Power Optimization Within Nanometer Designs

Dr. Danny Rittman
danny@tayden.com

Abstract

As the integrated circuits (ICs) are scaled into deep nanometer dimensions and operate in giga-hertz frequencies, power optimization is becoming critical in determining system performance and reliability. Power dissipation is becoming one of the most challenging design constraints in nanometer technologies. For example, among various design implementation schemes, standard cell ASICs offers one of the best power efficiency for high-performance applications. The flexibility of ASICs allow for the use of multiple voltages and multiple thresholds to match the performance of critical regions to their timing constraints, and minimize the power everywhere else. Typically, implementing nanometer-scale ICs begins and ends with wires. The ability of an IC to perform its function is dependent upon the transformation of that function into a specific configuration of wires and their connections to cell-level and, ultimately, to transistor-level behaviors. This paper discusses nanometer power aware design, optimization and management that operate upon a behavioral design description.

Introduction

The steady downscaling of transistor dimensions over the past two decades has been the main reason to the growth of silicon integrated circuits (ICs) and the electronic industry. The more an IC is scaled, the higher becomes its packing density, the higher its circuit speed, and the lower its power dissipation. These have been the key in the evolutionary progress leading to today's computers and communication systems that offer superior performance, dramatically reduced cost per function, and much-reduced physical size compared to their predecessors. As ICs marched down the technology path from 0.13 to 90 nm, designers and their customers enjoyed faster and more complex IC, without a significant increase in power consumption. But at 90 nm and below, the cost of developing nanometer designs significantly increased, particularly in masks and NRE costs. In addition, more designs are targeted for mobile applications running off batteries resulting in new design constraints for both dynamic and static power. Finally, the physics of the silicon (power related) started to work against engineering. However, the increase in leakage current was most critical. Thinner gate oxides and lower threshold voltages result in a significant increase in leakage current and static power consumption. There are, in total, seven additional secondary sources of leakage current, all of which get worse at 130 nm and even more so at 90 nm and 65 nm. We begin with an overview of a power fundamentals and behavioral design methodology. We will continue with a description of how power issues interact with the nanometer design space and describes how some of these issues can be influenced early in the design cycle. The industry direction is towards behavioral design that is becoming an absolute necessity in order to address the many nearly intractable issues that rise in nanometer design. In 2004, consumer electronics, mostly mobile, battery-powered applications is one of the major driving forces in the growth of fabless IC companies. Gartner/Dataquest forecasts that global revenue from semiconductors used in consumer electronics will increase by 20% in 2004 ($34.1 billion) and even more in 2005. Many of these devices will go into mobile and low power applications. No Doubt, the market is demanding an urgent solution to the nanometer power crisis.

Power Fundamentals

When discussing power fundamentals the two subjects to explore are two main effects that contribute to the total power dissipation on a chip. These effects are dynamic power and leakage power. Dynamic (also called active) power occurs when a transistor switches state and is due to capacitive charging and discharging. Leakage power arises due to leakage (also called static) current flowing through the transistor. Leakage current can even consume power in standby or sleep modes of operation. With each step down in process geometry, leakage power has more or less doubled in magnitude. For example, 90nm process technology leakage power contributes 40 to 50 percent of the total power budget, with active power dissipation from transistor switching making up the rest. The management of power effects in the design process should also be concerned with power integrity. This includes analysis of IR drop, which degrades performance and reduces noise immunity, and electro-migration (EM). EM is due to high current density causing metal electrons migration, resulting in open or short circuits. EM also causes performance and reliability issues over time. The key power management challenges are to reduce device power consumption by controlling dynamic and leakage effects, and eliminate power-related failures by addressing IR drop and EM.

Current approaches

The VLSI technology today comprises devices because of their unique characteristic of negligible standby power, which allows the integration of tens of millions of transistors on a processor chip with only a very small fraction (<1%) of them switching at any given instant. As the CMOS dimension, in particular the channel length, is scaled to the nanometer regime (<100 nm), however, the electrical barriers in the device begin to lose their insulating properties because of thermal injection and quantum-mechanical tunneling. This results in a rapid rise of the standby power of the chip, placing a limit on the integration level as well as on the switching speed. IC design companies are mainly attacking the issue of nanometer power consumption. (Fig. 1) While individually effective in their own target areas, the current solutions really only provide a partial solution to the problem.

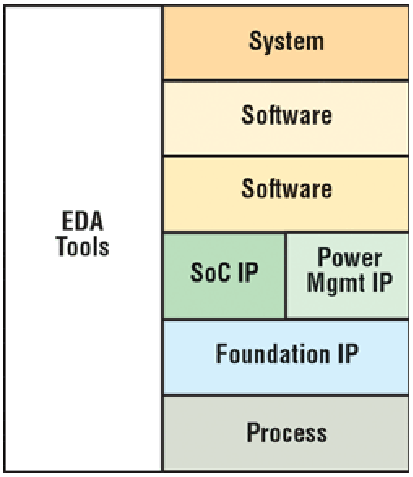


Fig. 1 – Nanometer power consumption issue.

Batteries have made significant strides in providing more power in a smaller space and with less weight. Yet these gains only offset the demand for more power as these products (laptops, PDAs, MP3 players, etc.) incorporate faster processors, advanced graphics and now micro hard drives in MP3 devices. Better batteries will not do it alone. EDA tools are used extensively throughout the design process. Analysis tools are doing excellent job of problem identification at the silicon level, but often do not provide a solution to the designer. The RTL optimization provides improvement in power, but as it is not tied to the actual silicon performance, it is at a fairly high level. Synthesis optimization for power is constantly improving, yet not enough for deep nanometer regime.

The wafer fabs have addressed the issue of power as well. Most fabs and independent foundries offer several process variations at 130 and 90 nm, each optimized for speed, power, or voltage. But all this really offers the designer is the classic power/speed tradeoff. They can have a high-speed high-leakage process or a low-speed low-leakage process. By raising the voltage threshold of the transistors, the leakage can be reduced, but this comes at a corresponding decrease in the speed of the transistor. The industry direction is to incorporate power management considerations throughout the design levels and not only at high level.

Efficient Power Management

In order to achieve effective power management an entire range of design techniques and tools must be applied, throughout the entire design process, to progressively diminish the main sources of power dissipation. Considering the system as a whole, applying power reduction measures through the process technology, circuit design, architecture, system and software, is the best strategy for effective power management. At the system level, it is clear that excessive clock speeds will dissipate power unnecessarily. The system must be clocked at a sufficient rate to meet the application software needs, and not higher. Along with determining the slowest possible clock speed, applying the lowest possible supply voltage also helps to reduce dynamic power. Conveniently, supply voltage scales with feature size.

However, lowering supply voltage requires that device thresholds are also lowered, which can increase leakage currents.

Partitioning the architecture to enable multi-voltage design can help to ensure that different parts of the chip operate at the optimum clock rates and supply voltages, ensuring both dynamic and leakage power are minimized. Due to the fact that many applications demand processing performance that varies over time between compute intensive operations, background processing and system idle, an optimized system would adjust voltage and frequency to meet each demand. In general, the decisions taken at the system level will have the greatest impact on overall power consumption. Such decisions include the functionality (for example, whether floating-point arithmetic is really necessary), hardware-software partitioning, bus strategy, clocking strategy, physical partitioning and selection of intellectual property (IP). Choice of memory architecture and implementation can have a significant effect on power budget. Applying well-known techniques during implementation such as RTL clock-gating, gate-level optimization and operand isolation can help to reduce dynamic power considerably. Clock-gating can be applied globally, or at a block level. It is important that application of clock-gating techniques does not impair testability or the ability to perform timing and formal verification. Leakage reduction techniques such as multi-Vt optimization, active well bias and state retention power gating can help to reduce leakage power considerably. Drawn feature sizes less than 100 nm (0.10 micron), dynamic power scaling trends can also lead to major packaging problems. To alleviate these concerns, techniques like thermal monitoring and feedback mechanisms can limit worst-case dissipation and reduce costs.

Low-power IP Methods

Low-power IP methods offer savings in power through the use of higher threshold transistors and by designing circuits for reduced power, but these methods come with a speed penalty. For low-speed applications, this technique produces acceptable results. The faster speed of the 130-nm process will offset the loss of performance of the low-power IP. For current-generation wireless baseband applications (cellular, Bluetooth) this works well. But for next-generation wireless products with greater graphics and video capabilities, and with high-speed wireless Internet connections, the performance trade-offs are unacceptable. What mobile electronic SoC designers require is a way to have their cake and eat it too.

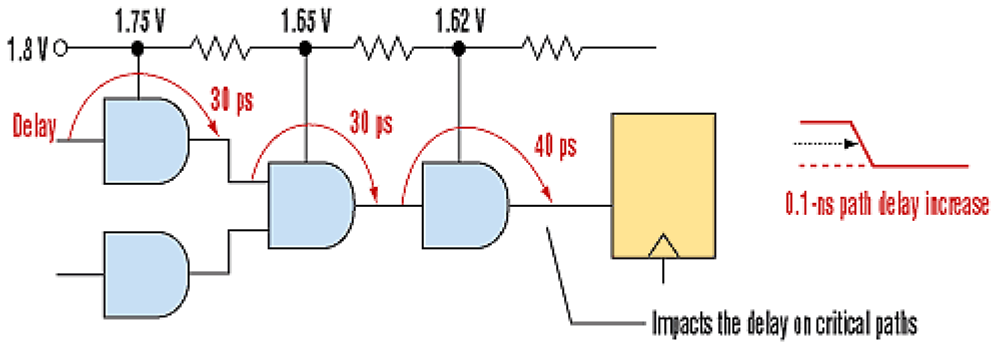
They cannot make any sacrifices to the performance of their designs and still meet the hard cost requirements that are dictated by the consumer. We all want our cell phones to have a camera and Internet access, but we still want them to fit in our pockets. As more complex and high-performance applications go mobile, SoC designers have no choice but to adopt the use of power management IP to control the increasingly severe dynamic power and leakage power issues.

Nanometer Power Obstacles

Design Size and Complexity – Larger and more complex circuitry requires more power – for many reasons. Typically, it is a good idea to reduce the number of transistors required to implement a design, which reduces the total switching activity and the nanometer main issue, leakage current. It also helps overall if the design is organized hierarchically. When a design is organized hierarchically, it can be treated as a modular structure of cells. The largest possible collection of cells can be optimized, with the lowest level of cells instantiated from a library having a selection of cells matching the optimization function: low power, specific voltage, specific clock speed, and so on. This provides a basis for planning and estimating wiring, and the opportunity to revise the architecture or the wiring plan to deal with wire-related issues.

Timing (Signal Integrity dependent) – Signal Integrity has major impact on power design. Optimizing the architecture prior to establishing the RTL description, can ensure optimum power and thus minimum opportunity for these effects. This is accomplished through simultaneously optimizing power and clock speed within appropriately segmented domains of the design; a library of high level (complex) cells characterized for various speed and voltage combinations supports this optimization process. If need be, the power constraints can then be selectively relaxed to improve performance or area considerations. An essential need in the entire nanometer-scale optimization process is access to cells characterized for signal integrity, switching power and other characterizations.

IR Drop – This issue is typically embedded within automated power optimization of architecture in cooperation with a characterized library of macro cells, or even of large IP blocks. EDA tools provide a successful solution predicting IR drop effects based on power grid requirements derived from an architectural design.

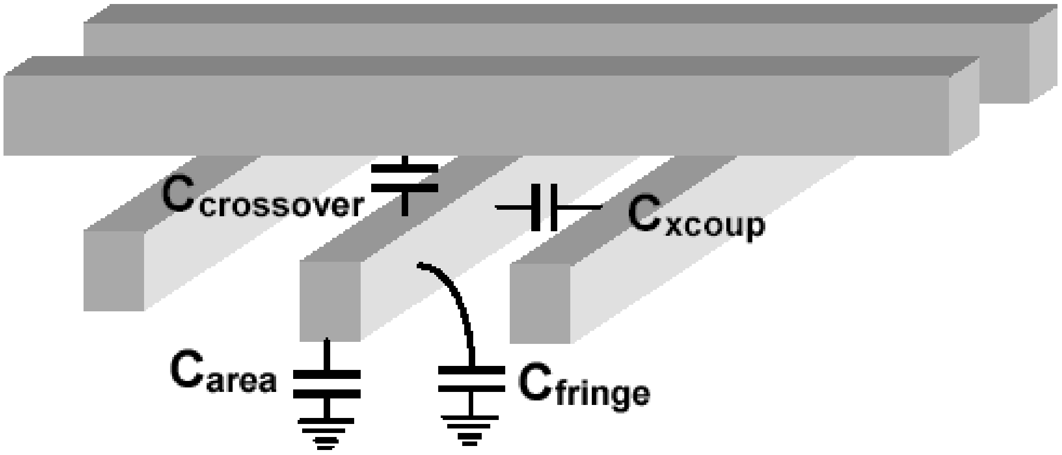

1.8 V
1.75 V
1.65 V
1.62 V
Delay
30 ps
30 ps
40 ps
0.1-ns path delay increase
Impacts the delay on critical paths

Voltage drops through a chain of logic gates can severely influence SoC timing closure. Here, 0.1 ns of additional delay, which may not have been accounted for in static timing analysis, is imposed by an aggregate drop of less than 200 mV.

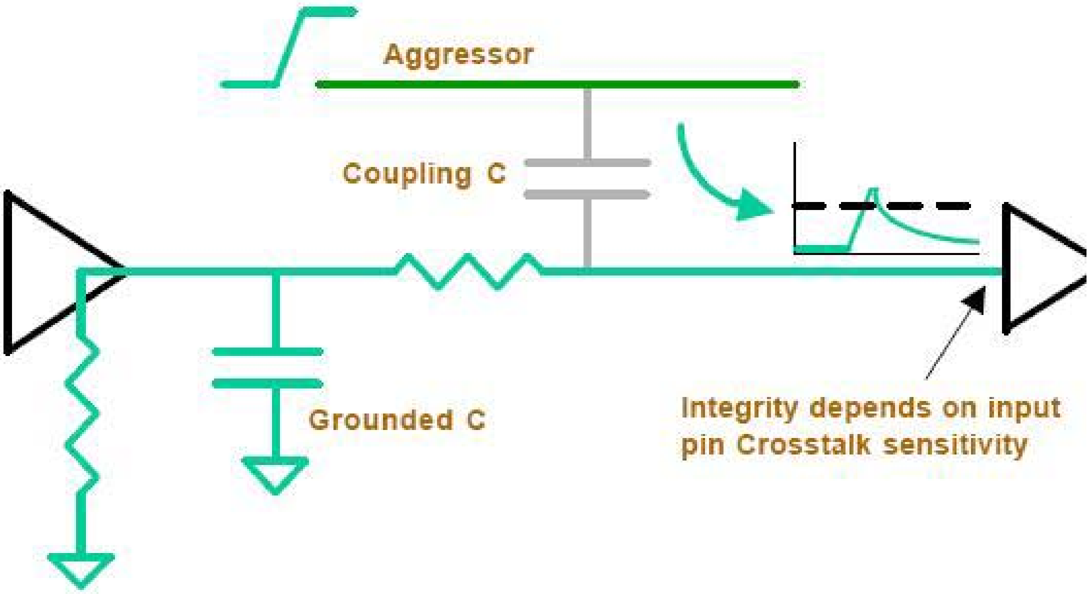
Image: Magma Design Automation

Crosstalk and Inductance – With the scaling of the horizontal dimensions of wires, the aspect ratio of the horizontal to vertical dimensions is reduced, resulting in increased ratios of coupling capacitance to ground capacitance (over or under crossovers or to substrate). A significant crosstalk noise may occur due to the relative rate of switching (rise and fall times of the signals) and the amount of

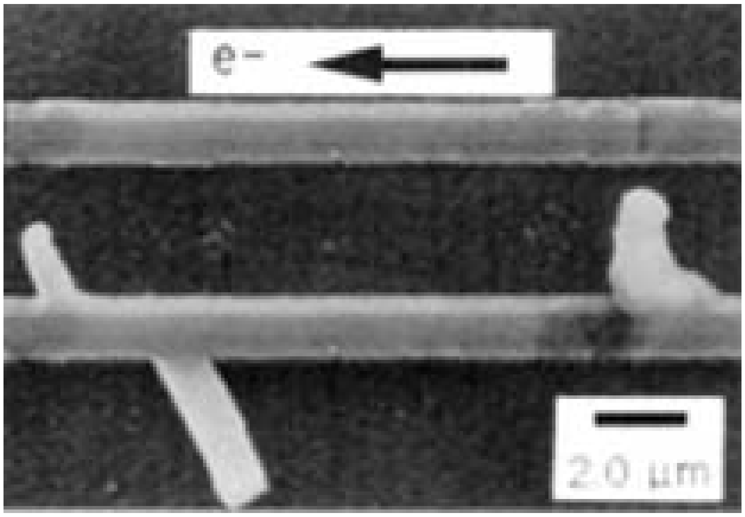
mutual capacitance. Crosstalk noise, depending on its amplitude and its timing may cause false switching or delays. The ability of a physical design environment concurrently to analyze and correct for these various signal integrity problems during a physical implementation flow is highly dependent on the design system architecture. An integrated design system is necessary to address VDSM phenomenon efficiently and to provide design closure in a timely manner.
This issue is also typically embedded within automated power optimization of architecture, performing global wiring. However, the optimization process can observe top-down constraints while growing a bottom-up design through the use of characterized cells. Behavioral design provides the means for minimizing the opportunity for crosstalk and other SI issues by managing voltage levels and clock speeds.



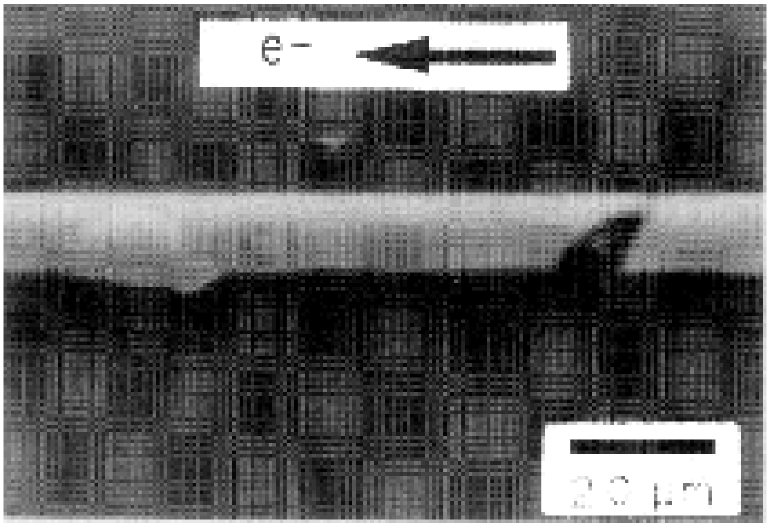
Crosstalk Noise
Image: Magma Design Automation



Electromigration – Very Deep Sub Micron designs contain millions of devices and operate at very high frequencies. The current densities (current per cross-sectional area) in the signal lines and power are consequently high and can result in either signal or power electromigration problems. The electron movement induced by the current in the metal power lines causes metal ions to migrate. That phenomenon of transport of mass in the path of a DC flow, as in the metal power lines in the design, is termed power electromigration. There are two types of electromigration. Uni-Directional, for example power and static signals and Bi-Directional, for example clocks and other switching signals. The most critical is the Uni-Directional electromigration type since the electron 'erosion' move constantly in one direction and can cause signal line failure. The power electromigration effect is harmful from the point of view of design reliability, since the transport of mass can cause open circuits, or shorts, to neighboring wires. Determination of voltage levels within an architecture or segment of same has a major influence on EM. Especially at nanometer scales, EM can be minimized through selection of connectivity and cells. Performing this at the behavioral level provides the maximum flexibility in meeting other constraints, such as performance. In power-driven design, the goal is to minimize clock speeds, voltage levels, transistor switching and capacitive loading. This is accomplished step-wise, by proposing architecture and then improving it iteratively. Each instance of architecture is marked by specific scheduling of computations, by a controller design, and by assignment of leaf cells. The leaf cells are the key. They are usually large blocks implementing multi-bit functions or even greater functionality. The cells are (ideally) available over a range of voltage levels and clock speeds to meet the possible needs of the architecture under design. It is even possible that such cells can be synthesized in response to need. However, the cells must be characterized. The result of the architecture synthesis then is an RTL design containing leaf cells of known activity characterization, which activity can be further applied in the context of a predicted physical design to support estimation of IR, EM and crosstalk effects, among others, which can then be optimized through iteration.



Electromigration Effect – Short Circuit
Image: Computer Simulation Laboratory



Electromigration Effect – Open Circuit
Image: Computer Simulation Laboratory

Digital-Analog Integration – These types of power issues typically are pre calculated within EDA tools results. Architectural optimization can account for voltages and clocks in order to minimize the effects of transitions in the area of analog blocks inside digital optimization cycle. It is unlikely to be practical to optimize it outside the cycle in real time.

Power Consumption – Power consumption is directly related to wiring concerns, for all the reasons already discussed. Addressing the magnitude of supply voltage, switching activity in the circuit, switching capacitive loads and clock frequency in a top-down, behavioral, iterative manner, can strongly affect many of the determinants of signal behavior, and in many cases optimize their effects achieving best results.

Power Integrity - Power integrity is a significant challenge as chip designs move to 90nm and below process technology. With the increasing of number of devise, technology scaling, increasing device integration, decreasing supply voltages, increasing leakage currents and the use of low-power techniques negatively impact current distribution gradients (Ldi/dt) and compound power integrity issues. At 90nm, static (or average) IR-drop analysis is insufficient to capture the power-ground noise-related chip failure issues. Static IR-drop takes into account only the power grid resistance (R), which can be used to identify gross design violations. At 90nm, it is important to consider the complete dynamic nature of the power-ground noise, including capacitive and/or inductive effects, simultaneous switching noise, and the effect of decoupling capacitors (de-caps). Undetected and uncorrected, these effects can lead to design failure. Power integrity analysis has become a must as part of the design sign-off process.

Power management IP

Power management and power rail design are critical issues for very deep submicron (VDSM) designs. With decreasing supply voltages, increasing demand for low power applications and increasing device densities, the challenge to minimize power consumption, minimize voltage drop effects, and maximize product reliability cannot be handled by traditional physical design techniques. Traditional back-end verification of power and rail design are necessary for sign-off, but the cost of detecting and repairing such problems at the end is extremely high. The IC layout need to be analyzed and optimized to provide a layout that meets the designer's power and reliability constraints. The entire circuit power consumption should be optimized and reduced if possible. Voltage drop and electromigration violations should be correct. Accurate timing for the full design is needed including each cell's timing to reflect the local voltage condition. With the physics of the semiconductor working against engineering at nanometer process technologies, the designer must turn to the voltage and clock frequency of the IC to find power savings. In the simplest sense, designers want to be able to treat individual blocks of an IC in much the same way they did when the different parts of the IC were separate chips on a pc board.

These chips could run at different clock rates and different voltages and could be turned off when not in use in order to save power. The designer needs to be able to partition the IC into electrically autonomous areas, called power islands, based on functionality (CPU, MPEG4, memory, analog, etc.) and power requirements. They then need to be able to dynamically scale the voltage and clock frequency of the power islands, so that each island is only running at the performance required at that time. By enabling designers to dynamically manage the voltage and frequency of individual power islands, designers can significantly lower both the dynamic power, and the quiescent or leakage power, of their IC's designs.

Dynamic voltage and frequency scaling can be additive to many of the other current power optimization techniques. But full implementation of this technique requires a vertically integrated approach within IC design supply chain, and power management IP. Power management IP is more than just low-power IP, it is IP that allows the designer to actively control and manage the power consumption of the IC during operation. Power management IP allows the CPU and software to issue commands to dynamically scale the voltage, frequency, and leakage current of the power islands of an IC. It allows the designer to leverage the enabling elements of the IC design supply chain—low-power IP (standard cells, memory, etc.), software, SoC IP (CPU, DSP, etc.), process technology and power supply IC—to dynamically manage the power of an IC while it is operating.

Dynamic Power Analysis

1. Power Heat Dissipation (Packaging)

As the semiconductor industry adopting the very deep-submicron (VDSM) technology, it creates a new demands on the packaging industry. Especially with high performance IC's the packaging is becoming a significant factor. Increased functionality, faster performance, lower operating voltages and reduced size are leading to increases in die density and I/Os, boosting package pin count and complexity. This has created the need for a new breed of high-density, multilayer, custom-designed packages. We will mention few; flip-chip, ball-grid-array (BGA), and pin-grid-array (PGA).

With chip's power rising, packaging technology must improve to meet heat dissipation demands. The reduction of thermal junction resistance requires advanced cooling techniques such as larger, more powerful fans, liquid/gas vapor cooling, etc'. Packaging experts believe cooled systems are the best solution for packaging high power density VDSM designs.

The advantages of cooling the ambient and junction temperatures are well known: improved voltage scalability due to reduced current leakage, higher carrier motilities, lower interconnect resistances, and improved reliability. Advanced cooling techniques like vapor compression are expensive and predicted to be used only for large IC's. Desktop applications are expected to use low power cooling methods.

Another approach to the packaging heat-constrain is dynamic thermal management. This concept involved thermal management technique that can be achieved in few ways. An example is Transmeta's approach to dynamically varies the supply voltage when the CPU is not heavily loaded. Another example is the thermal monitor in Intel's Pentium IV design which has an on-chip temperature sensor (The temperature sensor is a diode with a fixed voltage across it) along with a reference current source and current comparator to determine when the on-die temperature exceeds a given value. When the temperature (and power consumption) is exceeded the permitted level, the internal clock frequency is reduced, limiting power, throughput and performance. The immediate effect is a reduction in the chip thermal level to bring it to the permitted range.

The importance of dynamic thermal management techniques lies in their ability to reduce the chip power (Wattage) to the effective worst-case power dissipation rather than the theoretical worst-case. The effective worst-case power consumption, as found by running power-hungry applications, is about 75% of the theoretical worst-case, which is determined using synthetic input code sequences that are not realized in practice. This difference has major implications for packaging costs and design flexibility. Small increases in the maximum power can lead to significantly, expensive cooling techniques.

No Doubt, packaging will become more and more in the critical path of the high performance VDSM designs. The increase demand for larger and more powerful chips creates new challenges for the IC's packaging industry.

2. Global Signaling

Global signaling within high performance VDSM designs is one of the serious challenges in the nanometer arena. The propagation of global signals across a large die in a shrinking clock period creates an entire series of electrical phenomenon. It appears likely that global signaling will use a slower clock than localized logic such as datapaths (although multi-cycles nets can be broken up using latches). Even with relaxed timing constraints on global communication, significant power is consumed to achieve the desired global clock speeds. Based on the current signaling paradigm of inserting large CMOS buffers along an RC line, this requires over 50 W of power in the nanometer arena. The proliferation of repeaters (nearly 106 required at 50-nm compared to about 104 in a large 180nm microprocessor and controllers) heightens difficulties in power distribution and floorplanning2. One solution is to use advanced signaling strategies such as differential and/or low-swing drivers and receivers for global communication.

In many cases, these approaches can lead to power and tpd (time propagation delay) savings due to smaller voltage transitions as well as major reductions in the scale of power grid current transients. For instance, the Alpha chip uses differential low-swing buses to communicate between functional units. Worst-case power for these buses was reduced considerably by limiting the voltage swing to 10% of Vdd. Differential signaling increases routing area, but the increase may be less than the expected factor of 2 due to the use of shielding in global signaling to limit coupling from neighboring signals on long lines. In addition, shielding may be insufficient to limit inductively coupled noise, whereas low-swing differential signaling creates less noise and is more noise immune than single-ended full-swing CMOS. With the industry trends indicating rising power consumption for global communication, the use of alternative signaling strategies will most likely increase. Further study is necessary to provide an efficient solution to the global signaling concern.

3. Library Optimization

Silicon-proven libraries, give designers and fabless companies very high performance solutions, using some of the most advanced processes available. While most high performance microprocessors rely heavily on custom design, library optimization can significantly enhance performance in these applications. Advances in library generation, and synthesis tools that take advantage of improved libraries, can together yield more automated, less expensive design flows. Libraries are one important reason that custom designs are significantly faster (6-10X) than counterpart ASIC designs. For instance, asserts that the lowest performance level (smallest) gates in modern libraries are nearly 10X larger than minimum-sized gates, leading to major power increases due to overdriving small loads. However, most current libraries contain a large number of drive strengths, including some very near minimum size. As evidence, we refer to the same 180 nm library as the smallest standard cell inverter has an input capacitance of just 1.5fF and the smallest inverter with balanced rise/fall delays has an input capacitance of 6.6fF. Other leading-edge libraries contain a rich set of drive strengths (e.g. 11 2-input NANDs, 16 inverter sizes), dual output polarities, and single pin inverted inputs on NAND/NOR's. This recent increase in library complexity seems to be closing the gap slightly between custom designed cells and those from libraries.

Today EDA vendors provide an entire line of optimized VDSM libraries in order to help customers to achieve efficient designs. For example like Synopsis (using Avant! Tools) provide today an entire set of optimized libraries including standard cells, IO's and memory compilers. The prediction is that in the near future the entire industry will use an optimized libraries provided by the foundries.

4. Multiple Powers Vdd

One of the most efficient methods to rise of dynamic power in VDSM designs is to use multiple power supply lines. (Vdd's) The general idea called clustered voltage scaling (CVS). With two Vdd levels (Vdd_h and Vdd_l), the circuit is partitioned so that non-critical gates run at Vdd_l and only critical gates use Vdd_h. Level conversions, performed when gates running at Vdd_l fan-out to gates at Vdd,h, are reduced by clustering Vdd,l and Vdd,h gates together to minimize the number of such interactions. Analysis indicates that Vdd_l should be around 0.6 to 0.7 times Vdd_h to maximize power savings. The dynamic power reduction by using two Vdd levels is readily calculated if one can estimate the fraction of cells that can be assigned to Vdd_l. Existing media processor designs that use CVS report that ~75% of all gates can tolerate Vdd_l without altering the critical path delay.

The key challenges to the use of multiple supplies on a chip lie in minimizing area overhead and providing EDA tool support for Vdd cell selection, placement given new clustering constraints, dual power grid routing, and enhanced library generation capabilities. Using this system within EDA tools provides a powerful capability for VDSM high performance designs.

Static Power Analysis

1. Multiple Vth Approaches

In order to reduce CMOS static power consumption several approaches have been developed. In this section we will briefly discuss several of these techniques that use multiple thresholds on a single chip to limit Ioff.

A. MTCMOS Method

Multi-Threshold CMOS (MTCMOS) gates a high-Vth transistor with a sleep mode signal to virtually eliminate leakage current in idle states. The sleep transistor is placed between ground and fast low-Vth CMOS logic. As it is in series, it adds delay, which can be reduced by increasing its area. Disadvantages include no leakage reduction in active mode, increased device area, and additional overhead for routing sleep signals. Other similar techniques include dual-Vth domino logic, substrate biasing to modify Vth in standby, and using negative NMOS gate voltages to bias the devices further into cut-off. A single threshold leakage reduction technique combines the concepts of sleep transistors and state dependent leakage. All these techniques trade off area to limit static power and most only reduce leakage in standby mode. In fact, they are currently limited to portable applications such as notebook processors. Also, some of the proposed methods do not scale well – the use of domino logic for example, and substrate bias controlled Vth (body bias is less effective at controlling Vth in scaled devices). Dual Vth insertion, described next, is the only technique used in current high-end MPUs.

B. Dual-Vth Method

Today, circuit designers have access to multiple threshold voltages on a single IC to select between gates that use high or low thresholds. The impact of Vth on the delay and power of gates such as inverters and NANDs is significant. A reduction in Vth (with constant Vdd) exponentially increases off current and roughly linearly reduces the propagation delay. An additional threshold adjust ion implantation step allows designers to choose from a wider range within the power-performance design envelope. Gates located on critical paths can be assigned fast low Vth, while gates that are not timing critical can tolerate high Vth and slower response times. Typical results show leakage power reductions of 40-80% with minimal penalty in critical path delay compared to all low-Vth implementations. Figure 2 shows the increase in Ion for the low-Vth device. The relative difference in Ioff between the two devices will remain constant throughout the roadmap (at about a 15X increase in Ioff for 100 mV reduction in Vth). Given that the off current change is constant, the steady improvement in Ion with scaling demonstrates that the dual-Vth (or multi- Vth) approach to leakage reduction is inherently scalable. Figure 2 also shows the resulting Ioff increase for Ion to rise 20% beyond the high-Vth case. At 35 nm, just a 7X rise in Ioff is required to yield 20% drive current improvement, compared with a factor of 54X today. In Figure 3 we can see the Ioff decrement with the process shrink and the reduction of Vdd.

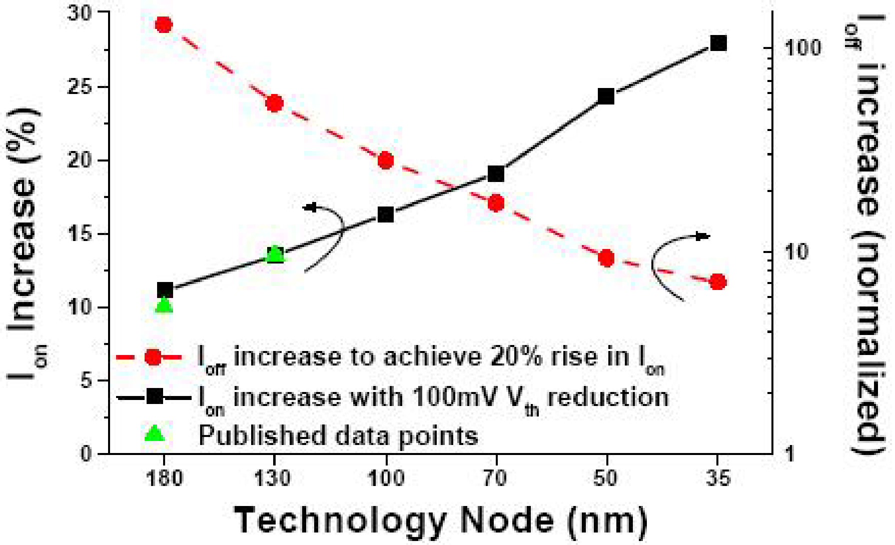


Figure 2. I_{on} increases more rapidly with a 100mV change in V_{th} for scaled technologies. I_{off} penalty for 20% I_{on} gain reduces with scaling.

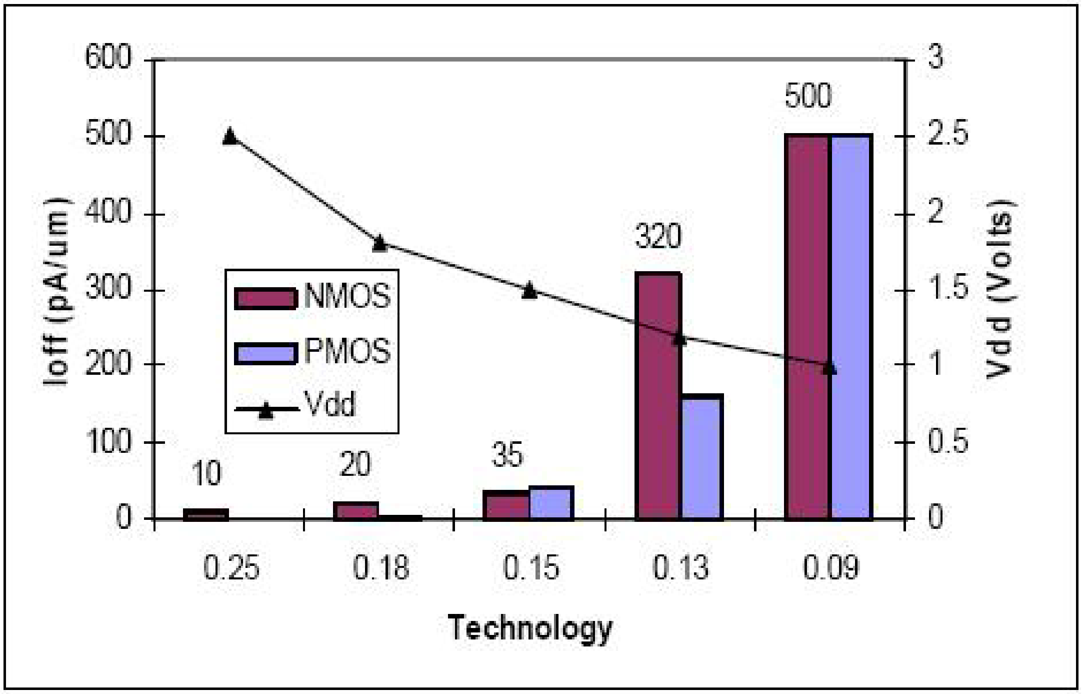


Figure 3 - Ioff reduction with the process and power

2. Scalable Dynamic/Static Power Approach

One of the most appealing approaches, in order to achieve scalable, flexible and cost effective design is a combination of multiple Vdd’s and multiple Vth’s. The combination of multiple Vdd’s, multiple Vth’s, and intra-cell size and Vth assignments points to a highly flexible, scalable, cost effective design approach to dynamic and static power minimization. With two voltage supply values available, different Vth’s will allow designers or EDA tools to choose to emphasize speed, standby power, or dynamic power.

Summary

The EDA industry continues to face new challenges as process continues to shrink into nanometer geometries. With each successive advancement of semiconductor technology a new VDSM challenge is born. Especially with high performance reliable designs the industry has to face a wide variety of phenomenon such as heat dissipation, electromigration, interconnect coupling and more. Many EDA tools have been enhanced to deal with these issues. The keys to a successful nanometer design are:

1. Efficient and reliable power management techniques such as on-chip temperature monitors and multiple voltage supplies will reduce dynamic power, enabling cheaper packaging and higher integration densities.

2. Power distribution will be manageable from the standpoint of IR drop – given changes in the ITRS to take advantage of technological advancements in flip-chip packaging. However, large current transients may be exacerbated by the use of sleep/standby modes.

3. Alternative techniques to CMOS repeaters for global signaling need to be studied and implemented within EDA tools to minimize power consumed in global communications.

4. A multi-layered approach to power reduction (both dynamic and static), combining multiple threshold and supply voltages with flexible gate layouts using different thresholds and device sizes within a gate. Non-critical gates are first assigned to a reduced Vdd, followed by sizing and Vth selection to reduce power most efficiently.

Taken together, the various aspects described in this document form a methodology that provides comprehensive power optimization. The power optimization aspect needs to be performed throughout the entire design stages particularly during architectural synthesis phase in order to solve critical nanometer issues.
While the power crisis in the nanometer arena is clearly a challenge, designers are beginning to see the introduction of a new breed of power management IP integrate with other elements of the design supply chain. In the coming years, we can expect more innovative EDA products that will deliver powerful solutions for the nanometer power crisis.

References

[1] J. Cong, L. He, C.-K. Koh, D. Z. Pan, and X. Yuan, "TRIO: Tree, repeater and interconnect optimization package." http://cadlab.cs.ucla.edu/_trio.
[2] J. Cong, L. He, A. B. Kahng, D. Noice, N. Shirali, and S. H.-C. Yen, "Analysis and justification of a simple, practical 2 1/2-d capacitance extraction methodology," in *Proc. ACM/IEEE Design Automation Conf.*, pp. 40.1.1–40.1.6, June,1997.
[3] J. Cong, L. He, C.-K. Koh, and P. H. Madden, "Performance optimization of VLSI interconnect layout," *Integration, the VLSI Journal*, vol. 21, pp. 1–94,1996.
[4] J. Cong, L. He, K.-Y. Khoo, C.-K. Koh, and D. Z. Pan, "Interconnect design for deep submicron ICs," in *Proc. Int. Conf. on Computer Aided Design*, pp. 478–485, 1997.
[5] Semiconductor Industry Association, *National Technology Roadmap for Semiconductors*, 1997.
[6] J. Cong and D. Z. Pan, "Interconnect delay estimation models for synthesis and design planning," in *Proc. Asia and South Pacific Design Automation Conf.*, pp. 97–100, Jan., 1999.
[7] J. Cong and D. Z. Pan, "Interconnect estimation and planning for deep submicron designs," in *Proc. Design Automation Conf*, June, 1999.
[8] C.-P. Chen and D. F. Wong, "Optimal wire sizing function with fringing capacitance
consideration," in *Proc. Design Automation Conf*, pp. 604–607, 1997.
[9] J. Cong, "Challenges and opportunities for design innovations in nanometer technologies," Dec. 1997. http://www.src.org/research/frontier.dgw.
[10] J. Cong, L. He, C.-K. Koh, and D. Z. Pan, "Global interconnect sizing and spacing with consideration of coupling capacitance," in *Proc. Int. Conf. on Computer Aided Design*, pp. 628–633, 1997.
[11] K. Usami and M. Horowitz, Clustered voltage scaling techniques for low-power Design, ISLPED 1995.
[12] C. Chen, A. Srivastava, and M. Sarrafzadeh, On gate level power optimization using dual-supply voltages, IEEE Trans. on VLSI Systems, vol.9, p.616-629, Oct. 2001.
[13] N. Rohrer, et al., A 480MHz RISC microprocessor in a 0.12_m Le_ CMOS technology with copper interconnects, ISSCC 1998, p.240-241.
[14] S. Sirichotiyakul, et al., Standby power minimization through simultaneous threshold voltage selection and circuit sizing, DAC 1999, p.436-441.
[15] Q. Wang and S.Vrudhula, Algorithms for minimizing standby power in deep submicron, dual-Vt CMOS circuits, IEEE Transactions on CAD, vol.21, p.306-318, 2002.
[16] M. Hamada, Y. Ootaguro, and T. Kuroda, Utilizing surplus timing for power reduction, CICC 2001, p.89-92.
[17] K. Usami, et al., Automated Low-Power Technique Exploiting Multiple Supply Voltages Applied to a Media Processor, IEEE JSSC, Vol.33, No.3, 1998.
[18] M. Hamada, et al., A top-down low power design technique using clustered voltage scaling with variable supply-voltage scheme, CICC 1998, p.495-498.
[19] D. Sylvester and H. Kaul, Future performance challenges in nanometer design, DAC 2001, p.3-8.
[20] A. Srivastava and D. Sylvester, Minimizing total power by simultaneous Vdd/Vth assignment, Proc. Asia-South Paci_c DAC 2003, p.400-403.
[21] C. Yeh, et al., Layout Techniques supporting the use of Dual Supply Voltages for Cell-based Designs, DAC 1999.

[22] D. Lackey, et al., Managing Power and Performance for SOC Designs using voltage islands, ICCAD 2002.
[23] S. Kosonocky, et al., Low Power Circuits and Technology for wireless digital Systems, IBM Journal of R&D, Vol. 47, No. 2/3, 2003.
[24] A. Correale, D. Pan, D. Lamb, D. Wallach, D. Kung, R. Puri, Generic Voltage Island: CAD Flow and Design Experience, Austin Conference on Energy E_cient Design, March 2003 (IBM Research Report)
[25] W. Donath, et al., Tranformational Placement and Synthesis, DATE, 2000.
[26] R. Puri, E. D'souza, L. Reddy, W. Scarpero, B. Wilson,
Optimizing Power-Performance with Multi-Threshold Cu11-Cu08 ASIC Libraries, Austin Conference on Energy E_cient Design, March 2003 (IBM Research Report).
[27] R. Puri, D. Pan, D. Kung, A Flexible Design Approach for the Use of Dual Supply Voltages and Level Conversion for Low-Power ASIC Design, Austin Conference on Energy E_cient Design, March 2003 (IBM Research Report).
[28] Y. Taur, CMOS Design near the limit of scaling, IBM Journal of R&D, Vol. 46, No. 2/3, 2002.
[29] J. P. Fishburn, Clock Skew Optimization," IEEE Transactions on Computers C-39, pp 945-951, 1990.
[30] T. G. Szymanski, Computing Optimal Clock Schedules," DAC 1992, p.399-404.
[31] C. Albrecht, B. Korte, J. Schietke and J. Vygen, Cycle Time and Slack Optimization for VLSI-Chips, ICCAD 1999, p.232-238.
[32] S. Bhattacharya, J. Cohn, R. Puri, L. Stok and D. Sunderland, Power reduction of Hardwired DSPs in standard ASIC methodology, Submitted to CICC, 2003.
[33] L. Stok, et al., BooleDozer Logic Synthesis for ASICs, IBM Journal of R&D, Volume 40, no. 3/4, 1996.
[34] F. Beeftink, P. Kudva, D. Kung, R. Puri, L. Stok, Combinatorial cell design for CMOS libraries, Integration: the VLSI Journal, V.29, p.67, 2000
[35] P. Kudva, D. Kung, R. Puri, L. Stok, Gain based Synthesis, ICCAD Tutorial, 2000.