Emerging Applications of 3D Integration and Approximate Computing in High-Performance Computing Systems: Unique Security Vulnerabilities

Pruthvy Yellu, Zhiming Zhang, Mohammad Mezanur Rahman Monjur, Ranuli Abeysinghe, and Qiaoyan Yu

Abstract—**High-performance computing (HPC) systems rely on new technologies such as emerging devices, advanced integration techniques, and computing architecture to continue advancing performance. The adoption of new techniques could potentially leave high-performance computing systems vulnerable to new security threats. This work analyzes the security challenges in the HPC systems that employ three-dimensional integrated circuits and approximating computing. Case studies are provided to show the impact of new security threats on the system integrity and highlight the urgent need for new security measures.**

Index Terms—**Computing systems, hardware security, approximate computing, machine learning, deep learning, reliability, fault tolerance, three-dimensional (3D) integrated circuit.**

I. Introduction

High-performance computing (HPC) is essential for advancing the study of nuclear energy, astrophysics, biology, chemistry, national security, and many other fields [1], [2]. HPC systems use supercomputers and computing clusters to solve large computational problems [3] and provide tremendous computing power for modeling or facilitating to make important decisions. Progress on the development of HPC systems relies on new algorithms for massive parallelism, new process unit architectures composed of new accelerators and general-purpose processors, new fast memory materials and new integration technologies [4].

Numerous industries are placing their trust and sensitive data in HPC systems. This fact underscores the need for secure HPC infrastructure in diverse fields such as disease eradication, biomedical research, and the geological and mining industry [2], [5], [6]. As security concerns on the semiconductor supply chain attract growing attention [7], trustworthy hardware for HPC emerges as an critical challenge for HPC development. Hence, the security issues of HPC systems' hardware components is the primary focus of this work.

The design of HPC involves the utilization of emerging technologies, like three-dimensional integrated circuits (3D ICs). The high device density, high bandwidth, and low power consumption qualifications of 3D architecture can perfectly help HPC systems achieve the goal of high performance at a low cost. However, the internal security issue of 3D ICs make the HPC systems built on them become vulnerable to malicious attacks. HPC systems require high computational capability to provide better performance. However, with the energy constraints and limited resources, traditional HPC systems still may not be able to provide the optimal energy-performance tradeoff. In order to solve this issue, approximate computing has emerged as a technique which improves the computational performance with acceptable error tolerance in the output. Nevertheless, approximate computing techniques could lead HPC systems to be vulnerable to new security threats.

The organization of the rest of this work is as follow: Section II provides examples of unique threats on HPC systems, Section III introduces the security threats in 3D ICs and systems, Section IV discusses potential attack surfaces in approximate computing systems. This work is concluded in Section VI.

II. Unique Security Threats on HPC Systems

As reliance on HPCs and their superior processing power becomes more ubiquitous in our nation's institutions, it is imperative to protect HPC systems from security threats [8]. Unlike with desktop computers, the major threat to HPCs (especially multi-user HPC systems) is escalation attacks, which exploit operating system vulnerabilities through acquisition of an administrator's privilege to eventually operate the entire system or damage it [9]. HPC computers have distinct systems, resources, and assets that an attacker could target. Thus, the security needs for HPCs are different from other communication systems [10]. For instance, an analysis of the Centre for Development of Advanced Computing HPC Lab revealed several security vulnerabilities in their system that could have been exploited in an attack [11]. A major flaw called "pam_tally" was intended as a defensive precaution. It functioned by locking users out of the system after too many failed password attempts. In reality, "pam_tally" exposed the system to a denial of service attack.

It emerges as a trend that IoT devices are connected with HPC systems [12]. However, the authentication protocol and middleware that permit safe and secure integration of IoT and HPC are not mature yet. Services like processing, storage, sensing, security, context awareness, and actuating are not working in the most cohesive manner [13]. Moreover, the inevitable integration of the IoT and HPC presents a new security challenge in that a virtual threat could impact a user's physical safety via any internet connected device [13]. Due to limited computational power and storage capacity in IoT devices, preventative security measures should be implemented in the HPC systems to assure the interconnection between the HPC and IoT.

Currently, there exist many software methods to assure the security of HPC systems. Unfortunately, software approaches could be bypassed eventually or lead to new attack surfaces. Moreover, HPC functionality is firmly grounded in hybrid computing that uses hardware accelerators and coprocessors to do parallel processing on a large scale [11]. As hardware is the root of trust, this work focuses on the security threats from the hardware perspective.

III. Security Challenges Due to 3D Integration

A. *3D ICs in High-Performance Computing Systems*

3D ICs play an important role in achieving high-performance computing. The natural advantages of the 3D architecture, including high device density, high bandwidth and low power consumption, fit them perfectly into HPC systems.

1) 3D Architecture for Increasing Memory Density: 3D architecture makes great contribution to tackle the communication bottleneck between memory and computational units in HPC systems [14]. The 3D architecture of DRAM, Hybrid Memory Cube (HMC) [15], integrates multiple DRAM layers plus a logic layer into a stacking memory cube, which significantly increases the memory density. This stacking structure also utilizes through-silicon vias (TSVs) to communicate between memory layers, reducing the system latency and power consumption simultaneously.

2) 3D Architecture for Expanding Bandwidth: In 3D systems, the utilization of TSVs is an effective strategy to expand the memory bandwidth and thus improve the data transmission speed. For instance, the 3D TSV packing technology introduced in [16] integrates over 1200 TSVs into a two-tier 3D structure, achieving a memory bandwidth of 12.8 GB/s.

3) 3D Architecture for Saving Power/Energy: Thanks to a short global wiring length, small chip size, and small pin capacitance, 3D ICs are promising to reduce the total power consumption of HPC systems [17], especially switching power for global interconnections [18]. In the work [19], an Intel Pentium 4 family microprocessor is divided into two dies and stacked together in face-to-face bonding with TSVs. This 3D architecture brings blocks closer in distance so that the inter-block interconnect is reduced, thus power and latency being reduced compared to the traditional planar structure. As reported in [19], both the power consumption and performance are improved by 15%.

B. *Security Threats in 3D Systems*

Despite its benefits on memory density, bandwidth and power consumption, 3D integration results in unique security challenges [20]. Under certain circumstances, it is even more challenging to address the security threats in 3D ICs than in 2D planar chips. Split manufacturing and outsourced fabrication may introduce threats either from untrusted single-die foundries or from untrusted vertical interconnect manufacturers. Untrusted foundries might insert malicious circuitry in 3D chips [21]. Unfortunately, the factors of high device density, limited probing capability and large PVT (power, voltage, and temperature) variation in 3D ICs increase the difficulty of applying functional testing to 3D chips [20]. In addition, poor thermal dissipation in 3D stacking structures may also be exploited by adversary to insert malicious component [21] or purposely accelerate device aging.

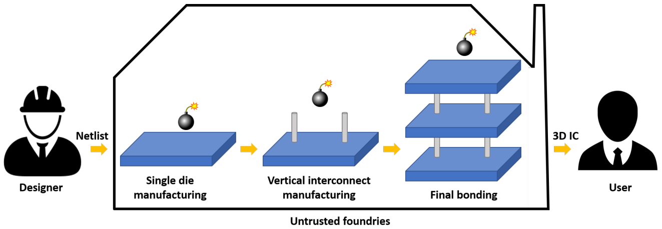


Fig. 1: 3D hardware Trojan insertion by untrusted foundries.

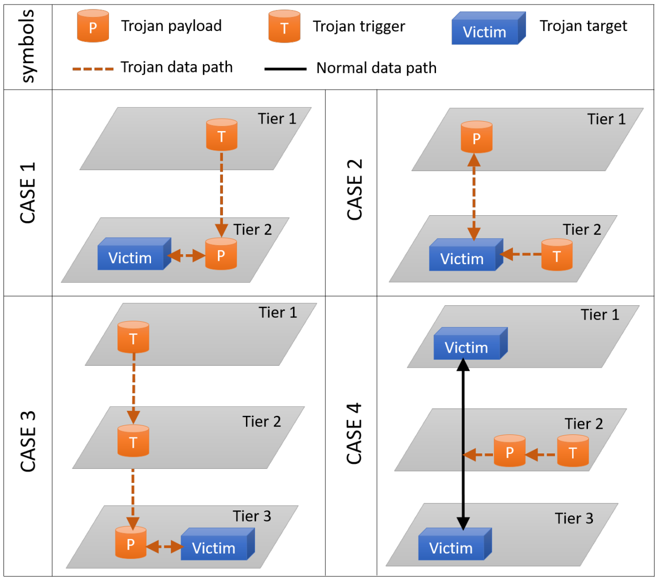


Fig. 2: 3D Trojans models [21].

Hardware Trojans are malicious modifications made on hardware to fulfill attackers' intentions such as sabotaging the original function carried by the target hardware, causing hardware performance degradation, and leaking confidential information embedded in the hardware. The increased number of transistors and the vertical dimension integration in 3D ICs leaves more potentially exploitable space for an attacker to implement hardware Trojans. Furthermore, split manufacturing and outsourced fabrication provide more opportunities for Trojan insertion in the long semiconductor supply chain, as shown in Fig. 1.

New hardware Trojans may also show up in 3D chips. The poor thermal conductivity in 3D chips leads to transition glitches, which could be exploited as Trojan triggers. As reported in the work [22], [23], thermal-triggered Trojans can be inserted by any malicious foundry with access to the layout of the design. Those Trojans are likely to be inserted in the middle tier, where the heat is harder to dissipate than in other tiers [23]. Our recent work [21], [24] envisions that new cross-tier Trojans might occur in the 3D systems. Either the trigger

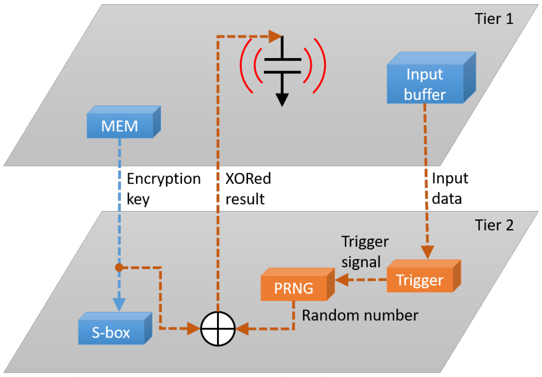


Fig. 3: MOLES Trojan in a 3D system.

circuit and payload circuit are separated and moved to different tiers, or the trigger circuit being split and relocated to multiple tiers jointly activates the payload [24]. Figure 2 summarizes the potential 3D Trojans.

Security threats from hardware Trojans cannot be ignored while we are pursuing better performance in our electronic devices and systems. As we introduced in our previous work [21], a Trojan mounted on 3D chips could possibly alter the original function of the chip or stealthily leak important information. If those chips with Trojan inserted are used in high-performance computing systems, those systems could suffer from more catastrophic effect in a more rapid manner than a personal computer (PC). This is because the HPC systems operate at a much faster speed and serve for more clients than a single PC. A breached storage node in the HPC system could leak a large amount of user data.

The detection of hardware Trojans is more difficult in the 3D environment compared to its 2D counterpart. A 2D Trojan detection approach, Temporal Self-Referencing (TeSR) approach [25], collects the current signatures of two consecutive time windows, in which the victim system runs the same logic transitions. If no Trojan is inserted in the system, the current signatures are identical. In contrast, the triggered Trojan leads to different current signatures. A metric, the Euclidean point-wise distance (EPWD) between two signatures collected from two consecutive time windows, was adopted to evaluate the consistency between the signatures [25]. The EPWD obtained from Trojan-free cased was used as a reference at runtime. If the EPWD for tested target system is greater than the reference, the presence of a hardware Trojan in the target system is detected.

We implemented the TeSR method and MOLES Trojans [26] in our transistor-level 3D IC model, which was built with a 45nm NCSU FreePDK technology [27]. The MOLES Trojan aims at leaking encryption key of crypto modules. It can be implemented in 3D system as shown in Fig 3. The Trojan trigger module monitors the input data of the target and generates a trigger signal, which initializes pseudo random number generator (PRNG). The generated random number is then XORed with the encryption key used in the cypto module (in our case study, it is an AES S-box). Finally, the XORed result will drive a capacitor to charge or discharge, assisting side-channel attacks. According to the experimental results shown in Fig. 4, TeSR fails to detect the MOLES Trojans in 3D system regardless of the Trojan size since the EPWD of the Trojan triggered cases is less than the reference. Next, we changed the size of the victim system (Trojan target) from a single AES S-box to eight S-boxes. As shown in Fig. 5, the TeSR method cannot detect most of the cases. From our case study, we conclude that the existing 2D hardware Trojan detection methods may not be as effective as when they operate in 3D scenarios.

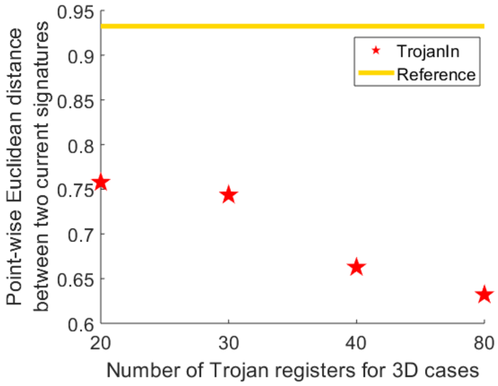


Fig. 4: TeSR effectiveness in 3D environment for different sizes of Trojans.

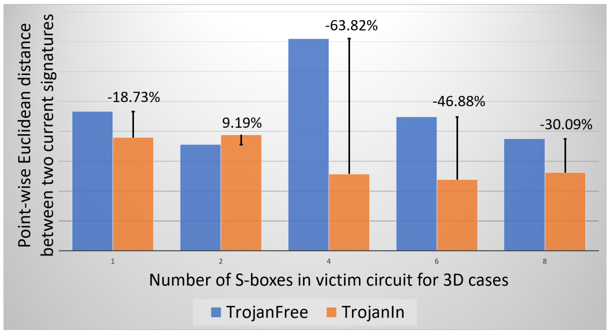


Fig. 5: TeSR effectiveness in 3D environment for different sized victim systems.

Other than hardware Trojans, 3D ICs may also face the challenge from other types of attacks. Many researchers consider 3D integration having natural defense to certain attacks due to their stacking structures. For example, the variation characteristics of 3D ICs can blur the relationship between side-channel signals and the data that attackers try to extract. However, if attackers can focus on the side-channel signal measurement of the target modules and mute other system operations, side-channel attacks are still applicable to 3D ICs. Split manufacturing is a secure mechanism for 3D IC fabrication. However, if the I/O definition and certain specifications of the commercial dies in a 3D stack are public, attackers can still reverse engineer the design and make counterfeit chips accordingly [28].

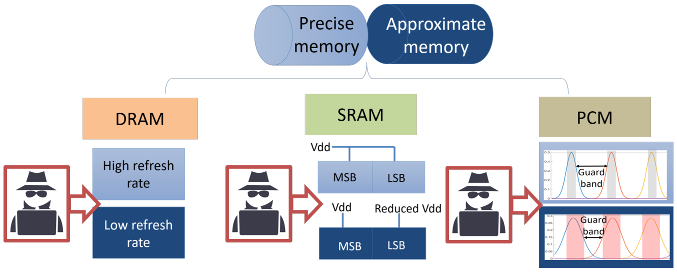


Fig. 6: Approximation strategies used in SRAM, DRAM and PCM.

IV. Security Challenges Due to Approximate Computing

A. Urgent Need for Approximate Computing

According to the International Data Corporation (IDC) [29], the amount of information managed by worldwide datacenters will grow by 50 times, while the number of processors will increase by only 10 times in a decade. The increase in digital data at the predicted rate, will cause a surge in demand for storage units which in turn will require additional resources. The electricity consumption of just U.S. datacenters, which was 61 billion kilo watt hour (kWh) in 2006, will increase to 140 billion kWh in 2020 [30]. Therefore, a solution is needed to accommodate more data with limited resources and power consumption. Approximate computing (AC) has emerged as a promising option to address this need. AC is able to improve energy efficiency at the cost of reduced accuracy [31], [32]. Since applications like image processing, machine learning, and computer vision can tolerate errors during computation or memory storage, approximate computing fits perfectly as a means of reducing power consumption and maintaining system quality [33], [34].

Applications like Recognition, Mining and Synthesis(RMS) requires high computations [35]. To improve the performance of the RMS applications researchers are considering to employ approximate computing techniques. For instance, the adders are replaced with approximate adders [36] and multipliers [37] are replaced with approximate multipliers. However, there are certain security threats imposed by using approximate computing techniques. Some of the security threats imposed by using approximate techniques in HPC systems are discussed in the below sub-section.

B. New Attack Surfaces in Approximate Computing Systems

Approximate computing can be implemented with four different strategies: approximate system, approximate software, approximate storage, and approximate arithmetic circuit. If approximate techniques are adopted in HPC systems, we need to be aware of the potential attacks induced by the use of approximate computing. In the following subsection, we discuss possible attack surfaces in approximate storage, approximate arithmetic circuits, and applications using approximate computing.

1) Memory: Figure 6 shows the approximate computing strategies employed in three types of memory: DRAM, SRAM, and phase-change memory (PCM). In DRAM, the power consumption for memory refresh is almost 50% of the total power consumption [38]. Moreover, write and read operations are prohibited during memory refreshing periods. This fact limits the throughput of DRAM. To improve energy efficiency and throughput, approximate DRAM selectively reduces the refresh rate. The DRAM controller issues the commands through the command bus to indicate if the DRAM memory cells of interest should be refreshed at the regular interval or a reduced rate. If approximate DRAM is deployed in HPC systems, it is critical to protect the memory refresh controller. Otherwise, once the adversary has control over the command bus and manipulates the refresh logic command, he/she could reduce the refresh rate for the precise DRAM cells to induce memory errors. Attacks on the hybrid precise/approximate DRAM will sabotage the memory integrity, thus harming the computation-intensive HPC applications.

In SRAM, approximation storage is achieved by reducing the supply voltage for the memory cells storing the least significant bits (LSB). An adversary with control of the voltage regulator could maliciously reduce the supply voltage of the approximate SRAM memory blocks, tampering with the stored data. The compromised SRAM blocks will lead the HPC system to experience some unexpected failures more often than usual, which will cause catastrophic consequences on the HPC users.

PCM is a non-volatile memory that is commonly used as a multi-level cell. It has great potential to be used in HPC systems storage. Approximation in PCM is obtained by reducing the guard band between digital levels. From Fig. 6 we can see that, the guard band for approximate PCM is narrower than that of precise PCM. Although approximate PCM is outfitted with advantages like a faster read and write speed and lower power consumption, the approximation mechanisms used in PCM could be exploited to develop new attack surfaces. We introduced possible attack scenarios in the work [39]. For instance, the definition of guard band could be altered such that the number of writing iterations for different logic levels is changed accordingly. The analog-to-digital level converter is prone to attack as well.

2) Arithmetic Circuit: Adders are the basic building blocks of arithmetic circuits. The power consumption and critical-path delay due to the carry-bit calculation is typically prominent in an adder. Since precise calculations are not necessary for all applications, approximation techniques employed in arithmetic circuits could help in achieving energy efficiency. For example, the use of inexact adders with acceptable accuracy loss in computation-intensive applications like machine learning can reduce the delay by 18.79% and area by 31.44% compared to using the precise arithmetic circuits [36].

Arithmetic circuits use approximation techniques like logic minimization in which the logical function is re-arranged such that the implementation requires the minimal number of logic gates. For example, in work [40], the truth table of the adder

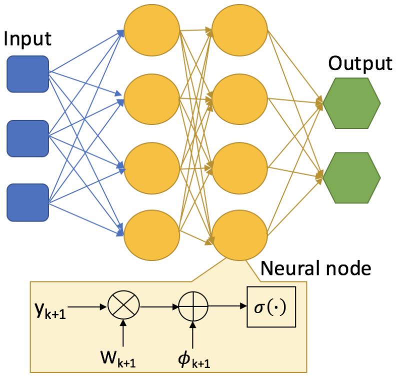


Fig. 7: General architecture for neural network computation.

is altered so that the sum and carry logic of a full adder is implemented with minimal logic gates. Other approximate techniques used in arithmetic circuits include ignoring the carry propagation logic for the LSB bits or separating the carry propagation logic for MSB and LSB bits.

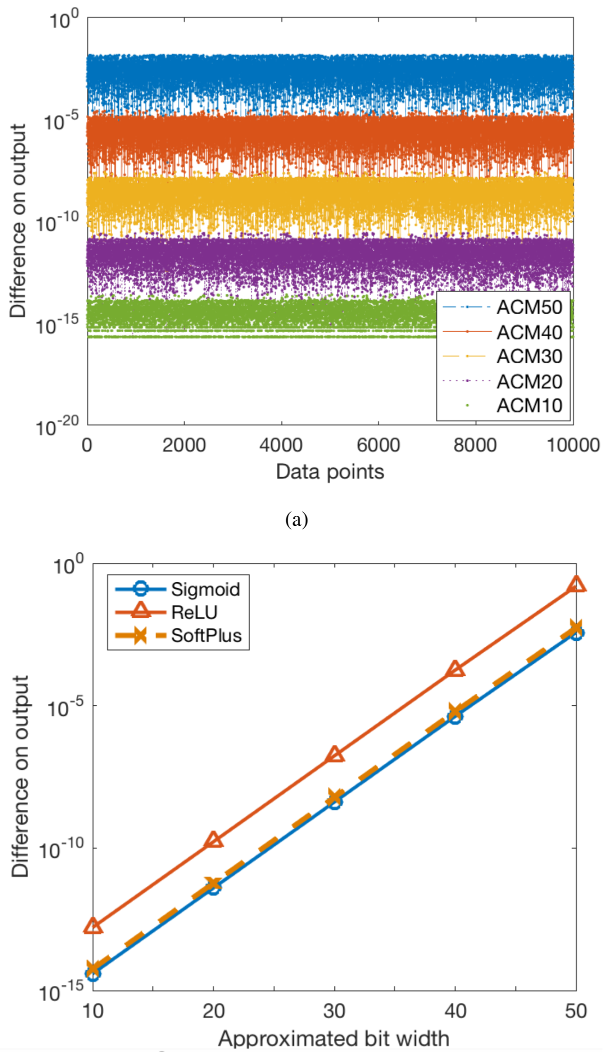
Approximate computing techniques employed in arithmetic circuits help in gaining energy efficiency. However, these approximate circuits could draw security threats. If approximate arithmetic circuits are used in HPC systems, the security threats possessed by arithmetic circuits could breach confidential information stored in HPC systems. For instance, the adversary could use the inaccuracy generated in the adder output to hide some malicious information within the inaccurate part. Since, the data in the inaccurate portion is not particularly important any changes could be easily bypassed during Trojan detection. An attacker could later use the data stored in the inaccurate portion to trigger the Trojan and thus damage or alter the functionality of the system.

3) Application: Recognition, mining, and synthesis (RMS) applications are considered to be emerging high performance and computation-intensive applications [35]. Approximate computing can be employed in RMS applications because they are inherently error tolerant as most of the inputs to these applications coming from the sensors, which often contain noise. Moreover, the output of RMS applications do not need to have high precision because humans have limited perceptual capabilities [36]. Artificial neural networks(ANN) are one of the most widely used machine learning techniques for RMS applications. In ANN, approximate computing is employed using approximate adders or multipliers [36], [37]. Other approximate techniques employed in ANN include memory access skipping [37] and precision scaling [41].

We used a general feedforward deep neural network as a case study to show the impact of manipulated approximate computing on the application output. Figure 7 shows the general architecture for a neural network. In the neural network, the function for the kth hidden layer can be represented using Eq.(1) [42].

$$y_{k+1} = \sigma_{k+1}\left(W_{k+1} * y_k + \phi_{k+1}\right) \quad (1)$$

The weight matrix W and bias co-efficiency ϕ are adjusted to train the network which adopts the best fit parameters to emulate the desired function. The common activation functions $\sigma()$ are the logistic function (a.k.a Sigmoid), Rectified linear unit function (ReLU) and SoftPlus. The approximate version of multiplication and addition facilitates the preservation of computation power for the neural network. However, the inexact arithmetic circuit will lead to network performance degradation.



(a)

(b)

Fig. 8: Impact of approximate computing in ANN.

We constructed the basic arithmetic circuit to implement the function expressed in Eq.(1). A series of random input, 10000 data points, was fed to the network. The uniformly distributed random weight and bias were applied to the network, as well. We manipulated the precision of multiplication or addition and compared the output of the trained network. The number of muted mantissa varies from 10 bits (i.e. ACM10) to 50 bits (i.e. ACM50) for double precision floating point numbers. As shown in Fig. 8(a), the difference in output increases with more ignored mantissa bits. We also examined the impact of different activation functions used in the neural network.

Figure 8(b) indicates that no matter which activation function is used, the difference on output due to approximate computing cannot be ignored when the number of approximate bits increases.

HPC systems are moving towards adopting AC techniques to improve performance and energy efficiency simultaneously. However, as indicated in the experimental results shown in Fig 8, the use of approximate computing will lead the system to be vulnerable to various security threats. Thus, while we employ AC techniques, security measures are needed to strengthen HPC systems against security threats analyzed above.

V. Conclusion

High performance is the primary focus of HPC system designers and users. However, since hardware for HPC systems suffers from supply chain attacks, it is imperative to investigate the security challenges on HPC systems. In particular, this work introduces the security threats from the hardware perspective. We first introduce the new hardware Trojans that could be implemented in the 3D-IC based HPC systems. The existing Trojan detection method for 2D systems are prone to fail in 3D systems now due to increased noise. We also envision that the use of approximate computing in HPC systems will lead to new attack surfaces. We expect this work will inspire researchers to develop effective countermeasures to improve the resilience of HPC systems against security threats on the hardware components used in HPC.

Acknowledgment

This project is partially supported by NSF CAREER Award #1652474 and NSF/SRC STARSS Award #1717130.

References

[1] "US Plans $1.8 Billion Spend on DOE Exascale Supercomputing." https://www.energy.gov/downloads/fact-sheet-collaboration-oak-ridge-argonne-and-livermore-coral. Accessed: 2019-05-22.

[2] S. Dewen and C. Wenlan, "Application of HPC technology in the building of a virtual geological visualization system," in *2010 2nd International Conference on Future Computer and Communication*, vol. 1, pp. V1–472–V1–476, May 2010.

[3] A. Prout, W. Arcand, D. Bestor, C. Byun, B. Bergeron, M. Hubbell, J. Kepner, P. Michaleas, J. Mullen, A. Reuther, and A. Rosa, "Scalable cryptographic authentication for high performance computing," in *2012 IEEE Conference on High Performance Extreme Computing*, pp. 1–2, Sep. 2012.

[4] J. Dongarra, S. Tomov, P. Luszczek, J. Kurzak, M. Gates, I. Yamazaki, H. Anzt, A. Haidar, and A. Abdelfattah, "With Extreme Computing, the Rules Have Changed," *Computing in Science Engineering*, vol. 19, pp. 52–62, May 2017.

[5] S. Gesing, J. Nabrzyski, and S. Jha, "Gateways to high-performance and distributed computing resources for global health challenges," in *2014 IEEE Canada International Humanitarian Technology Conference - (IHTC)*, pp. 1–5, June 2014.

[6] A. Wallqvist, N. Zavaljevski, R. Vijaya Satya, R. Bondugula, V. Desai, X. Hu, K. Kumar, M. S. Lee, I. Yeh, C. Yu, and J. Reifman, "Accelerating Biomedical Research in Designing Diagnostic Assays, Drugs, and Vaccines," *Computing in Science Engineering*, vol. 12, pp. 46–55, Sep. 2010.

[7] S. Hamdioui, J. Danger, G. Di Natale, F. Smailbegovic, G. van Battum, and M. Tehranipoor, "Hacking and protecting IC hardware," in *2014 Design, Automation Test in Europe Conference Exhibition (DATE)*, pp. 1–7, March 2014.

[8] A. Malin and G. Van Heule, "Continuous Monitoring and Cyber Security for High Performance Computing," in *Proceedings of the First Workshop on Changing Landscapes in HPC Security*, CLHS '13, (New York, NY, USA), pp. 9–14, ACM, 2013.

[9] T. Yamauchi, Y. Akao, R. Yoshitani, Y. Nakamura, and M. Hashimoto, "Additional Kernel Observer to Prevent Privilege Escalation Attacks by Focusing on System Call Privilege Changes," in *2018 IEEE Conference on Dependable and Secure Computing (DSC)*, pp. 1–8, Dec 2018.

[10] S. Peisert, "Security in High-performance Computing Environments," *Commun. ACM*, vol. 60, pp. 72–80, Aug. 2017.

[11] R. Bulusu, P. Jain, P. Pawar, M. Afzal, and S. Wandhekar, "Addressing security aspects for HPC infrastructure," in *2018 International Conference on Information and Computer Technologies (ICICT)*, pp. 27–30, March 2018.

[12] D. E. Baz, "IoT and the Need for High Performance Computing," in *2014 International Conference on Identification, Information and Knowledge in the Internet of Things*, pp. 1–6, Oct 2014.

[13] L. de Souza Cimino, J. E. E. d. Resende, L. H. M. Silva, S. Q. S. Rocha, M. de Oliveira Correia, G. S. Monteiro, G. N. de Souza Fernandes, S. G. M. Almeida, A. L. B. Almeida, A. L. L. de Aquino, and J. de Castro Lima, "IoT and HPC Integration: Revision and Perspectives," in *2017 VII Brazilian Symposium on Computing Systems Engineering (SBESC)*, pp. 132–139, Nov 2017.

[14] R. R. Tummala, "3D system package architecture as alternative to 3D stacking of ICs with TSV at system level," in *2017 IEEE International Electron Devices Meeting (IEDM)*, pp. 3.4.1–3.4.3, Dec 2017.

[15] J. Jeddeloh and B. Keeth, "Hybrid memory cube new DRAM architecture increases density and performance," in *2012 Symposium on VLSI Technology (VLSIT)*, pp. 87–88, June 2012.

[16] D. J. Na, K. O. Aung, W. K. Choi, T. Kida, T. Ochiai, T. Hashimoto, M. Kimura, K. Kata, S. W. Yoon, and A. C. B. Yong, "TSV MEOL (mid end of line) and packaging technology of mobile 3D-IC stacking," in *2014 IEEE 64th Electronic Components and Technology Conference (ECTC)*, pp. 596–600, May 2014.

[17] T. Tanaka, "3D-IC technology and reliability challenges," in *2017 17th International Workshop on Junction Technology (IWJT)*, pp. 51–53, June 2017.

[18] T. Fukushima, H. Kikuchi, Y. Yamada, T. Konno, J. Liang, K. Sasaki, K. Inamura, T. Tanaka, and M. Koyanagi, "New Three-Dimensional Integration Technology Based on Reconfigured Wafer-on-Wafer Bonding Technique," in *2007 IEEE International Electron Devices Meeting*, pp. 985–988, Dec 2007.

[19] B. Black, M. Annavaram, N. Brekelbaum, J. DeVale, L. Jiang, G. H. Loh, D. McCaule, P. Morrow, D. W. Nelson, D. Pantuso, P. Reed, J. Rupley, S. Shankar, J. Shen, and C. Webb, "Die Stacking (3D) Microarchitecture," in *2006 39th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'06)*, pp. 469–479, Dec 2006.

[20] J. Dofe, Q. Yu, H. Wang, and E. Salman, "Hardware security threats and potential countermeasures in emerging 3D ICs," in *2016 International Great Lakes Symposium on VLSI (GLSVLSI)*, pp. 69–74, May 2016.

[21] Z. Zhang and Q. Yu, "Modeling Hardware Trojans in 3D ICs," in *in Proc. ISVLSI'19*, pp. 483–488, July 2019.

[22] S. F. Mossa, S. R. Hasan, and O. Elkeelany, "Hardware trojans in 3D ICs due to NBTI effects and countermeasure," *Integration*, vol. 59, pp. 64–74, 2017.

[23] S. R. Hasan, S. F. Mossa, O. S. A. Elkeelany, and F. Awwad, "Tenacious hardware trojans due to high temperature in middle tiers of 3-D ICs," in *2015 IEEE 58th International Midwest Symposium on Circuits and Systems (MWSCAS)*, pp. 1–4, Aug 2015.

[24] J. Dofe, P. Gu, D. Stow, Q. Yu, E. Kursun, and Y. Xie, "Security threats and countermeasures in three-dimensional integrated circuits," in *Proceedings of the on Great Lakes Symposium on VLSI 2017*, pp. 321–326, ACM, 2017.

[25] S. Narasimhan, X. Wang, D. Du, R. S. Chakraborty, and S. Bhunia, "TeSR: A robust Temporal Self-Referencing approach for Hardware Trojan detection," in *2011 IEEE International Symposium on Hardware-Oriented Security and Trust*, pp. 71–74, June 2011.

[26] L. Lin, W. Burleson, and C. Paar, "MOLES: Malicious off-chip leakage enabled by side-channels," in *2009 IEEE/ACM International Conference on Computer-Aided Design - Digest of Technical Papers*, pp. 117–122, Nov 2009.

[27] S. M. Satheesh and E. Salman, "Power Distribution in TSV-Based 3-D Processor-Memory Stacks," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 2, pp. 692–703, Dec 2012.

[28] J. Dofe, Q. Yu, H. Wang, and E. Salman, “Hardware security threats and potential countermeasures in emerging 3d ics,” in *2016 International Great Lakes Symposium on VLSI (GLSVLSI)*, pp. 69–74, May 2016.

[29] “The Digitization of the World From Edge to Core.” https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf.

[30] “Energy Aware Virtual Machine Scheduling in Data Centers.” https://www.mdpi.com/1996-1073/12/4/646.

[31] A. K. Mishra, R. Barik, and S. Paul, “iACT: A Software-Hardware Framework for Understanding the Scope of Approximate Computing,” 2014.

[32] V. K. Chippa, S. T. Chakradhar, K. Roy, and A. Raghunathan, “Analysis and characterization of inherent application resilience for approximate computing,” in *2013 50th ACM/EDAC/IEEE Design Automation Conference (DAC)*, 2013.

[33] D. T. Nguyen, H. Kim, H. Lee, and I. Chang, “An Approximate Memory Architecture for a Reduction of Refresh Power Consumption in Deep Learning Applications,” in *Proc. 2018 ISCAS*, pp. 1–5, May 2018.

[34] F. Qiao, N. Zhou, Y. Chen, and H. Yang, “Approximate Computing in Chrominance Cache for Image/Video Processing,” in *Proc. 2015 IEEE Intl. Conf. on Multimedia Big Data*, pp. 180–183, April 2015.

[35] P. Dubey, “Recognition, mining and synthesis moves computers to the era of tera,” *Technology@ Intel Magazine*, vol. 9, no. 2, pp. 1–10, 2005.

[36] Z. Du, K. Palem, A. Lingamneni, O. Temam, Y. Chen, and C. Wu, “Leveraging the error resilience of machine-learning applications for designing highly energy efficient accelerators,” in *2014 19th Asia and South Pacific Design Automation Conference (ASP-DAC)*, pp. 201–206, Jan 2014.

[37] Q. Zhang, T. Wang, Y. Tian, F. Yuan, and Q. Xu, “ApproxANN: An approximate computing framework for artificial neural network,” in *2015 Design, Automation Test in Europe Conference Exhibition (DATE)*, pp. 701–706, March 2015.

[38] I. Bhati, Z. Chishti, S. Lu, and B. Jacob, “Flexible auto-refresh: Enabling scalable and energy-efficient DRAM refresh reductions,” in *2015 ACM/IEEE 42nd Annual International Symposium on Computer Architecture (ISCA)*, pp. 235–246, June 2015.

[39] P. Yellu, N. Boskov, M. Kinsy, and Q. Yu, “Security Threats on Approximate Computing Systems,” in *The ACM Great Lakes Symposium on VLSI (GLSVLSI), 2019*, pp. 387–392, May 2019.

[40] V. Gupta, D. Mohapatra, A. Raghunathan, and K. Roy, “Low-Power Digital Signal Processing Using Approximate Adders,” *IEEE Trans. on Computer-Aided Design of Integr. Circuits and Syst.*, vol. 32, pp. 124–137, Jan 2013.

[41] Y. Tian, Q. Zhang, T. Wang, F. Yuan, and Q. Xu, “ApproxMA: Approximate Memory Access for Dynamic Precision Scaling,” in *ACM Great Lakes Symposium on VLSI*, pp. 337–342, 2015.

[42] K. Hwang and W. Sung, “Fixed-point feedforward deep neural network design using weights +1, 0, and 1,” in *2014 IEEE Workshop on Signal Processing Systems (SiPS)*, pp. 1–6, Oct 2014.