Leakage Current in Sub-Micrometer CMOS Gates

Paulo Francisco Butzen and Renato Perez Ribas

Universidade Federal do Rio Grande do Sul

{pbutzen, rpribas}@inf.ufrgs.br

Abstract. Static power consumption is nowadays a crucial design parameter in digital circuits due to emergent mobile products. Leakage currents, the main responsible for static power dissipation during idle mode, are increasing dramatically in sub-100nm processes. Subthershold leakage rises due to threshold voltage scaling while gate leakage current increases due to scaling of oxide thickness. It means the static power dissipation should be considered as soon as possible in the design flow. Leakage mechanisms and reduction techniques will be reviewed, providing a minimum background about this issue.

INTRODUCTION

In the past, the major concerns of the VLSI designers were performance and miniaturization. With the substantial growth in portable computing and wireless communication in the last few years, power dissipation has become a critical issue. Problems with heat removal and cooling are worsening because the magnitude of power dissipated per unit area is growing with scaling. Years ago, portable battery-powered applications were characterized by low computational requirement. Nowadays, these applications require the computational performance similar to as non-portable ones. It is important to extend the battery life as much as possible. For these reasons power dissipation becomes a challenge for circuit designers and a critical factor in the future of microelectronics.

An integrated circuit is composed by sequential and combinational circuits, memories blocks and I/O devices. Each one gives its own contribution to the total power dissipation of the system. Fig. 1 shows the approximate power distribution in microprocessors [1-3]. Power consumption is concentrated in the logic circuits, 40 % in sequential blocks and 30 % in combinational parts. Memory blocks and I/O device represent approximately 30% of the total power.

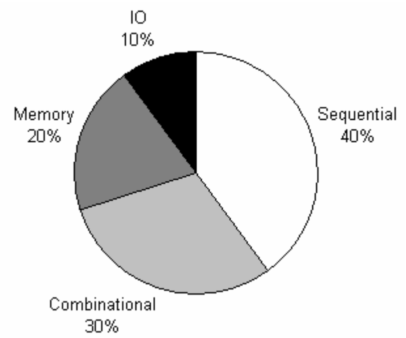

Fig. 1. Power distribution in microprocessors [1-3].

There are four components of power dissipation in digital CMOS circuits, as describe in equation (1).

$$P = P_{\text{dynamic}-\text{switching}} + P_{short-circuit} + P_{\text{static}-\text{biasing}} + P_{leakage} \qquad \textbf{(1)}$$

where P is the total power dissipation, $P_{\text{dynamic–switching}}$ is the dynamic switching power, $P_{\text{short–circuit}}$ is the short-circuit power, $P_{\text{static–biasing}}$ is the static biasing power and P_{leakage} is the leakage power.

Dynamic switching power dissipation is caused by charging capacitances in the circuit. During each low-to-high output transition, the load capacitance C_L, in Fig. 2, is charged through the PMOS transistor, and a certain amount of energy is drawn from the power supply. Part of this energy is dissipated in PMOS device and part is stored on C_L. It is discharged during the high-to-low output transition, and the stored energy is dissipated through the NMOS transistor.

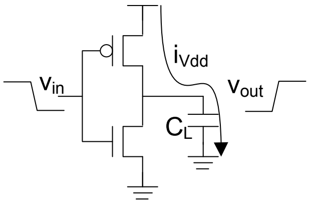

Fig. 2. Dynamic switching power dissipation scheme in CMOS inverter

Considering the CMOS inverter, shown in Fig. 2, and assuming that the input waveform has zero rise and fall times, the energy consumption during low-to-high output transition can be derived by integrating the instantaneous power over the period of interest. Equation (2) shows that it draws $C_L.V_{dd}^2$ Joules from the power supply.

$$E_{V_{dd}} = \int_0^{\infty} i_{V_{dd}}(t) V_{dd}\, dt = V_{dd} \int_0^{\infty} C_L \frac{dv_{out}}{dt} dt = C_L V_{dd} \int_0^{V_{dd}} dv_{out} = C_L V_{dd}^2 \quad \textbf{(2)}$$

The charge stored on the load capacitor is equals to $C_L.V_{dd}^2/2$ by equation (3). This means that only half of the energy supplied by the power source is stored in C_L. The other half had been dissipated by the PMOS transistor. The high-to-low output transition dissipates the energy stored on the load capacitance into the NMOS transistor.

$$E_{C_L} = \int_0^{\infty} i_{V_{dd}}(t) v_{out}\, dt = \int_0^{\infty} C_L \frac{dv_{out}}{dt} v_{out}\, dt = C_L \int_0^{V_{dd}} v_{out}\, dv_{out} = \frac{C_L V_{dd}^2}{2} \quad \textbf{(3)}$$

To compute the power consumption, it is necessary to take into account how often the circuit is switched. Given a gate switching frequency *f*, the power drawn from the supply is given by:

$$P_{\text{dynamic-switching}} = C_L V_{dd}^2 f \quad \textbf{(4)}$$

The dynamic switching power dissipation was the dominant factor compared with other components of power dissipation in digital CMOS circuits for technologies up to 0.18µm, where it is about 90% of total circuit dissipation [4].

Short–circuit power is the second source of total power dissipation described in equation (1). During a transient on the input signal, there will be a period in which both NMOS and PMOS transistor will conduct simultaneously, causing a current flow through the direct path existing between power supply and ground terminals. This short circuit current usually happens for very small intervals. In a static CMOS inverter this current flows as long as the input voltage is higher than a NMOS threshold voltage (V_{thn}) above ground and lower than a PMOS threshold voltage (V_{thp}) below the power supply, as shown in Fig. 3. It is proportional to the input ramp, the output load, and the transistors size. It can be approximated by [5], according to equation (5)

$$P_{short-circuit} = K\left(V_{dd} - 2V_{th}\right)^3 .\tau .f \quad \textbf{(5)}$$

where K is a constant that depends on the transistors size, and on the technology parameters, V_{dd} is the supply voltage, V_{th} is the threshold voltage, τ is the rise or fall time of the input signal, and f is the clock frequency.

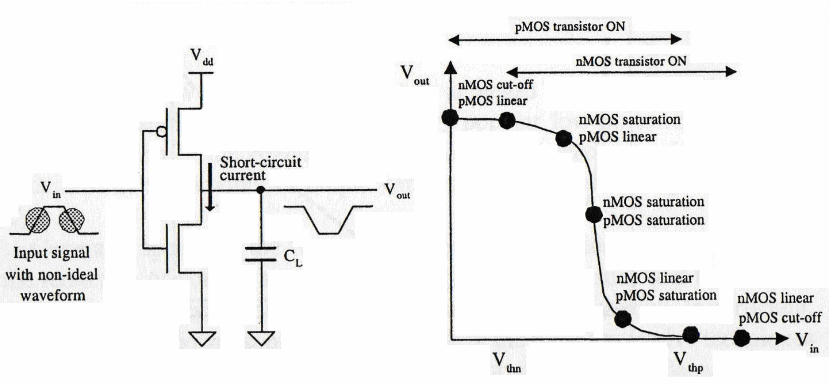

Fig. 3. CMOS inverter short-circuit current [6]

This component represents less than 20% of the dynamic switching power consumption if the NMOS and PMOS transistors are sized in order to balance the rise/fall signal slopes at input and output nodes [5].

Both of the above sources of power dissipation in CMOS circuits are related to transitions at gate terminals and for that reason are usually referred as dynamic power dissipation. On the other hand, the other two components of power dissipation, static biasing and leakage, are related to the current that flows when the gate terminals are not changing, and are therefore commonly referred as static power dissipation.

Ideally, in steady state, CMOS circuits do not present static power dissipation. That is the most attractive characteristic of CMOS technology. However, real systems present degraded voltage levels feeding CMOS gates and a current flow from the power supply to ground nodes is observed. This flow is known as static biasing current.

In Fig. 4, a NMOS pass-transistor drives an inverter. From basic CMOS circuit theory is known that the voltage in node A is degraded (V_{dd}-V_{th}). Since the inverter input is high (V_{dd}-V_{th}), its output should be low. However, the PMOS transistor is weakly ON and, consequently, presents a static biasing current from power supply to ground nodes.

The static biasing current only happen in specific conditions as reported above. Static current that flows from V_{dd} to ground nodes, without degraded inputs is known as leakage power. In past technologies the magnitude of leakage current was low and usually neglected. However, the devices have been scaling for decades to achieve higher density, performance. As a consequence, leakage current in the nanometer regime is becoming a significant portion of power dissipation in CMOS circuits, as depicted in Fig. 5

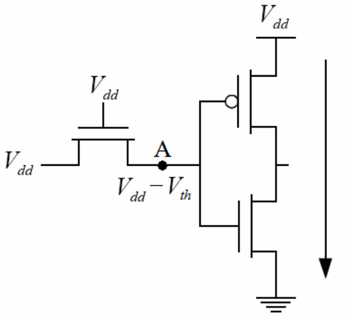

Fig. 4. Degraded voltage level at the input node of an CMOS inverter results in static biasing power consumption.

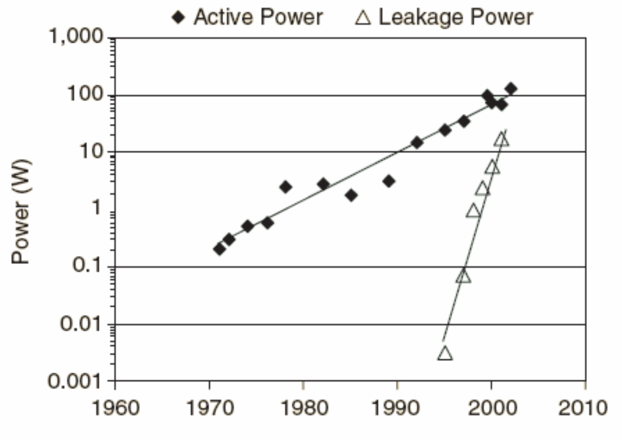

Fig. 5. Active and leakage power dissipation [7].

With the technology scaling, supply voltage needs to be reduce due to dynamic power and reliability issues. However, it requires the scaling of the device threshold voltage (V_{th}) to maintain a reasonable gate over drive [8]. The V_{th} reduction, result in an exponential increase in the subthreshold current. Moreover, to control the short channel effects (SCEs) and to maintain the transistor drive strength at low supply voltage, oxide thickness needs to be also scaled down. The aggressive scaling of oxide thickness results in a high tunneling current through the transistor gate insulator [8]. Furthermore, scaled devices require the use of the higher substrate doping density. It causes significantly leakage current through these drain- and source-to-substrate junctions under high reversed biasing [8].

These are the three major types of leakage mechanisms: subthreshold, gate oxide and reverse-bias pn-junction leakage (BTBT – band-to-band tunneling). In addition to these three major leakage components, there are other ones like gate-induced drain leakage (GIDL) and punchthrough current. Those components can be neglected in normal modes of operation [9].

VLSI circuit designers have to respect a power specification. Accurate and efficient power estimation during the design phase is required in order to meet the power specification without a costly redesign process. Precise simulators, such as HSPICE^{TM}, can accurately account for leakage current, but they are only proper for small circuits due to convergence, CPU time and memory issues. The physical models to treat the leakage mechanisms [9-10] are too complex to be used by circuit designers. Faster techniques to estimate the subthreshold and gate leakage current have been proposed in the literature [11-12]. It is important to estimate both average and maximum power in CMOS circuits at different levels of design abstraction. The average power dissipation is important to determine the battery life, while the maximum power demanded is related to circuit reliability issues.

The power consumption reduction in digital systems involves optimization at different design levels. This optimization includes the technology used to implement digital circuits, the logic style, the circuit architecture, and the algorithm that are being implemented.

Optimization in technology level are related to materials used in the fabrication process, like high-K gate dielectric and metal gates [14], its dimension and concentration, like oxide thickness and substrate profile, and device structure, like “halo” doping [15] and silicon-on-insulator (SOI) structures [16]. Design level involves optimization in physical and logic design. Placement, routing and sizing strategy are example of physical design optimization. Logic minimization and technology mapping are example of logic design optimization techniques [17]. Architectural level typically presents solutions based on parallel or pipelined structures to achieve the same performance with a reduced supply voltage [18]. Algorithm level explores the concurrency to be implemented in a parallel architecture and the minimization of the number of operations to reduce the switching activity, and consequently the dynamic consumption [17].

Leakage mechanisms, estimation and reduction techniques will be reviewed in the following sections, providing useful background to IC designers about leakage currents.

LEAKAGE CURRENT MECHANISMS

For nanometer devices, leakage current is dominated by subthreshold leakage, gate-oxide tunneling leakage and reverse-bias pn-junction leakage. Those three major leakage current mechanisms are illustrated in Fig. 6. There are still other leakage components, like gate induced drain leakage (GIDL) and punchthrough current, however those ones can be still neglected in normal operation of digital circuits [9].

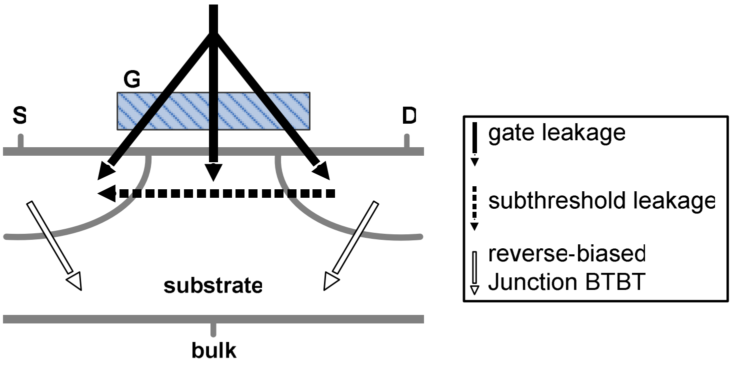

Fig. 6. Major leakage mechanisms in MOS transistor;

Subthreshold Current

Supply voltage has been scaled down to keep dynamic power consumption under control. To maintain a high drive current capability, the threshold voltage (V_{th}) has to be scaled too. However, the V_{th} scaling results in increasing subthreshold leakage currents. Subthreshold current occurs between drain and source when transistor is operating in weak inversion region, i.e., the gate voltage is lower than the V_{th}.

The drain-to-source current is composed by drift current and diffusion current. The drift current is the dominant mechanism in strong inversion regime, when the gate-to-source voltage exceeds the V_{th}. In weak inversion, the minority carrier concentration is almost zero, and the channel has no horizontal electric field, but a small longitudinal electric field appears due the drain-to-source voltage. In this situation, the carries move by diffusion between the source and the drain of MOS transistor. Therefore, the subthreshold current is dominated by diffusion current and it depends exponentially on both gate-to-source and threshold voltage.

Considering the BSIM MOS transistor model [10], the subthreshold leakage current for a MOSFET device can be expressed as:

$$I_{subthresho\ ld} = I_0 e^{\frac{V_{gs}-V_{th}}{nV_T}} \left[1 - e^{-\frac{V_{ds}}{V_T}} \right] \quad \textbf{(6)}$$

where $I_0 = \frac{W\mu_0 C_{ox} V_T^2 e^{1.8}}{L}$, $V_T = \frac{KT}{q}$ is the thermal voltage, V_{th} is the threshold voltage, V_{ds} and V_{gs} are the drain-to-source and gate-to-source voltages respectively. W and L are the effective transistor width and length, respectively. C_{ox} is the gate oxide capacitance, μ_0 is the carrier mobility and n is the subthreshold swing coefficient.

In short channel devices, source and drain depletion regions advances significantly into the channel influencing the field and potential profile inside that. These are known as short channel effects (SCE). Such effects reduce transistor threshold voltage due to the channel length reduction (V_{th} roll-off) and the DIBL increasing. This results in significant subthreshold current in short channel devices.

Gate Oxide Tunneling Current

As mentioned before, the aggressive device scaling in nanometer regime increases short channel effects such as DIBL and V_{th} roll-off. To control the short channel effects, oxide thickness must also become thinner in each technology generation. Aggressive scaling of the oxide thickness, in turn, gives rise to high electric field, resulting in a high direct-tunneling current through transistor gate insulator.

The tunneling of electrons (or holes) from the bulk and source/drain overlap region through the gate oxide potential barrier into the gate (or vice-versa) is referred as gate oxide tunneling current. This phenomenon is related with the MOS capacitance concept. There are three major gate leakage mechanisms in a MOS structure. The first one is the electron conduction-band tunneling (ECB), where electrons tunneling from conduction band of the substrate to the conduction band of the gate (or vice versa). The second one is the electron valence-band tunneling (EVB). In this case, electrons tunneling from the valence band of the substrate to the conduct band of the gate. The last one is known as hole valence-band (HVB) tunneling, where holes tunneling from the valence band of the substrate to the valence band of the gate (or vice- versa). Fig. 7 illustrates these three mechanisms.

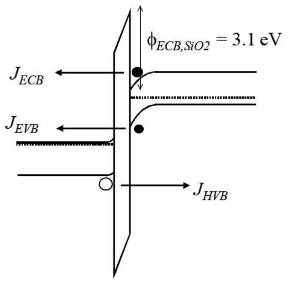

Fig. 7. Three mechanisms of gate dielectric direct tunneling leakage [11]

Each mechanism is dominant or important in different regions of operation for NMOS and PMOS transistors. For each mechanism, gate leakage current can be modeled by [8]:

$$I_{gate} = W.L.A\left(\frac{V_{ox}}{t_{ox}}\right)^2 \exp\left(\frac{-B\left(1-\left(1-\frac{V_{ox}}{\phi_{ox}}\right)^{3/2}\right)}{\frac{V_{ox}}{t_{ox}}}\right) \qquad (7)$$

where W and L are the effective transistor width and length, respectively, $A = q^3/16\pi^2 h\phi_{ox}$, $B = 4\pi\sqrt{2m_{ox}}\phi_{ox}^{3/2}/3hq$, m_{ox} is the effective mass of the tunneling particle, ϕ_{ox} is the tunneling barrier height, t_{ox} is the oxide thickness, h is $1/2\pi$ times Planck's constant and q is the electron charge.

Band-to-Band Tunneling Current

The MOS transistor has two pn junctions – drain and source to well junctions. These junctions are typically reverse biased, causing a pn junction leakage current. This current is a function of junction area and doping concentration. When ‘n’ and ‘p’ regions are heavily doped, band-to-band tunneling (BTBT) leakage dominates the reverse biased pn junction leakage mechanism.

A high electric field across a reverse biased pn junction causes a current flow through the junction due to tunneling of electrons from the valence band of the p-region to the conduction band of the n-region, as shown in Fig. 8.

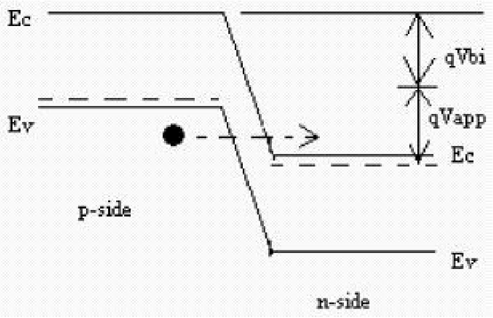

Fig. 8. BTBT in reverse-biased pn junction [8]

Tunneling current occurs when the total voltage drop across the junction, applied reverse bias (V_{app}) plus built-in voltage (ψ_{bi}), is larger than the band-gap. The tunneling current density through a silicon pn junction is given by [8]:

$$J_{BTBT} = A\frac{EV_{app}}{E_g^{1/2}}\exp\left(-B\frac{E_g^{3/2}}{E}\right) \quad \mathbf{(8)}$$

where $A = \sqrt{2m^* q^3}/4\pi^3 h^2$, and $B = 4\sqrt{2m^*}/3hq$. m^* is the effective mass of 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 electron charge; and h is $1/2\pi$ times the Planck’s constant.

Band-to-band tunneling leakage, negligible in current processes when compared to the subthreshold and gate oxide leakages, starts to be taken into account in 25nm technologies [19].

The junction tunneling current depends exponentially on the junction doping and the reverse bias across the junction. Forward body bias can be used to reduce the band-to-band tunneling leakage.

LEAKAGE ESTIMATION METHODS

The total leakage power in CMOS circuits is determined by the contribution of leakage currents in each transistor, which has two main sources: subthreshold leakage current and gate tunneling leakage current. Band-to-Band-Tunneling leakage is still very small in existing technologies and can be ignored [19].

Subthreshold Leakage

With technology scaling, the supply voltage needs to be scaled down to reduce the dynamic power and maintain reliability. However, this requires the scaling of the device threshold voltage (V_{th}) to maintain a reasonable gate over drive [8]. The V_{th} scaling, result in an exponential increase in the subthreshold current, according equation (6).

Subthreshold leakage current occurs only in turned-off transistors. For an individual device, leakage current can be calculated by equation (6). However, for a whole circuit, the leakage current is not simply the sum of leakage current of all devices. The circuit topology is determinant to the overall leakage. In particular, series connected devices, or stacked devices, have lower leakage than the sum of the leakage for each device taken in isolation. In a stack with two transistors, a slight reverse bias between the gate and source occurs in the pull-up transistor when both transistors are turned off. Because subthreshold current is exponentially dependent on gate-to-source voltage (V_{gs}), according to equation (6), a substantial current reduction is obtained. This phenomenon is referred as “stacking effect". Fig. 9 depicts the subthreshold leakage current for different stacks of off-transistors.

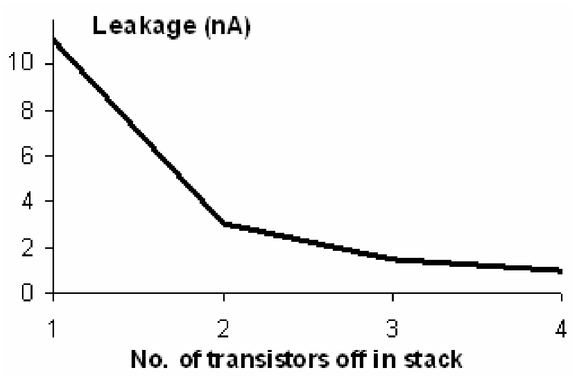

Fig. 9. Subthreshold leakage current in different stacks of off-transistors.

Subthreshold leakage estimation methods [17, 19, 20] compute the subthreshold leakage based on the following steps. The basic difference in each method is the equation used to model the subthreshold leakage.

I. Identify the off-plane, plane that does not have a conducting path from power rail and the output;

II. On-transistors in these off-planes are considered as ideal short-circuits;

III. Single devices have their subthreshold leakage current computed;

IV. Branches with stacks of turned-off transistors present the "stack effect" and are evaluated using the Kirchhoff Current Law (KCL). The follow analysis exemplifies the subthreshold leakage evaluation for a two off-transistors stack, as show in Fig. 10.

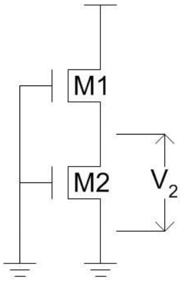

Fig. 10. Two off-transistor stack.

Considering the subthreshold leakage current model reported by [18]:

$$I_S = I_0 W e^{\frac{V_{gs} - (V_{t0} - \eta V_{ds} - \gamma V_{bs})}{nV_T}} \left[1 - e^{\frac{-V_{ds}}{V_T}} \right] \qquad \textbf{(9)}$$

where $I_0 = \frac{\mu_0 C_{ox} V_T^2 e^{1.8}}{L}$, $V_T = \frac{kT}{q}$, V_{t0} is the zero-bias threshold voltage, W is the effective transistor width, L is the effective channel length, n is the subthreshold slope coefficient, C_{ox} is the gate oxide capacitance, μ_0 is the mobility, η is the drain-induced barrier lowering coefficient and γ is the linearized body effect coefficient.

In a two-transistor stack the subthreshold leakage currents passing through the transistors is given by

$$I_{SM1} = I_0 W_1 e^{\frac{-V_2 - [V_{t0} - \eta(V_{dd} - V_2) + \gamma V_2]}{nV_T}} \qquad \textbf{(10)}$$

$$I_{SM2} = I_0 W_2 e^{\frac{-V_{t0} + \eta V_2}{nV_T}} \qquad \textbf{(11)}$$

and the intermediate node voltage V2 is expressed as

$$V_2 = \frac{\eta V_{dd} + nV_T \ln\left(\frac{W_1}{W_2} \right)}{1 + 2\eta + \gamma} \qquad \textbf{(12)}$$

The voltage V_2, calculated by equation (12), is used to determine the subthreshold leakage current in this branch using equantion (10) or (11) since $I_S = I_{SM1} = I_{SM2}$.

V. Subthreshold current in each branch and single device are then summed to provide the total subthreshold leakage;

Gate Oxide Leakage

The reduction of vertical dimensions has been harder than horizontal ones. An aggressive scaling of gate oxide thickness is required to provide large current drive capability at reduced voltages supplies and to suppress short-channel effects, such as drain induced-barrier lowering. This scaling increases the field across the oxide. The high electric field coupled with the low oxide thickness results in gate tunneling leakage current from the gate to the channel and source/drain overlap region, or from the source/drain overlap region to the gate. These mechanisms are depicted in Fig. 11 (a) and (b), respectively.

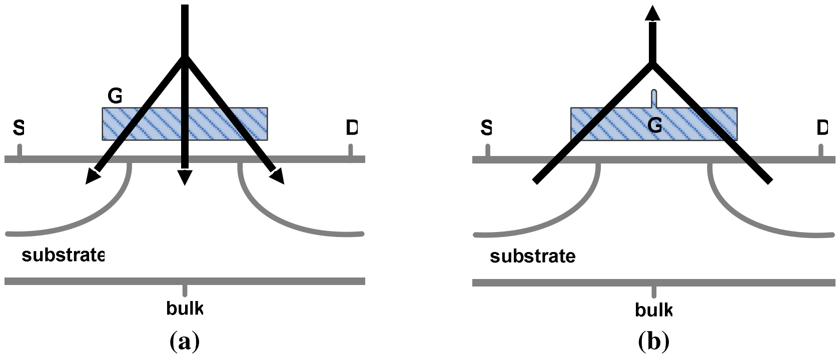

Fig. 11. Gate oxide leakage current (a) from gate to channel and source/drain overlap region and (b) from source/drain overlap region to gate

Gate leakage current increases exponentially with decreasing oxide thickness. When the gate oxide thickness reaches 3nm and below, gate tunneling current comes into the order of the subthreshold leakage [13]. It also increases exponentially with voltage across gate oxide. Fig. 12 shows the density of gate leakage current (A/m^2) in a NMOS device versus potential drop across the oxide for several oxide thicknesses.

Ignoring the variability in oxide thickness due to process variation, it is possible to consider only the voltage dependence in a gate leakage analysis.

Subthreshold leakage is evaluated only when transistor is turned off. Gate oxide leakage, on the other hand, occurs in both cases, when transistors are turned on and off. Gate leakage current is independently in both, turned on or off, transistor states. When transistor is turned off the current flows by the overlap source and drain regions. In the case where the transistor is turned on, the current uses the overlap source/drain regions and the transistor channel. For these reasons, gate oxide leakage is usually higher in such condition.

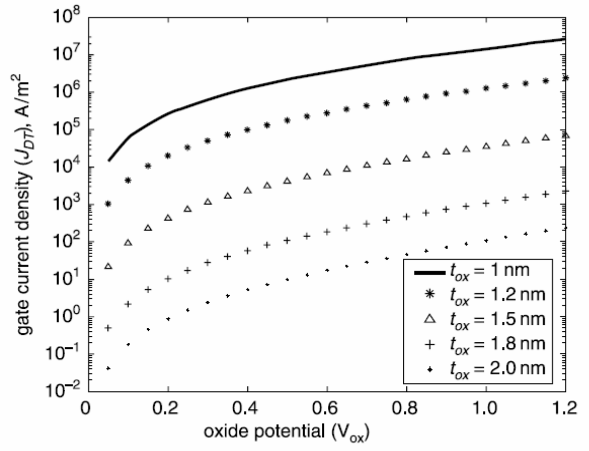

Fig. 12. Variation of tunneling current density with potential drop across the oxide [12].

Considering previous statement, the easy method to investigate gate leakage current is evaluating the transistor bias conditions. Fig. 13 presents all eight possible bias conditions for a NMOS transistor. Fig. 13 (f) and (g) can be ignored because they represent transient states and does not occur in steady state. In Fig. 13 (a) and (h) gate leakage is not present because all terminals have the same potential. In the other conditions gate leakage has to be computed.

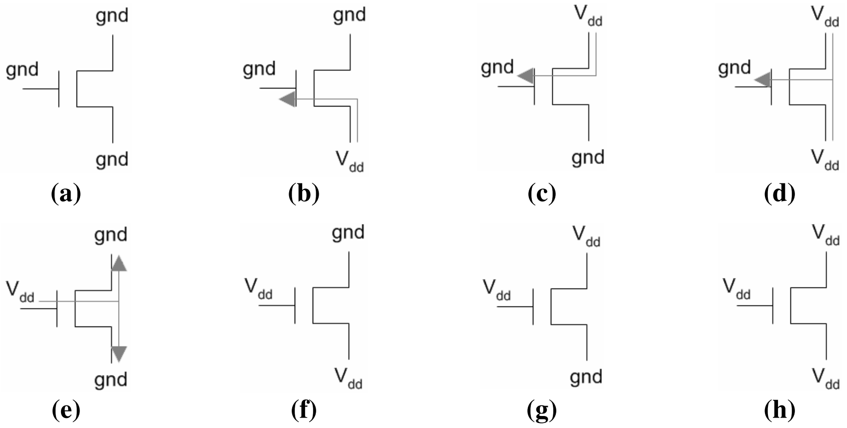

Fig. 13. Possible bias condition of NMOS transistors in CMOS logic circuits.

Estimation methods [12, 22, 23] evaluate gate leakage current based on transistor bias conditions, depicted in Fig. 13. The gate oxide leakage current is calculated using equation (7) after each device has its bias condition defined based on input vector. The total gate oxide leakage is the sum of the current in each device.

In [23], it is assumed that the internal nodes attain full logic levels (i.e., they are either at V_{DD} or gnd) and in a transistor stack the entire voltage drops across off devices. It does not treat the subthreshold leakage and the interactions between both mechanisms. These assumptions make the techniques really fast, but the accuracy for complex gates is compromised.

Both subthreshold and gate leakage mechanisms are explored in [13]. Subthreshold leakage current is calculated using the model proposed by [20] and the gate leakage is calculated using the same concept proposed in [23]. Both leakage currents are added to provide the total leakage. This approach is valid, however it does not considered the interaction between both leakage mechanisms since they are evaluated separately.

A complete, accurate, but some what complex analytical leakage estimation method is presented in [24]. The methodology proposed in such work uses numerical solvers to evaluate the interactions between both mechanisms, compromising the method performance.

LEAKAGE REDUCTION TECHNIQUES

In CMOS circuit, the total power dissipation includes dynamic and static components during the active mode of operation. In case of standby mode, the power dissipation is related to leakage currents. According to leakage mechanisms, described in previous section, leakage power increases dramatically in the scaled devices, becomes a significant component of total power consumption in both modes of operation.

To suppress power consumption in low-voltage circuits, it is necessary to reduce leakage power in both active and standby modes. Reduction in leakage current can be achieved by using both process and circuit level techniques. At process level, leakage reduction can be achieved by controlling the dimensions (length, oxide thickness, junction depth, etc.) and doping profile in transistor. At circuit level, several techniques to reduce leakage consumption have been proposed in the literature [4, 7, 24].

To reduce leakage currents, these techniques explore supply and threshold voltage leakage dependence, as well as the concepts of stacking effect and body biasing. Some of such are briefly presented below.

Dual Threshold CMOS

Dual threshold CMOS is a static technique that exploit the delay slack in non-critical paths to reduce leakage power. It provides both high and low threshold voltage (V_{th}) transistors in a single chip that are used to deal with the leakage problem.

Fabrication process can achieve a different V_{th} device by varying different parameters. Changing the channel doping profile, increasing the channel length, changing the body bias, and using a higher gate oxide thickness are examples of fabrications parameters that can be changed to achieve high V_{th} transistor. Each parameter has its own trade-off in terms of process cost, effect on different leakage components, and short channel effects.

High V_{th} transistors suppress the subthreshold current, while low V_{th} transistors are used to achieve high performance. For a logic circuit, the transistors in non-critical paths can be assigned high V_{th} to reduce subthreshold leakage current, while the performance is not sacrificed by using low V_{th} transistors in the critical paths [26]. It has the same critical delay as the single low V_{th} CMOS circuits, while leakage power is saved in non-critical paths. Therefore, no additional control circuitry is required and both high performance and low leakage power can be achieved simultaneously. Fig. 14 illustrates the basic idea of a dual V_{th} circuit.

With the increase in V_{th} variation and supply voltage scaling, it is becoming difficult to maintain sufficient gap among low V_{th}, high V_{th} and supply voltage required for dual V_{th} design. Furthermore, dual V_{th} design increases the number of critical paths in a die, decreasing both mean and standard deviation of the die frequency distribution, resulting in reduced performance [27].

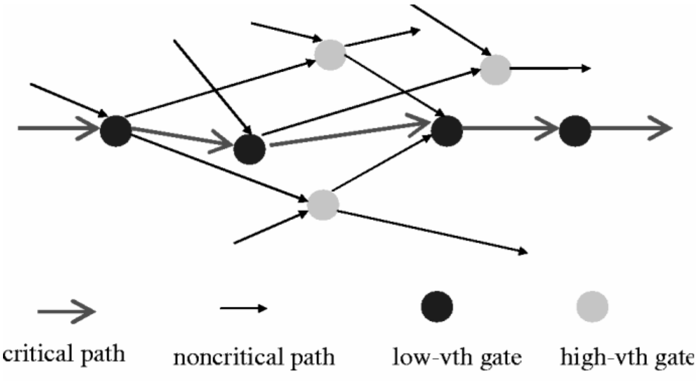

Fig. 14. Dual V_{th} CMOS circuit [8].

Supply Voltage Scaling

Supply voltage scaling is used to reduce dynamic and leakage power. It was originally developed for switching power reduction. It is an effective method of consumption reduction due to the quadratic dependence of the switching power in relation to supply voltage. Supply voltage scaling also provides leakage power savings.

Lowering supply voltage provides an exponential reduction in subthreshold current resulting from Drain-Induced Barrier Lowering (DIBL) effect. The DIBL effect tends to become more severe with process scaling to shorter gate lengths. For this reason, the achievable savings by this technique will increase with technology scaling.

Gate oxide leakage is also affected by this technique. Lowering V_{dd} will reduces gate leakage even faster than subthreshold leakage [28]. Fig. 15 shows how gate tunneling current reduces as V_{dd} decrease. Thus, this technique saves standby power by decreasing subthreshold and gate leakage currents.

In theory, the standby power supply for a circuit can decrease to zero, but the circuit loses performance and its logic states. The optimal point for power savings using this technique is the lowest voltage which the circuit retains its logic states and does not compromise performance [29].

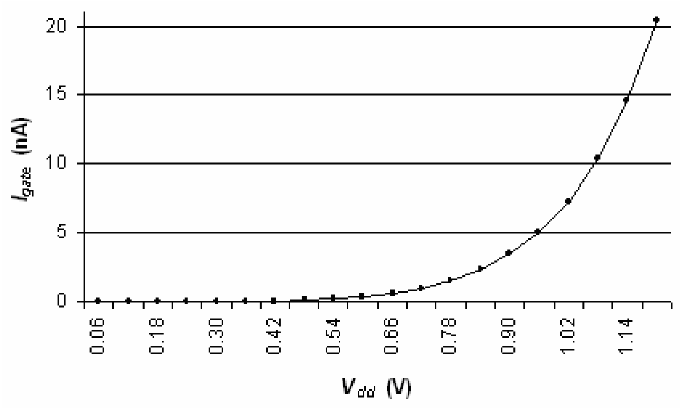

Fig. 15. Gate oxide leakage current versus power supply.

To achieve low-power benefits without compromising performance, two ways of lowering supply voltage can be employed: the static supply scaling and the dynamic supply scaling.

Static Supply Scaling

In static supply scaling, multiple supply voltages are used as shown in Fig. 16. Critical and non-critical paths and/or units of the design are clustered and powered by higher and lower voltages, respectively [30]. In an extreme case the combinational logic in a circuit can fall to zero when the circuit is in idle mode because it does not need to hold its logic state, increasing the power saving. 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. Furthermore, the secondary voltages could be generated off-chip [31] or regulated on-die from the core supply [32].

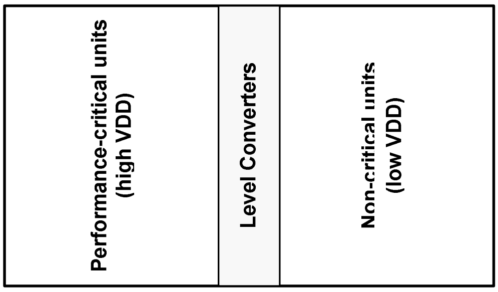

Fig. 16. Two-level multiple static supply voltage scheme.

Dynamic Supply Scaling

Dynamic supply scaling overrides the cost of using multiple supply voltages by adapting the single supply voltage to the performance demand. When performance demand is low, supply voltage and clock frequency are lowered, delivering reduced performance with substantial power reduction [33].

As mentioned before, this technique gets rid of the cost of using multiple supply voltages. However, the follow overheads are added when this technique is implemented:

- Circuit has to operate over a wide voltage range;
- Operating system to intelligently determine the processor speed;
- Regulator to generate the minimum voltage for specific speed.

Transistor Stack Effect

Subthreshold 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 “stacking effect”. It is better understood by considering a two transistor stack, as illustrated in Fig. 17. When both transistor M1 and M2 are turned off, the voltage at the intermediate node (V_X) is positive due to a small drain current. Positive potential at the intermediate node has three effects:

1. Due to the positive source potential V_X, gate-to-source voltage of transistor M1 (V_{gs1}) becomes negative; hence, the subthreshold current reduces substantially.
2. Due to $V_X > 0$, bulk-to-source potential (V_{bs1}) of transistor M1 becomes negative, increasing the threshold voltage (V_{th}) (larger body effect) of M1, and thus reducing the subthreshold leakage.
3. Due to $V_X > 0$, the drain-to-source potential (V_{ds1}) of transistor M1 decreases, increasing V_{th} (less DIBL) of M1, and thus reducing the subthreshold leakage.

The leakage of a two-transistor stack is about an order of magnitude less than the leakage in a single transistor, as already shown in Fig. 9.

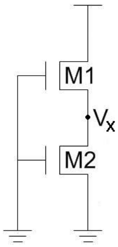

Fig. 17. Two NMOS off-transistor stack.

Input Dependence

Functional blocks such as NAND, NOR and other complex gates readily have stacks of transistors. Due to the stacking effect, the subthreshold leakage through a logic gate depends on the applied input vector. Maximizing the number of off transistors in a stack by applying proper input vectors can reduce the standby leakage of a functional block. Table 1 presents the input vector leakage dependence in a NAND gate for a standard 130nm CMOS process.

Table 1. Subthreshold leakage current for 2-input NAND gate.

Input Vector	Leakage current (nA)
00	*3.94*
01	*15.25*
10	*13.65*
11	*4.57*

Standby leakage power reduction resulted from minimum leakage input vector is a very effective way to control the subthreshold current in the standby mode of circuit operation. The most straightforward way to find a low leakage input vector is to enumerate all input combinations. For a circuit with ‘n’ inputs, there are 2^n input states combinations. Due to the exponential complexity, 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 vector.

Gate and band-to-band tunneling leakage are also important in scaled technologies, and can be a significant portion of total leakage.

The input vector control technique using a stack of transistors needs to be reviewed to effectively reduce the total leakage.

Researchers have shown that with high gate leakage, the traditional way of using stacked transistors fails to reduce leakage and in the worst case might increase the overall leakage [24]. In scaled technologies where gate leakage dominates the total leakage, using "10" might produce more savings in leakage as compared to "00". The gate leakage depends on the voltage drop across the transistor gate oxide. Applying "00" as the input to a two transistors stack reduces subthreshold leakage and does not change the gate leakage component. It has been shown that using "10" reduces the voltage drop across the terminals, where the gate leakage dominates, thereby lowering the gate leakage while offering marginal improvement in subthreshold leakage [24].

Band-to-band tunneling leakage is a weak function of input voltage and hence it can be neglect in this analysis [34].

Stacking Single Switch

In CMOS complex gates, a certain number of transistor stacking (branches), between the supply voltage or ground nodes and the output node, can be observed. Such branches have usually different amounts of transistors. The basic idea of this technique is to duplicate transistors without increasing the longest transistor path or branch, expecting that the worst-case delay of the logic cell remains the same [35]. This procedure is applied to both pull-up (PMOS network) and pull-down (NMOS network) separately. Fig. 18 (a) presents a circuit topology and Fig. 18 (b) illustrates the optimized circuit, resulted from the developed method described above.

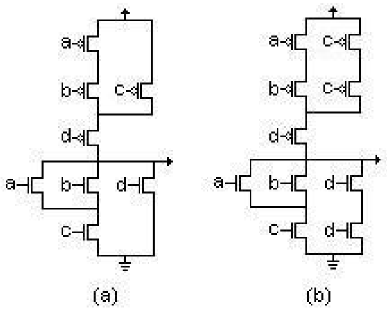

Fig. 18. (a) Original and (b) modified topology for leakage optimization CMOS gate.

Power Gating

Power gating technique uses the power supply voltage as the primary source for reducing leakage current. It refers to using a MOSFET switch (sleep transistor) to cut off, or gate, a circuit from the power rails (V_{dd} and/or gnd) during standby mode. The power gating switch typically is positioned as header between the circuit and the power supply or as footer between the circuit and the gnd. During active operation, the power gating switch remains on, supplying the current that the circuit uses to operate. During standby mode, turning off the power gating switch reduces the current dissipated through the circuit.

Turning off the sleep transistor provides leakage reduction for two primary reasons. First, the width of the sleep transistor is usually less than total width of transistors being gated. The smaller width provides a linear reduction in the total current drawn from supply node during standby mode. Secondly, leakage currents diminish whenever stacks of transistors are off due to the source biasing effect.

During active mode, the same effects cause degradation in circuit performance. Even though the on-resistance of the power gating switch is much less than its off-resistance, it still creates a small positive voltage at the virtual node. Again, these voltages reduce the drive capability and increase the threshold voltage of the NMOS devices through body biasing. Hence, this technique is typically used for paths that are non-critical.

MTCMOS

Multi-Threshold CMOS (MTCMOS) is a popular power gating approach that uses high V_{th} devices for power switches [36]. Fig. 19 shows the basic MTCMOS structure, where a low V_{th} computational block uses high V_{th} switches for power gating. Low V_{th} transistor in the logic gate provides a high performance operation. However, by introducing a series device to the power supplies, MTCMOS circuits incur a performance penalty compared to CMOS circuits.

In fact, only one type of high V_{th} transistor is sufficient for leakage reduction. The NMOS insertion scheme is preferable, since the NMOS on-resistance is smaller at the same width and hence it can be sized smaller than a corresponding PMOS [37].

However, MTCMOS can only reduce leakage power in standby mode and a large insertion of sleep transistors can increase significantly area and delay. Moreover, when data retention is required in standby mode, an extra high V_{th} memory circuit is needed to maintain the data [38].

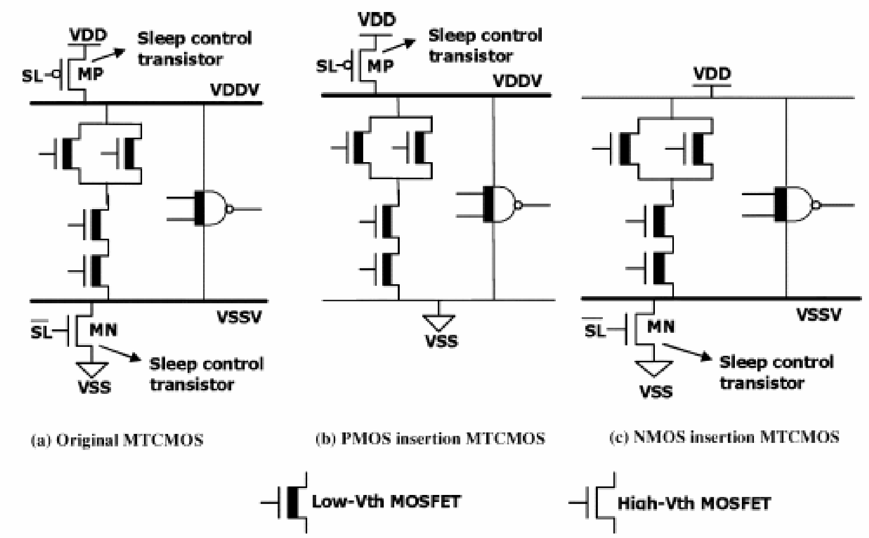

Fig. 19. Schematic of MTCMOS circuit [8].

Body Biasing

Reverse body biasing (RBB) has been used in memory chips since the years 70’s, to mitigate the risk of data destruction. In logic chips, on the other hand, the substrate and wells are typically biased stably to the ground and power supply. However, since the years 90’s, reverse body biasing has been applied in logic chips for a different reason: power reduction.

The original propose of the substrate biasing was utilized to reduce subthreshold leakage during standby mode in portable applications. More recently, it has been employed to reduce the maximum power dissipation by lowering V_{th} (forward body biasing) in active mode, and by compensating V_{th} variations.

Variable Threshold CMOS (VTCMOS)

Variable threshold CMOS is a body biasing based design technique [39]. Fig. 20 illustrates the VTCMOS scheme. To achieve different threshold voltages, it uses a self-substrate bias circuit to control the body bias.

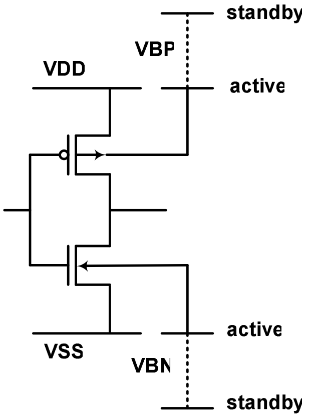

Fig. 20. Schematic of VTCMOS technique

In the active mode, VTCMOS technique applies a zero body bias (ZBB). As the subthreshold leakage current depends strongly on threshold voltage, in standby mode, a deep reverse body bias is applied to increase the V_{th}, saving thus leakage power. However, this reduction technique has an overhead in chip area due to additional signal routing required to provide the body bias voltage.

Reverse body bias can reduce circuit leakage by three orders of magnitude in a typical 0.35µm CMOS technology [40]. However, more recent data shows that the effectiveness of RBB decreases with technology scales due to the exponential increase in band-to-band tunneling leakage at the source/substrate and drain/substrate pn junctions [40]. Moreover, a small channel length and a low channel doping (to reduce V_{th}) worsen the short channel effect and mitigate the body effect. This, in turn, weakens the V_{th} modulation capability of RBB.

Recent design has been proposed using forward body biasing (FBB) to achieve better current drive with less short channel effect [41]. Circuit is designed using high V_{th} transistor (high channel doping) to reduce leakage in standby mode, while FBB is applied in active mode to achieve high performance. Both high channel doping and FBB reduce short channel effect relaxing the scalability limit of channel length due to V_{th} roll off and DIBL. This result in higher I_{on} compared to low V_{th} design for similar I_{off} worst case, improving performance. RBB can also be applied in standby mode together with FBB to further reduction in leakage current.

It has been shown that FBB and *high*-V_{th} along with RBB reduce leakage by 20X, as opposed to 3X for the RBB and *low*-V_{th} [41]. However, FBB devices have larger junction capacitance and body effect, which reduce the delay improvement mainly in stacked circuits.

Dynamic Vth Scaling

In many cases, applications do not require a fast circuit to operate at the highest performance level all the time. Active leakage techniques exploit this idea to intermittently slow down fast circuitry and reduce both leakage power and dynamic power consumption when maximum performance is not required.

Dynamic V_{th} scaling (DVTS) scheme uses body biasing to adjust V_{th} based on the performance demand [42]. This technique uses the same concept of previously discussed VTCMOS technique. However, VTCMOS changes V_{th} based on both active and standby mode, while DVTS modifies it based on circuit performance demand, i.e., the V_{th} is changed based on the system performance requirement.

The lowest V_{th} is delivered via zero body bias, when the highest performance is required. In the case of low performance demand, the clock frequency is lowered and V_{th} is raised via reverse body bias to reduce the run-time leakage power dissipation. In cases when there is no workload at all, the V_{th} can be increased to its upper limit to significantly reduce the standby leakage power. This scheme deliveries just enough throughput for the current workload by tracking the optimal V_{th}. A block diagram of the DVTS scheme and its feedback loop are presented in Fig. 21.

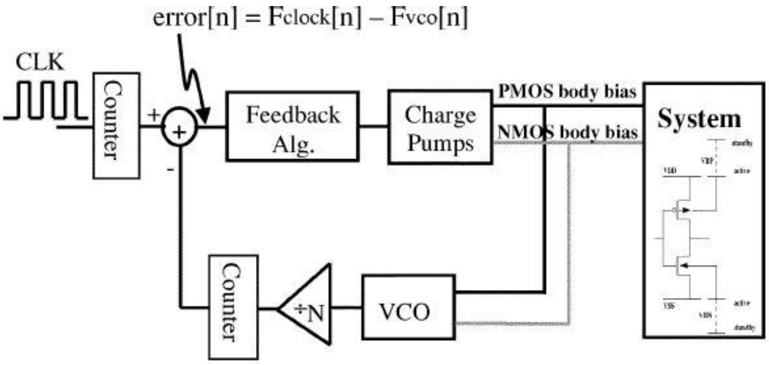

Fig. 21. Schematic of DVTS hardware [42].

CONCLUSION

Power dissipation of electronic products has become an important issue with the massive growth in portable computing and wireless communication in the last few years. As power consumption is directly proportional to the square of the power supply voltage, MOS transistor has been scaled to maintain performance at reduced supply voltage. Transistor threshold voltage is also reduced to avoid short channel effect, resulting in a substantial increasing in leakage currents when transistor scaling into nanometer dimensions. Standby current becomes a significant portion of the total IC power consumption, being a challenge for designers and a critical factor in low-power circuits. It means the static power dissipation should be considered as soon as possible in the design flow. Leakage mechanisms and reduction techniques have been reviewed, providing a minimum background about this issue.

Analysis of leakage current mechanisms was initially presented. Such review demonstrated that subthreshold leakage represents the main leakage mechanism, and with the transistor scaling into sub-100nm dimensions, the gate leakage achieves the same order of importance.

Estimation techniques for both subthreshold and gate leakage mechanisms have been discussed. Accurate models to estimate subthreshold leakage have to treat the stack effect. Gate leakage estimation, on the other hand, is usually based on transistor biasing.

Leakage reduction techniques were also revised to complete the basic leakage knowledge. Dual-threshold CMOS, which uses high V_{th} transistors in non critical path to achieve leakage reduction without performance penalties, was the first technique outlined. Supply voltage scaling, usually used to reduce active power, is also a good alternative to static power reduction. Other techniques that explore staking effect were also discussed, as well as techniques based on power gating and body biasing principle.

REFERENCES

[1] GRONOWSKI, P. E. et al. High-Performance Microprocessor Design. **IEEE Journal of Solid State Circuits**, New York, v.33, n.5, p. 676-686, May 1998.

[2] TAKAYANAGI, T. et al. A Dual-Core 64-bit UltraSPARC Microprocessor for Dense Server Applications. **IEEE Journal of Solid State Circuits**, New York, v.40, n.1, p. 7-17, Jan. 2005.

[3] LEON, A. S. et al. A Power-Efficient High-Throughput 32-Thread SPARC Processor. **IEEE Journal of Solid State Circuits**, New York, v.42, n.1, p. 7-16, Jan 2007.

[4] PARK, J. C.; MOONEY III, V. J. Sleepy Stack Leakage Reduction. **IEEE Transactions on VLSI Systems**, New York, v.14, n.11, p.1250-1262, November 2006.

[5] VEENDRICK, H.J.M. Short-Circuit Dissipation of Static CMOS Circuitry and its Impact on the Design of Buffer Circuits. **IEEE Journal of Solid State Circuits**, New York, v.SC-19, n.4, p. 468-473, Aug. 1984.

[6] SOUDRIS, D. et al. **Designing CMOS circuits for low power**. Boston: Kluwer Academic, 2002. 277p.

[7] MOORE, G. E. No exponential is forever: but "Forever" can be delayed! In: IEEE INT. CONF. SOLID STATE CIRCUITS, 2003. **Proceedings…** IEEE, 2003, p. 20-23.

[8] ROY, K. et al. Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits. **Proceedings of IEEE**, New York, v.91, n.2, p. 305-327, Feb 2003

[9] AGARWAL, A. et al. Leakage Power Analysis and Reduction: Models, Estimation and Tools, **Proc. IEE**, v.152, n.3, p 353-368, May 2005

[10] SHEU, B. J. et al. BSIM: Berkeley Short-Channel IGFET Model for MOS Transistors. **IEEE Journal of Solid State Circuits**, New York, v.SC-22, n.4, p. 558-566, Aug. 1987.

[11] CAO, K.M. et al. BSIM4 Gate Leakage Model Including Source-Drain Partition.In: INT. ELECTRON DEVICES MEETING, 2000. **Digest of Technical Papers**, Dec. 2000, p. 815-818.

[12] CHEN, Z. et al. Estimation of Standby Leakage Power in CMOS Circuits Considering Accurate Modeling of Transistor Stacks. In: INT. SYMP. LOW POWER ELECTRONICS

AND DESING, ISLPED, 1998. **Proceedings…**, New York: ACM SIGDA, 1998, p.239-244.

[13] YANG, S. et al. Accurate Stacking Effect Macro-modeling of Leakage Power in Sub-100nm Circuits. In: INT. CONFERENCE ON VLSI DESIGN, 2005. **Proceegins…** 2005, p. 165-170.

[14] SINGER, P. Intel and IBM Commit to High-K, Metal Gates. **Semiconductor International**. Available at: <http://www.reed-electronics.com/semiconductor/article/CA6410945?spacedesc= news>. Visited on: Jan. 2007.

[15] OGURA, S. et al. Design and Characteristics of the Lightly Doped Drain-Source (LDD) Insulated Gate Field-Effect Transistor. **IEEE Journal of Solid State Circuits**, New York, v. SC-15, n.4, p. 424-432, Aug 1980.

[16] DONAGHY, D. et al. A Simulation Study to Quantify the Advantages of Silicon-On-Insulator (SOI) Technology for Low Power. In: LOW POWER IC DESIGN SEMINAR, 2001. **Proccedings…** London: IEE, 2001, p. 11/1-11/6.

[17] CHANDRAKASAN, A. P.; BRODERSEN, R. W. Minimizing Power Consumption in Digital CMOS Circuits. **Proceedings of the IEEE**, New York, v.83, n.4, p.498-523, April 1995.

[18] ROY, K.; PRASAD, S. C. **Low-Power CMOS VLSI Circuit Design**. New York: Wiley Interscience, 2000. 359 p.

[19] MUKHOPADHYAY, S. et al. Accurate Estimation of Total Leakage in Nanometer-Scale Bulk CMOS Circuits Based on Device Geometry and Doping Profile. **IEEE Trans. on CAD of IC and Systems**, New York, v.24, n.3, p. 363-381, Mar. 2005.

[20] GU, R. X.; ELMASRY, M.I. Power Dissipation Analysis and Optimization of Deep Submicron CMOS Digital Circuits. **IEEE Journal of Solid State Circuits**, New York, v.31, n.5, p. 707-713, May 1996.

[21] NARENDRA, S. G.; CHANDRAKASAN, A. **Leakage in Nanometer CMOS Technologies**. New York: Springer, 2006. 307 p.

[23] RAO, R. M. et al. Efficient techniques for gate leakage estimation. In: INT. SYMP. LOW POWER ELECTRONICS AND DESING, 2003. **Proceedings…**, New York: ACM, 2003, p.100-103.

[24] MUKHOPADHYAY, S. et. al. Gate Leakage Reduction for Scaled Device Using Transistor Stacking. **IEEE Trans. on VLSI Systems**, New York, v.11, n.4, p. 716-730, Aug 2003.

[25] GUINDI, R. S.; NAJM, F. N. Design techniques for gate-leakage reduction in CMOS circuits. In: INT. SYMP. QUALITY ELECTRONIC DESIGN, 2003. **Proceedings…**, IEEE, 2003, p.61-65.

[26] WEI, L. et. al. Design and Optimization of Dual-Threshold Circuits for Low-Voltage Low-Power Applications. **IEEE Trans. on VLSI Systems**, New York, v.7, n.1, p. 16-24, March 1999.

[27] BOWMAN, K. A et al. Impact of Die-to-Die and Within Die Parameter Fluctuations on the Maximum Clock Frequency Distribution fo Gigascale Integration. **IEEE Journal of Solid State Circuits**, New York, v.37, n.2, p. 183-190, Feb. 2002.

[28] KRISHNARNURTHY, R. K. et al. High-Performance and Low-Power Challenges for Sub-70 nm Microprocessor Circuits. In: IEEE CUSTOM INTEGRATED CIRCUIT CONFERENCE, 2002. **Proceedings…** IEEE, 2002, p. 125-128.

[29] WANG, A.; CALHOUN, B.; CHANDRAKASAN, A. P. **Sub-Threshold Design for Ultra Low-Power Systems**. New York: Springer, 2006. 209 p.

[30] TAKAHASI, M. et al. A 60-mW MPEG4 video codec using clustered voltage scaling with variable supply-voltage scheme. **IEEE Journal of Solid State Circuits**, New York, v.33, n.11, p. 1772-1780, Nov. 1998.

[31] FUSE, T. et al. A 0.5 V Power-Supply Scheme for Low Power LSIs Using Multi-Vt SOI CMOS Technology. In: SYMPOSIUM ON VLSI CIRCUITS, 2001. **Digest of Technical Papers.** [S. l.]: IEEE, 2001. p. 219-220.

[32] CARLEY, L. R.; AGARWAL, A. A Completely On-Chip Voltage Regulation Technique for Low Power Digital Circuits. In: INT. SYMP. LOW POWER ELECTRONIVS AND DESING, 1999. **Proceedings…**, New York: ACM, c2000, p.109-111.

[33] BURD, T. D. et al. A Dynamic Voltage Scaled Microprocessor System. **IEEE Journal of Solid State Circuits**, New York, v.35, n.11, p. 1571-1580, Nov. 2000.

[34] AGARWAL, A. et al. Leakage Power Analysis and Reduction for Nanoscale Circuits, **IEEE Micro**, Los Alamitos, v.26, n.2, p 68-80, Mar. 2006

[35] BUTZEN, P. F. et al. Leakage reduction technique for CMOS complex gates. In: SOUTH SYMPOSIUM ON MICROELECTRONICS, 21., 2006, Porto Alegre. **Proceedings**… Porto Alegre: Instituto de Informática, UFRGS, 2006, p.111-114.

[36] MUTOH, S. et al. 1-V Power Supply High-speed Digital Circuit Technology with Multi-threshold Voltage CMOS. **IEEE J. of Solid State Circuits**, New York, v.30, n.8, p. 847-854, Aug. 1995.

[37] KAO, J. et al. Transistor Sizing Issues and Tool for Multi-Threshold CMOS Technology. In: DESIGN AUTOMATION CONFERENCE, 1997. **Proceedings…** [S.l.]: IEEE, 1997, p.409-414.

[38] SHIGEMATSU, S. et al. 1-V high-speed MTCMOS circuit scheme for power-down application circuits. **IEEE Journal of Solid State Circuits**, New York, v.32, n.6, p. 861-869, Jun. 1997.

[39] KURODA, T. et al. A 0.9-V, 150-MHz, 10-mW, 4 mm^2, 2-D Discrete Cosine Transform Core Processor with Variable Threshold-Voltage (VT) Scheme. **IEEE Journal of Solid State Circuits**, New York, v.31, n.11, p. 1770-1779, Nov 1996.

[40] KESHAVARZI, A. et al. Effectiveness of Reverse Body Bias for Leakage Control in Scaled Dual Vt CMOS ICs. In: INT. SYMP. LOW POWER ELECTRONIVS AND DESING, 1998. **Proceedings…**, New York: ACM 2001, p.207-212.

[41] NARENDRA, S. et al. Forward Body Bias for Microprocessors in 130-nm Technology Generation and Beyond. **IEEE Journal of Solid State Circuits**, New York, v.38, n.5, p. 696-701, May 2003.

[42] KIM, C. H.; ROY, K. Dynamic V_{th} Scaling Scheme for Active Leakage Power Reduction. In: DESIGN AUTOMATION AND TEST IN EUROPE CONFERENCE, 2002. **Proceedings…** IEEE, 2002, p. 163-167.