High Bandwidth Thermal Covert Channel in 3-D-Integrated Multicore Processors

Krithika Dhananjay, *Member, IEEE*, Vasilis F. Pavlidis, *Senior Member, IEEE*, Ayse K. Coskun, *Senior Member, IEEE*, and Emre Salman, *Senior Member, IEEE*

***Abstract*— Exploiting thermal coupling among the cores of a processor to secretly communicate sensitive information is a serious threat in mobile, desktop, and server platforms. Existing works on temperature-based covert communication typically rely on controlling the execution of high-power CPU stressing programs to transmit confidential information. Such covert channels with high-power programs are typically easier to detect as they cause significant rise in temperature. In this work, we demonstrate that by leveraging vertical integration, it is sufficient to execute typical SPLASH-2 benchmark applications to transfer 200 bits per second (bps) of secret data via thermal covert channels. The strong vertical thermal coupling among the cores of a 3-D multicore processor increases the rates of covert communication by 3.4× compared to covert communication in conventional 2-D integrated circuits (ICs). Furthermore, we show that the bandwidth of this thermal communication in 3-D ICs is more resilient to thermal interference caused by applications running in other cores. This reduced interference significantly increases the danger posed by such attacks. We also investigate the effect of reducing intertier overlap between colluded cores and show that the covert channel bandwidth is reduced by up to 62% with no overlap.**

***Index Terms*— Computer security, multicore processors, side-channel attacks, thermal management of electronics, three-dimensional integrated circuits, through-silicon vias (TSVs).**

I. Introduction

STRINGENT security protocols have been implemented in modern microprocessors to mitigate various vulnerabilities [1]. However, malicious attackers find ways to bypass these security measures and retrieve confidential information. Covert communication channel is one such attack that transfers data between two entities that are not authorized to communicate [2]. For example, in timing-based covert channel attacks, the transmitter delays the network transmission time to encode secret data. Alternatively, in storage-based covert channel attacks, sensitive information is communicated to the receiver by modifying a shared storage resource directly, such as a hard drive location or unused packets in headers [3], [4] or indirectly such as temperature, light, and sound [5], [6], [7].

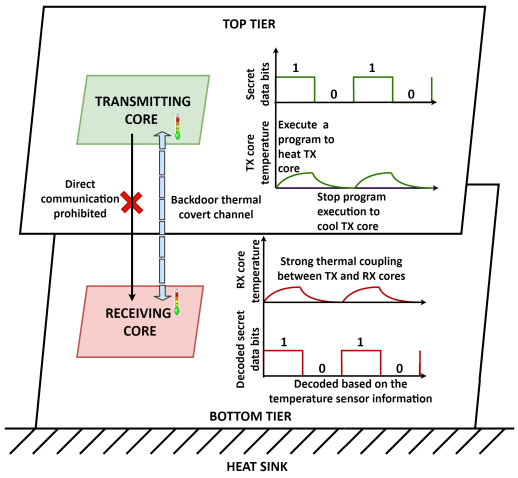

Fig. 1. Block diagram of TCC in 3-D multicore processors.

Recently, temperature-based covert channel communication has gained attention, where an adversary uses heat to communicate sensitive data between two unauthorized compute elements [5], [8]. For example, in a multicore processor, the thermal covert channel communication (TCC) is established between two cores of the processor by encoding sensitive information within the temperature profile of the transmitting core [5]. Specifically, an attacker application transmits a bit "1" by executing a program in the transmitting core to raise its temperature. In order to transmit a bit "0," the attacker stops program execution to lower the temperature of the transmitting core, as shown in Fig. 1. Due to thermal coupling among the cores, an application in a receiving core can retrieve the information by reading its temperature sensor. In a majority of the modern commercial processors, information provided by thermal sensors is accessible to user applications [9].

Several studies have been published during the past decade about TCC modeling [8], [10], detection [11], [12], and countermeasures [11], [12], [13] in multicore processors. Long *et al.* [14] and Huang *et al.* [11] demonstrated that a successful TCC with low error rates can be established, provided that the transmission frequency of the channel, i.e., TCC rate, is higher than the frequency band of power consumption of the other active cores. In conventional 2-D integrated

Manuscript received 21 February 2022; revised 13 June 2022 and 27 July 2022; accepted 18 August 2022. Date of publication 20 September 2022; date of current version 24 October 2022. This work was supported in part by the National Science Foundation under Grant CCF 1910075/1909027. *(Corresponding author: Krithika Dhananjay.)*

Krithika Dhananjay and Emre Salman are with the Department of Electrical and Computer Engineering, Stony Brook University (SUNY), Stony Brook, NY 11794 USA (e-mail: krithika.yethiraj@stonybrook.edu).

Vasilis F. Pavlidis is with the Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece

Ayse K. Coskun is with the Department of Electrical and Computer Engineering, Boston University, Boston, MA 02215 USA.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TVLSI.2022.3203430.

Digital Object Identifier 10.1109/TVLSI.2022.3203430

circuits (ICs), high TCC rates can be achieved by executing high-power programs (such as CPU stress tests). Thus, a majority of the existing works rely on high-power programs to sufficiently raise the temperature of the transmitting core, thereby reducing the error rates of covert communication [5], [8], [11], [14]. These programs, however, are likely to cause overheating, thus enabling the attack to be detected relatively easily. Furthermore, in these works, the fan used for cooling the cores works at the maximum capacity to mitigate the issue of overheating [5], [11]. In this work, we demonstrate that a high bandwidth TCC can be established with relatively low-power benchmark applications by leveraging vertical integration technologies, such as through-silicon via (TSV)-based die stacking [15] and monolithic 3-D (Mono3D) integration [16]. Unlike TCC in conventional 2-D integration, where heat flows between the cores of a processor in lateral fashion, the close proximity of tiers in 3-D ICs increases the vertical thermal coupling among the intertier functional blocks. Thus, TCC attacks are potentially more dangerous in 3-D multicore systems because larger blocks of sensitive data can be communicated at faster rates. The key findings from this work are given as follows.

1) We show that TCC established in 3-D processors can achieve negligible error rates (<1%) with transmission rates of 200 bits per second (bps) by executing commonly used SPLASH-2 benchmark applications. Therefore, the average power consumed during the attack is significantly reduced, making the attack more difficult to detect.
2) We characterize the bandwidth and bit error rates (BERs) of TCC among the cores of a flip-chip two-tier Mono3D and TSV3D (face-to-back bonding) integrated processor. The TCC bandwidth in Mono3D and TSV3D processors is, respectively, 3.4× and 4× greater than the TCC bandwidth in a conventional 2-D processor.
3) We investigate the effect of thermal interference from applications running on other cores. We observe that the TCC bandwidth in 3-D processors remains the same in the presence of interference from one of the cores. Alternatively, for a 2-D integrated processor, the bandwidth degrades by 12% with a minimum achievable error rate of 3%. Therefore, TCC in 2-D processors is highly sensitive to the heat generated by other active cores, whereas TCC in 3-D processors is comparatively more robust.
4) The thermal coupling between cores in 3-D integrated processors can be reduced by decreasing the amount of overlap between the core with access to sensitive information and other insecure cores. Therefore, we study the effect of having 50% and 0% overlap between transmitting and receiving cores. The maximum degradation in the TCC bandwidth for nonoverlapping cores is 62% for Mono3D and 58% for TSV3D processors, compared to the TCC bandwidth between fully overlapping cores.
5) We also study the effects of varying the distance between the transmitting core and the heat sink, the variations in the transient power profile of the TCC program, and having the transmitting and receiving cores on nonadjacent tiers of a 3-D system. For example, the TCC bandwidth increases by approximately 10% when the transmitting core is placed closer to the heat sink and the receiving core is above the transmitting core. A TCC program with a more stable power profile increases the TCC bandwidth for both 2-D and 3-D technologies, even though the average and peak power consumption is less. If the transmitting and receiving cores are on nonadjacent tiers, the TCC bandwidth in the Mono3D processor remains the same, whereas for TSV3D processor, the TCC bandwidth increases due to elevated temperature levels.

The rest of this article is organized as follows. Existing works on TCC are summarized in Section II. The attack model, methodology for TCC analysis and characterization framework, design models, and covert communication protocol are detailed in Section III. The results for both 2-D and 3-D systems are described in Section IV. This article is concluded in Section V.

II. Related Work

TCC has been identified as a significant threat for several hardware platforms such as cloud-based field-programmable gate arrays (FPGAs) [17], [18], the Internet of Things (IoT) devices [19], desktop [8], [11], [12], mobile [8], [20], and server processors [5]. Masti *et al.* [5] showed that the accessibility of temperature sensors in modern processors enables thermal covert communication between its cores. A maximum bandwidth of 1.33 bps with 11% BER was demonstrated via measurement results for an Intel Xeon processor [5]. Bartolini *et al.* [8] proposed an enhanced communication scheme that uses Manchester encoding and bitwise decoding with naive Bayes classifier. They demonstrated via measurements that a TCC bandwidth of 5 bps with less than 1% BER can be achieved between neighboring cores. They also proposed spectral techniques to characterize the maximum capacity of thermal covert channels for mobile and laptop platforms. Both of these works assume that cores other than the transmitter and receiver are idle to minimize thermal interference. Long *et al.* [14] considered the thermal interference from other cores and showed that the BER can be reduced by 75% and the transmission rate can be increased by 370% via two techniques: 1) by selecting a higher TCC transmission frequency than the frequency of the power consumption caused by applications running in other cores and 2) by adopting a return-to-zero encoding scheme.

Several works focused on detection and mitigation of TCC. Huang *et al.* [11], [12] proposed techniques for thermal covert channel detection based on scanning instructions per cycle (IPCs) of each processor core and the frequency spectrum of temperature profiles, respectively. Furthermore, Huang *et al.* [12] also proposed countermeasures based on dynamic voltage and frequency scaling (DVFS) to mitigate an active TCC attack. Wang *et al.* [21] proposed a channel-aware noise jamming technique to mitigate a TCC that dynamically changes its transmission frequency. Wang *et al.* [10] developed analytic models to efficiently determine the critical TCC parameters. A majority of these works leverage

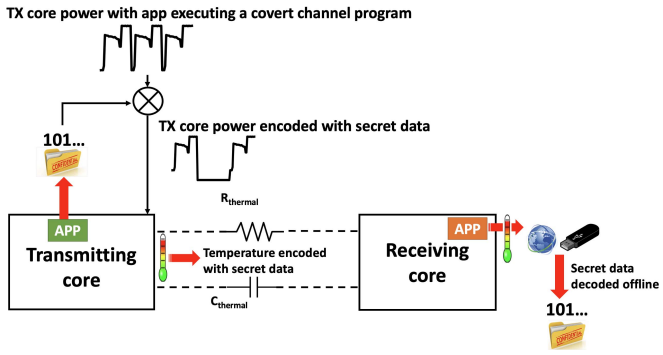

Fig. 2. Attack model of TCC between two cores of a multicore processor.

the lateral thermal coupling in 2-D technologies to establish TCC. For 3-D ICs, thermal side-channel (rather than covert channel) attacks have been studied in the past [22], [23]. Furthermore, Chen *et al.* [24] exploited the close proximity of system-on-chip (SoC) to the DRAM chip (fabricated using the package-on-package technology) to transmit secret information using heat. This temperature-based communication was achieved by generating heat patterns in one core of the SoC and indirectly decoding them at another core by measuring the decay rate of the DRAM cells. Finally, Huang *et al.* [11] presented TCC detection and DVFS-based countermeasure techniques for 2-D and TSV-based 3-D integrated processors.

To the best of the authors' knowledge, all of the previous works have established TCC with computationally intensive programs, which is relatively easier to detect. In this work, we demonstrate that by leveraging 3-D technologies, a moderate power SPLASH-2 benchmark application can be sufficient to establish a TCC attack with significant communication bandwidth. Unlike previous works, a nonreturn-to-zero (NRZ) encoding is used in this work since the temperature of the transmitting and receiving cores does not significantly rise (due to executing relatively low-power applications). Thus, NRZ is adopted to increase the bandwidth of thermal covert communication. We quantify the bandwidth and BER of TCC in both Mono3D and TSV-based 3-D processors. The robustness of TCC in the presence of thermal interference from applications running in other cores is also investigated for both 2-D and 3-D systems. Finally, we demonstrate that a TCC attack in 3-D processors can be mitigated by reducing the vertical overlap between the transmitting and receiving cores.

III. Methodology

The TCC attack model and analysis framework are described in this section. The processor architecture, 3-D floorplans, layer stack and thermal models, covert channel application, and communication protocol are also detailed.

A. Attack Model

In this work, we consider that TCC is established between two physical cores of a multicore processor that execute compromised software applications (henceforth, referred to as apps) concurrently, as shown in Fig. 2. Let us assume that the app executed on the transmitting core has access to confidential information. Some examples include a contact app on a mobile phone that has access to private list of contacts or personal finance management apps with access to confidential monetary data. In modern multicore processors, data packets from these secure applications can be protected by special enclaves using technologies such as Intel Software Guard Extensions (SGX) [25] and Arm TrustZone [26]. These technologies prevent sensitive data managed by these apps from being accessed by outside world. However, thermal coupling between the physical cores can be leveraged to leak sensitive data by bypassing these security measures [5], [11]. In order to achieve this covert channel, the transmitting app controls the execution of a program to raise and lower the power consumption of the transmitting core. A sample power profile of the transmitting core executing a covert channel program is shown in Fig. 2. Consequently, the temperature of the transmitting core is encoded with the sensitive information and is coupled to the receiving core through the thermal resistance (R_{thermal}) and capacitance (C_{thermal}) of the medium, as shown in the figure.

We assume that the app on the receiving core is not security enforced and hence does not have direct access to any confidential information. However, for this app to read the temperature profile of the receiving core, we assume that it has access to the temperature sensor of the core. This assumption is based on commonly used thermal management policies, where the user-installed apps can access temperature sensor data without special permissions [8]. The app either decodes the data bits on the receiving core or sends the temperature data for offline decoding, as shown in Fig. 2.

B. TCC Analysis Framework

The simulation framework of TCC for both 2-D and 3-D ICs is shown in Fig. 3. Each step of this flowchart is described in this section.

1) Processor Architecture: Our target system is a four-core CPU based on Intel Haswell architecture [27]. The 22-nm processor operates at a 3.4-GHz frequency with a supply voltage of 1.2 V. The specific architectural configurations, such as performance models of each core, L1, L2, and L3 caches, translation buffer, and reorder buffers, are adapted from published data for the processor, as listed in Table I [28], [29]. The workloads used for TCC are simulated on this architecture using SNIPER [30], which is an interval-based timing simulator designed specifically for multicore Intel processors. The transient power traces of the covert channel applications are obtained via the multicore power simulator, McPAT [31], which is integrated within SNIPER. The power consumption obtained from McPAT is calibrated with the real power measurements of a similar processor architecture [32].

2) Encoding the Secret Data: The transmitting core encodes the secret data bits in its transient temperature profile before communicating to the receiver core. In this work, the encoding is performed using the NRZ encoding technique. To transmit a bit "1," the compromised app on the transmitting core continuously executes a program to increase the power consumption

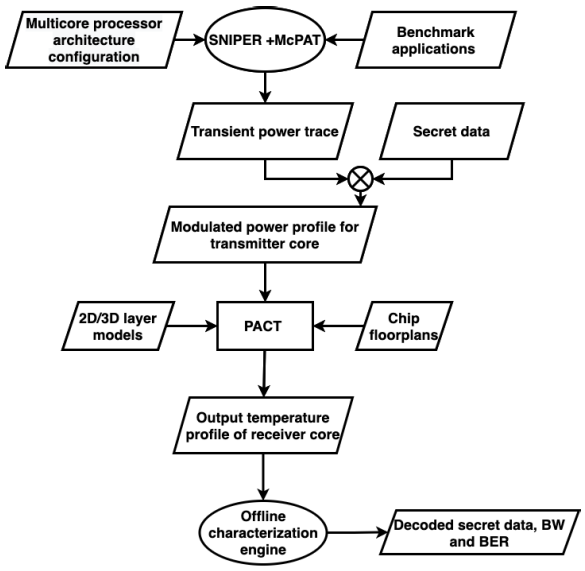

Fig. 3. Flowchart showing the simulation framework for characterizing TCC in 2-D and 3-D multicore processors.

TABLE I
ARCHITECTURAL CONFIGURATIONS OF FOUR-CORE HASWELL PROCESSOR

Instruction set architecture	x86-64
Clock frequency	3.4 GHz
Technology node	22 nm
Supply voltage	1.2 V
Issue width	4
Reorder buffer entries	192 entries
TLB entries	ITLB: 128, DTLB: 64, STLB: 1024
L1 cache	32 KB 8-way set associative L1I cache and L1D cache
L2 cache	256 KB 8-way set associative
L3 cache	8 MB shared

Algorithm 1 Generation of Modulated Power Trace Encoded With Secret Data

$number\_of\_ones \leftarrow 0$
$next\_bit$:
for bit in secret data **do**
  if $bit == 1$ **then**
    if $number\_of\_ones == 0$ **then**
      run program for bit-width duration
      update power trace
    else
      continue program execution for bit-width duration
      update power trace
    end if
    $number\_of\_ones \leftarrow number\_of\_ones + 1$
    goto $nextbit$
  else
    $number\_of\_ones \leftarrow 0$
    stop program if executing
    update power trace
    goto $next\_bit$
  end if
end for

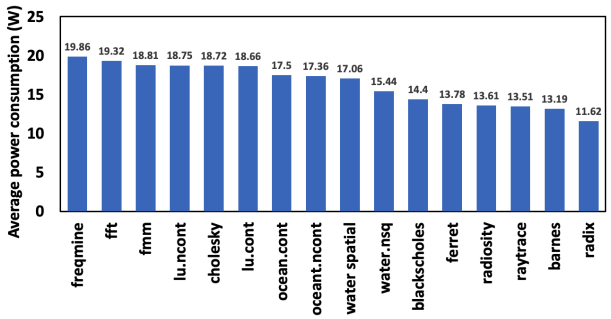

Fig. 4. Average power consumption of Haswell four-core processor running different applications from SPLASH-2 and PARSEC benchmark suites.

(and, hence, the temperature) of that core until the next bit to be transmitted is "0." Similar to transmit bit "0," the app stops the program execution and remains idle such that the temperature of the core decreases. The pseudocode of this encoding process is shown in Algorithm 1. We define the bit-width in the pseudocode as the duration during which a bit "1" or "0" is transmitted.

To transmit bit "1," we execute programs from the common SPLASH-2 and PARSEC benchmark application suites [33]. The applications *freqmine*, *ferret*, and *blackscholes* are from PARSEC and the other applications are from SPLASH-2. The total simulated power (obtained from McPAT) consumed by the Haswell processor executing these single-threaded applications ranges from 11 to 20 W, as shown in Fig. 4. From this figure, it can be observed that *FFT* from SPLASH-2 and *freqmine* from PARSEC consume the highest power and, therefore, are capable of producing relatively significant variations in the temperature profile of the transmitting core. Therefore, *FFT* is chosen as the target application program and is continuously executed in a single-threaded fashion for a bit "1" until the following bit is "0." The transient power trace of the transmitting core executing the *FFT* program is obtained from McPAT with a time step of 0.5 ms. To synchronize the start of the communication, the app in the transmitting core prefixes the beginning of every secret block of data with preamble bits. The preamble consists of a sequence of alternate bits of ones and zeros. This pattern is used since it ensures a symmetric and well-correlated temperature profile between the two communicating cores [5]. Sample secret bits prefixed with preamble bits are plotted in Fig. 5 along with the modulated transmitting core power profile generated based on Algorithm 1.

In previous works, other encoding techniques, such as return-to-zero amplitude-shift keying (ASK) or Manchester encoding, have been explored, even though the communication bandwidth is reduced compared to the NRZ encoding technique [8], [14]. The primary reason for using these alternative encoding schemes is that, when encoding a continuous stream of bit "1"s in the NRZ technique, the CPU stress test has to

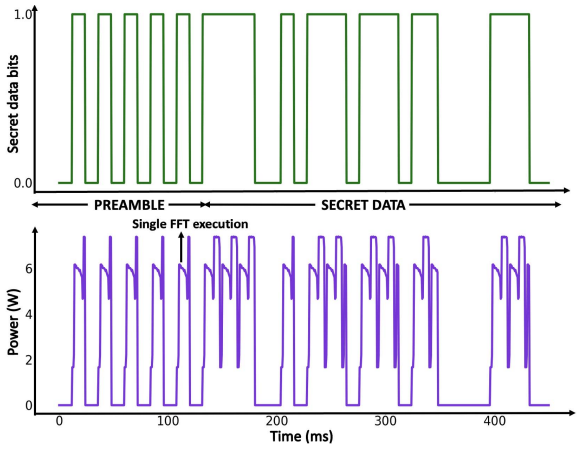

Fig. 5. Sample secret data bit-stream and the corresponding TCC application power profile of the transmitting core

be executed continuously. Thus, the temperature of the transmitting core can increase excessively, leading to overheating issues. However, this problem is mitigated in our work since we execute a normal *FFT* program that does not consume such high power and, hence, does not cause overheating even when transmitting continuous "1"s. Therefore, the need for alternative encoding schemes that reduce the communication bandwidth is eliminated.

3) Thermal Covert Communication Analysis: To characterize the impact of 3-D integration on thermal covert channel attacks, the modulated power trace of the transmitting core encoded with the secret message is given as an input to a thermal simulator. We perform the simulations in PACT, a modern parallel thermal simulator capable of efficiently handling multilayered chips with fine granularity of layer thickness [34]. Yuan *et al.* [34] validated the transient simulations of PACT and demonstrated it to be 186× faster than the well-known thermal simulator HotSpot while exhibiting a maximum error of 2.77% for steady state and 3.28% for transient thermal simulations compared to COMSOL, a finite-element method (FEM)-based simulator [35]. We model the cross section of 2-D, Mono3D, and TSV3D chip, as shown in Fig. 6. The 2-D chip floorplan for the quad-core processor is adapted from the published work on the Intel Haswell processor [27]. After analyzing the existing types of tier partitioning strategies for 3-D multicore processors [36], [37], [38], the 2-D floorplan in this work is partitioned into two tiers for both Mono3D and TSV3D systems where each tier has two cores, as shown in Fig. 6(b). We characterize the bandwidth and BERs of TCC among the cores of a flip-chip two-tier Mono3D and TSV3D (face-to-back bonding) integrated processor [39]. The TCC transmitting core (CORE_0) and receiver core (CORE_1) are highlighted in both floorplans in the figure.

The TSVs are modeled as two arrays of 10 mm × 0.2 mm at the top and bottom of the chip, as highlighted in Fig. 6. The diameter of each TSV is 10 μm [40], [41], [42] and the center-to-center pitch between the TSVs is 40 μm [15], producing 250 × 5 number of TSVs in each array. TSVs cross field-base dielectric layer, top-tier, adhesive benzocylcobutene (BCB) layer, and the bulk layers, as shown in the figure. The joint resistivity, ρ_{joint}, and specific heat capacity, c_{joint}, of the TSV array for each of these layers are calculated, assuming that the thermal resistance of the TSVs and the thermal resistance of the portion of the array that is not occupied by the TSVs are in parallel [43], [44], [45]. For example, for the TSV array within the BCB layer, ρ_{joint} and c_{joint} are

$$\rho_{\text{joint}} = \frac{A_{\text{TSVarray}}}{\frac{A_{\text{TSV}}}{\rho_{\text{TSV}}} + \frac{A_{\text{TSVarray}} - A_{\text{TSV}}}{\rho_{\text{BCB}}}} \tag{1}$$

$$c_{\text{joint}} = \frac{c_{\text{TSV}} A_{\text{TSV}} + c_{\text{BCB}} (A_{\text{TSVarray}} - A_{\text{TSV}})}{A_{\text{TSVarray}}} \tag{2}$$

where A_{TSVarray} is the area of the entire TSV array, A_{TSV} is the area occupied only by TSVs within the array, $\rho_{\text{TSV}} = 2.56\text{e}^{-3}$ mK/W and $\rho_{\text{BCB}} = 3.448$ mK/W are the thermal resistivities of TSV (copper) and BCB, respectively, and $c_{\text{TSV}} = 3.45\text{e}^{6}$ J/m^{3}K and $c_{\text{BCB}} = 1.75\text{e}^{6}$ J/m^{3}K are the specific heat capacities of TSV (copper) and BCB, respectively. The grid resolution for the 3-D IC is 16 × 16 where the size of each grid is 625 μm × 837.5 μm. Since the size of a TSV array is 10 mm × 200 μm, there are multiple grids within the array structure that partially overlap with the TSV array. The lateral heat conduction between those grids within the TSV array is considered in all of the simulations. However, heat flow among individual TSVs is not considered since we rely on joint resistivity and heat capacity. For Mono3D technology, the monolithic intertier vias (MIV) are modeled within the inter-layer dielectric (ILD) layer in Fig. 6. The diameter of each MIV is 50 nm and the center-to-center pitch is 170 nm [46]. Therefore, the overall number of the MIVs is 59 K × 76 K, which is still only 7% of the total ILD area. The joint thermal resistivity and specific heat capacity of the ILD layer are calculated similar to (1) and (2). Specifically, ρ_{joint} is 0.183 mK/W and c_{joint} is 3.25e^{6} J/m^{3}K. For thermal simulations, the grid size is the same for 2-D, TSV3D, and Mono3D systems and we use the default steady-state and transient solver options in PACT. The heat sink for all of the systems is modeled as a conventional pin fin heat sink with a heat spreader and a fan to mimic practical cooling mechanisms in processors [34], [47]. The heat transfer coefficient is the same for the three systems to ensure that the same amount of heat is removed by the heat sink per unit area.

A linear model for the temperature-dependent leakage power for the target processor is adapted and scaled based on the published data for Intel 22-nm processors [48]. We derive the leakage power model for every core of the Intel Haswell processor as $P_{\text{leak}} = 0.0137 \times T - 0.055$, where T is the temperature profile of the core from PACT in degree Celsius. The thermal simulations are typically performed in multiple iterations until the temperature variation is within 1 °C. In our experiments, we obtain this convergence in two iterations. The final output temperature of the receiver core is analyzed to characterize the covert communication channel and to decode the secret data.

4) Decoding the Secret Data: The secret data are embedded in the temperature profile of the transmitting core and are communicated to the receiving core through the thermal coupling between them. Since both of the cores are compromised by

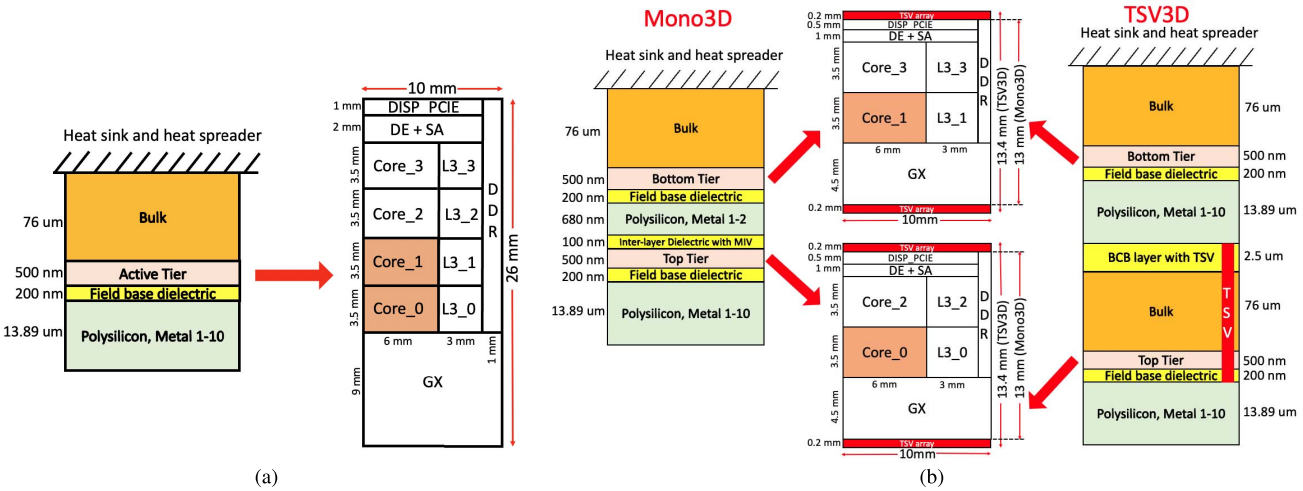

Fig. 6. Cross-sectional layers and active layer floorplan for Haswell processor integrated in (a) 2-D and (b) Mono3D and TSV3D technologies.

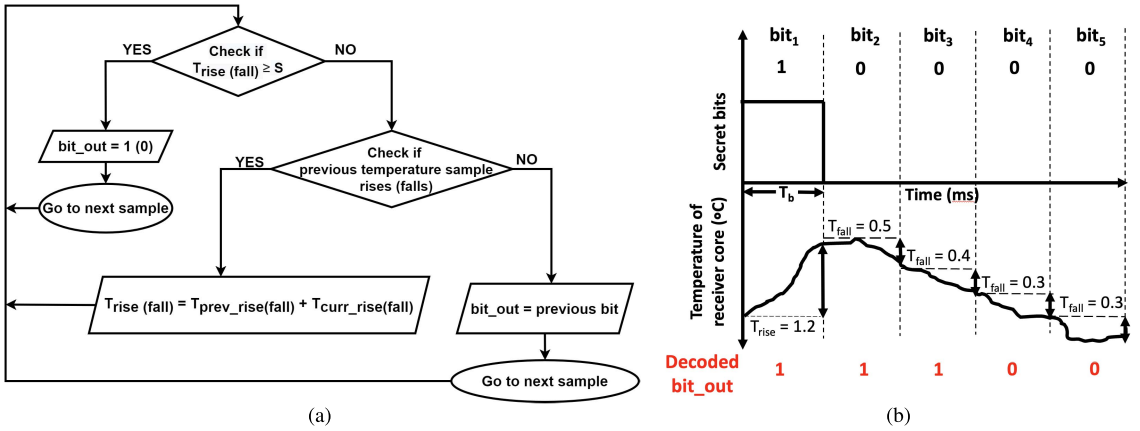

Fig. 7. Decoding process: (a) algorithm explained using flowchart, (b) example, where sensor resolution, $S = 1$ °C and $T_{\text{rise(fall)}}$ is the rise (fall) time in the temperature of the receiver core.

the attacker, the time of communication, and the bit-width, preamble bits are agreed upon as part of the communication protocol. Therefore, the receiving core records the temperature sensor information as soon as the transmitting core starts the encoding process. These data bits can either be decoded within the receiving core or sent offline for remote decoding since the receiving core has access to the network, as described in Section III-A. The decoding process used in this work is explained via the flowchart and example in Fig. 7.

The transient temperature profile of the receiving core is sampled at the end of every bit-width. The temperature rise (T_{rise}) or fall (T_{fall}) at every sample is recorded. The minimum difference in the temperature detectable by the sensor, also called the sensor resolution, is referred to as S. Since most modern processors have temperature sensors with a resolution of approximately 1 °C [49], [50], S is assumed to be 1 °C in this work. In Fig. 7(b), the first secret bit, bit_1, is "1" and the corresponding rise in the receiver core temperature is 1.2 °C. Based on the algorithm in Fig. 7(a), since this temperature rise is greater than $S = 1$ °C, the decoded output bit is also "1." However, for the following two bits (bit_2 and bit_3), the decoded output bit remains as "1" since the cumulative fall in temperature is less than S. For bit_4, cumulative fall $T_{\text{fall}} = 0.5 + 0.4 + 0.3 = 1.2$ °C and is greater than S, and therefore, the output bit is "0."

Based on the detected bit_out, the BER in percentage is characterized as

$$\text{BER} = \frac{n_{\text{corr}}}{N} \times 100 \tag{3}$$

where n_{corr} is the number of correct bits in bit_out that matches the transmitted bits and N is the total number of bits received. In the example discussed above, $n_{\text{corr}} = 3$ and $N = 5$ and, hence, BER = 60%. The transient temperature simulations are performed extensively for various transmission rates (1/bit_width). The effective bandwidth of the communication channel is estimated as the highest bitrate that can yield less than 1% BER. Note that this assumption for BER is based on previous works that show similar or higher TCC error rates [5], [8]. Error correction techniques can also be used to further reduce the BER of TCC [11], [12].

IV. Results

The experiments are performed by encoding ten blocks of secret data into the transient power profile of the transmitting core (Fig. 5). Each block comprises 10 bits of preamble and

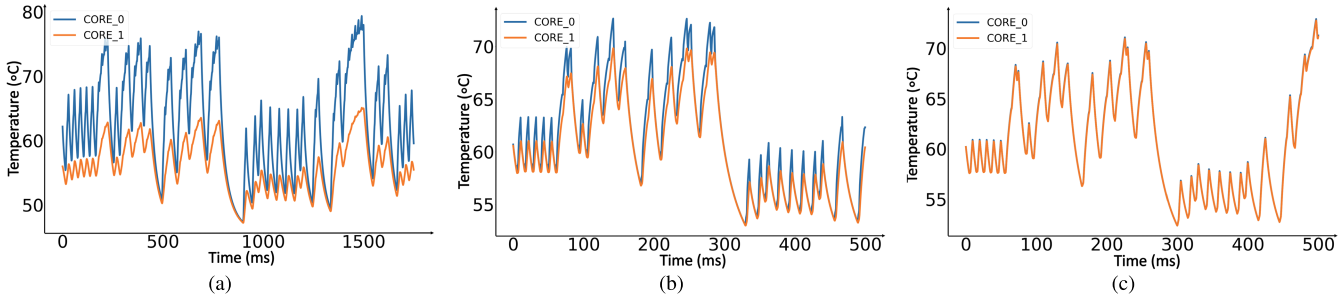

Fig. 8. Transient temperature profiles of the transmitting core (CORE_0) and receiver core (CORE_1) of the Haswell processor for transmitting one block of secret data for (a) 2-D for bit-width = 17 ms, (b) TSV3D for bit-width = 5.5 ms, and (c) Mono3D for bit-width = 5 ms. The time scale is different for (a) since the *FFT* application is executed for longer durations in 2-D to obtain reduced error rates.

100 random bits of data and the same set of random bits are encoded in 2-D, Mono3D, and TSV3D systems to ensure a fair comparison. The characterization of TCC is performed for six different scenarios: 1) without thermal interference from other cores (noise-free) to isolate the effect of thermal coupling between the transmitting and receiving cores on TCC; 2) with thermal interference to study the effect of other active cores on TCC bandwidth; 3) with partial and nonoverlapping placement of the transmitter and receiver cores in 3-D systems to analyze the effect of 50% and 0% overlap on TCC; 4) when the transmitting core is placed closer to the heat sink to analyze the effect of changing heat flow on TCC bandwidth, 5) with a lower power TCC program to analyze the effect of transient power variations on TCC; and 6) with a four-tier 3-D system to study the effect of more than two tiers on TCC. Note that for one of the scenarios, we performed an experiment with 30 blocks of 100 random data bits to determine whether the TCC bandwidth would change with a larger block size. The average BER, in this case, was still less than 1%, validating the TCC bandwidth obtained with ten blocks of data.

A. TCC Characterization Without Thermal Interference

In this scenario, the cores other than the transmitter and the receiver are assumed to be in the sleep state. The transmitting core (CORE_0) and the receiving core (CORE_1) are placed adjacent to each other on the 2-D system and are placed on top of each other on Mono3D and TSV3D processors, as shown in Fig. 6. The output transient temperature profiles of all the cores for transmitting one block of data is shown in Fig. 8(a) for 2-D processor for a bit-width of 17 ms, in Fig. 8(b) for TSV3D processor for a bit-width of 5.5 ms, and in Fig. 8(c) for Mono3D processor for a bit-width of 5 ms. Transmitting ten blocks of secret data at these bit-widths yields a BER of less than 1%. The differences in the thermal coupling between transmitter and receiver cores for 2-D, Mono3D, and TSV3D processors can be observed in the figure. For the 2-D system, a 10 °C increase in the CORE_0 temperature produces an increase of 2 °C in the CORE_1 temperature due to the lateral coupling. However, for TSV3D, a 10 °C increase in the CORE_0 temperature produces an increase of 8 °C because of the stronger vertical thermal coupling between the transmitting core located on the top tier and receiver core located on the bottom tier. Alternatively, the temperature profile of the transmitter and the receiver cores overlaps for the Mono3D system because of the highest intertier thermal coupling enabled by the sufficiently small thickness of the ILD layer. Note that the temperature range of the transmitting core in 2-D IC in Fig. 8(a) is higher than the corresponding temperatures for Mono3D and TSV3D systems because the *FFT* application is executed for an extended period of time in the 2-D system (bit-width of 17 ms) to ensure sufficiently low BER.

The BER of the covert communication is characterized for various bit-widths in order to determine the effective bandwidth. The variation of BER for different bit-widths for each block of data and the average BER of all the blocks are shown in Fig. 9. As observed from the figure, the error rate varies for each block due to the randomness of the secret data. The bit-width at which the average BER is less than 1% determines the effective bandwidth, as listed in columns 2 and 3 of Table II. From the table, it can be observed that the bandwidth of a Mono3D-based TCC attack is 200 bps and is 3.5× greater than the bandwidth achieved by a 2-D integrated processor. Furthermore, the bandwidth achieved by Mono3D and TSV3D technologies is similar even though the vertical thermal coupling is stronger in the Mono3D technology. The primary reason for this similarity is the higher thermal resistance between the top tier (where the transmitting core is located) and the heat sink for the TSV3D-based system, as shown in Fig. 6. Therefore, the steady-state and the peak-to-peak values of the temperature of the transmitting core for TSV3D processor are greater than those of Mono3D. Consequently, even though the intertier coupling is weaker in TSV3D technology, since the transmitting core temperatures are greater, the bandwidth of TSV3D and Mono3D processors is comparable, making both technologies more vulnerable to a TCC attack than a 2-D many-core processor.

B. TCC Characterization With Thermal Interference

In Section IV-A, we show that a typical benchmark program, such as *FFT* with a nominal power profile, is sufficient to transfer at 200 bps in a 3-D many-core processor. However, this bandwidth was achieved in a lightly loaded scenario when all of the other cores are in the sleep state. In this section, we investigate the effect of an active core (referred to as noise core) other than the transmitter and receiver cores, on TCC bandwidth.

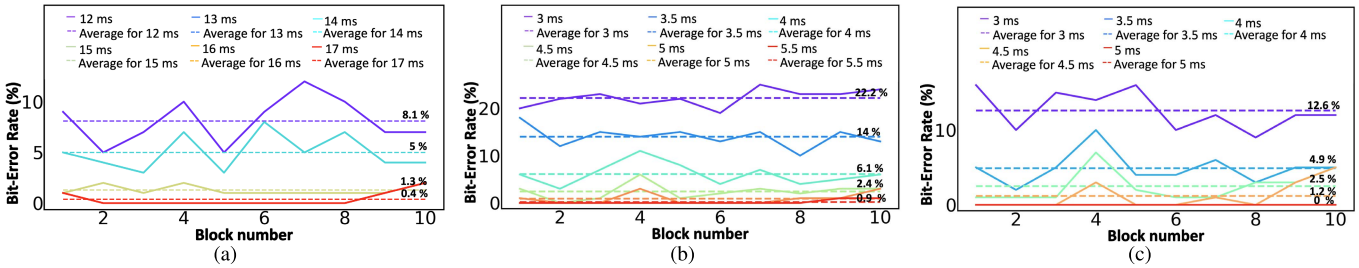

Fig. 9. BER versus number of blocks of secret data for different bit-widths of time for (a) 2-D, (b) TSV3D, and (c) Mono3D-based systems.

TABLE II

COMPARISON OF TCC BANDWIDTH (BW) FOR 2-D, MONO3D, AND TSV3D BASED MULTICORE PROCESSOR FOR THREE SCENARIOS WITH AND WITHOUT THERMAL INTERFERENCE (NOISE): (1) WITH 100% VERTICAL OVERLAP, (2) WITH 50% VERTICAL OVERLAP, AND (3) WITH 0% OVERLAP BETWEEN TRANSMITTING AND RECEIVING CORES. THE NOISE APPLICATION IS EXECUTED ONLY IN THE CORE ADJACENT TO THE RECEIVING CORE FOR THE SCENARIOS WITH NOISE.

Integration technology	With 100% overlap				With 50% overlap				With 0% overlap			
	Without noise		With noise		Without noise		With noise		Without noise		With noise	
	BW (bps)	BER (%)	BW (bps)	BER (%)	BW (bps)	BER (%)	BW (bps)	BER (%)	BW (bps)	BER (%)	BW (bps)	BER (%)
2D	59	<1	52	3	NA	NA	NA	NA	NA	NA	NA	NA
Mono3D	200	<1	200	<1	167	<1	142	<1	77	<1	67	<1
TSV3D	182	<1	182	<1	154	<1	100	<1	77	<1	67	6

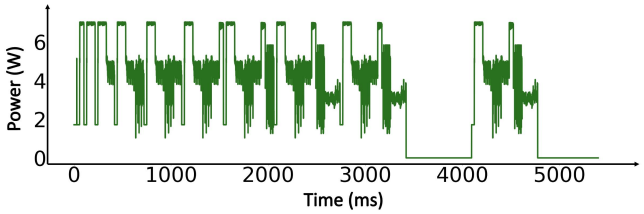

Fig. 10. Power profile of the noise application executing in cores other than the transmitter and the receiver cores.

During TCC between the transmitting and the receiving cores, the noise core sequentially executes random applications from the SPLASH-2 benchmark suite, with dispersed instants of idle time. These noise applications are executed in CORE_3 for the 3-D systems and in CORE_2 for the 2-D system because of the proximity of these cores to the receiving core (as shown in Fig. 6). The transient power consumption of the noise core is shown in Fig. 10. Thermal simulations in PACT are performed with the noise power trace. TCC error rates are estimated using a similar analysis as in Fig. 9 in order to characterize the bandwidth. These results are tabulated in columns 4 and 5 of Table II. As observed from the table, the TCC bandwidth in Mono3D and TSV3D processors remains approximately the same. Therefore, TCC in 3-D processors exhibits increased resistance to noise. The reason for this higher robustness can be better described with the transient temperature profiles of the transmitter, receiver, and noise cores, as shown in Fig. 11. First, the variation in the temperature profile of the noise core (CORE_3) executing SPLASH-2 applications is sufficiently slower compared to the temperature profile of the transmitting core (CORE_0) encoded with the secret data, as shown in Fig. 11(b) and (c). Furthermore, the temperature of the receiver core (CORE_1) has sharp rise and fall times compared to the temperature of the noise core (due to the strong vertical thermal coupling). Since the TCC bandwidth depends on these steep rise and fall times of the receiver core, the 3-D processors exhibit lower sensitivity to thermal interference from the noise core. Note that the noise applications from SPLASH-2 are executed for the total execution time, whereas the TCC program is executed only for the duration of the bit-width shown in Fig. 11. Therefore, the rise in temperature for the transmitting core (CORE_0) and the receiving core (CORE_1) is lower than the rise in temperature of the noise core (CORE_3).

In 2-D processors with thermal interference, the TCC bandwidth degrades to 52 bps with BER of 3%, as listed in column 5 of Table II. This degradation is due to similar rate of variation in the temperature profile of the noise core (CORE_2) and transmitting core for the 2-D system, as shown in Fig. 11(a). In other words, the transmission rate of the covert channel application overlaps with the rate at which the temperature of the noise core varies.

The results are further described in Fig. 12(a), which depicts the variation of the BER with TCC transmission rates in 2-D and 3-D processors in the presence of the noise core. The transmission rates are divided into three regions: B1 (yellow), B2 (pink), and B3 (green). B1 refers to the range of bitrates that overlap with the frequency of the temperature profile of the noise core. B2 refers to the range for which the BER is negligible for TSV3D and Mono3D processors. The BER in 3-D systems starts to increase with increase in the transmission rates in region B3. In B2, BER at 83 bps is 16% for the 2-D processor. When the transmission rate is decreased,

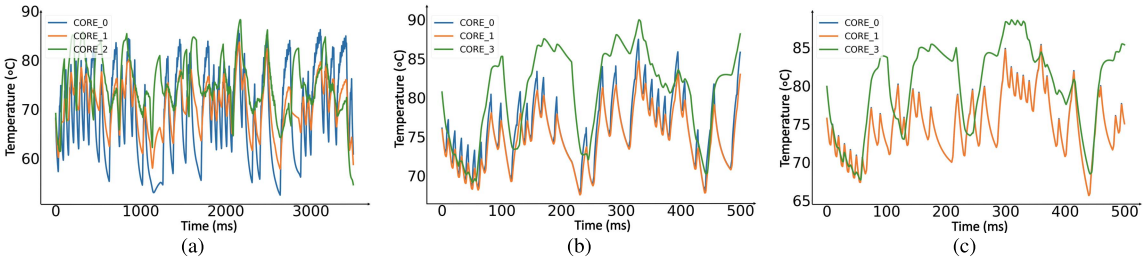

Fig. 11. Transient temperature profiles of the transmitting core (CORE_0), receiver core (CORE_1), and noise core of the Haswell processor for transmitting one block of secret data for (a) 2-D for a bit-width of 30 ms, (b) TSV3D for a bit-width of 6 ms, and (c) Mono3D for a bit-width of 5 ms. The time scale is different for (a) because the *FFT* application is executed for longer duration in 2-D to obtain reduced error rates.

BER also decreases, as expected. However, the BER does not decrease below 3% (at 52 bps). Reducing the TCC rate below 52 bps interferes with the frequency range of the temperature of noise core in B1 and this interference increases the BER monotonically. However, a TCC attack can be mounted with negligible error rates in 3-D systems at transmission rates up to 182 bps for TSV3D and 200 bps for Mono3D, without interfering with the frequency of the noise core temperature in B1. Thus, for 3-D systems, thermal interference does not affect the TCC bandwidth. Note that the execution of the noise application in CORE_2 (closer to the transmitting core) instead of CORE_3 does not change the TCC bandwidth results for the 3-D systems. However, when the noise application is executed in both CORE_2 and CORE_3, the TCC bandwidth of the Mono3D and TSV3D systems is reduced to 182 and 133 bps, respectively. Alternatively, for 2-D processor, the maximum TCC bandwidth is 50 bps with a relatively large BER of 7%. Although there is some degradation in the bandwidth for Mono3D and TSV3D systems, the effect of thermal interference is significantly weaker compared to TCC in the 2-D system.

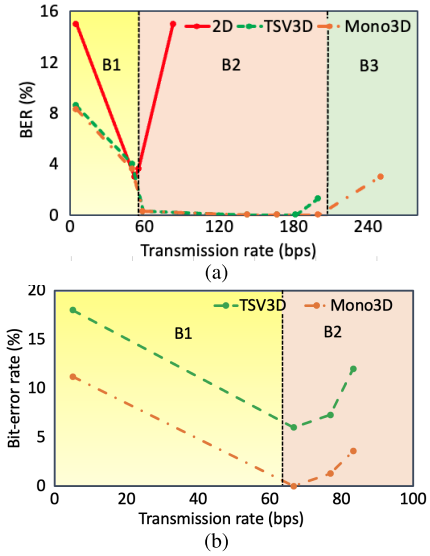

Fig. 12. BER versus transmission rate for (a) 2-D, TSV3D, and Mono3D integrated processor in the presence of a noise core (CORE_3 for 3-D processors and CORE_2 for the 2-D processor), and (b) TSV3D and Mono3D integrated processor with 0% overlapping transmitter CORE_0 and receiver CORE_3 and a noise core CORE_1.

C. Nonoverlapping Transmitting and Receiving Cores

As shown in Sections IV-A and IV-B, the TCC bandwidth in 3-D processors is significantly higher than 2-D processors due to the strong thermal coupling between the transmitter (CORE_0) and the receiver (CORE_1) cores that are located in different tiers, as shown in Fig. 6. In this section, we investigate the effect of reducing the overlap between the transmitter and receiver cores. Two different placement scenarios are considered, one with 50% overlap and the other one with 0% overlap. For the first scenario, the 2-D floorplan of the Haswell processor (see Fig. 6) is partitioned into two tiers such that there is a 50% overlap between the cores on each tier, as shown in Fig. 13. The transmitting core is CORE_0 and the receiving core is CORE_1. The results are listed in columns 6 and 7 of Table II. The TCC bandwidth yielding BER $< 1\%$ is 167 bps for Mono3D and 154 bps for TSV3D processor. These results are 17% and 15% lower than the bandwidths obtained with 100% overlap. The primary reason for this degradation is the reduced thermal coupling between the two cores in this scenario. Furthermore, the noise application was also executed in CORE_3 to study the effect of thermal interference and the results are tabulated in columns 8 and 9 of Table II.

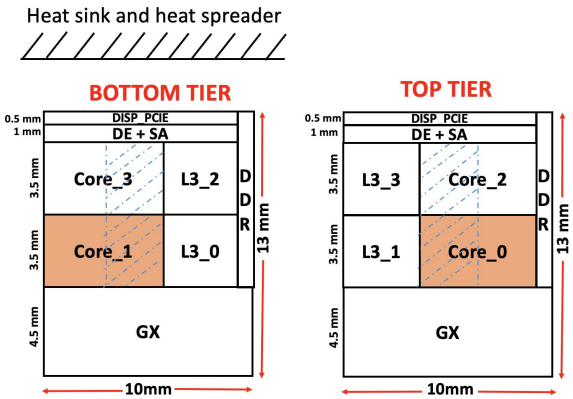

Fig. 13. Bottom tier and top tier floorplan for the scenario where the transmitting and receiving cores overlap by 50%.

The TCC bandwidths degraded further, by 15% and 35% for, respectively, Mono3D and TSV3D technologies.

For the second scenario, the transmitting and receiving cores do not overlap. The same floorplan shown in Fig. 6

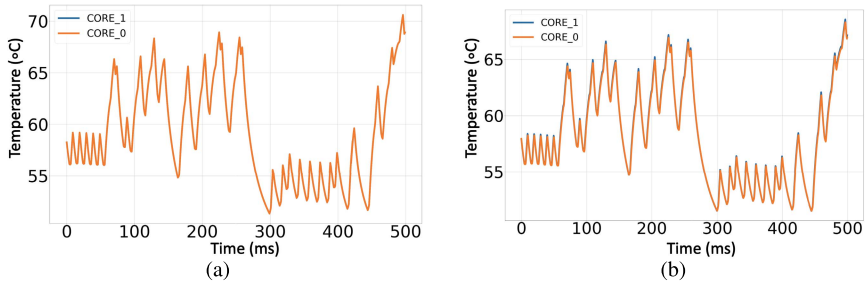

Fig. 14. Transient temperature profiles when the transmitting core is CORE_1 (on the bottom tier) and the receiving core is CORE_0 (on the top tier) for (a) Mono3D and (b) TSV3D systems.

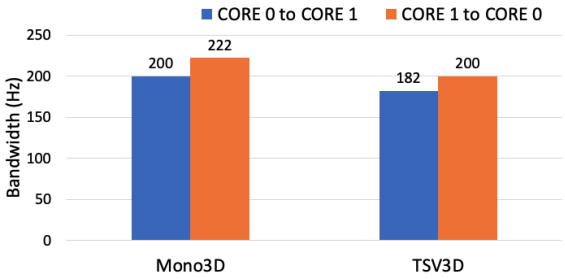

Fig. 15. Comparison of TCC bandwidth (with BER < 1%) for Mono3D and TSV3D systems with both CORE_0 to CORE_1 and CORE_1 to CORE_0 communication.

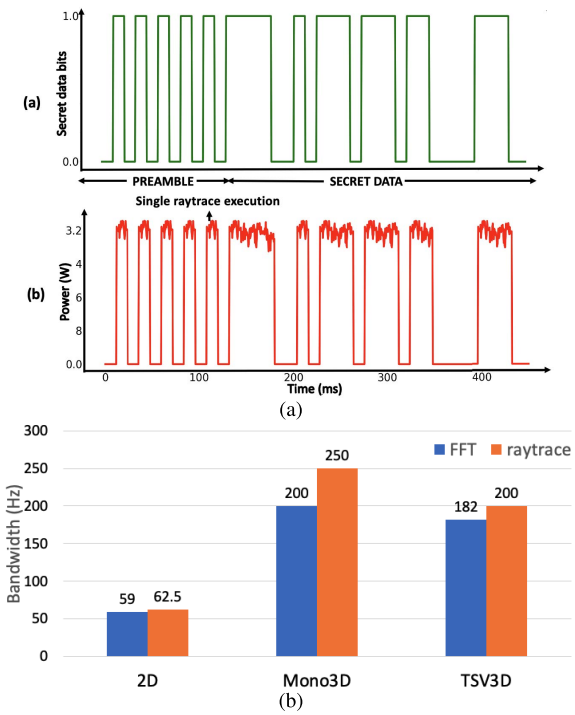

Fig. 16. TCC using lower power *raytrace* program from SPLASH-2. (a) TCC encoding showing (top) sample secret data bit-stream and (bottom) the encoded transmitting core power profile. (b) Comparison of communication bandwidths with *FFT* and *raytrace* as the TCC programs.

is assumed. The transmitting core is still CORE_0, whereas the receiving core is CORE_3. According to TCC results, when the transmitter and receiver cores do not have any

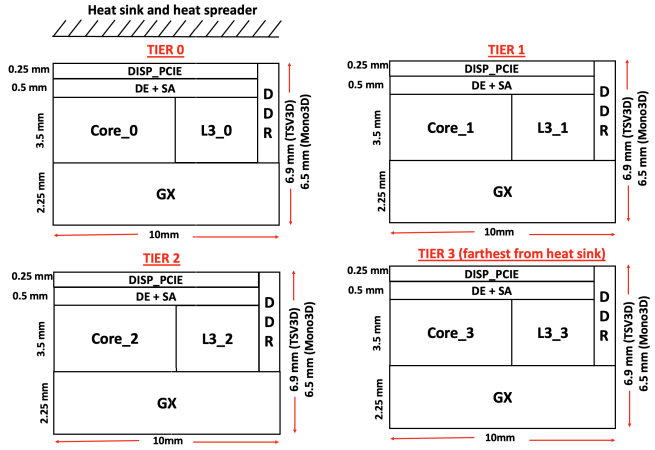

Fig. 17. Floorplan of the Haswell processor partitioned into four tiers using Mono3D and TSV3D integration.

overlap, TCC bandwidth degrades by 62% for Mono3D and 58% for TSV3D processor, as listed in columns 10 and 11 of Table II. Therefore, in 3-D ICs, it is preferable for apps that have access to sensitive information to execute on cores that are not overlapping with cores executing the insecure apps. Furthermore, the thermal interference from CORE_1 (the core closest to the receiver CORE_3) was also considered, similar to Section IV-B. In the presence of the noise power profile shown in Fig. 10, the TCC bandwidth in Mono3D and TSV3D processors is further reduced by approximately 13%, as listed in columns 12 and 13 of Table II. Note that the degradation in bandwidth in the presence of thermal interference for both of the scenarios is in contrast to the unaffected TCC bandwidth when the cores fully overlap. The reason for the degradation of bandwidth for this scenario can be explained with the help of Fig. 12(b) that illustrates the variation of BER with the transmission rates for 0% overlapping cores. The plots resemble the variation for 2-D processors observed in Fig. 12(a), and the transmission rate is similarly divided into regions B1 (yellow) and B2 (pink). Due to the weaker thermal coupling between the nonoverlapping cores, the minimum BER for TSV3D and Mono3D processors is achieved only at a transmission rate of 67 bps, as seen in region B2. For less than 67 bps, the transmission rates start overlapping with the frequency of temperature profile of the noise core, thus increasing the sensitivity to heat generated by other

cores. Note that when the thermal interference is considered from CORE_2, a more significant degradation in the TCC bandwidth was observed due to strong coupling between the noise core and receiving core. Specifically, the TCC bandwidth for Mono3D and TSV3D processors was 50 bps with a large BER of 12%.

D. Placement of Transmitting Core Closer to Heat Sink

All of the simulations in Sections IV-A–IV-C considered that the transmitting core is on the top tier, away from the heat sink, as shown in the Mono3D and TSV3D layer stack in Fig. 6. Alternatively, in this section, we consider a scenario where the transmitting core (CORE_1) is located on the bottom tier, closest to the heat sink, and the receiving core (CORE_0) is located on the top tier. The temperature profile of the transmitting core and the receiving core is shown in Fig. 14 for both Mono3D and TSV3D processors. Since the bottom tier is located closer to the heat sink, the majority of the heat flows from the transmitting core on the bottom tier to the heat sink. Thus, the core temperatures in both tiers overlap in both of the systems, as illustrated in the figure. This behavior is in contrast to the transient temperature profiles for CORE_0 to CORE_1 communication in Fig. 8, particularly for TSV3D processors where there is a significant difference between the transmitting and receiving core temperatures.

The bandwidth with less than 1% BER is characterized for this scenario, as shown in Fig. 15, where the bandwidths are compared for both CORE_0 (in top tier) to CORE_1 (in bottom tier) and CORE_1 to CORE_0 scenarios. We can observe that TCC bandwidths for Mono3D and TSV3D processors are greater for the CORE_1 to CORE_0 scenario by, respectively, 11% and 9.9%. The bandwidth increases since the receiving core temperature is almost identical to the transmitting core temperature into which the secret data are encoded. Therefore, an attacker can transmit a higher number of bits during the same time interval, thereby increasing the danger posed by this scenario.

E. Effect of Transient Power Variations on TCC Bandwidth

As discussed in Section III-B2, *FFT* program from the SPLASH-2 suite is used to establish a TCC attack for the results presented thus far. In this section, the effect of executing a lower power program on TCC bandwidth is explored. As shown in Fig. 4, *raytrace* is one of the low-power applications within SPLASH-2. When this application is used to encode the sensitive data bits, the encoded power profile of the transmitting core is shown in Fig. 16(a). The TCC bandwidth that yields a BER $< 1\%$ is shown in Fig. 16(b) for 2-D, Mono3D, and TSV3D systems. Although the peak power consumption of the transmitting core executing the *FFT* program is greater than the peak power consumed when executing the *raytrace* program, TCC bandwidth achieved by *raytrace* is greater by 6%, 25%, and 10%, respectively, for 2-D, Mono3D, and TSV3D systems. This increase in TCC bandwidth can be explained through the transient power profile. Specifically, when executing *FFT*, even though the peak power is higher, the power variation during execution is also high, as shown in Fig. 5. Alternatively, the transient power profile of *raytrace* is relatively more stable during execution (even though at lower power levels), as shown in Fig. 16(a). This transient stability of power profile results in higher TCC bandwidth since the corresponding temperature profile exhibits less noise.

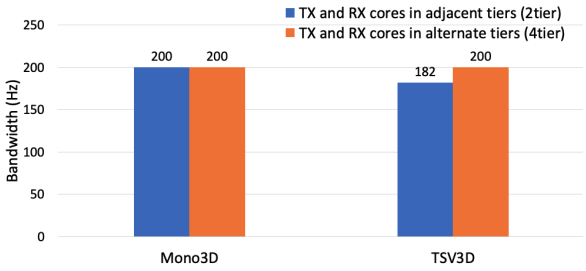

Fig. 18. Comparison of TCC bandwidth (with BER $< 1\%$) for Mono3D and TSV3D systems with transmitting and receiving cores on adjacent tiers and alternate tiers.

F. TCC in 3-D Processors With More Than Two Tiers

In this section, the TCC bandwidth is analyzed for a scenario with more than two tiers in Mono3D and TSV3D processors. The 2-D floorplan of the Intel Haswell processor in Fig. 6 is partitioned into four tiers and the floorplan of each tier is shown in Fig. 17. We perform two experiments with the four-tier processor. First, TCC between cores on nonadjacent tiers is studied, where CORE_3 on Tier 3 is the transmitting core and CORE_1 on Tier 1 is the receiving core. The objective is to quantify the impact of having an additional tier between transmitting and receiving cores. According to the results of this scenario, as shown in Fig. 18, the TCC bandwidth for Mono3D system is the same as the two-tier scenario where the communicating cores are located on adjacent tiers. Alternatively, the bandwidth of TSV3D is 10% greater for the four-tier scenario, despite the fact that transmitting and receiving cores are located farther apart on nonadjacent tiers. This behavior can be explained with the transient temperature profiles of the transmitting and receiving cores, as shown in Fig. 19. For TCC between nonadjacent tiers, the thermal coupling is still sufficiently strong for the Mono3D processor (due to thin cross-sectional layers), resulting in the same bandwidth. For the TSV3D processor, the peak temperature of CORE_3 is 75 °C, whereas for Mono3D, it is 70 °C. Furthermore, the rise and fall times are also steeper for CORE_3 in the TSV3D processor because the thermal resistance of the upper most tier (Tier 3) is the highest for TSV3D technology (due to thick bulk and BCB layers between consecutive tiers). Thus, due to higher temperatures and steeper rise/fall times, the TCC bandwidth for the TSV3D processor is greater for the four-tier scenario with nonadjacent transmitting and receiving cores compared to the two-tier scenario with adjacent transmitting and receiving cores.

For the second experiment with the four-tier model, CORE_3 is the transmitting core, CORE_2 is the receiving core, and noise application is executed in CORE_1. In this

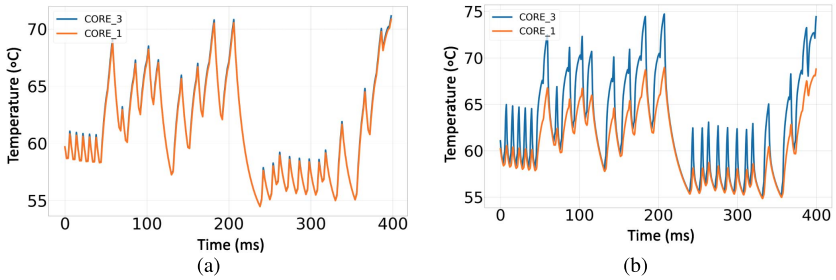

Fig. 19. Transient temperature profiles of the transmitting core (CORE_3) on Tier 3 and receiving core (CORE_1) on Tier 1 for (a) Mono3D and (b) TSV3D-based processor.

scenario, the maximum TCC bandwidth is calculated as 100 bps for both Mono3D and TSV3D processors with BER of 2% and 3%, respectively. Thus, when the noise application is executed within the tier directly beneath the receiving core, the TCC bandwidth is significantly degraded due to strong coupling of the interference caused by the noise application.

V. Conclusion and Future Work

In modern multicore processors, accessing sensitive information by using heat as a medium of communication between cores is a dangerous security threat. In conventional 2-D processors, heat flow among the cores occurs primarily in a lateral fashion. However, strong vertical thermal coupling in 3-D technologies increases the covert channel communication bandwidth, as shown in this work. We demonstrate that the thermal covert channel bandwidth between cores of a Mono3D- and TSV3D-based Intel processor is 3.4× and 4× greater than the bandwidth achieved in a conventional 2-D processor. Furthermore, unlike previous works that typically rely on computationally intensive CPU stress applications to encode the secret data, in this work, we establish a high bandwidth channel by using common SPLASH-2 benchmark applications such as *FFT* and *raytrace*. Furthermore, we also perform thermal covert channel characterization in the presence of thermal interference due to neighboring active cores and show that the covert channel bandwidth in 3-D systems is mostly unaffected from heat generated by the other cores while still achieving less than 1% BER. However, for 2-D systems, the thermal interference increases the minimum BER to 7% and the bandwidth degrades by 13%. The significant increase in covert channel communication bandwidth in vertically integrated processors is due to the overlapping transmitter and receiver cores, which maximizes the thermal coupling between them. Thus, we investigate the effect of reducing or eliminating the overlap between the cores on covert channel communication and show that the bandwidth degrades by up to 62% for the Mono3D processor and by 58% for the TSV3D processor.

A potential future direction is the accurate detection of TCC in 3-D ICs, executed with low and moderate power applications. The existing techniques for detecting TCC are based on comparing the temperature or CPU workload of each core in the frequency domain with a set threshold [11], [12]. If the power spectrum of the above metrics is greater than this threshold, TCC is said to be detected. This threshold is calculated based on a core executing SPLASH-2/PARSEC applications without a TCC. Since we propose that it is possible to establish a high bandwidth TCC by executing the same SPLASH-2/PARSEC applications, these detection techniques may not provide sufficiently accurate detection rates, as explored in future work.

References

[1] B. A. Smith and K. Curran, "Security vulnerabilities in microprocessors," *Semicond. Sci. Inf. Devices*, vol. 3, no. 1, May 2021.
[2] Orange Book, "Trusted computer system evaluation criteria," U.S. Nat. Comput. Secur. Center, Tech. Rep., 1985.
[3] R. A. Kemmerer, "A practical approach to identifying storage and timing channels: Twenty years later," in *Proc. 18th Annu. Comput. Secur. Appl. Conf.*, 2002, pp. 109–118.
[4] H. Okhravi, S. Bak, and S. T. King, "Design, implementation and evaluation of covert channel attacks," in *Proc. IEEE Int. Conf. Technol. Homeland Secur. (HST)*, Nov. 2010, pp. 481–487.
[5] R. J. Masti, D. Rai, A. Ranganathan, C. Müller, L. Thiele, and S. Capkun, "Thermal covert channels on multi-core platforms," in *Proc. 24th USENIX Secur. Symp.*, 2015, pp. 865–880.
[6] M. Guri, O. Hasson, G. Kedma, and Y. Elovici, "VisiSploit: An optical covert-channel to leak data through an air-gap," 2016, *arXiv:1607.03946*.
[7] L. Deshotels, "Inaudible sound as a covert channel in mobile devices," in *Proc. 8th USENIX Workshop Offensive Technol.*, 2014, pp. 1–9.
[8] D. B. Bartolini, P. Miedl, and L. Thiele, "On the capacity of thermal covert channels in multicores," in *Proc. 11th Eur. Conf. Comput. Syst.*, Apr. 2016, pp. 1–16.
[9] L. Brown and H. Seshadri, "Cool hand linux* handheld thermal extensions," in *Proc. Linux Symp.*, 2007, pp. 75–80.
[10] S. Wang, X. Wang, Y. Jiang, A. Singh, L. Huang, and M. Yang, "Modeling and analysis of thermal covert channel attacks in many-core systems," *IEEE Trans. Comput.*, early access, Mar. 17, 2022, doi: 10.1109/TC.2022.3160356.
[11] H. Huang, X. Wang, Y. Jiang, A. K. Singh, M. Yang, and L. Huang, "Detection of and countermeasure against thermal covert channel in many-core systems," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 41, no. 2, pp. 252–265, Feb. 2022.
[12] H. Huang, X. Wang, Y. Jiang, A. K. Singh, M. Yang, and L. Huang, "On countermeasures against the thermal covert channel attacks targeting many-core systems," in *Proc. 57th ACM/IEEE Design Autom. Conf. (DAC)*, Jul. 2020, pp. 1–6.
[13] Q. Wu, X. Wang, and J. Chen, "Defending against thermal covert channel attacks by task migration in many-core system," in *Proc. IEEE 3rd Int. Conf. Circuits Syst. (ICCS)*, Oct. 2021, pp. 111–120.
[14] Z. Long, X. Wang, Y. Jiang, G. Cui, L. Zhang, and T. Mak, "Improving the efficiency of thermal covert channels in multi-/many-core systems," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2018, pp. 1459–1464.
[15] P. Emma *et al.*, "3D stacking of high-performance processors," in *Proc. IEEE 20th Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2014, pp. 500–511.
[16] K. Dhananjay, P. Shukla, V. F. Pavlidis, A. Coskun, and E. Salman, "Monolithic 3D Integrated circuits: Recent trends and future prospects," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 68, no. 3, pp. 837–843, 2021.

[17] S. Tian and J. Szefer, "Temporal thermal covert channels in cloud FPGAs," in *Proc. ACM/SIGDA Int. Symp. Field-Program. Gate Arrays*, Feb. 2019, pp. 298–303.
[18] I. Giechaskiel, K. B. Rasmussen, and J. Szefer, "C3APSULe: Cross-FPGA covert-channel attacks through power supply unit leakage," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2020, pp. 1728–1741.
[19] T. Claeys, F. Rousseau, B. Simunovic, and B. Tourancheau, "Thermal covert channel in Bluetooth low energy networks," in *Proc. 12th Conf. Secur. Privacy Wireless Mobile Networks*, 2019, pp. 267–276.
[20] P. Rahimi, A. K. Singh, and X. Wang, "Selective noise based power-efficient and effective countermeasure against thermal covert channel attacks in multi-core systems," *J. Low Power Electron. Appl.*, vol. 12, no. 2, p. 25, May 2022.
[21] J. Wang, X. Wang, Y. Jiang, A. K. Singh, L. Huang, and M. Yang, "Combating enhanced thermal covert channel in multi-/many-core systems with channel-aware jamming," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 39, no. 11, pp. 3276–3287, Nov. 2020.
[22] J. Knechtel and O. Sinanoglu, "On mitigation of side-channel attacks in 3D ICs: Decorrelating thermal patterns from power and activity," in *Proc. 54th ACM/EDAC/IEEE Design Automat. Conf. (DAC)*, Jun. 2017, pp. 1–6.
[23] P. Gu, D. Stow, R. Barnes, E. Kursun, and Y. Xie, "Thermal-aware 3D design for side-channel information leakage," in *Proc. IEEE 34th Int. Conf. Comput. Design (ICCD)*, Oct. 2016, pp. 520–527.
[24] S. Chen, W. Xiong, Y. Xu, B. Li, and J. Szefer, "Thermal covert channels leveraging package-on-package DRAM," in *Proc. 18th IEEE Int. Conf. Trust, Secur. Privacy Comput. Commun./13th IEEE Int. Conf. Big Data Sci. Eng. (TrustCom/BigDataSE)*, Aug. 2019, pp. 319–326.
[25] F. McKeen *et al.*, "Innovative instructions and software model for isolated execution," in *Proc. 2nd Int. Workshop Hardw. Architectural Support Secur. Privacy (HASP)*, 2013, pp. 1–8.
[26] *Arm Trustzone*. Accessed: Aug. 25, 2022. [Online]. Available: https://www.arm.com/technologies/trustzone-for-cortex-a
[27] N. Kurd *et al.*, "Haswell: A family of IA 22 nm processors," *IEEE J. Solid-State Circuits*, vol. 50, no. 1, pp. 49–58, Jan. 2015.
[28] *Haswell—Microarchitectures—Intel*. Accessed: Oct. 21, 2021. [Online]. Available: https://en.wikichip.org/wiki/intel/microarchitectures/haswell_(client)
[29] *Intel's Haswell CPU Microarchitecture*. Accessed: Oct. 12, 2021. [Online]. Available: https://www.realworldtech.com/page/3/
[30] T. E. Carlson, W. Heirman, S. Eyerman, I. Hur, and L. Eeckhout, "An evaluation of high-level mechanistic core models," *ACM Trans. Archit. Code Optim.*, vol. 11, no. 3, pp. 1–25, Oct. 2014.
[31] S. Li, J. H. Ahn, R. D. Strong, J. B. Brockman, D. M. Tullsen, and N. P. Jouppi, "McPAT: An integrated power, area, and timing modeling framework for multicore and manycore architectures," in *Proc. 42nd Annu. IEEE/ACM Int. Symp. Microarchitecture*, Dec. 2009, pp. 469–480.
[32] T. Rauber, G. Rünger, and M. Stachowski, "Performance and energy metrics for multi-threaded applications on DVFS processors," *Sustain. Comput., Informat. Syst.*, vol. 17, pp. 55–68, Mar. 2018.
[33] S. C. Woo, M. Ohara, E. Torrie, J. P. Singh, and A. Gupta, "The SPLASH-2 programs: Characterization and methodological considerations," *ACM SIGARCH Comput. Archit. News*, vol. 23, no. 2, pp. 24–36, 1995.
[34] Z. Yuan, P. Shukla, S. Chetoui, S. Nemtzow, S. Reda, and A. K. Coskun, "PACT: An extensible parallel thermal simulator for emerging integration and cooling technologies," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 41, no. 4, pp. 1048–1061, Apr. 2022.
[35] E. J. Dickinson, H. Ekström, and E. Fontes, "COMSOL Multiphysics: Finite element software for electrochemical analysis. A mini-review," *Electrochem. Commun.*, vol. 40, pp. 71–74, Mar. 2014.
[36] G. H. Loh, Y. Xie, and B. Black, "Processor design in 3D die-stacking technologies," *IEEE Micro*, vol. 27, no. 3, pp. 31–48, May/Jun. 2007.
[37] B. Gopireddy and J. Torrellas, "Designing vertical processors in monolithic 3D," in *Proc. 46th Int. Symp. Comput. Archit.*, Jun. 2019, pp. 643–656.
[38] X. Zhou, J. Yang, Y. Xu, Y. Zhang, and J. Zhao, "Thermal-aware task scheduling for 3D multicore processors," *IEEE Trans. Parallel Distrib. Syst.*, vol. 21, no. 1, pp. 60–71, Jan. 2010.
[39] P. Shukla, A. K. Coskun, V. F. Pavlidis, and E. Salman, "An overview of thermal challenges and opportunities for monolithic 3D ICs," in *Proc. Great Lakes Symp. VLSI*, May 2019, pp. 439–444.
[40] S. K. Samal, D. Nayak, M. Ichihashi, S. Banna, and S. K. Lim, "Monolithic 3D IC vs. TSV-based 3D IC in 14 nm FinFET technology," in *Proc. IEEE SOI-3D-Subthreshold Microelectron. Technol. Unified Conf. (S3S)*, Oct. 2016, pp. 1–2.
[41] Y.-H. Gong, J. Kong, and S. W. Chung, "Quantifying the impact of monolithic 3D (M3D) integration on l1 caches," *IEEE Trans. Emerg. Topics Comput.*, vol. 9, no. 2, pp. 854–865, Apr. 2021.
[42] H. Wang, M. H. Asgari, and E. Salman, "Compact model to efficiently characterize TSV-to-transistor noise coupling in 3D ICs," *Integr., VLSI J.*, vol. 47, no. 3, pp. 296–306, Jun. 2014.
[43] J. Meng, K. Kawakami, and A. K. Coskun, "Optimizing energy efficiency of 3-D multicore systems with stacked DRAM under power and thermal constraints," in *Proc. 49th Annu. Design Autom. Conf. (DAC)*, 2012, pp. 648–655.
[44] A. K. Coskun, J. L. Ayala, D. Atienza, T. S. Rosing, and Y. Leblebici, "Dynamic thermal management in 3D multicore architectures," in *Proc. Design, Autom. Test Eur. Conf. Exhib.*, Apr. 2009, pp. 1410–1415.
[45] P. Budhathoki, A. Henschel, and I. A. M. Elfadel, "Thermal-driven 3D floorplanning using localized TSV placement," in *Proc. IEEE Int. Conf. IC Design Technol.*, May 2014, pp. 1–4.
[46] C. Yan and E. Salman, "Mono3D: Open source cell library for monolithic 3-D integrated circuits," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 65, no. 3, pp. 1075–1085, Mar. 2018.
[47] K. Skadron, M. R. Stan, W. Huang, S. Velusamy, K. Sankaranarayanan, and D. Tarjan, "Temperature-aware microarchitecture," *ACM SIGARCH Comput. Archit. News*, vol. 31, no. 2, pp. 2–13, 2003.
[48] *A Comparison of Intel's 32 nm and 22 nm Core i5 CPUs: Power, Voltage, Temperature, and Frequency*. Accessed: Oct. 12, 2021. [Online]. Available: http://blog.stuffedcow.net/2012/10/intel32nm-22nm-core-i5-comparison/
[49] J.-J. Horng *et al.*, "A 0.7 V resistive sensor with temperature/voltage detection function in 16 nm FinFET technologies," in *Proc. Symp. VLSI Circuits Dig. Tech. Papers*, Jun. 2014, pp. 1–2.
[50] T. Oshita, J. Shor, D. E. Duarte, A. Kornfeld, and D. Zilberman, "Compact BJT-based thermal sensor for processor applications in a 14 nm tri-Gate CMOS process," *IEEE J. Solid-State Circuits*, vol. 50, no. 3, pp. 799–807, Mar. 2015.

Krithika Dhananjay (Member, IEEE) received the B.Tech. degree in electronics and communication engineering from SASTRA University, Thanjavur, India, in 2011, and the M.S. and Ph.D. degrees in electrical engineering from Stony Brook University (SUNY), Stony Brook, NY, USA, in 2015 and 2022, respectively.

From 2011 to 2014, she worked at International Business Machines (IBM) Corporation, Bengaluru, India, as a Research and Development Engineer, where she was responsible for the physical design of common library logic blocks in 22-nm silicon-on-insulator (SOI) and 14-nm FINFET nodes. In 2015, she was with Marvell Semiconductors, Boise, ID, USA, as a Hardware Design Intern, where she was involved in custom clock tree design and timing closures for application-specific integrated circuits. She worked as a Research Assistant with the Brookhaven National Laboratory, Upton, NY, USA, from November 2015 to February 2018, where she custom-designed mixed-signal circuits for high-energy physics applications in cryogenic environments. She will be starting a job as a Senior Logic Design Engineer at Qualcomm Technologies, Boxborough, MA, USA, in September 2022. Her broad research interests include the design and analysis of circuits for energy-efficient and hardware secure computing applications.

Vasilis F. Pavlidis (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in electrical and computer engineering from the University of Rochester, Rochester, NY, USA, in 2003 and 2008, respectively.

He is currently an Associate Professor with the Department of Electronics and Computer Engineering, Aristotle University of Thessaloniki, Thessaloniki, Greece. He has 15 years of research experience in the areas of on-chip interconnect modeling and design and emerging technologies, such as 3-D integration and spintronics. He is the leading author of the book *Three-Dimensional Integrated Circuit Design* (First and Second Editions) and a contributor to the software tool, Manchester Thermal Analyzer.

Dr. Pavlidis offers editorial services for several IC design and VLSI journals.

Ayse K. Coskun (Senior Member, IEEE) received the M.S. and Ph.D. degrees in computer science and engineering from the University of California at San Diego, La Jolla, CA, USA, in 2006 and 2009, respectively.

She is currently a Professor with the Electrical and Computer Engineering Department, Boston University, Boston, MA, USA. Her research interests include energy and temperature awareness in computing systems, novel computer architectures, and management of cloud and HPC data centers.

Dr. Coskun was a recipient of the IEEE CEDA Ernest S. Kuh Early Career Award. She serves as the Deputy Editor-in-Chief for the IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS.

Emre Salman (Senior Member, IEEE) received the M.S. and Ph.D. degrees in electrical engineering from the University of Rochester, Rochester, NY, USA, in 2006 and 2009, respectively.

Since 2010, he has been with the Department of Electrical and Computer Engineering, Stony Brook University (SUNY), Stony Brook, NY, USA, where he is currently an Associate Professor. His broad research interests include analysis, modeling, and design methodologies for integrated circuits and VLSI systems with applications to low power and secure computing, the Internet of Things with energy harvesting, and implantable devices. He is the leading author of a comprehensive tutorial book *High Performance Integrated Circuit Design* (McGraw-Hill, 2012, Chinese translation, 2015).

Dr. Salman was a recipient of the National Science Foundation Faculty Early Career Development Award in 2013; the Outstanding Young Engineer Award from IEEE Long Island, NY, USA, in 2014; and the Technological Innovation Award from IEEE Region 1 in 2018. He received multiple outreach initiative awards from the IEEE Circuits and Systems Society. He served as the Chair for the VLSI Systems and Applications Technical Committee (VSA-TC) of the IEEE Circuits and Systems Society. He serves as an Americas Regional Editor for the *Journal of Circuits, Systems and Computers* and on the Editorial Board of IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTING and the organizational/technical committees of various IEEE and ACM conferences.