Energy-Delay Space Exploration of Clocked Storage Elements Using Circuit Sizing

Mustafa Aktan
Electrical Engineering
UT Dallas
Richardson, TX 75080
e-mail:aktanmus@utdallas.edu

Sivakumar Paramesvaran
Electrical Engineering
UT Dallas
Richardson, TX 75080
e-mail:sxp081000@utdallas.edu

Joosik Moon
Electrical Engineering
UT Dallas
Richardson, TX 75080
e-mail:jxm055000@utdallas.edu

Vojin Oklobdzija
Electrical Engineering
UT Dallas
Richardson, TX 75080
e-mail:vojin@utdallas.edu

Abstract— Rapid energy-delay exploration methodology based on circuit sizing as applied to clocked storage elements is presented. Circuit delay and energy are modeled using improved RC delay model of a transistor. The accuracy of the model is increased by using Logical Effort setup accounting for input signal slope and extraction of technology dependent parameters. The minimal energy-delay curve is generated by optimizing transistor sizes for minimum energy at given delay targets. Results show two orders of magnitude time improvement as compared to H-SPICE in order to generate such curves while the delay accuracy of the model used remains within 10 % as compared to H-SPICE.

I. Introduction

Comparing different circuit topologies in energy-delay (E-D) space is necessary in order to determine optimal trade-offs in terms of performance and power. Two different circuits may behave differently for different speed or energy targets, as illustrated in Fig.1. The E-D curves for the two circuits, shown in Fig.1, are obtained by tuning transistor sizes for the optimum delay for a given energy budget (or vice versa), under fixed input/output load and supply voltage assumption.

The best circuit topology of choice is determined by system specifications (input/output load, delay, etc). There are other factors influencing the choice such as reliability and process variation tolerance, but these are not subject of this paper. In any case, one needs to obtain E-D plots of various circuit topologies under various conditions and for the given system specifications [13]. How to generate these curves without elaborate simulation is important because there are many possible sizing solutions. An exhaustive search on every possible sizing is time consuming and for large circuits, not possible. Consider as an example a circuit composed of 10 transistors where each transistor can be assigned 10 different sizes. Then, there are 10^{10} possible cases to be simulated.

Clocked Storage Elements (CSE), namely flip-flops or latches, are employed in the pipeline of high-performance processors. The outputs of logic operations from the previous stage need to be stored and will be used as the inputs for the next stage in the next clock cycle [1]. Because of their excessive number it is important to employ high-speed but low-power CSEs to compensate for the hardware overhead accompanied by the pipelined architecture. The optimal selection of CSEs is determined by the system specifications [1]. For fixed input/output load and supply voltage, energy and delay can be traded against each other. Since the delay and power of a CSE depends on the size of each transistor, analysis of the Energy-Delay metric is required [2].

The problem involves transistor/gate sizing and developing a CAD tool for sizing digital CMOS circuits is not a new problem [3-5,10]. There has been much effort in sizing digital CMOS circuits for minimum delay, area, etc. Methods have been proposed on different levels of abstraction, mainly transistor [3,4] and gate-level [5,10]. Transistor-level sizing formulation depends on modeling the transistor as a switched R-C network. Each transistor is replaced by its corresponding network and writing the delay equations is merely writing the delay of a distributed R-C network. Gate-level solutions take a similar approach. Instead of the transistor, the gate is modeled as a switched R-C network. The gate-level approach reduces the number of variables involved in the optimization problem. The number of variables in optimization problems is critical since it has a direct impact on the solution time. Transistor-level methods

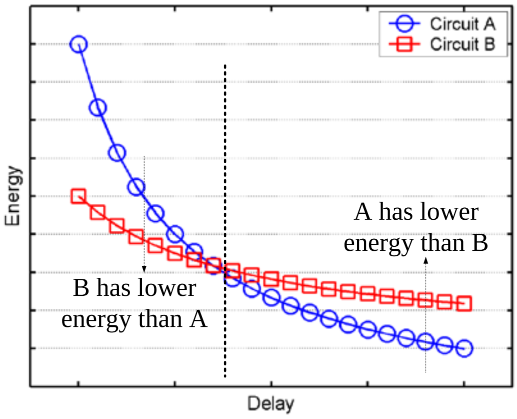


Fig. 1. Energy-delay space

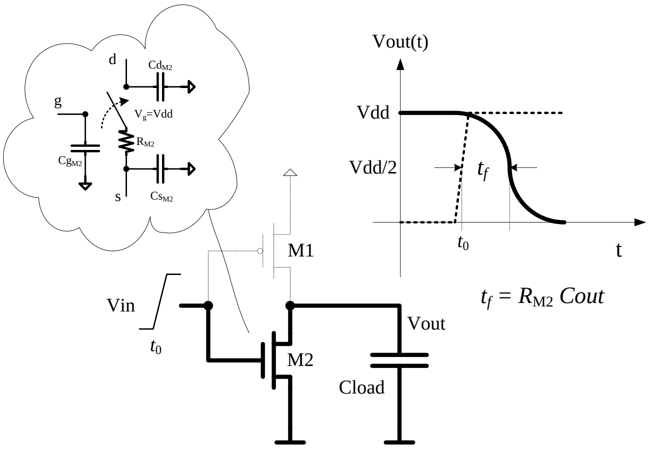


Fig. 2. Transistor R-C model.

have the advantage of providing a better solution due to the flexibility in sizing transistors independently as opposed to the gate-level approach where the p-n ratio is fixed.

In this work, a quick way of generating E-D curves for CSEs using transistor sizing is given. Logical Effort [10] is a simple sizing method; however it is difficult to decompose CSEs into simple gates (i.e. nand, nor, etc.) which involve feedback paths and different circuit structures exist. Therefore, sizing is done at the transistor level which avoids the need to decompose the circuit. Furthermore, transistor level sizing gives the flexibility to size each transistor independently (i.e. there is no fixed pn ratio). The RC delay model used is an enhanced model. Parameters are extracted from technology characterization by taking into account signal slope effects. The methodology is tested on different structures. It allows the right selection of topology and is used to quickly evaluate E-D space behaviors of various CSE elements.

II. TRANSISTOR MODEL USED FOR SIZING

A transistor can be modeled as a set of resistances and capacitances. Each transistor in the circuit is replaced by its equivalent R-C model. Delay is modeled as the calculated R-C network delay (Elmore delay) of the critical paths.

A transistor M of size w (size refers to width of the transistor we assume channel length is fixed and is the same for all transistors in the circuit) is modeled by its gate/drain/source capacitances and channel (drain-to-source) resistance as shown in Fig. 2. The gate and drain/source capacitances for a transistor can be written as Cg = Ct*w and Cd = Cs = Cp*w   where Ct and Cp are the gate and parasitic capacitances of a unit-size transistor. If M is NMOS (PMOS), the channel resistance is RM = Rn/wM (RM = Rp/wM) where Rn (Rp) is the channel resistance of a unit sized NMOS (PMOS) transistor.

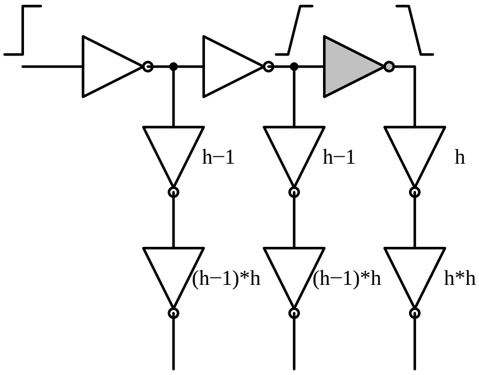


Fig. 3. Setup for determining $\tau_n = \hat{R}_n Ct$ considering input slopes and variable loads (h = 2,3,…).

Consider the fall delay of the inverter in Fig. 2. The delay associated with this operation is tf .The unit size transistor channel resistance   which is used to calculate RM2 can be analytically calculated from

$$\hat{R}_n = \left(\frac{2V_{Tn}}{V_{dd}-V_{Tn}} + \ln\left(\frac{4(V_{dd}-V_{Tn})}{V_{dd}} - 1\right)\right)\cdot\frac{L}{\mu Cox(V_{dd}-V_{Tn})} \quad (1)$$

[7]. Equation (1) does not account for the input slope and other second order effects. Instead of deriving more complex equations, technology characterization (i.e. measuring with simulation) can be used. The measured (and other parameters) will reflect second order effects. Fig. 3 shows a Logical Effort (LE) [11] characterization setup for determining $\tau_n = \hat{R}_n Ct$ by measuring the fall delay of the shaded inverter. Note that the input-output slopes are made equal by proper loading. Each inverter has a load of *h* inverters. By varying the loading factor *h* the parameter of interest is extracted for different input/output slope conditions. Parameters obtained from this setup account for the input slope effect on circuit delay. However, equal input-output slope assumption of LE is not correct when circuit is sized for minimum energy. However, in [14] it is shown that this will not introduce a significant error.

A. Modeling Delay

Consider the buffer in Fig. 4. Using the R-C models for transistors, the rise delay can be written as the sum of the stage delays as

$$T_{HL} = R_{M2}C1 + R_{M3}C2 \quad (2)$$

Plugging in the resistance and capacitance values

$$\begin{aligned} T_{HL} = {} & \hat{R}_n/w_{M2}\,((w_{M1}+w_{M2})\,Cp \\ & + (w_{M3}+w_{M4})\,Ct) \\ & + \hat{R}_p/w_{M3}\,((w_{M3}+w_{M4})\,Cp + Cload) \end{aligned} \quad (3)$$

Factoring out time constant $\tau_n = \hat{R}_n Ct$

$$\mathrm{T}_{\mathrm{HL}} = \tau_{\mathrm{n}} \left(\begin{array}{l} \left(1+\frac{w_{\mathrm{M1}}}{w_{\mathrm{M2}}}\right)\alpha_{\mathrm{C}} + \frac{w_{\mathrm{M3}}+w_{\mathrm{M4}}}{w_{\mathrm{M2}}} + \\ \alpha_{\mathrm{R}}\left(\left(1+\frac{w_{\mathrm{M4}}}{w_{\mathrm{M3}}}\right)\alpha_{\mathrm{C}} + \frac{w_{\mathrm{load}}}{w_{\mathrm{M3}}}\right) \end{array} \right) \tag{4}$$

where $\alpha_{\mathrm{C}} = \mathrm{Cp}/\mathrm{Ct}$, $\alpha_{\mathrm{R}} = \hat{\mathrm{R}}_{\mathrm{p}}/\hat{\mathrm{R}}_{\mathrm{n}}$, and w_{load} is the transistor width that gives Cload equivalent capacitance. τ_n, α_C, and α_R are technology dependent parametric constants and can be determined from technology characterization.

B. *Modeling Energy*

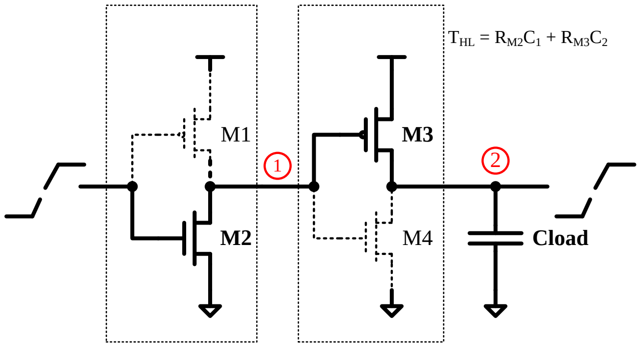
In determining optimal circuit size we consider dynamic energy because it has been shown that leakage energy is linearly proportional to transistor size [2]. Leakage energy shifts the E-D curve upwards without changing the shape of the curve. Short-circuit energy is consumed only when both pMOS and nMOS are conducting at the same time. It can be ignored for low supply-voltage values [15]. Therefore, minimization of energy over the dynamic energy is sufficient. Dynamic energy in a CMOS circuit is

$$E = \sum s_i C_i V^2 \tag{5}$$

where C_i is the node capacitance, s_i is the node switching activity, and V is the voltage swing (in general equal to supply voltage) at the node. Switching activity can be extracted by simulating the circuit and observing activity rates at all nodes. The node capacitance is composed of the parasitic and/or gate capacitances of transistors connected to the node. Capacitance of node *i* is

$$C_i = \sum_{j\in\mathrm{Node}_i} w_{\mathrm{M}_j}\cdot\mathrm{Ct} + \sum_{k\in\mathrm{Node}_i} w_{\mathrm{M}_k}\cdot\mathrm{Cp} \tag{6}$$

for all transistors M_j whose gate is connected to node *i* and all transistors M_k whose drain or source is connected to node *i*. Factoring out Ct gives



$\mathrm{R}_{\mathrm{M2}} = \mathrm{Rn}/w_{\mathrm{M2}}$

$\mathrm{R}_{\mathrm{M3}} = \mathrm{Rp}/w_{\mathrm{M3}}$

$\mathrm{C}_1 = \mathrm{Cd}_{\mathrm{M1}} + \mathrm{Cd}_{\mathrm{M2}} + \mathrm{Cg}_{\mathrm{M3}} + \mathrm{Cg}_{\mathrm{M4}}$

$= (w_{\mathrm{M1}} + w_{\mathrm{M2}})*\mathrm{Cp} + (w_{\mathrm{M3}} + w_{\mathrm{M4}})*\mathrm{Ct}$

$\mathrm{C}_2 = \mathrm{Cd}_{\mathrm{M3}} + \mathrm{Cd}_{\mathrm{M4}} + \mathrm{Cload}$

$= (w_{\mathrm{M3}} + w_{\mathrm{M4}})*\mathrm{Cp} + \mathrm{Cload}$

Fig. 4. Rising delay critical path of a buffer

$$C_i = \mathrm{Ct}\left(\sum_{j\in\mathrm{Node}_i} w_{\mathrm{M}_j} + \alpha_{\mathrm{C}}\sum_{k\in\mathrm{Node}_i} w_{\mathrm{M}_k}\right) \tag{7}$$

Assuming constant voltage swing for the whole circuit and re-writing (5) yields

$$E = \mathrm{Ct}V^2\sum s_i\left(\sum_{j\in\mathrm{Node}_i} w_{\mathrm{M}_j} + \alpha_{\mathrm{C}}\sum_{k\in\mathrm{Node}_i} w_{\mathrm{M}_k}\right) \tag{8}$$

It can be shown that (8) reduces to

$$E = \sum r_k w_k \tag{9}$$

where r_k is the appropriate weighing factor for transistor *k*.

C. *The optimization problem*

E-D curves can be generated by setting a delay target and optimizing for energy. For a CSE, the critical delay is the data-to-output delay: $t_{\mathrm{D\text{-}Q}}$ [8,12]. The final formulation for minimizing energy given a delay target T is

Minimize

$$\mathrm{Ct}V^2\sum s_i\left(\sum_{j\in\mathrm{Node}_i} w_{\mathrm{M}_j} + \alpha_{\mathrm{C}}\sum_{k\in\mathrm{Node}_i} w_{\mathrm{M}_k}\right)$$

Subject to rise/fall delay constraints

$$\mathrm{THL(D)}\rightarrow\mathrm{(Q)} \le \mathrm{T}$$

$$\mathrm{TLH(D)}\rightarrow\mathrm{(Q)} \le \mathrm{T} \tag{10}$$

where the objective function and delay expressions are functions of the transistor sizes $w_{\mathrm{M}i}$.

Using the RC modeling for delay equations, the problem turns to be a geometric programming problem which can be solved using convex optimization [4]. Application to a clocked storage element is straightforward: The critical paths for data-to-output delay are identified. Delays along these paths in terms of the transistors sizes are written. For energy, node capacitances are written in terms of transistor sizes. Then, given a delay target T, the problem of (11) is solved for transistor sizes with a geometric programming solver.

III. Design Examples

The effectiveness of the methodology is demonstrated on different state-of-the-art CSEs taken from literature, namely the UltraSparc edition of the semi-dynamic Flip-Flop (USPARC) [8] and the transmission gate master-slave latch (TGMS)[9]. The two examples, USPARC and TGMS, are both used in industry and are representatives of high-performance and low-power designs, respectively.

For the transistors models we used high-performance Predictive Technology Model (PTM) for 45nm technology node [12]. Some key features and extracted optimization parameters of the technology node are summarized in Table I. The CSEs are tested using the setup shown in Fig. 7 [1]. The size of the clock driver (w_{ck}) is adjusted to produce a FO2

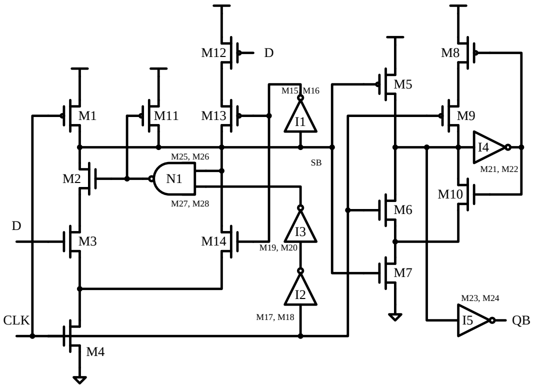


Fig. 5. Ultrasparc flipflop (USPARC)

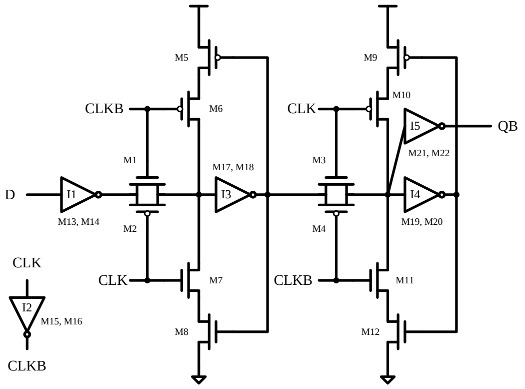


Fig. 6. Transmission-gate master-slave latch (TGMS)

slope at the clock input of the CSE. The output load (Ld_{out}) is set to 42 minimum transistors size (42x) equivalent capacitance. This is equivalent to 14 minimum sized inverters. For all experiments, the input loading (Ldin) is less than or equal to 9x.

TABLE I
Technology dependent parameters of PTM technology nodes

Tech.	Vdd (V)	w_{min}(nm)	τ_n (ps)	α_C	α_R	Ct(fF)
45nm	1.0	90	0.90	1.76	1.9	0.126

TABLE II.
Transistor sizing range for exhaustive search.

USPARC	TGMS
wM1 = [1:3]	wM1 = [1:4]
wM2 = wM3 = wM4 = [2:8]	wM2 = [1:5]
wM5 = [1:12]	wM3 = [1:8]
wM6 = wM7 = [1:3]	wM4 = [1:6]
wM24 = [1:6], wM23 = 2*wM24	wM14 = [1:3], wM13 = 2*wM14
	wM18 = [1:5], wM17 = 2*wM18
	wM22 = [1:6], wM21 = 2*wM22

E-D curves for USPARC and TGMS are generated by the sizing optimization methodology. For USPARC the delay target range was chosen to be 25ps to 65ps with a 1ps interval. For TGMS it was chosen to be from 60ps to 100ps. Optimizations resulted in transistor sizes for approximately 40 E-D points.

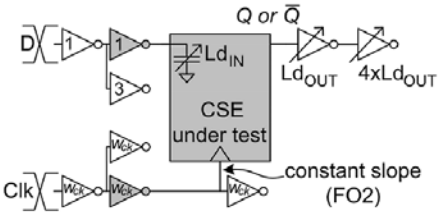


Fig. 7. Spice simulation set-up for CSEs

Using these sizes in spice simulation, the exact delay and energy values are obtained for both CSEs. To find out the correctness of our sizing solutions a comparison with exhaustive search is done. The sizing ranges used for transistors in exhaustive search are given in Table II. The sizes of transistors not given in the table are assigned unit size (i.e. w_{min}). The comparison of results produced by the two methods is shown in Fig. 8. The points for TGMS obtained by exhaustive search are the minimum energy solutions for each delay point selected among ~9000 points. Similarly, for USPARC ~5000 points are simulated among which the minimum energy solutions are plotted in Fig. 8. Both curves are H-SPICE simulation results. The results obtained from exhaustive search do not show a monotonic characteristic. There are jumps and gaps. This is because in order to avoid excessive run-time we have restricted the number of sizes a transistor can have. For example transistor M1 of USPARC has a sizing range of 1 to 3 minimum transistor (w_{min}) width with a step of 1 w_{min}. That is, it is allowed to have 3 possible values { w_{min},2* w_{min},3* w_{min} }. If the step size was set to 0.5* w_{min} each transistor would have twice the number of possible sizing values. However, if the size set of all transistor would have doubled the number of points to be simulated would have increased from ~5000 to ~160000. The search for optimal sizes in exhaustive search is done by simulating each possible sizing combination of the transistors in H-SPICE. For the proposed method, transistor sizes are obtained from optimization and then only the optimized sizings are simulated with H-SPICE.

The optimizer finds the optimum transistor sizes based on the delay and energy models described in Section 2. A comparison of the E-D plots with model estimates and spice simulation results for the same transistor sizes are given in Fig. 9. For both CSEs, the estimation agreement with H-SPICE at the low energy region is deteriorated. This is because at this region, transistors are close to minimum size and optimization is not effective. Nevertheless, for USPARC a maximum of 10% deviation is observed whereas it is only 5% for the TGMS. For the energy model, we expect more deviation, since we have not accounted for short circuit and leakage energy. Therefore, the model estimates are expected to show less energy as compared to H-SPICE as seen in Fig. 9. Furthermore, the unit size gate capacitance Ct is a non-linear function of the voltage. The measured Ct is an average value. Since it is a constant factor in the model, an estimation error in Ct is directly reflected to the estimated energy value. Regardless, the optimal transistor sizing is achieved and is not affected by these deviations.

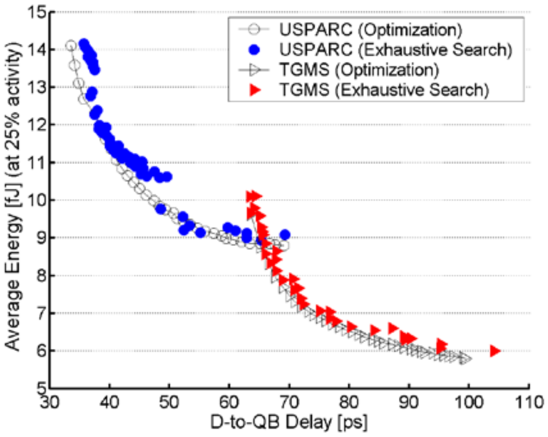


Fig. 8. Comparison of E-D plots generated by optimization and exhaustive search for USPARC and TGMS.

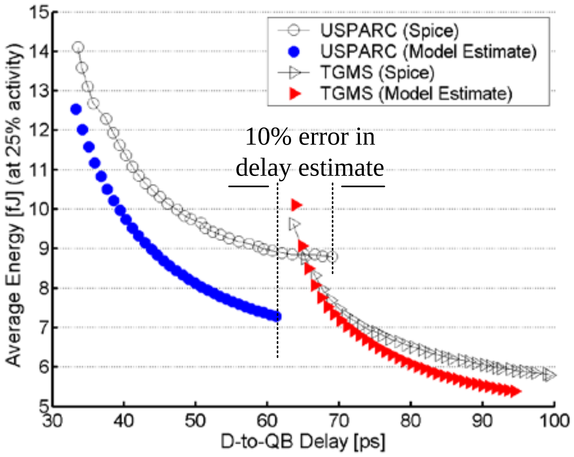


Fig. 9. Estimation error for the delay/energy model used in optimization.

IV. Conclusion

A methodology for rapid transistor sizing for optimal Flip-Flops configurations is presented. It is simple, fast, and accurate. It can be applied to any Flip-Flop or circuit topology. Despite the inaccuracies of the energy and delay models used in optimization, the transistor sizes found give comparable results to time consuming exhaustive search which is able to find optimum sizes when sufficient computing time is available. The total number of spice simulations is reduced by two orders of magnitude when compared to exhaustive search.

Acknowledgements

This work has been supported by SRC Research Grant No. 2008-HJ-1799.

References

[1] Giacomotto, C., Nedovic, N., Oklobdzija, V. G., “The Effect of the System Specification on the Optimal Selection of Clocked Storage Elements,” IEEE J. Solid State Circuits, vol. 42, no. 6, pp. 1392-1404, June 2007.

[2] Dao, H. Q., Zeydel, B.R., Oklobdzija, V.G., “Energy Optimization of Pipelined Digital Systems Using Circuit Sizing and Supply Scaling,” IEEE Transactions on VLSI Systems, vol. 14, no. 2, pp. 122-134, February 2006.

[3] Fishburn, J., Dunlop, A.: TILOS, “A Posynomial Programming Approach to Transistor Sizing,” In Proceedings of the IEEE International Conference on Computer-Aided Design, pp. 326-328, 1985.

[4] Sapatnekar, S. S., Rao, V. B., Vaidya, P. M., Kang, S. M. “An Exact Solution to the Transistor Sizing Problem for CMOS Circuits Using Convex Optimization,” IEEE Trans. Computer-Aided Design, vol. 12, pp. 1621-1634, Nov. 1993.

[5] Joshi, S., Boyd. S., “An Efficient Method for Large-Scale Gate Sizing,” IEEE Trans. Circuits Syst. I, vol. 55, no. 9, pp. 2760-2773, Oct. 2008.

[6] Uyemura, J.P.: CMOS Logic Circuit Design, 4th edition, Springer, 1999.

[7] Stojanovic, V., Oklobdzija, V. G., “Comparative Analysis of Master-Slave Latches and Flip-Flops for High-Performance and Low-Power Systems,” IEEE J. Solid State Circuits, vol. 34, pp. 536-548, 1999.

[8] Heald, R., et al., “A Third-Generation SPARC V9 64-b Microprocessor,” IEEE J. Solid-State Circuits, vol. 35, no. 11, pp. 1526-1538, Nov. 2000.

[9] Gerosa, G., et al., “A 2.2 W, 80MHz Superscalar RISC Microprocessor,” IEEE J. Solid-State Circuits, vol. 29, no. 12, pp. 1440-1454, Dec. 1994.

[10] Sutherland, I., Sproull, B., Harris, D., “Logical Effort: Designing Fast CMOS Circuits,” San Francisco, CA: Morgan Kaufmann, 1999.

[11] Oklobdzija, V. G., Stojanovic, V. M., Markovic, D. M., Nedovic, N. M., Digital System Clocking: High Performance and Low-Power Aspects. 1st edition New York: Wiley/IEEE Press, 2003

[12] Predictive Technology Model, http:// http://www.eas.asu.edu/~ptm/

[13] Zyuban, V., Strenski, P.N., “Unified Methodology for Resolving Power Performance Tradeoffs at the Microarchitectural and Circuit Levels,” In Proc. Int. Symp. Low Power Electronics and Design (ISLPED), 2002, pp. 166-171.

[14] Zeydel, B. R., Oklobdzija, V. G., “Methodology for Energy-Efficient Digital Circuit Sizing: Important Issues and Design Limitations,” 16th international Workshop on Power and Timing Modeling, Optimization and Simulation (PATMOS), Montpellier, France, 2006.

[15] Vratonjic, M., Zeydel, B. R., Oklobdzija, V. G., “Circuit Sizing and Supply-Voltage Selection for Low-Power Digital Circuit Design,” 16th international Workshop on Power and Timing Modeling, Optimization and Simulation (PATMOS), Montpellier, France, 2006.