Critical Path - Oriented & Thermal Aware X-Filling for High Un-modeled Defect Coverage

Fotios Vartziotis
Computer Science and Engineering, University of Ioannina, Greece
Computer Engineering, T.E.I. of Epirus, Greece
fvartzi@teiep.gr

Xrysovalantis Kavousianos
Computer Science and Engineering
University of Ioannina, Greece
kabousia@cs.uoi.gr

Abstract—The thermal activity during testing can be considerably reduced by applying power-oriented filling of the unspecified bits of test vectors. However, traditional power-oriented X-fill methods do not correlate the thermal activity with delay failures, and they consume all the unspecified bits to reduce the power dissipation at every region of the core. Therefore, they adversely affect the un-modeled defect coverage of the generated test vectors. The proposed method identifies the unspecified bits that are more critical for delay failures, and it fills them in such a way as to create a thermal-safe neighborhood around the most critical regions of the core. For the rest of the unspecified bits a probabilistic model based on output deviations is adopted to increase the un-modeled defect coverage of the test vectors.

I. INTRODUCTION

Excessive power consumption during test, increases the overall chip temperature, and in many cases, it creates localized overheating. Thermal aware testing resolves thermal−related issues by reducing the temperature at hotspots in a SoC [1], [5], [8], [10], [18], [20], [25], [26]. However, critical paths are not always affected by the hotspots. In some cases, the reduction of the temperature at a hotspot may even increase the temperature at other areas of the SoC. In such cases the temperature of critical paths increases, and the probability good dies to fail test increases too. Therefore, methods that are both thermal-aware and delay-aware are needed to avoid unnecessary yield loss.

Since thermal generation is proportional to the average power dissipation, many techniques reduce the switching activity of the core during scan-in/scan-out and/or capture operation [3], [6], [7], [9], [15], [21], [24]. X−fill methods are very effective in reducing shift and/or capture power by manipulating only the unspecified bits of test cubes [4], [12], [13], [16], [17], [19], [22], [23]. X-fill methods have negligible impact on the ATPG process, they affect neither the scan chain structure nor the circuit under test (CUT), and they can be easily combined with other test methods.

Since the thermal behavior of a core depends on both the heat generation and the heat dissipation mechanism, layout information is needed to tackle the combined problem of hotspot-temperature and critical-path delay minimization. However, existing X−fill methods are totally unaware of the core layout, and they consume all 'X' values for reducing power at the entire die area, while they neglect other test quality objectives, like the un−modeled defect coverage (UDC). An example is the popular fill−adjacent method [12], which minimizes the number of transitions at the scan chain during the scan−in process, but offers very low UDC [2], because it generates highly correlated test vectors (major parts of vectors are filled with long runs of 0s and 1s). The method in [2] offers a trade−off between power consumption and UDC, but it does not consider local thermal and delay effects of the tests.

We propose a thermal and delay aware X−fill method that offers high UDC. Layout information is used to identify the scan cells that are most important for removing hotspots at critical nets. Then, a thermal safe neighborhood is created around these nets. The rest of the scan cells are exploited to increase the UDC of the generated test vectors by the means of an output-deviation based metric [14]. The proposed method avoids any unnecessary yield loss because it reduces the delays at critical paths during testing, while it considerably increases the quality of the generated test vectors in terms of UDC.

II. PROPOSED METHOD

To avoid overheating of critical paths, the power dissipated at the critical areas, i.e., the areas around the critical paths, must be minimized. At the same time, proper testing conditions must be imposed to ensure that these areas are also protected from heat transferred from other areas of the chip. When two areas are geometrically close to each other, then a large thermal gradient causes significant heat transfer between them and quickly changes their temperatures [11]. Therefore, to protect a critical area, a thermal-safe zone must be created around it, that is able to absorb heat from this area and other neighboring areas that develop high temperatures.

The first step towards this goal is to identify critical paths geometrically, and specify, topologically, thermal-safe zones that surround them. Then, the heat generated inside these zones must be minimized. Since the largest portion of the power dissipated during testing is dissipated during scan-in, the transitions occurring inside the thermal-safe zones during this process must be minimized. Thermal-safe zones decrease the power dissipation internally, but they permit higher power dissipation externally at non-critical areas. Thus, thermal-safe zones reduce considerably the number of scan-cells exploited
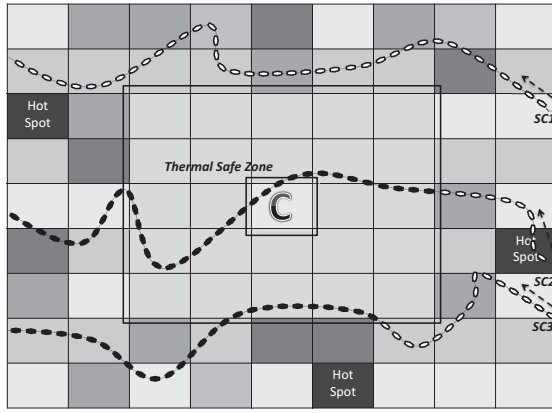
Fig. 1. Critical Area & Thermal-Safe Zone

to decrease the thermal activity of the core, and the remaining scan-cells can be used to achieve other test objectives.

Example 1. Let us consider the CUT shown in Fig. 1, which is partitioned into 9×9 equally sized blocks, and let C be the critical block. Test data are loaded using scan chains SC_1, SC_2 and SC_3 from right to left. The color of each block reflects the thermal profile of the block during testing (dark colors means high temperatures). The thermal-safe zone consists of two layers of blocks that surround block C. The test data of the black scan-cells pass through the zone during the scan-in, and thus these cells are filled using either power-oriented or defect-oriented criteria depending on their relative position at the scan-chain. The white scan-cells do not affect the thermal-safe zones and they are filled using defect-oriented criteria. ∎

The area-characterization of the core requires to identify critical blocks and create a thermal-safe zone around each one of them. To this end, the core's floorplan is first partitioned into a given number N of equally-sized rectangle blocks. Then, the critical paths are identified using post−layout timing analysis and the blocks are separated into critical blocks of zone Z_0 and non−critical blocks depending on whether they contain critical paths or not. Every non-critical block that is an immediate neighbor of a critical block is included in the first layer of the thermal-safe zone, the Z_1 zone. Every non-critical block that is an immediate neighbor of a block in the Z_1 zone is included in the second layer of the thermal-safe zone, the Z_2 zone. Every non-critical block that is included in the Z_1 zone of one critical block, is not included in the Z_2 zone of any other critical block. The thermal activity increases from zone Z_0 to zone Z_2. Outside these zones the thermal activity is left completely unconstrained. Each one of the blocks in zones Z_0, Z_1, Z_2 is assigned a different thermal weight $w_{Z_0} > w_{Z_1} > w_{Z_2} > 0$, based on the importance of thermal activity at this block (the higher is the weight, the lower must be the thermal activity of the block). Non-critical blocks are assigned zero weights.

During the loading of scan chain i scan-cell SC_j^i is considered as successor of SC_{j+1}^i and predecessor of SC_{j-1}^i. The thermal weight of each scan-cell SC_j^i depends on the block it belongs to, as well as on its relative position j in the scan chain i. The position j of SC_j^i in the scan chain i is used to measure the impact of SC_j^i on the overall number of transitions caused

by SC_j^i during the scan-in operation. Scan cells close to the scan output receive higher weights than those located close to scan inputs, because their test data need to travel long distances in the scan chains to reach them. As a result they affect the heat generation at the core during testing more than the scan-cells located close to the scan inputs. Therefore, the thermal weight of scan-cell SC_j^i that belongs to thermal zone Z_k is calculated as $W(SC_j^i) = w_{Z_k} \times \frac{1}{j}$ for $k = 0, 1, 2$.

After the scan-cells are assigned weights, the unspecified bits of the test cubes (namely test vectors with '0', '1' and 'X' logic values) are filled in such a way as to reduce the transitions inside the safe-zones. Every pair of complementary test bits at two successive scan-cells causes transitions at all their predecessor scan-cells during the scan-in process. Thus, even if a scan-cell SC_j^i is located inside a non-critical block, it impacts the thermal activity of blocks of thermal-safe zones, when these zones contain scan-cells that precede SC_j^i.

The impact IP of scan-cell SC_j^i to the power dissipation of the die is defined as the sum of the weights of SC_j^i, SC_{j+1}^i, SC_{j+2}^i, … according to the formula $IP(SC_j^i) = \sum_{m=j...L} W(SC_m^i)$, where L is the number of scan chains. A large value of $IP(SC_j^i)$ indicates that scan-cell SC_j^i should be set to the same logic value with scan-cell SC_{j-1}^i, else a logic transition will be introduced to the scan-in process, with a large impact on the thermal activity of the critical areas of the core. The value of $IP(SC_j^i)$ is normalized in the range $[0, 1]$ according to the formula $NIP(SC_j^i) = \frac{IP(SC_j^i)}{IP_{max}}$ where $IP_{max} = max\{IP(SC_j^i)\}, \forall i, j$. $NIP(SC_j^i)$ is used to fill the unspecified bits of each test cube in a probabilistic manner, starting from the scan-out cells SC_1^i and moving towards the scan-in cells SC_L^i. Specifically, let R_j^i be a random number generated in the range $[0, 1]$ for SC_j^i. If $R_j^i \leq NIP(SC_j^i)$ the unspecified value of scan-cell SC_j^i is set equal to the specified value of scan-cell SC_{j-1}^i, else it is set randomly to either logic value 0 or 1. In the case that SC_j^i does not affect much the thermal-safe zones the value of $NIP(SC_j^i)$ is low, and the probability that SC_j^i is set randomly is high. However, when SC_j^i affects thermal-safe zones the value of $NIP(SC_j^i)$ is high and there is a high probability that it is assigned the same logic value with its successor scan-cell SC_{j-1}^i.

Even though this formula achieves the required thermal objectives, there are cases that the X-fill process must be further biased towards more power friendly test vectors. To this end, a parameter $P \in [0, 1]$ is set by the designer according to the formula $R \leq NIP(SC_j^i) + P(1 - NIP(SC_j^i))$. As the value of P increases from 0 to 1, the right part of this formula increases from $NIP(SC_j^i)$ to 1, and thus the probability that this formula is true increases. As a result, more scan-cells are assigned the same logic values with their successor cells, and the power dissipation of the test vectors decreases even more.

Large values of P tend to generate test vectors with very low power dissipation. However, these test vectors are correlated due to the biased filling of their unspecified bits, and thus the UDC drops. Low values of P generate test vectors with many unspecified bits filled randomly, which increases the UDC.

Moreover, due to the probabilistic nature of the proposed method, a number of test vectors can be generated for each test cube, by repeatedly applying this X-fill technique on every test cube. Even though all the test vectors generated for one test cube are equally effective in term of power dissipation at the critical blocks and the thermal-safe zones for a given value of P, they offer different UDC.

In order to select the best test vector for each test cube, the test vectors are evaluated using the output-deviation metric proposed in [14]. Output deviations are probability measures at primary outputs and pseudo-outputs that indicate the likelihood of error detection. The output deviation for input pattern tp and an output/pseudo-output w is defined as the probability output w to receive the opposite than the error-free logic value. Each test vector is applied with two capture cycles r_1 and r_2 (i.e., we assume the Launch-On-Capture technique as it is common in industry). For each output w, the generated test vectors are partitioned into four groups: those producing fault-free response 0 and 1 at capture cycles r_1 and r_2. The output-deviation values of all generated test vectors are calculated and the largest value for every output w and for each fault-free response v at capture cycle r_k are used to evaluate the test vectors. The selection process ensures that one test vector is selected for every test cube, and the final set of selected test vectors maximizes the output deviations at all outputs.

III. EXPERIMENTAL RESULTS

To evaluate the proposed methodology, we run experiments on the largest *ISCAS'89* and *IWLS* benchmark cores. Each core was synthesized using the 45nm Nangate technology, and its layout was generated using commercial DFT-enabled tools. The floorplan of each core was partitioned into a number of blocks, which was determined based on the area and the number of scan cells of the core. The critical paths were identified using post-layout timing analysis based upon standard operating condition. All paths with delays within a margin of 90% of the worst path delay were classified as critical paths. A nearest-neighborhood search was applied to determine the blocks of zones Z_1, Z_2. The weights w_{Z_0}, w_{Z_1} and w_{Z_2} were set equal to 3, 1.5 and 1 respectively.

The proposed method was developed using C++. The Random-Fill (RF), the Fill-Adjacent (FA) [12] and the Modified-Fill-Adjacent (MFA) method [2] were also implemented for comparison purposes. All these methods were applied on compacted test sets generated for complete coverage of stuck−at faults. Similar to [2] these test sets were evaluated for UDC by using the transition-fault model as surrogate fault model (a fault model that is not targeted by the generated test sets). In both the proposed and the MFA methods, the same output-deviation based metric was used, and 30 test vector candidates were generated per test cube [14].

To evaluate the thermal activity of each core, the power profile and the floorplan of every core and every test set were given as inputs to the *Hotspot* tool [11]. Then, to measure the impact of the evaluated methods on the delay of critical paths, we used the steady profile generated by *Hotspot* for each core
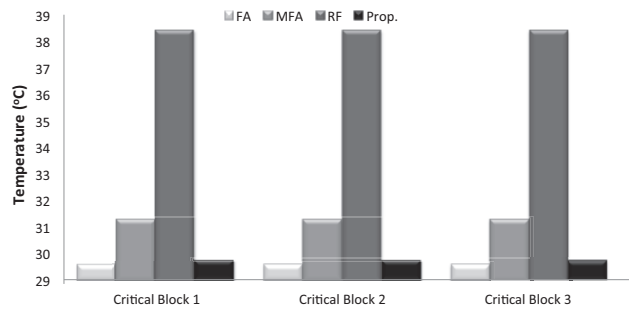

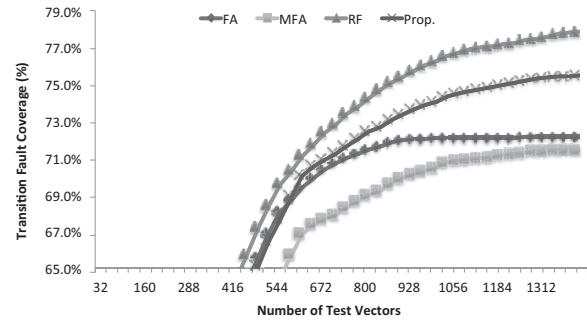Fig. 2.   Temperature at critical areas of Ethernet core

Fig. 3.   Transition-fault coverage ramp-up for Ethernet core

and each test set, to determine the operating condition (OC) of each block and every scan cell inside the block. This OC was provided to a commercial tool, to perform timing analysis using on-chip variations and two slow-corner libraries of the 45nm Nangate technology, the worst-low and the slow library. Both libraries were characterized at low power supply voltage equal to 0.95V, and temperatures set at −40 and 125 grades in Celsius scale respectively. Then, the path delays at every block were generated by the static analyzer of a commercial tool, which interpolated the timing information of the library to estimate the delays of standard cells at the given OC.

The average power consumption of the proposed method on the Ethernet core, which is the largest and more representative core, is equal to 9.54mW, and it is slightly higher than the 9.20mW consumed by the FA. The power consumption of the MFA is much higher, and it is equal to 12.73mW. The power consumption of the RF is equal to 27.65mW and it is almost three times higher than both of the proposed and the FA methods. The post-layout timing analysis identified 3 critical blocks, 7 blocks in the Z_1 zone and 11 blocks in the Z_2 zone (the rest 79 blocks are non-critical). The "steady" temperature at each one of the critical blocks is depicted in Fig. 2 for FA, MFA, RF and the proposed method. Clearly, the temperature at the critical blocks for both the proposed method and FA is the lowest one. MFA increases the temperature by 2 degrees and RF increases the temperature by 10 degrees. According to the thermal-aware timing analysis the delay of the critical path in the proposed and the FA methods was found to be equal to 1.61ns, while the worst path delays of MFA and RFA were found to be equal to 1.64ns and 1.72ns respectively.

To compare the four X-fill methods with respect to the UDC, we present in Fig. 3 the transition-fault coverage provided by

TABLE I
TEMPERATURE (CELSIUS DEGREES)

Benchmark Cores	RF	Proposed Method			MFA	FA
		$P = 0$	$P = 0.85$	$P = 1$		
ethernet	38.43	37.98	29.73	37.19	31.26	29.57
des3	35.38	32.63	31.13	31.75	32.10	31.8
aes_cipher	41.57	40.79	38.88	39.68	40.82	38.44
wb_conmax	32.09	31.49	30.11	30.72	31.78	29.75
tv80s	30.57	28.93	27.54	27.35	28.6	27.24
usbf	33.53	31.25	28.86	28.52	29.37	28.31
s38417	29.12	28.14	27.29	27.13	27.78	27.11
s38584	30.3	29.88	28.29	29.32	29.36	28.8

TABLE II
TRANSITION FAULT COVERAGE (%)

Benchmark Cores	RF	Proposed Method			MFA	FA
		$P = 0$	$P = 0.85$	$P = 1$		
ethernet	77.89	77.28	75.48	76.46	71.48	72.14
des3	90.61	88.80	84.01	80.67	80.55	80.21
aes_cipher	84.76	84.54	83.47	83.02	84.41	82.47
wb_conmax	93.8	93.37	92.13	92.25	93.68	91.35
tv80s	42.01	41.91	39.98	38.24	39.82	38.22
usbf	23.95	23.88	22.60	21.86	22.60	21.83
s38417	93.23	91.96	86.77	79.29	84.59	79.25
s38584	86.87	84.89	79.74	80.02	79.63	77.51

each one of them. The x-axis presents the number of vectors applied, and the y-axis the transition-fault coverage. Clearly, both FA and MFA provide lower transition-fault coverage than the RF and the proposed methods. The highest transition-fault coverage is provided by the RF, which however is comparable to the transition-fault coverage offered by the proposed method. In addition, both of them offer very high ramp-up, which offers further test time savings at abort-at-first-fail environments. However, the high power consumption and temperature of the RF do not permit its application in a thermal constrained environment. Therefore, the proposed method outperforms the other methods when all four parameters of power, temperature, delay and UDC are considered.

Tables I, II present the steady temperature and the transition-fault coverage of the RF, FA, MFA and the proposed method for $P = 0$, $P = 0.85$ and $P = 1$. When $P = 0$ only a few scan cells are power constrained, and thus the results are close to the RF method. When $P = 1$ case many scan cells are power constrained, and thus the results are close to the FA method. In all cases, the proposed method offers power consumption and temperature that is very close to the most power-efficient FA method, while at the same time it offers UDC that is very close to the RF method. Therefore, we conclude that the proposed method combines the advantages of both FA and RF method, and it offers high UDC without any adverse impact on power dissipation, temperature and critical-path delay during testing.

IV. CONCLUSIONS

In this paper we have presented a critical path−oriented and thermal-aware 'X'−fill method, which offers high un-modeled defect coverage. Extensive experiments on the large ISCAS'89 and the IWLS benchmark circuits have shown that the power dissipation and the thermal activity at the most delay-critical regions of the cores were considerably reduced. At the same time, the generated test sets offered high un-modeled defect coverage, which is similar to that provided by power-unaware and thermal-unaware approaches.

REFERENCES

[1] N. Aghaee, Z. He, Z. Peng, and P. Eles, "Temperature aware soc test scheduling considering inter-chip process variation," in *19th IEEE ATS*, Dec. 2010, pp. 395 –398.
[2] S. Balatsouka, V. Tenentes, X. Kavousianos, and K. Chakrabarty, "Defect aware x-filling for low-power scan testing," in *DATE*, 2010, pp. 873–878.
[3] A. Chandra and K. Chakrabarty, "Low-power scan testing and test data compression for system-on a-chip," *IEEE Trans. on CAD,*, vol. 21, no. 5, pp. 597 –604, 2002.
[4] A. Chandra and R. Kapur, "Bounded adjacent fill for low capture power scan testing," in *Proc. IEEE VTS*, 27 2008-may 1 2008, pp. 131 –138.
[5] M. Cho and D. Pan, "Peakaso: peak-temperature aware scan-vector optimization," in *24th IEEE VTS*, April-4 May 2006, p. 6 pp.
[6] D. Czysz, G. Mrugalski, J. Rajski, and J. Tyszer, "Low-power test data application in edt environment through decompressor freeze," *IEEE Trans. on CAD,*, vol. 27, no. 7, pp. 1278 –1290, 2008.
[7] D. Czysz, et. al., "Deterministic clustering of incompatible test cubes for higher power-aware edt compression," *IEEE Trans. on CAD*, vol. 30, no. 8, pp. 1225 –1238, 2011.
[8] D. R. Bild, et. al., "Temperature-aware test scheduling for multiprocessor systems-on-chip," in *ICCAD*, 2008, pp. 59 –66.
[9] P. Girard, L. Guiller, C. Landrault, and S. Pravossoudovitch, "A test vector inhibiting technique for low energy bist design," in *17th IEEE VTS,*, 1999, pp. 407 –412.
[10] Z. He, Z. Peng, and P. Eles, "Simulation-driven thermal-safe test time minimization for system-on-chip," in *ATS*, 2008, pp. 283 –288.
[11] Huang Wei, et. al., "Hotspot: a compact thermal modeling methodology for early-stage vlsi design," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 14, no. 5, pp. 501 –513, may 2006.
[12] K. M. Butler, et. al., "Minimizing power consumption in scan testing: pattern generation and dft techniques," in *ITC*, 2004, pp. 355 – 364.
[13] S. Kajihara, K. Ishida, and K. Miyase, "Test vector modification for power reduction during scan testing," in *IEEE VTS*, 2002, pp. 160–165.
[14] X. Kavousianos, V. Tenentes, K. Chakrabarty, and E. Kalligeros, "Defect-oriented lfsr reseeding to target unmodeled defects using stuck-at test sets," *IEEE Trans. on VLSI Systems*, vol. 19, no. 12, pp. 2330 –2335, 2011.
[15] J. Lee and N. Touba, "Low power test data compression based on lfsr reseeding," in *IEEE ICCD*, oct. 2004, pp. 180 – 185.
[16] J. Li, Q. Xu, Y. Hu, and X. Li, "ifill: An impact-oriented x-filling method for shift- and capture-power reduction in at-speed scan-based testing," in *DATE*, march 2008, pp. 1184 –1189.
[17] W. Li, S. Reddy, and I. Pomeranz, "On reducing peak current and power during test," in *IEEE Computer Society Annual Symposium on VLSI*, may 2005, pp. 156 – 161.
[18] C. Liu, K. Veeraraghavan, and V. Iyengar, "Thermal-aware test scheduling and hot spot temperature minimization for core-based systems," in *20th IEEE Int. Symp. on DFT in VLSI Systems*, 2005, pp. 552 – 560.
[19] S. Remersaro, X. Lin, Z. Zhang, S. M. Reddy, I. Pomeranz, and J. Rajski, "Preferred fill: A scalable method to reduce capture power for scan based designs," in *ITC*, oct. 2006, pp. 1 –10.
[20] P. Rosinger, B. Al-Hashimi, and K. Chakrabarty, "Rapid generation of thermal-safe test schedules," in *DATE*, March 2005, pp. 840 – 845.
[21] X. Wen, S. Kajihara, K. Miyase, T. Suzuki, K. Saluja, L.-T. Wang, K. Abdel-Hafez, and K. Kinoshita, "A new atpg method for efficient capture power reduction during scan testing," in *24th IEEE VTS*, 30 2006-may 4 2006, pp. 6 pp. –65.
[22] X. Wen, Y. Yamashita, S. Morishima, S. Kajihara, L.-T. Wang, K. Saluja, and K. Kinoshita, "Low-capture-power test generation for scan-based at-speed testing," in *ITC*, nov. 2005, pp. 10 pp. –1028.
[23] X. Wen, et. al,, "A highly-guided x-filling method for effective low-capture-power scan test generation," in *ICCD*, oct. 2006, pp. 251 –258.
[24] X. Wen, et. al., "A capture-safe test generation scheme for at-speed scan testing," in *13th ETS*, may 2008, pp. 55 –60.
[25] C. Yao, K. Saluja, and P. Ramanathan, "Partition based soc test scheduling with thermal and power constraints under deep submicron technologies," in *ATS*, nov. 2009, pp. 281 –286.
[26] ——, "Power and thermal constrained test scheduling," in *ITC*, nov. 2009, p. 1.