A Qualitative Security Analysis of a New Class of 3-D Integrated Crypto Co-processors

Jonathan Valamehr[1], Ted Huffmire[2], Cynthia Irvine[2], Ryan Kastner[3], Çetin Kaya Koç[1,4], Timothy Levin[2], and Timothy Sherwood[1]

[1] Univ. of California, Santa Barbara
{valamehr,koc,sherwood}@cs.ucsb.edu
[2] Naval Postgraduate School
{tdhuffmi,irvine,levin}@nps.edu
[3] Univ. of California, San Diego
kastner@cs.ucsd.edu
[4] Istanbul Şehir University

Abstract

3-D integration presents many new opportunities for architects and embedded systems designers. However, 3-D integration has not yet been explored by the cryptographic hardware community. Traditionally, crypto co-processors have been implemented as a separate die or by utilizing one or more cores in a chip multiprocessor. These methods have their drawbacks and limitations in terms of tamper-resistance, side-channel immunity and performance. In this work we propose a new class of co-processors that are "snapped-on" to the main processor using 3-D integration, and we investigate their security ramifications. These 3-D co-processors hold many advantages over previous implementations. This paper begins with an overview of 3-D integration and its prior applications. We then outline security threat models relevant to crypto co-processors and discuss the advantages and disadvantages of using a dedicated 3-D crypto co-processor compared to traditional, commodity, off-chip crypto co-processors. We also discuss the performance improvements that can be gained from using a 3-D approach.

1 Introduction

For many systems that require strong guarantees on the integrity and secure transfer of their data, cryptography provides ample protection. For example, servers use cryptography to transform their data into presumably unreadable formats before being transmitted through a network. However, not all organizations need the same level of protection, and the requirements of a security system that are capable of protecting against a state-sponsored attack are quite different than those needed to protect against amateurs. As the necessity for secure communication and computation increases, more and more powerful and exotic operations are needed. No single chip design will ever simultaneously satisfy both the cost needs of the mass market and the cryptographic demands of the most security-conscious users.

Off-chip hardware solutions have the performance benefits associated with dedicated crypto hardware, and allow co-processors to be designed with specialized physical properties (such as tamper-resistance) not possible with other approaches. However, even well thought-out high performance hardware cryptographic solutions are plagued by attacks that compromise the confidentiality of sensitive information such as the secret keys used in cryptographic algorithms [2]. One of the biggest problems in designing such a system is balancing security and performance.

We propose a novel method to combat the high-throughput needs of modern day cryptographic co-processors by leveraging 3-D integration, a technology that allows vertical stacking of multiple dies to compose a single chip. These separate dies are connected to each other through very short, very fine-pitch vias that travel through the bulk substrate of the chip, creating an incredibly high-speed interface between the two dies. 3-D integration can provide a framework for establishing a high-bandwidth channel of communication between a main processor and a cryptographic co-processor, to achieve gigabit performance of cryptographic algorithms. An additional benefit of the 3-D integration techniques and our proposal to place the cryptographic co-processor on a 3-D layer is that we can also address certain high-assurance requirements. For critical applications where a security compromise cannot be tolerated, for example, satellite communications, military or highly sensitive applications, we need cryptographic functionality beyond commodity crypto (such as the Intel AES instruction set) and much higher levels of assurance about the secrecy of the keys and the data. The National Security Agency's Suite B cryptography specification is a prime example [31, 32]. By implementing the cryptographic functionality in a separate plane from non-security hardware functions, we can offer both a larger set of cryptographic functions and higher levels of protection that will never be realized in a commercial product.

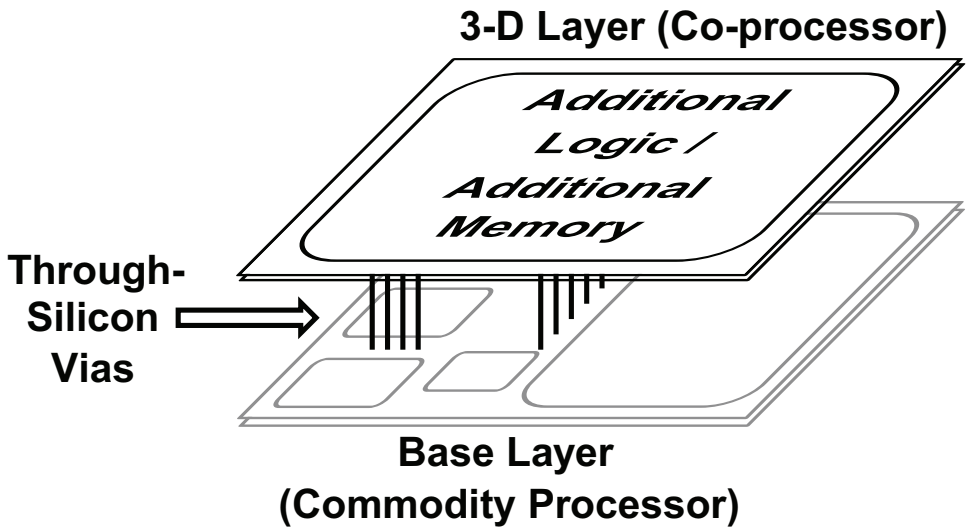

Fig. 1: This figure outlines the general architecture of 3-D integrated circuits, with multiple layers being connected with Through-Silicon Vias (TSVs). Almost all applications of 3-D chips have gravitated toward using the 3-D layer for additional logic space and full system-on-chip implementations, or using the 3-D layer to stack extra levels of cache or main memory.

While 3-D integration easily promotes high performance computing, it has the added benefit of protecting cryptographic processes and keys from malicious processes in the same system. In this paper we are the first to propose using 3-D integration to include a cryptographic co-processor on a single chip to address the growing performance and security needs of the industry (discussed in the remainder of this section), and set out to examine 3-D integration and its susceptibility to many popular attacks ranging from physical tamper to side-channel attacks. We then describe 3-D integration and its current and proposed applications. We follow with a summary of current hardware security attacks against information integrity, with qualitative analysis of each attack's threat to a system fabricated using 3-D integration. We also provide a brief discussion of the performance enhancements that can be gained from a 3-D approach.

1.1 Industry Motivations

In the past, cryptographic co-processors were used in military applications such as secure communication links. However, the proliferation of Automated Teller Machines (ATMs) in the '80s introduced them to commercial applications. Today many popular consumer devices have cryptographic processors in them, for example, smart-cards for pay-TV access machines and mobile phones, lottery ticket vending machines, and various electronic payment systems. The main reason for their use in such applications is that they hold decryption keys and provide *tamper-resistant hardware*. There was very little need for high performance (throughput) in such systems, and their most important function is tamper-resistance, i.e., the protection of the cryptographic keys from physical attacks [38].

However, the evolution of network security requirements in the '90s increased attention on performance. Cryptographic co-processors are expected to protect Secure Sockets Layer (SSL) keys used in web servers and provide the performance needed by several thousands of simultaneous network connections. Network security chip designs used in SSL boxes, enterprise VPN/firewall appliances and IPsec routers are primarily driven by three factors: silicon integration trends, speed, and security features. The integration trend actually started at the low end, i.e., embedded processors with cryptographic acceleration were used in relatively low-speed connections. Since about 2005, the next step in silicon integration arrived as "integrated cryptography processors"; they combine a CPU with memory and I/O subsystems, and gigabit encryption engines on a single die [44]. It is clear that while minimizing the complexity of cryptographic functionality is an important area of intellectual pursuit, in practice high performance is achieved by interfacing with the data in a fast and efficient manner. These highly integrated security processors include hardware blocks that accelerate packet processing, compression, and content inspection.

These gigabit-class cryptographic co-processors coupled with the "commoditization or commercialization of cryptography," i.e., fixing and accelerating the deployment of a subset (for example, RSA, RC4, AES, MD5, and SHA-1 used in SSL) of cryptographic algorithms for mainstream e-commerce, are the current industrial trends.

Since performance is the main objective, higher levels of integration are useful. The rapid evolution of emerging security applications (e.g., intrusion prevention, application-level firewalls, and anti-spam) will present challenges because such applications require inspection of Layer 7 content. Hardware integration must significantly increase in order to meet such challenges. Specifically, multiple CPUs will have to interoperate to be integrated with the gigabit cryptographic engines.

1.2 Security Considerations

While this new class of high performance cryptographic co-processors is needed to protect the confidentiality of information transmitted between computers and is designed to be resistant to attacks against the ciphers, side-channel attacks, which threaten the implementation of the cryptographic algorithm, are often exploited. Initially devised as a method of attack on cryptographic keys inside smart cards used in credit cards, side-channel attacks are now clearly understood to be applicable to computer systems. Smart cards do not have their own power source, and their architecture is quite simple; thus, they are an easy target for side-channel attacks. An adversary capable of (even passively) observing some of its physical and electrical properties (e.g., timing and instantaneous power) can learn significant portions of the secret keys. The security community did not believe these attacks could be applied to general computer systems, but a timing attack on a Web server [12] changed this perspective. Researchers showed that such an attack could compromise remote systems over a network, which is very different from performing side-channel attacks on smart cards that are in the attacker's possession. Improvements to the original remote timing attack made it even more practical [3].

More recent work on side-channel analysis has established a new field, micro-architectural analysis, which studies the effects of common processor components on system security [4]. Microprocessor components generate easily observable, data-dependent effects; a crypto algorithm's execution, for example, leaves "footprints" on the persistent state of data caches, instruction caches, and branch prediction units. These easy-to-see footprints depend on the operations performed during execution as well the data used in them, so an adversary could break a cryptosystem simply by running in parallel a so-called spy process to trace the footprints during or after the algorithm's execution. It is important to note that although spy processes run in full isolation and cannot directly read any data from the cryptosystem, leaked footprints can lead to dramatic security breaches.

In addition, as cloud computing and virtualization techniques bring processes of diverse trustworthiness together, micro-architectural and other types of side-channel attacks constitute serious threats. Therefore, we must design integrated cryptographic co-processors that operate in isolation from the processes running on the CPUs. To meet these security and performance needs, we propose using 3-D integration to develop crypto co-processors that can be attached to a main processor in a modular yet isolated fashion. Since cryptographic co-processors are subject to a wide variety of attacks, ranging from those that require physical access to the machine to those that can be performed remotely, it is important to investigate the feasibility of each of these attacks on a 3-D platform. In the following section we provide an introduction to 3-D integration, and follow with a discussion of each proven security threat to cryptographic co-processors and analyze the susceptibility of a 3-D co-processor solution to these attacks.

2 3-D Integration

3-D integration is a relatively new IC manufacturing technology that allows several layers of silicon to be vertically stacked to form a single chip. This provides many opportunities for system designers, as a chip can have several simultaneous active layers of computation, as opposed to traditional "2-D" chips that have one active layer of silicon. 3-D interconnect is one of a number of different competing technologies, including chip-bonding, Multi-chip Modules (MCM) [27], chip-stacking with vias [8, 13], or even wireless superconnect [29]. While chip-bonding and MCM technology are already used in a variety of embedded contexts [1, 7], more aggressive interconnect technologies are being heavily researched by several major industrial consortia. One of the more promising options is to connect separate layers to each other through 3-D integration by use of high-speed Through-Silicon Vias (TSVs). TSVs are very short, acting as a very high speed interconnect between the layers with a delay of only 12ps [26] when traveling through a 20-layer stack.

With current TSV pitches being under 10μm [25], a chip can support several thousands of TSVs between its layers, complementing the high speed with extremely high bus widths. The layers that make up the stack in a 3-D chip are each fabricated separately, and then joined using one of several techniques [10] discussed in the next section. Since each layer is printed on a different wafer, 3-D chips may include layers of mixed process technologies and

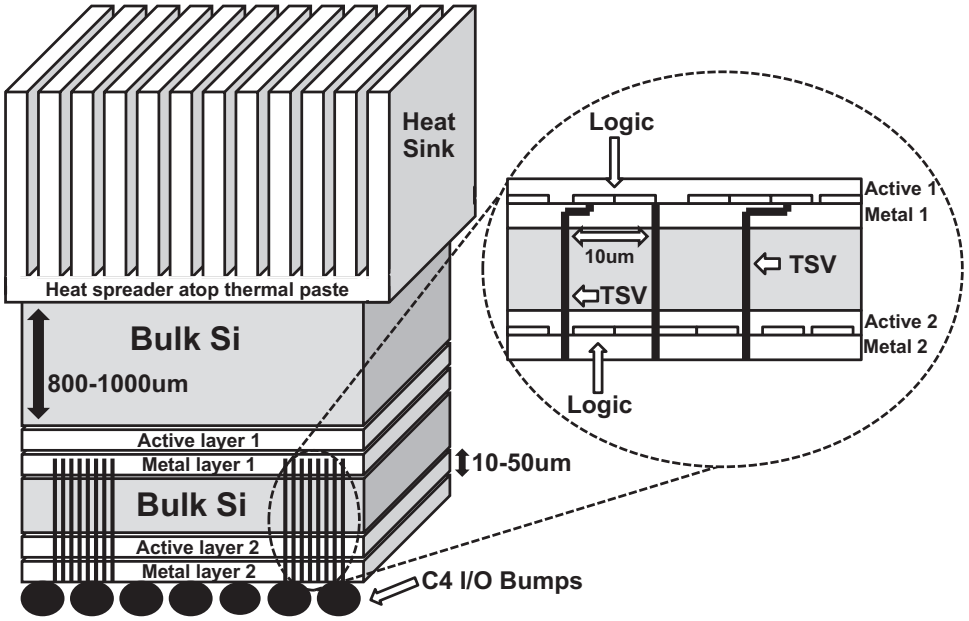

Fig. 2: A structural overview of a Face-to-Back 3-D configuration, complete with 2 separate metal layers and Through-Silicon vias (TSVs) traversing the bulk silicon to each die.

feature sizes. The main advantage of 3-D chips is the ability to exploit physical locality to shorten wire length, by utilizing the third dimension of placement. This allows designers to place circuits above other circuits, rather than being restricted to adjacent placements. Doing so allows system builders to place additional logic or resources such as cache directly above the area of the chip that needs that resource. Since TSVs are much shorter than cross-chip wires, 3-D chips benefit both from shortened latency and lower power consumption resulting from driving wires of decreased length.

2.1 3-D Manufacturing Techniques

3-D chips use new process technologies to make the bonding of several dies possible. To connect the separate dies in a 3-D chip, one of several bonding methods is used. One popular method is wafer-to-wafer bonding, where an entire wafer of homogenous dies is aligned and placed on top of another homogenous wafer containing the other dies that are to be stacked vertically, and the wafers are bonded before the individual dies are cut. This method is is very practical, as the alignment and bonding process is only performed once per wafer, rather than once per chip. However, wafer-to-wafer bonding usually results in a lower yield of working 3-D chips. Alternatively, die-to-die bonding can be performed, which does not suffer from the same yield issues but is more complex in nature and more difficult to complete.

In addition to yield, another area of concern surrounding 3-D integration lies in the thermal management of 3-D chips. Due to the close proximity of components on both layers, 3-D chips run at higher temperature densities [26] than their 2-D counterparts. Much research [35, 26, 28, 15, 18] has been conducted on thermal management for 3-D chips, and the consensus is that a more expensive cooling solution is required.

There are also several different die-stacking configurations available with 3-D integration. In face-to-face bonding, the active metal layers are bonded next to each other, with the additional metal layers on each side of the newly bonded active layers. In a face-to-face configuration, TSVs connect the joined dies to the external output pins. In face-to-back bonding, TSVs are used to connect the separate dies, and the lowest die retains its external I/O and power connections.

Wafer thinning is one manufacturing technique that is performed for improved electrical characteristics and physical construction of the 3-D chip [10]. Wafer thinning is performed by grinding off a large portion of the bulk Silicon to create a very thin die. While this sometimes damages the wafer, this is counteracted by chemically and mechanically polishing the wafer. With modern wafer thinning techniques, dies can be reduced from above 300μm to between 10μm and 50μm.

2.2 Applications of 3-D Integration

Many different uses of 3-D integration have been proposed, from stacking additional memory or extra levels of cache [10, 34, 43, 24, 47, 20, 19] to stacking multiple processors [6]. These two examples exploit the full advantages of 3-D chips, as attaching additional memory can provide lower latency compared to off-chip memory, and power can be saved because driving TSVs requires less power than long off-chip wires. One example of how 3-D integration has already been used in the commercial market is Toshiba's Chip Scale Camera Module (CSCM), which is a CMOS image sensor module that is able to leverage TSVs to satisfy high-speed I/O requirements [46] while realizing a significant reduction in chip size. Additionally, power consumption can be lowered through this approach, as long off-chip wires no longer have to be driven to communicate between the main processor and an off-chip sensor module.

3-D integration has been proposed for the development of a modular "snap-on" layer, that can be optionally placed on some chips requiring extra functionality while being omitted from other chips, specifically, an optional introspective 3-D layer for program analysis and performance profiling [30]. One major finding of this work was that less than a 2% increase in area on the base active layer is required to compensate for the TSVs needed for the introspection layer. The modular property exhibited by this architecture enables designers to create processors that optionally include a layer of logic when the consumer's application needs it, but omit the layer from systems when the consumer does not require this extra functionality. In particular, we propose the use of the optimal layer to support cryptographic functions.

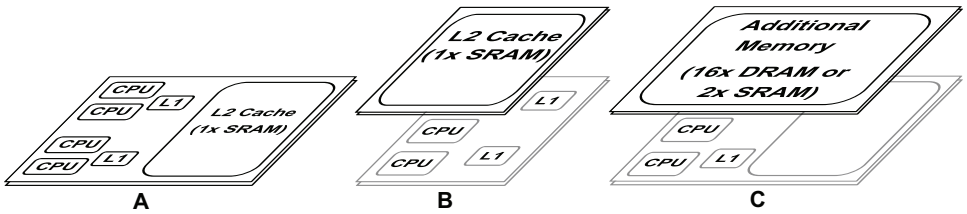

Fig. 3: Example memory-on-logic applications of 3-D integration. Figure A shows a baseline processor with an integrated L2 cache. Figure B shows how the footprint of the same chip can be decreased, while increasing cache performance by exploiting the physical locality of the cache and high-speed TSV interconnect. Figure C shows another configuration that places additional cache memory on the 3-D layer to enhance performance and lower cache miss rates.

2.3 Our Proposed 3-D Integrated Crypto Co-processor

For our analysis, we propose using a modular 3-D layer to act as a crypto co-processor, and additionally be able to safeguard against certain types of security threats. This design will also have dedicated memory in the 3-D layer (Figure 4) that will contain classified cryptographic state and keys during computation. This design will be similar to crypto co-processors [45] that have been proposed in the past, that are able to perform several standard crypto algorithms and support different key sizes. The next section outlines many threats and attacks that are associated with secure hardware implementations and crypto co-processors, and analyzes each threat and its effectiveness on a 3-D integrated crypto co-processor.

3 Security Ramifications

A challenge in the design of cryptographic hardware is to guard against various security threats, including explicit and implicit channels of information leakage. The traditional off-chip on-board model of cryptographic co-processors has the advantage of its optional use in a system. However, it is still prone to certain dangers. This section discusses the security threats faced by crypto hardware designers and provides an analysis of whether a 3-D integrated crypto co-processor alleviates such attacks. We consider the effects of integrating the crypto co-processor using 3-D integration, as well as the effects of the possible security measures that can be implemented on a co-processor, regardless of its location (whether it be on-chip, off-chip or as a 3-D plane). We base our comparison on a baseline

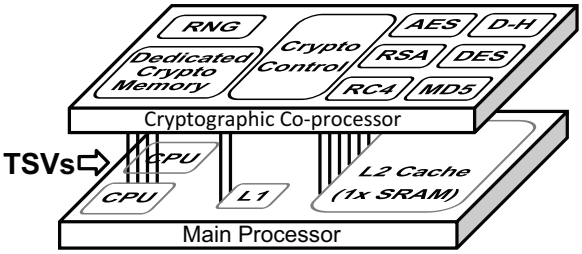

Fig. 4: Our proposed 3-D cryptographic co-processor, complete with dedicated memory for crypto keys and state.

off-chip crypto co-processor, and a novel 3-D integrated crypto co-processor (Figure 4) discussed earlier that has memory on the 3-D layer to hold keys and cryptographic state. An on-chip solution is not compared, as the modular nature of a 3-D co-processor and an off-chip solution both allow us to meet high-assurance security needs with few exceptions (e.g., fab milling and image capture). We review the following threats: physical tamper, memory remanence, access-driven cache side channels, time-driven cache side channels, fault analysis, electromagnetic analysis, power analysis, and thermal analysis. In general, all of these attacks can be mitigated with a 3-D crypto solution, although in several cases the rationale is one of impracticability as opposed to impossibility.

3.1 Physical Tamper

A certain class of security vulnerabilities and attacks is performed physically, with the device in possession or within reach of a malicious user. This can include smart cards, personal computers, and servers, where cryptography is performed for secure information exchange.

Threat Model: A specific physical threat to crypto co-processors is pin and bus probing, to intercept the unencrypted traffic between a main processor and a crypto co-processor. This provides an explicit channel of information, compromising secret information with practically 0% error rate. This, unlike other physical-retrieval attacks (discussed in next section), is performed while the device is fully powered and operating.

3-D Co-processor Advantages/Disadvantages: A 3-D integrated crypto co-processor can circumvent these types of attacks, as the 2 layers of computation are bonded very tightly and have no exposed shared busses or I/O pins to read from. The TSVs that connect the two layers are completely enclosed in the chip, giving the 3-D crypto co-processor complete physical isolation from the outside world during powered operation.

3.2 Memory Remanence

Memory remanence threats are applicable to data that is stored in locations that assume protection from a malicious user or volatility upon the loss of power to the system. Memory remanence threats can be classified into one of two types, based on the nature of retrieval of the data.

Physical Retrieval: Access to data stored in internal portions of a chip has been achieved through physical probing in a number of ways [40]. Modern devices with decreasing feature sizes, however, make this technique more difficult. An advanced method of retrieving data from an internal portion of a chip is through the use of a Focused Ion Beam (FIB), which can use ions to mill a very small layer of a chip ($\sim$0.1μm), exposing nanoscale devices to image capture. Milling a chip and capturing images using an FIB can yield the data stored in a chip (usually nonvolatile memory), regardless of its external connection design. This can be used maliciously to read a memory unit that stores confidential information that is inaccessible by traditional physical tampering alone. FIB milling can also be used to expose new parts of the chip that are easily probed by means of physical connections.

Electrical Retrieval: While many believe DRAM is a volatile memory element, it has been shown that DRAM retains its contents for a few seconds after a system has been powered down. Furthermore, this volatility is dependent on temperature, as DRAM exposed to very low temperatures (-50°C and lower) is readable even after several minutes, with low error rates [17]. This presents a serious security threat, as DRAM can be moved to a different system to have its contents read, which may include sensitive data or secret crypto keys. In addition to this threat, data stored in SRAM has the characteristic of retaining its information when data has been stored for a prolonged amount of time.

Threat Model: Here we will discuss the threat model and successful attacks and demonstrations with each of these types of memory remanence.

Physical: Attempts at using a physical retrieval method have been successful, such as the full recovery of data from a damaged nonvolatile EEPROM memory module in a crashed aircraft, using an FIB technique [42, 23].

Electrical: Several attacks [17] have been discovered, where DRAM inside a system is either read on a warm boot or removed from a system and placed in another system to be read on a cold boot. Because of the memory remanence properties discussed, the full contents of the DRAM can be extracted, where keys or classified data may be stored. In one version of the attack, drive encryption methods such as Microsoft Bitlocker can be compromised. The secret key used to encrypt the contents of the hard drive of the computer is extracted from DRAM and is used to decrypt the sensitive data stored in the system hard drive.

3-D Co-processor Advantages/Disadvantages: Here we will discuss the advantages and disadvantages of each of these types of memory remanence with respect to a 3-D co-processor solution.

Physical: 3-D integration does not seem to help alleviate the threat of physical retrieval techniques such as FIB milling and image capture, as a 3-D chip may only add more material that needs to be milled before an exposure occurs. Simply fabricating a chip using 3-D integration does not enhance its ability to thwart this type of attack.

Electrical: With the framework we outlined earlier, all operations done on the 3-D crypto co-processor will have exclusive memory to store data. This allows keys and sensitive state to be stored in a non-shared, non-removable resource. Since the memory used in the 3-D layer will be embedded and only interface to the base computation plane through TSVs, this threat is mitigated.

3.3 Access-Driven Cache Side-Channel Attacks

In most modern day processors, resource sharing is used to increase the throughput of the system by exploiting instruction-level or thread-level parallelism. Unfortunately, with the increase in performance comes vulnerabilities in the form of side-channel attacks. One attack uses the cache access patterns of cryptography software to extract portions of the secret key, until the whole key can be constructed.

Threat Model: This threat was made evident when an attack [33] on an implementation of the RSA encryption standard was successfully launched. The attack used shared cache memory inside a processor employing simultaneous multithreading to view the process-to-process interference to the cache. Cache line evictions dictate which lines are being accessed, which allows a malicious thread of code to extract the cache access patterns of a victim thread. The attack works on this premise, and was achieved by a malicious thread accessing enough data to occupy sufficient space in the cache, so that when the victim thread were to access its own data, it would have to evict some cache lines placed in the cache by the malicious code. On subsequent accesses to the cache, the malicious thread can observe which lines had been evicted by the victim thread simply by measuring the variable access times of each cache access. This is enough information to infer parts of the cryptographic key due of the nature of look-up tables used by some cryptographic algorithms. The whole key can eventually be compromised with a low rate of error.

3-D Co-processor Advantages/Disadvantages: As stated previously, this threat is entirely made possible because of resource sharing. A 3-D crypto co-processor has the advantage of being fabricated with its own dedicated memory to store cryptographic state and secret keys during its operations. This would eliminate the vulnerability of this information to cache-sharing attacks.

3.4 Time-Driven Side-Channel Attacks

In addition to observing cache access patterns of crypto software, the running time of said software can be used as a side channel for sensitive information. Time-driven attacks on cryptosystems revolve around the underlying fact that most crypto software has a variable execution time, due to many factors including architectural optimizations such as cache and branch predictors. This variable execution time is dependent on the inputs, allowing one to use this difference in execution time to aid in the retrieval of a secret key.

Threat Model: Many timing attacks have been successfully demonstrated. This concept of a time-driven implicit channel of information was introduced when Kocher [22] showed that key retrieval was possible through measurements of crypto execution time. The work successfully demonstrated this type of attack on several different crypto

algorithms including RSA and Diffie-Hellman. Another attack [9] was able to recover a full key from an AES implementation. Unlike the access-driven security threat, this attack can even be done remotely by merely invoking a crypto operation on another machine and measuring the varying execution time.

3-D Co-processor Advantages/Disadvantages: In order to relieve susceptibility to this threat, each crypto operation must be uniform in execution time. Some hardware crypto solutions such as the proposed Intel AES instructions [16] can thwart this type of attack because each AES encryption/decryption instruction has a fixed execution time. While this is not exclusive to a 3-D setup, A 3-D hardware implementation can hold the same property, while reaping the added benefits of 3-D integration.

3.5 Fault Analysis

A certain class of attacks on crypto systems takes advantage of hardware faults, which are errors in the computation of a processor. Hardware faults can appear due to a variety of reasons (including effects of high temperature, radiation, excessive power, or faulty hardware) and can even be introduced (or "injected") into an otherwise fault-free system. When these faults occur during a cryptographic operation, they introduce a variability in the computation that can be measured when compared to the same operation executed without error.

Threat Model: The first attack to use the principle of exploiting faults in a system was introduced theoretically in 1997 [11], when it was proven that a hardware fault could lead to the compromise of encryption schemes such as RSA and Rabin. More recently, a successful attack [39] on AES was discovered which utilizes faults that are induced in the state matrix. The particular faults used were caused by clock glitching, by momentarily speeding up the clock rate fast enough to produce an erroneous value. Once these faults gave rise to variability in the state matrix, the inter-relation between the columns of the matrix can be used to reduce the key space. The attack was proven with an AES hardware implementation on an FPGA platform using a clock rate increase to produce the faults. Once this is performed and the key space is reduced, a brute force attack is used to recover the full key in under 7 minutes.

3-D Co-processor Advantages/Disadvantages: Theoretically, fault analysis attacks can be performed on a 3-D integrated crypto co-processor. However, the very nature of fault analysis attacks relies on either a random hardware fault occurring, or a fault being injected into the hardware. Waiting for and detecting a random fault on any crypto co-processor seems to be very unlikely, and infeasible based on how often common AES implementions switch their secret keys. Fault injection attacks depend on the ability to inject the faults in the first place. However, to the best of our knowledge, a reliable method to inject a fault into a high-performance microprocessor does not exist. Given this fact, we suggest that a fault injection attack is not fully practical in a real world implementation on an ASIC.

3.6 Electromagnetic Analysis

Electromagnetic (EM) side-channel attacks have a long history and "folklore" associated with them. It is well-known and established that highly-sensitive antennas and sophisticated receivers can be used to capture data emanating from various equipment. Defense organizations use the codename "tempest" to refer to efforts to limit the leakage of data through EM channels. The first openly leakage of their data through EM channels. The first openly demonstrated EM attack on ICs and CPUs performing cryptographic computations was demonstrated in [36] and [14] in 2001. EM signals can be recorded and later analyzed by placing tiny antennas in close proximity to the chips and the boards being examined.

Threat Model: The early successful attacks are semi-invasive; they required the decapsulation of the chip packaging and careful placement of micro-antennas. More recent work [5] showed that EM attacks on CPUs and cryptographic chips were also possible at a distance (a couple of meters). Also, earlier work was more concentrated on direct emanations; it turns out such emanations from chips and boards are very hard to capture without invasive approaches. In reality, there are also unintentional emanations due to various electrical and electromagnetic couplings between components, depending on their proximity and geometry. These couplings manifest themselves as modulations of carrier signals generated, present or introduced within the device. If a modulated carrier can be captured, sensitive signals can be recovered by an EM receiver tuned to the carrier frequency. Experiments show that EM side-channels exist via Amplitude Modulation of a carrier signal. Similar to the other side-channels (particularly, power), the compromising EM signals can be extracted using AM demodulation, and provide details about the computation.

3-D Co-processor Advantages/Disadvantages: Since the 3-D will provide a much higher level of integration, bringing multiple CPUs, memory blocks (buffers, caches, registers, etc), and cryptographic engines together, we expect that the resulting EM signals will have higher levels of superimpositions. Most successful EM attacks were possible because compute-intensive cryptographic functions (such as modular exponentiations for RSA or point multiplication operations for elliptic curve cryptography) were dominating the entire device in terms of energy and time. However, in highly integrated 3-D systems multiple CPUs, memory blocks, and cryptographic engines will be competing for spectrum in terms of their signal strength. The resulting noisy channel would require very careful orchestration of cryptographic functions to extract a signal, reducing the likelihood of EM side-channel attacks on 3-D systems.

3.7 Power Analysis

Power analysis is by far the most successful form of side-channel attack. First, it is not invasive; a passive adversary collects and analyzes the instantaneous power usage traces of a cryptographic device, which is generally a smart card [21]. In the simplest form of the attack, the adversary collects and interprets the power traces, a technique called Simple Power Analysis (SPA). A more powerful and effective attack is Differential Power Analysis (DPA) in which power traces from multiple cryptographic operations are collected and statistically analyzed to gather information about intermediate results in a cryptographic computation. The practice of SPA and DPA over the last decade has shown that cryptographic keys (RSA, DES, etc.) can easily be compromised using power analysis [2].

Threat Model: However, power analysis is still very difficult to apply to computer systems for at least two reasons: 1) power traces are not generally available to a passive or remote adversary, 2) in a complex computer system, tens of processes run simultaneously and affect the instantaneous power. The resulting noisy channel makes it difficult to separate and analyze the signal.

3-D Co-processor Advantages/Disadvantages: As mentioned, power analysis is an attack that is usually launched on very simple hardware such as smart cards that use low-complexity hardware, whose power traces can easily be analyzed to provide useful information. However, a successful power analysis attack on a complex single-layer microprocessor has not been launched. Moreover, 3-D systems with their multiple CPU, memory, and cryptographic engines constitute an even more complex computer system than a single-layer system. Even if an adversary gains physical access to the device and measures and collects power traces, under normal operating modes, it will be very difficult to obtain meaningful data from this information. Similar to the EM channels, a careful orchestration of the processes may yield meaningful data (for example, to disable all other units except a cryptographic functional unit, and then collect power traces which would be dominated by this unit). Generally, this is very difficult to achieve.

3.8 Thermal Analysis

When high-performance processors are executing code, they expend energy in the form of heat all over the chip. Depending on the specific instructions being executed, workload size, execution time, and chip architecture, this can create hot spots, areas of the chip that are more active than others, and consequently at a higher temperature. Since a processor is executing a different program(s) with differing sets of inputs at any given time, the hot spots on the chip can vary, thus creating a "thermal profile" associated with the specific program and inputs. With high resolution thermal imaging capability, one can use these thermal profiles to analyze activity on the processor and infer information about what is being executed at any given time.

Threat Model: Thermal analysis is regarded as a theoretical attack; it is rarely used against cryptographic devices [37], as the diffusion of heat for processors is very limited and hard to measure accurately to launch a practical attack.

3-D Co-processor Advantages/Disadvantages: While in theory a 3-D co-processor may still be susceptible to thermal analysis attacks, it is unclear whether such an attack will be successful on a 3-D platform. The main processor coupled with the co-processor may introduce enough "thermal noise" to the profile to make a thermal analysis impractical.

4 Performance Ramifications

The locality of a 3-D crypto co-processor provides various performance and power benefits. In this section we outline the performance advantages of a 3-D crypto co-processor and quantify several relevant metrics including latency and clock speed of current competing co-processor approaches.

Co-Processor Architecture	Power	Latency	Bandwidth	I/O Resources
Off-Chip Co-processor	Power-hungry off-chip buses and an additional chip contribute to high power usage	Very large delay between off-chip co-processor and CPU (>200 cycles)	Data bus widths are limited (1- 8 bytes), at low external clock rates (~ 400 MHz)	I/O pins on the main processor need to be allocated to communicate with co-processor
3-D Integrated Co-processor	3-D only increases power usage by the addition of extra logic on an active layer, driving short TSVs	The latency of a TSV traversing a 20-layer stack is only 12ps (< 1 cycle delay)	3-D can accommodate large bus widths (up to 128 bytes), at core clock speed (>2 GHz)	I/O pin availability is unaffected, as TSVs are used as interconnect between the active layers

Fig. 5: This table compares traditional crypto co-processors and 3-D crypto co-processors, showing the advantages and disadvantages in terms of power, bandwidth, and delay. [26, 41].

In general, implementing security features in software is inexpensive and slow, compared to hardware solutions. However, these performance benefits differ greatly between different hardware solutions. An off-chip co-processor has one main crippling disadvantage for its performance: power-hungry, high latency buses running at much lower frequencies than the TSVs used in a 3-D chip. Long off-chip buses suffer from increased power usage, as driving such buses consumes much more power than driving short inter-die vias in a 3-D configuration. Also, these buses must run at decreased clock speeds (Figure 5) to compensate for their increased length. This increased length, in turn, introduces delays in the critical path of a cryptographic co-processor.

Adding to the low performance of an off-chip co-processor is the amount of available pins that may be used to interface with a main processor, as this is usually subtracted from the main I/O pins, which are very limited in quantity. A lower amount of pins means smaller bus widths; 3-D chips have the advantage of utilizing very fine-pitch inter-die vias and creating extremely high bandwidth buses between the dies. In fact, with chip-to-package I/O connections currently at a pitch of 500μm and current TSVs at a pitch of 5μm, in one "pin" worth of space you can fit 10,000 TSVs – indicating the inherent advantage of the 3-D approach.

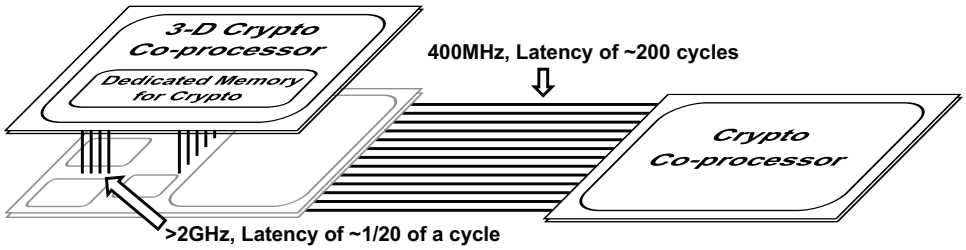

Fig. 6: A comparison of the different latency and delay characteristics of a 3-D crypto co-processor vs. an off-chip crypto co-processor.

5 Conclusions

In this paper we proposed a novel method of optionally including a cryptographic co-processor with a commodity microprocessor using 3-D integration, to meet the performance and security needs of government and industry, as well as to provide functionality beyond what has been proposed in "commodity crypto" hardware. We are the

first to propose using 3-D integration to meet these needs as well as mitigate several types of attacks to which traditional co-processor solutions have been vulnerable in the past. We outline a wide range of security threats, and analyze how a 3-D crypto co-processor mitigates these attacks. We find that a 3-D crypto co-processor can mitigate some types of board-level pin and bus probing attacks, memory remanence attacks, access-driven cache side channel attacks, time-driven side channel attacks, electromagnetic analysis attacks, power analysis attacks, and thermal attacks. We also outline the performance benefits that can be achieved from using 3-D integration. In the future, we hope that our work inspires new work on using 3-D integration for novel security purposes.

References

1. Cristinel Ababei, Yan Feng, Brent Goplen, Hushrav Mogal, Tianpei Zhang, Kia Bazargan, and Sachin Sapatnekar. Placement and Routing in 3D Integrated Circuits. *IEEE Design and Test of Computers*, 22(6):520–531, Nov/Dec 2005.
2. O. Acıiçmez, J.P. Seifert, and C.K. Koc. Micro-architectural cryptanalysis. *IEEE Security and Privacy Magazine*, 5(4), July-August 2007.
3. O. Aciicmez, W. Schindler, and Ç. K. Koç. Improving Brumley and Boneh timing attack on unprotected SSL implementations. In *Proceedings of the 12th ACM Conference on Computer and Communications Security*, pages 139–146, November 2005.
4. O. Aciicmez, J. P. Seifert, and Ç. K. Koç. Micro-architectural cryptanalysis. *IEEE Security & Privacy*, 5(4):62–64, July/August 2007.
5. D. Agrawal, B. Archambeault, J. R. Rao, and P. Rohatgi. The EM Side-Channel(s). In *Proceedings of the Workshop on Cryptographic Hardware and Embedded Systems (CHES)*, volume 2523, pages 29–45, August 2002.
6. A. Akturk, N. Goldsman, and G. Metze. Self-Consistent Modeling of Heating and MOSFET Performance in 3-D Integrated Circuits. *IEEE Transactions on Electron Devices*, 52(11):2395–2403, 2005.
7. Kaustav Banerjee, Shukri J. Souri, Pawan Kapur, and Krishna C. Saraswat. 3-D ICs: A Novel Chip Design for Improving Deep Submicron Interconnect Performance and Systems-on-Chip Integration. *Proceedings of the IEEE*, 89(5):602–633, May 2001.
8. Benkart et al. 3D Chip Stack Technology Using Through-Chip Interconnects. *IEEE Design and Test of Computers*, 22(6):512–518, Nov/Dec 2005.
9. Daniel J. Bernstein. Cache-timing attacks on AES. `http://cr.yp.to/antiforgery/cachetiming-20050414.pdf`, April 2005. Revised version of earlier 2004-11 version.
10. Bryan Black, Murali Annavaram, Ned Brekelbaum, John DeVale, Lei Jiang, Gabriel H. Loh, Don McCauley, Pat Morrow, Donald W. Nelson, Daniel Pantuso, Paul Reed, Jeff Rupley, Sadasivan Shankar, John Shen, and Clair Webb. Die Stacking (3D) Microarchitecture. *Proceedings of the 39th Annual IEEE/ACM International Symposium on Microarchitecture*, pages 469–479, December 2006.
11. Dan Boneh, Richard A. Demillo, and Richard J. Lipton. On the Importance of Checking Cryptographic Protocols for Faults. In *Proceedings of Advances in Cryptology - Eurocrypt*, pages 37–51, 1997.
12. D. Brumley and D. Boneh. Remote Timing Attacks Are Practical. In *Proceedings of the 12th USENIX Security Symposium*, 2003.
13. W.R. Davis, J. Wilson, S. Mick, J. Xu, H. Hua, C. Mineo, A.M. Sule, M. Steer, and P.D. Franzon. Demystifying 3D ICs: The Pros and Cons of Going Vertical. *IEEE Design and Test of Computers*, 22(6):498–510, Nov/Dec 2005.
14. K. Gandolfi, C. Mourtel, and F. Olivier. Electromagnetic analysis: Concrete results. In *Proceedings of the Workshop on Cryptographic Hardware and Embedded Systems (CHES)*, volume 2162, pages 251–261, August 2001.
15. Moishe Groger, Shadi M. Harb, Devin Morris, William R. Eisenstadt, and Sudeep Puligundla. High Speed I/O and Thermal Effect Characterization of 3D Stacked ICs. In *Proceedings of the IEEE International Conference on 3D System Integration (3D IC)*, pages 1–5, September 2009.
16. Shay Gueron. White paper: Advanced encryption standard (AES) instructions set, Intel corporation, July 2008.
17. J. Alex Halderman, Seth D. Schoen, Nadia Heninger, William Clarkson, William Paul, Joseph A. Calandrino, Ariel J. Feldman, Jacob Appelbaum, and Edward W. Felten. Lest we remember: Cold-boot attacks on encryption keys. In *Proceedings of the USENIX Security Symposium (Sec '08)*, June 2008.
18. Brent Hollosi, Tao Zhang, Ravi S. P. Nair, Yuan Xie, Jia Di, and Scott Smith. Investigation and Comparison of Thermal Distribution in Synchronous and Asynchronous 3D ICs. In *Proceedings of the IEEE International Conference on 3D System Integration (3D IC)*, pages 1–5, September 2009.
19. Philip Jacob, Okan Erdogan, Aamir Zia, Paul M. Belemjian, Russell P. Kraft, and John F. McDonald. Predicting the performance of a 3D processor-memory chip stack. *IEEE Design and Test of Computers*, 22(6):540–547, Nov/Dec 2005.
20. Michael B. Kleiner, Stefan A. Kühn, and Werner Weber. Performance Improvement of the Memory Hierarchy of RISC Systems by Applications of 3-D Technology. In *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 2305–2308, 1995.
21. P. Kocher, J. Jaffe, and B. Jun. Differential power analysis. In *Proceedings of Advances in Cryptology CRYPTO99*, volume 1666, page 388397, August 1999.
22. Paul C. Kocher. Timing attacks on implementations of diffie-hellman, rsa, dss, and other systems. In *Advances in Cryptology - CRYPTO 96*, pages 104–113. Springer-Verlag, 1996.
23. Jean Kumagai. Chip detectives. *IEEE Spectrum*, 37(11):43, November 2000.
24. Christianto C. Liu, Ilya Ganusov, Martin Burtscher, and Sandip Tiwari. Bridging the Processor-Memory Performance Gap with 3D IC Technology. *IEEE Design and Test*, 22(6):556–564, 2005.

25. Gabriel H. Loh. 3D-Stacked Memory Architectures for Multi-Core Processors. In *Proceedings of the 35th Annual International Symposium on Computer Architecture (ISCA)*, pages 453–464, June 2008.
26. Gian Luca Loi, Banit Agrawal, Navin Srivastava, Sheng-Chih Lin, Timothy Sherwood, and Kaustav Banerjee. A Thermally-Aware Performance Analysis of Vertically Integrated (3-D) Processor-Memory Hierarchy. In *Proceedings of the 43nd Design Automation Conference (DAC)*, June 2006.
27. Claude Massit and Nicolas Gerard. Three-dimensional multichip module. United State Patent, US 5373189, December 1994.
28. Keiji Matsumoto and Yoichi Taira. Thermal resistance measurements of interconnections and modeling of thermal conduction path, for the investigation of the thermal resistance of a three- dimensional (3D) chip stack. In *Proceedings of the 13th IEEE International Symposium on Consumer Electronics (ISCE 2009)*, pages 598–602, July 2009.
29. Miura et al. A 195Gb/s 1.2W 3D-Stacked Inductive Inter-Chip Wireless Superconnect with Transmit Power Control Scheme. In *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, pages 264–265, Feb 2005.
30. S. Mysore, B. Agrawal, S.C. Lin, N. Srivastava, K. Banerjee, and T. Sherwood. Introspective 3-D chips. In *Proceedings of the 12th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, San Jose, CA, October 2006.
31. National Security Agency (NSA). NSA Suite B Cryptography. `http://www.nsa.gov/ia/programs/suiteb_cryptography`.
32. National Institute of Standards and Technology (NIST). Suite B Implementer's Guide to NIST SP 800-56A. *CryptoBytes, RSA Laboratories*, 4(1):6–10, July 2009.
33. Colin Percival. Cache missing for fun and profit. In *Proceedings of the Technical BSD Conference (BSDCan 2005)*, Ottowa, Canada, May 2005.
34. Kiran Puttaswamy and Gabriel H. Loh. Implementing Caches in a 3D Technology for High Performance Processors. In *IEEE International Conference on Computer Design (ICCD) 2006*, pages 525–532, October 2005.
35. Kiran Puttaswamy and Gabriel H. Loh. Thermal analysis of a 3D die-stacked high-performance microprocessor. *Proceedings of the 16th ACM Great Lakes symposium on VLSI*, pages 19–24, May 2006.
36. J.-J. Quisquater and D. Samyde. Electromagnetic analysis (EMA): Measures and countermeasures for smart cards. In *Proceedings of the International Conference on Research in Smart Cards (e-Smart 2001)*, volume 2140, pages 200–210, 2001.
37. J-J. Quisquater and D. Samyde. Side Channel Cryptanalysis. In *Proceedings of the Workshop on the Security of Communications on the Internet (SECI)*, pages 179–184, September 2002.
38. Dhiman Saha, Debdeep Mukhopadhyay, and Dipanwita RoyChowdhury. Cryptographic processors - a survey. In *Proceedings of the IEEE*, volume 94(2), pages 357–369, 2006.
39. Dhiman Saha, Debdeep Mukhopadhyay, and Dipanwita RoyChowdhury. A Diagonal Fault Attack on the Advanced Encryption Standard. In *Cryptology ePrint Archive*, volume 581, 2009.
40. Jerry M. Soden and Richard E. Anderson. IC failure analysis: Techniques and tools for quality and reliability improvement. *Microelectronics and Reliability*, 35(3):429–453, 1995.
41. Hongbin Sun, Jibang Liu, Rakesh S. Anigundi, Nanning Zheng, Jian-Qiang Lu, Kenneth Rose, and Tong Zhang. 3D DRAM design and application to 3D multicore systems. *IEEE Design and Test of Computers*, 26(5), September 2009.
42. Pauline Tam. Ottawa firm rescues data from Swissair black box. *The Ottawa Citizen*, March 21, 2000.
43. Yuh-Fang Tsai, Yuan Xie, N. Vijaykrishnan, and Mary Jane Irwin. Three-Dimensional Cache Design Exploration Using 3DCacti. In *IEEE International Conference on Computer Design*. IEEE, October 2005.
44. B. Wheeler and L. Gwennap. A guide to security processors and accelerators. *Technical Report, The Linley Group*, June 2008.
45. Lisa Wu, Chris Weaver, and Todd Austin. CryptoManiac: A Fast Flexible Architecture for Secure Communication. In *Proceedings of the 28th Annual International Symposium on Computer Architecture (ISCA)*, pages 110–119, June-July 2001.
46. Hiroshi Yoshikawa, Atsuko Kawasaki, Tomoaki Iizuka, Yasushi Nishimura, Kazumasa Tanida, Kazutaka Akiyama, Masahiro Sekiguchi, Mie Matsuo, Satoru Fukuchi, and Katsutomu Takahashi. Chip scale camera module (CSCM) using through-silicon-via (TSV). In *Proceedings of the International Solid-State Circuits Conference (ISSCC)*, San Francisco, CA, February 2009.
47. Annie Zeng, James Lu, Kenneth Rose, and Ronald J. Gutmann. First-Order Performance Prediction of Cache Memory with Wafer-Level 3D Integration. *IEEE Design and Test of Computers*, 22(6):548–555, Nov/Dec 2005.