Technology Scaling And Low Power Design Techniques

T.Esther Rani[1], Dr. Rameshwar Rao [2] and Dr.M.Asha Rani[3]
[1]CVR College of Engineering, Department of ECE., Ibrahimpatan, R.R.District, A.P., India
Email: estherlawrenc@gmail.com
[2]Hon'ble VC, JNT University, Hyderabad, India
Email:Rameshwar_rao@hotmail.com
[3]JNT University, Department of ECE, Hyderabad, A.P., India
Email: ashajntu1@yahoo.com

Abstract —**Scaling the feature size of transistor made a remarkable advancement in silicon industry. The demand for power-sensitive design has grown significantly in recent years due to growth in portable applications. The need for power-efficient design techniques is increasing. Various efficient design techniques have been proposed to reduce both dynamic as well as static power in state-of-the-art VLSI applications. In this paper, different circuit design techniques both static and dynamic are discussed that reduce the power consumption.**

***Index Terms*—Scaling, feature size, power-efficient, design techniques.**

I. Introduction

In the past, the major concerns of the VLSI designer were area, performance, cost, reliability. Power consideration was mostly of secondary importance. In recent years this has begun to change. Power is being given comparable importance to area and speed. Several factors have contributed to this trend. The primary driving factor has been a remarkable success in terms of growth of the class of personal computing devices and wireless communications systems, which demand high-speed computation and complex functionality with low power consumption. The semiconductor industry has witnessed the growth in demand and supply of portable systems in consumer electronics market. High performance portable products, ranging from small hand-held personal communication devices, such as pagers and cellular phones, to larger and more sophisticated products that support multimedia applications, such as lap-top and palm-top computers, enjoyed considerable success among consumers.

According to ITRS, battery life for these devices peaked in 2004. However, battery life is reduced because additional features have been added faster. For all applications, reducing the power consumed by SoCs is essential in order to have better performance with additional features.

The scaling implies lower supply voltages. Since dynamic power is proportional to the square of the supply voltage, scaling provides an effective way to reduce power consumption. Unfortunately, supply voltage scaling adversely affects performance of any circuit. Reduction of the threshold voltage is the most intuitive solution to this problem, because it can be achieved as a natural by-product of technology scaling. The drawback of lowering the threshold voltage is that, it leads to an exponential increase in the sub-threshold leakage current, thus originating a remarkable increase in leakage power consumption. As a consequence, leakage power is expected to become larger than its dynamic counterpart. In future-nanometer designs, this is further boosted by increased process variations in this design space [5].

Scaling of transistor threshold voltage is associated with exponential increase in sub threshold leakage current [1]. Aggressive scaling of the devices not only increases the sub threshold leakage but also has other negative impacts such as increased drain-induced barrier lowering (DIBL), V_{th} roll-off, reduced on-current to off current ratio, and an increase in source–drain resistance [2]. A small variation in channel length might result in large V_{th} variation, which makes device characteristics unpredictable. To avoid these short channel effects, oxide thickness scaling and non uniform doping need to be incorporated [3] as the devices are scaled. The low oxide thickness gives rise to high electric field, resulting in considerable direct tunneling current [4]. Higher doping results in high electric field across the reverse biased p–n junctions (source–substrate or drain–substrate) which cause significant band-to-band tunneling (BTBT) of electrons from the valence band of the p-region to the conduction band of the n-region. Peak halo doping (P+) is restricted such that the BTBT component is maintained reasonably small compared to the other leakage components.

There are other leakage mechanisms apart from gate, junction BTBT and sub threshold leakage which are product of small geometries. For example, as drain voltage V_D increases, the drain to channel depletion region widens and significant drain current can result. This increase in I_{OFF} is typically due to channel surface current from DIBL or due to deep channel punch through currents [2,4,5]. Moreover, as the channel width decreases, the threshold voltage and the off current both get modulated by the width of the transistor, giving rise to significant narrow-width effect resulting in channel surface current. Gate-induced drain leakage (GIDL) is another significant leakage mechanism due to the depletion at the drain surface below the gate–drain overlap region. However, during normal mode of

DOI:10.32377/cvrjst0207

operation, the major leakage currents are gate-leakage, junction BTBT and sub threshold leakage.

In this paper, we describe different circuit design techniques to reduce both dynamic and leakage power. We review a spectrum of circuit techniques including transistor sizing, clock gating, multiple and dynamic supply voltage for reducing dynamic power. For low-leakage design, different circuit techniques including, dual Vth, forward/reverse bias, dynamically varying the Vth during run time, sleep transistor, natural stacking are reviewed. Based on these techniques, different leakage tolerant schemes for logic and memories are summarized.

II. Power dissipation in Integrated Circuits

The total power dissipation in a circuit conventionally consists of two components, namely, the static and dynamic power dissipation. Many circuit techniques have been proposed to reduce these components in VLSI circuit design. Some of these techniques are suitable for reducing the static components while some techniques are efficient to reduce the dynamic power dissipation of the circuit.

A. Dynamic power

For dynamic power dissipation there are two components one is switching power due to charging and discharging of load capacitance. The other is the short circuit power due to the nonzero rise and fall time of input waveforms. The switching power of a single gate can be expressed as

$$P_D = \alpha C_L V_{DD}^2 f \tag{1}$$

Where α is the switching activity, f the operation frequency, C_L the load capacitance and V_{DD} the supply voltage. The short circuit power of an unloaded inverter can be approximately given by

$$P_{SC} = \frac{\beta}{12}(V_{DD} - V_{th})^3 \frac{\tau}{T} \tag{2}$$

Where β is the transistor coefficient, t the rise/fall time and T (1/f) the delay.

Each of the dynamic power reduction techniques described in the latter sections optimizes the above parameters (e.g. α, V_{DD}, C_L) to achieve a low-power design.

B. Static power

There are three dominant components of leakage in a MOSFET in the nanometer regime:

(i) Subthreshold leakage, which is the leakage current from drain to source $I_{sub.}$

(ii) Direct tunneling gate leakage, which is due to the tunneling of electron (or hole) from the bulk silicon through the gate oxide potential barrier into the gate.

(iii) The source/substrate and drain/substrate reverse-biased p–n junction BTBT leakage. This leakage component is expected to be large for sub-50nm devices [6].

Other components of leakage current such as GIDL, impact ionization, etc. are not expected to be large for regular CMOS operations [7].

*(i). Subthreshold leakage:*Subthreshold or weak inversion conduction current between source and drain in a 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} that is used. The weak inversion current can be expressed based on the following equation.

$$I_{\text{subth}} = A e^{(q/nkT)(V_{GS} - V_{TH0} - \gamma' V_{SB} + \eta V_{DS})}\left(1 - e^{(-qV_{DS}/kT)}\right) \tag{3}$$

$$A = \mu_0 C'_{ox} \frac{W}{L_{eff}} \left(\frac{kT}{q}\right)^2 e^{1.8} \tag{4}$$

Where V_{GS}, V_{DS}, and V_{SB} are the gate voltage, drain voltage and body voltage of the transistor, respectively. Body effect is represented by the term γV_{SB}, where γ is the linearized body effect coefficient. η is the DIBL coefficient, representing the effect of V_{DS} on threshold voltage. C_{ox} is the gate oxide capacitance. μ_0 is the zero bias mobility and n is the sub threshold swing coefficient of the transistor. This equation shows the exponential dependency of sub threshold leakage on V_{th0}, V_{GS}, V_{DS} (due to DIBL), and V_{SB}. Each of the leakage reduction techniques described in the latter sections utilized these parameters in a MOSFET to achieve a low leakage state.

(ii). Gate leakage: Gate direct tunneling current is due to the tunneling of electron (or hole) from the bulk silicon through the gate oxide potential barrier into the gate. The direct tunneling is modeled as

$$J_{DT} = A(V_{ox}/T_{ox})^2 \exp\left(\frac{-B(1-(1-V_{ox}/\phi_{ox})^{3/2})}{V_{ox}/T_{ox}}\right) \tag{5}$$

Where J_{DT} is the direct tunneling current density, V_{ox} is the potential drop across the thin oxide, ϕ_{ox} is the barrier height of tunneling electron and t_{ox} is the oxide thickness [8]. The tunneling current increases exponentially with decrease in oxide thickness. It also depends on the device structure and the bias condition [9].

(iii). Source/substrate and drain/substrate PN junction leakage: Drain and source to well junctions are typically reverse-biased causing pn junction leakage current. A reverse bias p-n junction leakage 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. Figure 1 shows gate leakage and junction currents.

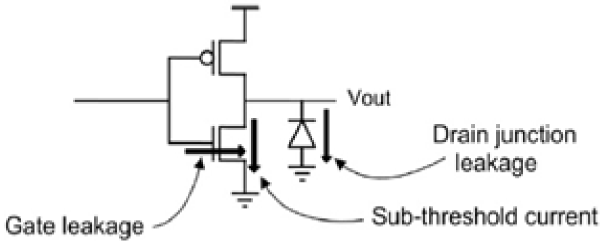

Figure1: Leakage currents

In the presence of a high electric field (4106 V/cm) electrons will tunnel across a reverse-biased p–n junction. A significant current can arise as electrons tunnel from

DOI:10.32377/cvrjst0207

the valence band of the p-region to the conduction band of the n-region [4]. Tunneling occurs when the total voltage drop across the junction is greater than the semiconductor band-gap. Since silicon is an indirect band-gap semiconductor the BTBT current in silicon involves the emission or absorption of phonons.

III. Power Reduction Methods

To reduce power, the semiconductor industry has adopted a multifaceted approach for attacking the problem on four fronts:

1. Reducing chip and package capacitance: This can be achieved through process development such as SOI with partially or fully depleted wells, CMOS scaling to submicron device sizes, and advanced interconnect substrates such as Multi-Chip Modules (MCM).This approach can be very effective but is also very expensive and has its own pace of development and introduction to the market.
2. Scaling the supply voltage: This approach can be very effective in reducing the power dissipation, but often requires new IC fabrication processing. Supply voltage scaling also requires support circuitry for low-voltage operation including level-converters and DC/DC converters as well as detailed consideration of issues such as signal-to-noise.
3. Employing better design techniques: This approach promises to be very successful because the investment to reduce power by design is relatively small in comparison to the other three approaches. Also it is relatively untapped in potential.
4. Using power management strategies: The power savings that can be achieved by various static and dynamic power management techniques are very application dependent.

IV. Dynamic Power Reduction Techniques

Though the leakage power increases significantly in every generation with technology scaling, the dynamic power still continues to dominate the total power dissipation of the general purpose microprocessors. Consequently, several device, circuit and architecture-level techniques have been proposed to reduce the dynamic power consumption of the circuit. Effective circuit techniques include transistor size and interconnect optimization, gated clock, multiple supply voltages and dynamic control of supply voltage. Incorporating the above approaches in the design of nanoscale circuits, the dynamic power dissipation can be reduced significantly. Other techniques such as instruction set optimization [11,12,13], memory access reduction [14,15] and low complexity algorithms [16,17] are also proposed to reduce the dynamic power dissipation in both logics and memories. However, in this paper we will focus mainly on low-power circuit techniques. In the following subsections, we will discuss the circuit techniques and their effectiveness in reducing the dynamic power dissipation.

A. Transistor sizing and interconnect optimization

The best way to reduce the junction capacitance as well as the overall gate capacitance is to optimize the transistor size for a particular performance. Several sizing techniques have been proposed to minimize the circuit area (hence, the power) while maintaining the performance [18-20]. Sizing techniques can be mainly divided into two types, path-based optimization and global optimization. In path-based optimization, gates in the critical paths are upsized to achieve the desired performance, while the gates in the off critical paths are down sized to reduce power consumption [18,19]. In global optimization [20], all gates in a circuit are globally optimized for a given delay.

With the scaling of technology, while local interconnect capacitances reduce every generation, the global interconnect capacitances, however, increase with scaling [21,22]. This is because the die size is increasing every generation resulting in larger global interconnect lengths. This increase in global interconnect lengths effectively increases the interconnect delay. To cope with this problem, wires which are wider than the minimum-sized global interconnects provided by the technology are used [23]. Increasing the width of interconnect proportionally reduces its resistance per unit length and also increases the line capacitance per unit length, which effectively increases the interconnect power. Several optimization techniques and algorithms have been proposed to reduce the interconnect delay as well as power [23-25]. These techniques provide the optimum width, height and the spacing between the wires. It is shown that by interconnect optimization, a significant amount of saving in power dissipation can be achieved [23,26].

B .Clock gating

Clock gating is an effective way of reducing the dynamic power dissipation in digital circuits [27-29]. In a typical synchronous circuit such as the general purpose microprocessor, only a portion of the circuit is active at any given time. Hence, by shutting down the idle portion of the circuit, the unnecessary power consumption can be prevented. One of the ways to achieve this is by masking the clock that goes to the idle portion of the circuit. This prevents unnecessary switching of the inputs to the idle circuit block, reducing the dynamic power. In addition, it saves the clock power by preventing any redundant switching in the clock tree. Figure2 shows a schematic diagram of gated clock design [30].

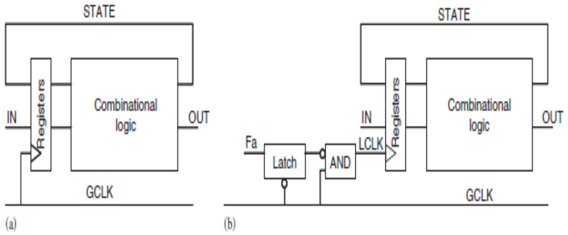

Figure2: A schematic diagram of gated clock design: (a) Single clock, flip-flop-based FSM, (b) gated-clock design.

DOI:10.32377/cvrjst0207

The inputs to the combinational logic come through the registers, which are usually composed of sequential elements, such as D flip-flops. A gated clock design can be obtained by modifying the clocking structure shown in Fig. 2(a). A control signal (fa) is used to selectively stop the local clock (LCLK) when the combinational block is not used. The local clock is blocked when fa is high. The latch shown in Fig. 2(b) is necessary to prevent any glitches in fa from propagating to the AND gate when the global clock (GCLK) is high. The circuit operates as follows. The signal fa is only valid before the rising edge of the global clock. When the global clock is low, the latch is transparent, however, fa does not affect the AND gate. If fa is high during the low-to-high transition of the global clock, then the global clock will be blocked by the AND gate and local clock will remain at low.

Power saving using gated clock technique strongly depends on the efficient synthesis and optimization of dedicated clock-stopping circuitry. Effective clock gating requires a methodology that determines which circuits are gated, when, and for how long. Clock-gating schemes that either result in frequent toggling of the clock-gated circuit between enabled and disabled states, or apply clock gating to such small blocks that the clock-gating control circuitry is almost as large as the blocks themselves, incur large overhead. This overhead may result in power dissipation to be higher than that without clock gating. There are many synthesis algorithm proposed to efficiently cluster the circuit module for clock gating [27-30] as well as for proper clock routing to minimize the switching capacitance of the clock tree [28,29].

An architecture-level technique called deterministic clock gating (DCG) is also proposed. This technique is based on the key observation that for many of the pipeline stages in a modern processor, a circuit block usage in a specific cycle in the near future is deterministically known a few cycles ahead of time.

C .Low-voltage operation

Supply voltage scaling was originally developed for switching power reduction. It is an effective method for switching power reduction because of the quadratic dependency of switching power on supply voltage. Supply voltage scaling also helps reduce leakage power since the sub-threshold leakage due to GIDL and DIBL decreases as well as the gate leakage component when the supply voltage is scaled down. In a 1.2 V, 0.13 mm technology, it is shown that the supply voltage scaling has impact in the orders of V3 and V4 on sub-threshold leakage and gate leakage, respectively. However, since the gate delay increases with decreasing V_{DD}, globally lowering V_{DD} degrades the overall circuit performance. To achieve low-power benefits without compromising performance, two ways of lowering supply voltage can be employed: static and dynamic supply scaling.

(i) Static supply voltage scaling schemes: In this technique, higher supply voltage is used in the critical paths of the circuit, while lower supply voltages are used in the off critical paths. Since the speed requirements of the non-critical units are lower than the critical ones, supply voltage of non-critical unit clusters can be lowered without degrading system performance. The secondary voltages may be generated off-chip [31] or regulated on-die from the core supply [32]. Depending on how many supply voltages are available, voltage scaling may be classified as multiple voltage approach or dual voltage approach.

In dual/multiple supply voltage technique, whenever an output from a low V_{DD} cluster has to drive an input to a high V_{DD} cluster, a level conversion is needed at the interface [33]. If a gate operating at a lower supply voltage directly drives a gate operating with a higher supply voltage, a large amount of static current is likely to flow through the PMOS transistors of the gate with higher supply voltage. This is because when the output of the low-voltage gate is high, its voltage level may not be sufficient to turn-off the PMOS of the succeeding high-voltage gate. To avoid this problem level converters are used to convert the low-voltage output to the high-voltage level. A typical level converter circuit is shown in Fig. 3. [34]. While the level converter eliminates the static or short circuit power dissipation, the power consumption within itself may be substantial. However, this technique is not very effective with tight timing constraints. The gate clustering problem for multiple supply voltage operation can be simplified if level converters are used. Several techniques have been proposed to synthesize the circuit with dual supply voltage [34,35,37,38] considering the power consumption and/or delay of the level converters.

Another important issue of multiple supply voltage design is to decide how many supply voltages are optimum for a design and what should be the supply voltages. Several techniques have been proposed to select the optimal set of supply voltages taking level converters into account [34,37]. Techniques to select the optimal low supply voltage in a dual supply design are also proposed [39,40].

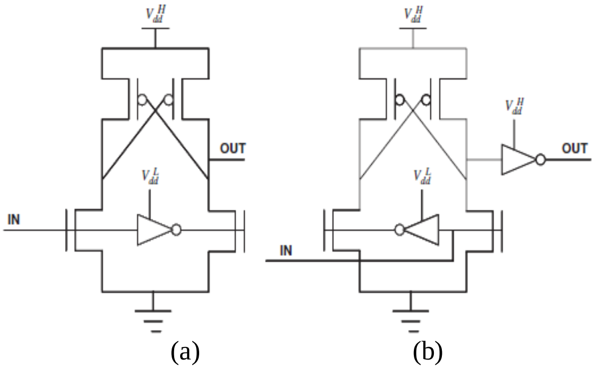

Figure3: (a) Level converter circuit for interfacing gates operating at a lower supply voltage to gates at a higher supply voltage, (b) can be used to interface low-voltage gates with multiple fanouts with high-voltage gates.

(ii) Dynamic supply voltage scaling schemes: Dynamic supply scaling overrides the cost of using two supply voltages, by adapting the single supply voltage to the 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, just delivering

DOI:10.32377/cvrjst0207

the required performance with substantial power reduction [41]. There are three key components for implementing DVS in a general-purpose processor: (1) an operating system that can intelligently vary the processor speed, (2) a regulation loop that can generate the minimum voltage required for the desired speed, and (3) a microprocessor that can operate over a wide voltage range. Fig. 4 shows a DVS system architecture [42]. 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-wire frequency–voltage feedback loop, using a ring oscillator as a critical path replica. All chips operate at the same clock frequency and same supply voltage, which are generated from the ring oscillator and the regulator.

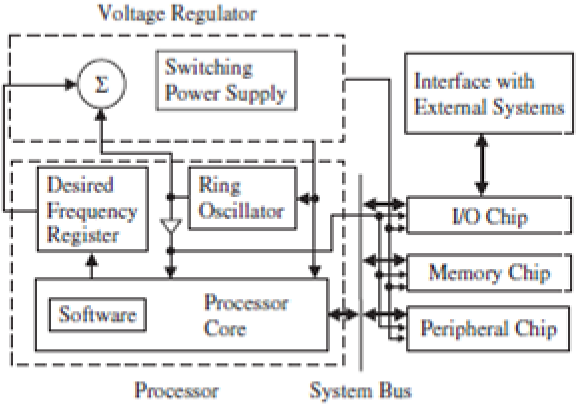

Figure4: Dynamic Voltage scaling architecture

V. Leakage Power Reduction Techniques

Since circuits are mostly designed for the highest performance and overall system cycle time requirements, they are composed of large gates, highly parallel architectures with logic duplication. As such, the leakage power consumption is substantial for such circuits. However, not every application requires a fast circuit to operate at the highest performance level all the time.

Modules, in which computation is less e.g. functional units in a microprocessor or sections of a cache, are often idle. It is of interest to conceive of methods that can reduce the leakage power consumed by these circuits. Different circuit techniques have been proposed to reduce leakage energy utilizing this slack without impacting performance. Standby leakage reduction techniques put the entire system in a low leakage mode when computation is not required. Active leakage reduction techniques slow down the system by dynamically changing the V_{th} to reduce leakage when maximum performance is not needed. In active mode, the operating temperature increases due to the switching activities of transistors. This has an exponential effect on sub threshold leakage making this the dominant leakage component during active mode and amplifying the leakage problem.

Design time techniques exploit the delay slack in non-critical paths to reduce leakage. These techniques are static; once it is fixed, it cannot be changed dynamically while the circuit is operating. Design time techniques include, dual threshold CMOS, changing doping profile, higher oxide thickness, higher oxide thickness.

Run time techniques include Standby leakage reduction, natural transistor stacks, sleep transistor, forward/reverse body biasing, active leakage reduction techniques, dynamic Vth scaling (DVTS), circuit techniques to reduce leakage in cache memories[5].

Many circuit techniques targeting the reduction of leakage power have appeared in the literature. Most of these techniques target the circuits during the standby operation mode and some target the circuits during the active mode of operation. Also, some of these techniques mainly aim to reduce sub-threshold leakage current while others tend to reduce gate leakage current. We can categorize leakage reduction circuit techniques into the following classes:

A. Transistor stacking techniques

All these techniques are based on the fact that when there are two or more stacked transistors which are switched OFF. Transistor stacking has a dual effect in reducing the sub-threshold leakage current. It increases the source bias of upper transistors in the stack and also lowers the gate-to-source voltages of these transistors, as shown in Fig.5. Both of these effects result in a lower sub-threshold leakage current.

Reducing leakage through the use of transistor stacks depends on the choice of input pattern during standby periods since it determines the number of OFF transistors in the stack. Leakage current dependencies on circuit state can be exploited and used to determine a low leakage state by using a heuristic search algorithm to find the minimum leakage input vector, which is fed into the circuit during sleep mode [43].

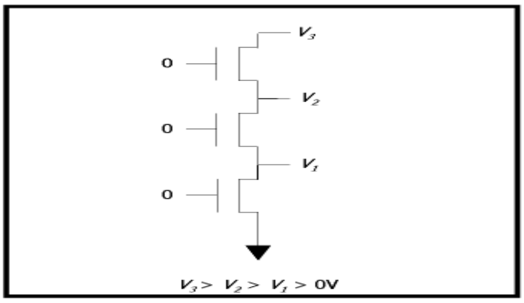

Figure5: The effect of transistor stacks in reducing Isub

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 ‘2n’ 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 a genetic algorithm to exploit historical information to speculate on new search points with

DOI:10.32377/cvrjst0207

expected improved performance to find a near optimal solution. The reduction of standby leakage power by application of an input vector is a very effective way of controlling the sub-threshold leakage in the standby mode of operation of a circuit.

After this input vector is found, the circuit is evaluated and additional leakage control transistors are inserted in the non-critical paths where only one transistor is originally turned OFF. A PMOS sleep transistor is used along with body bias and clock gating techniques to reduce the leakage power. A low block reactivation time after exiting the sleep mode is also maintained.

B. Multi-Vth techniques

This is one of the most common approaches to reduce leakage currents where two different types of transistors are fabricated on the chip, a high V_{th} to lower sub-threshold leakage current and a low V_{th} to enhance circuit performance by increasing its speed. Based on the multi-threshold technologies previously described, several multiple-threshold circuit design techniques have been developed [44, 45].

Multi-threshold voltage CMOS: reduces the leakage by inserting high-threshold devices in series to low V_{th} circuitry. Fig.6 (a) shows the schematic of an MTCMOS circuit. A sleep control scheme is introduced for efficient power management. In the active mode, Sleep 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 (Virtual Vdd and Virtual GND) almost function as real power lines. In the standby mode, Sleep 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.6 (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. NMOS can be sized smaller than corresponding PMOS. MTCMOS can be easily implemented based on existing circuits [46,47,48].

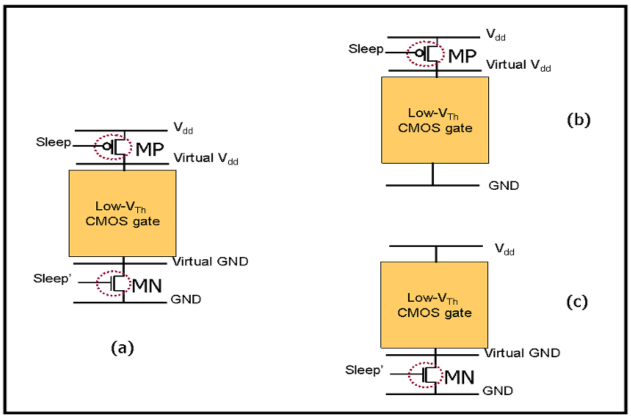

Figure6: MTCMOS Technique

However, MTCMOS can only reduce the standby leakage power, and the large inserted MOSFETs can increase the area and delay. Moreover, if data retention is required in the standby mode, an extra high V_{th} memory circuit is needed to maintain the data. Instead of using high V_{th} sleep control transistors as MTCMOS, super cutoff CMOS (SCCMOS) technique uses low V_{th} transistors with an inserted gate bias generator, as depicted in Fig.7. For the PMOS (NMOS) insertion, the gate is applied to 0V (Vdd) in the active mode, and the virtual Vdd (Virtual GND) line is connected to supply Vdd (GND). In the standby mode, the gate is applied to Vdd+ΔV (GND –ΔV) to fully cut off the leakage current. Compared with MTCMOS, SCCMOS circuits can work at lower supply voltages.

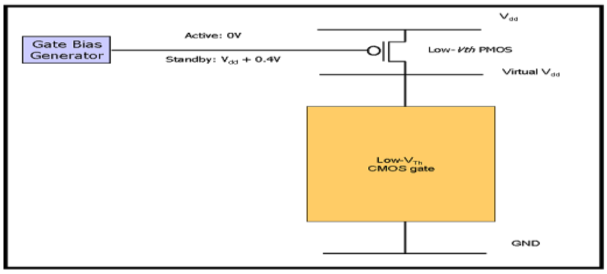

Figure 7: Super Cutoff CMOS (SCCMOS) Technique

However, this technique suffers from two main drawbacks: There are area overhead and some performance degradation. The overhead can be reduced if sleep transistor insertion is applied to cluster of gates, instead of single gates, as shown in Figure 8. Clustered sleep transistor insertion implies addressing a number of issues, including the granularity of the insertion .For large CMOS blocks, the size of the sleep transistors and the driving strengths of sleep signals may become large and for small CMOS blocks the number of sleep transistors and the size of the control logic may become large. The design of the sleep transistor cells .Which must have different sizes and driving strengths and that must be compliant with the cells in the library. The required area and delay control implying some constraints on the selection of gates to which sleep transistor insertion should be applied and the need of layout information, the automatic generation of the sleep signals which implies some area, timing and power overhead. Several of the issues above have been addressed in.

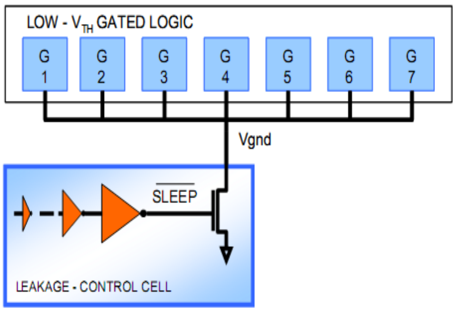

Figure 8: Clustered MTCMOS/SSCMOS technique

Another innovative circuit technique to reduce leakage current is the Smart Series Switch (Triple-S) technique. In this technique two parallel switches are connected in series with a leaky device. A low V_{th} transistor switch as a function of the operation mode (active/standby) and a high V_{th} transistor switch as a function of the state of the leaky device. The Triple-S technique suffers from an area overhead ranging from 20% to 40%.

DOI:10.32377/cvrjst0207

Dual-threshold voltage 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 paths. Therefore, no additional leakage control transistors are required, and both high performance and low power can be achieved simultaneously. Dual-threshold CMOS (DTCMOS) has the same critical delay as the single low CMOS circuit, but the transistors in non-critical paths can be assigned high to reduce leakage power. Dual threshold technique is good for leakage power reduction during both standby and active modes without delay and area overhead [46].

Dynamic threshold CMOS: For dynamic threshold CMOS (DTMOS), 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 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. Fig. 9 shows the schematic of a DTMOS inverter. DTMOS can be developed in bulk technologies by using triple wells to reduce the parasitic components. Stronger advantages of DTMOS can be seen in partially depleted SOI devices. The supply voltage of DTMOS is limited by the diode built-in potential in bulk silicon technology. The p-n 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.

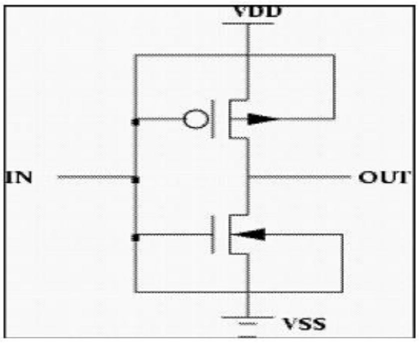

Figure 9: Schematic of a DTMOS inverter

Double-gate dynamic threshold SOI CMOS: The double-gate dynamic threshold voltage (DGDT) SOI MOSFET combines the advantages of DTMOS and double-gate FD SOI MOSFETs without any limitation on the supply voltage. Figure 10 shows the structure of a DGDT SOI MOSFET. A 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 sub-threshold characteristics. Compared with symmetric double-gate SOI CMOS, the power delay product of DGDT SOI CMOS is smaller [45].

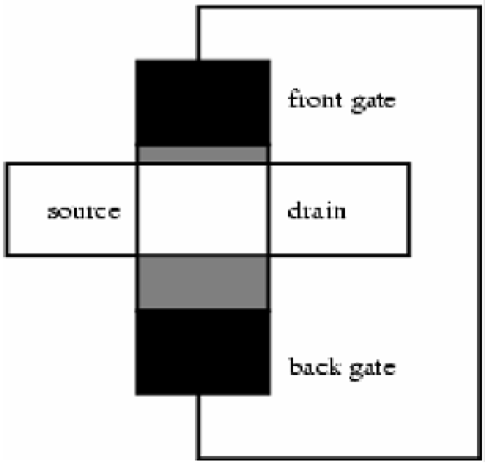

Figure 10: DGDT SOI MOSFET structure

C. Dynamic Vth techniques

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 (DVTS) have been proposed [46, 47].

V_{th} -hopping scheme: Figure 11shows the schematic diagram of the Vth-hopping Scheme Using the control signal (CONT), which is obtained from software, the power control block generates select signals, V_{th} –low-Enable and V_{th} –high-Enable, which in turn control the substrate bias for the circuit. When the controller asserts V_{th} –low-Enable, V_{th} in the target processor reduces to V_{th} –low. On the other hand, when the controller asserts V_{th} -high-Enable, the target processor Vth becomes V_{th} –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} –low-Enable, the frequency controller generates fCLK, and when the controller asserts V_{th} –high-Enable, the frequency controller generates, for example, fCLK/2.

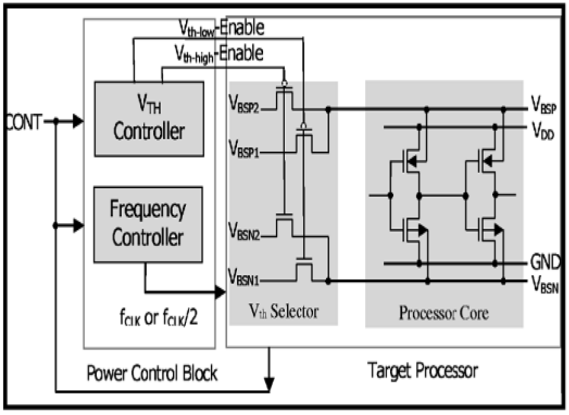

Figure11: Schematic diagram of Vth-hopping

Dynamic V_{th} -scaling scheme: A block diagram of the DVTS scheme and its feedback loop is presented in Figure 12. A clock speed scheduler, 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

DOI:10.32377/cvrjst0207

frequency of the voltage-controlled oscillator 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 [48,49].

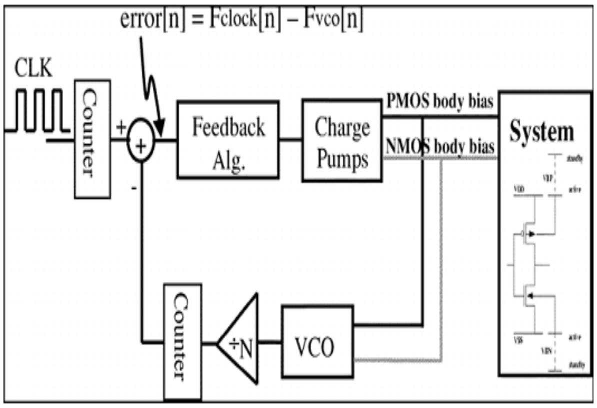

Figure 12: Schematic of DVTS hardware

D. Supply voltage scaling techniques

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 sub-threshold leakage due to DIBL decreases as the supply voltage is scaled down. For a 1.2-V 0.13µm technology, it is shown that the supply voltage scaling has significant impacts on sub-threshold leakage and gate leakage (reductions in the orders of V3 and V4, respectively). To achieve low-power benefits without compromising performance, two ways of lowering supply voltage can be employed: Static supply scaling and dynamic supply scaling. In static supply scaling, multiple supply voltages are used as shown in Figure13.

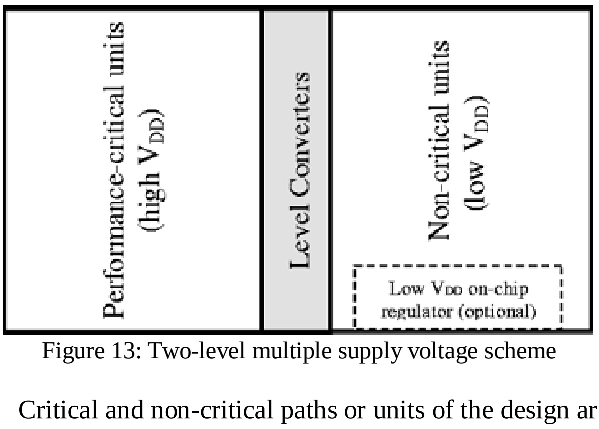
Performance-critical units (high V_{DD})
Level Converters
Non-critical units (low V_{DD})
Low V_{DD} on-chip regulator (optional)

Figure 13: Two-level multiple supply voltage scheme

Critical and non-critical paths or units of the design are clustered and powered by higher and lower supply voltages, respectively. Since the speed requirements of the non-critical units are lower than the critical ones, supply voltage of non-critical units can be lowered without degrading system performance. Whenever an output from a low Vdd unit has to drive an input of a high Vdd 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 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. Figure 14 shows DVS system architecture. 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 the critical path. All chips operate at the same clock frequency and same supply voltage, which are generated from the ring oscillator and the regulator.

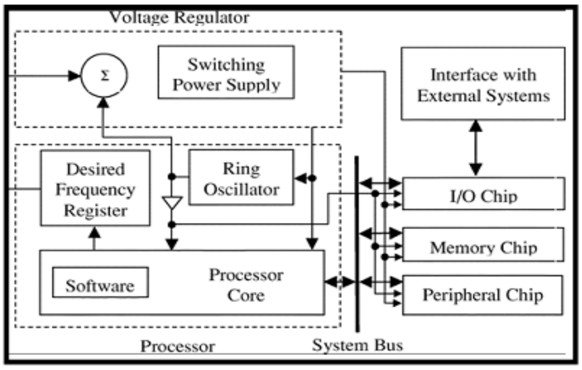

Figure14: DVS architecture

E. Input vector control techniques

The basic idea behind input vector control techniques for leakage reduction is to force the combinational logic of the circuit into a low-leakage state during standby periods. This low leakage state forces the largest number possible of transistors to be turned OFF so as to reduce leakage and make use of multiple OFF-transistors in stacks. Leakage current can be reduced to 10 times lower if input vector control is used. Inserting the low-leakage input vector during standby phases is done by either adding multiplexers or static latches at the inputs of the circuit. Due to the added circuitry (multiplexers/latches) switching, which consumes power, this technique is useful for circuits with standby periods greater than some predetermined threshold value so that the amount of saved power outweighs the power used by the added circuits.

Determining the best input vector that needs to be fed into the circuit depends on circuit structure, complexity, and size. There are three main methods for selecting the required input vector:

(1) Analyzing the circuit and looking for a good input vector.

DOI:10.32377/cvrjst0207

(2) Employing an algorithm that searches for the best input vector.

(3) Simulating the circuit with a large number of input test patterns and selecting the one that results in the lowest leakage power among these test patterns.

Another way is to apply an algorithm that uses the probabilistic theory to search a large number of random inputs looking for a “good” input vector based on a certain confidence and error tolerance. Another approach to finding the lowest leakage input vector pattern is achieved by considering the problem of finding that vector as a Boolean Satisfiability (SAT) problem. Thus algorithms developed for solving SAT problems may be used in finding the desired vector such as PBS and GRASP. The use of SAT-based algorithms to find the desired input vector was introduced in literature. All previous techniques reduced leakage power during sleep mode. Some other work has been done to reduce leakage power during circuit’s runtime. Leakage currents are reduced in buffers by changing the conventional two inverter CMOS buffer to an asymmetric buffer with three inverters in which two oppositely skewed inverters are used to drive the PMOS and the NMOS of the final inverter instead of driving both transistors simultaneously by the same signal. Up to 77% leakage power reduction was reported by selectively assigning high- V_{th} transistors.

F. Body biasing techniques

Adaptive body biasing is an effective way of reducing active leakage, as well as standby leakage through its effect in increasing the threshold voltages of MOS transistors. In this technique a high reverse body bias is applied so as to increase the threshold voltage of transistors to reduce sub-threshold leakage currents. Body biasing is also effective in reducing the negative effect of DIBL and V_{th} -Roll off to further reduction of leakage currents and improve circuit performance. One example of using the body bias technique is the Variable Threshold CMOS (VTMOS) technique, where a deep reverse body bias is applied during standby. A higher than Vdd bias for PMOS transistors and lower than GND bias for NMOS transistors, to further increase the threshold voltage and further push the transistors in the OFF region to achieve lower sub-threshold leakage currents .The VTMOS technique is shown in Figure15.

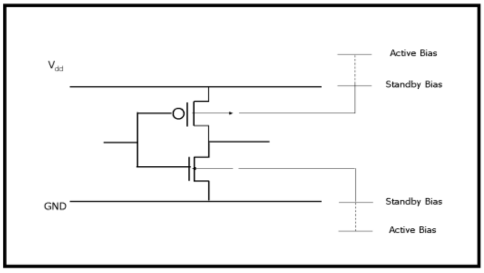

Figure15: The Variable-Threshold CMOS Body Biasing Technique

CONCLUSIONS

In this paper, we discussed different circuit techniques to reduce both dynamic and leakage power in deep submicron circuits. We describe design techniques such as transistor sizing and interconnect optimization, gated clock, multiple supply voltages and dynamic control of supply voltage. Incorporating the above approaches in the design of deep submicron circuits, the dynamic power dissipation can be reduced significantly. We also review different circuit and integrated architecture-level techniques to reduce leakage power in high-performance systems. To maintain the leakage power within bounds in logic, several techniques, e.g. dual V_{th} stacking, forward/reverse body bias and dynamic threshold voltage scaling are discussed. Different circuit techniques, e.g. gated-ground and dynamic V_{th} and dynamic Vdd, etc. are used to reduce leakage in memory.

REFERENCES

[1] S. Borkar, Design challenges of technology scaling, IEEE Micro 19 (4) (1999) 23.

[2] J. Brews, in: S.M. Sze (Ed.), High Speed Semiconductor Devices, Wiley, New York, 1990.

[3] K. Roy, S.C. Prasad, Low-Power CMOS VLSI Circuit Design, Wiley Interscience Publications, New York, 2000.

[4] Y. Taur, T.H. Ning, Fundamentals of Modern VLSI Devices, Cambridge University Press, New York, 1998.

[5] Bipul C. Paul, Amit Agarwal, Kaushik Roy ,”Low-power design techniques for scaled technologies,” *INTEGRATION, the VLSI journal 39 (2006)64–89.*

[6] S. Mukhopadhyay, K. Roy, Accurate modeling of transistor stacks to effectively reduce total standby leakage in nano-scale CMOS circuits, in: Symposium of VLSI Circuits, 2003.

[7] K. Roy, S. Mukhopadhyay, H. Mahmoodi-Meimand, Leakage current mechanisms and leakage reduction techniques in deep-submicron CMOS circuits, Proc. IEEE (2003).

[8] K. Schuegraf, C. Hu, Hole injection SiO2 breakdown model for very low voltage lifetime extrapolation, IEEE Trans. Electron Devices 41 (1994) 761.

[9] C. Choi, K. Nam, Z. Yu, R.W. Dutton, Impact of gate direct tunneling current on circuit performance: a simulation study, IEEE Trans. Electron Devices 48 (2001) 2823.

[10] K. Cao, et al., BSIM4 gate leakage model including source drain partition, in: IEDM Technical Digest, 2000, p. 815.

[11] S. Mukhopadhyay, A. Raychowdhury, K. Roy, Accurate estimation of total leakage current in scaled CMOS logic circuits based on compact current model, in: Design Automation Conference (DAC), 2003, pp. 169–174.

[12] T. Glokler, S. Bitterlich, H. Meyr, Increasing the power efficiency of application specific instruction set processors using datapath optimization, in: IEEE Workshop on Signal Processing Systems, 2000, pp. 563–570.

[13] L. Benini, G. De Micheli, A. Macii, E. Macii, M. Poncino, Reducing power consumption of dedicated processors through instruction set encoding, in: Proceedings of the Eighth Great Lakes Symposium on VLSI, 1998, pp. 8–12.

[14] K. Masselos, S. Theoharis, P. Merakos, T. Stouraitis, C.E. Goutis, Memory accesses reordering for interconnect power reduction in sum-of-products

DOI:10.32377/cvrjst0207

computations, IEEE Trans. Signal Process. 50 (2002) 2889–2899.

[15] R. Saied, C. Chakrabarti, Scheduling for minimizing the number of memory accesses in low power applications, in: Workshop on VLSI Signal Processing, 1996, pp. 169–178.

[16] J. Park, W. Jeong, H. Choo, H. Mahmoodi, Y. Wang, K. Roy, High performance and low power FIR filter design based on sharing multiplication, in: International Symposium on Low Power Electronics and Design (ISLPED'02), 2002, pp. 295–300.

[17] D. Kang, H. Choo, K. Roy, Floorplan-aware low-complexity digital filter synthesis for low-power & high-speed, in: International Conference on Computer Design, 2004, pp. 354–357.

[18] J.P. Fishburn, A.E. Dunlop, TILOS: a posynomial programming approach to transistor sizing, IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. (1985) 326–328.

[19] [19] S. Sapatnekar, V.B. Rao, P.M. Vaidya, S.M. Kang, An exact solution of the transistor sizing problem for CMOS circuits using convex optimization, IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. (1993) 1612–1634.

[20] C.P. Chen, C.C.N. Chu, D.F. Wong, Fast and exact simulations gate and wire sizing by Lagrangian relaxation, IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. 18 (1999) 1014–1025.

[21] M.T. Bohr, Interconnect scaling-the real limiter to high performance ULSI, in: International Electron Device Meeting (IEDM) Technical Digest, 1995, pp. 241–244.

[22] J.D. Meindl, Beyond Moore's law: the interconnect era, Comput. Sci. Eng. (2003) 20–24.

[23] M.L. Mui, K. Banerjee, A. Mehrotra, A global interconnect optimization scheme for nanometer scale VLSI with implications for latency, bandwidth and power dissipation, IEEE Trans. Electron Devices 51 (2004) 195–203.

[24] S. Katkoori, S. Alupoaei, RT-level interconnect optimization in DSM regime, in: IEEE Computer Society Workshop on VLSI, 2000, pp. 143–148.

[25] Y. Cao, C. Hu, X. Huang, A.B. Kahng, S. Muddu, D. Stroobandt, D. Sylvester, Effects of global interconnect optimizations on performance estimation of deep submicron design, in: International Conference on Computer Aided Design, 2000, pp. 56–61.

[26] A. Naeemi, R. Venkatesan, J.D. Meindl, System-on-a-chip global interconnect optimization, in: International SOC Conference, 2002, pp. 399–403.

[27] D. Garrett, M. Stan, A. Dean, Challenges in clock gating for a low power ASIC methodology, in: International Symposium on Low Power Electronics and Design, 1999, pp. 176–181.

[28] J. Oh, M. Pedram, Gated clock routing for low-power microprocessor design, IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. 20 (2001) 715–722.

[29] N. Raghavan, V. Akella, S. Bakshi, Automatic insertion of gated clocks at register transfer level, in: International Conference on VLSI Design, 1999, pp. 48–54.

[30] L. Benini, G.D. Micheli, Automatic synthesis of low power gated clock finite state machines, IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. 15 (1996) 630–643. 86 B.C. Paul et al. / INTEGRATION, the VLSI journal 39 (2006) 64–89

[32] T. Fuse, A. Kameyama, M. Ohta, K. Ohuchi, A 0.5V power-supply scheme for low power LSIs using multi-Vt SOI CMOS technology, in: Digest of Technical Papers of Symposium on VLSI Circuits, 2001, pp. 219–220.

[33] L.R. Carley, A. Aggarwal, A completely on-chip voltage regulation technique for low power digital circuits, in: Proceedings of International Symposium on Low Power Electronics and Design, 1999, pp. 109–111.

[34] Y. Kanno, H. Mizuno, K. Tanaka, T. Watanabe, Level converters with high immunity to power-supply bouncing for high-speed sub-1-V LSIs, in: Digest of Technical Papers of Symposium on VLSI Circuits, 2000, pp. 202–203.

[35] V. Sundararajan, K.K. Parhi, Synthesis of low power CMOS VLSI circuits using dual supply voltages, in: Design Automation Conference, 1999, pp. 72–75.

[36] K. Usami, M. Horowitz, Cluster voltage scaling techniques for low power design, in: International Symposium on Low Power Design, 1995, pp. 3–8.

[37] C. Yeh, M. Chang, S. Chang, W. Jone, Gate level design exploiting dual supply voltages for power driven applications, in: Design Automation Conference, 1999, pp. 68–71.

[38] M.C. Johnson, K. Roy, Optimal selection of supply voltages and level conversions during data path scheduling under resource constraints, in: Proceedings of the International Conference on Computer Design (ICCD), 1996, pp. 72–77.

[39] J. Chang, M. Pedram, Energy minimization using multiple supply voltages, IEEE Trans. VLSI Syst. 5 (1997) 1–8.

[40] C. Chen, A. Srivastava, M. Sarrafzadeh, On gate level power optimization using dual supply voltages, IEEE Trans. VLSI Syst. 9 (2001) 616–629.

[41] T. Kuroda, M. Hamada, Low power CMOS digital design with dual embedded adaptive power supplies, IEEE J. Solid State Circuits 35 (2000) 652–655.

[42] Jack Horgan, "Low Power Soc Design", EDA Weekly Review May 17 - 21, 2004.

[43] Cadence, "Low Power in EncounterTM RTL Compiler", Product Version 5.2, December 2005.

[44] Cadence, "Low Power Application Note for RC 4.1 and SoCE 4.1 USR3", Version 1.0,1/14/2005.

[45] V.Kursun and E. G. Friedman,*Multi-Voltage CMOS Circuit Design.*New York: Wiley, 2006.

[46] A. Chandrakasan and B. Brodersen, editors,"Low Power CMOS Design", IEEE Press, 1998.

[47] J.K. Kao and A. Chandrakasan,"Dual-Threshold Voltage Techniques for Low-Power Digital Circuits",IEEE Journal of Solid State Circuits, Vol. 35, No. 7,pp. 1009-1018, July 2000.

[48] Liqiong Wei, Zhanping Chen, Roy, K., Yibin Ye, De, V., "Mixed-Vth (MVT) CMOS Circuit Design Methodology for Low Power Applications" *Design Automation Conference, 1999. Proceedings. 36th*, Jun. 1999, pp. 430-435.

[49] M. Anis, S. Areibi, and M. Elmasry, "Design and Optimization of Multithreshold CMOS(MTCMOS) Circuits," *IEEE Transaction on Computer-Aided Design of Integrated Circuits and Systems*,vol. 22, no. 10, pp. 1324-1342, Oct. 2003.

[50] S. Sirichotiyakul and et al., "Stand-by Power Minimization through Simultaneous Threshold Voltage Selection and Circuit Sizing," *Proc. of the DAC*, pp. 436-441, 1999.

DOI:10.32377/cvrjst0207