Invariance Checking based Trojan Detection Method for Three-Dimensional Integrated Circuits

Zhiming Zhang and Qiaoyan Yu
Department of Electrical and Computer Engineering
University of New Hampshire, Durham, New Hampshire 03824, United States
Email: qiaoyan.yu@unh.edu

***Abstract*—Recently literature indicates that stack based three-dimensional (3D) integration techniques may bring in new security vulnerabilities, such as new attack surfaces for hardware Trojan (HT) insertion. Compared to its two-dimensional counterpart (2DHTs), a 3D hardware Trojan (3DHT) could be stealthily distributed in multiple tiers in a single 3D chip. Although the comprehensive models for 3DHTs are available in recent work, there still lacks 3DHT detection and mitigation methods, especially run-time countermeasures against 3DHTs. This work proposes to leverage the 3D communication infrastructure, 3D network-on-chips (NoCs), to tackle the cross-tier hardware Trojans in stacked multi-tier chips. An invariance checking method is further proposed to detect the Trojans that induce malicious NoC packets or facilitate information leak. The proposed method is successfully deployed in NoC routers and achieves a Trojan detection rate of over 94%. The synthesis result of a hardened router at a 45nm technology node shows that the proposed invariance checking only increases the area by 6.49% and consumes 3.76% more dynamic power than an existing 3D router. The NoC protected with the proposed method is applied to the image authentication in a 3D system. The case study indicates that the proposed security measure reduces the correlation coefficient by up to 31% over the baseline.**

I. Introduction

As the semiconductor manufacturing process is approaching the physical limit of silicon, it is difficult to continue the Moore's Law [1]. Innovative integration is one of the ways to achieve "More than Moore" [2]. Three-dimensional (3D) integration emerges as a strong candidate [3], which vertically integrates multiple independently fabricated integrated circuits (ICs) as 3D tiers [4]. The stacked 3D structure can effectively increase the device density. Furthermore, the utilization of through-silicon vias (TSVs) as inter-tier connections reduces the global wire length, thus improving system performance and saving power consumption on global interconnects.

However, 3D ICs may bring in unique and new security vulnerabilities [5]. Outsourcing fabrication of individual 3D tiers provides malicious foundries a chance to insert hardware Trojans. The 3D ICs' special stacking structure leaves attackers more exploration space to build new types of hardware Trojans [6]. Split manufacturing techniques separate a complete design into multiple incomplete portions, one for a untrusted foundry, thus thwarting reverse engineering attacks. Unfortunately, the heuristics of electronic design automation tools could nullify the split manufacturing effort [7], [8], by leaving attackers hints to recover the missing connections.

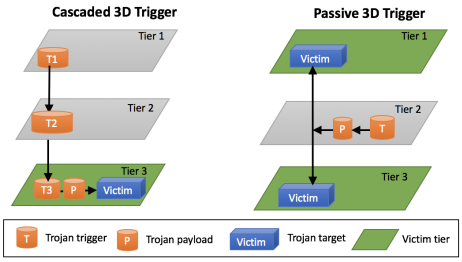
In the survey [9], hardware Trojan detection methods are categorized into logic testing, side-channel analysis, and image analysis on the Scanning Electron Microscope pictures of the de-metalized chips. Those existing Trojan detection methods are mainly designed for Trojans in 2D ICs. As 3D ICs usually have larger variation on process, voltage and temperature (PVT) [10], [11], [12], 2D side-channel signal based Trojan detection methods will result in high false-negative detection rate [3]. Compared to 2D chips, 3D ICs inherently make more resources available for attackers to design Trojan triggering mechanisms. The triggering probability of 3D Trojans could be even lower than that in 2D scenarios. As a result, it is more challenging to generate effective testing vectors to trigger 3D Trojans [6]. Due to limited probing space, Trojan detection via probing techniques is not scalable in stacked 3D ICs. Although each die and TSV can be examined during the pre-bond and mid-bond stages, the testing probe may damage some TSVs and thus harm downstream integration [13].

In this work, we propose a run-time Trojan detection and mitigation method to complement the existing countermeasures against 2D and 3D hardware Trojans. Our main contributions are as follows: (1) our method proposes to leverage the 3D communication infrastructure, 3D-Network-on-chips (3D-NoCs), to tackle the cross-tier hardware Trojans in stacked multi-tier chips, and (2) an invariance checking method is proposed to detect Trojans, which introduce malicious NoC packets or facilitate information leak among 3D tiers.

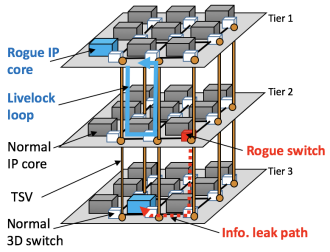
The rest of this paper is organized as below. Section 2 presents the attack model interested in this work. Section 3 proposes a novel invariance checking method to thwart 3DHT insertion attacks. Simulation and synthesis results are provided in Section 4. This work is concluded in Section 5.

II. Attack Model

The models for representable 3DHTs are introduced in the recent work [6]. As highlighted in [6], the most significant difference between 3DHTs and conventional 2DHTs is that the trigger and payload circuits of 3D Trojans are not located in the same 3D tier. Figure 1(a) shows examples of *Cross-Tier Trojans*. In the left case, the trigger circuits are distributed in the top and middle tiers, and they jointly trigger the payload in the bottom tier. This triggering mechanism can have a much lower triggering probability than the Trojan trigger in a single tier. In the right case, neither the trigger nor the payload circuit



(a)



(b)

Fig. 1. Attack scenarios considered in this work. (a) Characterization of 3DHTs, and (b) an example of the activated 3DHT effect [14].

is located in the same tier where the victim remains; the data transmission between victims is leaked due to the Trojan in the middle tier. This type of Trojans does not interrupt normal data communication. If 3DHTs described in Fig. 1(a) are placed in a 3D-NoC system shown in Fig. 1(b), that system may suffer from livelock and information leak [14]. In this work, we analyze the characteristics of these two types of 3DHTs and propose a mitigation method accordingly.

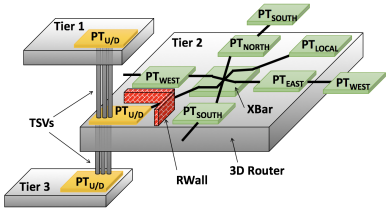
Our 3DHT detection and mitigation method is based on the following assumptions: (1) each tier is a completed die (rather than a die appeared in the middle of split manufacturing), (2) the communication between tiers is at IP core level, rather than functional block level, and (3) the routing rule used in 3D routers is public to attackers.

III. Proposed Invariance Checking Based 3D Hardware Trojan Detection and Mitigation

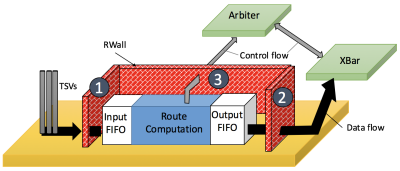
A. Proposed Hardened Router Architecture for 3D-NoCs

Cross-tier hardware Trojans (or multi-tier collaborative Trojans) emerge as a new hardware Trojan model for 3D ICs. Due to 3DHTs' unique threat characterizations, it is a pressing need to investigate new detection and mitigation methods specifically for 3DHTs. Moreover, the countermeasures against 3DHTs are expected to be compatible with the architecture of 3D systems. The defense mechanism should be integrated into the system specification, rather than an add-on component patched afterwards.

We propose to tackle cross-tier Trojans with a *router-hardened 3D-NoC*, which is the communication backbone for 3D integrated circuits and systems. The proposed security mechanism complements to the investigation on other 3D-NoC aspects (thermal issue [15], architecture [16], and usage in computationally intensive applications [17]). As 3D IC testing



(a)



(b)

Fig. 2. Proposed cross-tier Trojan detection. (a) Proposed router architecture for 3D-NoC, and (b) block diagram of vertical port $PT_{U/D}$ protected with invariance checking based hardware firewall.

is not as thorough as 2D IC verification, there will be residual hardware Trojans, especially cross-tier Trojans, harming 3D systems after testing [6]. To address this issue, we propose a run-time Trojan detection and mitigation method against cross-tier hardware Trojans.

Figure 2(a) shows the architecture of proposed 3D router, in which the five ports PT_{NORTH}, PT_{SOUTH}, PT_{WEST}, PT_{EAST} and PT_{LOCAL} are used for the intra-tier communication, and PT_U and PT_D are responsible for transferring data to the upper and lower tiers, respectively. To detect and mitigate potential 3DHT intrusion, we propose a *RWall*, an invariance checking based hardware firewall, to thwart unauthorized access to the other router ports and prevent 3D-NoCs from sniffing attacks. The zoom-in view for the proposed RWall is illustrated in Fig. 2(b). The RWall ① examines whether a NoC flit (i.e., a basic flow unit in NoCs [18]) is tampered during its propagation from other tiers. Such tampering could be induced due to malicious through-silicon-vias (TSVs) or compromised input FIFOs. The RWall ② terminates the requests from $PT_{U/D}$ to use other ports. The RWall ③ monitors the duplication of malicious NoC packets among multiple output ports. The combined effect of ② ③ blocks the illegal information leak and prevents the 3D communication infrastructure from being tampered at the router level.

B. Proposed Invariance Checking within NoC Router

Invariance checking is a cost-effective method for fault tolerance within NoC [19]. Following that footprint, we propose to leverage the invariance within 3D-NoCs to tackle cross-tier hardware Trojans. In this subsection, we first examine the suitable invariance at the flit, port, and router levels and then develop a practical implementation algorithm. Figure 3 provides the detailed view of a hardened NoC router. A typical router for 2D-NoCs consists of five bi-directional routing ports, each of which is composed of input/output FIFOs, a routing computation, a crossbar (XBar), and an arbiter. For 3D-

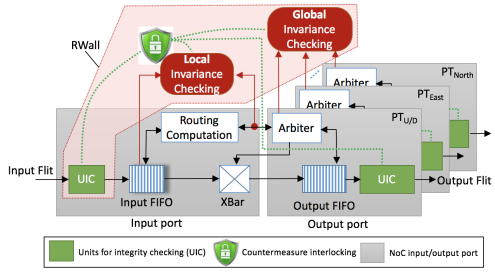


Fig. 3. Proposed invariance checking in NoC router.

NoCs, up-stream and down-stream ports ($PT_{U/D}$) are added to access other 3D tiers. Global invariance checking examines any violation of port access among seven bi-directional NoC router ports. As our defense target is cross-tier Trojans, we pay extra attention to $PT_{U/D}$ by adding local invariance checking. The complete Trojan detection and mitigation algorithm is shown in Algorithm 1. More precisely, the proposed algorithm is implemented at the flit, port, and router levels.

At flit level, tampered flits (router inputs) will be detected by the *Units for Integrity Check* (UIC). Error control coding (ECC) is a common low-cost approach for data integrity detection. We propose to use two-level ECC based integrity check as expressed in Equations (1) and (2). In addition to encode/decode the entire flit, limited configurations of critical flit fields will be encoded for another level integrity check.

$$UIC_{alert1} = DecFun_1(Flit) \tag{1}$$

In which, $Flit$ is a tuple of {flit type, flit source, flit destination, hopping path, routing priority, parity check}.

$$UIC_{alert2} = DecFun_2(flit_{type}, field_{sel}, parity_{2nd}) \tag{2}$$

Where, $flit_{type}$ indicates whether the flit is a header or payload, $field_{sel}$ is several selected fields for second-level integrity check (e.g., flit source and destination), $parity_{2nd}$ is the second level coding algorithm for integrity check. The two alert signals UIC_{alert1} and UIC_{alert2} from UIC will stop the malicious flit from entering or leaving the suspicious router port.

At port level, the invariance for Trojan detection includes illegal port requests and mismatched control-data flows. Only a header flit can request to reserve port-to-port connection. Any port-requests issued from other flits indicate Trojan intrusion. Since port-to-port communication is exclusive, each output port can accept one and only one request from all other input ports in the same router. Likewise, an input port cannot simultaneously issue multiple requests to access more than two output ports. Another invariance is originated from the routing history. The incoming and outgoing port request (RC_{req}) should match to packet source ($SRID$), destination ($DRID$) and the current router IDs ($CRID$). The routing inverse function expressed in Eq.(3) facilitates the detection of tampered routing history.

$$Local_{invar} = RInverse(SRID, DRID, CRID, RC_{req}) \tag{3}$$

Algorithm 1: Proposed multi-level invariance check.

Data: Packets through a 3D-NoC router
Result: Alert for 3DHT intrusion

UIC_{alert1} (Input flits);
UIC_{alert2} (Selective flit breakdown fields);
while *Cross-tier packets being transferred* **do**
  //Local invariance checking;
  if $\Sigma\left(RC_{req\ from\ PT_{U/D}}\right) > 1$ **then**
    Information leak detected;
  else
    if $\Sigma\left(PT_{U/D} \rightleftharpoons PortFIFOs\right) > 1$ **then**
      Intrusion attack detected;
      Terminate cross-tier communication;
    else
      //Global invariance checking;
      if *RInverse outputs mismatch* RC_{req} **then**
        Intrusion attack detected;
        Drop malicious flits;
      else
        Pass local invariance check;
      end
    end
  end
  Use encryption key to unlock arbiter tables;
end

As the information regarding the complete routing path varies with different NoC applications, the hardware Trojans inserted in 3D-NoC design time is not able to bypass all of the routing consistency check. Moreover, the invariance rules mentioned above are not mutable once the router is power up. Thus, our invariance checking does not only detect malfunctions but also monitors illegal behaviors triggered by 3DHTs.

At router level, our method examines the invariance available among arbiters. In the baseline, the arbiter grants one of the port requests based on even opportunity (i.e., round robin rule). Updating on the round-robin register tables has to satisfy the priority rule. Any interrupts appeared in the middle of packet transmission indicates the occurrence of an attack. Logic encryption [20] is adopted to harden the round-robin tables. In our case study, we use a 7-bit key to unlock the updating logic for arbiters in 3D routers. The incorrect encryption key will terminate the arbiter's normal function.

IV. Experimental Results

A. Area, Power, and Delay

We implemented the proposed 3D NoC router in Verilog HDL and synthesized the HDL code in Synopsys Design Compiler with a 45nm NCSU openPDK technology. The flit width for the NoC is 32 bits. The input and output FIFOs are 32-bit single-depth buffers. Round-robin arbitration was used in the router arbiter. We set the clock frequency to 1 GHz. The area, delay and power consumption for the baseline [5] and our method are compared in Table I. As shown, our method is

TABLE I
COMPARISON OF AREA, DELAY AND POWER

Metric under comparison	Baseline [5]	Proposed	Overhead
Area (μm^2)	19731	21005	6.46%
Delay (ns)	0.86	0.94	9.30%
Dynamic power (mW)	13.0733	13.5646	3.76%
Leakage power (μW)	108.0194	115.6355	7.05%

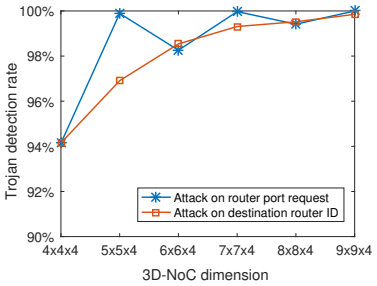


Fig. 4. Trojan detection rate of proposed method.

a lightweight countermeasure. The area is only increased by 6.49%. The overhead on dynamic power and leakage power are 3.76% and 7.05%, respectively. As we add invariance checking in the cross-bar unit, the worst-case delay of our router is 9.3% higher than that of the baseline.

B. Trojan Detection Rate

The proposed invariance checking examines the consistency between the port requests and the routing history to detect 3DHTs. We randomly altered the port request to access upper and lower tiers (i.e., attack on router port requests) or the destination router ID carried in the NoC header flit (i.e. attack on destination router ID). Each Trojan detection rate was obtained from 10,000 simulations. As shown in Fig. 4, the Trojan detection rate of our method is above 94%, no matter the Trojan attack is on the port request signals or the destination router ID.

C. Impact of Cross-tier Trojan Mitigation on Image Authentication in a 3D system

In our case study, we used a 3D-NoC to perform image based authentication. The experimental setup is shown in Fig. 5(a). Through the 3D NoC routers, tier 1 and tier 2 transmit two images to tier 3 for image authentication. Pearson correlation coefficient (PCC) is adopted as the metric to indicate whether the two images from tiers 1 and 2 depict the same person. Hardware Trojan insertion happens in the 3D router located in tier 2 or the TSVs connecting tiers 2 and 3. The activated Trojan tampers the header flits or payload flits of the image packets. The proposed method filters out the tampered flits. If a header flit is altered by a 3DHT, the entire targeted packet is replaced by a malicious packet (baseline) or dropped with notifications (proposed). If a payload flit is sabotaged by a 3DHT, only that flit is substituted by a dummy flit (baseline) or deleted (proposed) and the rest flits in that packet remain the same. The PCC between images from tiers 1 and 2 are computed in the victim unit (i.e., Corr in Fig. 5(a)). As shown in Fig. 5(b) and 5(c), our scheme removes malicious flits significantly and thus reduces the correlation coefficient.

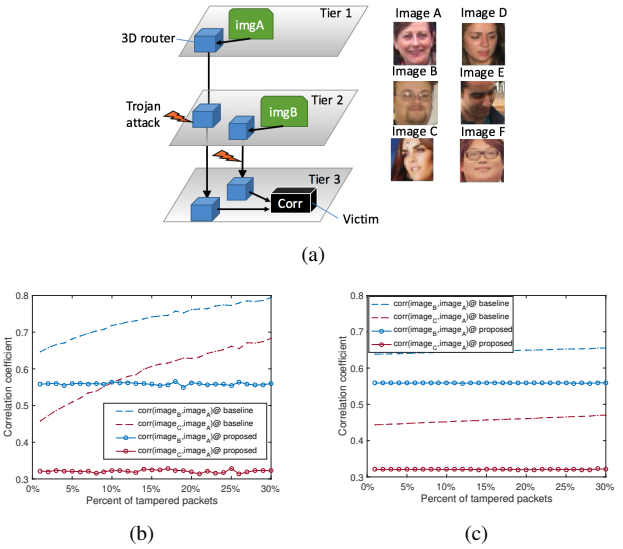


Fig. 5. Impact of Trojans on the application of 3D image authentication. (a) attack scenario, (b) impact of attacking header flit on correlation coefficient, and (c) impact of attacking payload flits on correlation coefficient.

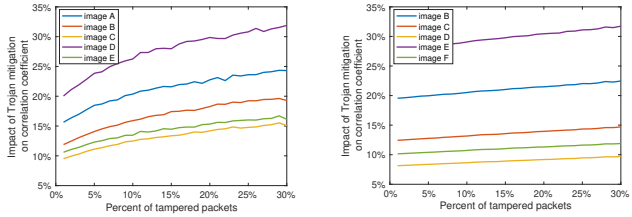


Fig. 6. Reduction on correlation coefficient achieved by Trojan mitigation.

This means that the tampered images are less likely to pass the authentication. For instance, the proposed method can reduce the PCC from 0.6755 to 0.3122. As each NoC packet is composed of one header flit and multiple payload flits, the baseline scheme is more sensitive to Trojan attacks aiming at header flits than at payload flits. In contrast, our Trojan mitigation overcomes that sensitivity.

We examined the Trojan mitigation effect with six images shown in Fig. 5(a). Images B, C, D, E, and F are correlated with image A (after Trojan detection and mitigation). As shown in Fig. 6(a), the proposed method can reduce the PCC by 31%. As the percent of tampered packets increases, our mitigation method will further reduce the correlation coefficient. The exact amount of reduction on correlation coefficient varies with the images used in authentication.

V. CONCLUSION

The emerging 3D integration techniques potentially bring in attack surfaces for new type of hardware Trojans, cross-tier 3D Trojans. Given the 3D Trojan models published in recent literature, this work proposes to leverage 3D-NoC architecture to detect and mitigate the newly characterized hardware Trojans. Invariance on port access and routing history is exploited in this work to perform run-time Trojan detection. Simulation results show that the proposed method achieves a high Trojan detection rate at minor cost on area and power consumption.

REFERENCES

[1] J. Knechtel, O. Sinanoglu, I. A. M. Elfadel, J. Lienig, and C. C. Sze, "Large-Scale 3D Chips: Challenges and Solutions for Design Automation, Testing, and Trustworthy Integration," *IPSJ Transactions on System LSI Design Methodology*, vol. 10, pp. 45–62, 2017.

[2] W. Arden, M. Brillouët, P. Cogez, M. Graef, B. Huizing, and R. Mahnkopf, "More-than-moore white paper." http://www.itrs2.net/uploads/4/9/7/7/49775221/irc-itrs-mtm-v2$_3$.*pdf*, 2010.

[3] Y. Xie, C. Bao, C. Serafy, T. Lu, A. Srivastava, and M. Tehranipoor, "Security and Vulnerability Implications of 3D ICs," *IEEE Transactions on Multi-Scale Computing Systems*, vol. 2, pp. 108–122, Apr 2016.

[4] F. Imeson, A. Emtenan, S. Garg, and M. Tripunitara, "Securing Computer Hardware Using 3D Integrated Circuit (IC) Technology and Split Manufacturing for Obfuscation," in *Proc. the 22nd USENIX Security Symposium*, pp. 495–510, 2013.

[5] J. Dofe, Q. Yu, H. Wang, and E. Salman, "Hardware Security Threats and Potential Countermeasures in Emerging 3D ICs," in *Proc. Great Lakes Symposium on VLSI (GLSVLSI)*, pp. 69–74, May 2016.

[6] Z. Zhang and Q. Yu, "Modeling Hardware Trojans in 3D ICs," in *Proc. IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, pp. 483–488, July 2019.

[7] J. J. Rajendran, O. Sinanoglu, and R. Karri, "Is Split Manufacturing Secure?," in *Proc. the Conference on Design, Automation and Test in Europe (DATE)*, pp. 1259–1264, Mar 2013.

[8] Y. Wang, P. Chen, J. Hu, G. Li, and J. Rajendran, "The Cat and Mouse in Split Manufacturing," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 26, pp. 805–817, May 2018.

[9] S. Bhunia, M. S. Hsiao, M. Banga, and S. Narasimhan, "Hardware Trojan Attacks: Threat Analysis and Countermeasures," *Proceedings of the IEEE*, vol. 102, pp. 1229–1247, Aug 2014.

[10] D. Juan, S. Garg, and D. Marculescu, "Statistical Thermal Evaluation and Mitigation Techniques for 3D Chip-Multiprocessors in The Presence of Process Variations," in *Proc. the Conference on Design, Automation Test in Europe (DATE)*, pp. 1–6, Mar 2011.

[11] S. Garg and D. Marculescu, "System-Level Process Variability Analysis and Mitigation for 3D MPSoCs," in *Proc. the Conference on Design, Automation Test in Europe (DATE)*, pp. 604–609, Apr 2009.

[12] J. Dofe and Q. Yu, "Exploiting PDN Noise to Thwart Correlation Power Analysis Attacks in 3D ICs," in *Proc. 2018 ACM/IEEE International Workshop on System Level Interconnect Prediction (SLIP)*, pp. 1–6, June 2018.

[13] E. J. Marinissen, "Challenges and Emerging Solutions in Testing TSV-Based 2 1 over 2D- and 3D-Stacked ICs," in *Proc. the Conference of Design, Automation Test in Europe Conference Exhibition (DATE)*, pp. 1277–1282, Mar 2012.

[14] Z. Zhang and Q. Yu, "Towards Energy-Efficient and Secure Computing Systems," *Journal of Low Power Electronics and Applications*, vol. 8, no. 4, p. 48, Dec 2018.

[15] A. Zia, S. Kannan, H. J. Chao, and G. S. Rose, "3D NoC for Many-Core Processors," *Microelectronics Journal*, vol. 42, pp. 1380 – 1390, Dec 2011.

[16] K. J. Chen, C. Chao, and A. A. Wu, "Thermal-Aware 3D Network-On-Chip (3D NoC) Designs: Routing Algorithms and Thermal Managements," *IEEE Circuits and Systems Magazine*, vol. 15, pp. 45–69, Nov 2015.

[17] B. K. Joardar, W. Choi, R. G. Kim, J. R. Doppa, P. P. Pande, D. Marculescu, and R. Marculescu, "3D NoC-Enabled Heterogeneous Manycore Architectures for Accelerating CNN Training: Performance and Thermal Trade-offs," in *Proc. the International Symposium on Networks-on-Chip (NOCS)*, pp. 18:1–18:8, Oct 2017.

[18] J. Frey and Q. Yu, "A Hardened Network-on-Chip Design Using Runtime Hardware Trojan Mitigation Methods," *Integration, the VLSI Journal*, vol. 56, pp. 15–31, Jan 2017.

[19] A. Prodromou, A. Panteli, C. Nicopoulos, and Y. Sazeides, "NoCAlert: An On-Line and Real-Time Fault Detection Mechanism for Network-on-Chip Architectures," in *Proc. the IEEE/ACM International Symposium on Microarchitecture*, pp. 60–71, Dec 2012.

[20] J. Dofe and Q. Yu, "Novel Dynamic State-Deflection Method for Gate-Level Design Obfuscation," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 37, pp. 273–285, Feb 2018.