Automatic OS Kernel TCB Reduction
by Leveraging Compile-Time Configurability

Reinhard Tartler\textsuperscript{1}, Anil Kurmus\textsuperscript{2},
Bernhard Heinloth\textsuperscript{1}, Valentin Rothberg\textsuperscript{1}, Andreas Ruprecht\textsuperscript{1}, Daniela Dorneanu\textsuperscript{2},
Rüdiger Kapitza\textsuperscript{3}, Wolfgang Schröder-Preikschat\textsuperscript{1}, and Daniel Lohmann\textsuperscript{1}

\textsuperscript{1}Friedrich-Alexander University Erlangen-Nürnberg
\textsuperscript{2}IBM Research - Zurich
\textsuperscript{3}TU Braunschweig

Abstract
The Linux kernel can be a threat to the dependability of systems because of its sheer size. A simple approach to produce smaller kernels is to manually configure the Linux kernel. However, the more than 11,000 configuration options available in recent Linux versions render this a demanding task. We report on designing and implementing the first automated generation of a workload-tailored kernel configuration and discuss the security gains such an approach offers in terms of reduction of the Trusted Computing Base (TCB) size. Our results show that the approach prevents the inclusion of 10\% of functions known to be vulnerable in the past.

1 Introduction
The Linux kernel is a commonly attacked target. In 2011, 148 Common Vulnerabilities and Exposures (CVE)\textsuperscript{1} entries for Linux have been recorded, and this number is expected to grow every year. This is a serious problem for system administrators who rely on a distribution-maintained kernel for the daily operation of their systems. On the Linux distributor side, kernel maintainers can make only very few assumptions on the kernel configuration for their users: Without a specific use case, the only option is to enable every available configuration option to maximize the functionality. The ever-growing kernel code size, caused by the addition of new features, such as drivers, file systems and so on, indicates that the risk of undetected vulnerabilities will constantly increase in the foreseeable future.

If the intended use of a system is known at kernel compilation time, an effective approach to reduce the kernel’s attack surface is to configure the kernel to not compile unneeded functionality. However, finding a fitting configuration requires extensive technical expertise about currently more than 11,000 Linux configuration options, and needs to be repeated at each kernel update. Therefore, maintaining such a custom-configured kernel entails considerable maintenance and engineering costs.

This paper presents a tool-assisted approach to automatically determine a kernel configuration that enables only kernel functionalities that are actually necessary in a given scenario. We quantify the security gains in terms of reduction of the Trusted Computing Base (TCB) size. The evaluation section (Section 3) focuses on an appliance-like virtual machine that runs a web server similar to those used to power large distributed web services in the cloud. Our approach exhibits promising security improvements for this use case: Compared with a default distribution kernel, 10\% of the kernel functions (i.e., 17 out of 179), for which in total 31 vulnerabilities have been reported, are removed from the tailored kernel.

The remainder of this paper is structured as follows: Section 2 presents the design and implementation of the first automated workload-specific kernel-build generation tool. Section 3 evaluates the usability of such an approach in a real-world scenario. Security benefits of the tailored Linux kernel are discussed in Section 4. Section 5 presents the related work. The paper concludes in Section 6.

2 Kernel-Configuration Tailoring
The goal of our approach is to compile a Linux kernel with a configuration that has only those features enabled which are necessary for a given use case. This section shows the fundamental steps of our approach to tailor such a kernel. The six necessary steps are depicted in Figure 1.

- **Enable tracing.** The first step is to prepare the kernel so that it records which parts of the kernel code are executed at run time. We use the Linux-provided \texttt{ftrace} feature, which is enabled with the \texttt{CONFIG} configuration option \texttt{CONFIG_FTRACE}. Enabling this configuration option modifies the Linux build process to include profiling

\textsuperscript{1}http://cve.mitre.org/
code that can be evaluated at runtime.

In addition, our approach requires a kernel built with debugging information so that any function addresses in the code segment can be correlated to functions and thus source file locations in the source code. For Linux, this is configured with the KCONFIG configuration option CONFIG_DEBUG_INFO.

Run workload. In this step, the system administrator runs the targeted application after enabling ftrace. The ftrace feature now records all addresses in the text segment that have been instrumented. For Linux, this covers most code, except for a small amount of critical code such as interrupt handling, context switches and the tracing feature itself.

To avoid overloading the system with often accessed kernel functions, ftrace’s own ignore list is dynamically filled with functions when they are used. This prevents such functions from appearing more than once in the output file of ftrace. We use a small wrapper script for ftrace to set the correct configuration before starting the trace, as well as to add functions to the ignore list while tracing and to parse the output file, printing only addresses that have not yet been encountered.

Correlation to source lines. A system service translates the raw address offsets to source line locations using the ADDR2LINE tool from the binutils tool suite. This identifies the source files and the #ifdef blocks that are actually being executed during the tracing phase. Technically, the tool stores its result to a text file with source-file names and line numbers on each line.

Establishment of the propositional formula. This step translates the source-file locations into a propositional formula. The propositional variables of this formula are the variation points the Linux configuration tool KCONFIG controls during the compilation process. This means that every C Preprocessor (CPP) block, KCONFIG item and source file can appear as a propositional variable in the resulting formula. This formula is constructed with the variability constraints that have been extracted from #ifdef blocks, KCONFIG feature descriptions and Linux Makefiles. The extractors we use have been developed, described and evaluated in previous work [5, 21, 22]. The resulting formula holds for every KCONFIG configuration that enables all source lines simultaneously.

Derivation of a tailored kernel configuration. A SAT checker proves the satisfiability of this formula and returns one concrete configuration that fulfills all these constraints. Note that finding an optimal solution to this problem is an NP-hard problem and was not the focus of our work. Instead, we rely on heuristics and configurable search strategies in the SAT checker to obtain a sufficiently small configuration.

As the resulting kernel configuration will contain some additional unwanted code, such as the tracing functionality itself, whitelists and blacklists are employed, allowing the user to specify additional constraints in order to force the selection (or deselection) of certain KCONFIG features. This results in additional constraints being conjugated to the formula just before invoking the SAT checker.

Compiling the kernel. The resulting solution to the propositional formula, obtained as described above, can only cover KCONFIG features of code that has been traced. As the KCONFIG feature descriptions declare non-trivial dependency constraints [25], special care must be taken to ensure that as many KCONFIG features as possible are not selected while still fulfilling all dependency constraints. We therefore use the KCONFIG tool itself to process this feature selection to a KCONFIG configuration that is both consistent and selects as few features as possible.

3 Practical Application

We evaluate the usefulness of our approach by setting up a Linux, Apache, MySQL and PHP (LAMP)-based web presence in a manner that is suited for deployment in a cloud environment. The system serves static webpages, the collaboration platform DOKUWIKI [7] and the message board system PHPBB3 [19] as an example for typical real-world applications. We use the distribution-provided packages from the Debian distribution without further specific configuration changes or optimization. Evaluation results are summarized in Table 1.

3.1 Kernel Tailoring

To derive a minimized kernel configuration, the first step consists of compiling a tracing-enabled Linux ker-
We use the standard Linux kernel source and configuration from the Debian distribution (version 2.6.32-41squeeze2) as a template for our tracing kernel (Step 1 in Figure 1). On this kernel, we enable the features CONFIG_FTRACE and CONFIG_DEBUG_INFO to include the ftrace tracing infrastructure and compile with debugging symbols. As our current prototype is not able to resolve functions from loadable kernel modules (LKMs) yet, we disable module support in the kernel configuration, which causes all compiled code to be loaded into the system at boot time.

Furthermore, a number of drivers cause compilation and linking errors when not compiled as LKMs. Most of these issues stem from drivers in the staging area. Also, when trying to boot this kernel, we observe kernel panics during the initialization of a range of watchdog drivers. As these drivers turn out to be unnecessary for this application scenario, we turn off the CONFIG options CONFIG_STAGING and CONFIG_WATCHDOG. These configuration changes account for the difference in size and features between the kernel shipped with Debian (~42 MB of code in the text segment) and the intermediary kernel that is used for collecting traces (~36 MB of code in the text segment).

With this intermediary tracing kernel, the system is tested against a test workload that covers all required functionality. We use the Skipfish [24] security analysis tool to systematically access all functionality of the appliance in an automated manner. This corresponds to Step 2 in Figure 1 and results in a total of 5,377 observed kernel functions.

These traced kernel functions correlate to 4,686 different source lines in 379 source files (Step 3). We use a modified version of the UNDERTAKER tool [22] to establish the propositional formula (Step 4) and to derive a solution for it (Step 5). To avoid unwanted functionality enabled in the resulting kernel, such as the ftrace infrastructure itself and LKM support, the UNDERTAKER tool obeys a blacklist that consists of the CONFIG options CONFIG_FTRACE and CONFIG_MODULES. Also, we add eight additional, use-case-agnostic CONFIG items to the whitelist in Step 6 to enable features that are used by the initialization startup scripts, which run before the system-wide tracing process starts. These steps take 69 sec on a commodity 2.8 GHz quad-core workstation with 4 GB of RAM.

Our tailored kernel is bootable and, as mentioned earlier, does not include any non-vulnerable functionality. This is achieved by removing all non-vulnerable functionality before we derive our solution (Step 5), as described in the subsection “Using Skipfish to derive the solution”. This true-to-source vulnerability scanning process allows us to apply the same process to various kernels. Furthermore, the necessary functions are included in our trace kernel for evaluation purposes.

3.2 Evaluation

To ensure the functionality of the appliance, we run the Skipfish [24] security scan again on the system with the tailored kernel, and compare the results with the previous run on the tracing kernel. The comparison of these two reports indicates no differences in the number of vulnerabilities or other issues.

The performance is tested with the httpperf tool [18]. The tool accesses a static website continuously, at a constant number of requests per second in each run. We did two setups of the same test scenario, both times using the same system, but once booted with the Debian standard kernel, and once with our tailored kernel. The data shows that our tailored kernel achieves a performance very similar to that of the original kernel provided by the distribution.
4 Discussion

After the presentation of a practical use case for our approach, this section now evaluates the security benefits. For this, we present an applicable security model to determine the TCB, and discuss security improvements in terms of TCB reduction.

Security Model. In the context of the web service presented, we assume both local and remote malicious attackers that target the kernel. However, we do not consider attackers that have physical access to the machine nor attacks that directly target hardware and firmware vulnerabilities.

The security goal is to prevent an attacker from gaining full control with arbitrary code execution in kernel mode, information leakage (e.g., recover uninitialized kernel memory content) to breach confidentiality, and denial-of-service attacks by crashing the kernel to reduce the availability of the system.

TCB sizes. Following the literature [13], we define the TCB as “the subset of components that need to be trusted to fulfill the security goals given in the security model”. Therefore, in the security model above, the TCB is solely composed of the kernel, including all LKM loaded during normal operation.

We apply three different metrics to measure the TCB reduction: a) the compiled code (text segment) size of the kernel, b) the total number of features that are enabled in CONFIG and c) the number of functions compiled into the kernel for which there has been an CVE entry in the past 10 years. More precisely, through a semi-automated process, we map a subset of 197 out of 873 CVE entries to vulnerabilities in 215 unique functions in the kernel, and use this dataset. The results for all three kernels used in the experiment in Section 3 are shown in Table 1.

Results. The data shows that the Linux kernel shipped by Debian loads 5.5 MB of program code into the memory for the virtual machine in the scenario described in Section 3. Compared with the code size of 4 MB for our tailored kernel, the total TCB size is reduced by 27%.

The number of features enabled is also reduced significantly, from 3,392 (with 1,093 features compiled statically into the kernel and 2,299 as LKM) to 379. The omission of functionality to load further LKMs constitutes an additional security benefit.

Finally, for each function in the TCB, we record the number of known vulnerabilities that have been reported in the past 10 years. When comparing the default distribution kernel to the tailored kernel, we observe a reduction of 10% of functions for which vulnerabilities have been reported in the past. However, this number is a lower-bound estimate, as the Linux kernel supports on-demand insertion of LKM, resulting in a higher initial TCB size, and therefore higher TCB reduction.

Sampling bias. Compared with the code size reduction results above, the CVE reduction numbers may seem lower than expected. We hypothesize that this impression can be attributed to sampling bias: code that is used more often is also audited more often, and better care is taken in documenting the vulnerabilities of such functions. A comparison of the average number of CVEs in kernel functions that are loaded and used (9.8‰) with the average number of CVEs in kernel functions that are not used (3.7‰) supports this hypothesis. Previous studies [3] have also shown that code in the drivers/ sub-directory of the kernel, which is known to contain a significant number of rarely-used code, on average contains significantly more bugs than any other parts of the kernel tree. Consequently, it is likely that unused features provided by the kernel still contain a significant amount of relatively easy-to-find vulnerabilities. This further confirms the importance of reducing the TCB size as presented in this work.

Unexpected impacts. The presented approach in this work could in turn cause a reduction of the security of the system – a drawback that is common to many security software but is often overlooked. Reviewing the process described in Section 2 (Step 9), we cannot rule out that for some application scenarios, performance-critical or security features might be removed from the base kernel. Possible reasons for this include that a) the feature was not triggered during the system-wide trace, b) the functionality has been excluded from the instrumentation with ftrace (e.g., for performance reasons), or c) the configuration options influence the resulting kernel in non-functional ways (e.g., different compilation flags, etc.). Although we were not able to find any results confirming this in this experiment — for example, we have verified that the CONFIG_CC_STACKPROTECTOR configuration option, which toggles the inclusion of the GCC flag for adding a stack frame canary, remains enabled, in future work we intend to further evaluate potential adverse impacts.

Applicability and Incomplete Traces. The presented approach relies on the assumption that the use-case of the system is clearly defined. Thanks to this a priori knowledge, it is possible to determine what kernel functionalities the application requires and therefore, what kernel configuration options have to be enabled. Still, the lack of guarantees that a trace of a given scenario is sufficient
and captures all required functionality may raise concerns. However, this does not invalidate the approach. Our approach works best for service providers that instantiate fairly homogeneous services, for which the chance to capture all necessary functionality is highest. Unfortunately, even the best tracing mechanism could still not guarantee catching all possible cases. Such a guarantee could only be provided with formal analysis and verification – a method that stands mathematical proof. However, for many scenarios organizations and system operators are likely to accept less costly approaches – such as the one presented in this paper. With the increasing importance of compute clouds, where customers employ virtual machines for very dedicated services such as the web server presented in Section 3, we expect that our approach can be easily applied to further use cases that are commonly deployed in the cloud.

Fortunately, incomplete traces do not lead to system crashes. With CONFIG, changing the static configuration of the Linux kernel statically restricts the provided functionality to applications that interact with the kernel via well-defined interfaces. These interfaces, and the careful maintenance of kernel developers, allows applications to check for the presence of kernel features that are required for correct operation. In the absence of kernel bugs, applications must not be able to crash the kernel. On the other hand, applications that crash because of missing kernel features fail to provide proper error handling – which alone makes them unsuitable for security sensitive environments. In summary, crashes that occur because of incomplete traces always stem from easy-to-address software bugs for the scenario presented in this paper.

Usability. Most of the steps presented in Section 2 require no domain specific knowledge of Linux internals. We therefore expect that they can be conducted in a straightforward manner by system administrators without specific experience in Linux kernel development. The system administrator, however, continues to use a code base that continuously receives maintenance in form of bug fixes and security updates from the Linux distributor. We therefore have confidence that our approach to automatically tailor a kernel configuration for specific use-cases is both practical and feasible to implement in real-world scenarios.

Extensibility. The experiment in Section 3 shows that the resulting kernel requires eight additional CONFIG options for proper operation. Alternatively to adding these features to the whitelist with distribution-specific knowledge, starting the application tracer at the start of the boot process would also capture the missing functionality. However, in this way we demonstrate the ability to specify wanted or unwanted CONFIG options independently of the tracing. This allows our approach to be assisted in the future by methods to determine kernel features that tracers such as ftrace cannot observe at all.

5 Related work

As we show below, this work relates to two research areas.

Kernel specialization. Several researchers have suggested approaches to tailor the Linux kernel, although security is usually not a goal, but improvements in code size or execution speed are: Lee et al. [14] manually modify the source code (e.g., by removing unnecessary system calls) based on a static analysis of the applications and the kernel. Chanet et al. [2], in contrast, propose a method based on link-time binary rewriting, but also employ static analysis techniques to infer and specialize the set of system calls to be used. Both approaches, however, do not leverage any of the built-in configurability of Linux to reduce unneeded code. Moreover, our approach is completely automated.

TCB reduction has always been a major design goal for micro-kernels [1, 15], which in turn facilitates a formal verification of the kernel [10] or its implementation in type-safe languages, such as OCaml [16].

Kernel attack surface reduction. The security model used in this paper is commonly used when building sandboxing or isolation solutions, in which each process must be contained within a particular security domain, such as [4, 9, 17], which are all based on the Linux Security Module (LSM) framework [23]. The idea of directly restricting the system call interface on a per-process basis has been previously explored as well, e.g., by Fraser, Badger, and Feldman [6], Ko et al. [11], or Provos [20], although not with specific focus on reducing the kernel’s attack surface. Seccomp [8] directly tackles this issue by allowing processes to be sandboxed at the system call interface. Ktrim [12] is a current research project which goes beyond simply limiting the system call interface, and explores the possibility of finer-granularity kernel attack surface reduction by restricting individual functions (or sets of functions) inside the kernel. In contrast, this work focuses on compile-time removal of functionality from the kernel at a system-wide level instead of a runtime removal at a per-application level.

6 Conclusion and Future Work

This paper presents an approach for automatically tailoring a Linux kernel configuration to a given use case. The result is a Linux kernel in which unnecessary functionality is removed at compile-time, hence significantly reducing
TCB size. The reduction can be quantified with 27% less code loaded and at least 10% fewer kernel functions which were previously vulnerable to attacks.

While the current prototype shows promising results, we intend to improve on the usability and applicability to additional use-cases. For instance, the current prototype unconditionally disables module loading support. As this may be undesirable in some cases, we intend to improve the handling of LKMs, as well as to remove the need for an intermediary tracing kernel.

Acknowledgments

This research has been partially supported by the TClouds project funded by the European Union’s Seventh Framework Programme (FP7/2007-2013) under grant agreement number ICT-257243.

References


