A Scalable Approach for Throughput Estimation of Timing Speculation Designs

Viraj Athavale, Jayanand Asok Kumar and Shobha Vasudevan
University of Illinois at Urbana-Champaign
Email: {athavale1, jasokkat2, shobvav}@illinois.edu

Abstract—Timing speculation is a 'better-than-worst-case' design methodology that times a digital circuit to its common-case delay. The average throughput of a speculation-based circuit can be estimated using the probability with which input patterns result in timing errors. In this paper, we present a scalable approach to compute the exact probabilities of the occurrence of timing errors at the gate level. We use Timed Characteristic Functions (TCFs) to compute the exact values of the probabilities. In order to improve the scalability, we decompose large circuits into smaller sub-circuits and restrict the TCF computation to these sub-circuits. Instead of substituting the expression for TCF of one sub-circuit into another, we propagate only the computed error probabilities. We demonstrate our technique on gate level combinational circuits from MCNC benchmarks.

I. INTRODUCTION

Traditional worst-case-timing design methodologies are insufficient to make digital circuits meet present day performance requirements. If the input patterns that trigger the worst-case delay are infrequently applied, it is overly pessimistic to design the circuit for error-free operation under the worst-case delay. With timing speculation techniques, the circuit can instead be tuned to a common-case delay and can be allowed to make timing errors for longer computations. These errors can then be corrected using circuit-level [1], [2] or microarchitecture-level [3], [4] techniques. Timing speculation achieves high single-cycle clock frequencies at the expense of extra clock cycles for input patterns causing delays longer than the clock period.

Critical probability of a signal is defined as the probability that the delay at that signal meets a specified time constraint. If the critical probability at the circuit output for a given clock period is low, the average throughput of the speculation-based design is significantly reduced due to the frequent clock cycle penalties. The average throughput of a speculation-based design can be estimated using the critical probability. If critical probabilities can be computed efficiently and accurately, circuit optimization techniques [5], [6] can use them to revise designs until the average throughput is maximized.

We propose a methodology to efficiently compute the exact critical probability at the circuit output for a specified time constraint. We use Timed Characteristic Functions (TCFs) [7] for this purpose. TCFs are Boolean expressions with the circuit inputs as variables. These expressions represent the set of input patterns that meet a specified time constraint at the circuit output. Since TCFs account for all such input patterns, the probabilities that we compute are exact.

The size of the TCF expression grows with the complexity of the circuit. Therefore, the size of the circuits that can be analyzed directly using this approach, is limited. Approximate formulations of TCF using Long Path Activation Functions (LPAF) [8] and Short Path Activation Functions (SPAF) [9] give smaller expressions. However, the probabilities that are computed using these approximate functions are not exact.

Circuits can be decomposed into independent subcircuits called supergates [10]. We instead decompose circuits into correlated subcircuits called relaxed supergates that are smaller than regular supergates. We then compute the TCF expressions for the outputs of the relaxed supergates. These TCF expressions are used to compute probabilities at outputs of the relaxed supergates which are propagated to compute the exact critical probability at circuit output. Since we do not propagate the TCF expressions themselves, our approach bounds the size of the TCF expressions computed at any stage.

In this paper, our main contribution is the proposal of a scalable and accurate method to compute critical probabilities at the circuit output for a specified time constraint.

Fig. 1. Block diagram showing our methodology

II. BACKGROUND

A. Probability concepts

The probability distribution (called PMF) of a Boolean variable $v$ is defined collectively by the probabilities with which $v$ is assigned a value 1 or 0. The joint probability of a set of variables is the probability with which a set of values are collectively assigned to the variables.

Two variables $a$ and $b$ are independent if the probability of an assignment to $a$ is not affected by the assignment to $b$. Variables that are not independent are called correlated variables. The joint PMF of independent variables is simply the product of their individual PMFs.

Let $V1$ and $V2$ be two sets of variables. Assume that the values assigned to the variables in $V2$ is fixed. The joint PMF of $V1$ under this assumption is called the conditional joint PMF. If the variables in $V1$ are independent of all the variables in $V2$, the conditional joint PMF of $V1$ is the same as the joint PMF of $V1$.

In Boolean logic, a literal [11] is a variable ($v$) or its complement ($\overline{v}$). A minterm for a set of $N$ variables is a product of $N$ literals such that there is exactly one literal for each variable. Two distinct minterms are always independent. A cube is a product of $M \ (M \leq N)$ literals and can be represented uniquely as a sum of minterms. Two cubes are not independent if and only if they cover one or more common minterms, which we call overlap terms.

We assume knowledge of the distribution of circuit inputs and that they are independently distributed.
B. Timed Characteristic Functions

$T CF(y,t)$ is a Boolean function that represents all input vectors whose evaluation arrives at node $y$ with a delay less than $t$. $T CF(y,t)$ evaluates to a logical ‘1’ for each of these inputs vectors and to a ‘0’, otherwise.

The $T CF$ expression uses the circuit inputs as its variables. For efficient computation of the $T CF$ at a node, a recursive formulation can be used [7]. In this formulation, the $T CF$ of a gate output is defined as a function of the $T CF$s of the gate inputs. For example, for a two-input OR gate with $a,b$ as inputs, $c$ as output and $d$ as delay of the gate,

$$T CF(a1,t)=T CF(a1, (t-d)+T CF(b1, (t-d))$$

$$T CF(a0, t)=T CF(a0, (t-d)+T CF(b0, (t-d)))$$

(1)

where $+$ and $*$ are the Boolean disjunction and conjunction operators respectively. Combining the cases where $c$ is evaluated to ‘1’ and ‘0’,

$$T CF(c, t-c)=T CF(c, (t-d)+T CF(c=1, (t-d)))$$

(2)

For each circuit node $c$, the fanin cone of $c$ is traversed recursively until circuit inputs are reached. All circuit inputs are assumed to be applied at time $t=0$. If $c$ is a circuit input, the expressions for $T CF(c=1,t)$ and $T CF(c=0,t)$ reduce to $c$ and $c'$ respectively for $t \geq 0$ and $0$, otherwise.

III. METHODOLOGY

In this section, we describe each step of our methodology in detail (Figure 1).

A. Decomposition into relaxed supergates

Regular supergates satisfy the following properties [10]:

- Each supergate can have multiple inputs but only one output
- Supergate output cannot have a fanout greater than 1
- If a node within a supergate has a reconvergent fanout, these fanout paths are also contained within the same supergate

These properties collectively ensure that the inputs to each supergate are independent. Figure 2(left) shows the decomposition of an example circuit into supergates. SG6 is the largest sub-circuit with nine gates and five input nodes.

Each supergate is an independent sub-circuit that can be analyzed separately. However, these sub-circuits themselves may be quite large. In order to obtain simpler sub-circuits, we refine the notion of supergates by relaxing some of the constraints imposed during their construction.

We now allow sub-circuit outputs to have a fanout greater than 1. This relaxed decomposition results in sub-circuits called relaxed supergates that are smaller than supergates. For the example circuit, we obtain 6 relaxed supergates (Figure 2(right)). RSG6 is the largest relaxed supergate with five gates and three input nodes.

A Stage-$i$ relaxed supergate is recursively defined as a relaxed supergate at least one input of which is an output of a Stage-$(i-1)$ relaxed supergate and rest of its inputs are circuit inputs or outputs of Stage-$k$ relaxed supergates where $k \leq (i-1)$. All inputs of a Stage-$1$ relaxed supergate are circuit inputs. The last stage relaxed supergate has the circuit output as its output.

Since relaxed supergate outputs are allowed to have multiple fanouts, relaxed supergates in the same stage may share one or more common inputs. We label each relaxed supergate input with the names of the circuit inputs that are in its fanin cone. In the example circuit, $f2$ is labeled with $i2$ and $i3$, $f3$ is labeled with $i2$, $i3$ and $i4$ and so on. Except for the last stage relaxed supergate, the number of labels on the output of a relaxed supergate is constrained to a small value (less than 10) during decomposition. We show later (Section III.C) that this helps in simplifying the computation of critical probabilities for the relaxed supergate.

Unlike supergates, we do not intend the relaxed supergates to be canonical. We only require a systematic way to produce sufficiently small sub-circuits for which computation of $T CF$ and critical probability is tractable. In practice, this requirement is fulfilled by relaxed supergates.

B. Formulating the $T CF$s for relaxed supergates

We now describe the formulation of $T CF$ for relaxed supergates. Using the recursive formulation for $T CF$, we express the $T CF$s of each output of a relaxed supergate as a function of the $T CF$s of its inputs. In the example circuit, $T CF(f2,t)$ is defined as a function of the $T CF$s of $f1,t1$ and $T CF[i3,t2]$. $T CF$s are represented in Conjunctive Normal Form (CNF).

We start with the last stage relaxed supergate. For a given time constraint at the circuit output, we derive the $T CF$ of the relaxed supergate output as a function of the $T CF$s of the outputs of relaxed supergates in the next lower stage. We now shift the $T CF$ formulation to the lower stage. We repeat this until we reach Stage-1 relaxed supergates.

We now have the $T CF$ expressions for the outputs of the Stage-1 relaxed supergates as a function of circuit inputs. In a regular recursive $T CF$ formulation, these $T CF$s are then plugged into the $T CF$ expressions of the outputs of Stage-2 relaxed supergates. Instead in our approach, we retain the $T CF$ expressions for each relaxed supergate as a function of the $T CF$s of its inputs. In the example circuit, $T CF(f2,t)$ is retained as a function of the variables $T CF[f1,t1]$ and $T CF[i3,t2]$.
C. Computing critical probabilities

1) Minimization of TCFs: We use logic minimization to compact the TCF expressions. These TCFs are now expressed as a sum of minimum number of cubes. In the minimized form, two cubes may share one or more minterms. In that case, the cubes are not independent. In order to compute the critical probability from a sum of cubes, the cubes need to be independent [12]. To achieve this, we start with the minimized set of cubes and generate independent cubes from them by ensuring that each minterm is contained in at most one cube.

Let $C$ be the set of minimized cubes. We pick a cube $c \in C$. From the remaining cubes in $C$, we identify the cubes that overlap with $c$. We repeatedly remove the overlapping terms from $c$ to form a set $C_1$ of smaller cubes. Any two cubes in $C_1$ are independent. Additionally, a cube in $C_1$ is also independent of every cube in $C - \{c\}$. At this stage, it may be possible to recombine two cubes in $C_1$ to form a larger cube. Clearly, this larger cube is also independent of the rest of the cubes in $C_1$ and all the cubes in $C - \{c\}$. We recursively do this recombination to get smaller number of independent cubes in $C_1$ until no more recombination is possible. $c$ can now be removed from $C$ and we repeat this process for each remaining cube of $C$. The algorithm terminates when $C$ is empty.

We illustrate this with an example over four variables. Figure 3 shows a Karnaugh map [11] for a Boolean function with variables $a$, $b$, $c$ and $d$. Suppose that the minimized TCF for an output is the sum of the three cubes: $a'c'$, $bd$ and $c'd$ (Figure 3 (ii)). $C = \{a'c', bd, c'd\}$. We start with the cube $a'c'$, $a'c'$ overlaps with both $bd$ and $c'd$. The intersection of $bd$ with $a'c'$ is $a'bd$. We remove $a'bd$ from $a'c'$ resulting in the cubes $a'bd'$ and $a'c'd'$. Now we look at the overlap of these cubes with $c'd$. $a'bc'd$ is the only cube that intersects with $c'd$ and the resulting cube is $a'bc'd'$ (Figure 3 (iii)). The cubes $a'bc'd'$ and $a'bd'$ are recombined to form $a'bd'$ (Figure 3 (iii)). Now, we move to the next cube in $C$ which is $bd$ and check the intersection with the only remaining cube $c'd$. The cube after removal of the intersection is $bd$. In this way, the final set of independent cubes is given by $a'c'd'$, $bd$ and $c'd$ (Figure 3 (iv)).

2) Propagating critical probabilities: We first consider the outputs of Stage-1 relaxed supergates. After minimization, the TCF of the relaxed supergate output is a sum of independent cubes that are defined over the set of circuit inputs. Since we assume that circuit inputs are independent, we compute joint probability of a cube as the product of the individual input probabilities. Moreover since cubes are independent, we compute critical probability at the relaxed supergate output by adding the joint probabilities of all cubes.

For Stage-2 relaxed supergates, we treat TCFs of Stage-1 relaxed supergate outputs as the “input variables”. We obtain the PMF of these variables from the critical probabilities of the corresponding TCFs. Let $TCF(f_1, t_1)$ be denoted by the variable $F_1(t_1)$ and $P(F_1(t_1))$ be the probability that $F_1(t_1)$ is evaluated to 1. $P(F_1(t_1))$ is equal to the critical probability corresponding to $TCF(f_1, t_1)$. We minimize the TCF of the Stage-2 output and convert it into a set of independent cubes. These cubes are functions of the Stage-2 input variables, i.e., the TCFs of the Stage-1 outputs or circuit inputs. However, in this case, the input variables are not necessarily independent.

During TCF formulation for a Stage-2 relaxed supergate output, multiple time constraints can be propagated back to the same output node of a Stage-1 relaxed supergate. Each of these constraints is represented by a Stage-2 input variable. For example, $F_1(t_1)$ and $F_1(t_2)$ are two input variables that correspond to time constraints $t_1$ and $t_2$ respectively. A cube for $F_2(t_1)$ may contain both of these correlated variables. In this case, we consider only the input variable corresponding to the lower time constraint.

Another form of correlation arises from shared circuit inputs present in the fanin cone of variables. We now analyze the labels corresponding to each variable in the cube. If two variables share one or more common circuit inputs in their labels, the variables are correlated. If there are no correlated variables, joint probability of the cube is computed as the product of individual probabilities of the variables. If there are correlated variables, we use the law of total probability [13] to compute the joint probability of the cube.

If a cube $c$ depends on $k$ circuit inputs $a_1, a_2, \ldots, a_k$, the joint probability of $c$, $P(c)$ is given by

$$P(c) = \sum_{v} P(c \mid v)P(v)$$

(3)

where $v$ is a bit-vector that denotes the values $v_i$ assigned to each $a_i$ and $P(c \mid v)$ is the conditional joint probability of $c$ given that $a_i = v_i$ for every $i$. Since circuit inputs are assumed to be independent, we have

$$P(v) = \prod_{i=1}^{k} P(a_i = v_i)$$

(4)

We then compute the critical probability at the relaxed supergate output by adding the joint probabilities of the cubes computed as shown in Equation 3. Since we consider all possible values of $v$, the computed joint probabilities are exact.

In the example, consider the cubes of the TCF at out. The nodes $f_3$ and $f_4$ share the labels $i2$ and $i3$. We fix the values of both $i2$ and $i3$ to be equal to 0. The joint probability of the cubes are

<table>
<thead>
<tr>
<th>Circuit</th>
<th>Critical Probabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>alu2</td>
<td>0.7090</td>
</tr>
<tr>
<td>eml50</td>
<td>0.5468</td>
</tr>
<tr>
<td>parity</td>
<td>0.0133</td>
</tr>
<tr>
<td>t481</td>
<td>0.3136</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time constraint</th>
<th>Critical Probabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.5156</td>
</tr>
<tr>
<td>8</td>
<td>0.9736</td>
</tr>
<tr>
<td>10</td>
<td>0.9950</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
</tr>
</tbody>
</table>
obtained under this condition. Each of these joint probabilities is scaled by $P[ar{z}_2 = 0] + P[z_3 = 0]$. We perform this for the 4 possible combinations of values that can be assigned to $z_2$ and $z_3$ together.

We repeat this computation for the higher stages of relaxed supergates until the critical probability of the circuit output is obtained.

Our algorithm to decompose input circuit into relaxed supergates is linear in the size of the circuit. Computation of critical probabilities at output of a relaxed supergate is exponential in the number of shared circuit inputs. However, in practice this number is kept small.

IV. EXPERIMENTAL RESULTS

We apply our technique to a set of MCNC benchmark circuits using the unit delay model. We use the tool SIS [14] for logic minimization of the computed TCF expression. All experiments are performed on a machine with 4GB RAM.

In Tables I and II, we demonstrate the accuracy of critical probabilities computed by our approach. We do not have access to the approximate TCF implementations to generate the probabilities for our time constraints and delay model. Moreover, comparison to results from approximate TCF implementations would not be reasonable since the critical probabilities computed by our approach are exact. Therefore, we compare our estimates against those obtained using high-confidence gate level random timing simulations.

Table I shows the comparison of critical probabilities for a set of MCNC benchmark circuits. A time constraint which is 75% of the worst case time for each of the circuits is used. Table II shows the comparison of critical probabilities for circuit cordic for different time constraints.

The small deviation between critical probabilities computed by our approach and those using gate level simulations can be attributed to the fact that with large number of inputs, simulations do not cover the entire input space. We observed that as we increase the number of simulation cycles, gate level estimates converge to the probability values computed by our approach.

In Table III, we compare decomposition into relaxed supegates and supergates. It can be seen that maximum size of a relaxed supergate is at most 60% (which is observed in alu2) of the total number of gates in a circuit. Moreover, for almost all the circuits, decomposition into relaxed supergates results in smaller sub-circuits than supergates. The only exception is the parity circuit which has a good decomposition into supergates due to the presence of very few reconvergent fanouts.

Table IV shows the scalability of our approach as compared to the case where circuit decomposition is not used. We compare the size of TCF in terms of CNF clauses, the number of cubes for the TCF expression obtained using SIS and the number of independent cubes formed for each case. It can be seen that our approach reduces the maximum size of the TCF that has to be computed by up to 90% so that larger circuits can be processed.

In this paper, we have proposed a scalable approach to find the exact critical probabilities of a circuit using TCFs. With the increased scalability afforded by our approach, we plan to extend it to sequential circuits by first converting them to equivalent combinational circuits using time unrolling.

REFERENCES


