Pre-Patched Software

Jianing Guo Jun Yuan Rob Johnson
Stony Brook University

Abstract

Developing and deploying software patches is currently slow and labor-intensive. After software vendors discover a security bug in their product, they must manually write a patch, test it thoroughly, and distribute it to users, who may peform further testing at their site before finally installing the patch. These manual steps take time, leaving users vulnerable for days or even weeks after a bug is discovered. Pre-patched software removes these time-consuming steps from the vulnerability response critical path, reducing the window of vulnerability to hours or even minutes. With pre-patched software, applications ship with latent run-time checks that are automatically inserted during the compilation process. The compiler emits checks to cover any potentially-unsafe operation in the code. When the software vendor discovers a new vulnerability in its product, it can issue an alert informing its customers that they should activate one or more of the checks. Generating the run-time checks in advance removes the manual patch-development and testing processes from the vulnerability response critical path. Thus, when the vendor discovers a new vulnerability, it can immediately issue an alert and users can act on that alert without hesitation. By default, the run-time checks are disabled and hence incur little or no overhead. We have developed a CIL-based program-transformation that pre-patches C programs for memory-safety bugs. Early experiments suggest that pre-patched software may incur little measurable run-time overhead.

1 Introduction

Software makers currently have two imperfect ways to deal with security bugs in deployed code: run-time checks and patches. Run-time checks, such as those used by CCured[3] or Java, prevent attackers from exploiting security bugs but can significantly reduce run-time performance. Because of the performance costs, most applications are shipped without run-time checks. Patches, on the other hand, are created and applied after a defect is discovered, so they impose no run-time overhead, but creating patches is time-consuming and error-prone. Furthermore, many users, such as server administrators, test patches before installing them because patches may contain new bugs. This leaves a large window-of-vulnerability between the time a bug is discovered and the time that users are protected.

This paper describes pre-patched software, a new technique that combines the advantages of patching and run-time checks for dealing with defects in deployed systems. With pre-patched software, a program will ship with a set of latent run-time checks generated at compile-time and embedded in its code. The program can subsequently be "patched" simply by activating one or more of the latent checks. Until a check is activated, it will incur little or no run-time overhead.

The primary benefit of this approach is to move patch development and testing out of the critical path for responding to a newly discovered vulnerability. Since the run-time checks are generated at compile time and shipped with the original software, the vendor and user can test the patches in advance. When the vendor discovers a new vulnerability, it only needs to issue an alert informing its users to activate one of the latent checks. The users can act on the alert immediately and without hesitation. If users configure their computers to automatically respond to alerts, then only a few minutes may elapse between vulnerability discovery and patch installation.

2 Related Work

Numerous researchers have developed program transformations that insert run-time checks to enforce security properties: CCured[3] enforces type-safety, CRED[7] enforces memory safety, dynamic taint-tracking[9] prevents input validation exploits, RICH[1] prevents integer-overflows. These transformations can add significant run-time overhead; transformed programs may run 1.5 to 10 times slower.

Other researchers have developed methods for automatically generating network filters for observed attacks[6, 5, 8]. All such systems face the challenge of constructing as precise and general a filter as possible from only a few observed attacks: too narrow a filter may miss future variants of the worm, too broad will reject valid traffic. This has led to the notions of attack-based detectors, i.e. filters specific to a particular worm or attack, and vulnerability-based detectors, i.e. filters that can detect any worm that targets a particular vulnerability. Obviously, vulnerability-based filters are preferred. Brumley, et al. produced the first vulnerability-based filter-generator by slicing the vulnerable application to construct a pared down program that acts as a recognizer for inputs that excercise the relevant application execution-path[2]. Our system also provides vulnerability-based defense but, since the patches will be generated in advance, will offer faster response times and increased reliability.

Wang, et al's Shield project uses network filters to prevent attackers from exploiting known vulnerabilities[8]. Shield was specifically designed in response to the unreliability and irreversibility of software patching. Pre-patched software also addresses these issues, but goes even further. With pre-patched software, patches are reliable, reversible, and *predictable*. Pre-patched software offers other advantages, as well. Shield filters must be generated after a vulnerability is discovered and often duplicate a substantial amount of the vulnerable application's logic. Pre-patched software will generate patches in advance and will not duplicate application logic, offering faster response times and potentially lower overhead.

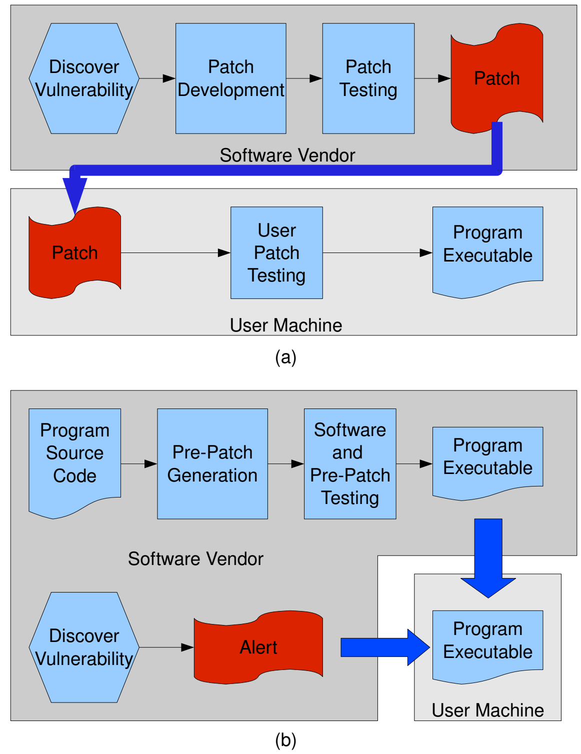

Figure 1: (a) The steps of traditional patch development and deployment. (b) Software development and "patch" deployment with pre-patched software.

3 Pre-Patched Software

Figure 1 presents the major differences between pre-patched software and conventional software patches. The traditional software distribution model makes no preparations for handling vulnerabilities discovered in deployed software. Thus, once the vendor discovers a vulnerability, it must develop a patch from scratch, test it thoroughly to ensure that the patch causes no regressions, and distribute the patched binary to its customers. Cautious customers may then conduct their own testing before finally installing the patch on production machines. These manual steps may take weeks or even months, during which time customers are vulnerable to attack. This is especially problematic when hackers discover the vulnerability before the vendor does.

Pre-patched software removes these steps from the critical path for responding to a newly-discovered vulnerability. A pre-patching compiler generates patches at compile time for potential bugs that may be discovered in the future. The vendor can test these patches, in isolation or combination, before shipping the software. The patches are shipped to customers with the software. Customers can perform further testing on the patches to ensure that the patches will not break functionality used at the customer's site. Once customers are satisfied that the patches are safe, they can configure their computers to automatically activate patches upon receiving an alert from the software vendor.

Generating patches at compile time. Several research projects have produced program transformations that insert run-time checks to catch security bugs at run time[4, 3, 10, 7, 9, 1]. A pre-patching compiler can use similar techniques, but must leave the inserted instrumentation latent until it is needed. The compiler can generate latent checks by guarding each instrumentation site with a branch conditioned on the value of some global variable. Each instrumentation site can have its own guard variable, enabling individual instrumentations to be activated separately. Alternatively, the compiler can generate the instrumentations, copy them to an unused portion of the executable, and replace them with NOPs in the original program. Instrumentations could then be activated by copying their code back into place in the main body of the code.

In either case, the overhead of checking guard variables or executing NOPs may be non-trivial. Since almost all the instrumentation will be disabled at any given time, we can optimize this common case by generating fast-path and slow path versions of each function. If the function has no active instrumentation, then it will execute the fast-path version, which will contain no instrumentation or NOPs and hence will run at full speed. If the function contains at least one active instrumentation site, then it will execute the slow-path version, which will contain full instrumentation.

Alert generation. The software vendor can use pre-patched software to respond quickly to zero-day worms and other exploits against their application. Upon discovering such an exploit, the vendor must generate an alert informing its customers to activate a check in the software. To determine which check must be activated, the vendor can simply run the exploit against an instance of the application with all instrumentation activated. The software will abort on some failing check. The vendor then issues an alert for that check. Thus, in the presence of a zero-day worm, the vendor can discover the relevant patch and issue an alert very quickly, perhaps in just a few minutes.

Limits of pre-patched software. Since the run-time checks must be generated automatically in advance, pre-patched software is not applicable to all kinds of security bug. For example, it would be difficult to create a program transformation to defend against high-level logic errors in the application. Pre-patched software is therefore most applicable to low-level security bugs that many applications wish to avoid, e.g. buffer overflows, format-string bugs, SQL injection bugs, cross-site-scripting bugs, etc. Also, pre-patched software is only as good as the pre-patching compiler. If the compiler only generates patches for one type of bug, then that is the only bug that can be handled with pre-patches in the deployed software. The vendor must respond to any other kind of bug with standard patches.

Pre-patched software is not intended as the final response to discovered vulnerabilities. Developers can almost always write a better patch, with lower overhead and better error response, than an automated compiler. Pre-patched software prevents worms and other exploits while the vendor develops a manually-crafted response to the vulnerability.

4 A C Memory-Safety Pre-Patcher

This section describes a prototype pre-patching program transformation we are developing to protect against buffer overflows in C. Our transformation follows the general approach of the Jones and Kelly[4] bounds checker, although we have made several enhancements described below. We chose to start with the Jones and Kelly bounds checker, as opposed to CCured[3] or the memory-safe C compiler[10], because Jones and Kelly's approach requires no changes to the input source code and has a high degree of compatibility between transformed and untransformed code. The Jones and Kelly bounds-checker's meta-data also has a particularly simple organization, making it easier to activate individual instrumentation sites.

The Jones and Kelly Bounds Checker. A bounds-checking transformation for C must track the bounds for pointers in the program and instrument certain pointer operations to ensure that no out-of-bounds pointer dereference occurs. The Jones and Kelly transformation stores all bounds information in an interval tree indexed by pointer values. As long as all pointers remain in their correct region, then lookups in this tree will return the correct bounds for any given pointer.

The Jones and Kelly bounds-checker instruments memory allocation and deallocation to register the newly allocated regions in the interval tree. Pointer arithmetic operations are annotated to verify that the result of the arithmetic points to the same region as the original pointer. Pointer dereferences are instrumented to ensure that the entire area read or written lies within the pointer's target region. Pointer assignments, including argument passing, require no instrumentation. Casts from pointer types to integer types require no instrumentation. Casts from integer types to pointer types and casts from one pointer type to another are un-checked.

Unfortunately, many C programs perform pointer arithmetic computations that yield intermediate results outside the original pointer's region, although the final result is within the region. Jones and Kelly do not support such programs. The CRED bounds-checker extended the Jones and Kelly bounds-checker to support such out-of-bounds pointers. With CRED, whenever a pointer goes out of bounds, its value is changed to point to a data structure that holds its true value and bounds. All other C pointer operations are modified to handle these OOB pointers.

Memsafe. Our transformation, which we call Memsafe, follows the Jones and Kelly approach with a few modifications. For each local pointer variable that does not have its address taken (which we call "solid" pointers), Memsafe creates a corresponding bounds variable that holds the bounds for that pointer. The bounds variable is updated whenever the pointer is. We cannot perform a similar optimization on non-solid pointers because they may change at any time due to aliasing and multithreading. For solid pointers, though, this optimization eliminates many lookups in the interval tree, which are the source of most overhead in the Jones and Kelly bounds checker. Whenever a solid pointer is assigned to a non-solid pointer, e.g. a global variable, we first check that the solid pointer is in bounds. This guarantees that all non-solid pointers are in-bounds and hence we can look up their bounds in the interval tree. Whenever a non-solid pointer is assigned to a solid pointer, we look up the non-solid pointer's bounds and store them in the solid pointer's bounds variable.

We also modify functions to accept bounds parameters corresponding to each pointer parameter. For backwards compatibility, we generate a wrapper function that expects no bounds parameters. This function looks up any missing bounds and calls the real function. This enables transformed code to inter-operate with untransformed code.

Storing bounds information for solid pointers reduces run-time overhead by eliminating many lookups and also enables our transformation to handle temporarily-out-of-bounds pointers in many cases. Source programs can generate and manipulate out-of-bounds pointers as long as those pointers are not dereferenced and are stored in solid variables. Experimental results in the next section show that many programs satisfy this requirement. This enables our transformation to support most programs without the complexity of CRED-style OOB structures.

Our current prototype performs several peephole and loop optimizations on the inserted instrumentation. Further optimizations, such as a CCured-like type-based proofs that some pointer-dereferences are safe, are also possible and may offer significant performance gains, although we have not implemented them yet.

Latent check implementation. Memsafe assigns each instrumentation site a unique index within a global bit-vector. The instrumentation at that site only executes if its corresponding entry in the vector is set. The transformation generates inter-instrumentation dependencies and encodes them as a data-structure inside the resulting object file. The global bit-vector is loaded from a file at program startup. The run-time system uses the dependency information to enable all supporting checks after initializing the bit-vector from disk.

Our transformation generates fast-path/slow-path code, as described in Section 3. Each function decides whether to run its fast path or its slow path upon entering the function. The function can also switch between fast and slow paths at the beginning and end of every loop. Thus, for example, if the function is executing along its slow path and reaches a loop that contains no active instrumentation, then the function will temporarily switch to the fast path for the duration of the loop. The optimal placement of switching points depends on the program's run-time behavior, so we chose the above heuristic since it is likely to give a good pay-off for relatively few switching points.

As with dependencies between instrumentations, there are also control-flow dependencies between switch points and instrumentations. Memsafe computes these dependencies at compile time and embeds them in the same dependency data structure as the inter-instrumentation dependencies. The run-time dependency resolution algorithm thus activates the correct set of switch points for any set of active instrumentations.

If a function contains no intrumentation, then we only generate a single path for it.

5 Evaluation

We implemented Memsafe as a source-to-source transformation using the CIL program analysis framework. Our evaluation focuses on two aspects of the transformation: correctness and performance.

Correctness. For backwards compatibility, a correct transformation should allow program executions that do not exhibit a memory error to execute normally. For security, a correct transformation should cause program executions with a memory error to halt as soon as the error occurs. Furthermore, we must verify that transformed code runs correctly with no checks enabled, all checks enabled, and arbitrary subsets of checks enabled.

Since memory errors taken from real-world programs can be brittle and highly-dependent on architecture and compiler details, we have chosen to evaluate the correctness and security of our transformation using a suite of simple test programs. Each test program accepts a command-line argument indicating whether it should execute code with a buffer overflow. The test programs are designed so that, when compiled with a normal compiler, the buffer overflows are all silent and harmless. We run each of these programs through the Memsafe transformation and verify that the resulting executables all satisfy the following requirements

- With all checks disabled and with no buffer overflow, the program executes normally.
- When run with all checks enabled and no buffer over-

	GCC	CIL	Memsafe	
			All off	All on
	Time	Ratio	Ratio	Ratio
bh	1.26	1.119	1.08	3.976
bisort	0.91	1.022	1.21	21.56
em3d	1.21	1.05	1.05	57.96
health	0.25	1.12	0.96	77.76
perimeter	0.63	1.206	2	15.75
power	0.95	1.074	1.22	1.547
treeadd	0.17	1.176	1.59	13.29
tsp	1.41	1.057	1.01	3.504
gzip	2.29	1	0.96	3.393
gunzip	1.33	0.737	1.06	1.496
Average	N/A	1.056	1.21	20.02

Table 1: Overhead of our transformation when compiled with optimization, i.e. gcc's -O3 mode.

	GCC	CIL	Memsafe	
			All off	All on
	Time	Ratio	Ratio	Ratio
bh	2.05	0.985	1.1	3.405
bisort	0.91	1.099	1.2	21.64
em3d	1.54	1.013	0.99	48.29
health	0.31	0.935	1.03	66.97
perimeter	2.07	0.932	1	5.657
power	0.95	1.063	1.22	1.558
treeadd	0.13	1.385	1.77	16.69
tsp	1.41	1.021	1.04	3.411
gzip	3.12	0.987	1.12	4.196
gunzip	1.15	1.296	1.04	2.565
Average	N/A	1.072	1.15	17.44

Table 2: Overhead of our transformation when compiled without optimization, i.e. gcc's -O0 mode.

flow occurs, the program completes its execution normally.

- When run with all checks enabled and a buffer overflow occurs, the program aborts due to a failed runtime check.
- We then turn off all checks except the failing check from the previous test (and any supporting checks), and re-run the program with a buffer overflow. It must abort as before.
- Finally, we run the program several times with random subsets of 10%, 20%, 30%, and 40% of the runtime checks enabled, but with no buffer overflow, and confirm that the program always executes normally.

Our test-suite currently contains 54 different tests, including several hand-written tests, tests derived from programs written by the authors for un-related projects, and gzip 1.2.4, which contains a known buffer overflow. All tests pass. The benchmarks used for the performance analysis described below serve as further evidence of the correctness of our transformation.

Performance. In the common case, there will be no known vulnerabilities in an application, so users will run the application with all checks disabled. Thus overhead in this configuration is the most important. Occasionally, the user will have a single check activated, along with its supporting checks, because of an outstanding vulnerability in the application. This overhead should be as low as possible, but it is less important than overhead with all checks off. Finally, during testing, the vendor or user may wish to run the application with all checks enabled. This overhead is not too important unless it is so high that the application is unusable. Thus, we measure the overhead of our pre-patching transformation in these three configurations: all checks off, one check enabled, and all checks enabled.

Table 1 shows the overhead of our transformation on the Olden benchmark when all checks are disabled or all checks are enabled. The mst benchmark from the Olden suite stores out-of-bounds pointers in non-solid variables, and so is not supported by our transformation. The transformation appears to incur non-trivial overhead (about 21%) when all checks are disabled, but we suspect that this overhead is primarily caused by interference with the GCC optimizer. To confirm this, we re-ran the experiments with GCC optimizations disabled. Table 2 shows that, in this configuration, the overheads are much lower, averaging 15%. It is also interesting to note from these tables that simply running the code through CIL can produce significant changes to the running time of a program – from nearly 30% faster to nearly 30% slower. From these experiments we conclude that a pre-patcher integrated with the compiler's optimizer can have very low overhead – as little as 8%.

Tables 1 and 2 show that the overhead when all checks are enabled can be very high. Although lower overhead would be better, this is acceptable because the software is rarely run in this mode.

Finally, Figure 2 shows the overheads when a single check is activated in our benchmark programs. Since the overhead will vary depending on which check is ac-

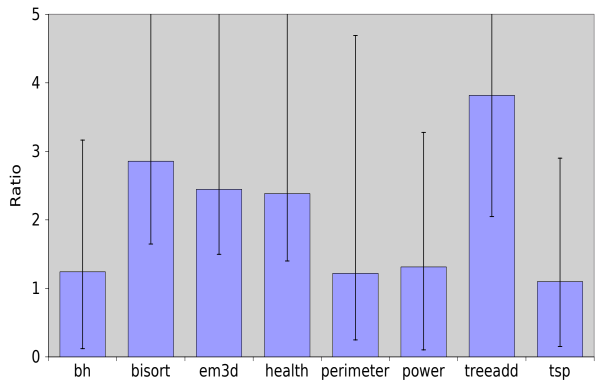

Figure 2: Memsafe overhead with one check enabled.

tivated, we re-ran each benchmark 100 times, activating a randomly chosen check on each execution. We report the average execution time ratio, along with error bars indicating the 95% confidence interval. The benchmarks are compiled without GCC optimizations. Overheads can vary significantly depending on the activated check, but on average the program runs twice as slow as when compiled with gcc. Although this is a relatively high overhead, the benefit is that, in the common case when all checks are off, the overhead is much lower, e.g. 8-15%, as shown in Table 2.

6 Conclusion

Pre-patched software turns the normal patching model on its head. By generating run-time checks in advance, but leaving them disabled until necessary, vendors can react quickly to newly-discovered bugs and worms without incurring a high run-time overhead.

Our prototype implementation demonstrates that pre-patching is a feasible mechanism for dealing with low-level bugs, such as memory-safety errors in C. The techniques developed for our prototype can be used to create a pre-patching compiler that addresses other security bugs.

References

[1] David Brumley, Tzi cker Chiueh, Robert Johnson, Huijia Lin, and Dawn Song. Rich: Automatically protecting against integer-based vulnerabilities. In *Proceedings of the 14th Annual Network & Distributed System Security Symposium*, 2007.

[2] David Brumley, James Newsome, Dawn Song, Hao Wang, and Somesh Jha. Towards automatic generation of vulnerability-based signatures. In *Proceedings of the 2006 IEEE Symposium on Security and Privacy*, 2006.

[3] J. Condit, M. Harren, S. McPeak, and etc. Ccured in the real world. In *Proceedings of the ACM SIGPLAN 2003 Conference on Programming L anguage Design and Implementation*, June 2003.

[4] Richard W. M. Jones and Paul H. J. Kelly. Backwards-compatible bounds checking for arrays and pointers in c programs. In M.K. Kamkar and D. Byers, editors, *Proceedings of the Third International Workshop on Automated Debugging*, volume 2–9 of *Linköping Electronic Articles in Computer and Information Science*, pages 13–26, Linköping, Sweden, May 1997. Linköping University Press.

[5] Zhenkai Liang, R. Sekar, and D. C. DuVarney. Automatic Synthesis of Filters to Discard Buffer Overflow Attacks: A Step Towards Realizing Self-Healing Systems. In *Proceedings of the USENIX Annual Technical Conference (to appear)*, 2005.

[6] J. Newsome and D. Song. Dynamic Taint Analysis for Automatic Detection, Analysis, and Signature Generation of Exploits on Commodity Software. In *The* 12^{th} *Annual Network and Distributed System Security Symposium*, February 2005.

[7] O. Ruwase and M. S. Lam. A practical dynamic buffer overflow detector. In *Proceedings of the Network and Distributed System Security (NDSS) Symposium*, pages 159–169, February 2004.

[8] Helen J. Wang, Chuanxiong Guo, Daniel R. Simon, and Alf Zugenmaier. Shield: Vulnerability-driven network filters for preventing known vulnerability exploits. In *Proceedings of ACM SIGCOMM*, August 2004.

[9] Wei Xu, Sandeep Bhatkar, and R. Sekar. Taint-enhanced policy enforcement: A practical approach to defeat a wide range of attacks. In *15th USENIX Security Symposium*, Vancouver, BC, Canada, August 2006.

[10] Wei Xu, Daniel C. DuVarney, and R. Sekar. An efficient and backwards-compatible transformation to ensure memory safety of c programs. *SIGSOFT Software Engineering Notes*, 29(6):117–126, 2004.