[8] P. Pant, R. K. Roy, and A. Chatterjee, "Dual-threshold voltage assignment with transistor sizing for low power CMOS circuits," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 9, no. 2, pp. 390–394, Apr. 2001.

[9] Y. S. Dhillon, A. U. Diril, A. Chattejee, and H.-H. S. Lee, "Algorithm for achieving minimum energy consumption in CMOS circuits using multiple supply and threshold voltages at the module level," in *Proc. ICCAD*, 2003, pp. 693–700.

[10] S. H. Choi, B. C. Paul, and K. Roy, "Novel sizing algorithm for yield improvement under process variation in nanometer technology," in *Proc. DATE*, 2004, pp. 454–459.

[11] A. Srivastava, D. Sylvester, and D. Blaauw, "Statistical optimization of leakage power considering process variations using dual-$V_{\rm th}$ and sizing," in *Proc. DAC*, 2004, pp. 773–778.

[12] O. Neiroukh and X. Song, "Improving the process-variation tolerance of digital circuits using gate sizing and statistical techniques," in *Proc. DATE*, 2005, pp. 294–299.

[13] M. R. Guthaus, N. Venkateswaran, C. Visweswariah, and V. Zolotov, "Gate sizing using incremental parameterized statistical timing analysis," in *Proc. ICCAD*, 2005, pp. 1029–1036.

[14] K. Chopra, S. Shah, A. Srivastava, D. Blaauw, and D. Sylvester, "Parametric yield maximization using gate sizing based on efficient statistical power and delay gradient computation," in *Proc. ICCAD*, 2005, pp. 1023–1028.

[15] A. Keshavarzi, S. Ma, S. Narendra, B. Bloechel, K. Mistry, T. Ghani, S. Borkar, and V. De, "Effectiveness of reverse body bias for leakage control in scaled dual V_t CMOS ICs," in *Proc. ISLPED*, Aug. 2001, pp. 207–212.

[16] J. Tschanz, J. T. Kao, S. G. Narendra, R. Nair, D. A. Antoniadis, A. P. Chandrakasan, and V. De, "Adaptive body bias for reducing impacts of die-to-die and within-die parameter variations on microprocessor frequency and leakage," *IEEE J. Solid-State Circuits*, vol. 37, no. 11, pp. 1396–1402, Nov. 2002.

[17] A. Keshavarzi, S. Narendra, S. Borkar, C. Hawkins, K. Roy, and V. De, "Technology scaling of optimum reverse body bias for standby leakage power reduction in CMOS IC's," in *Proc. ISLPED*, 1999, pp. 252–254.

[18] T. Kuroda and M. Hamada, "Low-power CMOS digital design with dual embedded adaptive power supplies," *IEEE J. Solid-State Circuits*, vol. 35, no. 4, pp. 652–655, Apr. 2000.

[19] C. H. Kim, K. Roy, S. Hsu, A. Alvandpour, R. K. Krishnamurthy, and S. Borkar, "A process variation compensating technique for sub-90-nm dynamic circuits," in *Symp. VLSI Circuits*, 2003, pp. 205–206.

[20] K. T. Cheng, S. Devadas, and K. Keutzer, "Robust delay-fault test generation and synthesis for testability under a standard scan design methodology," in *DAC*, 1991, pp. 80–86.

[21] M. Abramovici, M. A. Breuer, and A. D. Friedman, *Digital Systems Testing and Testable Design*. Murray Hill, NJ: AT&T, 1990.

[22] M. Ashouei, M. M. Nisar, A. Chatterjee, A. D. Singh, and A. U. Diril, "Probabilistic self-adaptation of nanoscale CMOS circuits: Yield maximization under increased intra-die variations," in *Proc. VLSI Des. Conf.*, 2007, pp. 711–716.

Accurate Predictive Interconnect Modeling for System-Level Design

Luca P. Carloni, Andrew B. Kahng, Swamy V. Muddu, Alessandro Pinto, Kambiz Samadi, and Puneet Sharma

***Abstract*—We propose new accurate predictive models for the delay, power, and area of buffered interconnects to enable a more effective system-level design exploration with existing and future nanometer technology processes. We show that our models are significantly more accurate than previous models—essentially matching sign-off analyses. We integrate our models in the COSI-OCC communication synthesis infrastructure and show how they impact the feasibility and optimality of the network-on-chip architectures that are synthesized by this tool.**

***Index Terms*—Communication synthesis, interconnect modeling, networks-on-chip (NoCs), system-level design.**

I. Introduction

Due to the increasing complexity of systems-on-chip (SoCs) and the poor scaling of interconnects with technology, on-chip communication is becoming a performance bottleneck and a significant consumer of power and area budgets [14], [34]. Decisions made in the early phase of the design process have the maximum potential to optimize the system for important objectives such as minimizing power dissipation [29]. Hence, in order to drive effective optimizations and reduce design guard band, it is crucial to account for global interconnects during system-level design by modeling their performance, power, and area.

In recent years, packet-switched networks-on-chip (NoCs) have been proposed as a new paradigm to design efficient and scalable on-chip communication fabrics [3], [8], [11], [12]. A NoC is obtained by combining multiple point-to-point data links (i.e., buffered interconnects) with routers and network interfaces [10], [13]. We focus on deriving closed-form models to predict the delay, power, and area of global buffered interconnects. Our goal is to provide system-level designers with fast and accurate models that can be used in the early phase of a SoC design process.

To date, there have not been any accurate yet simple models available to system-level designers. Current models are either quite accurate but too complex to be employed at the system level or, else, too coarse and inaccurate which leads to incorrect architectural design decisions. Furthermore, there has not been any study of the sensitivity of system-level decisions to the accuracy of these models. To this point,

Manuscript received June 02, 2008; revised December 14, 2008. First published November 17, 2009; current version published March 24, 2010. This work was supported by the Microelectronics Advanced Research Corporation Gigascale Systems Research Center.

L. P. Carloni is with the Department of Computer Science, Columbia University at New York, NY 10027 USA (e-mail: luca@cs.columbia.edu).

A. B. Kahng is with the Department of Computer Science and Engineering and the Department of Electrical and Computer Engineering, University of California at San Diego, La Jolla, CA 92093 USA (e-mail: abk@ucsd.edu).

S. V. Muddu is with AMD, Sunnyvale, CA 94088-3453 USA (e-mail: swamy@vlsicad.ucsd.edu).

A. Pinto is with the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA 94720 USA (e-mail: apinto@eecs.berkeley.edu).

K. Samadi is with the Department of Electrical and Computer Engineering, University of California at San Diego, La Jolla, CA 92093 USA (e-mail: ksamadi@ucsd.edu).

P. Sharma is with Freescale Semiconductor, Austin, TX 78735 USA (e-mail: sharma@vlsicad.ucsd.edu).

Digital Object Identifier 10.1109/TVLSI.2009.2014772

TABLE I
FITTING COEFFICIENTS FOR THE PREDICTIVE MODELS ACROSS SIX TECHNOLOGIES

Tech.	α_0	α_1	α_2	β_0	β_1	γ_0	γ_1	γ_2	η	κ_0^n	κ_1^n	κ_0^p	κ_1^p	τ_0	τ_1
90nm	0.013	0.217	-0.088	3.008	1.494	0.015	5.553	0.128	0.0015	-6.128	29.313	1.261	13.274	1.312	1.099
65nm	0.008	0.234	-0.144	2.219	1.252	0.012	4.162	0.142	0.0011	-6.034	26.561	1.238	27.082	0.657	0.866
45nm	0.007	0.157	-0.032	1.709	0.219	0.007	2.903	0.277	0.0008	54.636	29.432	86.21	34.131	0.678	0.022
32nm	0.024	0.084	-0.065	1.536	0.189	0.0042	2.301	0.452	0.0005	75.236	32.065	104.302	41.012	0.452	0.162
22nm	0.011	0.065	-0.087	1.132	0.122	0.0031	1.856	0.401	0.00018	92.325	38.842	131.36	44.812	0.386	0.089
16nm	0.007	0.188	-0.097	1.095	0.108	0.0011	1.015	0.341	0.00011	129.031	41.024	168.064	43.259	0.289	0.078

our work shows that accurate models can still be simple and that improved models lead to different optimization results.

We first define the requirements that a system-level model for global buffered interconnects should satisfy, and then, we discuss the shortcomings of the models available in literature. We present our predictive models together with a *reproducible* methodology to derive them. Different from previous work in the literature, we build our predictive models through accurate experimentations and calibrations against industry technology files and provide necessary explanations of the models and associated parameters. We apply linear and quadratic regressions to obtain the fitting coefficients of our predictive models, and we report coefficient values for six different nanometer technologies, from 90 to 16 nm. Since the accuracy of our models relies on the accuracy of the underlying technology parameters, we also highlight reliable sources that are available to system-level designers for present and future nanometer technologies. We compare predictions from our models with existing models and validate their accuracy against PrimeTime SI [25], an industry golden tool. Finally, we show the impact of the improved modeling accuracy on system-level design choices by comparing the NoC topologies that are synthesized by a system-level tool for the automatic synthesis of on-chip communication (COSI-OCC [22], [23]) when using different models.

The remainder of this paper is organized as follows. Section II reviews previous work and describes modeling requirements. In Section III, we develop accurate physical models for interconnect wires and repeaters. In Section IV, we validate the accuracy of our buffered interconnect-delay model against PrimeTime SI and also show the impact of the new models on the optimal NoC configurations that can be achieved with COSI-OCC.

II. RELATED WORK

System-level designers require *accurate yet simple* models of library elements (i.e., communicating entities and interconnections among them) to bridge planning and implementation and to enable meaningful system design optimization choices. Existing methods for on-chip communication synthesis [4], [22] and analysis [12] primarily use "classic" delay and power models, such as the one proposed by Bakoglu [2] or, more recently, by Pamunuwa *et al.* [20]. These models do not consider the impact of input-slew change on effective driver on-resistance or that of electron scattering and barrier thickness on interconnect resistance. The aforementioned deficiencies in gate- and wire-delay models are addressed to some extent in the large body of work on gate-delay [1], [9] and interconnect-delay [21], [30] modeling. While being very accurate, such models (e.g., asymptotic-waveform-evaluation (AWE)-based approaches [30] and post-AWE approaches [19], [33] which are mainstream) need detailed interconnect parasitic information which is unavailable during the system-level design phase.[1]

The delay of buffered interconnects is the sum of the wire and repeater delays.[2] For gate delays, previous works model input voltage as a piecewise-linear function and choose the value of series resistance more elaborately. Such approach has the drawback that the drive resistance is modeled as independent from the input transition time (slew). In reality, drive resistance (R_d) varies with input slew. This also affects the output slew. Both the drive-resistance dependence on input slew and the output-slew dependence on load capacitance and input slew must be considered to derive an accurate gate-delay model. Moreover, Shao *et al.* [31] propose a gate-delay model that relies on a second-order RC model of the gate. They propose analytical formulas for computing the output voltage waveform for a given ramp input waveform. However, they do not address gate loading during model construction.

III. BUFFERED INTERCONNECT MODEL

In this section, we describe our models and present a methodology to construct them from reliable and easily accessible sources for existing and future technologies. Our models are, by construction, calibrated against SPICE and contain well-defined parameters. We apply linear and quadratic regressions to obtain the fitting coefficients of our predictive models.

A. Repeater-Delay Model

For brevity, we present our repeater-delay model and describe its derivation only for the case of rise transitions in inverters. The derived functional forms are identical for fall transitions, and for buffers, only the function coefficients change. Table I lists the coefficients derived for TSMC 90- and 65-nm high-performance technologies, a foundry 45-nm low-power technology, as well as Predictive Technology Model (PTM) [24] 32-, 22-, and 16-nm high-performance technologies.

The repeater delay $d_r = i + r_d \cdot c_l$ can be decomposed into the sum of a load-independent part (or intrinsic delay of the gate) i and a load-dependent part that is the product of the drive resistance r_d and the load capacitance c_l. The intrinsic delay i can potentially depend on the input slew of the gate and the gate size. However, as shown in Fig. 1, i is practically independent of the gate size while it depends nearly quadratically on the input slew. The independence of intrinsic delay from gate size can be understood as follows. As the inverter size increases, the drain capacitance increases and the gate resistance decreases. Hence, the overall impact on intrinsic delay is negligible. For buffers, the intrinsic delay additionally comprises of the delay of the inverter in the first stage which drives the inverter in the second stage. As the buffer size increases, the size of the second-stage inverter increases but the size of the first-stage inverter is also increased to maintain a small intrinsic delay. Consequently, the total intrinsic delay of a buffer is nearly independent of the buffer size. The quadratic dependence of the intrinsic delay on input slew is captured by $i(s_i) = \alpha_0 + \alpha_1 \cdot s_i + \alpha_2 \cdot s_i^2$, where s_i denotes the input slew and α_0, α_1, and α_2 are the coefficients determined by quadratic regression.

We observe that the drive resistance r_d is nearly linear with input slew particularly for larger input-slew values. We also note that both the intercept and slope vary with repeater size; hence, we can write $r_d = r_{d0} + r_{d1} \cdot s_i$, where r_{d0} and r_{d1} are coefficients that both depend on the repeater size. Both r_{d0} and r_{d1} can readily be calculated using linear regression for a few repeater sizes. Previous works (e.g., [2]) have assumed r_d to be inversely proportional to the repeater size. We have confirmed this relationship to be sufficiently accurate for sub-90-nm

[1] PrimeTime SI uses such post-AWE methods.

[2] We use the term "repeater" to denote both an inverter and a buffer.

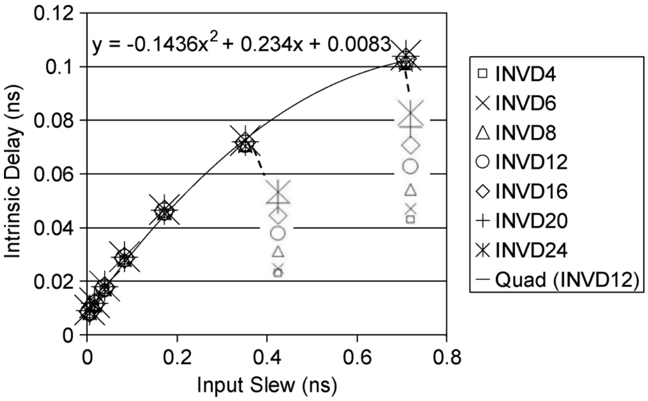

Fig. 1. Dependence of repeater intrinsic delay on input slew and inverter size. Intrinsic delay is essentially independent of repeater size and depends quadratically on input slew.

technology modeling. To be precise, we use the pMOS (nMOS) device width as the repeater size for rise (fall) transitions. Both r_{d0} and r_{d1} are inversely proportional to the repeater size, and the exact coefficients can be calculated using linear regression with zero intercept as $r_{d0}(w_r) = \beta_0/w_r$ and $r_{d1}(w_r) = \beta_1/w_r$, respectively, where w_r is the repeater size that is equal to the pMOS (nMOS) width for rise (fall) transitions and β_0 and β_1 are the fitted coefficients. Since our gate-delay model depends on input slew, we also model the output slew of the previous stage of the buffered interconnect. As with gate delay, slew depends on repeater size, input slew, and load capacitance. Slew depends strongly on the load capacitance, and we have found a linear relationship to be a good tradeoff between simplicity and accuracy. We note that the slope is nearly independent from the input slew, while the intercept depends linearly on it. Hence, the output slew for a given repeater is $s_o(c_l, s_i) = s_{o0} + s_{o1} \cdot s_i + s_{o2} \cdot c_l$, where s_o is the output slew and s_{o0}, s_{o1}, and s_{o2} are the fitting coefficients readily derived from multiple linear regressions. Furthermore, we consistently observe that s_{o0} and s_{o2} are independent of the repeater size, but s_{o1} varies inversely with repeater size. Hence, output slew can be calculated as $s_o(c_l, s_i, w_r) = \gamma_0 + (\gamma_1 \cdot s_i / w_r) + \gamma_2 \cdot c_l$, where γ_0, γ_1, and γ_2 are fitting coefficients.

In addition, the input capacitance of a repeater is required to calculate the load capacitance of the previous stage. As expected, the input capacitance is proportional to the repeater size. Typically, the P/N ratio is kept constant for repeaters of all sizes, and the previous models (e.g., [2]) are sufficient. When it does change with repeater size, the input capacitance can be evaluated as $c_i = \eta \times (w_p + w_n)$, where w_p and w_n are pMOS and nMOS widths, respectively, and η is a coefficient derived using linear regression with zero intercept.

B. Wire-Delay Model

For the wire delay, we start from the model proposed by Pamunuwa *et al.* [20] which accounts for cross-talk-induced delay $d_w = r_w(0.4c_g + (\lambda_i/2)c_c + 0.7c_i)$, where d_w, r_w, c_g, c_c, and c_i denote the wire delay, wire resistance, ground capacitance, coupling capacitance, and input capacitance of the next-stage repeater, respectively. The coefficient λ_i accounts for the switching patterns of the neighboring wires and is equal to 1.51 for worst case switching. We enhance the accuracy of the model by considering two important factors that affect wire resistance: 1) electron scattering and 2) interconnect barrier. For the scattering effect, we adopt the closed-form width-dependent resistivity equation proposed in [32]. To incorporate the impact of barrier thickness on interconnect resistance, we use the model presented in [27] and [28].

C. Power and Area Models

Power is a first-class design objective and must be modeled early in the design flow [29]. In current technologies, leakage and dynamic power are the main components of power dissipation. In repeaters, leakage occurs in both output states. nMOS devices leak when the output is high, while pMOS devices leak when the output is low. This also applies to buffers because the second-stage devices are the primary contributors due to their large sizes. Leakage power has two main components: 1) subthreshold leakage and 2) gate-tunneling current. Both components depend linearly on device size. Thus, leakage power can be calculated using $p_s = (p_s^n + p_s^p)/2$, where $p_s^n = \kappa_0^n + \kappa_1^n \cdot w_n$ and $p_s^p = \kappa_0^p + \kappa_1^p \cdot w_p$ are the leakage power for nMOS and pMOS devices, respectively, and κ_0^n, κ_1^n, κ_0^p, and κ_1^p are coefficients determined using linear regression. The dynamic power is given by the well-known equation $p_d = \alpha \cdot c_l \cdot v_{\mathrm{dd}}^2 \cdot f$, as a function of activity factor α, load capacitance c_l, supply voltage v_{dd}, and clock frequency f. The load capacitance is the sum of the input capacitance of the next repeater (c_i) and the ground (c_g) and coupling (c_c) capacitances of the driven wire.

Since repeaters are composed of several fingered devices connected in parallel, the repeater area grows linearly with the repeater size. For existing technologies, the repeater area a_r can be calculated as $a_r = \tau_0 + \tau_1 \cdot w_n$, where τ_0 and τ_1 are coefficients determined using linear regression. For future technologies, area values may not be available for performing linear regression. Hence, we propose the use of feature size, contact pitch, and row height—all of which become available early in process and library development and are also predictable—to estimate the area. The number of fingers can be calculated as $N_f = (w_p + w_n)/(h_{\mathrm{row}} - 4 \cdot p_{\mathrm{contact}})$, where h_{row} and p_{contact} are the row height and contact pitch, respectively, and cell width can be derived using $w_{\mathrm{cell}} = (N_f + 1) \times p_{\mathrm{contact}}$. Hence, the repeater area is $a_r = h_{\mathrm{row}} \times w_{\mathrm{cell}}$. The area of global wiring can be calculated as $a_w = n \times (w_w + s_w) + s_w$, where a_w denotes the wire area, n is the bit width of the bus, and w_w and s_w are the wire width and spacing computed from the width and spacing of the layer on which the wire is routed, considering the design style.

D. Buffering Schemes

Delay-optimal buffering optimizes the size and number of repeaters and has been addressed under simple delay models in previous works including [2], [7], and [20]. However, delay-optimal buffering results in extremely large repeaters having sizes that are never used in practice due to area and power-consumption considerations.

Our buffering optimization technique is based on binary search to optimize a given objective function (i.e., a weighted product of delay and power) for a given number and size of repeaters. Similarly to the approach in [5], we exhaustively search for the best combination of the size and number of buffers that minimizes a linear combination of the delay and power based on a specific weighting factor (i.e., the weighting factor allows us to emphasize either power or delay depending on the application). We use our proposed delay and power models to compute the necessary metrics in the objective function. The advantage of our approach with respect to the one in [5] is that we do not need to run SPICE simulations for each technology node as delay and power models are already calibrated for multiple technology nodes (i.e., 90, 65, and 45 nm). We also support the use of staggering buffer insertion to avoid the cross-talk effect on the signal delay by setting the Miller factor to zero in our delay equation. We note that, for these technologies, power can be reduced by 20% at the cost of just above 2% degradation in delay.

E. Modeling Infrastructure and Usage

We have developed a set of tools and application programming interfaces that allow us to abstract the buffered interconnect cost-performance tradeoffs from detailed SPICE simulations.[3] Our delay, power, and area models can be mathematically derived from the following inputs. For *repeater-delay calculation*, delay and slew values for a set of input-slew and load-capacitance values, along with input-capacitance values, are required for a few repeaters. Since the coefficients are derived using regression, a larger data set improves accuracy. The required data set is available from Liberty library files or can be generated using SPICE simulations for existing technologies. Since libraries are not available for future technologies, SPICE simulations must be used along with SPICE netlists for repeaters and predictive device models such as PTM. To construct the repeater netlists, a pMOS/nMOS ratio is assumed (from previous technology experience or from expected pMOS/nMOS drive strengths and is kept constant for all repeaters), and a variety of repeaters are constructed for different device sizes.

For *wire-delay calculation*, we require the wire dimensions and interwire spacings for global and intermediate layers. These values are available in LEF [16] and ITF [18] files for existing technologies and in the International Technology Roadmap for Semiconductors (ITRS) [15] for future and existing technologies. For *power calculations*, input capacitance (computed in repeater-delay calculation) and wire parasitics (computed in wire-delay calculation) are used. Additionally, device leakage is required and can be computed from the Liberty library files or SPICE simulations. For *area calculations*, wire dimensions used in wire-delay calculation are used to compute wire area. Repeater area is readily available for existing technologies in Liberty or LEF files or from layouts. For future technologies, ITRS A-factors can be used or equations developed in Section III can be used along with the feature size, row height, and contact pitch, all of which values are available early in process and library development. Finally, the total delay of a buffered interconnect is the sum of the delays of all repeaters and wire segments in it.

TABLE II
EVALUATION OF MODEL ACCURACY

Tech.	L	DS	PT	B	P	$Prop.$	RT
	(mm)		(ns)	(%)	(%)	(%)	(X)
90nm	1	SW-SS	0.144	89.9	26.3	-11.2	2.2
		shielding	0.108	84.2	22.3	-8.6	2.3
	3	SW-SS	0.411	91.1	36.2	-2.3	2.1
		shielding	0.398	89.6	31.2	-1.8	2.2
	5	SW-SS	0.670	97.0	66.4	-8.2	2.3
		shielding	0.659	92.4	65.2	-6.7	2.3
	10	SW-SS	1.394	85.6	52.3	-10.4	2.3
		shielding	1.344	79.5	47.6	-7.1	2.3
	15	SW-SS	2.170	105.9	59.1	-6.2	2.4
		shielding	1.630	99.2	55.3	-5.8	2.3
65nm	1	SW-SS	0.116	6.1	53.2	-4.3	2.2
		shielding	0.107	5.1	50.9	-3.1	2.2
	3	SW-SS	0.318	-2.3	45.3	-3.5	2.2
		shielding	0.302	-3.4	41.3	-2.9	2.1
	5	SW-SS	0.505	-6.9	33.7	-5.0	2.2
		shielding	0.489	-4.5	31.9	-3.9	2.3
	10	SW-SS	1.061	-3.1	39.6	-4.9	2.1
		shielding	1.012	-4.5	29.8	-2.9	2.3
	15	SW-SS	1.641	-7.1	44.2	-4.2	2.3
		shielding	1.531	-5.1	41.7	-3.9	2.3
45nm	1	SW-SS	0.107	16.3	33.8	6.3	2.1
		shielding	0.098	11.2	31.2	6.2	2.1
	3	SW-SS	0.301	17.4	26.6	8.5	2.2
		shielding	0.291	14.2	26.1	7.9	2.1
	5	SW-SS	0.485	23.4	29.3	9.7	2.2
		shielding	0.474	24.2	26.7	7.8	2.2
	10	SW-SS	0.990	21.2	32.6	9.9	2.2
		shielding	0.962	24.4	23.8	9.1	2.3
	15	SW-SS	1.607	31.3	29.4	8.8	2.4
		shielding	1.479	29.2	28.7	8.2	2.4

IV. VALIDATION AND SIGNIFICANCE ASSESSMENT

To assess the accuracy of our model with respect to previously proposed models ([2] and [20]), we consider buffered interconnects of lengths 1, 3, 5, 10, and 15 mm for three technology choices (90, 65, and 45 nm), two design styles (single-width–single-spacing and shielding), and global wiring regime against physical implementation.[4]

To create the layout of a buffered interconnect, we first define the placement area in *Cadence SOC Encounter (version 6.1)*. Repeaters are then placed at equal distances along the wire length to buffer the interconnect uniformly. Connections between the inputs, outputs, and buffers are created by *Cadence NanoRoute*. The values of minimum wire spacing and width are chosen from the input LEF file. Parasitic extraction on the buffered lines is performed using *SOC Encounter's* built-in extractor. To perform timing analysis, we read in the parasitics output from *SOC Encounter* in Standard Parasitic Exchange Format (SPEF) and the timing library (Liberty format) into *PrimeTime SI (version 2006.12)* for sign-off delay calculation. The results of our accuracy studies are presented in Table II as a function of the wire length L and design style DS. The columns denoted as B, P, and $Prop.$ report the errors in delay prediction using Bakoglu's model [2], the model of Pamunuwa *et al.* [20], and our proposed model with respect to the delay of the buffered line evaluated using PrimeTime $(input\,transition\,time = 300\text{ ps})$, which is reported in column PT. We observe that the prediction with our proposed method matches the value from PrimeTime within 12%. In comparison, previous models have errors in the range of -7%–106%. Finally, the column denoted as RT reports the ratio of the CPU runtime of our proposed model versus PrimeTime (the runtimes of the Bakoglu and Pamunuwa models are similar to ours since they are all simple analytical models). To perform runtime comparison, we use the following approach. For PrimeTime, we measure the time from when it starts calculating the interconnect delay (i.e., when "report timing" is called) until it returns the delay value.[5] For our model, we only measure the computation time (i.e., from when inputs are available until the delay estimate is returned). Our models are implemented in C++. We report the average runtime values over 50 trials. Our proposed model is computationally at least 2.1 times faster than PrimeTime when both are run on a computer with a 2.4-GHz Intel Xeon processor. More importantly, our models avoid the significant setup time, license management, etc., required for PrimeTime. In summary, our new models achieve significant accuracy and runtime improvement compared with the previous models and PrimeTime, respectively.

We also verify the accuracy of our leakage-power and repeater-area models. With respect to the cell leakage-power values reported in the Liberty files for 90-, 65-, and 45-nm technologies, the maximum error of our predictive model is less than 11%.[6] With respect to the cell area values of the corresponding cells in the Liberty files, the maximum error of our predictive model is less than 8%.

To assess the impact of improved accuracy on system-level design-space exploration, we integrate our models in COSI-OCC, a system-level

[3] SPICE simulation solves circuit equations based on device-level compact models [26]. These models capture several CMOS phenomena required to calculate device-level electrical metrics such as terminal-to-terminal currents.

[4] Since delay changes linearly with respect to length for buffered interconnects (Table II), 1, 3, 5, 10, and 15 mm are representative of other lengths that require buffering.

[5] To run PrimeTime, we need several components including the netlist, SPEF, and Liberty files which all require a significant amount of time to generate. We consider these as one-time runtime costs and do not include them in our runtime analysis.

[6] The repeater sizes used in our experiments include INVD4, INVD6, INVD8, INVD12, INVD16, INVD20, and INVD20.

TABLE III
MODEL IMPACT ON NoC SYNTHESIS

SoC		P_{dyn} $(m\mathrm{W})$		P_{leak} $(m\mathrm{W})$		A_d (mm^2)	
		Orig.	Prop.	Orig.	Prop.	Orig.	Prop.
VPROC	90nm	117.3	364.8	38.1	99.6	0.070	0.009
	65nm	51.1	179.9	69.9	86.7	0.036	0.007
	45nm	18	231	49	291	0.02	0.003
dVOPD	90nm	63.4	88.0	14.2	32.5	0.026	0.003
	65nm	27.3	73.2	25.7	33.2	0.013	0.003
	45nm	9.6	98	18.1	142	0.007	0.002
SoC		A_{tot} (mm^2)		Ave. # of hops		Max. # of hops	
		Orig.	Prop.	Orig.	Prop.	Orig.	Prop.
VPROC	90nm	0.370	0.346	3.09	3.01	4	5
	65nm	0.217	0.223	3.10	3.42	4	6
	45nm	0.138	0.137	3.1	3.2	4	6
dVOPD	90nm	0.141	0.162	1.76	1.76	3	3
	65nm	0.082	0.085	1.76	1.91	3	4
	45nm	0.053	0.029	1.76	2.12	3	5

tool for the synthesis of NoCs. We use two representative SoC designs as test cases. The first design (VPROC) is a video processor with 42 cores and 128-b data widths. The second design is based on a dual video object plane decoder, where two video streams are decoded in parallel by utilizing 26 cores and 128-b data widths. Table III compares the interconnect power, delay, and area when the model originally used in COSI-OCC [22] and the proposed model are used. The original model uses the Bakoglu delay model and does not consider any of the improvements that we have discussed. It also obtains its technology inputs from PTMs which are not calibrated compared with industry library files. The clock frequencies used are 1.5, 2.25, and 3.0 for 90-, 65-, and 45-nm technology nodes, respectively. Hop count, which captures the communication latency, is also reported. The main differences between the NoC architectures obtained using the original and the proposed models are in the power and hop-count figures across all technology processes. The dynamic power consumption estimated by the proposed model is up to three times as large as the dynamic power consumption estimated by the original model for 90- and 65-nm technology nodes. The difference depends on the coupling capacitance that is neglected by the original model and the different sizes and numbers of repeaters used by the two models. For the proposed model, we observe an increase in dynamic power going from 65 to 45 nm. This is due to the supply voltage increase in the library files from 1 to 1.1 V, respectively. This difference also widens the gap in dynamic power between the original and proposed models. The leakage power is also different, mainly as a consequence of the number and size of the repeaters that are optimistically estimated by the original model. Moreover, the original model turns out to be very optimistic in allowing the use of excessively long wires. This is an example of a nonconservative abstraction that leads to design solutions that are actually not implementable. Finally, we note that the difference in area estimates between the original and proposed models is very large because of the simplistic assumption on the area occupation in the original model.

V. CONCLUSION

The accurate estimation of the delay, power, and area of global interconnects in the early phases of the design process can drive effective system-level exploration. We have proposed new accurate predictive models, integrated them in the COSI-OCC communication synthesis infrastructure, and found that their use substantially improves the quality of the NoC synthesis results.

REFERENCES

[1] R. Arunachalam, F. Dartu, and L. Pileggi, "CMOS gate delay models for general RLC loading," in *Proc. ICCD*, 1997, pp. 224–229.

[2] H. Bakoglu, *Circuits, Interconnections and Packaging for VLSI*. Reading, MA: Addison-Wesley, 1990.

[3] L. Benini and G. De Micheli, "A new SoC paradigm," *Computer*, vol. 35, no. 1, pp. 70–78, Jan. 2002.

[4] D. Bertozzi, A. Jalabert, S. Murali, R. Tamhankar, S. Stergiou, L. Benini, and G. De Micheli, "NoC synthesis flow for customized domain specific multiprocessor systems-on-chip," *IEEE Trans. Parallel Distrib. Syst.*, vol. 16, no. 2, pp. 113–129, Feb. 2005.

[5] Y. Cao, C. M. Hu, X. J. Huang, A. B. Kahng, S. Muddu, D. Stroobandt, and D. Sylvester, "Effects of global interconnect optimizations on performance estimation of deep submicron design," in *Proc. IEEE ICCAD*, 2000, pp. 56–61.

[6] L. Carloni, A. B. Kahng, S. Muddu, A. Pinto, K. Samadi, and P. Sharma, "Interconnect modeling for improved system-level design optimization," in *Proc. ASPDAC*, 2008, pp. 258–264.

[7] J. Cong and D. Z. Pan, "Interconnect delay estimation models for synthesis and design planning," in *Proc. IEEE ASPDAC*, 1999, pp. 507–510.

[8] W. J. Dally and B. Towles, "Route packets, not wires: On-chip interconnection networks," in *Proc. ACM/IEEE DAC*, 2001, pp. 684–689.

[9] F. Dartu, N. Menezes, and L. Pileggi, "Performance computation for precharacterized CMOS gate with RC load," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 15, no. 5, pp. 544–553, May 1996.

[10] G. De Micheli and L. Benini, *Networks on Chip*. San Mateo, CA: Morgan Kaufmann, 2006.

[11] A. Hemani, A. Jantsch, S. Kumar, A. Postula, J. Oberg, M. Millberg, and D. Lindqvist, "Network on chip: An architecture for billion transistor era," in *Proc. 18th IEEE NorChip Conf.*, Nov. 2000, pp. 166–173.

[12] S. Heo and K. Asanovic, "Replacing global wires with an on-chip network: A power analysis," in *Proc. ISLPED*, 2005, pp. 369–374.

[13] A. Jantsch and H. Tenhunen, *Networks on Chip*. Norwell, MA: Kluwer, 2003.

[14] K. W. Mai, R. Ho, and M. A. Horowitz, "The future of wires," *Proc. IEEE*, vol. 89, no. 4, pp. 490–504, Apr. 2001.

[15] ITRS, "International Technology Roadmap for Semiconductors," 2007. [Online]. Available: http://www.public.itrs.net/

[16] *"LEF/DEF Language Reference,"* Cadence, San Jose, CA, 2004 [Online]. Available: https://www.si2.org/openeda.si2.org/projects/lefdef

[17] *"Liberty File Format, Liberty NCX User Guide,"* ver. B-2008.06-SP2, Synopsys, Inc., Mountain View, CA, 2004.

[18] *"Star-RCXT User Guide,"* ver. B-2008.06-SP2, Synopsys, Inc., Mountain View, CA, 2008.

[19] A. Odabasioglu, M. Celik, and L. T. Pileggi, "PRIMA: Passive reduced-order interconnect macromodeling algorithm," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 17, no. 8, pp. 645–654, Aug. 1998.

[20] D. Pamunuwa, L.-R. Zheng, and H. Tenhunen, "Maximizing throughput over parallel wire structures in the deep submicrometer regime," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 11, no. 2, pp. 224–243, Apr. 2003.

[21] L. Pillage and R. Rohrer, "Asymptotic waveform evaluation for timing analysis," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 9, no. 4, pp. 352–366, Apr. 1990.

[22] A. Pinto, L. P. Carloni, and A. L. Sangiovanni-Vincentelli, UCB/EECS, Berkeley, CA, A methodology and an open software infrastructure for constraint-driven synthesis of on-chip communications Tech. Rep. UCB/EECS-2007-130, Nov. 2007.

[23] A. Pinto, L. P. Carloni, and A. L. Sangiovanni-Vincentelli, "COSI: A framework for the design of interconnection networks," *IEEE Des. Test Comput.*, vol. 25, no. 5, pp. 402–415, Sep./Oct. 2008.

[24] ASU, Tempe, AZ, "Predictive technology model," 2008. [Online]. Available: http://www.eas.asu.edu/~ptm/

[25] *"PrimeTime SI User Guide,"* ver. B-2006.12, Synopsys, Inc., Mountain View, CA, 2004.

[26] UC Berkeley, Berkeley, CA, "BSIM models," 2005. [Online]. Available: http://www-device.eecs.berkeley.edu/bsim3/

[27] N. Lu, M. Angyal, G. Matusiewicz, V. McGahay, and T. Standaert, "Characterization, modeling and extraction of Cu wire resistance for 65 nm technology," in *Proc. CICC*, 2007, pp. 57–60.

[28] Y. Travaly, M. Bamal, L. Carbonell, F. Iacopi, M. Stucchi, M. Van Hove, and G. P. Beyer, "A novel approach to resistivity and interconnect modeling," *Microelectron. Eng.*, vol. 83, no. 11/12, pp. 2417–2421, Nov./Dec. 2006.

[29] A. Raghunathan, N. K. Niraj, and S. Dey, *High-Level Power Analysis and Optimization*. Norwell, MA: Kluwer, 1998.

[30] C. Ratzlaff and L. Pillage, "RICE: Rapid interconnect circuit evaluation using AWE," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 13, no. 6, pp. 763–776, Jun. 1994.

[31] M. Shao, M. D. F. Wong, H. Cao, Y. Gao, L.-P. Yuan, L.-D. Huang, and S. Lee, "Explicit gate delay model for timing evaluation," in *Proc. ACM/IEEE ISPD*, 2003, pp. 32–38.

[32] S. X. Shi and D. Z. Pan, "Wire sizing and shaping with scattering effect for nanoscale interconnection," in *Proc. IEEE ASPDAC*, 2006, pp. 503–508.

[33] L. M. Silveria, M. Kamon, I. Elfadel, and J. White, "A coordinate-transformed Arnoldi algorithm for generating guaranteed stable reduced-order models of RLC circuits," in *Proc. ICCAD*, 1996, pp. 288–294.

[34] D. Sylvester and K. Keutzer, "A global wiring paradigm for deep submicron design," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 19, no. 2, pp. 242–252, Feb. 2000.

An Approach for Adaptive DRAM Temperature and Power Management

Song Liu, Yu Zhang, Seda Ogrenci Memik, and Gokhan Memik

***Abstract*—High-performance DRAMs are providing increasing memory access bandwidth to processors, which is leading to high power consumption and operating temperature in DRAM chips. In this paper, we propose a customized low-power technique for high-performance DRAM systems to improve DRAM page hit rate by buffering write operations that may incur page misses. This approach reduces DRAM system power consumption and temperature without any performance penalty. We combine the throughput-aware page-hit-aware write buffer (TAP) with low-power-state-based techniques for further power and temperature reduction, namely, TAP-low. Our experiments show that a system with TAP-low could reduce the total DRAM power consumption by up to 68.6% (19.9% on average). The steady-state temperature can be reduced by as much as 7.84 °C and 2.55°C on average across eight representative workloads.**

***Index Terms*—DRAM, power, temperature.**

I. Introduction

Technological advances in microprocessor architectures enable high performance with an underlying assumption on increasing utilization of memory systems. On the other hand, increasing memory densities and data rates lead to higher operating temperatures in DRAM systems. Moreover, several techniques have been proposed to place DRAM closer to processor cores, such as 3-D ICs [10], and embedded DRAM [5]. With increasing power consumption and closer physical proximity to hot processor cores, modern DRAMs are operated under increasing temperatures. Prior studies have shown that DRAM temperature control has become a practical and pressing issue [4].

In this paper, we propose a DRAM architecture enhancement, which could harvest the largest peak temperature reduction without incurring any performance overhead. Specifically, we propose a customized method to reduce DRAM power consumption by improving DRAM page hit rate. Moreover, higher page hit rate also leads to less average DRAM access latency and thus improves system performance.

In our previous works, we have designed and analyzed the page-hit-aware write buffer (PHA-WB) [12] and the throughput-aware PHA-WB (TAP) [11]. PHA-WB provides a buffering mechanism to hold write operations that may cause a page miss. The TAP scheme was designed to dynamically adjust the tradeoff between the aggressiveness of the power optimization mechanism at the expense of more storage for buffering the data and orchestrating the buffer–DRAM coordination.

In this paper, we extend our work in two main aspects.

We take DRAM refresh operations into consideration. Experiments show that refresh operations have a strong impact on DRAM page hit rate. However, this impact decreases as DRAM traffic increases.

We also extend TAP with low-power-state-based techniques into TAP-low. We demonstrate that PHA-WB can actually increase the utilization of low-power states. PHA-WB also reduces average read delay by improving page hit rate, which reduces the performance penalty of low-power-state-based techniques.

Our experiments show that the TAP-low approach reduces the total DRAM system power consumption by as much as 68.6% (19.9% on average) and DRAM steady-state temperature by as much as 7.84 °C (2.55°C on average) for eight different workloads based on 20 SPEC CPU 2000 benchmarks running on a four-core CMP [18].

The remainder of this paper is organized as follows. Section II discusses related work. Our proposed technique is described in Section III. Section IV describes TAP-low, which is a combination of TAP and a low-power-state-based technique.Section V presents the experimental methodology and results. Finally, we conclude this paper in Section VI.

II. Related Work

DRAM power consumption can comprise a large portion of the total system power. Modern DRAMs provide various power-saving states. Various low-power DRAM techniques focus on utilizing these idle states efficiently to achieve the best energy-delay product [2], [3], [6]. Our goal, however, is to tie the active periods of DRAM operation to power consumption. These low-power-state-based techniques can be used as complementary to our approach.

Memory controller reordering is a widely used technique in stream processors [7], [15]. In these systems, the memory controller reorders memory accesses, so that there are more chances to use efficient page and burst modes. On the other hand, our technique, which targets general-propose processors with write-back cache and fully buffered dual inline memory module (FB-DIMM), is a further enhancement for burst-accessed DRAM.

Dynamic thermal management (DTM) of DRAM has become a pressing issue in mobile systems [4]. In order to cool down the DRAM while keeping the performance penalty small, Lin *et al.* [8] proposed adaptive core gating and dynamic voltage and frequency scaling (DVFS) to CMP systems. However, DTM and DVFS are known to introduce system performance penalties. We refer to these techniques as memory-traffic-control-based temperature-aware techniques since they handle DRAM thermal emergencies by reducing DRAM access density.

Existing power- and temperature-aware techniques focus on two special cases. Power-state-based techniques are designed for applications

Manuscript received August 01, 2008; revised January 23, 2009. First published June 16, 2009; current version published March 24, 2010. This work was supported in part by the US Department of Energy (DOE) under Award DEFG02-05ER25691 and in part by the National Science Foundation (NSF) under Awards CNS-0546305, CCF-0747201, CNS-0720691, IIS-0613568, CCF-0541337, CNS-0551639, IIS-0536994, CCF-0444405. The work of G. Memik was supported in part by Wissner-Slivka Chair funds.

The authors are with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208 USA (e-mail: sli646@eecs.northwestern.edu; yzh702@eecs.northwestern.edu; seda@eecs.northwestern.edu; memik@eecs.northwestern.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVLSI.2009.2014842