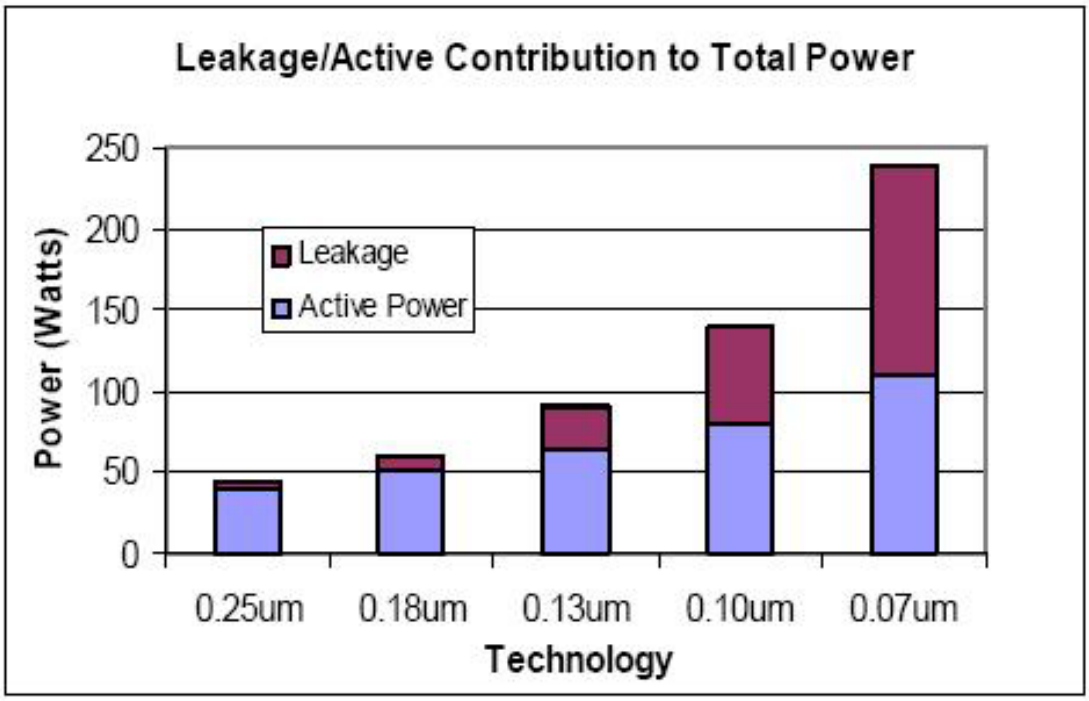
Challenges & Solutions in Physical Design for High-performance IC Design in Very-Deep-Sub-Micron (VDSM) Era

Dr. Danny Rittman, Jan 2004

Abstract: One year in IC Design Technology evolution is equivalent to 10 years in aviation progress. This fact was actually proven in 1995 by a team of researches from MIT by comparing new technologies from both fields. As fabrication technology moves to the nanometer range profound VDSM effects become critical in developing IC's with hundreds of millions of transistors. Recognizing the problems EDA vendors and foundries have introduced products and services directly aimed at simplifying both front-end and back-end chip design. Are we capable to meet the challenges of the nanometer world such as performance closure, power, reliability, manufacturability, and cost? What are the challenges and the solutions in the VDSM physical design arena? Can we maintain productivity within these new constraints? The future of design technology is most uncertain in the physical implementation space, where many complexities must be managed without sacrificing system value or turnaround time. In this article we'll discuss the arising issues and solutions in the high performance VDSM physical design field.



Background – High Performance VDSM physical design

As process technology shrinks, designers are able to place larger numbers of transistors on single silicon chip. However shrinking process technology increases the chip complexity and creates more complicated design verification tasks. With the trend toward 65-nanometer process technologies, designers face major barriers arising from increased electrical and physical effects associated with dense interconnects and closely packed transistors. Already, advanced designs at 90 nm and below exhibit increasing non-digital behaviors such as dynamic IR drop, leakage current, electromigration, antenna and cross-coupling effects that can erode timing margins, introduce reliability problems and lead to circuit failure. The problems become significant and even dominant with high performance designs. The main issues of VDSM high performance physical designs are current density and power distribution, synchronization, manufacturing variability, high-frequency and coupling noise. Dynamic analysis techniques are the best approach to understand and resolve these challenging VDSM problems.

Key Challenges in Power Design:

High-Performance Design = Low-Power Design

- Reliable Power Distribution
- Efficient Power Management
- Power Heat Dissipation (Packaging)

Reliable Power Distribution

Power distribution is one of the major issues with VDSM high performance designs. Grid array and Flip-chip packaging allows distribution of Vdd/GND and signals throughout a die, rather than just at the periphery. This increased flexibility makes power grid IR drops substantially more manageable. However, current ITRS projections for power/grid pad connectivity in nanometer designs do not fully take advantage of grid array capabilities and lead to power distribution problems.

Hot-spots are considered since uniform power density assumptions are overly optimistic. A hot-spot is defined to have a localized power density four times larger than a uniform power density approximation.

(Specified by Pchip / Achip)
Let's take a look at IR scaling trends graph. (Figure 1)

Figure 1 - Shows the necessary power rail width (normalized to minimum top-level metal width) to ensure <10% IR drop in "hot-spots" of a design in scaled technologies using the minimum permissible bump pitch. This figure considers top-level routing only, assuming that the remainder of the power grid is under the designers control whereas the top-level granularity is technology-limited. As seen from the graph 35 nm is less restricted than 50 nm due to a reduction in power density at 35 nm. In general, while the trend seems alarming (approximately quadratic increase in power rail linewidth, normalized to minimum allowable linewidth), even 35 nm results are manageable, in that Vdd and GND rails that are 16X minimum width will consume less than 4% of top-level routing resources (based on 80 micron bump and power-grid pitch). The total routing resources consumed due to power routing is around 17-20% as a constant factor of 16% is used to reflect the need for large metal "landing pads" for the bumps. The continued reductions in bump pitch allow Vdd/GND to be supplied at finer granularities where it is most needed.

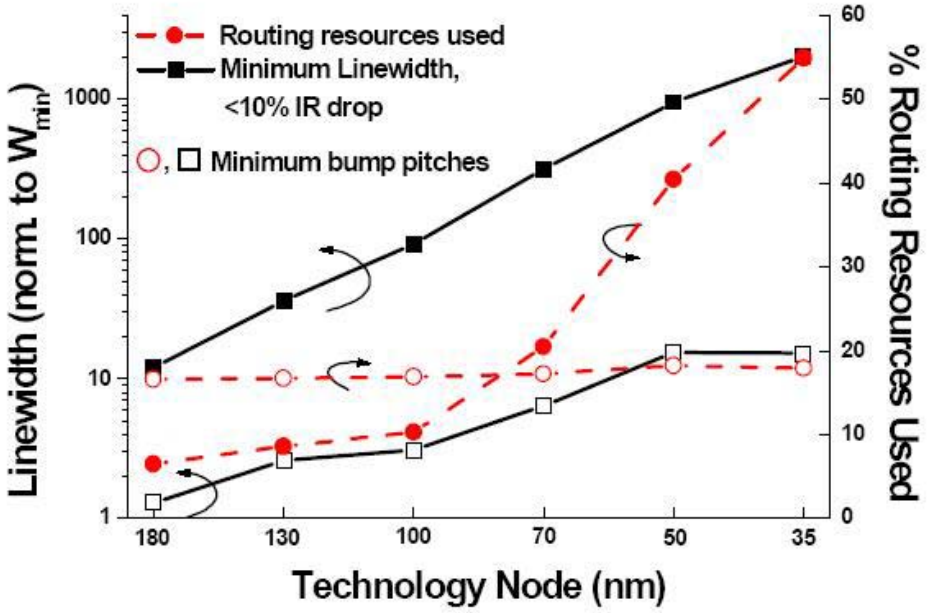


Figure 1 -IR drop scaling trends based on minimum allowable bump pitcl (open symbols) and ITRS bump/pad count projections (solid symbols).

Efficient Power Management

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.

Let’s take a look at the necessary steps for efficient power management.

- Power efficient physical layouts are a must for high performance VDSM designs. Power rails optimization is needed to avoid deadly voltage drop and current density violations.
- Accurate timing is needed! EDA vendors provide a nice set of tools to cover this issue.
- Power pad placement optimization is required.

Placement; Power-Driven

Power-driven placement reduces the loading effects on wires that have high-frequency switching activity. An optimized power placement minimizes the interconnect length between driving gates and loads. This optimization results in significant dynamic power savings on nets with high switching activity.

Placement; Rail Driven & Thermal

Rail-driven placement is essential to minimize voltage (IR) drop along the power and ground rails. This type of placement is important so that the current demand from high power cells is spread evenly across

the power grid --- avoiding "hot spots" in the design. This effectively minimizes IR drop, current density and temperature problems.

With each new process generation, power supply voltage levels decrease to meet the increasing need for power-efficient designs. Coupled with decreasing feature sizes and increasing circuit densities, voltage drop levels that were acceptable in a previous generation now cause failures, and temperature gradients cause significant performance degradation.

1. Dynamic Power Analysis

1.1 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.

1.2. 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.

1.3. 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.

1.4. 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.

2. Static Power Analysis

2.1 Multiple V_{th} 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 [37]. 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.

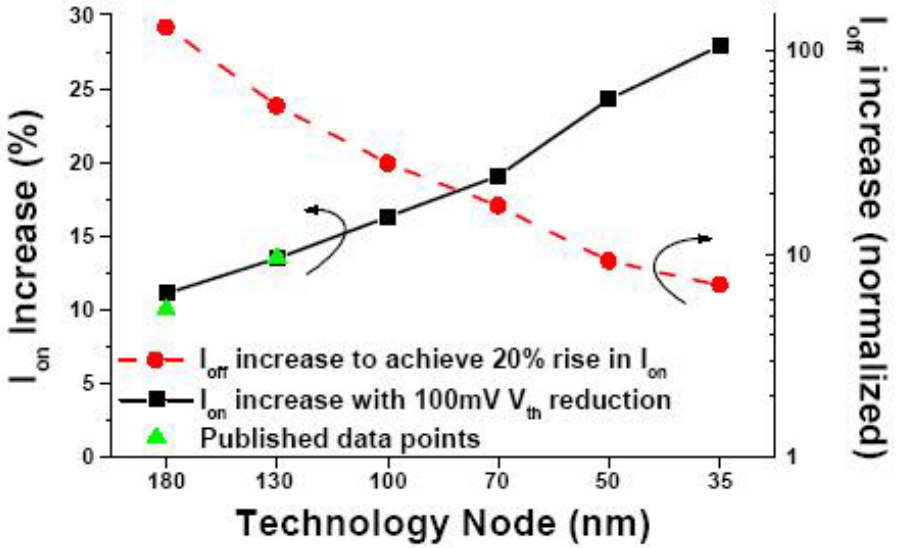


Figure 2. I_{on} increases more rapidly with a 100mV change in V_{th} for scaled technologies. I_{off} penalty for 20% I_{on} gain reduces with scaling.

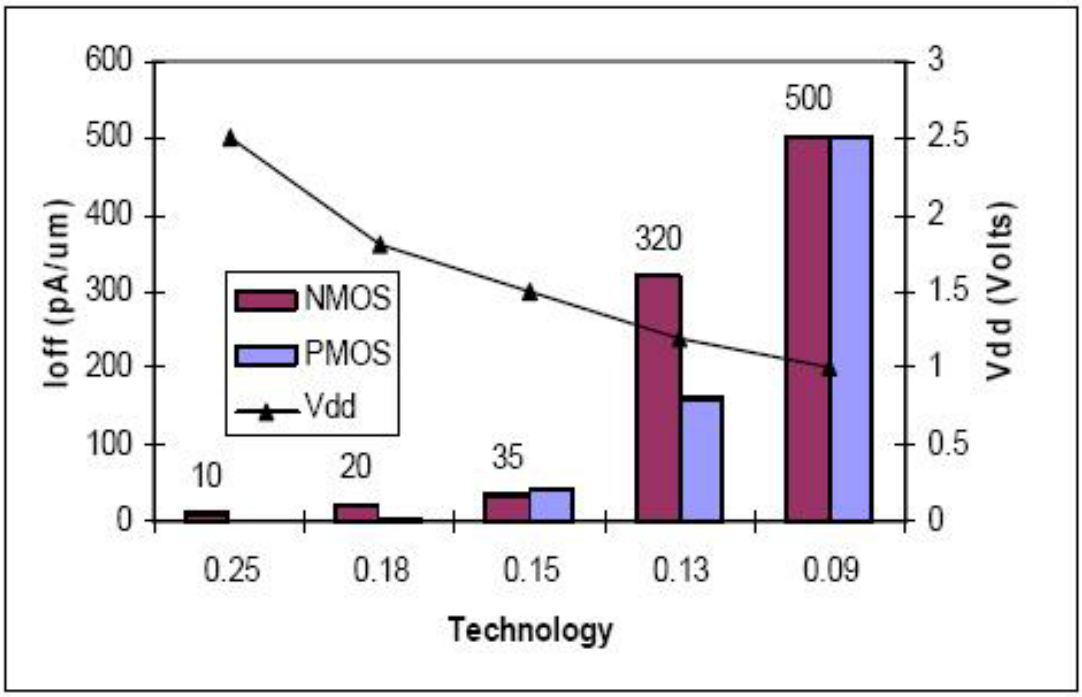


Figure 3 - Ioff reduction with the process and power

2.2. Scalable Dynamic/Static Power Approach

One of the most appealing approach, 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.

CONCLUSIONS

The EDA industry continues to face new challenges as process continues to shrink into deep-submicron 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 leader EDA tools companies like Cadence, Synopsys and Mentor Graphics provide a nice set of tools to solve these issues. Nevertheless in many cases of high performance designs current EDA technology does not have the full power to provide the best solution. Next are the conclusions of this paper:

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) is proposed, 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.

REFERENCES

[1] J.T. Kao and A.P. Chandrakasan, "Dual-threshold voltage techniques for lowpower digital circuits," IEEE J. Solid-State Circ., pp. 1009-1018, Jul. 2000.
[2] T. Kuroda, et al., "A 0.9V, 150MHz, 10mW, 4mm2, 2-DCT core processor with variable VT scheme," IEEE J. Solid-State Circ., pp. 1770-1778, Nov. 1996.
[3] H. Kawaguchi, et al., "A CMOS scheme for 0.5V supply voltage with picoampere standby current," Proc. ISSCC, pp. 192-193, 1998.
[4] M.C. Johnson, et al., "Leakage control with efficient use of transistor stacks in single threshold CMOS," Proc. DAC, pp. 442-445, 1999.
[5] L. Wei, et al., "Design and optimization of dual-threshold circuits for lowvoltage low-power applications," IEEE T. VLSI Sys, pp. 16-24, Mar. 1999.
[6] S. Tyagi, et al., "A 130nm generation logic technology featuring 70nm transistors, dual-Vt transistors and 6 layers of Cu interconnects," Proc. IEDM, pp. 567 - 570, 2000.
[7] http://www-device.eecs.berkeley.edu/~dennis/BACPAC, see also: http://vlsicad.cs.ucla.edu/GSRC/GTX
[8] J.M. Musicer and J. Rabaey, "MOS current mode logic for low power, low noise CORDIC computation in mixed-signal environments," Proc. ISLPED, pp. 102-107, 2000.
[9] D. Sylvester and K. Keutzer, "Getting to the bottom of deep submicron II: A global wiring paradigm," Proc. ISPD, pp. 193-201, 1999.
[10] D. Sylvester and K. Kaul, "Future Performance Challenges in Nanometer Design,", 2001.
[11] R. Ho, K. Mai, H. Kapadia, and M. Horowitz, "Interconnect scaling implications for CAD," Proc. ICCAD, pp. 425-429, 1999.
[12] R. McInerney, et al, "Methodology for repeater insertion in the Itanium microprocessor," Proc. ISPD, pp. 99-104, 2000.
[13] H. Zhang, et al., "Low-swing on-chip signaling techniques: effectiveness and robustness," IEEE Trans. VLSI Systems, pp. 264-272, Jun. 2000.
[14] Y. Massoud, et al., "Differential signaling in crosstalk avoidance strategies for physical synthesis," Proc. TAU, 2000.
[15] D.G. Chinnery and K. Keutzer, "Closing the gap between ASIC and custom: an ASIC perspective," Proc. DAC, pp. 637-641, 2000.
[16] W.J. Dally and A. Chang, "The role of custom design in ASIC chips," Proc. DAC, pp. 643-647, 2000.
[17] IBM SA-27E ASIC standard cell datasheet.
[18] P. Hurat, "Beyond physical synthesis," SNUG Europe 2001.

[19] K. Usami, et al., "Automated low-power technique exploiting multiple supply voltages applied to a media processor," IEEE J. Solid-State Circ., pp. 463-472, Mar. 1998.
[20] M. Takahashi, et al., "A 60-mW MPEG4 video codec using clustered voltage scaling with variable supply-voltage scheme," IEEE J. Solid-State Circ., pp. 1772-1780, Nov. 1998.
[21] K. Usami and M. Horowitz, "Cluster voltage scaling technique for low power design," Proc. ISLPED, pp. 3-8, 1995.
[22] C. Akrout, et al, "A 480MHz RISC microprocessor in a 0.12μm Leff CMOS technology with copper interconnects," IEEE J. Solid-State Circ., pp. 1609- 1616, Nov. 1998.
[23] S. Sirichotiyakul, et al., "Standby power minimization through simultaneous threshold voltage and circuit sizing," Proc. DAC, pp. 436-441, 1999.
[24] S. Borkar, "Design challenges of technology scaling," IEEE Micro, pp. 23-29, Jul-Aug 1999.
[25] R. Chau, et al., "30nm physical gate length CMOS transistors with 1.0ps NMOS and 1.7ps PMOS gate delays," Proc. IEDM, pp. 45-48, 2000.
[26] S. Song, et al., "CMOS device scaling beyond 100nm," Proc. IEDM, pp. 235-238, 2000.
[27] H. Wakabayashi, et al., "45-nm gate length CMOS technology and beyond using steep halo," Proc. IEDM, pp. 49-52, 2000.
[28] M.Mehrotra, et al., "A 1.2V, sub-0.09μm gate length CMOS technology," Proc. IEDM, pp. 419-422, 1999.
[29] I.Y. Yang, et al., "Sub-60nm physical gate length SOI CMOS," Proc. IEDM, pp. 431-434, 1999.