CISS 2024 Report

Organised by:

The Digital and Intelligence Service, Ministry of Defence, Singapore

iTrust, Centre of Research in Cyber Security, SUTD, Singapore

Sponsored by:

Cyber Security Agency of Singapore

Date:

17 Sep to 24 Sep 2024

Report by:

ME4 Francisco FURTADO, Andy TAY, Ivan CHRISTIAN

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN | iTrust Centre for Research in Cyber Security

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. Executive Summary

The eighth iteration of iTrust's annual Critical Infrastructure Security Showdown (CISS) 2024 was held from 17 to 24 September 2024 at the Singapore University of Technology and Design (SUTD). CISS' objectives are to enhance the understanding of cyber threats, and develop and validate technologies and capabilities for defending critical infrastructure against cyber-attacks.

A New Format

CISS 2024 introduced a new 48-hour Capture-The-Flag (CTF) format. This change aimed to provide a more continuous and immersive experience for participants. Additionally, CISS introduced a new script submission challenge, requiring teams to develop and execute scripts to achieve specific attack objectives on iTrust's physical critical infrastructure testbeds.

New Collaborations

To enrich the competition, CISS 2024 collaborated with local and international partners, including iTrust's sister lab, the National Cybersecurity R&D Laboratory, Deutschlands Bester Hacker, Illinois ARCS, and IllusionIQ. These collaborations led to the introduction of innovative challenges, such as honeypots and advanced cyber-physical testbeds.

Enhanced Evaluation Framework

An enhanced 5-metric evaluation framework was implemented to assess the performance of Intrusion Detection Systems (IDS). This framework considers factors such as correctness, explainability, accuracy, responsiveness, and disruptiveness.

Looking Forward

CISS 2024 has demonstrated the value of continuous innovation and collaboration in such global nature as the field of cyber security. By adapting to evolving threats and technologies, CISS will continue to be a valuable platform for gathering like-minded cyber professionals to work towards the goal of critical infrastructure protection. iTrust would like to express its gratitude to the Cyber Security Agency of Singapore (CSA) and the Digital Intelligence Service (DIS) for their continued support and guidance.

2. Event Overview

The eighth iteration of CISS was held from 17 September 2024, 0900 hrs to 19 September 2024, 1500 hrs, at the Singapore University of Technology and Design. The showdown was designed to enhance the understanding of composite Tactics, Techniques and Procedures (TTPs) for operation security, validate and assess the effectiveness of technologies developed by researchers associated with iTrust, and develop capabilities for defending critical infrastructure against cyber- attacks.

This year's 48-hour CTF format introduced a new challenge category involving script submissions. 46 Red Teams, each comprising up to eight members, competed to solve a series of Operational Technology (OT) focused challenges. Points were awarded for successfully completing challenges. In one of the challenges, the Red Teams were tasked to develop scripts to achieve an attack objective in iTrust's physical testbeds.

In addition to the Red Teams, a group of Intrusion Detection Systems (IDS) teams, comprising commercial products and iTrust's in-house solutions, were tasked with detecting the attacks launched by the scripts submitted by the Red Teams. An enhanced 5-metric evaluation framework was introduced to assess the performance of these IDS systems.

3. Capture-The-Flag (CTF)

CISS 2024 saw a new record number of 46 Red Teams (11 government, 18 industry and 17 academia) comprising 322 participants. The competition, themed "The Orthanc Obstacles," continued the Lord of the Rings narrative from 2023.

The CTF format featured 53 challenges across 7 categories, totalling 4,010 points. Table 1 shows the breakdown of these categories. Teams were awarded points for successfully completing challenges. In the event of ties, the time of completion the challenge was used as a tiebreaker.

Table 1: CTF Category Breakdown

Category	Number of Challenges
Dunharrow Brigade HQ	12
IllusionIQ	12
Master Challenge Questions	10
Orthanc Systems	8
Palantir Control	2
Rohan Powerstone Network	7
Script Submission	2

3.1. Dunharrow Brigade HQ

The Dunharrow Brigade HQ category were contributed by Deutschlands-Bester-Hacker and presented 12 challenges that tested a wide range of cyber security skills, including web exploitation and forensics.

3.2. IllusionIQ

The IllusionIQ category, contributed by IllusionIQ, featured 13 challenges. These challenges were aimed at testing the participants' ability to handle advanced cyber-physical testbeds and honeypots which included Modbus, OpenPLC and SCADASim.

3.3. Master Challenge Questions

The Master Challenge Questions category presented 10 foundational multiple-choice questions, each worth 20 points. Teams were allowed two attempts per question. These questions covered fundamental OT protocols like Modbus, Ethernet/IP, IEC 61850, and BACnet. Figure 1 illustrates a sample question involving a state machine for a motorised valve.

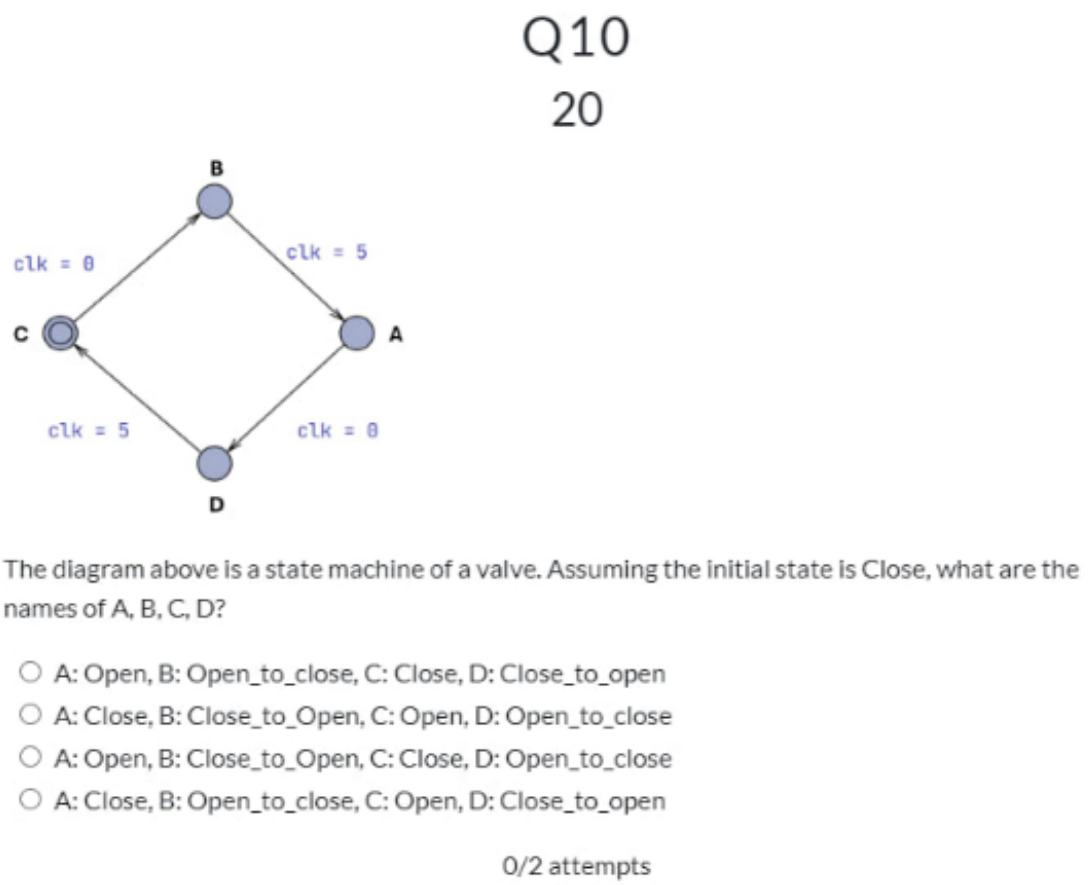

Figure 1: Sample question from the Master Challenge Questions category

3.4. Orthanc Systems

The Orthanc Systems category presented eight challenges focused on understanding and manipulating ICS protocols. Participants were tasked with decoding and analysing ICS traffic, identifying vulnerabilities, and executing attacks.

Control Process Challenges

Participants were required to control physical systems, such as flooding a tank or manipulating valves, by understanding the underlying control logic and exploiting vulnerabilities in the system.

Historian Analysis

Participants were challenged to analyse historical data to identify anomalies, security incidents, and potential threats. For example, they might be asked to determine the exact time of an attack on a specific sensor or device.

Packet Capture Analysis

Participants were tasked with analysing network traffic to identify communication patterns, protocol usage, and potential security risks.

Reverse Engineering

Participants were challenged to reverse engineer binary files to understand their functionality and identify potential vulnerabilities.

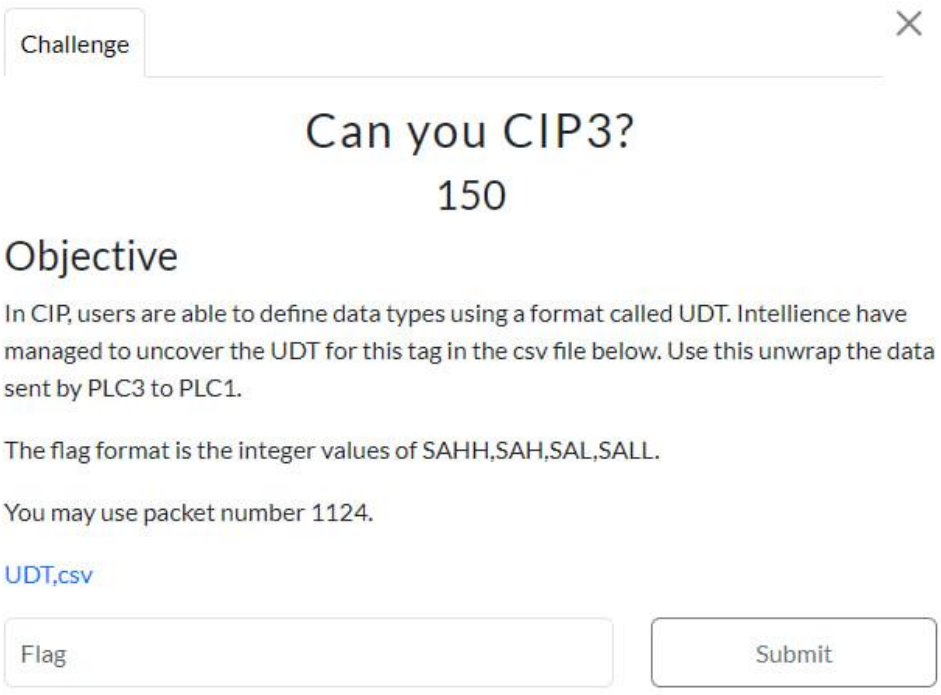

Figure 2: Sample ICS Protocol Decoding question from the Orthanc Systems category

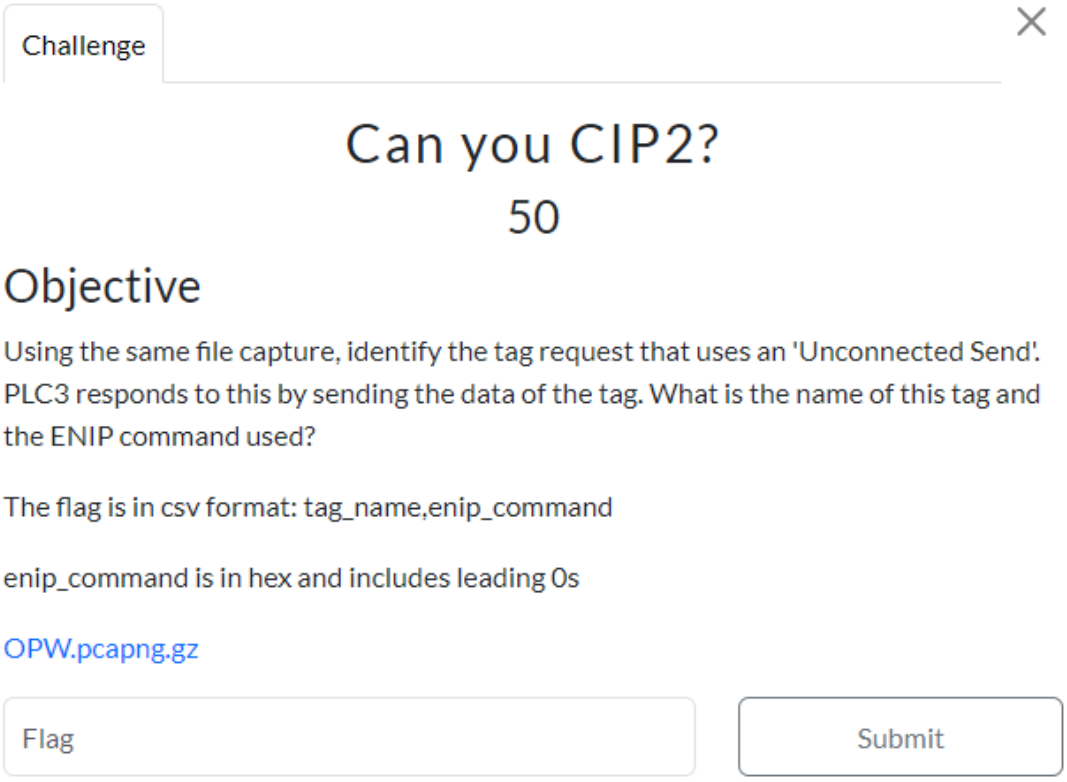

Figure 3: Sample of an ICS Protocol communication question

3.5. Palantir Control

The Palantir Control category, contributed by National Cybersecurity R&D Laboratories (NCL), presented a challenging scenario involving a mini IT-OT environment as shown in Figure 4. This environment, consisting of a Programmable Logic Controller (PLC) and a remote controller programme, emulated a simplified SCADA system.

Participants were tasked with navigating a series of interconnected challenges:

1. Network Reconnaissance: Identifying vulnerable systems and services through network scanning.
2. Protocol Exploitation: Exploiting vulnerabilities in UDP and HTTP protocols to gain unauthorised access.
3. Web Application Attack: Compromising web-based interfaces using techniques like brute-force attacks.
4. OT System Compromise: Injecting false data into the PLC to disrupt its operation.

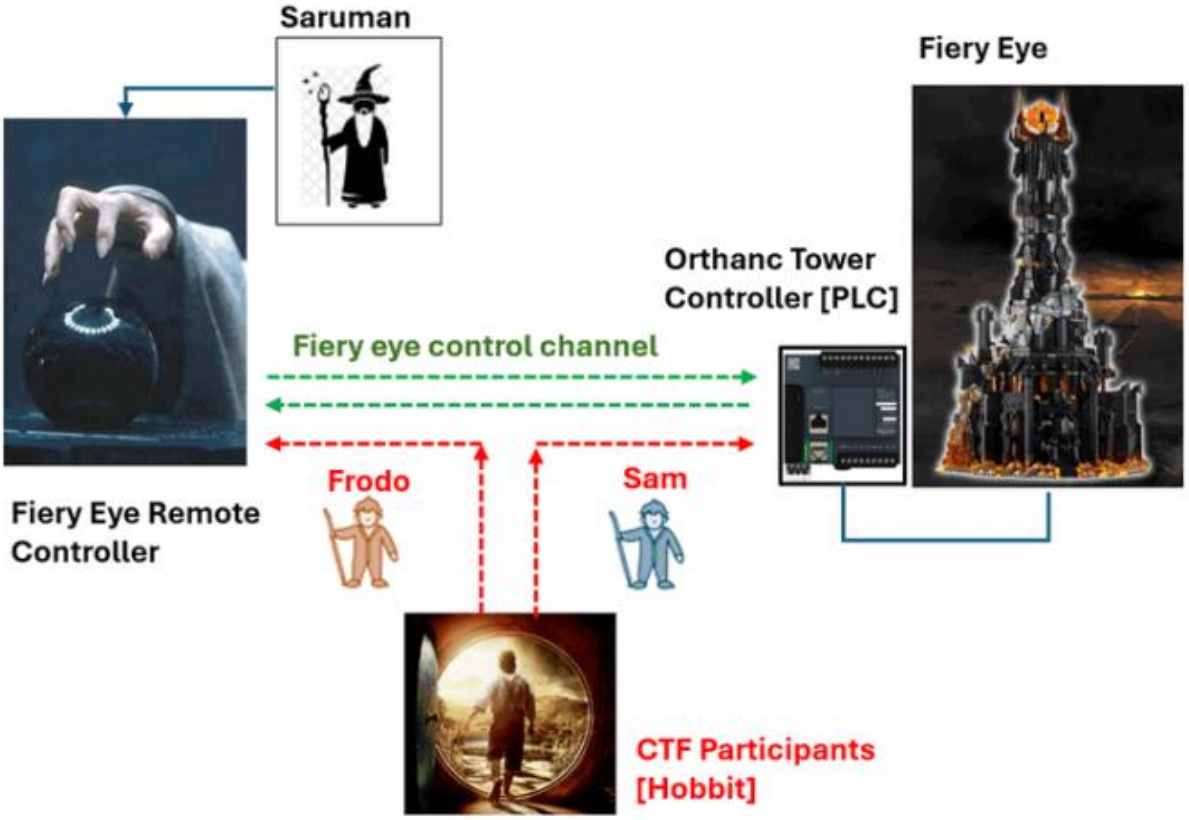

Figure 4: Palantir Control mini IT-OT environment.

By successfully completing these challenges, participants demonstrated their ability to bridge the gap between IT and OT security and execute effective cyber attacks.

3.6. Rohan Powerstone Network

The Rohan Powerstone Network category featured a sophisticated 7-substation PowerGrid Twin, developed by Illinois ARCS and iTrust under NSOE. This realistic simulation, seen in Figure 5, challenged participants to navigate complex OT scenarios and apply their knowledge of IEC 61850.

Seven challenging tasks were presented, requiring participants to exploit vulnerabilities, analyse network traffic, and conduct forensics of incidents. One particularly challenging task involved exploiting a known CVE in OpenPLC61850 to gain remote access to the PLC.

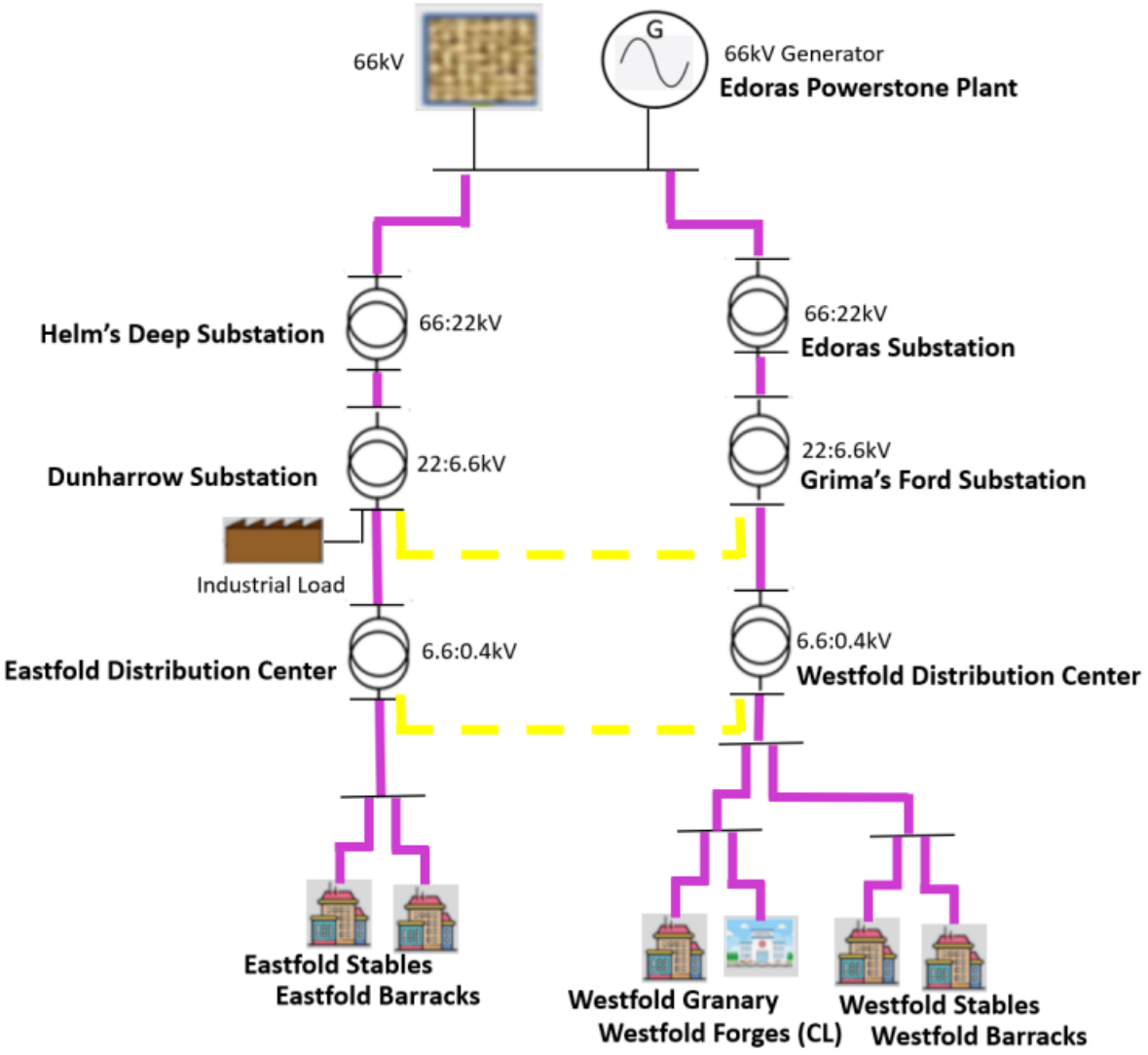

Figure 5: Overview of Rohan Powerstone Network

3.7. Script Submission

The Script Submission category presented two challenging tasks that required participants to develop and submit scripts to achieve specific attack objectives against the Secure Water Treatment (SWaT), Water Distribution (WaDi), and Electric Power Intelligent Control (EPIC) testbeds in iTrust.

The Red teams submitted ten scripts targeting the SWaT and WaDi testbeds, and another two for the EPIC testbed. Each script was executed for 30 minutes from 20 September to 24 September 2024. A 15-minute break was scheduled between each execution for system reset.

The performance of each script was evaluated based on two primary metrics and using the formula below:

1. **Number of Signatures Detected:** The ability of the script to trigger detections from the intrusion detection systems.

2. **Number of Packets Generated:** The efficiency and effectiveness of the script in achieving its objectives.

A weighted scoring system was used to rank the teams, with 70% of the score based on the number of signatures detected and 30% based on the number of packets generated. A lower overall score indicated a higher ranking, reflecting the effectiveness of the attack.

$$R_i = \left(\frac{S_i}{\sum_{i=1}^{n} S_n} \cdot 0.7 + \frac{P_i}{\sum_{i=1}^{n} P_n} \cdot 0.3 \right)$$

Where:

- S_i are the signatures that an individual red team generated
- p_i are the packets that an individual red team generated
- S_n are the total signatures that the red teams generated
- p_n are the total packets that the red teams generated
- R_i is the Ranking Score of the individual red team

To maintain security and integrity, the screens of the script host and the plant Human-Machine Interfaces (HMIs) were recorded during the execution process. This provided a detailed record of each script's behaviour and potential impact on the system. A neutral judge, Matthias Yeo, CEO of CyberXCenter, oversaw the execution process to ensure fairness and transparency.

3.8. Results

Table 2 presents the final scores and rankings of the top 10 teams, while Figure 6 illustrates the leaderboard for CISS 2024. Notably, Team Sesame demonstrated exceptional efficiency in solving challenges within a limited timeframe. From 18 September 2024, 1200 hrs onwards, they maintained a consistent lead in points through to the conclusion of the CTF.

In contrast, Team UncleCY showed a steady pace progressing consistently throughout the competition.

Table 2: Score and placing of the top 10 teams of the CTF

Placing	Team Name	Score
1	Sesame	2710
2	UncleCY	2660
3	KrautStrike	2610
4	FCC	2560
5	T-Lao-Sec	2540
6	Tomatoes	2420
7	UncleWY	2240
8	PWNed	2170
9	404: APT NOT FOUND	2160
10	LaKopi	2160

Figure 6: CISS 2024 leaderboard

4. IDS Evaluation

4.1. DRACE (Original)

DRACE is a framework developed by iTrust that introduces standardised metrics — Disruptivity, Responsiveness, Accuracy, Correctness, and Explainability (DRACE) — to objectively measure and compare the effectiveness of Intrusion Detection Systems (IDS) deployed in OT environments. By addressing gaps in current evaluation practices, DRACE helps ensure reliable and informed IDS deployment in critical infrastructure. The original DRACE was first published in July 2023 and used in CISS 2023.

4.1.1. Disruptivity

This metric evaluates the impact of an IDS on the normal operations of a system. An ideal IDS, as managed by its owners or operators, minimises system downtime caused by its own deployment or actions, ensuring smooth and uninterrupted plant operations.

4.1.2. Responsiveness

Responsiveness is the time taken by the IDS to detect and alert on an anomaly from the moment it begins. Faster detection improves the system's ability to respond to threats effectively.

4.1.3. Accuracy

This metric measures the IDS's ability to correctly identify true positives while minimising false positives and negatives. High accuracy ensures reliable threat detection without overwhelming operators with unnecessary alerts.

4.1.4. Correctness

This metric evaluates how precisely the IDS identifies the source, type, and cause of an anomaly, providing actionable insights for engineers and IT specialists to address issues promptly.

4.1.5. Explainability

This metric assesses how clearly and comprehensibly the IDS presents its alerts, ensuring that plant engineers and IT specialists can easily understand and act on the provided information.

4.2. Performance modification and Summary

This year's IDS evaluation used a modified DRACE format. Due to a change in CISS 2024's format, the IDS was only deployed during the final stage where it observes the attacks given by the finalists. The modifications to the DRACE format are as follows:

4.2.1. Disruptivity

The disruptive metrics were not measured as the IDS was deployed in a separate network that did not influence the network used in CISS 2024. In this case, the IDS was a passive observer wherein the plant data was duplicated and sent to the IDS. It was set up such that the IDS will not be able to actively add anything towards the network traffic or modify any parts of the plant network. Hence, disruptively was be measured in CISS 2024.

4.2.2. Responsiveness

CISS 2024 limited the execution of the attacks to a maximum of 20 minutes. The script submission was also limited to script that could run within the plant network. This meant that the lowest grade of the metric was given when an alert came in after the 2-minute time limit. Details are provided in Table 5.

Table 3: Responsivity Scoring Metric Table

Score	Responsiveness
4	0 - 10 seconds
3	11- 30 seconds
2	31 - 60 seconds
1	61 - 120 seconds
0	121 - 900 seconds

4.2.3. Accuracy

Due to the limitations of the attacks launched by the participants, correspondingly, there were limited observable impacts to the plant. Because of this, it would have been difficult to assess the IDS based on the DRACE metrics along. As such, if the IDS were able to detect other actions taken by organisers - such as resetting or redownloading of the PLC codes – and raised alerts because of those actions, these detections were also counted towards the IDS' accuracy.

Table 4: Accuracy Scoring Metric Table

Score	Accuracy
4	90 - 100 % Action Identified Correctly
3	70 - 89.9 % Action Identified Correctly
2	50 - 69.9 % Action Identified Correctly
1	30 - 49.9 % Action Identified Correctly
0	0 - 29.9 % Action Identified Correctly

4.2.4. Limitations

The team was able to launch the scripts submitted by the red teams, but a majority of scripts could not affect the plant in a meaningful manner. This created a limit on what the IDS can detect meaningfully during the event. As the IDS is able to detect the changes during the reset and standardisation process between the attacks launched, we will be considering the other actions that can be detected by the detector.

In total, there were two IDS teams including 1 from commercial vendors and 1 from Institutes of Higher Learning (IHLs). Only one of the commercial teams was finally evaluated.

Notes: The Flesch-Kincaid test measures reading ease and grade level of text based on sentence length and word complexity. The FK score is used to measure the IDS Explainability.

Tables 5 and 6 summarise the updated IDS performance metrics under the modified DRACE framework. The five metrics — Disruptivity, Responsiveness, Accuracy, Correctness, and Explainability — are detailed in their respective columns, including descriptions of how scores are awarded.

Table 5: IDS Performance Metrics Part A

Score	Disruptivity	Responsiveness	Accuracy
4	Solution does not disrupt plant operations	Alarm is generated within 0s - 10s of the anomaly	Model Detects - 90.0% – 100% of attacks
3	Solution disrupts plant operations 1% - 20% of the time	Alarm is generated within 11s - 30s of the anomaly	Model Detects - 70.0% – 89.9% of attacks
2	Solution disrupts plant operations 21% - 50% of the time	Alarm is generated within 31s - 60s of the anomaly	Model Detects - 50.0% – 69.9% of attacks
1	Solution disrupts plant operations 51% - 99% of the time	Alarm is generated within 61s - 120s of the anomaly	Model Detects - 30.0% – 49.9% of attacks
0	Solution disrupts plant operations 100% of the time	Alarm is generated after 121s or not generated	Model Detects - 0.0% - 29.9% of attacks

Table 6: IDS Performance Metrics Part B

Score	Correctness	Explainability
4	Identified - Specific Stage - Specific Component - Component failure type - Reason for failure	Identified - Severity identified - Protocol identified - SIP and DIP identified FK Score: >= 80
3	Identified - Specific Stage - Specific Component - Component failure type	Identified - Severity identified - Protocol identified - SIP and DIP identified FK Score: < 80
2	Identified - Specific Stage - Specific Component	Identified - Severity identified - Protocol identified FK Score: < 80
1	Identified - Specific Stage	Identified - Severity identified FK Score: < 80
0	None identified	None identified

4.3. Evaluation of IDS

The deployed IDS is a passive system that takes in two inputs: network traffic from a SPAN port and process data from the Historian. These inputs are used to detect anomalies in the system. Below are 3 anomaly samples and Figure 7 shows the performance of IDS 1.

1. Anomaly 1
    a. ID: epxbDZIBozbdHb_TSbEJ
    b. Timestamp: Sep 20, 2024 @ 09:20:52.000
    c. Protocol: OPCUA
    d. FAULT CODE: HMI_3_AIT_004.PV, -999
    e. Type: FAULT
    f. Alert: Unusual change in sensor value detected
    g. Type: alert-fault_detection_2024-09-20
2. Anomaly 2
    a. ID: d5xbDZIBozbdHb_TSLG5
    b. Timestamp: Sep 20, 2024 @ 09:20:52.000
    c. Protocol: OPCUA
    d. FAULT CODE: HMI_3_AIT_003.PV, -999
    e. Type: FAULT
    f. Alert: Unusual change in sensor value detected
    g. Type: alert-fault_detection_2024-09-20
3. Anomaly 3
    a. ID: HMkvDpIBozbdHb_TP-V-
    b. Timestamp: Sep 20, 2024 @ 09:21:33.238
    c. SIP: 192.168.1.201
    d. DIP: 192.168.1.60
    e. Protocol: CIP
    f. Type: CMD
    g. Alert: Unauthorized command detected
    h. Asset: SWAT-HMI
    i. Type: alert-abnormal_command_2024-09-20

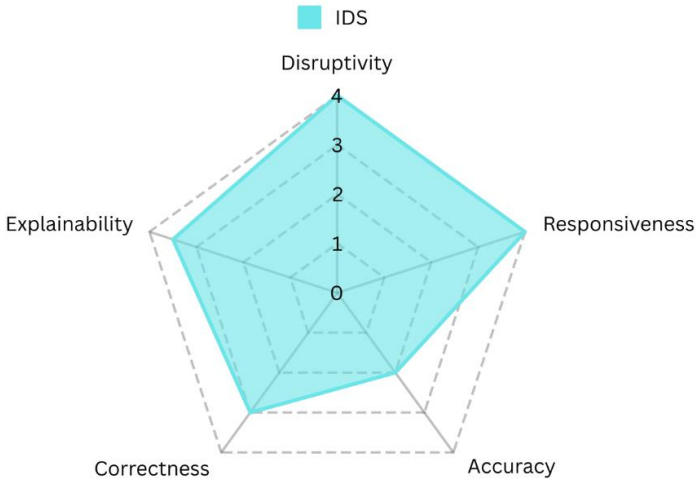

Figure 7: IDS performance.

Responsiveness (4)

The IDS was able to respond within 10 seconds after the action was executed. Based on the detailed report generated by the IDS, it was able to respond to minute actions from the plant. Changes from the commands as well as various states of the plants were observed and reported within 10 seconds of the changes done on the system. This nets the IDS 4 points in the responsiveness category.

Accuracy (2)

There seem to be a lot of false alarms in the detailed report as the IDS was detecting abnormal values at certain timings, when in reality, there was only a reset action within the time period. It led to a larger number of alerts to be processed and as such a larger percentage of false alarm when the plant was running normally.

It was able to accurately identify when certain network-based actions were launched and as such based on a weighted analysis, it could get 62.2% accuracy. For this, it was given 2 points.

Correctness (3)

The components were correctly identified in terms of their timing, devices, and results of the attack. The IDS did not elaborate the possible reason as to why the results came up and as such was not given the maximum point in this assessment. The IDS was able to detail the affected components and could highlight the type of anomaly in occurring in the system at that point in time.

Explainability (3.5)

The IDS report presented the alerts in both a visual and concise textual manner which allowed for clear identification of the alerts. It was able to present the SIP, DIP, severity clearly concisely and clearly.

However, with the short alerts such as unusual change in sensor value detected and unauthorised command detected, the FK score for the alerts was 12.32 and 20.98 respectively. As the FK score was below 80, a score of 3 was awarded. A bonus 0.5 points was awarded because the report and alerts contained visualisation that aided in the presentation of the alerts to clarify the possible alerts within the report. As such, the IDS was given a total of 3.5 points.

5. Feedback

After 54 hours of gameplay, most of the challenges were solved. Owing to the new CISS format and the introduction of the challenging task of script submission, only one Red Team managed to submit a workable script. Based on participants' feedback, teams who participated in the previous years' CISS preferred the 2023's format.

This was expected due to the lack of interaction with, and feedback from, the physical testbeds this year. Participants found the script submission challenge to be demanding, requiring a different skillset and a deeper understanding of the testbed's processes.

The limited time frame and lack of a testing environment as a feedback loop made it difficult for many teams to develop effective scripts. If the script submission format was to be retained, more detailed guidelines and support would be required to help participants develop effective scripts.

<End of document>