Fast and Accurate Workload-Level Neural Network Based IC Energy Consumption Estimation

N. Cucu Laurenciu, S.D. Cotofana
Computer Engineering Laboratory,
Delft University of Technology, The Netherlands.
{N.CucuLaurenciu, S.D.Cotofana}@tudelft.nl

***Abstract*—A fast, yet accurate nanoscale IC energy estimation is a design-time desideratum for area-delay-power-reliability optimized circuits and architectures. This paper introduces an IC energy estimation approach, which instead of sequentially propagating workload vectors throughout the circuit, relies on an one time propagation of the workload statistics. To this end, the basic gates need be SPICE pre-characterized with respect to (w.r.t.) static and dynamic energy consumption per input transition type and Neural Network based gate models constructed and trained in order to estimate gate output statistics and consumed energy based on gate input statistics, i.e., the $'0' \rightarrow '0'$, $'0' \rightarrow '1'$, $'1' \rightarrow '0'$, and $'1' \rightarrow '1'$ transition probabilities. Both pre-characterization and training are done once per technology node and do not contribute to the actual evaluation time. In this way, regardless of n, the number of workload input vectors, by propagating signal statistics instead of logic values the overall circuit energy consumption is evaluated in one instead of n circuit traversals. Moreover, as opposed to the constant and equal gate delay assumption utilized in state of the art energy estimation methods, the proposed approach takes into account the real gate propagation delays, which yields estimates that are closer to the actual energy figures. We evaluated with the proposed method the static and dynamic energy consumption for a set of ISCAS'85 circuits and a $10,508$-gate hashing circuit, using TSMC 40nm CMOS technology, and $50,000$-vector workloads. The experiments indicate that our method provides an estimation error below 2.6% and 1.5% for dynamic and static energy, respectively, when compared to the accurate SPICE measurements, while providing an estimation speedup in the order of $50,000$x.**

***Index Terms*—Energy Estimation, Neural Networks.**

I. Introduction

Energy consumption estimation and analysis forms an integral constituent of effective nanoscale IC design and optimization, and in order to enable a shorter development cycle and thus to reduce the IC time-to-market, a fast and yet accurate energy estimation mean is desirable.

In practice, if one is interested in evaluating the energy consumed by a circuit under specific workloads one has to either employ (i) a low-level method in a transition by transition evaluation fashion (approach extremely accurate, though rather tedious, especially if the circuit under analysis is large), or (ii) a less accurate, high-level, fast and tractable approach, that exploits the workload characteristics or other high-level information and preferably uses as little circuit passes as possible. In the (ii) context, Neural Networks (NNs) were employed , e.g., [1], [2], [3], [4], [5], [6], however, with a few exceptions, e.g., [3], the NN based energy estimation techniques are applied directly at the circuit level, requiring the re-building of the entire NN macro-model for every new circuit. Furthermore, existing NN-based power estimation methods preponderantly assume either a unitary gate delay model (i.e., the gate propagation delay is constant for all circuit gates, that is one unit of time), e.g., [3], or a zero gate delay model (i.e., no gate delays are considered at all), e.g., [4], [6]. These assumptions however lead to a loss of power/energy estimation accuracy as in practice a gate propagation delay is for instance a non-linear function of gate fan-in, fan-out, and gate inputs slew rate. Consequently, such approaches cannot capture spurious transient currents - glitches - which are possibly accounting for $10\% - 40\%$ of the total power consumption [7], [8], [9], depending on e.g., circuit topology, technology, logic style. Another potential inaccuracy source of previous NN based power estimation approaches, relates to the fact that they disregard short-circuit currents (which can account for $10\% - 20\%$ of the total power consumption [9]), as power contributors, e.g., [6]. Not considering the circuit topology, e.g., [5], can also affect the estimation accuracy, as circuits with different gate types and counts can result in significantly different power/energy estimates.

In this paper we introduce in Section II a fast yet highly accurate method, which allows for static and dynamic energy evaluations by propagating through the circuit topology input data statistics instead of signal transitions. To this end our approach relies on: (i) static and dynamic energy consumption SPICE based pre-characterization of each and every basic Boolean gate for each possible input-output transition type and (ii) Neural Network based estimation models able to derive gate energy consumption and output data statistics as a function of gate type and input stimuli statistics. Given that workload statistics instead of individual input vectors are propagated through the circuit a single circuit traversal from its primary inputs to its primary outputs suffices to determine the static/dynamic energy consumption of all gates in the circuit. During this propagation the static/dynamic energy is computed as a sum of static/dynamic energies (corresponding to the gate output transition types), weighted by the afferent gate input statistics for each circuit gate and the total static/dynamic energy consumed by the circuit is estimated by summing up the static/dynamic energy values calculated for all circuit gates. Moreover, in contrast with the state of the art zero/unitary gate delay assumption, the proposed methodology accounts for actual gate propagation delays (as during the SPICE based gates energy pre-characterization we measure the energy per transition for each gate type under different inputs and gate loading),

978-1-5090-5052-9/17/$31.00 ©2017 IEEE

which positively impact its estimation accuracy. We note that: (i) the pre-characterization and the NN model training are done only once per technology node for each Boolean gate type, and thus they are not inducing any overhead into the actual energy evaluation process, (ii) the NN complexity is related to the standard cells set size and complexity and bears no relation with the circuit and workload sizes, and (iii) the proposed framework can easily accommodate the effects of process and environmental variations during the standard cells energy pre-characterization phase, by including into the SPICE simulation setup the aggression profiles to which circuits are subjected.

We evaluated the proposed methodology on a set of ISCAS'85 circuits, and a bigger $10,508$-gate Pearson hashing circuit. The simulation results, presented in Section III, indicate that when analyzing the static/dynamic energy estimation error variation with the workload size, which was varied from $0.5k$ to $50k$ input vectors, we obtained an average value of 0.67% and 0.14%, standard deviation of 0.40% and 0.12%, and maximum value of 1.32% and 0.35% for the c6288 circuit dynamic and static components, respectively. For the considered circuits and $50k$-size workload, a fairly good agreement of the dynamic/static energy estimates with the accurate SPICE measured reference energy counterparts was observed. Specifically, we obtained a maximum energy estimation error below 2.6% and 1.5% for dynamic and static energy, respectively, for all considered circuits.

II. Workload based Energy Evaluation

A CMOS gate total dynamic energy consumption can be written as:

$$E_{dyn}^{GATE} = E_{capacitive} + E_{short-circuit} + E_{glitch}, \quad (1)$$

which can be approximated as the weighted sum of the dynamic energies afferent to the 4 possible gate output transitions (i.e., $'0' \rightarrow '0'$, $'0' \rightarrow '1'$, $'1' \rightarrow '0'$, and $'1' \rightarrow '1'$), as:

$$\hat{E}_{dyn}^{GATE} = N_{00} \cdot E_{00}^{dyn} + N_{01} \cdot E_{01}^{dyn} + N_{10} \cdot E_{10}^{dyn} + N_{11} \cdot E_{11}^{dyn}, \quad (2)$$

where N_{jk} with $j, k \in \{0, 1\}$ denotes the number of times the gate output is transitioning from state j to state k and E_{jk} is the dynamic energy consumed by the gate during a j to k transition. Consequently, the circuit dynamic energy consumption can be computed as the sum of the dynamic energy consumed by each of its gates:

$$E_{dyn}^{CIRC} = \sum_{i=1}^{N_{GATES}} E_{dyn}^{GATE_i}. \quad (3)$$

For the static energy component, (2) becomes:

$$\hat{E}_{st}^{GATE} = N_{00} \cdot E_{00}^{st} + N_{01} \cdot E_{01^{st}} + N_{10} \cdot E_{10}^{st} + N_{11} \cdot E_{11}^{st}, \quad (4)$$

and (3) rewrites $E_{st}^{CIRC} = \sum_{i=1}^{N_{GATES}} E_{st}^{GATE_i}$. To evaluate E_{st}^{CIRC} and E_{dyn}^{CIRC} a 2-phase approach is undertaken: (i) a standard cells SPICE pre-characterization phase, which is a one-time per technology process, detailed in Section II-A, and (ii) a circuit static/dynamic energy evaluation phase, detailed in Section II-B.

TABLE I: Gate Statistics.

Statistic	Expression
Prob. of logic $'1'$	$P_{H_x} = \frac{\sum x_i}{N}$
Prob. of $'0' \rightarrow '0'$	$P_{00_x} = \frac{\sum (x_i \vee x_{i+1} = '0')}{N-1}$
Prob. of $'0' \rightarrow '1'$	$P_{01_x} = \frac{\sum ((x_i = '0') \wedge (x_{i+1} = '1'))}{N-1}$
Prob. of $'1' \rightarrow '0'$	$P_{10_x} = \frac{\sum ((x_i = '1') \wedge (x_{i+1} = '0'))}{N-1}$
Prob. of $'1' \rightarrow '1'$	$P_{11_x} = \frac{\sum (x_i \wedge x_{i+1} = '1')}{N-1}$

A. Standard Cells Pre-Characterization

Standard cells pre-characterization is performed one-time for a given technology, and involves for each gate type: (i) the measurement of the values of static/dynamic energy per transition type (E_{00}^{st}, E_{01}^{st}, E_{10}^{st}, and E_{11}^{st} in (4) and E_{00}^{dyn}, E_{01}^{dyn}, E_{10}^{dyn}, and E_{11}^{dyn} in (2)), by means of SPICE simulation, and (ii) the construction and training of a Neural Network (NN) able to estimate the gate output signal transition counts (N_{00}, N_{01}, N_{10}, and N_{11} from (2) and (4)), based on a reduced set of input signals statistics. As the accuracy of the NN estimates highly depends on the training process effectiveness, a variety of configurations ought to be considered for each gate type, as to account for the energy dependency on the gate varying input signals slopes (from upstream gates) and load capacitance (downstream gates). Judiciously choosing during the NN construction of a relevant and limited set of statistical input/output parameters, as well as generating - during the NN training - gate input vectors which span as much as possible the range of values of the input statistical parameters, are essential for the NN estimation accuracy. Subsequently, we detail the main aspects of the NN construction and training.

1) NN Input/Output Statistical Parameters

As the gate input signals exhibit a correlation profile both temporal (i.e., dependencies between the same or different input signals at different time moments), as well as spatial (i.e., dependencies between input signals at the same time moment), several signal statistic parameters can be devised (e.g., probability of logic $'1'$, probability of switching, auto-correlation, cross-correlation). To guide the selection of the parameters which are most relevant from the energy estimation point of view, a Principle Component Analysis (PCA) [10] was conducted, serving as an NN front-end to reduce the dimensionality of the nominal input dataset, and thus to reduce the NN learning time, and to avoid over fitting the training data. However, one may note that if the dataset analyzed by PCA exhibits outliers, or nonlinear relationships between variables, the PCA analysis may not succeed in exposing the underlying connections. Based on the PCA results, trials, and convenience from the computational standpoint, 5 statistics, as summarized in Table I, were selected to characterize a gate input/output signal x (with x_i, $i = 1 \ldots N$ denoting its N samples): (i) P_{H_x} - probability of logic $'1'$, (ii) P_{00_x} - probability of $'0' \rightarrow '0'$ transitions, (iii) P_{01_x} - probability of $'0' \rightarrow '1'$ transitions, (iv) P_{10_x} - probability of $'1' \rightarrow '0'$ transitions, and (v) P_{11_x} - probability of $'1' \rightarrow '1'$ transitions. It should be mentioned

here that while gate input signals temporal correlations are taken into consideration by the NN statistical parameters, our simulation results indicate that the NN model was able to inherently account for spatial correlations, alleviating thus the need of additional explicit NN input/output parameters. As an example, for a 2-input gate, there are 10 input statistics (5 for each gate input signal) and 5 output statistics, the NN defining a mapping as follows:

$$f : \left\{ P_{H_{in_j}}, P_{00_{in_j}}, P_{01_{in_j}}, P_{10_{in_j}}, P_{11_{in_j}} \right\} \longrightarrow \{ P_{H_{out}}, P_{00_{out}}, P_{01_{out}}, P_{10_{out}}, P_{11_{out}} \}, \qquad j = 1, 2.$$

2) NN Parameters & Training

Having selected the most relevant statistical parameters we now need to create the gate model able to operate on them rather than on standard Boolean values. As NNs allow for an automated abstraction of the relationship between outputs and their inducing factors, they exhibit direct beneficial implications over an analytical approach, especially for more complex, non-linear relationships. For illustration we targeted 2-input gates only. We employed a 2-layer feed-forward neural network, with 10 inputs (corresponding to the 10 input statistics for a 2-input gate), 10 neurons in the hidden layer, and 5 neurons in the output layer (corresponding to the 5 output statistics). For gates with larger fan-in, the same NN structure can be employed, but with a different number of input statistics (5 for each gate primary input) and possibly more neurons in the hidden layer. As concerns the activation functions, we employed a radial basis sigmoid function in the hidden layer, and a linear transfer function in the output layer. To derive an input dataset for the neural network, $M = 1000$ samples of $N = 5000$ bits per sample per gate primary input, were generated. The values of M and N are selected for the purpose of illustration, and for simulation convenience (as they impact directly the training process convergence duration), without restricting their generality (M and N can be chosen with regard to the required workload dimensions). We note that bigger M and N values increase the NN training required time, however they do not impact the actual computation time for the circuit energy evaluation. Based on these M samples, the 10 input statistics from Table I were computed. In order to derive the 5 reference output statistics that we would like to approximate, the gate N-bit output sequence was first computed. For a 2-input gate, the NN input dataset of size $M \times 10$ (as each of the M N-bit workload samples generates 10 input statistics) was partitioned into 80% for training, 10% for validation, and 10% for testing.

B. Circuit Energy Evaluation Phase

For a given circuit topology, and workload of size n, both the consumed static and dynamic energies can now be evaluated in a single traversal pass through the circuit, instead of the n iterations. To this end, the circuit gates are traversed one time in their processing order, and for each gate, (2) (and (4)) are evaluated with (i) the energy per transition values E_{00}^{dyn}, E_{01}^{dyn}, E_{10}^{dyn}, and E_{11}^{dyn} (and E_{00}^{st}, E_{01}^{st}, E_{10}^{st}, and E_{11}^{st}), known from the standard cells pre-characterization phase in Section II-A, and with (ii) the transition counts N_{00}, N_{01}, N_{10}, and N_{11} determined for the primary inputs and estimated for all other gates (based on the gate input statistics via the gate characteristic NN model developed in Section II-A).

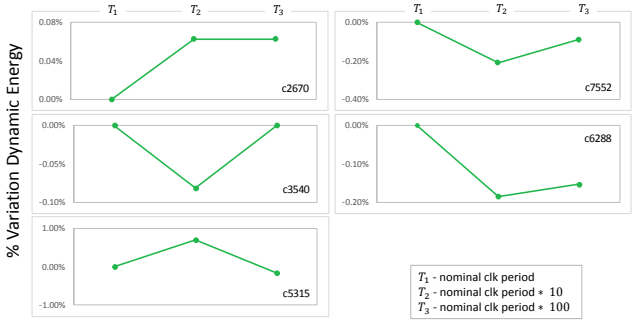

Fig. 1: Dynamic Energy Variation vs. Clock Period.

III. Simulation Results

To evaluate the accuracy of the proposed method we utilized as simulation vehicles a set of combinational ISCAS'85 circuits with gate cardinality ranging from 817 to 2837, and a bigger 10,508-gate circuit, which implements a Pearson hashing function [11].

To obtain the power consumption reference values, we performed transistor level SPICE simulation, via Synopsis NanoSim [12], which allows for a faster full chip power profile characterization with $2\% - 5\%$ loss of accuracy when compared to the industry's "gold standard" HSPICE [12]. We opted for transistor level reference values, as gate level power evaluation tools that employ switching activity statistics, allow for power evaluation in a single pass, but are less accurate. For instance, Synopsis Power Compiler [12] yields 37% post-synthesis and 30% post-layout total power consumption estimation error with respect to the NanoSim values, for the ISCAS'85 c6288 circuit.

We note that depending on the circuit topology and environmental conditions (e.g., supply voltage variation) an energy per transition value can contain both dynamic (capacitive + short-circuit + glitch) and static (leakage + other DC paths) components. Thus, even if under most circumstances only one component is predominant, considering all 4 gate output transitions to reflect the dynamic component leads to more accurate estimates.

As energy is time dependent, the 8 measured energy per transition values E_{00}^{dyn}, E_{01}^{dyn}, E_{10}^{dyn}, and E_{11}^{dyn} (E_{00}^{st}, E_{01}^{st}, E_{10}^{st}, and E_{11}^{st}) also depend on the clock period T_{clkG} utilized during the SPICE gate evaluations. Given that in practice the circuit under evaluation clock period T_{clkC} differs from the clock period used during the gate pre-characterization phase, arises the question of whether energy per transition values re-measurements are necessary. In our experiments we assume that the E_{00}^{dyn}, E_{01}^{dyn}, E_{10}^{dyn}, and E_{11}^{dyn} do not need to be readjusted with respect to the clock period change, as the main dynamic energy contributors are E_{01}^{dyn} and E_{10}^{dyn}, which

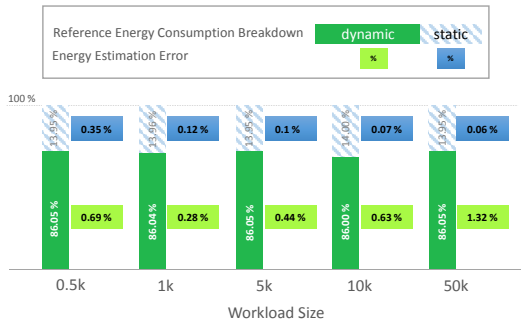

Fig. 2: c6288 Energy Estimation Error vs. Workload Size

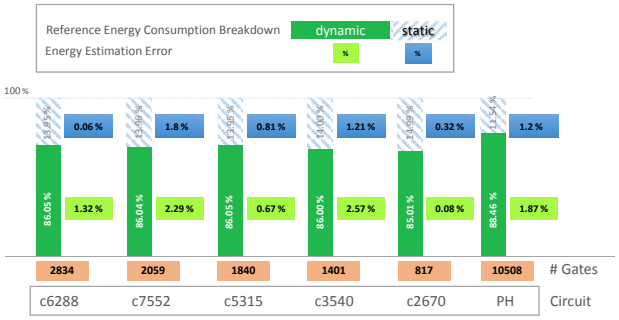

Fig. 3: Energy Estimation Error for $50k$-Size Workload.

may depend on the input signal slew rate but not on the clock period. To gain insight into the implications of this assumption, we plot in Fig. 1 the percentage of variation of the dynamic energy measured in SPICE for each considered ISCAS'85 circuit, when the input data are sampled with the clock period $T_2 = T_{clkC} \times 10$ and $T_3 = T_{clkC} \times 100$ relative to the dynamic energy corresponding to the circuit minimum clock period $T_1 = T_{clkC}$. One can observe in the Figure a small percentage (less than 1%) of total dynamic energy deviation w.r.t. the energy measurements for the minimum clock period, which supports our assumption. The static counterparts E^{st}_{00}, E^{st}_{01}, E^{st}_{10}, and E^{st}_{11} vary proportionally with the clock period and the main contributors are E^{st}_{00} and E^{st}_{11}. This is taken into account during the evaluation by adjusting the static value as, e.g., $k \cdot E^{st}_{00}$, for $T_{clkC} = T_{clkG} \cdot k$. Thus, in principle neither the dynamic, nor the static energy per transition values, do not require re-measurement regardless of the circuit evaluation clock period.

Fig. 2 illustrates the static and dynamic energy estimation error variation with respect to the workload size (i.e., the number of input vectors), for the c6288 circuit. The estimation errors are caused by: (i) the transition counts N_{00}, N_{01}, N_{10}, and N_{11} estimation with the NN model, and (ii) the approximations (e.g., of (1) with (2) when assuming the correct values of the 4 transition counts). To assess the estimation error susceptibility to the circuit workload size, we considered input vector sizes ranging from 500 to $50,000$, and measured in SPICE the c6288 energy reference values. When estimating the energy with the proposed methodology, we obtained an average value of 0.67% and 0.14%, standard deviation of 0.40% and 0.12%, and maximum value of 1.32% and 0.35% for the c6288 circuit dynamic and static components, respectively. As indicated in Fig. 2, the estimation error obtained with the proposed method is relatively constant when increasing the workload size, as long as its statistics are similar, which implies a dynamic energy evaluation speed-up potential of $n\times$, n being the workload size (i.e., for a larger workload the circuit energy can be evaluated in a single pass instead of n passes).

Fig. 3 depicts the energy estimation errors for the considered combinational circuits, under $50k$ samples workloads. The results indicate that for all circuits the maximum static and dynamic estimation error is below 2.6% and 1.5%, respectively. One may note that the number of gates has little influence on the energy estimates accuracy, but it is rather the NN model training effectiveness that significantly affects the estimation errors.

IV. Conclusions

In this paper we proposed a fast and accurate Neural Network (NN) based IC energy estimation methodology, which instead of propagating every workload vector through the circuit topology, relies on workload statistics propagation. To this end basic gates are one-time SPICE pre-characterized w.r.t. static and dynamic energy per transition type, and NN models are constructed to estimate gate output statistics based on its input statistics. The method is fast (one circuit traversal per workload regardless of its cardinality), and accurate as it accounts for the real gate propagation delays. We evaluated the dynamic energy of a set of ISCAS'85 circuits and a $10,508$-gate hashing circuit, using TSMC 40nm CMOS technology, and $50k$-vector workloads. We obtained a fairly small estimation error below 2.6% and 1.5% for the dynamic and static energy estimates, respectively, when compared to the accurate SPICE measured counterparts.

References

[1] Hsu, C. Y., and Hsieh, W. T., and et al., "A Tableless Approach for High-Level Power Modeling Using Neural Networks." in *Journal of Information Science and Engineering*, 2007, pp. 71–90.

[2] Cao L., "Circuit Power Estimation Using Pattern Recognition Techniques." in *IEEE/ACM International Conference on Computer-Aided Design*, 2002, pp. 412–417.

[3] Macii, E., and Poncino, M., "Estimating Power Consumption of CMOS Circuits Modeled as Symbolic Neural Networks." in *IEE Proceedings on Computers and Digital Techniques*, 1996, pp. 331–336.

[4] Gupta, S., and Najm, F., "Analytical Model for High Level Power Modeling of Combinational and Sequential Circuits." in *IEEE Alessandro Volta Memorial Workshop on Low-Power Design*, 1999.

[5] Hou, L., and Zheng, L., and Wu, W., "Neural Network Based VLSI Power Estimation." in *International Conference on Solid-State and Integrated Circuit Technology*, 2006, pp. 1919–1921.

[6] Bernacchia, G., and Papaefthymiou, M. C., "Analytical Macro-Modeling for High-Level Power Estimation." in *IEEE/ACM International Conference on Computer Aided Design*, 1999, pp. 280–283.

[7] Favalli, M., and Benini, L., "Analysis of glitch power dissipation in CMOS ICs." in *International Workshop on Low Power Design*, 1995.

[8] Bellaouar, A., and Elmasry, I., *Low-Power Digital VLSI Design: Circuits and Systems.* MA: Kluwer, 1995.

[9] Rabaey, J. M., and Pedram, M., *Low Power Design Methodologies.* MA: Kluwer, 1996.

[10] Byrne, C. L., *Signal Processing: A Mathematical Approach.* CRC Press.

[11] Pearson, P., "Fast Hashing of Variable-Length Text Strings." in *Communications of the ACM 30(6)*, 1990.

[12] Synopsys. [Online]. Available: http://www.synopsys.com/