Leakage Power," International Symposium on Computer Architecture, pp. 148 -157, 2002.
[34] H. Mizuno, et. al., "An 18 μA Standby Current 1.8-V, 200-MHz Microprocessor with Self-Substrate-Biased Data-Retention Mode", IEEE Journal of Solid-State Circuits, vol. 34, pp. 1492-1500, November 1999.
[35] C. H. Kim and K. Roy, "Dynamic Vt SRAM: A Leakage Tolerant Cache Memory for Low Voltage Microprocessors," International Symposium on Low Power Electronic Design, August 2002.