DARPA Cyber Grand Challenge:
CQE Scoring Document

Defense Advanced Research Projects Agency
Information Innovation Office

July 7, 2014
Version 1.1

Document Change Summary

Section	Description	Date
1.1.1	"Original CB" → "Reference Patched CB" in table showing examples	July 7, 2014
References	Fixed reference to CGC Rules document	July 7, 2014
	Initial release	April 21, 2014

Introduction

This document describes the scoring algorithm for Cyber Grand Challenge Qualifier Event (CQE). DARPA held a public comment period for the Cyber Grand Challenge (CGC) scoring algorithms and integrated said feedback into the final scoring algorithms.

1 CQE Scoring Method

CQE scoring is the product of three assessed quantities: **Availability Score**, **Security Score**, and **Evaluation Score**. CQE Scores will be assessed per Challenge Binary (CB); the total score for a Cyber Reasoning System (CRS) at the end of CQE shall be the sum of that CRS's Replacement CB scores. A team is eligible for a non-zero CB Score only if they submit a Replacement CB that mitigates at least one Reference Proof of Vulnerability (PoV). Each CB Score will be calculated as follows:

$$\text{CB Score} = \text{Availability Score} \times \text{Security Score} \times \text{Evaluation Score}$$

1.1 Availability Score

Availability Score measures performance in Area of Excellence (AoE) 4, as described in the CGC Rules, Section 1.3 [1].

This quantity shall vary as a multi-step function between 0 and 1, with 1 being a perfect score. Performance and retained functionality will be measured, with Availability being set to the minimum of these quantities. Competitors are advised that there is a faster-than-linear Availability score dropoff; for example, a 50% impact will more than halve the availability score.

$$\text{Availability Score} = min(\text{PerfScore}, \text{FuncScore})$$

1.1.1 Performance Score

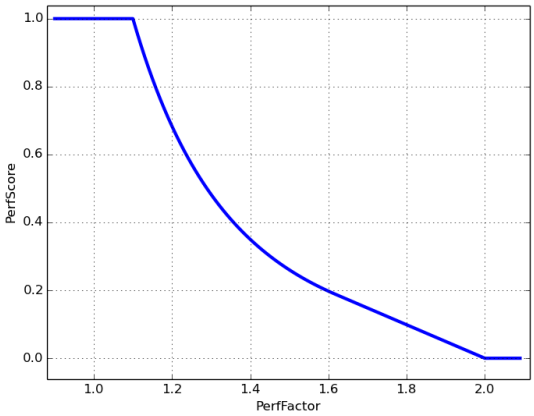
Performance score will be the worst negative impact of file size, execution time, or memory usage with a 40%, 10%, and 10% grace factor against each respective metric. Performance measures will compare Replacement CB against Reference Patched CB.

$$\text{FileSizeOverhead} = \frac{file\_size(\text{Replacement CB})}{file\_size(\text{Reference Patched CB})} - 1$$

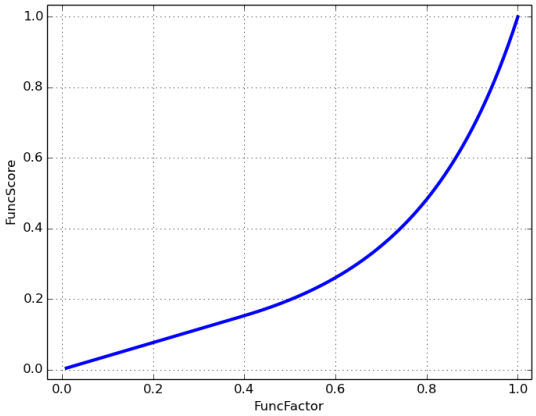
$$\text{MemUseOverhead} = \frac{mem\_use(\text{Replacement CB})}{mem\_use(\text{Reference Patched CB})} - 1$$

$$\text{ExecTimeOverhead} = \frac{exec\_time(\text{Replacement CB})}{exec\_time(\text{Reference Patched CB})} - 1$$

$$\text{PerfFactor} = 1 + max(0.25 \times \text{FileSizeOverhead}, \text{MemUseOverhead}, \text{ExecTimeOverhead})$$



(a) PerfFactor to PerfScore conversion curve.



(b) FuncFactor to FuncScore conversion curve.

Figure 1: Curves illustrating conversions for Performance and Functionality.

$$\text{PerfScore} = \begin{cases} 1 & 0 \leq \text{PerfFactor} < 1.10 \\ (\text{PerfFactor} - 0.1)^{-4} & 1.10 \leq \text{PerfFactor} < 1.62 \\ (-0.493 \times \text{PerfFactor} + 0.986) & 1.62 \leq \text{PerfFactor} < 2 \\ 0 & 2 \leq \text{PerfFactor} \end{cases}$$

The performance-related factors will assess the following measurements:

$file\_size(CB)$ Size of a CB as stored on disk.

$mem\_use(CB)$ Size of maximum memory used by a CB while running.

$exec\_time(CB)$ CPU time taken by a CB when responding to service polls drawn from the functionality test suite for CB.

For CBs consisting of multiple binaries, the value used for computing score factors will be the sum of the corresponding measurements for constituent binaries.

For a graphical representation of the PerfScore function, please see Figure 1a. As an example, consider the following Replacement CBs:

Name	$file\_size$	mem_use	$exec\_time$	**PerfFactor**	**PerfScore**
Reference Patched CB	100	100	100	1.00	**1.00**
Replacement CB 1	120	105	105	1.05	**1.00**
Replacement CB 2	200	105	105	1.25	**0.57**
Replacement CB 3	120	120	120	1.20	**0.68**
Replacement CB 4	100	160	130	1.60	**0.20**
Replacement CB 5	140	120	175	1.75	**0.12**

1.1.2 Functionality Score

Retained functionality will be a measure of how many test cases the Replacement CB passes. Note that original CBs and Reference Patched CBs will pass all tests in the functionality test suite.

$$\text{FuncFactor} = \frac{count(passed\_test\_cases(\text{Replacement CB}))}{count(passed\_test\_cases(\text{Reference Patched CB}))}$$

$$\text{FuncScore} = \begin{cases} 1 & \text{FuncFactor} = 1 \\ (2 - \text{FuncFactor})^{-4} & 0.40 \leq \text{FuncFactor} < 1 \\ 0.381 \times \text{FuncFactor} & 0 < \text{FuncFactor} < 0.40 \\ 0 & \text{FuncFactor} = 0 \end{cases}$$

For a graphical representation of this function, please see Figure 1b.

1.2 Security Score

Security Score measures performance in AoE 2, as described in the CGC Rules, Section 1.3 [1]. Security Score will be computed as follows:

$$\text{Security Score} = \begin{cases} 0 & \text{if ReferenceScore} = 0 \\ 1 + \frac{1}{2} \times (\text{ReferenceScore} + \text{ConsensusScore}) & \text{otherwise} \end{cases}$$

$$\text{ReferenceScore} = 1 - \frac{\text{Number of Reference PoVs which prove vulnerability in Replacement CB}}{\text{Number of Reference PoVs for this CB}}$$

$$\text{ConsensusScore} = \begin{cases} 0 & \textbf{Any} \text{ Submitted PoV proved vulnerability in Replacement CB} \\ 1 & \textbf{No} \text{ Submitted PoV proved vulnerability in Replacement CB} \end{cases}$$

1.3 Evaluation Score

Evaluation Score measures performance in AoE 3, as described in the CGC Rules, Section 1.3 [1].

$$\text{Evaluation Score} = \begin{cases} 1 & \text{Submitted PoV } \textbf{did not} \text{ prove vulnerability in CB} \\ 2 & \text{Submitted PoV } \textbf{did} \text{ prove vulnerability in CB} \end{cases}$$

2 Additional Information

CGC teams will be able to interact with the CQE scoring system during Scored Events, scheduled to precede CQE – see CGC Rules, Section 3.1.1 for details and dates [1].

In addition to the CQE scoring algorithm presented in this document, DARPA will release a set of example CBs with corresponding Reference PoVs prior to the first Scored Event. Each set will include several Reference Patched CBs with data for associated FuncScore and PerfScore component measurements.

To mitigate an unlikely event of a tie, DARPA will release a tie-breaker algorithm before CQE.

References

[1] Cyber Grand Challenge Rules, `https://cgc.darpa.mil/documents.aspx`

Glossary

AoE Area of Excellence.

CB Challenge Binary.

CGC Cyber Grand Challenge.

Challenge Binary A vulnerable network service that accepts remote network connections, composed of one or more communicating binaries.

CQE Cyber Grand Challenge Qualifier Event.

CRS Cyber Reasoning System.

Cyber Reasoning System Unmanned systems that autonomously reason about novel program flaws, prove the existence of flaws in networked applications, and formulate effective defenses.

PoV Proof of Vulnerability.

Proof of Vulnerability An input that activates and proves the existence of a hidden flaw in a CB.

Reference Patched CB DARPA's solution for a CB.

Reference PoV PoV supplied by CB author.

Replacement CB Solution for a CB supplied by a team's CRS.

Submitted PoV PoV supplied by a team's CRS.