Security versus Energy Tradeoffs in Host-Based Mobile Malware Detection

Jeffrey Bickford
Department of Computer Science
Rutgers University
jbickfrd@cs.rutgers.edu

H. Andrés Lagar-Cavilla
AT&T Labs – Research
Florham Park, NJ
andres@research.att.com

Alexander Varshavsky
AT&T Labs – Research
Florham Park, NJ
varshavsky@research.att.com

Vinod Ganapathy
Department of Computer Science
Rutgers University
vinodg@cs.rutgers.edu

Liviu Iftode
Department of Computer Science
Rutgers University
iftode@cs.rutgers.edu

ABSTRACT

The rapid growth of mobile malware necessitates the presence of robust malware detectors on mobile devices. However, running malware detectors on mobile devices may drain their battery, causing users to disable these protection mechanisms to save power. This paper studies the security versus energy tradeoffs for a particularly challenging class of malware detectors, namely rootkit detectors. We investigate the security versus energy tradeoffs along two axes: attack surface and malware scanning frequency, for both code and data based rootkit detectors. Our findings, based on a real implementation on a mobile handheld device, reveal that protecting against code-driven attacks is relatively cheap, while protecting against all data-driven attacks is prohibitively expensive. Based on our findings, we determine a sweet spot in the security versus energy tradeoff, called the balanced profile, which protects a mobile device against a vast majority of known attacks, while consuming a limited amount of extra battery power.

Categories and Subject Descriptors. C.5.3 **[Computer System Implementation]**: Microcomputers—*Portable devices (e.g., laptops, personal digital assitants);* D.4.6 **[Operating Systems]**: Security and Protection—*Invasive software (e.g., viruses, worms, Trojan horses)*

General Terms. Experimentation, Measurement, Security

Keywords. Mobile malware, rootkits, security, energy

1. INTRODUCTION

We have come to rely on mobile devices as an integral part of our everyday lives. We entrust our smartphones, netbooks and laptops with personal information, such as email, friend lists, current location, and passwords to online banking websites. The future holds an even greater role for mobile devices, *e.g.,* as interfaces for wireless payments [8] or smart home control [6]. Mobile devices are thus swiftly becoming prized bounties for malicious entities: while the quantity and diversity of mobile malware available today pales in comparison with malware available for desktops, the incentives available to attackers point to a large and thriving future underground economy based on infected mobile devices. This has motivated recent research on both attacks against and defenses for mobile devices [14, 15, 19, 21, 28, 34, 38, 41].

The main goal of this paper is to study how the energy-constrained nature of mobile devices impacts their ability to run malware detection tools. Conventional wisdom holds that executing malware detectors on resource-constrained mobile devices will drain their battery [36], causing users to disable malware detection to extend battery life, and in turn exposing them to greater risk of infection. We present a framework to quantify the degree of security being traded off when prolonging battery life, and the ways in which such tradeoffs can be implemented. Specifically, we study security tradeoffs along two axes: (1) the surface of attacks that the malware detector will cover, and (2) the frequency with which the malware detector will be invoked.

Some emerging proposals for malware detection have sought to sidestep the energy constraints that we formalize and quantify in this study using *offloaded architectures* [9, 18, 37, 41], in which the malware detector itself executes on a well-provisioned server and monitors mobile devices. Unfortunately, malware detection offload either incurs significant power expenditures [41] due to data upload, or has limited effectiveness because it is best suited to traditional signature-based scanning. Such signature scanning is easily defeated with encryption, polymorphism and other stealth techniques. For this reason, there is growing consensus that signature-based scanning must be supplemented with powerful host-based agents that, for example, employ behavior-based detection algorithms [17]. Host-based detectors execute on and share resources such as CPU time and battery with the host device, thereby making the *security versus energy tradeoff* germane to the design of such detectors.

In this paper, we focus on security versus energy tradeoffs for host-based *rootkit detection*. Rootkits are a class of malware that infect the code and data of the operating system (OS) kernel. By infecting the kernel itself, they gain control over the layer that is traditionally considered the trusted computing base (TCB) on most systems. Rootkits can therefore be used to evade user-space malware detectors (including most commercial solutions that employ

Funded in part by NSF grants CNS-0831268, CNS-0915394, CNS-0931992 and CNS-0952128, and by the US Army CERDEC. Part of this work was done during Bickford's internship at AT&T Labs – Research, Florham Park, NJ. Bickford is currently affiliated with both Rutgers University and the AT&T Security Research Center in New York, NY.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
MobiSys'11, June 28–July 1, 2011, Bethesda, Maryland, USA.
Copyright 2011 ACM 978-1-4503-0643-0/11/06 ...$10.00.

signature-based scanning). Further, rootkits enable other attacks by hiding malicious processes, and allow attackers to stealthily retain long-term control over infected devices. Recent work has argued that the increasing complexity of mobile device OSes offers a vast attack surface of code and data that makes rootkits a realistic threat [14]. Indeed, rootkits have recently been developed for iPhones [34] and Android phones [38]. As a consequence of the variety of ways in which a kernel can be exploited, and rootkit detection implemented, we show that rootkit detectors can be *modulated* to explore a rich space of configuration options. Varying these configurations allows us to explore, in a general manner, the tradeoff between the security provided by the detection agent and the energy consumption of the host.

We conduct our study by adapting two complementary rootkit detectors proposed in prior work, namely Patagonix [30] and Gibraltar [10,11], to work on a mobile phone-like platform. We measured their security guarantees and energy footprint under several configurations that varied the surface of attacks the detectors covered, and the frequency with which they performed checks. Patagonix offers protection against malicious code in the kernel, by checking the integrity of static code pages (kernel inclusive). Gibraltar offers protection against malicious data in the kernel, by scanning the kernel's data segment and ensuring that its data structures satisfy certain integrity properties, which are normally violated in rootkit-infected kernels. For both rootkit detectors, trading security for energy savings, *e.g.,* by reducing the attack surface monitored or by reducing the frequency of checks, introduces a window of time during which rootkits can infect the mobile device. We recognize this might open a new class of attacks in which malware exploits the periodic nature of the system. We aim to mitigate this through the use of randomization. In this paper, we focus only on detecting the infection and not on recovering the infected device.

We outline here the results of our study and the contributions of our paper:

- The energy impact of checking the integrity of kernel code pages is minimal, as low as 3% after a bootstrap phase. Therefore, a rootkit detector that offers kernel code integrity can do so while draining minimal energy, and can potentially be an "always-on" tool.
- Checking the integrity of all kernel data structures can place a significant strain on the mobile device's battery life. However, the energy consumption of the detector can be significantly reduced by a factor of three to five, if integrity checks are performed on selected high-risk data structures, or if data structure checks are only performed periodically.
- Based upon these measurements, we identify a sweet spot in the security versus energy tradeoff, one that provides the best compromise between energy consumption and the window of vulnerability opened as a result. This *balanced profile* is able to detect a vast majority of known attacks which work against code and selected kernel data structures, while consuming a limited amount of extra battery in our testbed, with an energy overhead between 6 to 9%.

To summarize, our main contribution is *to quantitatively explore, for the first time, the tradeoffs between security monitoring and energy consumption on mobile devices.*

2. ROOTKITS: ATTACKS AND DEFENSES

The stealthy nature of rootkits allows them to retain long-term control over infected devices, and to serve as a stepping stone for other attacks such as key-loggers or backdoors. It is no surprise then that a 2006 study by MacAfee Avert Labs [4] reported a 600% increase in the number of rootkits in the three year period from 2004-2006. The explosive growth of rootkits continues; MacAfee's 2010 threat predictions report also contains several examples of rootkit-aided Trojan horses that were used to commit bank fraud [5]. The increasing complexity of the hardware and software stack of mobile devices, coupled with the increasing economic value of personal data stored on mobile devices, point to an impending adoption of rootkits in the mobile malware arena [14]. The recent development of proof-of-concept rootkits for Android-based phones [38] and the iPhone [34] only reinforces these predictions.

2.1 Attack vectors

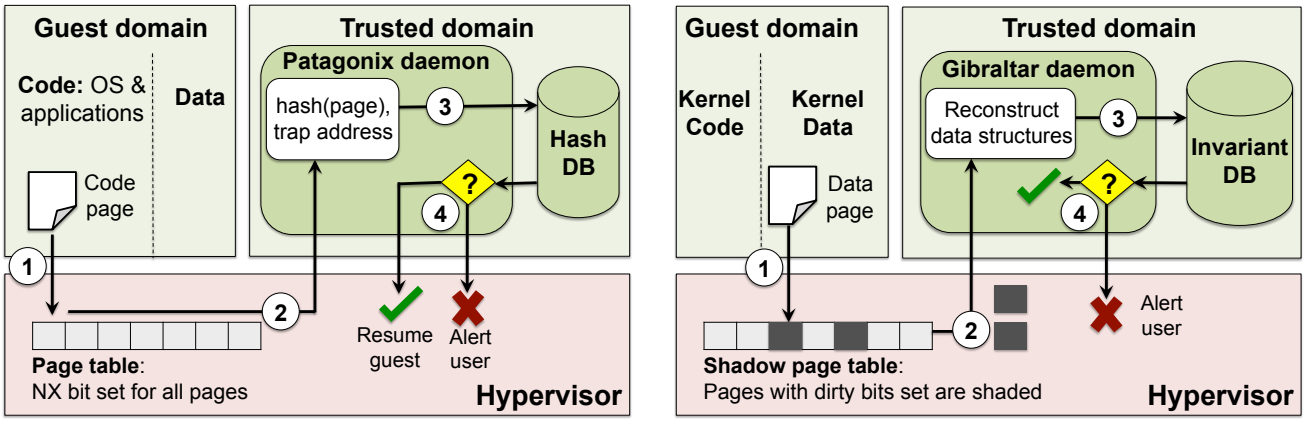
Rootkits remain stealthy by compromising the integrity of entities that belong to the trusted computing base (TCB) of victim devices. On most devices, these include OS code and data, as well as key user-space processes and files. We briefly survey the evolution of rootkit attack vectors, from those that are easiest to detect to those that are most challenging to detect.

- *System utilities.* Early rootkits attempted to hide the presence of malicious processes by compromising system utilities that are used for diagnostics. For example, a rootkit that replaces the `ls` and `ps` binaries with trojaned versions can hide the presence of malicious files and processes. Such rootkits are easy to detect by an uncompromised TCB that certifies the integrity of user-space utilities with checksums.
- *Kernel code.* The next generation of rootkits attempted to evade detection by affecting the integrity of kernel code. Such corruption is most usually achieved by coercing the system into loading malicious kernel modules. Once a rootkit has gained kernel execution privileges, it can mislead all detection attempts from user- or kernel-space. Successful detection of such rootkits is achieved instead by components located outside the control of the infected kernel. The two main approaches involve use of external hardware which scans the kernel memory using DMA (*e.g.,* [10, 11, 27, 47]), or introspection from the vantage point of a different virtual machine (*e.g.,* [22, 30, 39]).
- *Kernel data structures.* A large majority of rootkits in use today corrupt *kernel control data* by modifying function pointers in data structures such as the system call table or the interrupt descriptor table. This attack technique allows rootkits to redirect control to attacker code when the kernel is invoked. For example, the Adore rootkit [31] hides user-space processes from reporting tools like `ps`, by hijacking the function pointer for `readdir()` in the root inode of the `/proc` file system. More recently, research has shown that attacks against *non-control kernel data* are realistic threats [12, 16]. For example, a rootkit can subvert key cryptographic routines by affecting kernel parameters controlling pseudo-random number generation.

2.2 Defenses

In this section, we discuss the design of two prior techniques to rootkit detection. The two techniques are representative of the algorithms used in most rootkit detectors, and complement each other. The first technique, based on Patagonix [30], detects rootkits by monitoring code integrity; the second technique, based on Gibraltar [10, 11], monitors kernel data integrity.

Both tools use hypervisors to achieve isolation from the kernels they monitor. The hypervisor guarantees isolation between a monitored system (the *untrusted guest domain*) and the monitoring tool (the *trusted domain*); functional correctness of such guarantees has been formally proven [29]. The hypervisor and the trusted domain therefore comprise the TCB of the system. When the trusted domain detects a compromise, the TCB is capable of taking over the



1(a) Checking code integrity with Patagonix. ① When a code page in the guest is first scheduled for execution, it results in a trap to the hypervisor and suspends the guest. ② The hypervisor forwards this page to the Patagonix daemon. ③ Patagonix hashes the page and authorizes it. ④ If the execution of the code page is authorized, Patagonix informs the hypervisor, which resumes execution of the guest; otherwise, Patagonix raises an alert.

1(b) Checking data integrity with Gibraltar. ① When the guest kernel modifies a data page, the dirty bit of the corresponding entry in the shadow page table is set. ② The Gibraltar daemon consults the shadow page table and fetches dirty pages. ③ It reconstructs the data structures in these pages and checks whether they satisfy integrity constraints. ④ Gibraltar allows the untrusted guest to execute only if integrity constraints are satisfied.

Figure 1: The design of Patagonix (Section 2.2.1) and Gibraltar (Section 2.2.2).

UI to alert the user and provide containment options – the specifics of this mechanism are outside the scope of this paper.

2.2.1 *Checking code integrity*

Patagonix [30] is a rootkit detection tool whose design typifies that of most code integrity monitoring systems. It provides mechanisms to ensure that all code executing on a system belongs to a whitelist of pre-approved code. Rootkits that modify system utilities or kernel code can be detected if the modified code is not in the whitelist. Patagonix can also detect certain data-modifying rootkits, *e.g.,* those that modify kernel data pages that should not be modified during normal operation. Figure 1(a) presents the design of Patagonix.

Patagonix uses the capabilities of the hypervisor and the non-executable (NX) page table bit to detect and identify all executing code, including kernel code, system utilities, and other user processes. It modifies the code in the hypervisor to first set the NX-bit on all pages in the guest domain. When a page is first scheduled for execution, the NX bit causes a processor fault. The hypervisor receives the fault, pauses the guest domain and places information about the fault in a shared page that can be accessed by the Patagonix daemon executing in the trusted domain. The daemon hashes and compares the executing code to a whitelist of known software, comprised of the hashes of all approved code pages.

Patagonix enforces the **W⊗X** principle: pages are either modifiable, or executable. The hypervisor manipulates permission bits to enforce mutual exclusion between the two states. Pages will thus always be re-checked after modification. However, for code pages that are kept resident in the system and never change, Patagonix will not need to perform any further work. Thus, beyond an initial bootstrapping phase, the kernel working set and long-lived processes represent no additional work for Patagonix.

Patagonix uses optimizations to ensure fast verification of code pages. It remembers pages of code that have been blessed and have not changed. Thus, short-lived but recurring processes (*e.g.,*`grep`) will result in hypervisor work as new page tables are created, but no daemon work, due to reuse of resident unmodified code pages. Patagonix knows the entry point of each binary in its whitelist – the first trap on a new binary should match an entry point in the whitelist. For approved binaries, it stores the associated address space (defined by the base address of the current page table) and the segment of the address space the binary occupies: pages within the same segment should only match pages of the same binary.

Though Patagonix is not representative of *all* code-integrity monitoring systems, its design is similar to several state-of-the art rootkit detection tools that have recently been proposed in the research literature. For example, NICKLE [42] and SecVisor [43] are similar in overall design to Patagonix. Grace *et al.*'s paper on commodity operating system protection [24] implements a subset of techniques used by Patagonix [45]. Our results on security versus energy tradeoffs for Patagonix will therefore also be applicable to these tools.

2.2.2 *Checking data integrity*

Rootkits that modify arbitrary kernel data structures are challenging to detect because of two reasons. First, the kernel manages several thousand heterogeneous data structures, thereby providing a vast attack surface. Second, unlike code, kernel data is routinely modified during the course of normal execution. Distinguishing benign modifications from malicious ones requires intricate specifications of data structure integrity. In this section, we describe Gibraltar [10, 11], a tool that monitors the integrity of kernel data structures to detect malicious changes.

Figure 1(b) shows the design of Gibraltar: a daemon executes on the trusted domain, and periodically fetches data pages from the untrusted guest kernel. The daemon reconstructs kernel data structures in a manner akin to a garbage collector. It starts at a set of kernel *root symbols* whose memory locations are fixed. Using the OS type definitions, it identifies pointers in these root symbols, and recursively fetches more pages that contain data structures referenced by these pointers.

Once data structures have been reconstructed, *data structure invariants* that specify kernel integrity constraints are verified. Some invariants are simple to verify: the values of function pointers must be addresses of known functions; the entries of the system call

table should remain constant during the execution of the kernel. Other more complex invariants span sophisticated data structures, *e.g.,* each process that is scheduled for execution must have an entry in the linked list of active processes on the system.

Data structure invariants can be specified by domain experts [40], but this approach can be labor-intensive. Instead, Gibraltar leverages the observation that a large number of data structure invariants can be automatically inferred by observing the execution of an uninfected kernel. Such inference is performed during a controlled *training* phase, when a clean OS executes several benign workloads. Prior work [10, 11] shows that high-quality invariants can be obtained with a relatively short training phase using the Daikon invariant inference tool [20].

Rootkits that affect kernel data integrity (and the corresponding detection tools) are a relatively recent development in contrast to rootkits that affect code integrity. The overall design of Gibraltar substantially resembles those of other data integrity monitoring tools, such as SBCFI [39] and Petroni *et al.*'s specification-based rootkit detection architecture [40]. HookSafe [46] prevents data-oriented rootkits (whereas Gibraltar can only detect them), but only protects a proper subset of Gibraltar's detection space. Other systems that detect rootkits by checking data invariants, such as OSck [25] and Co-Pilot [27], check for simpler invariants and thus may miss rootkits that Gibraltar can detect.

Shadow page table optimization.

The original design of Gibraltar ran the daemon on a physically isolated machine and fetched memory pages from the monitored machine via DMA (using an intelligent NIC [10, 11]). For the study in this paper, we adapted Gibraltar to execute on a hypervisor, which allowed us to implement novel performance optimizations. Notably, we implemented a *shadow page table optimization* that allows the Gibraltar daemon to focus the application of integrity constraints on just those data pages that were modified by the guest. This optimization relies on the use of shadow page tables by modern hypervisors, which grant to the TCB fine-grained control over the permission bits of virtual-to-physical memory translation. In particular, they can be used to cause faults on the first attempt to modify a page. The hypervisor catches these faults and records them in a "log-dirty" bitmap. The Gibraltar daemon consults this bitmap and only focuses on pages whose dirty bits are set, and are known to contain data-structures of interest subject to integrity constraints.

The shadow page table optimization has a substantial effect on the number of checks Gibraltar has to perform. In experiments using the lmbench [33] workload executing for 144 seconds, 25 rounds of checks are performed by the optimized version of Gibraltar, as opposed to 5. By avoiding unnecessary checks to unmodified data, Gibraltar asserts the integrity of the kernel data structures 5 times more frequently, for the same power-budget and the same length of a user workload. We observed similar benefits in the other workloads employed in this paper.

3. THE SECURITY/ENERGY TRADEOFF

Security mechanisms have traditionally focused on well-provisioned computers such as heavy-duty servers or user desktops. Mobile devices present a fundamental departure from these classes of machines because they are critically resource-constrained. While advances throughout the last decade in mobile processor, GPU and wireless capabilities have been staggering, the hard fact is that mobile devices utilize batteries with a limited amount of stored power.

In this context, some fundamental tenets of security mechanism design need to be reconsidered. Without the limit of resource constraints, security mechanisms will check *everything they can, all the time*. In a mobile device, aggressively performing checks on large sets of security targets will inexorably lead to resource exhaustion and the inability to carry on useful tasks. Arguably, a certain amount of engineering could be added to any given security mechanism to make it marginally more efficient in terms of resource usage. We have just shown one such example with our shadow page table-based optimizations for Gibraltar. But we counter-argue that nothing short of a fundamental transformation will make security monitors palatable for mobile environments because energy must be a core consideration when designing tools for such environments.

The primary contribution of our work is in acknowledging that security needs to be traded off for battery lifetime in a mobile device, and in providing a framework to classify the choices a designer will face when modulating her security mechanism for a battery-constrained environment. Furthermore, we provide means for measuring the amount of security being traded off. To the best of our knowledge, neither such a framework nor such metrics were deemed necessary before the widespread adoption of smartphones and other mobile devices.

3.1 What to check and when to check it

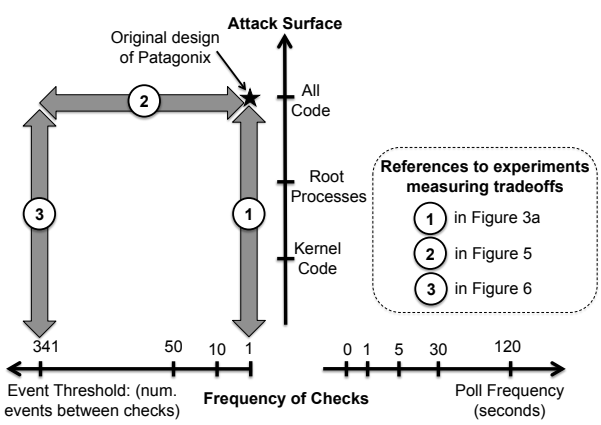
Energy-oblivious security mechanisms will check everything they can as frequently as they can. In a mobile setting, one must decide upon the attack surface to monitor (*i.e.,* what to check) and the frequency with which to perform monitoring (*i.e.,* when to check). These two factors must be incorporated as design parameters of the security mechanism itself to allow the mobile device to flexibly navigate the security versus energy tradeoff. We apply these concepts to the two rootkit detection systems discussed in the previous section.

• *What to check?* Operating system kernels provide a vast attack surface and a rootkit detection mechanism that monitors the entire attack surface will soon exhaust the mobile phone's battery. Both Patagonix and Gibraltar can be configured to check various subsets of the attack surface.

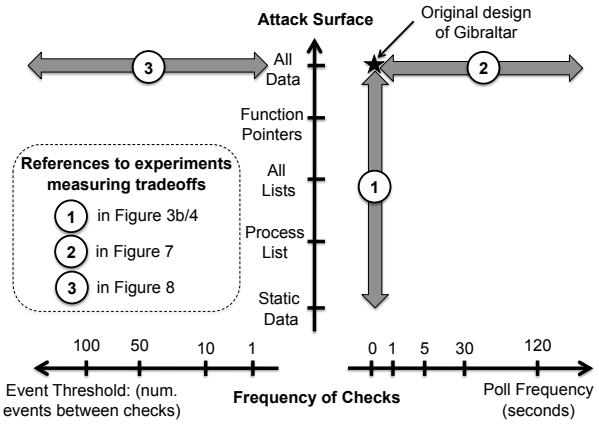
The Patagonix system can be configured to check (a) only the execution of kernel code pages; (b) kernel code pages and the execution of key binaries, such as root processes; and (c) kernel code, root processes, and selected kernel data structures, such as the system call table. Option (c) provides the highest security, whereas option (a) is the most efficient. The Patagonix daemon can trivially differentiate between kernel and user-space code due to the virtual addresses used: in all commodity OSes the kernel resides, on all address spaces, in a high band of virtual addresses. Root processes are manually classified and tagged in the whitelist; this does not prevent Patagonix from checking libraries linked in the address space of a root process (*e.g.,* OpenSSL for sshd).

Gibraltar also offers a wide variety of configurations, ranging from checks on selected kernel data structures, such as those that store control data (including function pointers), to checks on all kernel data structures. In between these two extremes, we can tune Gibraltar to check additional classes of data structures: static data; the process list and runnable queue, which are a common target of attacks that hide the existence of a malicious user-space process [31]; all linked lists beyond the previous two; and more.

• *When to check?* Independently of the size of the attack surface being checked, one must tune how often to perform the checks. The design space for the frequency of checks ranges from an approach that uses periodic polling (where the period is configurable) to one that uses event-based or interrupt-based notifications to trigger the



2(a) Design space for code integrity checks.



2(b) Design space for data integrity checks.

Figure 2: The security versus energy tradeoff. These figures illustrate various points in the design space of Patagonix (Figure 2(a)) and Gibraltar (Figure 2(b)). The *y*-axis of each figure shows various subsets of the attack surface, while the *x*-axis considers the parameters used to decide the frequency of checks. The shaded portions of each figure show the portions of the design space that we explored in our experiments (Section 5).

rootkit detector. Choosing the appropriate approach is a fairly well-understood dichotomy prevalent in systems design. With proper hardware support, one can implement event-based checks relatively efficiently, preventing the use of busy loops that burn too much CPU, or sleep timers that ignore momentous events.

The original design of Patagonix uses an event-based approach to pause the guest domain each time a new page is scheduled for execution and check the page. However, Patagonix can also be modified to batch and perform these checks *en masse*. The former option detects and prevents malicious code execution in an online fashion, but may frequently pause the guest domain. In contrast, the latter option may detect malicious code only after it has executed on the system, but is likely to be more efficient.

The Gibraltar daemon traverses the guest OS's data pages in rounds, pausing for an interval of time after each round. During this interval, Gibraltar does not scan the kernel's data structures. The frequency of this traversal impacts energy efficiency and security. Frequent traversal minimizes the vulnerability of the guest operating system, while infrequent traversal conserves energy. The original version of Gibraltar has a T of zero as it continuously scanned all kernel pages: once the daemon completed traversal of all relevant kernel data structures, it immediately started a new round of traversals.

Our implementation of Gibraltar for mobile devices also incorporates an event-based mechanism in which the hypervisor interrupts the Gibraltar daemon when the guest has modified a certain number of pages, N, so that the data structure checks for these pages can be batched and performed *en masse*. We added a new interface to allow the Gibraltar daemon to instruct the hypervisor about which pages are relevant and which are not. The daemon prepares a bitmap indicating pages in which data structures of importance reside. The hypervisor will only wake up the daemon once N relevant pages in the bitmap have changed. This prevents the hypervisor from accounting for frequent stack or page cache modifications as relevant.

Figure 2 summarizes these concepts, and also shows the portions of the design space that we considered in our experiments to quantify the security versus energy tradeoffs for mobile device rootkit detection.

3.2 Measuring the security we give away

Reducing the attack surface monitored or the frequency of checks introduces the possibility of evasion. A rootkit could evade detection by infecting the kernel between checks or by modifying unmonitored data structures. Therefore, the security provided to a system is intimately related to the frequency of checks and the attack surface monitored.

• *Impact of attack surface size.* It is challenging to measure the impact of varying the attack surface on the security of the system. This is because (1) different entities in the attack surface impact system security to varying degrees; and (2) not all entities in the attack surface can be compromised with equal ease. With rootkits, attacks that modify kernel code, static data, and data structures that hold control data (*e.g.,* the system call table) are more abundant and easy to program than attacks that modify arbitrary kernel data structures.

For lack of a good metric, we default here to manual expert curation. Prior studies have shown that: (1) kernel code is far more important to rootkit detection than user-space code [30, 42, 43]; (2) among rootkit-based attacks that modify kernel data, function pointers are the prime target [39] as opposed to other data structures. A 2007 study of 25 popular rootkits by Petroni *et al.* [39] showed that 24 of these rootkits modified function pointers to achieve their malicious goals. Rootkits that do not use function pointers as an attack vector typically either modify different data structures [12] or inject malicious code into the kernel [25].

• *Impact of check frequency.* Constant vigilance is likely more effective than daily overnight checks at catching exploits before they have done much harm. To quantify how the frequency of checks impacts the security provided by a detection system, we introduce the concept of *window of vulnerability*.

The window of vulnerability for a given object is defined as the time elapsed between two consecutive checks on that object. For example, if we check the kernel system call table every two seconds, a rootkit has a maximum of two seconds to hijack the system call table, steal user information written to a file via `write()`, and optionally restore the table to its pristine state to avoid detection. Our window of vulnerability is therefore two seconds. For security systems that check multiple components, the window of vulner-

ability metrics of each component (each code page or each data structure in our case) can be statistically aggregated into a system-wide value (*e.g.,* as the window of vulnerability averaged over all components).

The window of vulnerability is the time period during which a system is vulnerable to attack. The greater the period of time between checks, the more time an attacker has to perform a sophisticated attack. For example, with a large window of vulnerability, a rootkit might have time to steal and transmit a user's personal information, *e.g.,* gathered during a secure browsing session using a key logger, to a malicious server. In general, a smaller window of vulnerability will expose fewer user interactions to the rootkit. For instance, with a smaller window of vulnerability, it may be possible to detect the presence of a rootkit and raise an alert before the user completes the secure browsing session, thereby protecting at least some of the user's personal information.

Periodic polling systems have a clearly defined set of windows of vulnerabilities that they expose for each object they check. For a polling period T, the average window of vulnerability will be *at least* T, plus the processing time involved within each round. However, event-based systems can provide a greater degree of assurance. If the hardware, can immediately alert the monitor of a potential threat even before it is allowed to happen, then the system can provide an effective window of vulnerability of zero. Doing so effectively requires the system to react to a potentially large volume of events. For this reason, it is common for an event-based system to perform event merging or coalescing, *e.g.,* interrupt batching for a processor, or signal handling for UNIX processes. In this case the window of vulnerability widens again depending on the amount of merging performed.

3.3 Mitigating a new class of timing attacks

Resourceful and knowledgeable adversaries will immediately recognize a new opportunity. By learning the timing mode and parameters of the system, they can craft attacks that break in, exploit, and clean up within the period of time during which security checks are inactive. We mitigate this by randomizing the timing parameters. For example, if Gibraltar is to be configured to check data structures every T seconds, we instead trigger checks at intervals pulled from a uniform distribution in the interval (T-M,T+M], with $M \leq T$. To generate a proper uniform distribution, the hypervisor can tap from sources of entropy that are protected from the guest kernel, such as the count of hardware interrupts. Because the checking intervals are uniformly distributed in the interval (T-M,T+M], windows of vulnerability and checking overhead will converge to the same values as if a fixed period of T seconds had been chosen.

For event-based timing modes, we can apply the same randomization to the number of events that will trigger security checks. However, we have to further augment the approach with an explicit timeout (which itself could be randomized, if necessary). The reason for this is that the system may enter a steady state in which the selected threshold of events (e.g. page executions or modifications) is not reached, thus granting the attacker an unlimited window of vulnerability.

In this paper we are focused on measuring and characterizing the tradeoffs between security checking and energy footprint. For those reasons, we use fixed intervals and event thresholds throughout the evaluation section. This removes an additional layer of experimental noise from our measurement goals.

A similar concern arises if we reduce the surface of coverage for our checks. Similarly, we might choose to catch the attacker off-guard by randomly triggering coverage of a wider attack surface.

4. EXPERIMENTAL SETUP

We now describe the experimental platform and the workloads employed for our study. Our goal is to illuminate various aspects of the security/energy tradeoff for host-based rootkit detection:

- *Impact of attack surface size.* If a malware detector provides greater security by monitoring a larger attack surface (*i.e.,* classes of attack), its detection algorithm will likely be more complex, CPU-intensive, or will take longer to execute. How does the size of the monitored attack surface impact battery life?

- *Impact of malware scanning schedule.* Malware detectors can be configured to be "always-on" tools that continuously monitor for malicious activity, or can periodically scan the mobile device. The former option provides increased security while the latter option improves battery life. How does the schedule of scanning impact energy consumption?

In Section 5 we report our findings relating to the above questions. In Section 6 we build upon our results to further address:

- *Adaptation.* Given the conventional wisdom that executing malware detectors reduces battery life, can we develop a strategy that maximizes security while minimizing battery consumption?

- *End user involvement.* Can we further expose such strategy and its inherent tradeoffs and options to end users? Can we do so in an intelligible manner similar to that used with traditional performance versus power-savings strategies?

4.1 Platform

We used a Viliv S5 mobile device [2] as our experimental platform. It is equipped with an Intel Atom Z520 1.33 GHz processor rated at 1.5 W, 4.8" touch screen, 32GB hard drive, 1GB of memory, WiFi, Bluetooth, a 3G modem, GPS, and a battery rated at 24,000 mWh. Since our rootkit detection tools are dependent on running in a virtualized environment, the ability to install a hypervisor on the device was a key requirement. The limited availability of mobile virtualization options dictated our platform choice. With its x86 Atom processor, the Viliv supports Xen paravirtualization [13], and is one of few such devices most resembling a smartphone that we could purchase in North America. Other virtualization platforms either require VT extensions [35], which are available only on a few higher-powered Atom models; are not available commercially and/or in open-source form (*e.g.,* VMware Mobile [3]); or cannot be installed on commodity smart phones available today (*e.g.,* the Xen port to the ARM platform [26] and the OKL4 Microvisor [1]). In spite of its slightly larger form-factor, the Viliv is functionally equivalent to a phone. Further, the Menlow chipset used by the Viliv is the precursor to the to-be-released Moorestown platform for Intel-based smartphones such as the LG GW990.

On the device, we used the Xen 3.4.2 hypervisor. Xen relies on a trusted domain (*i.e.,* dom0) to manage VM lifecycles and execute device drivers, which in our case was a Fedora 12 stack running a version of Linux 2.6.27.42 with appropriate Xen patches. We enhanced the hypervisor on the device with support for Patagonix and Gibraltar, and added the respective daemons to the dom0 stack. Our guest domain ran Linux 2.6.27.5 with Xen paravirtualization patches under a CentOS 5.5 distribution.

To measure power, we used a Tektronix TDS-3014 oscilloscope with a Hall effect current probe. When performing power measurements, we disconnected the battery from the Viliv S5 device and supplied power directly from a 5V source. We used this approach to ensure that the current we measure is directly powering the device. The current probe was attached to the charging cord and a

Operation	Energy (mWh)
Send/Receive Phone Call	1.1 ± .03 / second
Send/Receive 160-char SMS	6.3 ± 1 / SMS
Send/Receive 5-char SMS	6.2 ± 1.2 / SMS

Table 1: Energy spent for common mobile phone operations.

laptop connected to the oscilloscope recorded the current readings over the time of an experiment.

4.2 Workloads

Experimental workloads that have traditionally been used to evaluate the performance of security tools, such as members of the SPEC family, often fail to capture the dynamics of the mobile experience. We therefore created our own workload for our evaluation. This workload aims to replicate standard mobile usage by loading a series of popular web pages and checking email.

Our workload is driven by a script that starts up the Firefox browser by pointing it to the desired site via a command line argument. It then monitors the CPU usage of the browser until it settles into reasonably low utilization; many popular sites employ Flash animations that never quite stop consuming resources. Once the site has quiesced, the script discards the Firefox instance, and moves on to the next site on the list. By pointing the browser to a Youtube clip, Flash playback will prevent quiescing of the browser throughout the duration of the clip, thus allowing full playback. The script similarly launches an email client and discards it after email checkout has finished and the process has quiesced.

Our workload is highly customizable and independent of a specific platform, needing just the ability to launch browser and email client instances from a script. We plan to augment our workload with fetching and uploading data to a social networking site and release it to the mobile computing community.

Throughout the experiments in this paper, we loaded google.com, cnn.com, gmail.com using an open account, youtube.com pointing to a 60-second video, and Thunderbird configured to check email from one IMAP account with several hundred messages in its inbox. We ran this workload on the Viliv using both 3G and WiFi connectivity, and for simplicity refer to the results respectively as "3G Browsing" and "WiFi Browsing."

For completeness, we also used lmbench [33], a CPU intensive workload designed to measure OS performance. We used the first six stages of lmbench because it thoroughly exercises multiple OS interfaces, thereby stressing our rootkit detectors.

4.3 Rootkit detector configuration

To generate a database of invariants, Gibraltar must first execute a training phase. Since lmbench modifies many data structures in the operating system, we trained Gibraltar against multiple complete executions of lmbench. The result is a database of 131,201 invariants across 2209 data structure types with a size of 7MB.

Patagonix, requires a database of hashes for all binaries running on a system. To generate this database, we generated an ELF parsing tool to output a hash of each code page. We parsed all binaries located in the official CentOS repository, resulting in a database size of 36 MB. The database stores 10929 different binary files and 509709 hashes. We also store a database of 627 kernel code pages resulting in a size of less than 1 MB.

5. EXPERIMENTAL RESULTS

In this section, we present experiments that illustrate the security versus energy tradeoff faced by kernel code-integrity and data-integrity monitors. In each experiment, we report the total energy dissipated by the Viliv as it executed one of three workloads (lmbench, 3G and WiFi Browsing) and the value of the corresponding security metric (attack surface or window of vulnerability). Unless otherwise noted, we report the average and standard deviations obtained from three experiment runs for each data point.

We start this section by quantifying the overhead of executing the hypervisor on the Viliv. We compare the execution of our workloads on a native bare-metal kernel, to the execution of our workloads inside a virtual machine, with no host-based rootkit detectors activated. As measured by total energy dissipation, the overhead is negligible in all cases. We observed the maximum overhead in the 227-second 3G browsing workload, with the energy footprint going from 333 to 335 mWh in virtualized mode (1.29%).

Table 1 presents the energy dissipated by two operations that are common to mobile phones: placing/receiving a 60-second phone call, and sending/receiving SMS messages. We observe that varying the SMS message size had relatively no effect on the total energy required to send the message. In the rest of this section, we will refer back to Table 1 to place the energy overheads in context by comparing them to the cost of common phone operations.

5.1 Impact of security on energy and runtime

The introduction of host-based rootkit detection has an immediate impact on the time to completion for a given workload, and a strongly correlated impact on the total energy consumed by the workload. Table 2 shows the measurements for our three workloads; we compare the original implementations of Patagonix and Gibraltar to runs without security checks. Pearson correlation coefficients between energy and time overheads exceed 0.97 for all workloads.

Primarily, host-based rootkit detection competes for CPU cycles with the workload, prolonging the time to completion. Patagonix's contention is a function of the amount of different code executed, and for our workloads the resulting overhead does not exceed 33%. Gibraltar is constantly asserting kernel data integrity, and thus constantly contending for CPU cycles. The overhead becomes dependent on the hardware in use. With a CPU-bound workload (lmbench) the overhead nears 100%. With network IO involved, there is no workload slowdown during periods in which the system stalls waiting for IO – Gibraltar occupies otherwise unused CPU cycles. With faster IO hardware (WiFi), there are fewer stall periods, and the relative overhead is higher (64% versus 43% for 3G). With slower and less-energy efficient hardware (3G), the longer fractions of IO stall occupied by Gibraltar yield a higher absolute overhead (146 mWh versus 89 mWh for WiFi).

The strong correlations between energy footprint and workload runtimes observed here also hold throughout the experiments in this paper. We focus on energy measures and do not report runtimes due to space considerations.

5.2 Modifying the attack surface

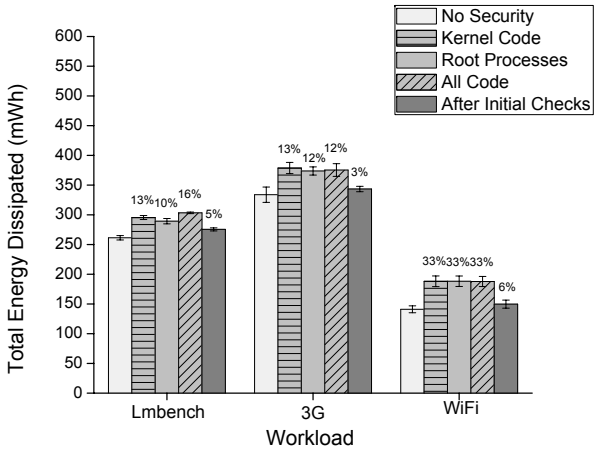
The energy dissipated by a security tool can be reduced by decreasing the fraction of the attack surface monitored. We quantify this observation considering the attack surface of code and data.

Impact on code integrity.

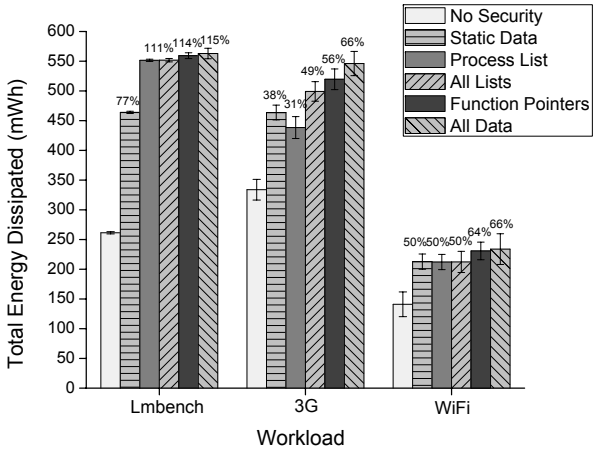
We configured Patagonix to monitor three subsets of the attack surface: (a) kernel code only; (b) kernel code and root processes; and (c) all code on the system, including kernel code, root and non-root processes. We set up Patagonix to check each code page as soon as it was scheduled for execution. On average, the Patagonix daemon verified 309 pages of kernel code as it executed the WiFi and 3G Browsing workloads; this number rose to 749 code pages when we

	No Security		Patagonix		Gibraltar		r
	Time (s)	Energy (mWh)	Time (s)	Energy (mWh)	Time (s)	Energy (mWh)	
lmbench	217.5 ± 0.2	261.39 ± 3.7	243.5 ± 2.1	303.31 ± 1.3	374.6 ± 13.4	473.56 ± 5.2	0.9995
3G Browsing	227.2 ± 9.9	333.84 ± 12.9	269.7 ± 5.8	375.36 ± 10.7	317.5 ± 14.1	479.95 ± 13.1	0.9779
WiFi Browsing	144.3 ± 5.3	141.08 ± 5.8	180.9 ± 8.9	187.84 ± 8.5	262.1 ± 7.5	230.38 ± 2.9	0.9707

Table 2: Runtime versus Energy correlation for security checks. Runtime and energy footprint for our workloads, ran with no host-based rootkit detection, and with the original versions of our code (Patagonix) and data (Gibraltar) integrity checkers. Column r shows that the Pearson correlation coefficient between energy and runtime is strong.



3(a) Code-integrity checks. Each time a page of code is executed, an integrity check is performed. Percentages indicate the energy overhead with respect to the case with no security.



3(b) Data-integrity checks. Each data class includes the previous class on the x-axis. Verifying the integrity of all kernel data is expensive when compared to code integrity checks.

Figure 3: Impact of varying the attack surface on the total energy dissipated by code and data integrity checks.

included user-space code as well, 90 pages of which corresponded to root processes. During the execution of the lmbench workload, Patagonix verified 301 kernel code pages and 1602 user-space code pages, 11 of which belonged to root processes.

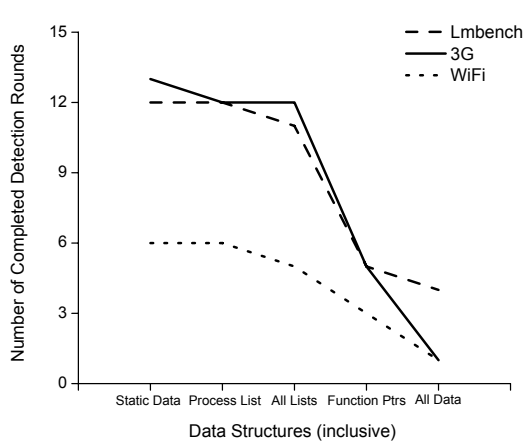
Figure 3(a) illustrates the energy dissipated by each of the three workloads. We present results for each subset of the attack surface considered. For each workload, the leftmost column is the baseline, which shows the energy dissipated by the workload executing in an environment with no security checks enabled (*i.e.,* in a hypervisor without Patagonix). Percentages reported above columns represent the extra energy dissipated (over the baseline value) when monitoring the corresponding subset of the attack surface. Comparing these results to Table 1, the extra energy dissipated by Patagonix when eagerly checking all code for the 144-second WiFi workload is the same as placing a 38 second phone call or sending 7 SMS messages.

Once Patagonix verifies the integrity of a code page, if the running process remains resident in memory and the code is not modified, Patagonix will never need to verify this page again, and it will therefore incur in no further overhead. This is particularly true for the kernel, which after boot remains resident and unchanged (save for module additions). The rightmost column in Figure 3(a) ("After Initial Checks") depicts the Patagonix overhead for the common case of recurring processes after bootstrap: the energy measurements were obtained by running the workloads a second time, after the initial execution. In this case, extra hypervisor work is necessary to enforce the $\mathbf{W}\otimes\mathbf{X}$ principle on the new page tables created. However, no additional daemon work is needed because the resident code pages have already been checked. The hypervisor overhead is small, and equivalent to a 10 second phone call or sending 2 SMS messages.

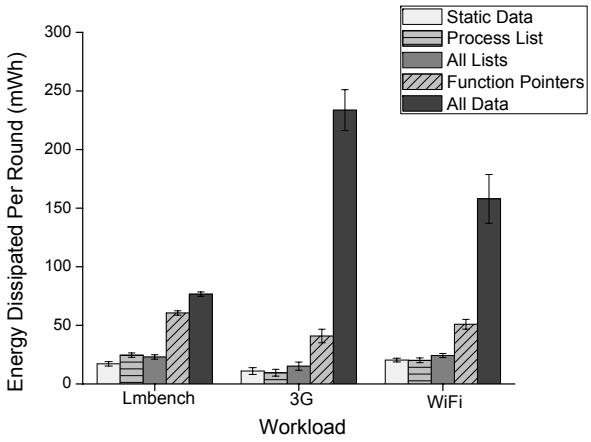
Impact on data integrity.

We configured Gibraltar to monitor five classes of kernel data, containing: (a) static kernel data, *i.e.,* data that is initialized during kernel boot-up and persists throughout the execution of the kernel; (b) data structures representing the process list; (c) all linked lists; (d) all kernel data structures that store function pointers; and (e) all data structures. Each class is inclusive, *i.e.,* data structures verified in each class also include the previous class as a subset. We set up Gibraltar to continuously monitor data integrity, and we refer to one complete traversal of the kernel's data segment as a *detection round*.

Figure 3(b) illustrates the total energy dissipated while Gibraltar monitors each of our three workloads. Our first observation is that the energy dissipated by Gibraltar is significantly higher than in the Patagonix case. As explained before, the Gibraltar daemon continuously contends for CPU cycles with the user workload. Our second observation is that the energy dissipated by Gibraltar varies with the attack surface being monitored. This is despite the fact that irrespective of the attack surface, Gibraltar executes continuously without any periods of dormancy. We hypothesize that cache pollution effects determine the overhead differences: with larger attack surfaces to cover, Gibraltar traverses a larger volume of data, thus



4(a) Number of rounds.



4(b) Energy dissipated per round.

Figure 4: Impact of varying the attack surface for kernel data integrity checks on the energy dissipated per detection round and the number of completed detection rounds. Each data class includes the previous class. As data classes are added to the attack surface being monitored, the number of detection rounds completed decreases, and the energy per round increases. Verifying the integrity of all kernel data is expensive, but restricting the amount of data structures verified can decrease the energy dissipated per round by up to 95%.

effectively behaving as an adversarial workload in terms of memory locality. Decreased memory locality impacts processor cache performance efficiency, and thus energy efficiency. We plan to further investigate this effect, to potentially adjust Gibraltar's behavior to cache pollution rates.

Figure 4 presents the number of detection rounds Gibraltar completed throughout a workload, as well as the energy overhead per detection round. Both metrics are essentially locked in a zero-sum game: as the surface of attacks covered increases, more energy is spent in proportionally fewer rounds. When monitoring a smaller attack surface, data structures are checked more frequently (see Figure 4(a)), presenting a smaller window of vulnerability to attackers. The energy per round spent during the verification of kernel static data, linked lists and function pointers is significantly lower than that spent checking all data (see Figure 4(b)). This is fact is especially evident for the 3G and WiFi Browsing workloads, which dissipate approximately 3×-5× less energy than when Gibraltar monitors all data structures. This result is significant because a recent study of 25 rootkits [39] shows that 24 operate by violating the integrity of static data, linked lists or function pointers. As a consequence, Gibraltar can protect against most known attacks with modest energy dissipation per round: in the next section we show the limited amount of security we trade off by spacing these rounds and preventing continuous checking (and energy dissipation).

The number of data structures is one of two parameters that determine Gibraltar's coverage. The other is the number of invariants that are checked on these data structures. Decreasing the number of invariants from the original 131,201 to zero resulted in virtually no energy savings per round. We conclude that the dominant factor in the overhead per round for Gibraltar is the cost of reconstructing data structures.

5.3 Modifying the frequency of checks

Rootkit detection can be *event-based*, as in the original design of Patagonix, or *polling-based*. Patagonix can be adapted to batch events, while Gibraltar can be configured to poll kernel memory in different ways. In this section, we explore the effects of changing the frequency of checks, while measuring the security that we give away.

Impact on code integrity.

In the Patagonix experiments we observed that there were, on average, 50050 hypervisor notifications for lmbench, 13803 for 3G Browsing, and 15825 for WiFi Browsing. Each notification triggers a context switch to the trusted domain, where the page of code attempting to execute is checked. To decrease the number of context switches, Patagonix can be configured to add pages to a queue maintained in the hypervisor, notifying the daemon in the trusted domain only when the queue is full. Recall from Section 2.2.1 that the hypervisor places information about a faulting executable page (page number, address space, and instruction address) on a page shared with the trusted domain. The maximum number of entries a single 4 KB x86 memory page can hold is 341, thus dictating the size of our queue.

Figure 5 shows that batching code integrity checks results in a net decrease in energy dissipation for Patagonix. We attribute this primarily to the decrease in context switches. We observed 440 context switches while executing lmbench, 75 while executing the 3G Browsing workload, and 70 while executing the WiFi Browsing workload. This 99% decrease in the number of context switches yields the most impact for the WiFi workload: the Patagonix overhead decreases from the equivalent of placing a 38 second phone call to the equivalent of placing a 22 second phone call. Figure 6 shows that these results hold as we vary the coverage surface of our code integrity checks. Finally, as in the original case, subsequent executions of the workloads require no additional code verifications by the daemon, resulting in decreased energy expenditures up to a minimum of 3% for 3G browsing.

Batching code execution notifications fundamentally alters the security guarantees of Patagonix. By design, the original version of Patagonix offers a zero window of vulnerability: no code ex-

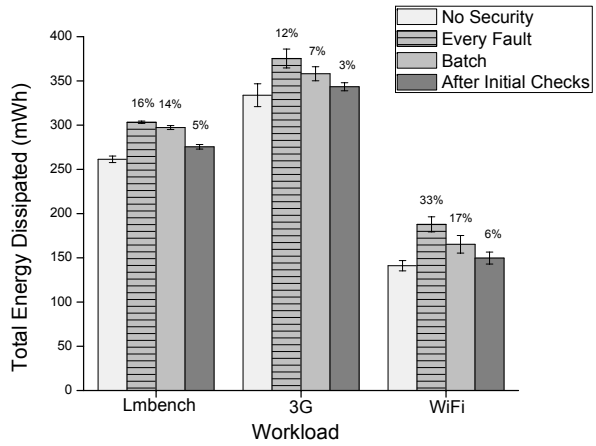


Figure 5: Impact of varying the frequency of code integrity checks. Batching code integrity checks reduces energy overhead up to a maximum of 3% for 3G browsing.

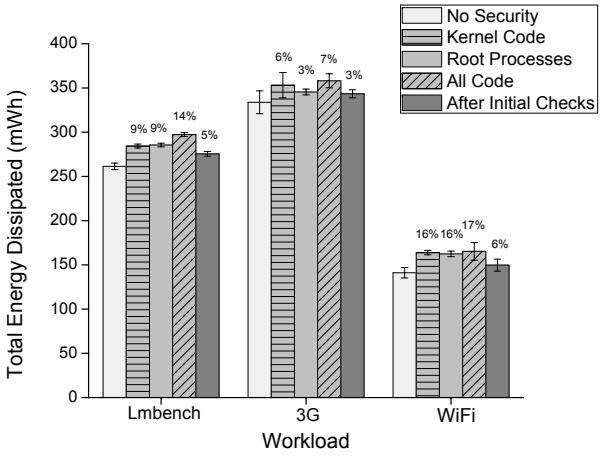


Figure 6: Impact of batching code integrity checks for various attack surfaces. Energy reductions are also observed when batching code integrity checks for different attack surfaces.

Workload	Window of Vulnerability (s)
lmbench	2.5066 ± 3.39
3G Browsing	0.8766 ± 1.63
WiFi Browsing	0.7233 ± 1.79

Table 3: Window of vulnerability for batched code integrity checks. While batching saves energy, it also allows code to execute for a brief period of time without being checked.

ecutes without prior inspection. Batching allows code to execute for a period of time without being modified, opening up a window of vulnerability. Table 3 shows that the windows of vulnerability are fairly small (under a second for browsing workloads), although quite variable because the rate at which new code pages execute is not at all uniform.

As discussed in Section 3.3, event-based queues need to be complemented with a timeout. Otherwise, the queue may never completely fill up, allowing in our case for rootkit code to remain undetected for arbitrarily long. We have not addressed this in this paper as our focus was in studying mechanisms in isolation. From Table 3 we conclude that a timeout of five seconds would suffice.

Impact on data integrity.

Gibraltar can be configured to poll kernel data structures in one of two ways. The first configuration option uses a *polling period T* (in seconds) between detection rounds – the Gibraltar daemon starts a fresh traversal of kernel data structures T seconds after finishing the previous traversal. The second configuration option is event-based: the Gibraltar daemon is woken up after the guest kernel has modified N pages containing data structures of importance.

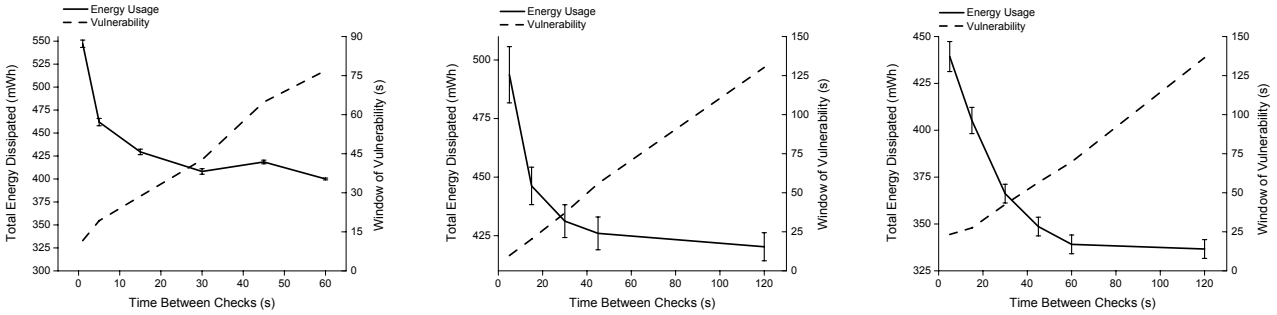
Figure 7 succinctly captures the security versus energy tradeoff. The solid lines represent the energy dissipated by the workloads executing in a guest domain monitored by Gibraltar, with different polling periods T varying between 0, 5, 15, 30, 45, 60, and 120 seconds. The broken lines represent the average window of vulnerability in the system. Increasing T results in less frequent rounds of verification – it increases battery life but decreases the overall security of the system, opening up wider windows of vulnerability. Recall that the average window of vulnerability is the mean of the times elapsed between consecutive integrity checks for each kernel data structure. The lower bound for the window of vulnerability is thus T, plus a small quantity derived from the time spent within each verification round.

Figure 8 presents the result of using the second configuration option to vary the frequency of checks. We configured Gibraltar to trigger integrity checks after N data pages have been modified, with N varying between 10, 50, 75, 100 and 120 pages. Both the lmbench and 3G Browsing workloads do not trigger a detection round for N=100 and N=120 pages. This is because these workloads repeatedly modify the same set of 75 to 99 kernel data pages. As the value of N increases, the amount of time between detection rounds also increases, and we observe the same phenomenon as in the polling case: energy overhead is traded off for an increase in the window of vulnerability in the kernel.

6. SECURITY/ENERGY PROFILES

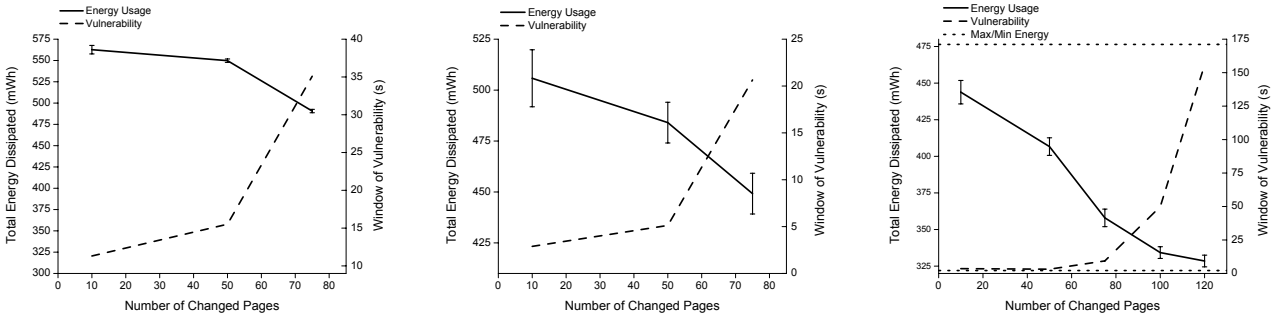
This section discusses how the results of the experiments from the previous section can be used to construct profiles that end-users can leverage to make educated decisions on how best to protect their mobile platforms. In that regard, security versus energy tradeoffs must be similar to performance/energy tradeoffs, which have existed since early laptop models and are therefore familiar to end-users. Such performance/energy tradeoffs are typically expressed succinctly, as a set of pre-defined power management profiles, in keeping with the conventional wisdom that a vast majority of end-users will steer away from both too much data and too many options.

Consider, for instance the power management profiles available on an iPhone running iOS 4.0. The only options exposed are: (1) screen brightness in standard mode; (2) the ability to automatically dim the screen brightness if inactive (but not parameters such as the dim gradient), and (3) the timeout period before the phone locks and the screen is turned off. Standard Windows 7 installations expose two power-management profiles, a *balanced* and *power saver* profile. Inquisitive users can find a third *high performance* profile. Further control is allowed by tailoring user-specific profiles. For security management to be useful, similar profiles must be made available to users. Power management profiles rep-



(a) **Lmbench** (b) **3G Browsing** (c) **WiFi Browsing**

Figure 7: Impact on energy and security when varying the period between data integrity checks. Gibraltar is triggered every time T seconds elapse between detection rounds, where T is represented on the x-axis. As the energy spent decreases, the window of vulnerability increases at a quasi-linear rate dependent on the time between detection rounds.



(a) **Lmbench** (b) **3G Browsing** (c) **WiFi Browsing**

Figure 8: Impact on energy and security when varying the threshold of dirty pages for data integrity checks. Gibraltar is triggered every time N pages change, with N represented on the x-axis. Because the rate of kernel page modification is not linear, windows of vulnerability increase non-linearly as well.

resenting the extremities of the security versus energy tradeoff are, of course, easy to synthesize. In the rest of this section, we explain the reasoning behind a "best compromise" profile.

The results from the previous section show that energy cost association with checking code integrity is much lower than the cost of checking data integrity. The cost of Patagonix reduces further after code pages have been verified once and the system settles into a relatively stable working set of code pages. However, checking the integrity of kernel code alone is not sufficient to detect rootkits. Rootkits can perform their nefarious activities without installing new code to do so, *e.g.,* by using existing binary streams [44] or directly modifying kernel data structures by exploiting kernel buffer overflows. Static checking of code, as performed by Patagonix, cannot prevent potential hijacking of JiT code regions. Ultimately, control-flow is often governed by data structures such as function pointers, whose tampering could lead to subtle compromises.

Figure 4 shows that the power consumption of Gibraltar rises sharply when one increases its coverage to include all kernel data. However, checking the integrity of kernel data objects that are typically attacked by rootkits (function pointers, process list, plus static objects) is three to five times cheaper energy-wise. The decision on which set of data structures to check is based on typical rootkit behavior [39] and is independent of the workload used in our experiments. Further, Figure 7 shows that a reasonable tradeoff between energy efficiency and window of vulnerability can be achieved by observing the intersection point between the corresponding curves. At the intersection point, there is a balance between energy consumption and the window of vulnerability of the system. Generally, this intersection point will be dependent on the workloads used to determine the tradeoff. From our browsing workloads, we determine the "sweet spot" as a polling mode with a period of $T = 30$ seconds. It is worth pointing out that for three fairly different workloads (from a system point-of-view), the intersection point lies in the neighborhood of $T = 30$ seconds. For the case of checking code integrity, Figure 5 shows that batching integrity checks reduces the energy overhead when using Patagonix.

Using this evidence, we construct a *balanced profile* for moderate energy consumption with a high degree of assurance against most rootkit attacks. This profile combines batched checks of kernel code pages, with polling-based integrity checks of static kernel data, linked lists and data structures containing function pointers, using the $T = 30$ second period identified as the sweet spot.

Figure 9 presents the energy dissipation of this balanced profile. Windows of vulnerability for kernel code vary between 2.5 and 0.7 seconds, while windows of vulnerability for kernel data structures monitored are on average 40.74 seconds. The energy overhead re-

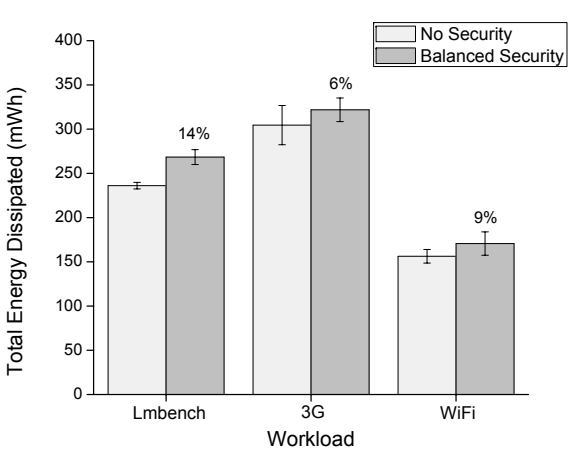


Figure 9: Energy dissipated for balanced security profile. Code and data integrity checks are executed simultaneously in a balanced setting. Patagonix batches daemon notifications while Gibraltar checks function pointers, lists, and static symbols every 30s. This figure shows that the energy overhead remains manageable while verifying 96% of the rootkit attack space [39].

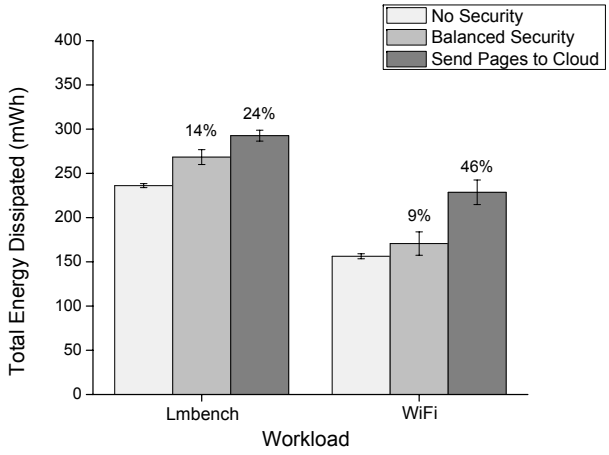


Figure 10: Cloud-based feasibility test. An increasing trend is to offload detection mechanisms to the cloud. The figure presents the total energy dissipated while offloading pages to an idealized cloud-based rootkit detector. For browsing workloads, cloud-based rootkit detection would require a significantly higher amount of energy compared to host based detection mechanisms.

mains manageable at a maximum increase of 14%. Web browsing over 3G or WiFi incurs less overhead at 6% and 9%, respectively. The latter overhead is equivalent to a 15 second phone call or 2 SMS messages, for a workload originally taking 144 seconds. We note that the lower-energy mode into which Patagonix transitions after checking the kernel working set and resident processes is not included here.

7. RELATED AND FUTURE WORK

In this section, we discuss prior work on detecting malware on resource-constrained mobile devices. Although these works have developed new detection approaches tailored for mobile devices, some of which are resource-aware, we are not aware of prior work on quantifying the security versus energy tradeoff. However, we leverage some observations from the literature to guide our future work plans.

Offloading detection.

As discussed in the Introduction, one way to sidestep the security versus energy tradeoff for detecting certain kinds of malware is to offload detection to a well-provisioned machine. Maui [18] and CloneCloud [9] approach general cloud offloading, while Paranoid Android [41] focuses on security. The latter performs user-space operation record and replay, at the granularity of system calls and signals. Replay on well-provisioned servers allowed offloading of security checks, as they are executed on a conceptually identical environment. However, host-based operation record, and the uploading of these operations to a server, resulted in an energy overhead of 30%.

Given the increasing popularity of cloud-offload, we conducted a feasibility study to investigate whether rootkit detection can be similarly offloaded. Instead of running the host-based rootkit detection logic locally, we perform a straight-forward partition. The same hypervisor logic executes in the device, selecting the same kernel code and data pages for checking. However, these pages are sent to a well-provisioned cloud server, which is idealized in two aspects. First, it replies immediately to the client: processing an arbitrary amount of rootkit detection logic consumes zero CPU cycles on the server. Second, we placed the server in the same LAN as the WiFi access point, resulting in small Internet RTT latencies.

Figure 10 compares the energy dissipation of the offloaded architecture with that of the balanced profile, and the case in which no security is enabled. We elided 3G from these experiments because: (a) previous work on cloud offload points to marginal energy gains at best using 3G [9, 18], and (b) 3G RTTs are substantially higher than WiFi RTTs in a LAN. For the browsing workload, cloud-offload presents a substantially higher energy overhead, due to the high frequency and volume at which kernel pages are sent. In spite of the idealized speed of the cloud server, network latency results in no gains in terms of windows of vulnerability. The results lead us to conclude that cloud-based rootkit detection is in principle more expensive than host-based detection, barring a fundamentally different approach.

Collaborative and behavior-based detection.

In keeping with the recent interest on behavior-based techniques for malware detection, researchers have investigated techniques tailored for mobile phones. The work of Bose *et al.* [15] and Kim *et al.* [28] are two such examples, which use a host-based agent that observes activities on the phone and reports anomalies such as forwarding SMS messages to external phone numbers or the deletion of important system files. The work of Kim *et al.* is an interesting complement to the security versus energy tradeoff studied in our work. They proposed a security tool that generates power signatures for applications running on a handheld device to detect energy-greedy anomalies caused by mobile malware such as Bluetooth worms.

The techniques developed in these works can possibly be used to inform the design of an *adaptive security versus energy profile* to complement the balanced profile discussed in Section 6. The key shortcoming of predefined profiles (*e.g.,* the balanced profile)

is that they rely on a set of assumptions about past rootkit behavior. If such profiles were to become standard in a software distribution, a vast user population will be bound to well-known profiles. Anecdotal observation indicates that most users never switch energy/performance profiles, pointing to a similar behavior for security versus energy profiles. Malware writers will thus be given the gift of a high-payoff and easy to study target.

An adaptive rootkit detection tool could leverage the techniques mentioned here to detect anomalous behavior unique to the mobile platform. We leave to future work a scenario where security transitions occur based on the perceived risk of user interactions. For example, a user browsing the internet might be considered a high risk scenario compared to a user placing a typical phone call. In this case, the tool can use this web browsing activity to automatically transition to a high security state to protect against malicious websites. During a normal phone call, the state may be transitioned to a low security mode. By automatically transitioning between power-saving and high-security modes, such an adaptive approach can protect against threats while also conserving battery power. It also provides the added benefit of not binding the device to fixed security versus energy profiles.

Smartphone app security.

Recent research advocates for *preemptive* approaches that aid distributors verify the security of smartphone apps before they are deployed. Although not a panacea to the security problem, preemptive certification as implemented by Kirin [19] and ScanDroid [21] can detect and discard a large fraction of malware before it reaches the phone. Such techniques can complement host-based detectors, which can run using conservative security versus energy profiles when "trusted" apps are downloaded and executed.

Trusted computing.

Recent proposals have suggested using Trusted Platform Module (TPM) hardware to certify the integrity of code executing on mobile devices [23]. TPMs can digitally sign the software stack executing on a mobile device, and transmit this signature as a proof of the security of the mobile device. While TPMs can complement host-based detectors by proving their presence on a mobile device, they cannot supplant them, because it is challenging to provide continuous integrity guarantees with a TPM (even with techniques such as Intel's TXT mode) [32].

Hardware advances.

ARM processors have a superior performance to Atom processors in terms of energy footprint. Once virtualization is available, porting our experiments to an Atom platform will likely result in smaller energy expenditures for host-based rootkit detectors. We have shown that the overheads imposed by security are tightly dependent on the hardware used and go beyond the processor. Security causes CPU contention, and this in turn prolongs workload runtime and keeps IO devices other than the processor, such as WiFi and 3G transceivers, turned on for longer. Multi-core mobile architectures [7] could altogether eliminate CPU contention and runtime overhead. It is unclear, however, what level of efficiency multi-core architectures will present, and what kinds of overhead will result from the aggressive execution of security checks on multiple cores.

8. CONCLUSIONS

This paper explored, for the first time, the tradeoff between security monitoring and energy consumption on mobile devices. We studied security versus energy tradeoffs for host-based detectors, focusing on rootkits, a particularly challenging class of malware. We proposed a framework to investigate security versus energy tradeoffs along two axes, attack surface and malware scanning frequency, and to measure the security being traded off. We applied our framework to complementary hypervisor-based code- and data-based rootkit detectors on a phone-like device. Our results show that protecting against code-driven attacks is relatively cheap, while protecting against all data-driven attacks is prohibitively expensive. We identified a sweet spot in the security versus energy tradeoff, one that minimizes both energy consumption and the window of vulnerability opened as a result. This *balanced profile* is able to detect a vast majority of known attacks, which work against code and selected kernel data structures, while consuming a limited amount of battery power. We conclude by motivating the need for new mechanisms to enable cloud-offload of rootkit detection, and by proposing the use of mobile-specific behavior-based anomaly detectors to transition between power-saving and high-assurance security modes.

Acknowledgments.

We thank Eyal de Lara, Ramón Cáceres, the reviewers of our MobiSys submission, and our shepherd Alec Wolman for useful comments on earlier drafts of this paper. We are also grateful to Ulrich Kremer and members of the EEL Lab in the Department of Computer Science at Rutgers University for help with the power measurement setup and experiments.

9. REFERENCES

[1] OKl4 microvisor. http://www.ok-labs.com/products/okl4-microvisor.

[2] Viliv S5 Real Pocket PC. http://www.myviliv.com/v4/product/s5/s5.asp.

[3] VMware Mobile Virtualization Platform. www.vmware.com/technology/mobile/.

[4] Rootkits, Part 1 of 3: A Growing Threat. http://download.nai.com/Products/mcafee-avert/whitepapers/akapoor_rootkits1.pdf, April 2006. MacAfee AVERT Labs Whitepaper.

[5] 2010 threat predictions. http://www.mcafee.com/us/resources/reports/rp-threat-predictions-2010.pdf, December 2009. MacAfee AVERT Labs Whitepaper.

[6] ABC News. Use Your Cell to Monitor Your Smart Home. http://abcnews.go.com/Technology/video/monitor-home-cell-phone-9887403.

[7] ARM. Cortex-a9 processor. http://www.arm.com/products/processors/cortex-a/cortex-a9.php.

[8] AT&T. AT&T, T-Mobile and Verizon Wireless Announce Joint Venture to Build National Mobile Commerce Network. http://www.att.com/gen/press-room?pid=18767&cdvn=news&newsarticleid=31369&mapcode=corporate|financial.

[9] B. Chun and P. Maniatis. Augmented Smartphone Applications Through Clone Cloud Execution. In *Proc. 12th Workshop on Hot Topic in Operating Systems*, May 2009.

[10] A. Baliga, V. Ganapathy, and L. Iftode. Automatic Inference and Enforcement of Kernel Data Structre Invariants. In *Proc. Annual Computer Security Applications Conference*, December 2008.

[11] A. Baliga, V. Ganapathy, and L. Iftode. Detecting kernel-level rootkits using data structure invariants. *IEEE Transactions on Dependable and Secure Computing*, 8(4), July/August 2011.

[12] A. Baliga, P. Kamat, and L. Iftode. Lurking in the Shadows: Identifying Systemic Threats to Kernel Data. In *Proc. IEEE Symposium on Security and Privacy*, May 2007.

[13] P. Barham, B. Dragovic, K. Fraser, S. Hand, T. Harris, A. Ho, R. Neugebauer, I. Pratt, and A. Warfield. Xen and the

Art of Virtualization. In *In Proc. 19th ACM Symposium on Operating Systems Principles*, 2003.

[14] J. Bickford, R. OHare, A. Baliga, V. Gannapathy, and L. Iftode. Rootkits on Smart Phones: Attacks, Implications and Opportunities. In *Proc. Workshop on Mobile Computing Systems and Applications*, February 2010.

[15] A. Bose, X. Hu, K. G. Shin, and T. Park. Behavioral Detection of Malware on Mobile Handsets. In *Proc. 6th Mobisys*, 2007.

[16] S. Chen, J. Xu, E. C. Sezer, P. Gauriar, and R. K. Iyer. Non-control-data Attacks Are Realistic Threats. In *Proc. USENIX Security Symposium*, August 2005.

[17] M. Christodorescu. *Behavior-based Malware Detection*. PhD thesis, University of Wisconsin-Madison, August 2007.

[18] E. Cuervo and A. Balasubramanian and D. Cho and A. Wolman and S. Saroiu and R. Chandra and P. Bahl. MAUI: Making Smartphones Last Longer with Code Offload. In *Proc. 8th Conference onMobile Systems, Applications and Service*, June 2010.

[19] W. Enck, M. Ongtang, and P. McDaniel. On Lightweight Mobile Phone Application Certification. In *Proc. ACM conference on Computer and Communications Security (CCS)*, 2009.

[20] M. Ernst, J. Perkins, P. Guo, S. McCamant, C. Pacheco, M. Tschantz, and C. Xiao. The Daikon System for Dynamic Detection of Likely Invariants. In *Science of Computer Programming*, 2006.

[21] A. P. Fuchs, A. Chaudhuri, and J. S. Foster. SCanDroid: Automated Security Certification of Android Applications. Manuscript, Univ. of Maryland, http://www.cs.umd.edu/~avik/projects/scandroidascaa.

[22] T. Garfinkel and M. Rosenblum. A Virtual Machine Introspection Based Architecture for Intrusion Detection. In *Proc. Network and Distributed Systems Security Symposium*, 2003.

[23] P. Gilbert, L. Cox, J. Jung, and D. Wetherall. Toward Trustworthy Mobile Sensing. In *Proc. Workshop on Mobile Computing Systems and Applications*, February 2010.

[24] M. Grace, Z. Wang, D. Srinivasan, J. Li, X. Jiang, Z. Liang, and S. Liakh. Transparent Protection of Commodity OS Kernels Using Hardware Virtualization. In *Proc. 6th Conference on Security and Privacy in Communication Networks*, 2010.

[25] O. S. Hofmann, A. M. Dunn, S. Kim, I. Roy, and E. Witchel. Ensuring operating system kernel integrity with osck. In *Proc. 16th Conference on Architectural Support for Programming Languages and Operating Systems*, March 2011.

[26] J. Hwang, S. Suh, S. Heo, C. Park, J. Ryu, S. Park, and C. Kim. Xen on ARM: System Virtualization Using Xen Hypervisor for ARM-Based Secure Mobile Phones. In *Consumer Communications and Networking Conference*, January 2008.

[27] N. L. Petroni Jr., T. Fraser, J. Molina, and W. A. Arbaugh. Copilot - a Coprocessor-based Kernel Runtime Integrity Monitor. In *Proc. USENIX Security Symposium*, 2004.

[28] H. Kim, J. Smith, and K. G. Shin. Detecting Energy-greedy Anomalies and Mobile Malware Variants. In *Proc. 6th Conference on Mobile Systems, Applications and Services*, 2008.

[29] G. Klein, K. Elphinstone, G. Heiser, J. Andronick, D. Cock, P. Derrin, D. Elkaduwe, K. Engelhardt, R. Kolanski, M. Norrish, T. Sewell, H. Tuch, and S. Winwood. seL4: Formal Verification of an OS Kernel. In *Proc. 22nd ACM Symposium on Operating Systems Principles*, October 2009.

[30] L. Litty, H. A. Lagar-Cavilla, and D. Lie. Hypervisor Support for Identifying Covertly Executing Binaries. In *Proc. 17th USENIX Security Symposium*, August 2008.

[31] LWN.net. A New Adore Root Kit. lwn.net/Articles/75990/.

[32] J. McCune, B. Parno, A. Perrig, M. Reiter, and H. Isozaki. Flicker: An execution infrastructure for TCB minimization. In *Proc. European Conference on Computer Systems*, April 2008.

[33] L. McVoy and C. Staelin. lmbench: Portable tools for performance analysis. In *Proc. USENIX Annual Technical Conference*, 1996.

[34] E. Monti. iPhone Rootkit? There's an App for That. http://sandiego.toorcon.org/index.php?option=com_content&task=view&id=48&Itemid=9, October 2010.

[35] G. Neiger, A. Santoni, F. Leung, D. Rodgers, and R. Uhlig. Intel Virtualization Technology: Hardware Support for Efficient Processor Virtualization. 10(3), August 2006.

[36] J. Oberheide and F. Jahanian. When Mobile is Harder Than Fixed (and Vice Versa): Demystifying Security Challenges in Mobile Environments. In *Proc. Workshop on Mobile Computing Systems and Applications*, February 2010.

[37] J. Oberheide, K. Veeraraghavan, E. Cooke, J. Flinn, and F. Jahanian. Virtualized In-Cloud Security Services for Mobile Devices. In *Proc. Workshop on Virtualization in Mobile Computing*, June 2008.

[38] N. Percoco and C. Papathanasiou. This Is Not the Droid You're Looking For... http://www.defcon.org/images/defcon-18/dc-18-presentations/Trustwave-Spiderlabs/DEFCON-18-Trustwave-Spiderlabs-Android-Rootkit-WP.pdf, July 2010.

[39] N. Petroni and M. Hicks. Automated Detection of Persistent Kernel Control-Flow Attacks. In *Proc. ACM Conference on Computer and Communications Security*, pages 103–115, October 2007.

[40] N. L. Petroni, T. Fraser, A. Walters, and W. A. Arbaugh. An Architecture for Specification-based Detection of Semantic Integrity Violations of Kernel Dynamic Data. In *Proc. USENIX Security Symposium*, August 2006.

[41] G. Portokalidisi, P. Homburg, K. Anagnostakis, and H. Bos. Paranoid Android: Versatile Protection For Smartphones. In *Proc. 26th Annual Computer Security Applications Conference*, 2010.

[42] R. Riley, X. Jiang, and D. Xu. Guest-Transparent Prevention of Kernel Rootkits with VMM-based Memory Shadowing. In *Proc. 11th Symposium on Recent Advances in Intrusion Detection*, September 2008.

[43] A. Seshadri, M. Luk, N. Qu, and A. Perrig. SecVisor: A Tiny Hypervisor to Provide Lifetime Kernel Code Integrity for Commodity OSes. In *Proc. 21st ACM Symposium on Operating Systems Principles*, November 2007.

[44] H. Shacham. The Geometry of Innocent Flesh on the Bone: Return-into-libc without Function Calls (on the x86). In *Proc. ACM Conference on Computer and Communications Security*, pages 552–561, October 2007.

[45] P.C. van Oorschot, A. Somayaji, and G. Wurster. Hardware-assisted circumvention of self-hashing software tamper resistance. *IEEE Transactions on Dependable and Secure Computting*, 2:82–92, April 2005.

[46] Z. Wang, X. Jiang, W. Cui, and P. Ning. Countering Kernel Rootkits with Lightweight Hook Protection. In *Proceedings of the ACM Conference on Computer and Communications Security*, November 2009.

[47] X. Zhang, L. van Doorn, T. Jaeger, R. Perez, and R. Sailer. Secure Coprocessor-based Intrusion Detection. In *Proc. 10th workshop on ACM SIGOPS European workshop: beyond the PC*, 2002.