Unencumbered by Success: The Usenix Security Grand Challenge Competition

Nicholas Weaver

TR-10-004

April 2010

Abstract

During Usenix Security 2009, I was part of a team (consisting of myself, Anup Ghosh, and Giovanni Vigna) which hosted a competition for an "Unhackable Server," sponsored by the National Science Foundation and BAE systems. Overall, I would rate the competition a failure, but a useful failure: what could have been a major disaster turned out to be only a minor embarrassment. This was due to multiple factors, including misjudging the difference between a competition of skill verses a competition of artifacts, lack of publicity, a poorly chosen prize amount, neglecting to account for the coolness vs. money tradeoff, and some competition logistic difficulties. Yet at the same time, the final results were not purely negative: the competition could be perceived as a success and we learned critical lessons for future competitions.

This work was made possible by National Science Foundation grant CNS: 0749648 ("Architecting Effective Computer Security Grand Challenge"). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors or originators and do not necessarily reflect the views of the National Science Foundation or the other contest organizers.

1 Introduction

"Failure is always an option" -Adam Savage

During Usenix Security 2009, I was part of a group which hosted an "Unhackable Server" competition, sponsored by the National Science Foundation with prize money provided by BAE systems. This competition intended to serve two purposes: improve the state of the art in computer security and to better understand problems faced in running competitions. Unfortunately, when it came to improving the state of the art the competition was a gross failure. But there were many lessons learned that need to be considered in future competitions.

2 The Competition Format

The format was reasonably simple, a "hands-off" Capture the Flag type competition, where the participants had first three hours and then overnight to secure a Linux virtual machine running a suite of custom services, including a PHP web application, a C server, and services written in other languages. Although the general setup was known in advance, the specific applications were not revealed to the competitors until the competition started, and the competitors were provided with a virtual machine which included all these services.

The scoring was based on sustained availability: a process would continuously write one or more "flags" (a key-value pair) and then later check that the flag was still present and reachable. The services themselves had multiple vulnerabilities that could be exploited, and a second process randomly selected exploits to launch at the services. Availability, rather than just exploitability, was measured because it represented both a harder metric and meant that we only needed to write crash exploits for the services.

The goal was to foster *automatic* techniques, since the short time limit would hopefully preclude human analysis of the code.

3 Problem: Competitions of Skill vs Competitions of Artifacts

An important distinction that we overlooked is the difference between a competition of skill and a competition of artifacts. A competition of skill, such as the Defcon CTF competition [2], is focused on the skill set of the competitors.

A competition of artifacts, however, is focused on the participants providing a device capable of performing the task. A classic example is the DARPA urban challenge [1], where participants needed to provide a self-driving car.

The intent of our competition was to develop a artifact-based competition for security, one where the participants provided the ability to quickly and automatically harden a base system, running custom services, from a wide variety of attacks.

Although the intent was sound, our implementation was lacking. For although we intended to create an artifact-based competition, our experience was in conducting a skill based competition, which lead to three significant errors: a far too weak problem specification, insufficient early publicity, and a far too small prize pool.

3.1 Specifications

A contest of skill can have very fuzzy specifications, as it can be adapted and changed based on human behavior. For example, the Defcon "Capture the Flag" competition has almost no specifications in advance on the services that the contestants are protecting, and may not even specify the base operating system ahead of time. Often a one or two page rule sheet is sufficient.

Artifact-based competitions, however, need very precise specifications, both for intent and implementation. Implementation is obvious, for example, the NASA space elevator tether competition doesn't require making a super-strong cable, but a specific cable: a 1m loop, weighing no more than 2 grams, that is destructively tested on a specific apparatus.

But even intent needs to be specified precisely. The DARPA Urban Challenge wasn't simply "Make a self-driving car", but make a driving system which would accomplish a host of challenges. As important were items the vehicle did NOT have to do, such as search for pedestrians or traffic signals. Thus the basic rulebook, not even including the scoring criteria, was 25 pages long![1]

In retrospect, we should have provided a much narrower and more tightly defined scope. One possibility would be rather than trying to protect an arbitrary set of services, instead limit the test service to a custom PHP application running on a specified web server, specified PHP version, and with a specified back-end

[1]Even then, criteria were changed dynamically to accommodate gaps in the rules , as the blue cars were removed from the background traffic in the competition because the blue paint was not detectable by some competitors' laser rangefinders, and the rules never specified the colors of cars.

database. This would have reduced the scope of the problem if it was a skills-based competition, but would make a artifact-based competition practical.

3.2 Preplanning and Publicity

Artifact-based competitions also require substantially more lead-time for participants. Where a skills competition may require a variable amount of preplanning (eg, some participants may train considerably or may develop tools for the task at hand), it is at least possible for participants to come in almost entirely unprepared.

Artifact-based competitions, however, require much longer lead time. Even a simple artifact competition needs several months of preplanning. Even in elementary school, kids don't just walk up to the science fair unprepared, as that is effectively an artifact-based competition.

As a consequence, it is critical that an artifact-based competition be publicized well in advance. In particular, we would suggest that advertising begin *at least a year* before the competition is scheduled.

3.3 Prize Selections

An artifact-based competition also involves considerably more effort on the participant's part, and thus requires a commensurably greater prize. Prizes actually take three forms, the direct compensation, the possibility of future compensation, and the "coolness benefit", and it is the total "prize package" which matters.

Direct compensation is obvious, as this is the money provided by the contest organizers or sponsors. Although critical, such prizes are not necessarily essential depending on the competition, but, as it is the only portion controlled by the organizers, it may need to be substantial in order to create a total prize package.

The competition itself also influences indirect future compensation, if the competition will involve developing a saleable technology. A good example is the NASA space elevator tether competition, as any winning entry will be able to earn 10x to 100x the prize money by having proven a fiber with 50% better strength to weight when compared with all commercially available fibers. This was also a significant motivator in the DARPA urban challenge, as winning competitors would also be demonstrating near salable technology for self-driving vehicles, especially for military convoy contracts.

The final portion is the "coolness benefit", a combination of both bragging rights and how fun the contest is. An example of a competition where almost the entire prize package is "coolness" was the Defcon Bots competition, where

participants needed to make a control system for a self-aiming gun, which would shoot a series of targets.

The total prize package also affects the type of participants. If only enthusiastic amateurs (such as the Defcon Bots competition) are targeted, a small prize package is acceptable. But if targeting professionals, the prize may need to be substantially larger.

In retrospect, we would argue that a good total prize package for an artifact based competition targeting professional entrants (such as a security competition) should probably be roughly $100,000.

4 Problems Encountered

Overall, the competition was almost a disaster, as several critical mistakes, including a too-vague definition, too-small prize money, too little publicity, and logistic difficulties all plagued the competition.

4.1 Problem: Problem-Space Definition

The first problem was a too vague and too broad problem definition. The "Hackproof computer" was specified as running linux and a broad suite of custom services, all which needed to be defended, and detailed in only the vaguest ways. In retrospect, it is probably impossible for an automated tool suite to be effective, given the very broad and overly vague problem definition.

In retrospect, the scope should have been much smaller and much better defined, eg, something like "a large scale, multi-user web bulletin board system, written in PHP, using a MySQL database and apache web servers. The only vulnerabilities are contained within the PHP application." And the problem statement should have included a more detailed specification. Although much narrower in scope, such a problem definition would have a reasonable chance of being addressed by an automatic tool suite.

4.2 Problem: Prize Magnitude

A related problem was the prize magnitude, both overall and with regard to the indirect prize component. By being rather "uncool", there was no substantial bragging rights for the winner. Since the problem description was too vague, this

was unlikely to produce a valuable tool. Thus the total prize pool was effectively limited to the monetary prize provided by the sponsor.

As such, $10,000 total compensation was simply too low for a significant artifact-based content. Assume it only takes four weeks to create an entry (which is low by the standards of many such contests), and there are 10 contestants. This translates to an expected hourly compensation of only $6.25/hr, which is too small for the expertise we envisioned.

4.3 Problem: Pre-Competition Publicity

A third problem was simply a lack of early publicity. We started publicizing the event only a few months before it was scheduled to be held. Although this timeframe would have been sufficient for a sills-based competition, it was simply too short for an artifact-based competition. In retrospect, we should have had publicity and prizes arranged a year in advance.

4.4 Problem: Competition Logistics

The competition itself was also beset with logistic difficulties involving participants transferring large (¿1 GB) Virtual machine images. Out of the four participating teams, one team effectively forfeited the second round because of a corruption issue on their image and an inability to contact them for correction.

Although we tested the scoring server and test applications, we never tested the ability to transfer these large files, simply taking it for granted. In retrospect, we should have allowed participants to log into the virtual machines directly to modify them in place and we should have tested every aspect of the competition: the only aspect we didn't test was the one which proved difficult.

5 Success: Contest Transformation

If the competition remained an artifact-based competition, the result would have been complete and embarrassing failure, with no valid participants even making an attempt. Any one of the three major mistakes, a lack of timely publicity, an insufficient total prize pool, or a too vague description, would have been sufficient to make the competition a total failure.

Fortunately, a happy coincidence changed the fate of the competition. Although intended as an artifact-based contest, we constructed it like a skills-based

competition, and it became a skills-based contest: using the participant's ingenuity, they had first 3 hours and then overnight to harden the applications from attack. Four teams participated, and three successfully completed the competition, and all three completing teams outscored the "null" entry (which was run simultaneously for calibration purposes).

The difference in preparation requirements was substantial. In fact, the winning team, *ad hoc*, was formed at the start of the conference! Likewise, what was too small a prize pool for an artifact competition was excellent motivation for a skill-based competition: an all-night hacking session for a chance to win $10,000 is a very attractive proposition. And the vague nature of the challange is not a handicap for human-competitors, but rather a feature.

Thus I have to conclude that the result was only a minor disaster: we didn't get the competition we desired, but we got a competition, with mulitple participants, and successful winners. The state of the art in computer security was not furthered, but we learned important lessons that future competitions should apply.

6 Disclaimer

Although the competition was funded by NSF grant 0749648 all opinions in this document are those of the author and not the funding institution or the other contest organizers.

References

[1] The DARPA Urban Challenge, http://www.darpa.mil/grandchallenge/index.asp.

[2] DEFCON, http://www.defcon.org/.