Low Power VLSI CMOS design by DCG Technique

Dr C. M. Jadhao

Abstract-**The demand for power-sensitive design has grown significantly in recent years due to tremendous growth in portable applications. Consequently, the need for power efficient design techniques has grown considerably. Several efficient design techniques have been proposed to reduce both dynamic as well as static power in state-of-the-art VLSI circuit applications. With the scaling of technology and the need for higher performance and more functionality, power dissipation is becoming a major bottleneck for microprocessor designs. Clock power is significant in high-performance processors. Deterministic Clock Gating (DCG) technique effectively reduces the clock power. DCG is based on the key observation that for many of the pipelined stages of a modern processor, the circuit block usage in the near future is known a few cycles ahead of time. DCG exploits this advance knowledge to clock-gate the unused blocks. Because individual circuit usage varies within and across applications, not all the circuits are used all the time, giving rise to power reduction opportunity. By ANDing the clock with a gate-control signal, clock-gating essentially disables the clock to a circuit whenever the circuit is not used, avoiding power dissipation due to unnecessary charging and discharging of the unused circuits. Results show that DCG is very effective in reducing clock power. Power consumption is reduced by using this method. As high-performance processor pipelines get deeper and power becomes a more critical factor, DCG's effectiveness and simplicity will continue to be important.**

***IndexTerms*—VLSI, CMOS, DCG, Power reduction, Clock gating.**

I. INTRODUCTION

In recent years, the demand for power-sensitive designs has grown significantly. This tremendous demand has mainly been due to the fast growth of battery-operated portable applications such as notebook and laptop computers, personal digital assistants, cellular phones, and other portable communication devices. Semiconductor devices are aggressively scaled each technology generation to achieve high-performance and high integration density. Due to increased density of transistors in a die and higher frequencies of operation, the power consumption in a die is increasing every technology generation. Supply voltage is scaled to maintain the power consumption within limit.

However, scaling of supply voltage is limited by the high-performance requirement. Hence, the scaling of supply voltage only may not be sufficient to maintain the power density within limit, which is required for power-sensitive applications. Circuit technique and system-level techniques are also required along with supply voltage scaling to achieve low-power designs.

Dr C. M. Jadhao, Principal, Mauli College of Engineering, Shegaon Shegaon-444203, India , cmjadhao@gmail.com

In the nano-meter regime, a significant portion of the total power consumption in high performance digital circuits is due to leakage currents. Because high-performance systems are constrained to a predefined power budget, the leakage power reduces the available power, impacting performance. It also contributes to the power consumption during standby operation, reducing battery life. Hence, techniques are necessary to reduce leakage power while maintaining the high performance. Moreover, as different components of leakage are becoming important with technology scaling, each leakage reduction technique needs reevaluation in scaled technologies where sub-threshold conduction is not the only leakage mechanism. New low-power circuit techniques are required to reduce total leakage in high-performance nano-scale circuits. A spectrum of circuit techniques including transistor sizing, clock gating, multiple and dynamic supply voltage are there to reduce the dynamic power. For low-leakage design, different circuit techniques including, dual V_{th}, forward/reverse bias, dynamically varying the V_{th}during run time, sleep transistor, natural stacking are there.

II. POWER DISSIPATION IN VLSI CIRCUITS

The total power dissipation in a circuit conventionally consists of two components, namely, the static and dynamic power dissipation.

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

$$P_D = \alpha \, CL \, V_{DD}^2 \, f \qquad (1)$$

Where α is the switching activity,
f is the operation frequency,
CL is the load capacitance,
V_{DD} is the supply voltage.

The short circuit power of an unloaded inverter can be approximately given by

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

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

B.Leakage power

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

(1) Sub-threshold leakage, which is the leakage current from drain to source (I_{sub}).

(2) 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.

(3) The source/substrate and drain/substrate reverse-biased p-n junction leakage.

III. 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. Effective circuit techniques to reduce the dynamic power consumption 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 nano-scale circuits, the dynamic power dissipation can be reduced significantly. Other techniques such as instruction set optimization, memory access reduction and low complexity algorithms are also there to reduce the dynamic power dissipation in both logics and memories.

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. Sizing techniques can be mainly divided into two types.

- Path-based optimization.
- 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.

In global optimization, all gates in a circuit are globally optimized for a given delay.

B. Clock gating

Clock gating is an effective way of reducing the dynamic power dissipation in digital circuits. In a typical synchronous circuit such as the general purpose microprocessor, only a portion of the circuit is active at any given time.

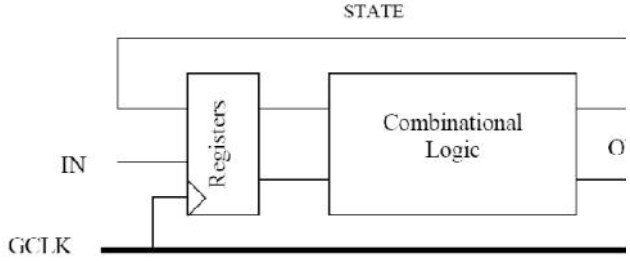

Figure1. Single clock, flip-flopbased FSM.

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. The input to the combinational logic comes through the registers, which are usually composed of sequential elements, such as D flip-flops (Fig.1). A gated clock design can be obtained by modifying the clocking structure shown in Fig.1. A control signal (f_a) is used to selectively stop the local clock (LCLK) when the combinational block is not used. The local clock is blocked when $f_a$is high. The latch shown in Fig.2 is necessary to prevent any glitches in $f_a$from propagating to the AND gate when the global clock (GCLK) is high. The circuit operates as follows.

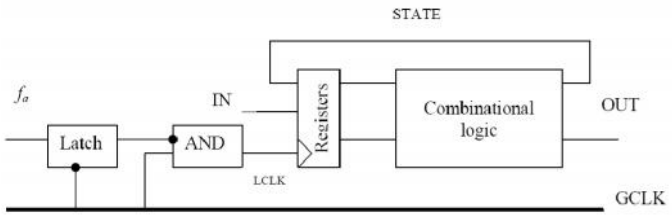

Figure2. Schematic diagram of gated clock design

The signalf_a is only valid before the rising edge of the global clock. When the global clock is low, the latch is transparent, however, $f_a$does not affect the AND gate. If $f_a$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.

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. 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. In Static supply voltage scaling schemes, higher supply voltage is used in the critical paths of the circuit, while lower supply voltages are used in the off critical paths. In Dynamic supply voltage scaling schemes, 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 the required performance with substantial power reduction.

IV. LEAKAGE POWER REDUCTION TECHNIQUES

The techniques to reduce leakage energy utilizing the slack without impacting performance can be categorized based on when and how they utilize the available timing slack e.g. dual V_{th}statically assigns high $V_{th}$to some transistors in non-critical paths at the design time so as to reduce leakage current. The techniques, which utilize the slack in run time, can be divided into two groups depending on whether they reduce standby leakage or active leakage. 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.

A. Design time techniques

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.

B. Dual threshold CMOS logic:

In this logic, a high V_{th}can be assigned to some transistors in the non-critical paths so as to reduce sub-threshold leakage current, while the performance is not sacrificed by using low V_{th}transistors in the critical path(s). No additional circuitry is required, and both high performance and low leakage can be achieved simultaneously.

Different Dual threshold CMOS techniques are

- Changing doping profile.
- Higher oxide thickness
- Large channel length

C. Run time techniques

Standby leakage reduction techniques place certain sections of the circuitry in standby mode (low leakage mode) when they are not required.

Different Standby leakage reduction techniques are

- Natural transistor stacks.
- Sleep transistor (forced stacking).
- Forward/reverse body biasing.

Active leakage reduction techniques are intermittently slows down the faster circuitry and reduces the leakage power consumption as well as the dynamic power consumption when maximum performance is not required. Dynamic V_{th}scaling (DVTS) scheme uses body biasing to adaptively change V_{th}based on the performance demand. The lowest $V_{th}$is delivered, if the highest performance is required. When performance demand is low, clock frequency is lowered and $V_{th}$is raised 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.

D. Cache memories

Circuit techniques to reduce leakage in cache memories are

- Source biasing scheme.
- Forward/reverse body biasing scheme.
- Dynamic VDD scheme
- Leakage biased scheme
- Negative word line scheme

V. Low Power CMOS Logic Design

VLSI designers have different options to reduce the power dissipation in the various design stages. For example, supply voltage may be reduced through fabrication technology, circuit design or dynamically through the system level. Switched load capacitance may be reduced through technology scaling, efficient layout.Circuit design, gate level optimization, and system level. Over the last decade, researchers have developed many techniques to reduce power dissipation in CMOS circuits. The gain obtained from each of these techniques depends solely on the application. Some of this technique may degrade the performance or increase the area to reduce power. Other techniques may degrade the performance or increase the area to reduce power. Many modifications may be applied to the process technology in order to reduce power dissipation. These modifications include reducing the threshold voltage, reducing minimum gate length, and increasing the number of metal layers. Power dissipation may also reduced by using alternative fabrication technology other than the CMOS process. This section reviews low-power CMOS technologies and present some of the alternative fabrication technologies for low – power design.

A. Threshold Voltage Reduction

Until recently, the V_{th} in most CMOS processes has been set to fairly high potential: 0.7V to 1.0V. For 5V circuit operation, this has little impact on circuit delay, which is inversely proportional to $(V_{dd}-V_{th})^2$ andthe main benefit of such a large threshold is that the sub threshold leakage is reduced exponentially. While the total leakage current of a chip is still well below the average supply current under operation, the reduced sub threshold current prolongs the duration the stored charge in dynamic circuits, providing more robust operation due to longer leakage times. Thus, there have been fewer tendencies to reduce the thresholds until recently, with the decrease of supply voltages to 3.3V, and emphasis on low-power design. The reduction of V_{th} enables VLSI designers to lower the supply voltage. This maintains circuits speed and results in a power reduction. However, the limitation of this technique is that at low thresholds, the sub threshold currents become significant, if not dominant, portion of the average current drawn from the supply. Previous work has shown the optimal V_{th} to range from 0.3V down to below 0.1V depending on the conditions of the circuit operation.

B.Technology Scaling

With every new process generation, the entire lateral and some of the vertical dimensions of the transistor are scaled down. This has an immediate impact on reducing power dissipation, as well as increasing circuit speed. The primary effect of process scaling is the reduction of all the capacitances, which provides a proportional decrease in power and circuit delays. Device sizes may be reduced to keep the delay constant over process scaling, which yields an even larger power reduction. Both gate capacitance and interconnect capacitance may be expressed as $C= W.L (1/t_{ox})$. The width W, length L, and oxide thickness t_{ox} all scale almost equally by a factor s, so that the total capacitance scales down by the same factor s. Diffusion capacitance is a more complex function of process scaling; however it is reduced by factor between s and $s^{3/2}$ for a constant supply voltage, both the power and circuit delays scale down approximately by the factor s. Thus, power reduction is accomplished with no alterations in the circuit design. As mentioned earlier, not all the vertical dimensions scale down. In particular, the thickness of the interconnect metal is roughly the same across the processes, due to fundamental processing requirements. This increases the fringing capacitance from the side of the metal to the substrate, and increases the capacitance between adjacent interconnect segments. With these secondary effects considered, the

overall capacitance scaling is somewhere below the factor s, and is difficult to accurately characterize without using a three-dimensional simulation model. New technologies from IBM and Motorola use copper instead of aluminum for interconnect, because copper has better conductivity and scales down better than aluminum.

C. Increasing Number of Metal Layers

Further power reduction may be achieved by using some features in today's more advanced processes; namely, an increased number of metal layers and a trend towards allowing stacked vias. If these are used wisely, not only can the power be reduced, but also the circuit area, and delay times. However, utilizing these advancements requires special circuit redesign methodologies. In old fabrication technologies with two metal layers, polysilicon has been used extensively in intercell signal routing, as second-level metal has been reserved for intercell routing to allow the CAD tools to perform global routing. In present technologies with more metal layers, the second-level, and perhaps higher metal layers, can be used for intercell routing. Since the capacitance per unit area decreases with each higher metal level, using the higher metal layers helps reduce the interconnect capacitance, which already contributes around 30% to the overall capacitance. That percentage is expected to increase with future VLSI generations. Also, by allowing stacked vias, the areas of the different gates can be compressed. Moreover, this reduces both the intercell and global routing because the terminal connections will be closer. Consequently, most interconnect routes will be reduced in length. However, condensed routing increases the coupling capacitance between interconnects, and cancels part of the power saving previously achieved.

D. Alternative Technologies

If the current rate of scaling MOSFET's were to continue, devices with lengths of 1nm would be in use in the year 2040. A nano-meter of oxide consists of only a few layers of atoms which approaches fundamental limits. Therefore, some new device structure will eventually replace the devices being used today. Predictions of when this will happen have consistently underestimated the ingenuity of fabrication engineers, and the conclusion is that the end of CMOS scaling is still too far away to accurately predict. However, the limits which are driving current device scaling can give some insight into what new technology might eventually replace today's devices. The following technologies seek to address the limitations of CMOS MOSFET's by allowing further reductions in the supply voltage and therefore further scaling of the device dimension, by providing improved performance at same device dimensions, or both.

1) Silicon on Insulator (SOI)

The elimination of junction capacitance gives SOI improved performance at the same device dimensions, and improved sub threshold slope allows for further device Scaling. The floating body of SOI devices is a concern for reliability and circuit simulation. It also causes these devices to have low breakdown voltages which have been a major roadblock to their use in the past, but as supply voltages continue to scale down this may present less of a problem.

2) Multi Threshold Voltage (MTCMOS) Devices

By using low threshold devices for circuits on the critical path and high threshold device elsewhere, it may be possible to achieve high performance, while maintaining reasonable leakage currents. However, very low threshold devices may not be suitable for the dynamic circuits, which are usually used in high speed applications.

3) Low Temperature CMOS (LTCMOS)

Reducing the Chip's operating temperature can enhance the performance due to improved carrier mobility and reduced wire resistance. It also reduces leakage current which increases exponentially with temperature. This would allow technologies to be scaled to smaller dimension. The disadvantage is the increased size and cost of the system.

4) Dynamic Substrate Biasing

Designing an on-chip circuit which dynamically varies the substrate voltage in order to compensate for threshold variations helps reduce the leakage current. Dealing with temperature and process variations across a single die would require several of these circuits to control isolated substrate regions. In such chips substrate noise would be a major concern.

5) New Gate Oxide Material

Replacing SiO_2 with a higher permittivity material would allow thicker gate oxide films to provide the same control of the channel. Fields the oxide layer would be reduced, making breakdown less of a concern, and tunneling currents would be reduced. The difficulties with this approach are depositing a very thin very high quality oxide layer rather than simply growing a thermal oxide. The Si- SiO_2 interface has been the subject of intensive study for decades and moving to a different material would be a very significant change in the fabrication process

VI. CONCLUSION

Deterministic clock-gating (DCG) methodology is based on the key observation that for many of the stages in a modern pipeline, a circuit block's usage in a specific cycle in the near future is deterministically known a few cycles ahead of time. Using this advance information, DCG clock-gates unused execution units, pipeline latches. Results show that DCG is very effective in reducing clock power. Power consumption is reduced by using this method. As high-performance processor pipelines get deeper and power becomes a more critical factor, DCG's effectiveness and simplicity will continue to be important. Effective clock-gating, however, requires a methodology that determines which circuits are gated, when, and for how long. Care to be taken while designing the clock-gating control circuitry; otherwise the circuitry may become an overhead. This overhead may result in power dissipation to be higher than that without clock-gating.

VII. REFERENCES

[1]. Bipul C. Paul, AmitAgarwal, Kaushik Roy, "Low-power design techniques for scaled technologies", INTEGRATION, the VLSI journal, science direct 39(2006).

[2]. L. Hai, S. Bhunia, Y. Chen, K. Roy, T.N. Vijay Kumar, " DCG : deterministic clock gating for low-power microprocessor design ", IEEE Trans. VLSI Syst. 12 (2004), pp.245-254.

[3]. Yan Luo , Jia Yu , Jun Yang , LaxmiBhuyan, Low power network processor design using clock gating, Proceedings of

the 42nd annual conference on Design automation, June 13-17, 2005, San Diego, California, USA.

[4]. Hai Li, SwarupBhunia, Yiran Chen, T. N. Vijaykumar, and Kaushik Roy, "Deterministic Clock Gating for Microprocessor Power Reduction",1285 EE Building, ECE Department, Purdue University.

[5]. Enoch O. Hwang, "Microprocessor Design Principles and Practices", Brooks / Cole, 2004.

[6]. Hai Li , Chen-Yong Cher , Kaushik Roy , T. N. Vijaykumar, Combined circuit and architectural level variable supply-voltage scaling for low power, IEEE Transactions on Very Large Scale Integration (VLSI) Systems, v.13 n.5, p.564-576, May 2005.

[7]. D. Folegnani and A. Gonzalez, "Energy-effective issue logic", in Proc.28th Int. Symp. Computer Architecture (ISCA), July 2001, pp. 230 - 239.

[8]. 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.

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

[10]. 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.

[11]. 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.

[12]. R. I. Bahar and S. Manne, "Power and energy reduction via pipeline balancing," in Proc. 28th Int. Symp. Computer Architecture (ISCA), July 2001, pp. 218–229.

[13]. D. Brooks, V. Tiwari, and M. Martonosi, "Wattch: a framework for architectural-level power analysis and optimizations," in Proc. 27th Int.Symp. Computer Architecture (ISCA), July 2000, pp. 83–94.

[14]. S. Palacharla, N. P. Jouppi, and J. E. Smith, "Complexity-effective superscalar processors," in Proc. 24th Annu. Int. Symp. Computer Architecture (ISCA), June 1997, pp. 206–218.

[15]. S. Manne, A. Klauser, and D. Grunwald, "Pipeline gating: speculation control for energy reduction," in Proc. 25th Int. Symp. Computer Architecture (ISCA), June 1998, pp. 132–141.

[16]. D. Folegnani and A. Gonzalez, "Energy-effective issue logic," in Proc. 28th Int. Symp. Computer Architecture (ISCA), July 2001, pp. 230–239.

[17]. D. Brooks and M. Martonosi, "Value-based clock gating and operation packing: dynamic strategies for improving processor power and performance," ACM Trans. Comput. Syst., vol. 18, no. 2, pp. 89–126, May 2000.

[18]. J. C. Monteiro, "Power optimization using dynamic power management," in Proc. XII Symp. Integrated Circuits Systems Design (ICSD), Sept. 1999, pp. 134–139.

[19]. MassoudPedram "Power Minimization in IC Design: Principles and Applications", Department of EE-Systems, University of Southern California, Los Angeles CA.

[20]. William M. Johnson "Super-Scalar Processor Design", Computer Systems Laboratory, Stanford University Stanford, CA, June 1989.

[21]. Ricardo E. Gonzalez "LOW-POWER PROCESSOR DESIGN", Computer Systems Laboratory, Stanford University Stanford, CA, June 1997.

[22]. Qing WU "Clock-Gating and Its Application to Low Power Design of Sequential Circuits", Department of Electrical Engineering-Systems, University of Southern California, Los Angeles, CA, USA.

Dr C M Jadhao, Principal, Mauli Group of Institutions College of Engineering and Technology, Shegaon. 24 years of Teaching, research and administrative experience, Specializations in Fiber optics, VLSI design and Microprocessors.