Adaptive Input Filtering: An Approach for Immunizing Servers from Repetitive Attacks

Zhenkai Liang, R. Sekar and Daniel C. DuVarney
Department of Computer Science
Stony Brook University
{zliang, sekar, dand}@cs.sunysb.edu

Buffer overflows have become the most common target for network-based attacks. Over 75% of the advisories issued by CERT Coordination Center last year were related to this class of attacks. Of particular significance is the choice of buffer overflows as the primary propagation mechanism used by worms. Although many techniques (such as StackGuard) have been developed to protect against buffer overflow attacks, all these techniques cause the victim process to crash. In the face of repetitive attacks such as those due to worms and "zombie machines," these protection mechanisms lead to repeated restarts of the victim application, in effect rendering its service unavailable. In contrast, we develop a novel approach that is inspired by the biological immune system. This approach is based on learning the characteristics of a particular attack, and developing a specialized "immune response" that is targeted at this attack. Since attacks on servers are launched via network packets, our approach is implemented as a protective layer that monitors inputs, learns the characteristics of inputs associated with attacks, and filters them out in the future. It can be implemented without changing the server code, or even having access to its source. Since attack-bearing inputs are dropped even before they corrupt the victim process, there is no need to restart the victim; as a result, recovery from attacks is extremely fast. Our experiments show that immunized servers can withstand repetitive attacks at a rate that is at least 10 times faster than previous approaches that relied on restart-based recovery. Our immunization technique is effective against buffer overflow attacks prevalent today, and produces very few false positives. Even when false positives were artificially introduced, our experiments show that the availability of the service is not significantly impaired, as each false positive appears like a (transient) network error from the perspective of the servers and clients.