Test Scheduling for 3D Stacked ICs under Power Constraints

Sengupta, Breea; Ingelsson, Urban; Larsson, Erik

2011

Link to publication

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
Abstract- This paper addresses test application time (TAT) reduction for core-based 3D Stacked ICs (SICs). Applying traditional test scheduling methods used for non-stacked chip testing where the same test schedule is applied both at wafer test and at final test to SICs, leads to unnecessarily high TAT. This is because the final test of 3D-SICs includes the testing of all the stacked chips. A key challenge in 3D-SIC testing is to reduce TAT by co-optimizing the wafer test and the final test while meeting power constraints. We consider a system of chips with cores equipped with dedicated BIST-engines and propose a test scheduling approach that reduces TAT while meeting the power constraints. Depending on the test schedule, the control lines that are required for BIST can be shared among several BIST engines. This is taken into account in the test scheduling approach and experiments show significant savings in TAT.

I. INTRODUCTION

Integrated circuits (ICs) with multiple chips (dies), so called 3D Stacked ICs (SICs), have recently attracted a fair amount of research [1-5]. A 3D-SIC is obtained by stacking and bonding individual chips, which are connected by Through-Silicon Vias (TSVs). Due to imperfections in IC manufacturing, each individual chip must be tested. Recent research has addressed test architecture design for 3D-SICs [6], testing the TSVs [1-6] and 3D-SIC-specific defects [1, 2], but no previous work has addressed test scheduling under power constraints for 3D-SICs, which is the topic of this paper.

Testing each individual chip is required for both 3D-SICs and traditional non-stacked ICs. While IC packaging is costly [7], each chip is tested twice: (1) in wafer sort test, where the bare die is tested (pre-bond test), and (2) in final test where the packaged IC is tested (post-bond test). For non-stacked chips the same test schedule is applied in both pre-bond and post-bond test. However, for a 3D-SIC the process is different. As will be discussed in this paper, applying the same test schedule for both pre-bond and post-bond tests in a 3D-SIC leads to sub-optimal Test Application Time (TAT). TAT is defined as the sum of the testing times for pre-bond tests and post-bond tests. TAT is a major part of the overall test cost. Hence it is important to schedule the tests for 3D-SIC such that TAT is minimized, which is addressed in this paper.

Much work has addressed test scheduling for non-stacked chips with the objective of minimizing TAT [8, 9]. For core-based systems where each core is to be tested, the main method of reducing TAT is to perform core tests concurrently. However, performing tests concurrently leads to higher power consumption than performing them sequentially. The test power consumption must be kept under control [9], to avoid false test positives due to voltage drop and damage due to overheating. For core-based systems, Chou et al. [9] proposed a method to schedule tests in sessions while taking resource conflicts and power consumption into account. A session is a group of tests that start at the same time. In the context of systems where each core has an dedicated Built-In Self Test (BIST) engine, all the core tests that are scheduled in the same session can be initiated using a single control line. As a rule, a low number of sessions is beneficial, since it leads to a low number of control lines and implies that several tests are performed concurrently, leading to a low TAT [8, 9, 10]. The studies in [8, 9, 10] address test scheduling for non-stacked chips under power constraints. However, no work has yet addressed test scheduling for 3D-SICs under power constraints, which is the topic of this paper. We propose a power constrained test scheduling approach, which considers a two-chip 3D-SIC design, consisting of cores, each equipped with a dedicated BIST-engine. In this context we present an analysis of the test scheduling problem in Section II leading to an approach in Section III. The experimental results are in Section IV and the conclusions are in Section V.

II. PROBLEM ANALYSIS

Figure 1 shows a chip with three cores where each core is tested by its BIST test. Associated with each test are the parameters test time and power consumption. The test controller, which is a Finite State Machine (FSM) determines when the test for each core is initiated. Figure 2 shows a test schedule for the tests of the three cores in Figure 1, which have been scheduled as per [8] where the TAT is minimized and the power consumption at any moment is less than the maximal allowed power consumption $P_{\text{max}}$, which is indicated by a horizontal line. The test schedules are represented with blocks for the core tests, where the height of a block is the power consumption for the test and the width of the block is the test time. The x-axis shows the time taken to perform the tests, and the y-axis marks the power consumption. Two types of constraints are considered for the test schedule. The first constraint type is resource constraints which determine that two tests are not to be performed concurrently and the second constraint type is a constraint regarding the maximum power consumption, $P_{\text{max}}$, which cannot be exceeded. The test schedule contains three sessions: Session1, Session2 and Session3, as marked in the figure. This chip is a single-die IC, so the same test schedule is applied at pre-bond test (wafer sort) and post-bond test (final test). $TAT = C_1 + C_1$, as the same test schedule is run twice.

Figure 3 shows a 3D-SIC where Chip1 (from Figure 1) is stacked on top of Chip2. The testing of the 3D-SIC requires pre-bond tests of Chip1 and Chip2 and a post-bond test of the stacked chip that tests the whole SIC by including tests for the cores in Chip1 and Chip2. While testing of TSVs is important,
the actual test time is fixed and relatively low. In this paper, the testing of TSVs is not addressed. The test durations and power consumption values for each core tests are provided in Table 1. The power constraint value is $P_{\text{max}} = 20$ units.

Prior to bonding chips into 3D-SIC each chip can be considered as individual non-stacked chips and the methods in [8, 9] apply for generating the pre-bond test schedules. Figure 4 shows an example of the pre-bond test schedules for the two chips, Chip1 and Chip2, from Table 1. The test schedule for Chip1 contains three sessions (Session1, Session2 and Session3) and the test schedule for Chip2 contains two sessions (Session4 and Session5) as shown in the figure. The test time for the schedules as obtained by [8] are $C_1$ and $C_2$ for Chip1 and Chip2, respectively.

Table 1: Test time and power consumption for core tests in Chip1 and Chip2.

<table>
<thead>
<tr>
<th>Chips</th>
<th>Tests</th>
<th>Duration</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chip1</td>
<td>T1</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>Chip2</td>
<td>T4</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>T2</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>T3</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>T5</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>T6</td>
<td>5</td>
<td>9</td>
</tr>
</tbody>
</table>

Once the chips have been stacked, each core of the chips again requires testing. We define three different approaches for test scheduling depending on the available knowledge from the pre-bond test. In this paper, the three approaches are called Serial Processing (SP), Partial Overlapping (PO) and ReScheduling (RS).

In case the only knowledge of the pre-bond test schedules consist of the test time for the schedules and the fact that the pre-bond test schedules are within the power constraint, the limited knowledge available restricts the test schedules that are possible. In this case the post-bond tests are scheduled by Serial Processing, which is illustrated in Figure 4. With Serial Processing we mean that the test schedules of individual chips are run serially during post-bond testing. It should be noted that, no tests from different chips are performed concurrently, because otherwise we would risk exceeding the power limit. For Serial Processing, the time taken to run the post-bond test schedule is equal to the sum of the time taken to test the individual chips, which has been denoted by $TAT_{SP}$ in the figure. For the schedule in Figure 4, $TAT_{SP} = C_1 + C_1 + C_2 + C_2$. 

Figure 3: 3D-SIC of Chip1 and Chip2.

Figure 2: Test schedule of Chip1.

Figure 4: Serial Processing.

Figure 5: Partial Overlapping
shown in Figure 6, Session4 is test T5, Session5 is test T6 and results in splitting Session4 and renumbering the sessions, as test T6 concurrently because they are power compatible. The pre-bond reduction in the TAT equal to the length of test C2 from Session4, there is an increase in the pre-bond test time for (Session3) and test T6 of Chip2 (Session4) are performed concurrently because they are power compatible. The pre-bond schedule of the chips remain unchanged, but there is a reduction in the TAT equal to the length of test T6 (Session4) and the resulting TATPO is lesser than TATSP.

When full knowledge is available concerning individual tests and sessions of the pre-bond test schedules, ReScheduling of the existing schedules can be performed. In the ReScheduling approach, knowledge of the pre-bond test schedules is utilized to create a post-bond test schedule to reduce test time. ReScheduling may cause changes to the pre-bond schedules. In this context, changing the pre-bond schedule means to split a session and replace it with two new sessions. The benefit of splitting a session is that the two new sessions can be scheduled concurrently with sessions of the other chip during post-bond test, if that reduces TAT. Figure 6 depicts the result of the ReScheduling approach. In the original pre-bond test schedule (Figure 4), Session4 consisted of tests T5 and T6. In the post-bond test schedule, after rescheduling, test T5 is performed serially with test T1, while test T3 is performed together with test T2. This results in a reduction of the post-bond test time equal to the duration of test T3. ReScheduling results in splitting Session4 and renumbering the sessions, as shown in Figure 6. Session4 is test T5, Session5 is test T6 and Session6 is test T4. But because of the splitting of the original Session4, there is an increase in the pre-bond test time for Chip2 from C2 to C2RS. The increase is equal to the duration of test T6 which is now performed serially with test T1. Compared to SP, the reduction in TAT is equal to the sum of the durations of tests T5 and T6, minus the duration of test T5. From the above, it can be seen that ReScheduling leads to lower TAT as compared to Serial Processing and Partial Overlapping, as is shown in Figure 6. However, in contrast to Serial Processing and Partial Overlapping, ReScheduling can lead to an increase in the number of control lines, as a result of splitting sessions. Our approach, detailed below, takes control lines into account.

III. PROPOSED APPROACHES

In this section we first detail the two approaches, Partial Overlapping (PO), and ReScheduling (RS), and then discuss the complexity of the approaches.

PO can be considered as a special case of RS. PO considers only the knowledge of individual sessions, and no sessions are split in the process. Hence, PO can be obtained by disregarding Step1 of RS.

RS is an approach in two steps as is described in the following. Before the first step, the initial pre-bond test for each chip is generated by the heuristic from [8]. Each session of the pre-bond test schedules is given a unique number.

Step1: Figure 7 shows an 11 stage process for implementing Step1 of RS. The key idea is to group the tests of two pre-bond test sessions from different chips in two post-bond test sessions such that the long tests are grouped together and the short tests are grouped together. This way, there will be one long test session and one short test session, instead of the previous two long sessions. The sum of the newly formed session duration is less than for the two original test sessions. In stage 1 of Figure 7, we consider two sessions, Sx and Sy, from the original pre-bond test schedules of two different chips, ChipX and ChipY, respectively. In stage 2, the tests of Sx and Sy are arranged in descending order of length in a list called M. A post-bond session, Sa, is produced in stage 3. Starting from the first test in M, i.e. the test with the longest test time, move the test from M to the post-bond session Sa, as shown in stage 4, and stage 5 checks if the total power of Sa is within the power constraint. Stage 4 and stage 5 are iterated until the power constraint is met. As soon as Pmax is exceeded as a result of moving a test from M to Sa, that test is moved back to M (stage 6). A new post-bond session, Sb, is created with the remaining tests of list M, which is shown in stage 7. Post-bond sessions Sa and Sb contain all the tests from M. The pre-bond sessions Sx and Sy are split into test sessions (say Sxa, Sxb, Sya and Syb) according to how the tests were allocated in Sa and Sb. The modified TAT is calculated in stage 8. Stage 9 considers the TAT. If the new test schedule (pre-bond and post-bond) is shorter than the test schedule for SP, the value is included in Table 2 as in stage 10, as the entry for session Sx and session Sy. Otherwise, if there is no reduction the value is set to be zero.

The process described above is repeated for all possible combinations of two sessions from the pre-bond test schedules.
of the two chips, as is shown in stage 11 of Figure 7.

A key observation from the above is that pairs of sessions can be handled independently. If combining a pair of sessions as described by stage 1 to stage 11 leads to a reduction in TAT compared to the test schedule in SP, a new test schedule can be constructed by combining several such session pairs. The total reduction in TAT can be summed up from the reductions in test time when all session pairs have been considered, while each session has been taken into account only once.

**Step 2:** This step involves the calculation of the maximum reduction in TAT from SP to RS (or PO) by considering all possible session pairs. Table 2 shows the possible reduction in TAT as a result of rescheduling a session of ChipX, as denoted by the column number, with a session of ChipY of the corresponding row number. Given Table 2, a schedule is generated by rescheduling each session of one chip with different sessions of the other chip, such that every session is considered only once. The sessions that are not rescheduled are added to the final schedule without any modification. The objective is to find the combination of rescheduled session pairs, which give the minimum TAT. The values in Table 2 are obtained by rescheduling sessions of the example used during the problem formulation. For example, with respect to Figure 4, considering Session2 from Chip1 and Session4 from Chip2 results in a reduction of 3 time units on rescheduling (as discussed in Step 1), compared to the time required to perform the original Session2 of Chip1 and Session4 of Chip2 sequentially, as in SP. In case of PO, where no sessions are split, the values in the table would either be zero (when the sessions are not power compatible), or equal to the length of the smaller session. For example, it was not possible to reduce TAT by combining Session1 with Session4 as marked by 0 in Table 2. The test schedule and the total reduction in TAT are obtained by rescheduling each session of ChipY, (Chip2 in the example) with sessions of ChipX (Chip1 in the example). As discussed before, tests from Session2 of Chip1 and tests from Session4 of Chip2 upon rescheduling, result in a reduction of 3 time units, while, Session5 of Chip2 with Session3 of Chip1 give a reduction of 5 time units. The sessions that result from the marked session pairs are included in the post-bond test schedule with the summed total of test time reduction adding up to 3 + 5 = 8 time units. The test time of the rescheduled session pairs are added to the remaining sessions to give TAT. Thus, the final post bond schedule has Session1 in series with the combination of Session2 with Session4 and Session3 with Session5. Thus TAT is 54 time units, obtained by reducing 8 time units from TAT SP, which has 31 time units in both pre-bond and post-bond.

**Table 2:** Maximum possible time reduction of sessions where ChipX and ChipY refers to the algorithm and Chip1 and Chip2 refers to the example.

<table>
<thead>
<tr>
<th>ChipX(Chip1)</th>
<th>ChipY(Chip2)</th>
<th>Session1</th>
<th>Session2</th>
<th>Session3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Session4</td>
<td>0</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Session5</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
</tbody>
</table>

Finding the combination of session pairs that give the minimum TAT on rescheduling requires comparison of all possible
combinations of session pairs, which is complex. To arrive at
the complexity of exploring all possible schedules from Table
2, say ChipX and ChipY have \(x\) and \(y\) number of sessions
respectively, and that \(x \geq y\). Then there are \(x\) columns and \(y\)
rows. The first row has \(x\) values to choose from. Once a value
is chosen, the row and column to which the value belongs are
ignored and there remains \(x-1\) values to choose from in the
second row. Thus, as we process each row, the number of
choices decreases by one. This accounts for a factorial function
that describes the number of possible sets of session pairs. But,
when \(y-1\) rows have been traversed the last value can be
chosen from the remaining \(x-y+1\) columns. Thus, the total
number of ways, \(N\), in which values can be selected from Table
2, with each value from a unique row or column, is given by
\(N = (x - y + 1) y!\). Hence, for a total number of ten
sessions each in two chips, \(N\) becomes as large as 3628800.
Thus, it can be seen that the problem of selecting session pairs
from Table 2 to explore all possible test schedules is difficult.
Existing heuristics can be applied to obtain a schedule from
Table 2. In the following we describe the greedy heuristic that
has been used. Prior to applying the heuristic, the rows are first
sorted in descending order of the highest value in each row and
then the same procedure is applied for the columns. After the
table has been sorted, an arbitrary starting value is chosen from
the table. The highest value from the neighboring row is then
considered along with it. The process is continued until all rows
are exhausted. The sum of all the values corresponding to the
session pairs considered for rescheduling give the net reduction
in test time. Sessions that were not joined with other sessions
are added to the list of session pairs to form the schedule. The
particular combination of session pairs that lead to the schedule
correspond directly to the pre-bond and post-bond test
schedules for the stacked 3D design. The combination of
session pairs that gives the largest reduction in terms of TAT
and an acceptable number of BIST control lines, as determined
by the designer of the stacked 3D chip, can be considered as the
final schedule.

To arrive to the final schedule, the heuristic is iterated \(K\) times,
where \(K\) is the sum of the number of rows and columns, with
different session pairs as starting point to produce a number
of solutions that can be evaluated by the designer of the stacked
3D chip with regard to the acceptable number of control lines.
This results in Table 3 for the considered example. Schedule1
is the result of combining Session2 with Session4 as well as
Session3 with Session5. Schedule 2 is the result of combining
Session2 with Session5.

### Table 3: TAT reduction versus increase in number of BIST control lines

<table>
<thead>
<tr>
<th>Schedule</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAT Reduction</td>
<td>8</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>BIST Control Line Increase</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

ReScheduling of solutions resulting in a reduction of TAT can
lead to a corresponding increase in the number of BIST control
lines due to splitting of sessions. Table 3 shows an example
providing the reduction in TAT and the number of additional
control lines for five of the test schedules produced by the
proposed RS approach.

**Complexity of the approach:** Here we study the complexity of
the RS approach. The approach consists of two steps, Step1 and
Step2.

In Step1 of the RS approach, the tests from two sessions are
initially sorted by their test durations and stored in the list \(M\).
The average time complexity for quick-sort is \(O(N \log N)\) for
\(N\) tests.

Step2 of the problem involves obtaining the maximum sum of
individual elements from the matrix, taking one element from
each row or column. As discussed, the solution space is large,
so the greedy heuristic has been applied, which has a average
time complexity of \(O(T \log T)\), where \(T\) is the number of
elements in the matrix. Prior to applying the heuristic, the rows
and columns were sorted in descending order of the value of
individual elements. The sorting here is also done by quick-
sort, which has a complexity of \(O(\log T)\). Step 2 is iterated \(K\)
times, where \(K\) is the sum of the number of rows and columns
of the corresponding matrix.

Thus the overall complexity of the approach is \(O(T \log T)\).

### IV. EXPERIMENTAL RESULTS

To demonstrate the benefits of the proposed test scheduling
approach, this section describes an experiment to compare TAT
achieved by Partial Overlapping (PO) and ReScheduling (RS)
with TAT achieved by the straightforward Serial Processing
(SP) approach, which is used as baseline. In the experiment,
the power constraint is met and the number of BIST control lines
required by different test schedules is taken into account. As
the RS approach yields a table such as Table 3 with several
different test schedule solutions where the acceptable number
of control lines determines the final test schedule selection, the
experiment is performed with the test schedule that results in
the largest TAT reduction (8 time units in the case of Table 3).
The initial pre-bond test schedules were generated by the
approach in [8] and our approaches where applied for
generating the post-bond test schedule. The approach proposed
in Section III was used to find the maximum reductions in TAT
while considering the number of BIST control lines as the
number of sessions in the example designs were in a reasonable
range.

The experiments are performed with the circuits ASIC Z [10],
System L [10] and Muresan [8] (marked by Z, L and M
respectively in Table 4) and these circuits were used to create
3D-SICs. These designs are seen as single-die chips and have 9
[10], 14 [11] and 10 [8] cores, respectively. To make a 3D-SIC,
two of the three single-die chips are combined. Groups of
columns marked Chip1 and Chip2 respectively, show the chips
that are combined into 3D-SICs, and groups of columns marked
Chip1 & Chip2 and TAT contain experimental results
regarding the combined 3D-SICs. To combine the Muresan
design with ASIC Z or System L, such as in the fifth row of
results in Table 4, adjustments were made on parameters
because the values of those parameters in the original designs
were given in different orders of magnitude. In the cases
marked M* and M**, the parameter values (test time, test
power consumption and power constraint) were scaled such
that the pair of single-die chips that are combined into 3D-SICs

The initial pre-bond test schedules were generated by the
approach in [8] and our approaches where applied for
generating the post-bond test schedule. The approach proposed
have their parameter values in the same order of magnitude. It should be noted that this scaling of parameter values is only performed to enable the experiments. The results are collected in Table 4. The first group of four columns marked Pre-bond test for Chip1 show the test times for SP, PO and RS for the pre-bond schedules for Chip1. The fourth column in the group shows the relative reduction in pre-bond test time of RS compared to SP. It should be noted that a negative reduction is an increase. Similarly, the second group of four columns shows the Pre-bond test for Chip2. The third group of four columns marked Chip1 & Chip2, Post-bond test, show test time for the post-bond test schedule generated by the three approaches, and gives the relative amount of post-bond test time reduction achieved comparing the result for SP with the result for RS. The group of columns marked TAT includes the sum of the pre-bond test times and post-bond test times. The first three columns in the group of four show TAT for the SP, PO and RS approaches, respectively. The relative reduction in TAT is shown in the last of the four columns where RS is compared against SP. The right-most column of Table 4, shows the relative increase in the number of control lines that result from splitting sessions in the RS approach. The number of control lines for the SP approach is shown in parenthesis. From Table 4, it can be seen that RS can achieve up to 42.1% reduction in the post-bond test schedule time in comparison to SP, when two chips of System L are stacked to form the 3D-SIC. This is for the 3D-SIC consisting of two System L chips. This result can be explained by a high power constraint, which enables a beneficial post-bond test schedule where many core tests are performed concurrently. In particular for the design with two System L chips, one session was split, resulting in an additional control line and an increase in the pre-bond test schedule duration. The reduction in TAT was 17.1%. It should be noted that other 3D-SICs consisting of two identical chips (such as the pair of ASIC Z chips) does not lead to the same result. For the 3D-SIC design made up by a pair of ASIC Z chips, TAT was reduced by 3.3% and RS and PO achieved the same result. This corresponds to a case when it is not possible to reduce TAT by splitting sessions. In the six experiments, only two experiments led to splitting of sessions, which increased the number of BIST control lines. For the other four experiments, the reduction in TAT was achieved without splitting sessions and the best result achieved without splitting sessions was 12.2% reduction in TAT.

V. CONCLUSION

In this paper, the problem of power-constrained test scheduling for 3D Stacked Integrated Circuits (SICs) has been addressed for the first time. It is shown that the test planning for 3D-SICs is different, compared to the test planning for non-stacked ICs, and requires specific test scheduling solutions. The paper proposes two test scheduling approaches, Partial Overlapping and ReScheduling that minimize test application time while taking power-constraints and the number of BIST control lines required to implement a test schedule into account. The two scheduling approaches and a straightforward approach (Serial Processing) have been implemented and experiments with several benchmarks show up to 17.1% reduction in test application time and an average reduction of 7.7% in test application time with a 3.8% average increase in the number of BIST control lines over the Serial Processing scheme.

REFERENCES