Nanometer Power Leakage

Dr. Danny Rittman
November 2005
danny@tayden.com

Abstract

The need to reduce IC's power consumption long ago recognized as a significant design issue and became more critical as larger, faster ICs go into portable applications. As a direct result, power management techniques are evolved throughout the design flow in order to assure the product reliability. One of the most significant power related subjects that arise for the past decade is power leakage. At sub-100 nm technologies, leakage power forms a significant component of the total power dissipation, especially due to within die and die-to-die variations in process, temperature and supply voltage. Since leakage power and operating temperature are electro-thermally coupled to each other, increasing power dissipation and thermal problems are becoming key concerns not only from a thermal management point of view but also because most reliability phenomenon are highly temperature sensitive. The demand for low-power designs for portable electronics has never been higher, even as the management of power dissipation has become one of the most challenging design constraints for 65nm and below. This paper presents an overview of the various components of leakage power, which became one of the most significant factors in nanometer power management. We also discuss temperature and reliability tradeoffs in leakage dominant nanometer designs.

Introduction

In the past decade, power dissipation and thermal management have been identified as key factors for nanometer designs. Leakage power, which is rapidly becoming a significant contributor to the total chip power, is strongly affected by the technology scaling, on-chip process, temperature and voltage variations. Furthermore, sub-threshold leakage power, which is the dominant leakage source for high performance nanometer designs, increases exponentially with die (junction) temperature rise. The die temperature in turn, is determined by the total chip power dissipation and system packaging/cooling technology. The total power consumed by a chip equals dynamic power plus static power. Dynamic power is the power consumed in switching logic states, both internal to the cells (internal power) and for driving the chip's nets and external loads (switching power). Dynamic power can be generally described via the next formula:

$$P_{DYMANIC} = C*V^2*F$$

Where C is the load, V is the voltage swing and F is the number of logic-state transitions.

As semiconductor structures become smaller, device and interconnect capacitances decrease, allowing for higher performance and lower power. Countering these factors are power increases due to larger designs and higher switching rates. Static power (leakage power) is consumed while transistors are not switching. This type of power is a derivative of DC current and can be generally described via the next formula:

$P_{STAT} = V*I_{STAT}$ (Not time domain dependent)

Although transistors have some reverse-biased diode leakage from drain to substrate, the larger portion of leakage power is due to the sub-threshold current through a transistor that is turned off. This sub-threshold current results from the conduction between source and drain through the transistor channel. The sub-threshold leakage current is a major issue because it increases as transistor threshold voltages (Vth) decrease. In fact, the move to 65nm and below may boost leakage power as high as 50 percent of the total chip power (Figure 4). Increased leakage power causes to exponentially increase reliability related failures in chips (even in standby mode).

Leakage power is strongly affected by the technology scaling, on-chip process, temperature and voltage variations. It has significant implications on IC performance, power management and reliability.

The impact of Nanometer Technology on leakage power

As IC's are deeply scaled into deep nanometer dimensions and operate in giga-hertz frequencies, power density and on-chip temperature in circuits have been rising steadily, determining the system's performance and reliability. Increasing the power trend in IC's creates significant rise in die temperatures, forming local hotspots on the substrate and therefore creates reliability issues. Due to the higher on-chip temperatures, the power leakage considerably increases. Another factor that causes the power leakage rise is within-die and die-to-die parameter variations. Die to-die parameter variations which result from lot-to-lot, wafer-to-wafer and a portion of within-wafer variations impact every element on a chip equally. On the other hand, within-die parameter fluctuations consisting of both random and systematic components produce non-uniformity of electrical characteristics across the chip. Within-die parameter variations can be categorized into two segments: environmental variations (temperature, power supply) and physical variations which include all process variations. These factors are all included within the static power dissipation domain. While dynamic power dissipation was the largest consumer of energy, with each step down into deep nanometer geometries, static power dissipation has become a key factor for IC's reliability and performance. Let's take a look at the source of static power dissipation.

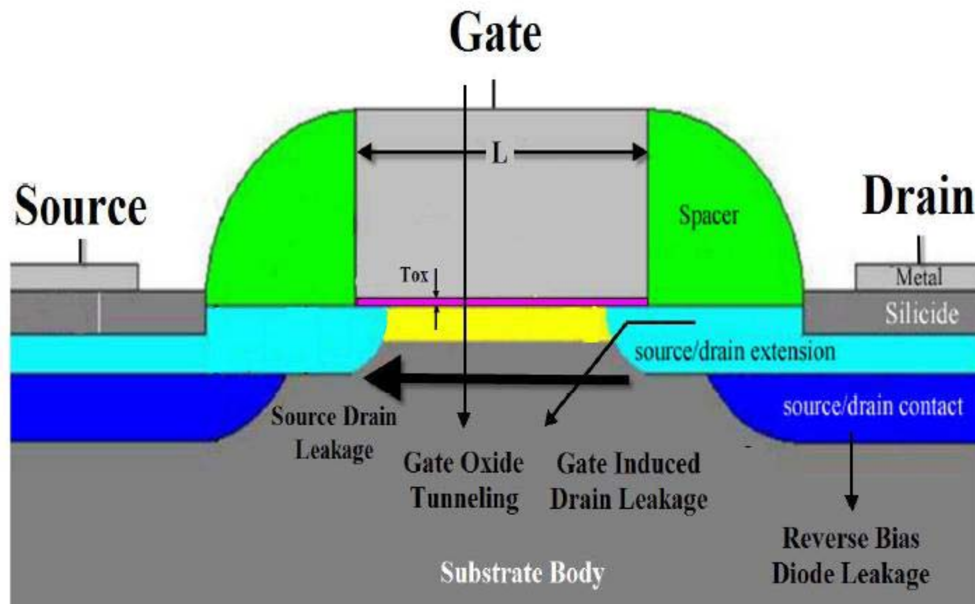

Figure 1: Leakage Current Source
Image source: eAsic

Figure 1: Components of leakage current that cause static power dissipation. The four leakage components are shown as arrows in the transistor cross section.

ILEAKAGE = ISUB + IOX + IGIDL + IDIODE

The dominant source of transistor leakage current at the 130 nm process node is **ISUB** the sub-threshold current in the channel between the source and drain. **ISUB** increases exponentially with a reduction in the transistor threshold voltage or an increase in die temperature. There are second-order effects such as Drain Induced Barrier Lowering (DIBL) and narrow width effects that are modeled as a shift in the transistor threshold voltage. These effects are becoming more pronounced at process nodes below 100 nm and will be discussed later in this section. The gate oxide tunneling current, **IOX**, is caused by electrons passing through the very thin gate oxide. At 130 nm, this current is less than 10% of **ISUB** but it begins to approach sub-threshold leakage current in magnitude in process nodes below 100 nm. At the 65 nm process node, when only 5 or 6 layers of atoms form the gate oxide, **ISUB** and IOX will be about equal. Unlike the other leakage currents, IOX is not heavily dependent on temperature but on the thickness of the gate oxide **TOX**. The leakage component **IGIDL** is Gate Induced Drain Leakage current. Since the transistor is turned off, there is voltage difference between the

gate and the drain terminals. This electric field causes some current to flow from the drain into the substrate. **IGIDL** increases as **TOX** becomes thinner and it also increases with temperature. The last component is the reverse bias diode leakage from the drain into the substrate. It also increases exponentially with temperature. Of the four leakage currents, three are heavily dependent on the die temperature. If the static power dissipation is large enough to increase the die temperature then the leakage current will increase, which causes the die temperature to increase even more. The resulting vicious cycle can lead to "Thermal Runaway" which can literarily cause the chip to melt. The heavy dependence of leakage current on die temperature and different threshold voltages can be seen in Figure 2.

Standard libraries are usually characterized at operating temperatures of **25O**C and **125O**C. (for ASIC's) Temperature de-rating factors are used to calculate the active and standby currents at the various die temperatures in between. In reality, the die is rarely at either of these two extremes and since some sections of the die are hotter than others there can be thermal gradients across the chip. Low-power design flows require accurate characterization of ASIC libraries at additional Process, Temperature and Voltages (PVT) operating points and EDA analysis tools need the ability to handle voltage drops, temperature gradients and process variations across the die. Once considered a second-order effect, temperature must be integrated into timing and power analysis going forward due to its increasing influence on power dissipation.

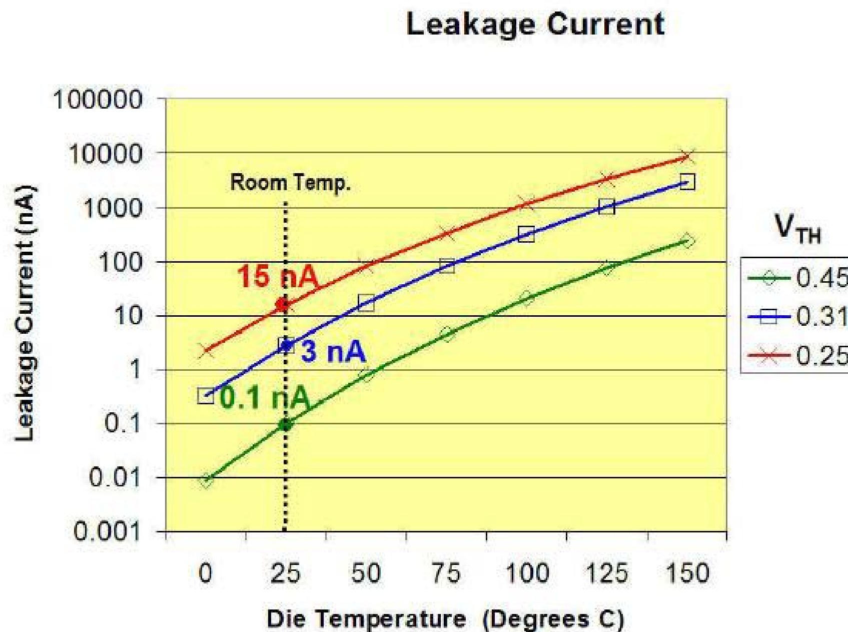

Figure 2: Leakage Current dependency on Vth and Temperature
Image source: eAsic

Sub-threshold Leakage

Comparing nanometer technologies, we are witnessing a significant increase in the sub-threshold current leakage power as a direct result of new geometries. At the 130 nm process node, sub-threshold leakage power is about 18% of total power dissipation but at the 65 nm node, sub-threshold leakage power becomes 54% of the total power unless radical steps are taken to reduce it. When a MOS transistor is turned off, a small amount of leakage current still flows. With millions of transistors on a chip, this leakage power adds up quickly. With each generation of technology, the threshold voltage will fall and with each 100 mV drop in threshold voltage the leakage current will increase by 15 times.

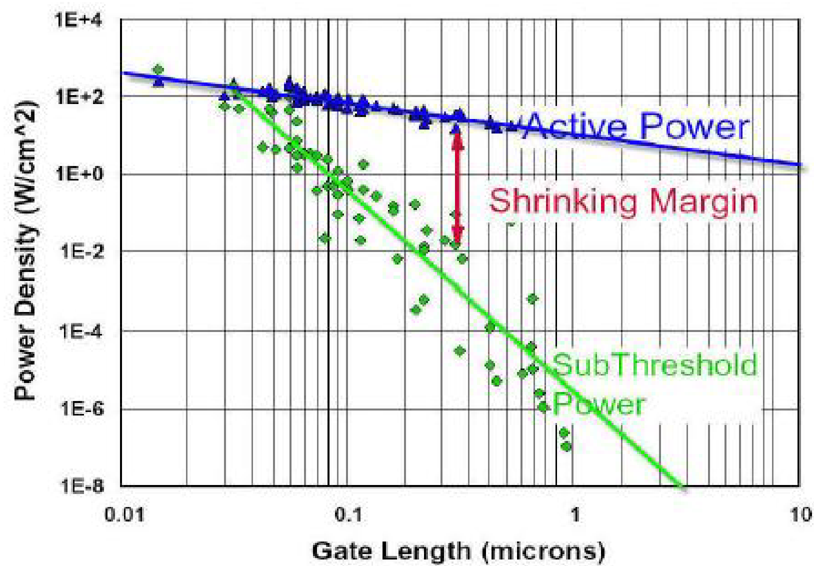

Figure 3: Leakage Power Matches Active Power
Image source: eAsic

Figure 3 shows the shrinking margin between active power and leakage power that has been a cause of industry concern for several years. Due to technology scaling and parameter variations, leakage power dissipation, which is dominated by sub-threshold leakage for high-performance ICs, becomes a significant component of total chip power consumption.
Also, sub-threshold leakage power dissipation is exponentially dependent on temperature and the dependence gets stronger with scaling.

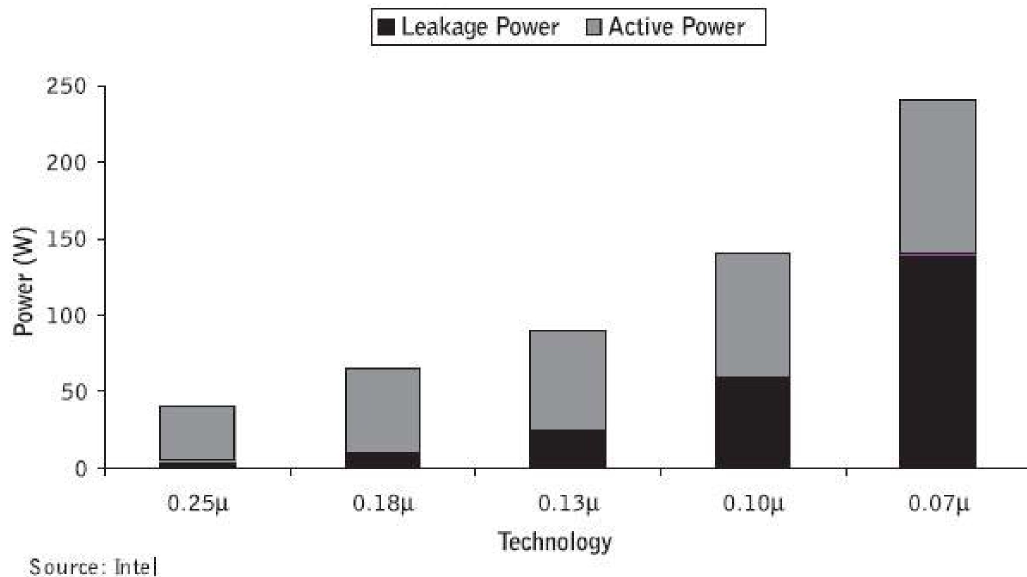

Figure 4: Increase in leakage power; bringing down transistor threshold voltages helps decrease dynamic power but increases sub-threshold leakage current.

desirable to scale supply voltage (**Vdd**) to reduce active power consumption. Although, scaling **Vdd** will degrade the performance of the circuit, it can be partially compensated by lowering threshold voltage (**Vth**) at the cost of increased leakage power. Thus, for applications, where both performance and amount of computation that can be done for a given energy budget are of importance, energy-delay product (EDP) is an appropriate metric to optimize and compare different designs. Also, system reliability is directly related to the operation temperature. For instance, reliability mechanisms such as electromigration (**EM**) and time-dependent gate-oxide breakdown (**TDDB**) are known to have an inverse exponential dependence on temperature. Therefore, it is crucial to generate a reliability and thermal aware design space for evaluating various power-performance-reliability tradeoffs and also for comparing different circuit designs under different reliability constraints.

Multiple voltage islands

Using multiple supply and/or threshold voltages may help manage leakage power. The use of voltage islands or voltage domains offers a way to meet both power consumption and performance requirements. In this scheme, sections of logic are grouped physically into separate regions according to their functionality. The logic regions that must operate at the highest speed use the highest supply voltage, while less timing-critical regions use lower

supply voltages. Frequency scaling is thus necessary along with the voltage scaling, so the voltage island approach works well with clock gating. The logic in a clock-gated block constantly consumes leakage power, but reducing the supply voltage to this block reduces the leakage. Multiple supply voltages must be provided through separate power pins or analog voltage regulators integrated into the device. The efficiency of these voltage regulators must be included in power calculations for the device. If only a small portion of the design will operate at a lower voltage, more power may be lost in the voltage regulator than is saved in the lower-voltage logic. Note that voltage island design may require level-shifter cells to ensure a proper rail shift for signals traveling between voltage domains.

In addition to reduce supply voltages, it is possible to vary the supply voltage of an island depending on system requirements. Among other challenges, this method requires the use of cells that have been characterized at all voltages.

A SoC can also be designed to power-down certain voltage islands to eliminate their leakage power. Such islands require the use of power isolation cells, which can be simple AND gates. The outputs from a powered-down section into an active power domain should never be allowed to float. Power isolation logic ensures that all inputs to the active power domain are clamped to a stable value. Additionally, a state retention technique may be required so that the blocks can resume operation when powered-up. Powering-down various islands' voltages or scaling their voltages dynamically may also require power sequencing circuitry to ensure correct operation of the chip. No doubt, using multiple supply and/or threshold voltages can be a viable solution to help manage leakage power.

Multiple-threshold design

Multiple supply-voltage islands work well with multi-threshold synthesis. Optimization meets timing goals by using low-Vth cells on critical timing paths and high-Vth cells on non-critical paths. Note that better leakage quality of results can be obtained by using state-dependent leakage models, if the silicon vendor provides such models. A one or two-pass synthesis flow can be used for multi-threshold designs, depending on the design team's methodology or preference. Initial synthesis may be performed with the low-Vth, high-performance library, followed by an incremental compile using multi-Vth libraries to reduce leakage current. For designs in which both timing and leakage are important, one-pass synthesis uses multi-Vth libraries simultaneously. The design is first optimized for timing, then leakage power optimization is performed without affecting the achieved timing (i.e., the worst negative slack, or WNS). The timing optimization is not degraded by power optimization. The power optimization is followed by area optimization. The use of multi-Vth libraries is recommended in the synthesis environment (using Power Compiler with Design Compiler or Physical Compiler) when optimizing for leakage power for either the one- or two-pass flow.

Market Approach

Electronic design automation (EDA) tools are extensively used throughout the design process, focusing on power management and optimization. Portable devices batteries have made significant progress 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 all of this alone. EDA tools focus on power savings and are generally divided into two categories: analysis tools focused on the actual layout of the IC, and RTL level and synthesis optimization at the more abstract end of the spectrum. The analysis tools do an 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 not a new feature in most tools, but the improvement at this level is only 5% to 10% and is dependent on the power modeling of intellectual property (IP), which has traditionally been poor in commodity free libraries. Wafer fabs have addressed the issue of power as well. Most fabs and independent foundries offer several process variations at 130nm and 90nm, each optimized for speed, power or voltage. However, all these processes are really offering designers is the classic power/speed trade-off. They can have a high-speed, high leakage process, or a low-speed, low-leakage process. The process performance operates on the fundamental physics of the semiconductor. 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. Low-power IP offers much the same proposition. Savings in power are frequently achieved through the use of higher threshold transistors and by designing circuits for reduced power, but both methods come with a speed penalty. This technique produces acceptable results in low-speed applications. The faster speed of the 130 nanometer process will offset the loss of performance of the low- power IP. This technique works well for current-generation wireless base-band applications (cellular, Bluetooth) but for next-generation wireless products with greater graphics, video capabilities and with high-speed wireless Internet connections, the performance trade-offs are unacceptable. In addition, the leakage current and dynamic power become so severe that the inherent increase in speed will not be able to compensate at 90nm and below. 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, which are dictated by the consumer. The consumer is also unwilling to accept an increase in the size of the product due to a larger battery. 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, high-performance applications go mobile, designers have no choice but to adopt the use of power management IP to control the increasingly severe dynamic power and leakage power issues.

Conclusions

As technology moves into deep nanometer arena, the impact on leakage power is becoming a critical design issue. Leakage power is also strongly affected by on-chip process, temperature and voltage variations. It has significant implications on IC performance, power management and reliability. In the past decade EDA tools are evolving towards complete power management solution, taking into effect various electro-thermal couplings between supply voltage, threshold voltage, frequency, and junction temperature. Electro-thermal tools can be applied for making various power/performance/reliability/cooling-solutions tradeoffs in leakage dominant nanometer technologies, and to optimize the performance of ICs. It can also be used to generate a reliability and thermally aware design environment. Many companies, in all areas of the design supply chain, are attacking the nanometer leakage power issue. While individually effective in their own target areas, current solutions only provide a partial solution to the problem.

REFERENCES

[1] Synopsys Technical Publications. www.synopsys.com
[2] Virtual Silicon - http://www.virtual-silicon.com
[3] V. De and S. Borkar, "Technology and Design Challenges for Low Power and High Performance," in *Proc. ISLPED*, 1999, pp. 163-168.
[4] K. Banerjee, S-C. Lin, A. Keshavarzi, S. Narendra, and V. De, "A Self-Consistent Junction Temperature Estimation Methodology for Nanometer Scale ICs with Implications for Performance and Thermal Management," in *IEDM Tech. Dig.*, 2003, pp. 887-890.
[5] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, V. De, "Parameter Variations and Impact on Circuits and Microarchitecture," *DAC*, 2003, pp. 338-342.
[6] S. Borkar, "Low-Power Design Challenges for the Decade," *ASP-DAC,* 2001, pp. 293-296.
[7] International Technology Roadmap for Semiconductors (ITRS), http://public.itrs.net
[8] P. P. Gelsinger, "Microprocessors for the New Millennium: Challenges, Opportunities, and New Frontiers," in *Proc. ISSCC,* 2001, pp. 22-25.
[9] P. Gelsinger, *41st DAC Keynote, Design Automation Conference (DAC), 2004.* (www.dac.com)
[10] www.intel.com
[11] R. Mahajan et al., "The Evolution of Microprocessor Packaging," *Intel Technology Journal 3rd quarter*, 2000.
[12] *Intel Pentium 4 Processor Thermal Design Guidelines*
[13] K. Nose, T. Sakurai, "Optimization of Vdd and Vth for Low Power and High Speed Applications," *Proc. ASP-DAC*, 2000, pp. 469-474.
[14] R. Gonzalez, B.M. Gordon, and M.A. Horowitz, "Supply and Threshold Voltage Scaling for Low Power CMOS," *IEEE Journal of Solid-State Circuits*, Vol. 32, pp.1210-1216, 1997.
[15] T. Sakurai, and A.R. Newton, "Alpha-Power Law MOSFET Model and its Applications to CMOS Inverter Delay and Other Formulas," *IEEE Journal of Solid-State Circuits*, Vol. 25, 1990, pp.584-594
[16] J. R. Black, "Electromigration—A Brief Survey and Some Recent Results," *IEEE Trans. Elec. Dev.*, vol. ED-16, pp. 338–347, 1969.
[17] C-K. Hu et al., "Scaling Effect on Electromigration in On-Chip Cu Wiring," in *Proc. IITC*, 1999, pp. 267-269.
[18] R. Blish, T. Dellin, S. Huber, M. Johnson, J. Maiz, B. Likins, N. Lycoudes, J. McPherson, Y. Peng, C. Peridier, A. Preussger, G. Prokop, and L. Tullos, "Critical Reliability Challenges for The Internatinonal Technology Roadmap for Semiconductors," *International Sematech Technology Transfer Document 03024377ATR*, 2003.
[19] A.M. Yassine, H.E. Nariman, M. McBride, M. Uzer, and K.R. Olasupo, "Time Dependent Breakdown of Ultra-Thin Gate Oxide," *IEEE Trans. Elec. Dev.*, Vol. 47, pp. 1416–1420, 2000.
[20] A. Basu, S-C. Lin, V. Wason, A. Mehrotra and K. Banerjee,

“Simultaneous Optimization of Supply and Threshold Voltages for Low-Power and High-Performance Circuits in the Leakage Dominant Era,” *Proc. DAC*, 2004, pp. 884-887.
[21] S-C. Lin, A. Basu, A. Keshavarzi, V. De and K. Banerjee, “Impact of Off-state Leakage Current on Electromigration Design Rules for Nanometer Scale CMOS Technologies,” *Proc.* IRPS, pp. 74-78, 2004.
[22] http://gigaic.com, Giga Scale’s InCyte product
[23] http://www.ggtcorp.com Golden Gate Technology’s GoPower
[24] K. Veendrick, “Short circuit dissociation of static CMOS circuitry and its impact on the design of buffer circuits., IEE Jorurnal of Solid-State Circuits vol. SC-19 pp. 468-473, Aug.1984
[25] K. Nose and T. Sakurai http://lowpower.iis.utokyo. ac.jp/SPN/1998/19980331.pdf
[26] K. Usami and M. Horowitz “Cluster Voltage Scaling Technique for Low Power Design, Proc. ISLPED, pp 3-9, 1995
[27] D. Sylvester and H. Kaul, “Future Performance Challenges in Nanometer Design” DAC 2001, June 18-22, 2001
[28] K. Usami *et al*., “Design Methodology of Ultra Low Power MPEG4 Codec Core Exploiting Voltage Scaling Techniques
[29] D. Lackey *et al*. “Managing power and performance for System-on-Chip designs using Voltage Islands”, ICCAD pp 195-202, 2002,
[30] Keshavarzi, et al, Intrinsic Leakage in Low-Power Deep Submicron CMOS ICs. ITC 1997
[31] S. Narendra *et al* “Leakage Issues in IC Design: Trends , Estimation and Avoidance”, Proc. ICAD 2003
[32] S. Oks, “Dual Vth Design Flow for High Performance Designs”, SNUG, San Jose 2001
[33] P. Butpa, A. Kahng, P. Shrama, D. Sylvester, “Selective Gate-Length Biasing for Cost-Effective Runtime Leakage Control. DAC 2004
[34] http://www.magma-da.com - Magma Design Automation
[35] http://www.easic.com