Analysis of Low Power Submicron Circuits

Prajoona Valsalan
Research Scholar
Karpagam University
Coimbatore

K Sankaranarayanan, PhD.
Dean
EASA College of Engineering and Technology,Coimbatore

ABSTRACT
With the advent of the integrated circuits, greater emphasis was given on performance and miniaturization. But with the increasing prominence of portable and battery operated appliances the key factor that requires attention is power consumption. The feature size is shrinking due to the advancement in fabrication technology which causes integration of more transistors in the integrated circuit. As a result, the magnitude of power per unit area is growing and the accompanying problem of heat removal and cooling is worsening. To maintain the chip temperature at an acceptable level the dissipated heat must be removed effectively, the cost of heat removal and cooling becomes a significant factor in these circuits. The reliability of the chip will be greatly degraded with high power dissipation due to silicon failure mechanism such as electro migration. The linear scaling of supply voltage with the features size was started from half-micron technology. But the power supply scaling affects the speed of the circuit. The need of the time is to put efforts in designing low power and high speed circuits. In this paper, we investigate and analyse the causes and solutions of power dissipation and delay in different topologies.

General Terms
Low power VLSI Design

Keywords
VLSI, Low Power Techniques, Power Estimation.

1. INTRODUCTION
In VLSI circuit design, two major concerns in optimization have been delay and area. Research has been done on the circuit level to examine the trade-offs between them. In particular, transistor sizing has been well established as a good way to achieve reduction in delay of circuits, while the resultant increase in rectangular area from transistor sizing can be minimized by special layout techniques. As wireless communication and mobility of the equipment becomes increasingly desirable, power dissipation of circuits has become a major concern in the circuit synthesis. In performance driven synthesis of VLSI circuit, low-power design has joined the ranks of area and delay as major motivation in optimization. Hence in today's VLSI circuit design, there is a need to ensure low power dissipation while satisfying delay constraints.

2. BACKGROUND

2.1 Transistor Sizing
Transistor sizing is well established as an effective way to speed up circuits. Numerous studies have been done in this area. For static CMOS, the delay of a transition can be modeled as dependent on RC, where R is the effective resistance of the transistors in the pull-up or pull-down circuitry and C is the capacitive load driven by these transistors. R is inversely proportional to the width of the transistors while C is proportional to the size of the transistors in the next stage. Hence by increasing the width of the transistor at the current stage delay can be reduced [14].

In most of the literature on sizing, the length of each transistor is kept at a set value while the width is treated as a continuous variable. Linear Programming methods or other numerical simulation methods are then applied to find the optimal size for the transistor in a circuit.

2.2 Power Estimation
As power dissipation is becoming an increasingly important issue, accurate power estimation models are needed. Recent developments of probabilistic techniques have produced a fast and efficient way of estimating power, which is proportional to the average switching probability of a node. The power dissipation of a gate is approximated by the change in energy for charging and discharging the output capacitance of the gate. Since a gate does not necessarily switch at every clock cycle, the frequency of switching is estimated by clock frequency multiplied by the expected number of switches per cycle[4,6]. Average power is given by

$$Pavg = (\tfrac{1}{2} \times Cload \times Vdd) \times (E(transitions/cycle) / Tcyc)$$

where Pavg denotes average power, Cload is capacitance,Vdd is supply voltage, Tcyc global clock period. E(transitions) – Expected value of the number of gate output transitions per global clock cycle.
A general delay model is used to that the Boolean conditions that cause glitches in circuits can correctly computed to be included in calculations of switching probabilities. This statistical method of power estimation provides a simple way of examining power dissipation in terms of sizing, since gate capacitance is proportional to transistor width.

2.3 Input Ordering
Delay through gates with multiple inputs is dependent on the arrival times of the input transition. The time between the latest switching of the inputs to the switching of the output is minimal if the input that switches last is closest to the output node, due to body effects.The inputs to a gate can be ordered such that the latest arriving input is placed at the fastest pin of a gate, so as to achieve a speedup of up to 60% decrease in delay.Input ordering does not cause any increase in the power dissipation due to capacitive load increase, and may actually decrease power due to shorter delay times and hence possible reduction in glitches. It provides yet another means of delay and power optimization.

3. NEED FOR OPTIMIZATION
Previous work has to size individual transistors in custom designed circuits. However, commercially, product to market times must be small, and hence much circuit design is done with standard cells as the target technology. Since standard cell libraries are widely used, it is feasible to have a library of

gates with transistors that are previously sized to give good power and delay trade-offs, and then the problem of optimization is to choose the best version of each cell to use. This provides a method with larger granularity, and because of the early binding of the transistor widths, a computationally simple method of optimization.

In many available standard cell libraries, cell transistors are not minimum-sized, and only one version of each gate is available. Examination of the power and delay curves of various gates show that delay can be reduced significantly with sizing without much increase in power. Hence, a library where several sized versions of a gat are available, where each gate is sized with area minimization an input ordering in mind such that they give good delay response without high power dissipation, would be very useful when designing of power and delay performance. Since the aim is to have low power dissipation, circuits are first mapped with minimum-sized gates and then changed to larger gates as necessary to satisfy delay constraints [15]. In those ways, we start with minimum power, and gates that are nit dominant in determining the delay of the circuit remain minimum-sized, thereby ensuring low-power dissipation. Having mapped the circuit, the switching probability of each node can be calculated, and this information, as well as the delay parameters of each gate, can be utilized in optimization routines that select the version that is best suited for each node, such that a given delay constraint is satisfied with minimum power.

4. CIRCUIT LEVEL LOW POWER TECHNIQUES: CHALLENGES AND SOLUTIONS

4.1 Power Dissipation

Along with the rapidly evolving silicon technology, the transistor size keeps decreasing dramatically. While transistors are getting smaller and faster, low power issue of VLSI circuits built with these transistors is getting more serious. The average power of a digital CMOS device can be conceptually modeled as

$$Pavg = Pstatic + Pdynamic$$

The dominant component of Pavg so far has been the dynamic power which is composed of the switching power Pswitching (charging and discharging of the load capacitance) and the short circuit power Pshort (when both P and N transistors are turned on during the switching). In the normal operation, switching power dominates the dynamic power and Pswitching = kCLVdd2fclk, where k is the switching activity factor, CL is the load capacitance, Vdd is the supply voltage and fclk is the clock frequency.

It might appear that the switching power will decrease when transistor size decreases since the load capacitance is proportional to transistor size. However, the die area does not shrink with the technology. It actually means a 2X increase in number of transistors packed in a chip per generation [4,6]. At the same time, power dissipation of a single CMOS gate does not decrease fast enough to compensate for such an increase. The increasing power density (power dissipation per unit area) and related heat removal have been getting increasingly problematic.

4.1.1 Total capacitance

The minimum feature size (MFS) scales down by 30% every generation. If we consider a constant die area (with 2X more transistors), then the total transistor capacitance actually increases by 40% (for each transistor, 0.5X gate area, 0.7X gate oxide thickness). On the other hand, silicon technology is getting increasingly interconnect dominant. While gate capacitance for individual transistor decreases with the minimum feature size, interconnect capacitance per unit length decreases slower due to sidewall contribution. Then, there are more global/local interconnect layers due to the higher integration and complexity of a chip for every generation. Therefore, total capacitance of a chip keeps increasing.

4.1.2 Clock speed

The second fact contributing to the increasing of the power (density) is that clock frequency is scaled faster than the technology. The clock frequency has doubled with every technology generation in the past. This is a much faster scaling than the technology, which should have been just 43% for constant power density. The need for performance drives such scaling and makes the power density increase for each generation.

4.1.3 Leakage power

Third, as technology enters deep-submicron era, leakage current becomes more serious. Since the supply voltage decreases every generation, threshold voltage needs to decrease accordingly to avoid increased delay. However, the leakage current increases exponentially when threshold voltage decreases. The dramatically increased leakage current in the off state leads to the increase of static power Pstatic.[13] The leakage current consists of both source-to-drain leakage due to sub-threshold conduction and drain-to-gate leakage due to electron tunneling across the ultra thin gate oxide. The latter is not yet a significant portion of leakage, but it is getting more noticeable and could dominate as technology scales further. The increase of leakage current has a direct impact on the standby power consumption of a chip.

4.1.4 Architecture trend

Another important factor that contributes to the increasing power density is that the hardware architecture trend is toward more flexible (programmable) and reusable cores. Comparing to application specific architecture, it is much less energy efficient. Programmability is a common requirement for designing a large scale SoC (system-on chip). It ensures function flexibility, and helps in post-fabrication bug fixing and tuning.

However, flexibility and programmability impose an energy burden for the chip. The power performance ratio required by a processor to carry out a given task can be several orders of magnitude higher than that achieved by an application specific architecture.

4.1.5 Challenges and solution

It is obvious that the power issue has become an increasingly important factor in VLSI design. As the wireless communication, PDAs, mobile computing, sensor networks, etc. become popular; mobile devices supplied by battery will be widely employed. Increases of switching power and leakage power consumption have direct impact on the battery life. For some of those devices (e.g., sensors in a sensor network), it is even difficult to frequently replace the battery. Higher power consumption impairs their lifetime and application. For other devices not using battery as power supply, the increase of power density adds difficulties to the power supply distribution and thermal management to remove the massive amount of heat generated by a chip.

Numerous low power techniques have been developed since last decade. They can be classified into device, circuit, logical and architecture levels. For example, transistor sizing is a

device level technique that optimizes the size of transistors in a circuit. Different circuit design styles like dynamic logic, pass transistor logic, etc., are circuit level techniques. Power optimized synthesis of logic structure is a logic level technique. Instruction set optimization to reduce the switching ratio is a good example of architecture level technique. Logic/architecture level low power techniques have a significant influence on the total power consumption of a system. However, device/circuit level techniques are more fundamental and can always be applied with any logic/architecture level technique. In this chapter, we concentrate on device/circuit level techniques and give a survey of those techniques. For each technique, we introduce the idea and discuss its effectiveness in power reduction.

4.2 Low Power Techniques

4.2.1 Architecture/logic level

One approach to power reduction at the architecture level is to build a power manageable architecture so that one can eliminate idle power consumption (power consumed when the hardware is not in use), and run time slack by controlling the clock activity, voltage, frequency, and even the device threshold voltage. One way to manage power at the architecture level is to use multiple voltages and clock frequencies. In multiple-voltage circuits, two or more supply voltages are distributed on chip according to the criticality of the path. Time-critical paths are supplied by a higher voltage and a lower supply voltage is used to reduce the power for non-critical paths. In variable-voltage circuits, supply voltages are modulated during the system operation. It is a very powerful technique because it can trade off power for speed at run time to fine tune performance and power according to the workload. In practice, however, it requires smart design techniques because voltage change requires non-negligible time and clock speed must be varied accordingly when supply voltage changes.

Clock gating is another common power management technique that allows turning off clock for idle modules in a circuit. Power savings are achieved in the registers (whose clock is halted) and in the combinational logic gates where signals do not propagate due to the freezing of data in registers. Clock gating is widely used because it is conceptually simple, has a small overhead in terms of additional circuitry and often has a zero performance overhead because the component can transit from an idle to an active state in one (or few) cycles. The main design challenges in the implementation of clock gating are: (i) to construct an idleness-detecting circuit which is small and accurate and (ii) to design gated-clock distribution circuitry that introduces minimum routing overhead and keeps clock skew under a tight control.

Leakage is a major concern in idle-power consumption. Most leakage-reduction techniques, e.g., dual-Vt, Variable Threshold CMOS (VTCMOS) and power supply gating, etc., can be considered as architectural level power management techniques. For the dual-Vt technique, the basic idea is to use low threshold transistor (fast and leaky) on time-critical paths and high threshold transistor (slow and less leaky) on non-critical paths. The dual-Vt technique tends to lose its effect when more paths become critical. VTCMOS allows dynamic control of threshold voltage via substrate biasing. It has a better leakage reduction effect than the dual-Vt but requires standby control circuit to detect the idleness of a module and then apply the biasing. Finally, the ultimate solution to avoid leakage is to shutdown the power supply during the standby time. An advantage of this approach is the wide applicability to all kind of electronic components, i.e., digital and analog units, sensors, and transducers. A major disadvantage is the wake-up recovery time, which is typically higher than in the case of clock gating because of the re-initialization of components.

Yet another approach for architecture level power reduction is the application dependent specialization, which is an ad-hoc way to specialize hardware platform for an application without compromising the reuse and design flow streamlining.

4.2.2 Device/circuit Level

We have listed the specific low power techniques in Table 2.1 and power optimization techniques for CMOS in Table 2.2. Most of these technologies/techniques are at device/circuit level. However, we have included leakage reduction techniques because they are getting increasingly important now.

Complimentary MOS (CMOS) was first proposed by Wanlass and Sah in 1963. The CMOS process is more complex than the NMOS process because it provides both n-channel and p-channel transistors on the same chip. However, CMOS circuits can achieve low power consumption by eliminating (most if not all) static power. Domino CMOS is a dynamic logic family originally suggested in, which combined the speed and power advantage of the dynamic logic circuit and the stability and ease of use of static logic (full Complementary MOS) circuit. Compared to static CMOS, domino CMOS reduces the dynamic power because it has a smaller switching capacitance (having fewer transistors), no spurious transitions (glitches) and no short circuit current as in CMOS [8,12,17]. However, domino CMOS does have some serious drawbacks that lead to additional power consumption like contention, which can consume additional power. In addition, the operation of domino CMOS requires pre-charge and evaluation phases, which means some nodes are charged and discharged unnecessarily. Overall, domino CMOS still appears to have a better time power trade-off than the static CMOS.

The difference between Pass Transistor Logic (PTL) and CMOS logic is that the source of the pass transistor network is connected to some input signal instead of the power lines and ground. Pass transistor logic is attractive because it can reduce the number of transistors in implementing XOR gate, multiplexers, registers, and other key building blocks [1]. However, the threshold voltage drop at the output requires level restoration, which means extra circuitry must be added. There are still debates about the power efficiency of PTL and CMOS. In practice, application of PTL/CMOS mixed logic has achieved considerable power reduction.

Self-timed Logic is an asynchronous design that utilizes handshake signals to synchronize the data exchange between asynchronous elements. One major advantage of self-timed logic is the elimination of the clock generator and distribution network, which could otherwise consume a significant portion of power. In addition, self-timed logic inherently powers down the unused modules and saves power consumption by them.

Table 2.1: Low power technologies: specific technologies and theirs power components. Except the CMOS technology, power reduction by each technology is compared to the static CMOS (clocked) counterpart

Low Power Techniques		Switching Power	Short-circuit Power	Leakage Power	Static Power	Power Reduction
CMOS		Yes	Yes	Yes	No	Nanowatt standby power [214]
Domino CMOS		Yes (no glitches, has wasted discharges)	Maybe (caused by the contention)	Yes	No	128% speed up with 41% more power consumption [48].
PTL	Pure PTL	Yes (reduced by a smaller capacitance)	Almost no (only exist at output inverter)	Yes	Maybe (due to V_{th} drop)	30-50% reduction by [1, 116, 156, 158, 227, 228], 44% of the power-delay product in [176]
	Mixed PTL/ CMOS	Yes (reduced)	Yes	Yes	Maybe (due to V_{th} drop)	more than 20% reduction in [41], about 50% of power delay product in [42]
Asyn. Design	Self-timed Logic	Yes (glitches eliminated)	Maybe (caused by contention)	Yes	No	25% reduction in [108]
	Others	Yes (no clock signals and clock networks)	–	Yes (reduced by a shorter signal quiescent time)	–	5 times saving with 20% area overhead in [207], up to 5 times less power in [149, 150]
Adiab. Switch. & Engergy Recov.	Fully Adiab.	No (asymptotically zero)	No	No (negligible)	No	up to 6 times less energy per addition CLA in [122]
	Partial Adiab.	Yes (asymptotically nonzero)	No	No (negligible)	No	10-20 times power gain inverter chain in [57, 127, 138]

One disadvantage of self-timed logic is that for certain logic families it may suffer from the "contention" problem as in domino CMOS [20]. With the requirement of dual-rail encoding (for completion signal), the energy consumption per transition could be high, which limits its application in a continuously active data path. Except for self-timed logic, asynchronous designs have recently drawn resurgent attention because of their low powers feature. A major advantage of an asynchronous design is that it does not require a power consuming clock network. New design technique (e.g. Angram framework) supports the plug and play composition of asynchronous components into systems, which significantly simplifies the design task for a large system. Many asynchronous designs exhibit dramatic power reduction especially for applications that have significant computation load fluctuation and large disparity between average performance and peak performance, e.g., general purpose microprocessors, error correctors, etc.

The fundamental cause of energy dissipation in a CMOS circuit is the charge transportation from Vdd to load capacitance and to GND. The principal idea of adiabatic switching is to minimize the energy dissipation during this process by slowing down the charging/discharging operation. Combining with reversible computation (no information loss during computation), one can build a "fully adiabatic" circuit, which has asymptotically zero power consumption.The limitation of fully adiabatic circuit is that the function of the circuit has to be reversible, which limits its application. "Charge recycling" or "energy recovery" are terms used more recently for describing circuit techniques that do not require reversible logic but "recycle" the information representing charges and use adiabatic switching to reduce the energy dissipation. In practice, the power saving is dramatic, sometimes as high as one order of magnitude. However, the major drawback of these techniques is that the operating frequency cannot be very high (due to the adiabatic switching principle).

As CMOS was prevailing in the last few decades, numerous low power techniques have been proposed to enhance the power performance of CMOS based circuit. Transistor/gate sizing is a technique determining the sizes of transistor/gate in a circuit. In the past, optimizations were primarily for circuit delay and area. With the growing concern for low power dissipation, new transistor/gate sizing techniques for power optimization have been proposed. Glitch (spurious transitions before a signal reaches the steady state value) reduction is another important topic in CMOS low power design. To eliminate glitches, the basic idea used is to balance the paths (path balancing) and/or filter the glitch by gate inertial delay (hazard filtering). Transistor/gate sizing can be used for the optimization.

One can also use linear programming algorithms to derive the delay assignments in the circuit and then realize these delay assignments by buffer insertion or gate level design.

Table 2.2: Low power technologies: optimization techniques for CMOS

Low Power Technology			Basic Idea	Power Reduction Effects
Dynamic Power Reduction	Transistor/Gate Sizing		Size the transistor/gate to optimize switching power and short-circuit power under the delay constraint	40% to 50% by TILOS [66], 50% to 65% reduction in [79, 223]
	Glitch-reduction Techniques	Balanced Delay	balance differential path delay	30% power reduction in [104]
		Hazard Filtering	Increase gate inertial delay to filter out glitches	27% power reduction in [4]
		Transistor Sizing	size the transitor/gate to balance the path	32% to 46% reduction in [220]
		Linear Programming	Use linear programming to derive gate delay assignments	up to 62% reduction in [170], 25% to 54% reduction in [171]
Leakage Reduction	Input-vector Control		switch input vector to low leakage pattern during standby	up to 2x leakage reduction [229]
	Body-bias Control	VTCMOS	dynamically change threshold voltage to high Vt at standby	100μA I_{leak} at active mode; 0.1μA I_{leak} at standby [115]
		DTCMOS	the floating body and the gate of a high Vt transistor are tied together	up to 5.5 time higher current drive when device is on (low Vt) [7]
		Dual-Vt	Low Vt for critical path; High Vt for non-critical paths	up to 80% leakage reduction [215]
	Power-supply Gating		Insert "sleep" transistor at pull down path or use power supply regulator to turn off the power supply during standby	virtually 100% Leakage reduction [59]

5. CONCLUSION AND FUTURE WORK

Low power dissipation has become a crucial factor in the modern VLSI circuit design. Linear programming techniques have been proposed to reduce the switching power of a circuit by removing glitches. However, the prior techniques have limitations due to the fixed-delay assumption. In this analysis, we provide new optimization methods considering the effects of process variation. It might be possible to devise a linear program using the linear gate delay model proposed by Berkerlaar. This linear program will be able to minimize the dynamic power dissipation considering both total capacitance and glitch reduction. Process-variation resistance will be one more feature that can be added. The leakage current variation has an exponential relationship with these two parameters, threshold voltage Vt and thermal voltage VT. The reduction of the variation in dynamic power has direct impact on the variation of the operating temperature. Thus, leakage power variation is suppressed when dynamic power variation is suppressed. Further research incorporating our technique into the reduction of leakage power variation might be possible.

6. REFERENCES

[1] R. Zimmermann and W. Fichtner, "Low-power logic styles: CMOS versus pass-transistor logic," IEEE J. Solid-State Circuits, vol. 32, pp. 1079–1090, July 1997.

[2] T. Nikoubin, , S. Pouri, P. Bahrebar and N. Navi, "A New Transistor Sizing Algorithm for Balanced XOR/XNOR Circuits," CSICC. Faculty of Electronical and Computer Engineering, SBU, Tehran, Iran February 20-22, 2007, [12th International CSI Computer Conference, CSICC 2007, Tehran, Iran] (in Persian).

[3] T. Callaway and E. Swartzlander, "Low power arithmetic components,"in Low Power Design Methodologies, J. Rabaey and M. Pedram,Eds. Boston, MA: Kluwer Academic, 1996.

[4] M. Nemani and F. N. Najm, "High-level area and power estimation for VLSI circuits," IEEE Transactions on Computer-Aided Design, vol. 18, no. 6, pp. 697–713, June 1999. [Online]. Available: http://www.eecg.toronto.edu/ najm/papers/tcad99-mahadev.pdf

[5] K. Yano, T. Yamanaka, T. Nishida, M. Saito, K. Shimohigashi, and A.Shimizu, "A 3.8-ns CMOS 16_16-b multiplier using complementary pass-transistor logic," IEEE J. Solid-State Circuits, vol. 25, pp. 388–395, Apr. 1990.

[6] P. Landman, "High-level power estimation," in Proc. 1996 Int. Symp. Low Power Electronics and Design, 1996, pp. 29-35.

[7] V. Tiwari, S. Malik, and A. Wolfe, "Power analysis of embedded soft- ware: A first step toward software power minimization," IEEE Trans. VLSI Syst., vol. 2, pp. 437-445, Dec. 1994.

[8] D. Liu and C. Svensson, "Power consumption estimation in CMOS VLSI chips," IEEE J. Solid State Circuits, vol. 29, no. 6, pp. 663-670, 1994.

[9] F. N. Najm, R. Burch, P. Yang, and I. Hajj, "Probabilistic simulation for reliability analysis of CMOS VLSI circuits," IEEE Trans. Computer- Aided Design, vol. 9, pp. 439-450, Apr. 1990.

[10] C.-Y. Tsui, M. Pedram, and A. M. Despain, "Efficient estimation of dynamic power dissipation under a real delay model," in Proc. IEEE Int.Conf. Computer Aided Design, Nov. 1993, pp. 224-228.

[11] A. A. Ghosh, S. Devadas, K. Keutzer, and J. White, "Estimation of average switching activity in combinational and sequential circuits," in Proc. 29th Design Automation Conf., June 1992, pp. 253-259.

[12] A. Chandrakasan and R. Brodersen, "Minimizing power consumption in digital CMOS circuits," Proc. IEEE, vol. 83, pp. 498-523, Apr. 1995.

[13] K. Roy and S. Prasad, Low-Power CMOS VLSI Circuit Design. New York: Wiley Intersci., 2000.

[14] J. P. Fibshburn and A.E. Dunlop, "TILOS: A Posynomial Programming Approach to Transistor sizing, in IEEE International Conference on Computer Aided Design, pg. 326-328, November 1985.

[15] D. Marpel, "Performance Optimization of Digital VLSI Circuits", Technical report:CSL-TR-86-308, Stanford University, 1985.

[16] S. Powell and P. Chau, "A model for estimating power dissipation in a class of DSP VLSI chips," IEEE Transactions on Circuits and Systems, vol. 36, no. 6, pp. 646–650, 1991.

[17] A. Chandrakasan and R. Brodersen, "Minimizing power consumption in digital CMOS circuits," Proc. IEEE, vol. 83, pp. 498–523, Apr. 1995.

[18] K. Roy and S. Prasad, Low-Power CMOS VLSI Circuit Design.New York: Wiley Intersci., 2008.

[19] J. Wang, S. Shieh, C. Yeh, and Y. Yeh, "Pseudo-Footless CMOS Domino Logic Circuits for High-Performance VLSI Designs," in Proc. Int. Symp. on Circuits and Systems, vol. 2, pp. 401-404, 2004

[20] S.Govindarajulu, Dr.T.Jayachandra Prasad "Energy efficient Reduced Swing Domino Logic Circuits in 65 nm Technology" International Journal of Engineering Science and Technology, Vol.2 (6), 2010, pp.2248-2257, ISSN:0975-5462.