

Low-Leakage Repeaters for NoC Interconnects

Arkadiy Morgenshtein, Israel Cidon, Avinoam Kolodny, Ran Ginosar

Electrical Engineering Department, Technion – Israel Institute of Technology, Haifa, Israel
arkadiy@tx.technion.ac.il

Abstract - Several low-leakage repeater circuits for Network-on-Chip (NoC) interconnects are presented and analyzed for various utilization rates. The recently proposed Staggered-Vt (SVT) repeater is compared with novel Dual-Vt Domino (DTD) repeaters and Sleep Repeaters (SR). These circuits are compared with standard Low-Vt (LVT) repeaters in a 32-bit link. Up to 70% and 61% power reduction was obtained in SVT and DTD repeaters, respectively. DTD repeaters are the most area-efficient ones, showing 40% reduction in total area of repeaters. Sleep Repeaters are most area-consuming and less effective in high and moderate utilization rates, but comparable to SVT in terms of power for utilization rates below 2%, showing 72% power reduction.

I. INTRODUCTION

Large Systems-on-Chip (SoC) can employ packet-switched Networks-on-Chip (NoC) [1]. NoC is based on module connection via a network of routers. The links between routers contain long wire interconnects. Interconnect optimization has become one of the major design considerations in state-of-the-art systems. Timing optimization of global wires is typically performed by repeater insertion. However, the usage of repeaters implies a significant cost in area and power. Recent studies indicate that in the near future a large portion of chip resources will be used by inverters operating as repeaters [2]. This dictates a need for power optimization of the global links with repeaters. Many previous works investigated techniques for reducing dynamic power, as it was dominant in older fabrication technologies [3][4][10][11][12]. One approach to dynamic power reduction in repeaters is activity reduction by encoding [5][6]. Other techniques include wire spacing and using fewer and smaller repeaters.

In modern nanometer CMOS, leakage current per device grows dramatically with technology scaling. Thus, the leakage portion of power becomes dominant, especially for circuits with reduced activity rate. In NoC the link utilization rates vary and in many cases are very low, reaching a few percents [7]. Networks are designed to operate at low link utilization in order to meet stringent latency requirements, and link over-capacity helps reduce packet collisions. However, when NoC links are idle they still consume leakage power in repeaters.

Leakage current reduction in CMOS circuits is an issue of numerous studies. The main techniques that were proposed are MTCMOS [8] and Dual-Vt [9] allowing a significant power reduction when used in logic circuits. However, the nature of lumped logic circuits is different from that of distributed links. The specifics that make leakage minimization in repeaters unique and challenging are: (a) large device sizes, (b) no transistor stacks, and (c) extremely high output loads of the inter-device wires.

Previous works on power minimization in repeater insertion for timing optimization [10][11][12] focus on sizes and numbers of regular inverter-based repeaters without modification of the circuit structure. Others [13][14] propose different structures of circuits regardless of reduction in leakage power. The timing optimization concept in [15] is based on early transition detection for total power reduction, using

complex receivers as repeaters. Recent research [16] proposes a new design approach of staggered threshold voltage (SVT) buffers with selective use of high-threshold transistors for power minimization. SVT buffers are based on inverters combining high-Vt and low-Vt transistors as shown in Figure 1. During operation in standby state the active devices are all low-Vt while the off-state devices are high-Vt with lower leakage. SVT technique is effective both in standby and in active mode, when the data is encoded so that most symbols incur minimal leakage. Although power was reduced with SVT buffers, the basic inverter-based structure remained unchanged. The area of SVT repeaters increased because of sizing of high-threshold devices.

Repeater design techniques have to be characterized for various link utilization rates. Average activity rate of micro-processor nets is about 4.5% [17], while some links may exhibit high utilization rates approaching 100%. A single Network-on-chip can contain numerous link types with various utilization rates and area resources [7]. Thus, different optimization and design techniques can be applied to different links at the same chip depending on the utilization and area characteristics of each link. This paper presents novel techniques for leakage minimization in NoC links and utilization and area comparisons of these and other design techniques.

We propose a new Dual-Vt Domino (DTD) repeater circuit which combines significant leakage power reduction with area reduction. We also present new Sleep Repeaters (SR) based on Multi-Threshold design [20], showing best leakage reduction at low utilization rates. The new circuits are compared with SVT buffers and low-threshold (LVT) repeaters under various utilization conditions of NoC links.

II. LOW-VT AND STAGGERED-VT REPEATERS

The novel techniques in this paper are compared to the existing alternatives presented in this section. Common high-performance repeaters employ Low-Vt (LVT), where all repeaters have low threshold voltage for the purpose of speed. However, low Vt causes high leakage current. Replacing LVT repeaters with high threshold voltage (HVT) repeaters provides significant leakage saving, but leads to performance degradation.

The recently proposed alternative is Staggered-Vt (SVT) repeaters [16]. The basic structure of the link with SVT repeaters is shown in Figure 1. Links using SVT buffers employ alternating inverters with high-Vt and low-Vt transistors. During most of the operation time the low-Vt devices are active while the high-Vt devices are in off-state with lower leakage. This is achieved by encoding the input signals to states that result in lowest leakage. Although power is reduced using SVT buffers, the basic inverter-based structure remains unchanged. Thus, in order to meet the delay target, the area of SVT repeaters is increased because of sizing of high-threshold devices.



Figure 1. SVT in 0-state low leakage [16]

III. DUAL-VT DOMINO (DTD) REPEATERS

The new Dual-Vt Domino (DTD) technique presented in this section offers a low-area alternative to SVT. The basic structure of the DTD repeaters is presented in Figure 2. Similar to SVT, the DTD repeaters are comprised of transistors with both high and low threshold voltages. However, instead of using regular inverters, DTD repeaters operate in domino structure. Data is applied to high-Vt transistors dedicated for *evaluation*. Clock is used for setting the standby state on the line by driving low-Vt transistors during the *pre-charge* phase.

The standby state of all the nodes in the repeater chain is known and can be predefined, similar to [18]. Standby state can be set along the line in which the leakage currents in the repeaters will be minimal, as it is done in SVT. To achieve this, the disconnected transistors in standby should have high Vt. As the logic chain contains only inverters, the logical values in the standby state will be mutually exclusive in each pair of adjacent nodes.

The standby state of the Data and the Clock at the input to the link is '0'. This state causes a logical pattern of '1010...' in the wire segments, so that the "off" transistors are the high-Vt ones and leakage is minimized. When Data signal arrives it can be 0→1→0 pulse, or 0→0→0 (when no transition occurs). The Clock starts slightly before and performs a 0→1→0 transition. When Clock rises the pre-charge transistors are disconnected and the evaluation transistors drive the data over the line. Note that a certain time margin is maintained to avoid short circuit current in case of (Data=1, Clock=0). This setup time period should be defined by worst-case skew scenario along the link due to delay uncertainty.

The wires and segmentation by repeaters are identical in the Data and Clock lines, so that the pulses are simultaneous and the propagation along the link is synchronized. In this way we assure that only one repeater stage is active at a time while the others are in standby mode.

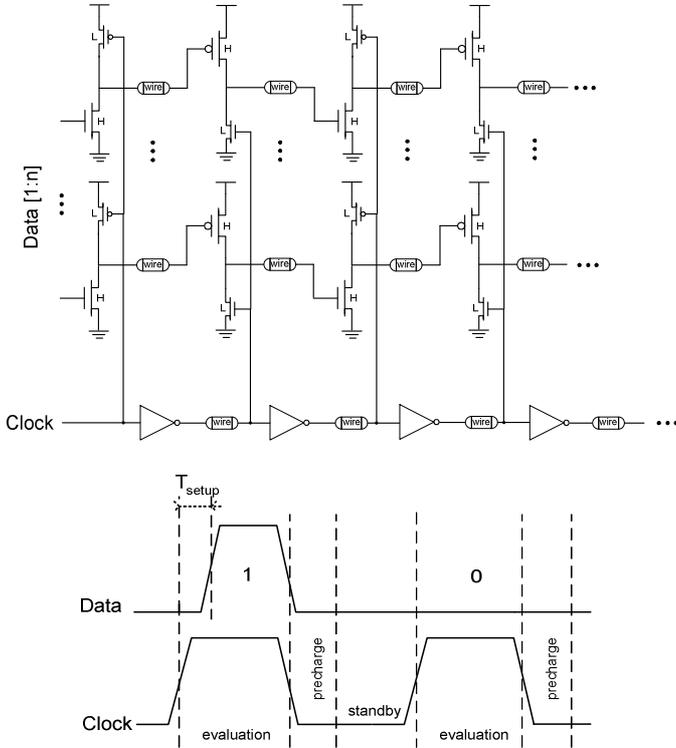


Figure 2. DTD repeaters and time diagram.

The Clock signals are assumed to be associated with the Data – no clock is transmitted over the link when no data exist. This can be easily achieved by either clock gating or local clock generation.

The low-leakage domino operation involves a penalty in dynamic power consumption of DTD because of additional signaling during two-phase operation. However, DTD repeaters require less area than LVT and SVT. Data transportation in DTD is performed while only one transistor gate of each pair is connected to the wire. This is different than in standard inverter-based repeaters used in LVT and SVT, where gates of both transistors are constantly connected to the interconnect. This means that each repeater drives a smaller load. This results in reduction of sizing factor of repeaters needed for timing minimization. In this work it allowed $\times 0.6$ scaling of the Data transistors in DTD to meet the delay target of LVT.

An additional characteristic of DTD is the use of high-threshold evaluation transistors in the critical path of evaluation. In LVT and in SVT mixed-Vt and low-Vt transistors are used for fast data evaluation, respectively. This makes the devices vulnerable to Vt variations. Although high-Vt devices in DTD are slower and are scaled to meet the delay target, they have increased immunity to Vt variations and to noise which, in turn, contributes to reduced delay uncertainty.

Note that in this configuration the Data link has to be accompanied by Clock signal to control the pre-charge. The parallel links in NoC typically include this additional signal for synchronization purposes anyway. In this paper a parallel link is investigated and is accompanied by a separate Clock line. In this case the repeaters drive a high load of multiple pre-charge transistors in addition to wires. Thus, the repeaters in the Clock line are scaled to target the required delay.

IV. SLEEP REPEATERS (SR)

The alternative to Dual-Vt circuits is the Multi-threshold (MT) design based on the addition of high-Vt sleep transistors to low-Vt logic structures for reduced leakage. MTCMOS is proposed in numerous works for power reduction in CMOS logic [19]. Repeaters, however, are different from logic circuits because of the increased sizes of the devices, minimal transistor stacks and much higher loads of the wire segments. Thus, in this section we investigate the application of sleep transistors to repeaters, combining the Sleep Repeaters (ST) configuration.

The analysis of sleep transistors application in global interconnect repeaters is illustrated by Figure 3. At first stage common sleep transistors insertion is assumed, while both NMOS and PMOS are used to achieve more effective leakage reduction for both pull-up and pull-down transistors in "off" state.

However, due to the long delays of the global link, it is beneficial to distribute the sleep transistors along the link so that each stage of repeaters has a separate pair of sleep transistors controlled by signals from the adjusted Clock line. In this manner, assuming a similar propagation delay of the Data and the Clock, leakage is minimized. Only one stage of repeaters is active while the others are in low-leakage standby.

As opposed to lumped logic, after exiting the standby state, the sleep transistor in the repeater is loaded by highly capacitive interconnect segments. In order to maintain the delay target, the sleep transistors have to be scaled by a factor equal to number of Data bits in the link, so that the final scaling factor of the sleep transistor can easily be over $\times 1000$. These numbers are unrealistic for implementation in terms of area and dynamic power, and the cascades needed to drive the sleep transistors. Problems related to routing complexity and wiring resources in case of a sleep transistor common to multiple repeaters also have to be accounted.

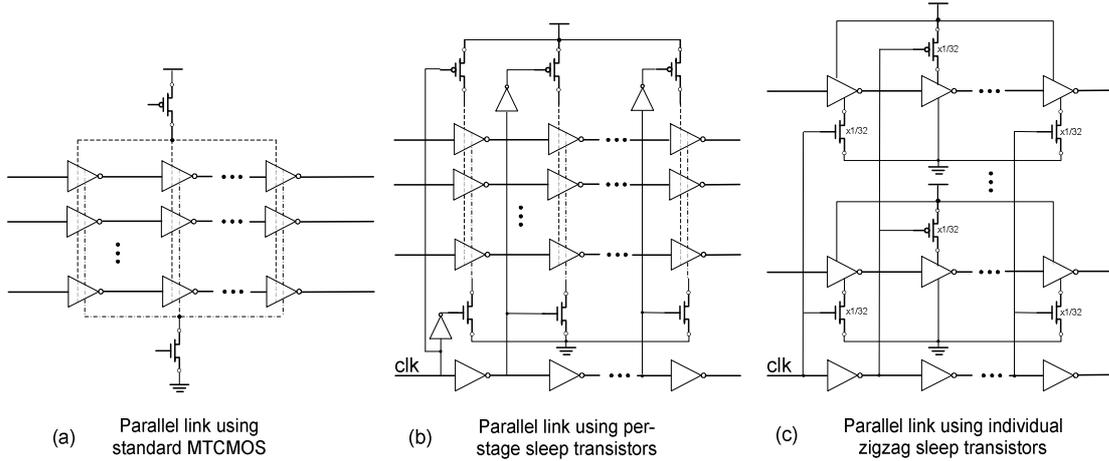


Figure 3. Sleep Repeaters evolution – from MTCMOS to SR.

The principal benefit is the application of smaller individual sleep transistors for each repeater with simpler routing. Additional benefit can be derived from the mutually exclusive pattern of the repeaters if the sleep transistors are connected in zigzag manner similar to [20], using fewer transistors than the original approach.

Multi-Threshold technique has been proven as very efficient in leakage power reduction [19], and with the proposed Sleep Repeaters methodology it can be efficiently used in links with repeaters. It should be noted that as compared to the DTD technique, the area overhead and the dynamic power consumption at high utilization rates are expected to be higher for SR repeaters due to the addition of sleep transistors.

V. COMPARATIVE SIMULATIONS

In order to complete the characterization, the proposed techniques were compared with regular high-performance LVT repeaters and with the low-leakage alternative of SVT buffers.

The comparison was based on circuit simulations using 65nm BPTM models for transistors and interconnect [22]. For HVT devices the V_t was 0.1902V and -0.213V for NMOS and PMOS respectively, while for LVT the values were 0.1502V and -0.183V. The transient simulations were performed for a 32-bit link operating at 105°C temperature. Several cases of link length and repeater insertion were investigated and values of area, delay and energy consumption were obtained for each of the compared techniques. Simulations were performed for several link utilization rates varying from 100% to 0.2%, as shown in Table 1.

LVT design was used as baseline for repeater insertion. The number of repeaters was derived according to [21], while scaling of LVT was modified to minimize the power [11][23]:

$$k_{rep} = \sqrt{\frac{0.4 \cdot C_{int} R_{int} \cdot L^2}{0.7 \cdot R_{inv} C_{inv}}}, \quad h_{rep} = a \cdot \sqrt{\frac{C_{int} R_{int}}{C_{inv} R_{int}}} \quad (1)$$

where k is count and h is scaling factor of devices with combined coefficient $a=0.29$ for reduced power consumption. C_{inv} and R_{inv} are the input capacitance and resistance of minimal-size inverter and C_{int} and R_{int} are the wire capacitance and resistance per unit of length. The number of repeaters was the same in all the techniques, while the scaling factor was adjusted for SVT, DTD and SR to meet the delay target equal to LVT. Additional simulations were performed for a reduced number of repeaters in order to characterize the techniques for longer wire segment.

The results of comparison can be seen in Table 1. A significant reduction of leakage power manifests itself at low utilization rates, where leakage power is dominant. At high utilization rates SVT shows the best results with power reduction even in 100% utilization case. The power saving in SVT reaches up to 71% for minimal utilization. A 15% penalty

is observed in area as compared to LVT in 8mm link repeaters as can be seen in Figure 4. The area penalty was caused by sizing of high- V_t transistors driving long interconnect segments.

The DTD technique also shows a significant power saving reaching up-to 61% of leakage power reduction. Though the SVT repeaters are more power-efficient than DTD, the DTD is much more area-efficient while being 40% smaller than LVT and up-to 48% smaller than SVT.

The SR technique consumes the biggest area due to addition of sleep transistors. The SR technique is the most effective in leakage power minimization, which is seen at low utilization rates. Total power saving in SR reaches 72% for minimal utilization rate, as shown in example of a 8mm link in Figure 5. SR repeaters consume more power at high utilization rates, but the characteristics improve significantly as the utilization reduces leading to break-even points with other techniques at about 2% utilization.

		Power - absolute [mW] & (normalized to LVT)				
		10mm link 8 repeaters D = 820ps	5mm link 4 repeaters D = 380ps	5mm link 2 repeaters D = 340ps	8mm link 2 repeaters D = 540ps	
Utilization Rate of the Link	100%	DTD	18.50 (1.08)	8.58 (1.03)	7.67 (1.03)	12.00 (1.03)
		SR	22.17 (1.30)	10.83 (1.30)	8.58 (1.16)	13.00 (1.12)
		SVT	16.00 (0.94)	7.83 (0.94)	7.08 (0.95)	11.30 (0.97)
		LVT	17.08 (1.00)	8.33 (1.00)	7.42 (1.00)	11.60 (1.00)
	60%	DTD	10.85 (1.00)	5.25 (0.99)	4.70 (1.02)	7.30 (1.03)
		SR	13.45 (1.25)	6.55 (1.24)	5.20 (1.13)	7.90 (1.11)
		SVT	9.75 (0.90)	4.80 (0.91)	4.30 (0.93)	6.80 (0.96)
		LVT	10.80 (1.00)	5.30 (1.00)	4.60 (1.00)	7.10 (1.00)
	20%	DTD	3.97 (0.88)	1.92 (0.86)	1.65 (0.94)	2.50 (0.96)
		SR	4.73 (1.04)	2.32 (1.05)	1.80 (1.03)	2.70 (1.04)
		SVT	3.50 (0.77)	1.72 (0.77)	1.50 (0.86)	2.40 (0.92)
		LVT	4.53 (1.00)	2.22 (1.00)	1.75 (1.00)	2.60 (1.00)
10%	DTD	2.24 (0.76)	1.12 (0.77)	0.89 (0.85)	1.30 (0.87)	
	SR	2.54 (0.86)	1.24 (0.85)	0.94 (0.90)	1.40 (0.93)	
	SVT	1.95 (0.66)	0.96 (0.66)	0.80 (0.76)	1.20 (0.80)	
	LVT	2.96 (1.00)	1.46 (1.00)	1.05 (1.00)	1.50 (1.00)	
2%	DTD	0.87 (0.51)	0.43 (0.51)	0.28 (0.57)	0.40 (0.67)	
	SR	0.80 (0.47)	0.39 (0.46)	0.26 (0.53)	0.30 (0.50)	
	SVT	0.70 (0.41)	0.35 (0.41)	0.24 (0.49)	0.40 (0.67)	
	LVT	1.70 (1.00)	0.85 (1.00)	0.49 (1.00)	0.60 (1.00)	
0.2%	DTD	0.56 (0.39)	0.28 (0.39)	0.15 (0.42)	0.15 (0.41)	
	SR	0.40 (0.28)	0.20 (0.28)	0.11 (0.31)	0.12 (0.32)	
	SVT	0.42 (0.30)	0.21 (0.30)	0.11 (0.31)	0.15 (0.41)	
	LVT	1.42 (1.00)	0.71 (1.00)	0.36 (1.00)	0.37 (1.00)	

Table 1. Simulation results for various lengths and utilization rates in 65nm technology: normalized values are in parentheses.

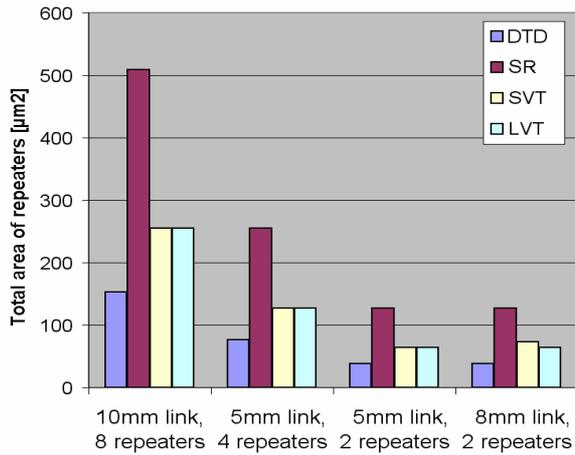


Figure 4. Total repeater area comparison.

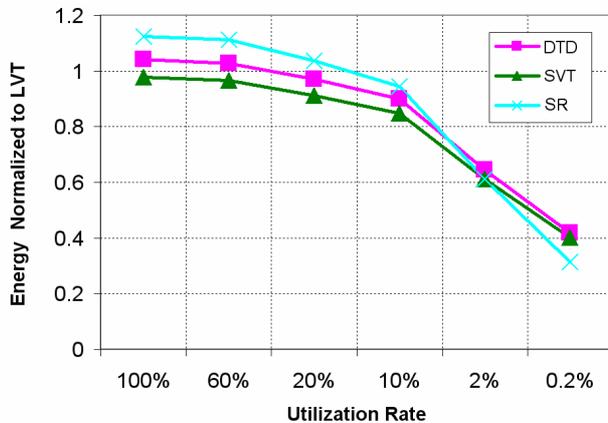


Figure 5. Energy comparison for 8mm link.

VI. DISCUSSION AND SUMMARY

Comparative analysis of interconnects in NoC revealed significant contribution of low-leakage repeaters to power minimization in NoC as compared to standard LVT repeaters, showing up-to 71% power reduction. The best power reduction is achieved at low utilization rates, which are typical for most of NoC links [7]. New DTD and SR repeaters were proposed as design alternatives for low-power links. The presented results allow the designer of NoC interconnects to choose the most suitable repeaters according to expected utilization and area parameters of a given link.

The SVT repeaters showed the most significant power reduction as compared to LVT repeaters for the widest range of utilization rates. However, SVT consumes more area. SVT repeaters will be effective in interconnects with variable utilization rates and in generic IP cores for parallel links without tight area limitations.

Novel DTD technique was developed introducing a change in the basic structure of repeaters. DTD shows significant power reduction with the best area characteristics, allowing a 40% area reduction. The technique suffers a penalty of dynamic power consumption due to signaling in domino protocol as compared to SVT and LVT. The design is more complex than SVT due to adding a Clock signal for pre-charge. However, thanks to significant leakage reduction and superior area characteristics, the DTD is candidate for design of NoC interconnects with limited area resources.

SR technique was optimized and adapted for repeaters in global interconnects. Although the addition of sleep transistors leads to increased area and dynamic power consumption at high utilization, SR

repeaters proved to be equivalent to SVT in leakage minimization at ultra low utilization rates below 2% with up-to 72% power reduction. This makes it a viable alternative for low-utilization links in NoC.

REFERENCES

- [1] W.J. Dally, B. Towles, "Route Packets, Not Wires: On-Chip Interconnection Networks", *DAC Conference*, pp. 684-689, 2001.
- [2] P. Saxena, M. Menezes, P. Cocchini, D.A. Kirkpatrick, "Repeater scaling and its impact on CAD", *IEEE Tran. CAD*, pp. 451 - 463, April, 2004.
- [3] V. Adler, E. G. Friedman, "Repeater Design to Reduce Delay and Power in Resistive Interconnect", *IEEE Tran. CAS*, no. 5, pp. 607-616, May 1998.
- [4] J. Lillis, C.K. Cheng, T.T.Y. Lin, "Optimal Wire Sizing and Buffer Insertion for Low Power and a Generalized Delay Model," *Proc. IEEE Int'l. Conf. on CAD*, pp. 138-143, Nov. 1995.
- [5] P.P. Sotiriadis, A.P. Chandrakasan, "Bus energy reduction by transition pattern coding using a detailed deep submicrometer bus model", *IEEE Tran. CAS*, pp. 1280 - 1295, Oct. 2003.
- [6] L. Macchiarulo, E. Macii, M. Poncino, "Low-energy encoding for deep-submicron address buses", *ISLPED Conf.*, pp. 176 - 181, Aug. 2001.
- [7] E. Bolotin, I. Cidon, R. Ginosar and A. Kolodny, "QNoC: QoS Architecture and Design Process for Network on Chip", *J. Systems Architecture*, vol. 50, pp. 105-128, Feb. 2004.
- [8] S. Mutoh, T. Douseki, Y. Matsuya, T. Aoki, S. Shigematsu, and J. Yamada, "1-V Power Supply High-Speed Digital Circuit Technology with Multi-Threshold Voltage CMOS", *IEEE J. Solid-State Circuits*, vol. 30, pp. 847-854, Aug. 1995.
- [9] L. Wei, Z. Chen, M. Johnson, K. Roy, Y. Ye, and V. De, "Design and Optimization of Dual Threshold Circuits for Low-Power Operation of CMOS Circuits", *IEEE Tran. VLSI Syst.*, pp. 16-24, March, 1999.
- [10] K. Banerjee, A. Mehrotra, "A Power-Optimal Repeater Insertion Methodology for Global Interconnects in Nanometer Designs", *IEEE Tran. Electron Devices*, vol. 49, pp. 2001-2007, November 2002.
- [11] G. Chen and E. G. Friedman, "Low Power Repeaters Driving RC Interconnect with Delay and Bandwidth Constraints," *Proceedings of the IEEE International SOC Conference*, pp. 335-339, Sep. 2004.
- [12] P. Kapur, G. Chandra, K.C. Saraswat, "Power Estimation in Global Interconnects & its Reduction Using a Novel Optimization Methodology", *DAC Conference*, pp. 461-466, 2002.
- [13] I. Dobbelaere, M. Horowitz, A.E. Gamal, "Regenerative Feedback Repeaters for Programmable Interconnections", *IEEE J. Solid-State Circuits*, vol. 30, pp. 1246-1253, Nov., 1995.
- [14] A. Nalamalpu, S. Srinivisan, W.P. Bursleson, "Boosters for Driving Long On-chip Interconnects - Design Issues, Interconnect Synthesis, Comparison with Repeaters", *IEEE Tran. CAD*, vol. 21, pp. 50-62, Jan. 2002.
- [15] H. Kaul, D. Sylvester, "Low-Power On-Chip Communication Based on Transition-Aware Global Signaling (TAGS)", *IEEE Tran. VLSI Syst.*, vol. 12, pp. 464-476, May, 2004.
- [16] H. Deogun, R. Rao, D. Sylvester, D. Blaauw, "Leakage- and Crosstalk-Aware Bus Encoding for Total Power Reduction". *IEEE/ACM Design Automation Conf.*, 779-782, June, 2004.
- [17] N. Magen, A. Kolodny, U. Weiser, N. Shamir, "Interconnect-power dissipation in a microprocessor", *SLIP Conf.*, Feb., 2004.
- [18] H. J. Yoo, "Dual-Vt Self-Timed CMOS Logic for Low Subthreshold Current Multigigabit Synchronous DRAM", *IEEE Tran. Circuits and Systems*, vol. 45, no. 9, pp. 1263-1271, Sep. 1998.
- [19] J.T. Kao, A.P. Chandrakasan, "Dual-Threshold Voltage Techniques for Low-Power Digital Circuits", *IEEE J. Solid-State Circuits*, vol. 35, pp. 1009-1018, July, 2000.
- [20] J.C. Park, V.J. Mooney III, P. Pfeifferberger, "Sleepy Stack Reduction of Leakage Power", *PATMOS*, pp. 148-158, 2004.
- [21] H.B. Bakoglu, "Circuits, Interconnections and Packaging for VLSI", Addison-Wesley, pp. 194-219, 1990.
- [22] Berkeley Predictive Technology Model (BPTM), www.device.eecs.berkeley.edu/~ptm/introduction.html.
- [23] Y. Cao, C. Hu, A.B. Kahng, S. Muddu, D. Stroobandt, D. Sylvester, "Effects of Global Interconnect Optimizations on Performance Estimation of Deep Submicron Designs", *IEEE/ACM CAD Conf.*, pp. 56-61, 2000.