Software-Based Self-Test of Processors under Power Constraints

Jun Zhou, Hans-Joachim Wunderlich

Institute of Computer Architecture and Computer Engineering, University of Stuttgart
Pfaffenwaldring 47, D-70569 Stuttgart, Germany

Abstract

Software-based self-test (SBST) of processors offers many benefits, such as dispense with expensive test equipments, test execution during maintenance and in the field or initialization tests for the whole system. In this paper, for the first time a structural SBST methodology is proposed which optimizes energy, average power consumption, test length and fault coverage at the same time.

Key words: Test program generation, processor test, low power test

1 Introduction

The principle of the SBST involves generation, storage and execution of a test program. Its main advantages are reusability at all the stages of the system life cycle, dispense with expensive test equipment, and low design for testability (DfT) overhead.

Originally, SBST was implemented as a purely functional test, which did not consider the structure of the core under test (CUT) but relied merely on the instruction set architecture (ISA) [1, 2]. It was mainly applied to systems with discrete components. However, a structural fault model was not explored, and test quality could not be assessed sufficiently. The advent of embedded processors and core-based designs exhibits more options. With the availability of the structure of a processor, structural automatic test pattern generation (ATPG) can be combined with SBST. Thus, either core vendors may deliver test programs validated by fault simulation, or users of soft cores may generate such programs.

SBST techniques based on emulated LFSR schemes were proposed and analyzed in [3]. In [4], a method for translating module level structural test patterns into processor level instructions was presented, and appropriate control and observation sequences were generated. In [5] a method for test program synthesis maximizing structural coverage was proposed. Recently, Li et al. introduced a deterministic functional self-test scheme based on emulated LFSRs [6, 7]. The previous work was improved leading to shorter test times and less memory requirement in [8]. Corno et al. presented a method for test program synthesis based on genetic algorithms [9, 10]. Most of the previous work targeted stuck-at faults, only recently were SBST schemes developed for detecting delay faults, e.g. [11, 12].

Structural testing maximizes switching activities within a circuit, which leads to an increased power consumption. Special care has to be taken in order to avoid reliability problems, reduced yield or even overheating. As a result, numerous DfT schemes are proposed to limit test power dissipation [13, 14, 15]. Without circuit modification, power can also be reduced through an appropriate test scheduling, proper test pattern optimization or system frequency decrease. The SBST operates under the functional mode so that it meets the peak power specification of the system. Nevertheless, it is of great significance for reliability reasons to optimize programs causing above-average switching activities. If such a test is deployed autonomously, not only average power but also energy consumption has to be optimized.

On the other hand, there are many research efforts on compiler techniques, which optimize code to restrict or minimize power [16, 17]. These techniques are inadequate for test programs, as they are implemented to retain program semantics. Program semantics is only subordinate from a test viewpoint where the main objective is to ensure or even maximize fault coverage, while power is reduced. For the first time, the presented paper combines structural SBST and software optimization for testing processors under power constraints. A novel method for test program synthesis is proposed which tackles fault coverage, test length, energy and average power consumption at the same time.

We arrange the rest of this paper as follows. The next section presents an overview of the methodology and describes the method of automatic test program generation. Section 3 details an algorithm on software power optimization which does not affect fault coverage. We apply the approach to a 32-bit RISC processor in Section 4. A short conclusion is provided in Section 5.
2 Test program synthesis

The overall methodology contains two procedures. First, based on the gate-level structure of a target processor, a test program is generated to provide maximal fault coverage with a short test length and low switching activities. Afterwards, the instruction sequence is optimized in order to minimize power consumption without sacrificing fault coverage. Though the stuck-at fault model is adopted in this paper, the method itself is applicable for any other combinational fault models, e.g. delay faults which require test pairs or sequences.

Fig. 1 elaborates the steps of test program synthesis. A standard ATPG generates structural test patterns for combinational modules, which are subject to a compaction process. The greedy algorithm takes fault coverage and correlation into account: at each time a pattern is selected which detects the largest number of faults yet undetected. If more alternatives are available, the one with the highest correlation measured by bit transitions is picked. The compaction ends once the fault coverage reaches that of the original test set. This step results in both test data reduction and potential energy and power savings without loss of fault coverage. In the end, the test set is mapped into instruction sequences using a template-based test program synthesis.

Fig. 1: Methodology for test program synthesis

The synthesis targets the modules which contribute to the major part of the processor area and structural fault coverage. For a standard processor, these are the arithmetic logic unit (ALU), the register file (RF) and the program counter (PC), as part of the control unit. We leave out test generation for the other components of the control unit which constitute deeply sequential logic and sufficient fault coverage for these parts requires special design for testability means, e.g. [18]. In the following paragraphs, we explain the essential code and the template structures similar to [7].

ALU:

Structural patterns for the ALU can be directly parsed into equivalent instructions. Suppose an ALU with two 32-bit inputs and 3 control bits leads to the format (A:<1:32>, B:<33:64>, S:<65:67>). Following this format, we extract an instruction whose operands are defined by the first 64 bits and the opcode is specified by the last 3 bits. Results are saved in general-purpose registers and fully observable. However, observing status signals needs particular instructions. For instance, instruction “JZ” (jump-if-zero) is able to access the status signal “Z”, which indicates whether or not the relevant input is zero, and hence is used to make “Z” observable. Fig. 2 represents the common structure for applying an ALU pattern using instruction sequences with parameters in brackets.

Fig. 2: Generating test instructions for the ALU

Program Counter (PC):

Program addresses, used as targets for branch instructions, are patterns for this unit. The generalized code in Fig. 3 applies a provided PC pattern for testing.

Fig. 3: Generating test instructions for the PC

Register file (RF):

A standard memory test, for instance the modified algorithmic test sequence (MATS) designed for stuck-at faults of memories [19], is able to achieve high fault coverage for the RF. A short MATS type test is implemented this way:

1) set \$x, [PC] ;load the target
2) jump \$x ;apply the pattern
3) .org [val_pc] ;[val_pc] parsed from the pc pattern
PC:
  save \$x, ([res]) ; make the target observable

ISA

Netlist
1: ATPG
2: Pattern compaction
3: Template-based test program synthesis
ISA
Test program
as a by-product of the ALU and the PC test if we allocate registers in a uniform way.

The above mentioned tests are altogether embedded into templates, each of which applies an ALU and a PC pattern. The coverage for the RF is maximized by regulating register allocation during template instantiation. PC responses are accumulated to reduce memory accesses [20]. In the head template in Fig. 4, whose instance is at the beginning of the final program, an initial value is loaded into a register specially for the PC test. Then an ALU pattern is applied using the code conforming to the structure in Fig. 2. The last two instructions apply the next pattern to the PC.

![Fig. 4: pseudo code of the head template](image)

An intermediate template bears the similar structure as that of the head, only replacing the first instruction with the one, in our case “Add”, for accumulation of PC responses. Its instances appear in the middle of the test program. Eventually, the accumulated response is stored in the instance of an tail template.

We generate the test set, containing both deterministic and random patterns, for the PC with the same size as that of the ALU. The former are designed in a way to cause transitions of the PC as many as possible, while the latter are only generated when the number of deterministic PC patterns is still less. The test program is generated by stepwise template instantiation where all the values for parametric fields are determined either by patterns or by register allocation.

### 3 Power optimization

The proposed method reaches high fault coverage based on knowledge about the gate level structure which can also be used for estimating switching activities of internal nodes of a circuit. If the layout is known as well, we can even take into account capacitances and specify power more accurately by weighted switching activities. In both cases we get a much better estimation of dynamic power dissipation than by merely considering transitions in registers at the program model level [16].

The test program synthesis described above considers already switching activities of the registers. Now, we tune two more factors to optimize a test program. One is the order of the instructions and the other one is defining unspecified bits in instruction words.

#### 3.1 Test-oriented reordering

Test-oriented reordering of instructions offers more degrees of freedom than traditional optimization where program semantics plays an important role. We explain this difference with the help of the code below:

```plaintext
1) set $0, 512
2) set $1, 65535
3) add $2, $1, $0
4) save $2, (result_0)
5) set $2, 324
6) set $4, 790
7) sub $2, $2, $4
8) save $2, (result_1)
```

The dependence graph (DG) of the shown code in Fig. 5 (a) displays the dependencies imposed by program semantics. Since most of the order is fixed, there are only few degrees of freedom for reordering, for example exchanging instruction 1 and 2, and putting 6 somewhere before 7.

![Fig. 5: program-oriented and test-oriented reordering](image)

However, viewing the same code as part of the ALU test, we divide instructions into units, where a pattern is loaded, applied, and relevant responses are stored. Such a unit forms a test behaviour. For the above code, two test behaviours are extracted in Fig. 5 (b). Obviously, fewer restrictions exist, as reordering is possible within or even across test behaviours if no conflict is introduced.

For computing the power consumption during optimization, we have two options: simulation-integrated or model-based. The former uses a gate level power analysis tool for controlling test program optimization. This approach is computationally expensive, and as an alternative, the second method builds a reference model for a target ISA in advance [21]. Once created, the model is reusable for code optimization, and the real power savings rely much on the quality of the model.

Here we illustrate steps towards constructing the power model. We use Hamming Distances (HD) as a measure characterizing transitions due to an instruction pair (i, j). The switching activities of the instruction register (IR), the RF and the ALU are easily predictable. For example, transitions TIR at the output of IR are related to instruction coding, and those for the RF, TRF, and the
ALU, $T_{ALU}$, depends on runtime data. For an n-bit instruction word, $T_{IR}$ is:

$$T_{IR}(i, j) = \text{HD}(C_i, C_j) = \sum_{m=0}^{i-1} (C_m \oplus C'_m)$$  \hspace{1cm} (1)

For a RF with two outputs, Q0 and Q1, let Q0 and Q1 be the values of output registers of instruction i, $Q_{0i,j}$, and $Q_{1i,j}$, be the temporary values due to (i, j). Then we model $T_{RF}$ as:

$$T_{RF}(i, j) = \text{HD}(Q_{0i,j}, Q_{0i,j}) + \text{HD}(Q_{0i,j}, Q_{0j})$$

$$+ \text{HD}(Q_{1i,j}, Q_{1i,j}) + \text{HD}(Q_{1i,j}, Q_{1j})$$ \hspace{1cm} (2)

Transitions of the ALU depend not only on the inputs (A, B, S) but also on the function it operates. For this reason, we consider transitions both at the inputs and the output (Q):

$$T_{ALU}(i, j) = \text{HD}(A_i, A_j) + \text{HD}(B_i, B_j)$$

$$+ \text{HD}(S_i, S_j) + \text{HD}(Q_i, Q_j)$$ \hspace{1cm} (3)

Hence, the problem to model average power consumption $A(i, j)$ due to inter instruction effects [17] for a given instruction pair (i, j) is finding relations between $A(i, j)$ and the above depicted factors. We can base the work on regression analysis, which is a standard statistical method to investigate relationships between dependent and independent variables. In our case, the dependent variable is $A(i, j)$, and independent variables are $T_{IR}$, $T_{RF}$ and $T_{ALU}$, that is:

$$A(i, j) = f(T_{IR}, T_{RF}, T_{ALU}) + \varepsilon$$  \hspace{1cm} (4)

$\varepsilon$ represents approximation discrepancy. If a linear relationship is assumed, we further specify the above formula as:

$$A(i, j) = \beta_0 + \beta_1 T_{IR}(i, j) + \beta_2 T_{RF}(i, j) + \beta_3 T_{ALU}(i, j) + \varepsilon$$  \hspace{1cm} (5)

$\beta_0, \ldots, \beta_3$ are coefficients to be determined by the process of multiple regression analysis based on samples. A sample is an observation of the dependent and independent variables. Equation (1) to (3) facilitate calculation for the values of independent variables. According to [17, 21], we explore Equation (6) to obtain values for the dependent variable, where $A_{ij}$ stands for the measured average power consumption of the loop (i, j), while $A_{basr}$ and $A_{base}$ are the basic average power when instruction i or j executes stand-alone:

$$A(i, j) = A_{ij} - \frac{1}{2}(A_{basr} + A_{base})$$ \hspace{1cm} (6)

The reference model is built by the following steps below:

- simulate programs consisting loops of individual instruction to generate basic power costs $A_{basr}$;
- simulate programs consisting loops of instruction pairs to get values for power costs $A_{ij}$;
- use Equation (1), (2), (3) and (6) to create samples by considering each possible instruction pair;
- perform multiple regression analysis to determine values for coefficients $\beta_0, \ldots, \beta_3$.

Since we work at the gate level, the model accordingly describes gate-level power consumption, offering closer estimation than the bit-transition model at the program level. A comprehensive statistical analysis for validating the model and parameter was done in [21].

### 3.2 Unused bit setting

Many existing publications use don’t-care bits to reduce transitions in test patterns [22]. A similar idea is explored in our work at the instruction level. Fig. 6 exemplifies a specification for an Add operation in SPARC v8 [23].

<table>
<thead>
<tr>
<th>Assembly Language Syntax</th>
</tr>
</thead>
<tbody>
<tr>
<td>add</td>
</tr>
</tbody>
</table>

**Format 1:**

<table>
<thead>
<tr>
<th>10 rd</th>
<th>000000</th>
<th>rs1</th>
<th>i=0</th>
<th>unused(zero)</th>
<th>rs2</th>
</tr>
</thead>
<tbody>
<tr>
<td>31 29 24 18 13 12</td>
<td>4</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Format 2:**

<table>
<thead>
<tr>
<th>10 rd</th>
<th>000000</th>
<th>rs1</th>
<th>i=1</th>
<th>imm</th>
</tr>
</thead>
<tbody>
<tr>
<td>31 29 24 18 13 12</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Fig. 6: Specification of Add in SPARC v8*

According to its format specification, ADD computes “r[rs1]+r[rs2]” if the 13th instruction bit is set to 0, else an immediate is used, and “r[rd]” holds the result. Every bit in Format 2 has its particular meaning, specifying either the operation code or relevant operands. On the contrary, in Format 1, 8 bits starting from Bit 5 to Bit 12 are defined as “unused” and assigned by default zeros. However, in some cases it is more power efficient, if other values are assigned to those bits. Our second optimization factor takes advantage of this observation.

Given the ISA, we are able to identify don’t-cares (DCs) of each instruction under consideration. For an instruction pair (i, j), where i is fully specified, we annotate DCs of j in a way that switching activities due to (i, j) are minimized. The reference model built previously can be used once again during DC specification.

### 3.3 The overall optimization methodology

Before optimization starts, the raw code is partitioned into blocks, each of which is an instance of templates. Then the overall algorithm works in two phases on a block basis:

- **Test-oriented greedy reordering.** First, test constraints are extracted with the guidance of template structures. Then we identify unscheduled in-
instructions that do not violate extracted test behaviors and for each of them, based on the power model, we compute their costs with respect to the last scheduled code. Finally the instruction with the minimal cost gets scheduled.

- DC specification for the instruction to be scheduled.

The overall process ends if all instructions are scheduled.

4 Experimental results

We apply the entire methodology to a 32-bit RISC processor [24]. It is mainly made up of the following functional components: the ALU, the RF with 32 general-purpose registers, the PC, the IR, the control unit. The design contains 14,244 logic gates including 1,088 sequential cells. Fault simulation reports 48,784 uncapped stuck-at faults.

The evaluation framework is an extension of [6] to support measurements of gate-level fault coverage and energy consumption in terms of switching activities. Test bench integrates the gate-level processor core and the test program binary. During simulation, two kinds of information are recorded. Primary inputs of the processor are captured and translated into patterns which are later fault simulated to evaluate the structural fault coverage. Switching activities during simulation are tracked in a file named Value Change Dump (VCD). Based on such file, we can used commercial tools, e.g. Primepower [25], for energy estimation, if layout information of the technology is available. As an alternative, it is straightforward to sum up recorded transitions and treat the final number as energy consumption for the simulated code directly.

Moreover, we set up our experiments as follows:

1) Perform ATPG for the ALU alone to provide the deterministic patterns. The test set achieves 99.90% fault coverage of this module;

2) Implement a MATS test for the RF, which is used together with programs generated in step 3 to 6;

3) Synthesize the test set generated in step 1 into a test program (Ori.TP), which reflects the rudimentary qualities of our algorithm;

4) Compact ALU patterns in step 1, and then synthesize the new set into a program (Compacted);

5) Reorder the code generated in step 4 under test constraints. The new code is later referred as Reordered;

6) Specify DCs of the code in step 5. The new code is termed as X-filling.

7) In the absence of program synthesis algorithms e.g. [8], we conduct a sequential ATPG for comparison. Since the fault coverage is too low to accept, a program with randomized instructions is added. The number of test cycles for them together is equal to that of Ori.TP generated in step 3. Throughout the steps above, ATPG and program synthesis last only a few seconds, while ATPG alone leads to 18.14% processor fault coverage with 1,457 test cycles, while the improvement by the randomized program, which contributes to a much larger portion of cycles, is not remarkable.

### Tab. 1: Comparison in fault coverage and test cycles

<table>
<thead>
<tr>
<th>Fault coverage</th>
<th>Ori. TP</th>
<th>Optimized</th>
<th>ATPG + randomized</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALU</td>
<td>99.90%</td>
<td>99.90%</td>
<td>2.87%</td>
</tr>
<tr>
<td>PC</td>
<td>92.38%</td>
<td>92.38%</td>
<td>66.40%</td>
</tr>
<tr>
<td>RF</td>
<td>98.46%</td>
<td>98.03%</td>
<td>17.96%</td>
</tr>
<tr>
<td>Proc.</td>
<td>96.70%</td>
<td>96.33%</td>
<td>19.64%</td>
</tr>
<tr>
<td>Test cycles</td>
<td>No.</td>
<td>13,886</td>
<td>9,628</td>
</tr>
<tr>
<td>Ratio</td>
<td>1.0</td>
<td>0.6935</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Tab. 2 details the outcome regarding energy and average power consumption. The last four rows clearly show a persistent reduction of these two parameters through every optimization step. In particular, row three indicates that compaction not only results in shortening test length, which means energy reduction as well, but also in minimizing average power dissipation by concerning switching activities of registers throughout the process. Moreover, as shown in the first row, our methodology also outperforms the combination of the sequential ATPG and the randomized program in this aspect. According to the experiment, the sequential ATPG patterns causes 1,215,280 transitions in total, which consequently bring the average power consumption (transitions per test cycle) to 834.10. Meanwhile, the randomized program consumes average power of 525.40, which is higher than any of our test programs but much lower than those structural patterns. The reason for this observation is, the structural test activates signal transitions as many as possible, while high correlation among instructions ensures much lower power consumption even for the case where random instructions with random data are used. Hence we come to the conclusion that the SBST is low power in nature.
5 Conclusion

In this paper, we present a method to synthesize test programs, where fault coverage, test length, energy and average power dissipation are addressed at the same time. Our experimental results indicate that this methodology achieves high structural fault coverage by targeting at a subset of modules of a processor. The program optimization algorithm reduces energy and average power without sacrificing fault coverage. Our future work focuses on adapting the algorithm to the prevailing superscalar architecture, where features like out-of-order execution pose challenges to program synthesis as well as software optimization for low power.

6 References


