ABSTRACT

The increasing demand for portable electronic devices has led to emphasis on power consumption within the semiconductor industry. As a result, chip designers are now forced to consider the impact of not only speed and area, but also power throughout the entire design process. Power reduction has to be addressed at every design level, like system, RTL, gate and transistor-level where most power can be saved at the highest level. Good power estimation is essential for successful power design. In order to evaluate how well a particular design type meets power constraints, designers have to often rely on CAD tools for power estimation. While tools have long existed for analyzing power consumption at the lower levels of abstraction like gate level and circuit level (PowerMill and SPICE) only recently tools with high-level power estimation capability are being developed. While system-level power estimation is new, power estimation from RTL-level and down have had many years to develop and mature. The RTL-power estimation can be divided into statistical- and simulation-based estimation. This paper surveys the various methods in high level power estimation, addressing techniques that operate at the RT Level of abstraction.

Index Terms: Low power, gate level estimation, CAD tools, statistical analysis, RTL power estimation, analytical methods

1. Introduction

RTL power estimation approaches can be classified into three broad categories, namely analytical techniques, characterization based macro-modeling, and fast synthesis based estimation. Different estimation techniques may be best suited to different parts of a design (such as arithmetic macro blocks, control logic, memory, clock network, and I/O). Analytical power modeling techniques attempt to correlate power consumption to measures of design complexity, using very little information from the functional specification. In characterization-based macro-modeling, a lower-level implementation of various RTL macro block characterized. A gate-level or transistor-level tool is used to estimate the power consumption of the macro block under various “training” input sequences and based on this data, a macro-model is constructed, which describes the power consumption of the block as a function of parameters, like, the signal statistics of its inputs and outputs. Fast synthesis refers to the process of performing a limited synthesis of the RTL description.

The design is mapped to a “meta-library” that typically consists of a small number of primitive cells smaller than a complete standard cell library. The resulting netlist is used for power estimation through gate-level simulation or probabilistic techniques.

Power estimation for a digital circuit involves two factors: how to model the circuit itself and how to model input signals. There are different techniques for both of them. Input signals can be generated as test vectors or they can be modeled probabilistically or statistically. Circuit macro-modeling techniques can also be further classified into two categories: those that use power coefficients measured on sample circuits, and those that analyze current and voltage from an equivalent circuit model. Given an input signal models, the macro-model estimates maximum power dissipation, or average power dissipation, or total energy consumed during certain cycle. The main challenge in establishing RTL power estimation methodology is the construction of efficient and accurate macro-models of the power dissipation. Such macro-models should be automatically built and should produce reliable average power estimates. At RT level designs are usually described hierarchically. The main challenge in estimating the power dissipation of a hierarchical design is the construction of accurate black-box power models for the leaves of the hierarchy, when only functional descriptions are available.

2. RT-Level Power Estimation Modeling

The lowest level is register-transfer level of abstraction. At this level the primitives are functional blocks such as adders, multipliers, controllers, register files, and SRAM’s. The difficulty in estimating power at this level stems from the fact that the gate, circuit, and layout level details of the design may not have been specified. Moreover, a floor plan may not be available, making analysis of interconnect and clock distribution
networks difficult by a table lookup with necessary interpolation equations.

Dynamic Estimation

2.1 MACROMODELLING

Macro modeling consists of generating power models for some given input data statistics which are obtained by RTL simulation. Most existing approaches of statistical power estimation consider the signal probability and average switching activity of the input signal and use signal probability propagation methods. But there is no guarantee that the estimated power has any relation to what the circuit will actually dissipate. The average power is estimated by simulating the circuit by a large number of samples drawn from the population. Signal statistics collected from simulation runs.

Analytical macro-modeling In addition to input signal probability Pin and input transition density Din, the input parameters of equation-based macro-model also include a spatial correlation S metric that further improves the model accuracy, without considering temporal correlation Tin for IP based system. But also output metrics such as output signal probability Pout, output transition density Dout, spatial correlation metric S and a temporal correlation metric T models for circuit modules not considered. Such techniques, which are commonly known as power macro-modeling, consist of generating circuit capacitance models for some assumed data statistics or properties. The statistics of input data are gathered during behavioral simulation of the circuit.

3. POWER ESTIMATION METHODS


3.1.1 Analytical methods

Analytical methods attempt to relate the power consumption of a particular RTL description to fundamental quantities that describe the physical capacitance and activity of a design. Since design complexity is a measure of physical capacitance we can divide the techniques into complexity-based and activity-based models.

3.1.2 Complexity-based models

One method is to describe the complexity of chip architecture in terms of “gate equivalents”. Which specifies the approximate number of reference gates i.e. 2-input NAND’s that would be required to implement a particular function like a 16x16 multiplier. This number can be specified in a library database or provided by the user. The power required for each functional block can then be estimated by multiplying the approximate number of gate equivalents by the average power consumed by each gate. The expression for average power is given as

$$P = \sum_{i.e(fins)} GE_i (E_{yp} + C^2_{i} V_{dd}) f A_{int}$$
where \( GE_i \) is the gate equivalent count for functional block \( i \), \( E_{typ} \) is the average energy consumed by an equivalent gate when active, \( C_i \) is the average capacitive load per gate including fan-out and wiring, \( f \) is the clock frequency, and \( A_{int} \) is the average percentage of gates switching each clock cycle within functional block. One disadvantage of this technique is that all power estimates are based on the energy consumption of a single reference gate. This does not take into account different circuit styles, clocking strategies, or layout techniques. The approximation is particularly inaccurate for specialized blocks such as memories. By applying customized estimation techniques to the different design entities: logic, memory, interconnect, and clock the power consumed by a memory cell array is modeled as:

\[
P_{memcell} = \frac{2^k}{2} (C_{int}l_{column} + 2^{n-k}C_{tr})V_{dd}V_{swing}f
\]

where \( 2^k \) is the number of cells in a row, \( C_{int} \) is the wire capacitance per unit length, \( l_{column} \) is the memory column length, \( 2^{n-k} \) is the number of cells in a column, \( C_{tr} \) is the minimum size drain capacitance, and \( V_{swing} \) is the bitline voltage swing. The total chip logic power is estimated by multiplying the estimated gate equivalent count by the basic gate energy and the activity factor. The activity factor is provided by the user and assumed fixed across the entire chip and the interconnect length and capacitance is modeled based on Rent’s Rule. The clock capacitance is based on the assumption of an H-tree distribution network. The power dissipation of the various components of a typical processor architecture, including on-chip memory, busses, local and global interconnect lines, H-tree clock net, off-chip drivers, random logic, and data path, are expressed as a function of a set of parameters related to the implementation style and internal architecture of these components.

Typical on-chip memory (a storage array of 6-transistor memory cells) consists of four parts: the memory cells, the row decoder, the column selection, the read/write circuits. The power model for a cell array of \( 2n-k \) rows and \( 2k \) columns in turn consists of expressions for: (1) the power consumed by \( 2k \) memory cells on a row during one pre-charge or one evaluation; (2) the power consumed by the row decoder; (3) the power needed for driving the selected row; (4) the power consumed by the column select part; and (5) the power dissipated in the sense amplifier and the readout inverter.

Memory cell power Equation:

\[
P_{memcell} = 0.5V_{swing}2^k(C_{int} + 2^{n-k}C_{tr})
\]

where \( V_{swing} \) is voltage swing on the bit/bit line (which may be different for read versus write), \( C_{int} \) gives the wiring-related row capacitance per memory cell, and \( 2^{n-k}C_{tr} \) gives the total drain capacitances on the bit/bit line.

The advantage of these complexity-based estimation techniques is that they require very little information. One disadvantage of the complexity-based methods is that they do not model circuit activity accurately. An overall fixed activity factor is typically assumed and, be provided by the user. In reality, activity factors will vary with block functionality and with the data being processed. So even if the user provides an activity factor that results in a good estimate of the total chip power, the predicted breakdown of power between modules is likely to be incorrect.

### 3.1.3 Stochastic power analysis techniques

This method is based on an activity-sensitive macro model which maintains that switching activities of high order bits depend on the temporal correlation of data, whereas lower order bits behave randomly. The module is thus completely characterized by its capacitance models in the most significant bit and least significant bit regions.

Input-output data model;

\[
Power = 0.5V^2f(C_iE_i + C_oE_o)
\]

where \( C_i \) and \( C_o \) represent the capacitance coefficients for the mean activities of the input and output bits, respectively.

#### i. Activity-based models
Activity-based models use the concept of entropy from information theory as a measure of the average activity in a circuit. The basic idea is to relate the power that a functional block consumes to the amount of computational work it performs. Entropy is used for measuring computational work on the assumption that power is proportional to the product of physical capacitance and activity; hence use area as a measure of physical capacitance and entropy as a measure of activity:

### ii. Power Factor Approximation models

The energy models are parameterized in terms of complexity parameters and PFA proportionality constant.

For the memory, the storage capacity in bits is used and for the I/O drivers the word length alone is adequate. The weakness of fixed-activity models is that they do not account for the influence that data activity can have on power consumption. PFA uses an experimentally determined weighting factor to model the average power consumed by a given module over a range of designs. Dual bit type model [4] exploits the fact that, in the data-path, switching activities of high-order bit behave similarly to white noise data. Thus a module is completely characterized by its capacitance models in the most significant bit (MSB) and least significant bit (LSB) regions. The break-point between the two regions is determined based on the signal statistics collected from simulation runs. The simplest power macro-model, is a constant type model which uses an experimentally determined weighting factor to model the average power consumed by a given module per input change. But this technique does not account for the data dependency of the power dissipation.

\[
\text{Power} = 0.5V^2n^2C_{\text{factiv}}
\]

where \( V \) is the supply voltage level, \( C \) is the capacitive regression coefficient, and \( f_{\text{activ}} \) is the activation frequency of the module.

### iii. Transition-sensitive energy model

Considers input transitions than input statistics. Then the energy model is provided for each functional unit, and the power consumed for each input transition is given as a table. Closely related input transition vectors and energy patterns are collapsed into clusters, thereby reducing the size of the table. After the energy models are built, it is not necessary to use any knowledge of the unit’s functionality or to have prior knowledge of any input statistics during power analysis. [12]

#### 3.1.4 Look-up-table (LUT) based macro-model:

The LUT stores the estimates for equi-spaced discrete values of the input signal statistics. The interpolation method can be used for estimates, if the input statistics do not correspond to LUT. Another inputs/outputs (I/O) based model was presented in [4] to capture the relation between power and input signal probability, input transition density, and output transition density. These methods only provide information about the average power consumption over relatively large number of clock cycles. The individual models built in this way are relatively accurate but overall accuracy is affected due to the incorrect input statistics or the inability to model the interactions correctly. [8]

**i. 3-D table, power macro-modeling technique:**

- Captures the dependence of power dissipation in a combinational logic circuit on the average input signal probability, the average switching activity of the input lines, and the average (zero-delay) switching activity of the output lines. The latter parameter is obtained from a fast functional simulation of the circuit.

#### 3.1.5 Regression-Based Models

The flow consists of the following steps:

- Characterize every component in the high-level design library by simulating it under pseudorandom data and fitting the power macro-model equation to the power dissipation results using a least mean square error fit.
- Extract the variable values for the macro-model equation from either static analysis of the circuit.
structure and functionality, or by performing a
behavioral simulation of the circuit using a co-
simulator linked with RT simulator to collect input
data statistics for various RT-level modules in the
design.c. Evaluate the power macro-model equations
for high-level design components in the library by
plugging the parameter values into the corresponding
macro-model equations. d. Estimate the power
dissipation for random logic or interface circuity by
simulating the gate-level description of these
components, or by performing probabilistic power
estimation. The low-level simulation can be of
statistical sampling techniques or automata-based
compaction techniques. Most RT-level power
estimation techniques use regression-based, switched
capacitance.

i. Bitwise data model:

\[
\text{Power} = 0.5V^2f \sum_{i=1}^{n} C_i E_i
\]

where \( n \) is the number of inputs for the module in
question, \( C_i \) is the (regression) capacitance for input
pin \( i \), and \( E_i \) is the switching activity for the \( i \)th pin of
the module. This equation can produce more accurate
results by including, spatial-temporal correlation
coefficients among the circuit inputs.

4. CENSUS MACRO MODELING:

RT level power evaluation can be implemented in the
form of a power co-simulator for standard RT level
simulators. The co-simulator is responsible for
collecting input statistics from the output of the
behavioral simulator and producing the power value at
the end. If the co-simulator is invoked by the RT level
simulator every simulation cycle to collect activity
information in the circuit it is called census macro
modeling. Evaluating the macro model equation at
each cycle during the simulation is actually a census
survey. The overhead of data collection and macro
model evaluation can be high.

Sampler macro modeling

To reduce the run time overhead Hsieh et al use simple
random sampling to select a sample and calculate the
macro model equation for the vector pairs in the sample
only. The sample size is determined before simulation The
 sampler macro modeling randomly selects \( n \) cycles and
marks those cycles. When the behavioral simulator
reaches the marked cycle the macro modeling invokes the
behavioral simulator for the current input vectors and
previous input vectors for each module. The input
statistics is only collected in these marked cycles. The
macro model equation is developed by using a training set
of input vectors. One way to reduce the gap between the
power macro model equation and the gate level power
estimation is to use a regression estimator. The adaptive
macro modeling thus invokes a gate level simulator on a
small number of cycles to improve the macro model
equation estimation accuracy. In this manner the bias of
the static macro models is reduced or even eliminated.

5. CYCLE ACCURATE POWER
ESTIMATES:

Clustering approach:

This approach relies on the assumption that closely
related input transitions have similar power dissipation. Hence, each input pattern is first mapped
into a cluster, and then a table lookup is performed to
obtain the corresponding power estimates from pre-
calculated and stored power characterization data for
the cluster. But the number of clusters has to be
relatively small, which would introduce errors into the
estimation result. And also the assumption that closely
related patterns i.e, patterns with short Hamming
distance result in similar power distribution may be
quite inaccurate, especially when the mode-changing
bits are involved; that is, when a bit change may cause
a dramatic change in the module behavior. The
abovementioned macro-models are (multi-cycle)
cumulative, in the sense that they can be used to
predict the average power under a sequence of input
vectors. In some applications, it is essential to estimate
the circuit power on a cycle-by-cycle basis. Addressing
this need, a cycle-accurate power macro-model is introduced in [23]. We can write Equation as:

\[ P_{wrk} = 0.5V^2 \sum_i C_i E_{i,k} \]

where \( P_{wrk} \) denotes the power consumption of the module at cycle \( k \), \( E_{i,k} \) is the switching activity (it can assume a value of either 0 or 1) for the \( i \)-th input of the module at cycle \( k \) and is obtained from functional simulation of the system in which the module is placed. The above equation illustrates that macro-modeling can be used to estimate the power consumption at each cycle; [23]

5.1 4D TABULAR MACRO MODEL:

All the approaches discussed above are limited to only combinational circuits. A power macro modeling approach for both combinational and sequential circuits that (1) takes into account the effect of the circuit input switching activity and does not treat the circuit inputs as white noise, (2) takes into account input correlation, both spatial and temporal and (3) is based on a single fixed macro model template which does not depend on the type of circuit being analyzed. The model is equation-based as a quadratic or cubic equation in the following four variables: average input signal probability \((P_{in})\), average input switching activity \((D_{in})\), average input spatial correlation coefficient \((S_{cin})\), and average output zero delay switching activity \((D_{out})\). The main advantage of this approach is that all types of circuits are treated in the same way, i.e., it does not use different model equation types for different modules. As a result, the method is very easy to use, and requires no user intervention.

CONCLUSION:

It is widely recognized that power consumption has become a critical issue in the development of digital systems. Since gate-level power estimation can be time-consuming and because power estimation from a high level of abstraction is desirable so as to reduce design time and cost, power macro modeling approaches for combinational and sequential circuits have been modeled. This paper provided a non-exhaustive review of existing methodologies for high-level power modeling and estimation.

REFERENCES:

[4] Power Analysis and Optimization Prof. Kurt Keutzer EECS University of California With thanks to D. Chinnery, UCB, S. Devadas, MIT, and D. Sylvester, U of Michigan


[19] High Level Power Modeling Estimation and Optimization Enrico Macii Massoud Pedram Fabio Somenzi