This supplementary material is associated with the article available at https://doi.org/10.1016/j.cose.2023.103120

Capability Assessment Tool v1.0

1. Sandbox Definition Capability Assessment Criteria

1.1. Functional Criteria

SD-1: Overall customization capabilities for sandboxes

- □ Level 1 – Instructors can only select from predefined sandbox definitions made available by platform developers.
- □ Level 2 – Instructors can define custom sandboxes, but only from building blocks (such as hosts, types of networks) provided by platform developers.
- □ Level 3 – Instructors can define custom sandboxes from building blocks provided either by platform developers or by themselves, although constraints on complexity and size may apply.

SD-2: Supported categories of hosts in sandboxes

- □ Level 1 – Full-fledged virtual or bare-metal hosts only.
- □ Level 1 – Container-based hosts only.
- □ Level 2 – Both full-fledged virtual or bare-metal hosts, and container-based hosts.

SD-3: Constraints on the CPU architectures and OSs of hosts in sandboxes

- □ Level 1 – Only hosts of the same CPU architecture and same family of OSs with the hypervisor, such as only Linux x86 hosts on a Linux x86 hypervisor, or only Windows x86 hosts on a Windows x86 hypervisor.
- □ Level 2 – Only hosts of the same CPU architecture, but more than one type of OS, such as Linux x86 and Windows x86 hosts on an x86 hypervisor.
- □ Level 3 – Hosts of more than one type of CPU architecture and more than one type of OS, such as Linux hosts with x86 CPU and Android hosts with ARM CPU.

SD-4: Mechanisms for customizing hosts in sandboxes[1]

- □ Level 1 – Instructors must provide final binary images of custom hosts in a specific format used by the underlying infrastructure (such as VMware, VirtualBox, KVM).
- □ Level 2 – Instructors do not need to build final binary images of custom hosts in a specific format, but must learn the *proprietary system* used by the platform for host provisioning.
- □ Level 2 – Instructors do not need to build final binary images of custom hosts in a specific format, but must be proficient in a particular *third-party configuration management system* (such as Ansible, Puppet, Chef) used by the platform for host provisioning.
- □ Level 3 – Instructors do not need to build final binary images of custom hosts or be proficient in a platform-specific configuration management system; the platform provides a library of fine-grained building blocks (such as vulnerable services, application stacks) that can be combined together as needed.

SD-5: Support for security-related sandbox configuration features

- □ Level 1 – Specific security-related *configurations*, such as for firewalls, can be performed via the sandbox definition.
- □ Level 1 – Security-related training content can be *created* via sandbox definition features, such as network attack emulation, traffic capture, malware emulation.
- □ Level 2 – Both security-related configurations and training content can be created via the sandbox definition.

[1] Customization features may include account management, package installation, content copy, program execution, etc.

SD-6: Support for physical devices (IoT, SCADA/ICS) in sandboxes

- □ Level 1 – No native support exists, but physical devices can be connected to other hosts in the sandbox at IP level.
- □ Level 2 – Native support exists for physical devices via a dedicated host type recognized by the platform.

SD-7: Constraints on the network topologies in sandboxes

- □ Level 1 – Only basic types of network topologies are supported by the platform, namely bus and star (see Figure 1).

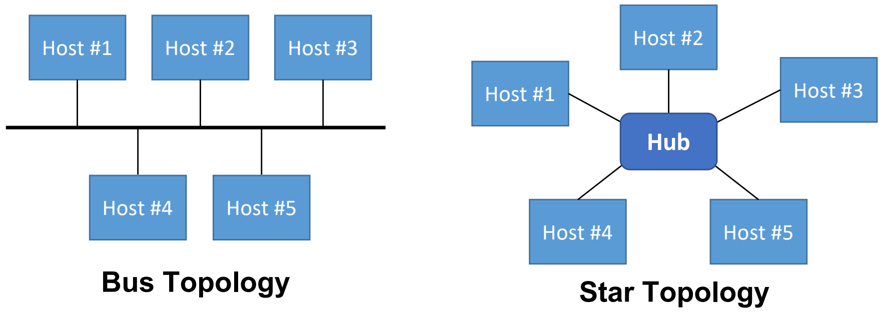

Figure 1: Bus and star network topologies (SD-7 Level 1).

- □ Level 2 – Basic types of network topologies are supported, as well as tree topologies, also known as star bus topologies, with a depth of more than two (see Figure 2).

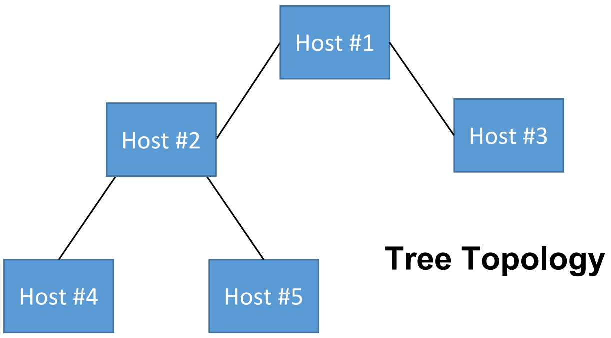

Figure 2: Tree (aka star bus) network topology (SD-7 Level 2).

- □ Level 3 – The platform supports both simple and complex network topologies, such as mesh, ring, or hybrid networks (see Figure 3).

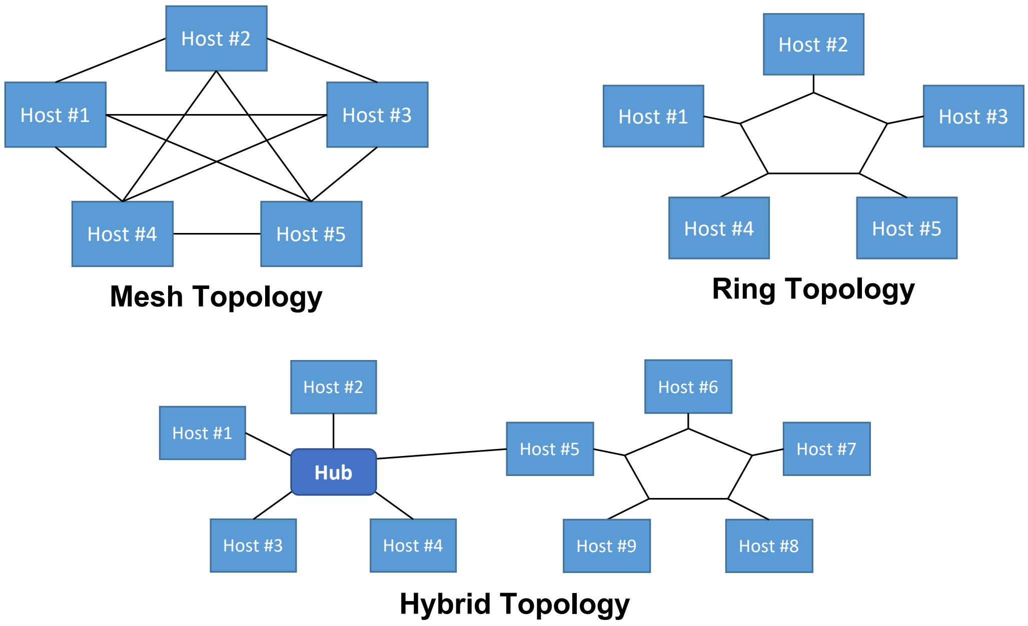

Figure 3: Mesh, ring, and hybrid network topologies (SD-7 Level 3).

1.2. Usability Criteria

SD-8: Mechanisms for creating/editing sandbox definitions

□ Level 1 – Creating/editing sandbox definitions is done by modifying directly the files.

□ Level 2 – Instructors can create, edit or preview the sandbox definition via a GUI.

SD-9: Security measures for isolating sandboxes

□ Level 1 – Hosts in individual sandbox instances cannot reach each other by default.

□ Level 1 – Hosts in sandbox instances are not directly reachable from the Internet.

□ Level 2 – Hosts in individual sandbox instances cannot reach each other by default, and are not directly reachable from the Internet.

SD-10: Representation format for the sandbox definition

□ Level 1 – Hard-coded descriptions of sandboxes are included in the platform source code.

□ Level 2 – Sandboxes definitions use a machine-readable, binary format.

□ Level 3 – Sandboxes definitions use a machine- and human-readable format.

SD-11: Validation mechanism for the sandbox definition

□ Level 1 – The platform only checks whether the definition contains the required attributes, but does not validate their values.

□ Level 2 – The platform validates both the required attributes and their values in order to prevent security threats, such as code injection.

SD-12: Availability of documentation for the sandbox definition format

□ Level 1 – Documentation exists, but it is not publicly available.

□ Level 2 – Documentation is publicly available.

SD-13: Availability of sample sandbox definitions and building blocks

□ Level 1 – Samples exist, but they are not publicly available.

□ Level 2 – Samples are publicly available.

2. Training Definition Capability Assessment Criteria

2.1. Functional Criteria

TD-1: Types of training supported in training definitions

- □ Level 1 – Task-based training: questions are presented ordered or unordered via a GUI, and trainees' answers are validated, but the GUI does not interact with the sandbox; similar to a jeopardy CTF game.
- □ Level 1 – Milestone-based training: the platform checks the status of the sandbox (network services or data stored on hosts) to automatically determine whether milestones defined by instructors are met or not; similar to attack-defense or King of the Hill CTFs, or cyber defense/offense exercises.
- □ Level 2 – Both task-based and milestone-based modes are supported.

TD-2: Structuring of individual tasks/milestones

- □ Level 1 – No relationships can be defined, and all tasks/milestones are available at once.
- □ Level 1 – Task/milestone relationships can be defined, so that completing one task/milestone is a prerequisite for presenting another task/milestone.
- □ Level 2 – Both modes above are supported.

TD-3: Types of questions supported in training definitions

- □ Level 1 – Only short-answer questions are supported.
- □ Level 1 – Only multiple-choice questions are supported.
- □ Level 2 – Both short-answer and multiple-choice questions are supported.

TD-4: Types of answers supported in training definitions

- □ Level 1 – Only one fixed answer is supported.
- □ Level 2 – Several alternative answers are supported (e.g., different capitalization for words).
- □ Level 3 – Answers expressed as regular expressions are supported.

TD-5: Support for specifying a trainee's role in a training

- □ Level 1 – Trainees' membership to a team can be specified.
- □ Level 2 – Trainees' roles, such as attacker or defender, can be specified.

TD-6: Support for rich formatting in training definitions

- □ Level 1 – Only text formatting is supported, and no built-in editor exists.
- □ Level 2 – Both text formatting and image embedding/linking are supported, but no built-in editor exists.
- □ Level 3 – Both text formatting and image embedding/linking are supported, and an editor is included.

2.2. Usability Criteria

TD-7: Mechanisms for creating/editing training definitions

- □ Level 1 – No GUI exists for training definition creation/editing, and the training definition files must be modified directly.
- □ Level 2 – Instructors can create, edit or preview the training definition via a GUI.

TD-8: Support for sharing/hiding the training content between instructors within the platform

- □ Level 1 – Only content sharing is possible via the platform.
- □ Level 2 – Both content sharing and hiding are available.

TD-9: Representation format for the training definition

- □ Level 1 – A hard-coded description is included in the platform source code.
- □ Level 2 – A machine-readable, binary format.
- □ Level 3 – A machine- and human-readable format.

TD-10: Validation mechanism for the training definition

- □ Level 1 – The platform only checks whether the definition contains the required attributes, but does not validate their values.
- □ Level 2 – The platform validates both the required attributes and their values in order to prevent security threats, such as code injection.

TD-11: Availability of documentation for the training definition format

- □ Level 1 – Documentation exists, but it is not publicly available.
- □ Level 2 – Documentation is publicly available.

TD-12: Availability of sample training definitions

- □ Level 1 – Samples exist, but they are not publicly available.
- □ Level 2 – Samples are publicly available.

3. Network Environment Management Capability Assessment Criteria

3.1. Functional Criteria

EM-1: Flexibility of network environment creation for parallel training activities

- □ Level 1 – Only multiple instances of the same network environment can be created for training activities conducted in parallel.
- □ Level 2 – Different network environments can be created even for training activities conducted in parallel.

EM-2: Access control features for network environment management

- □ Level 1 – Only instructors can manage the network environments, but each instructor can create, access, or delete any of them.
- □ Level 2 – Instructors can manage only the network environments they create.
- □ Level 3 – Instructors can manage the network environments they create, and also allow other instructors to manage those network environments.

EM-3: Support for automated action execution in a network environment *after* its creation

- □ Level 1 – Built-in features exist for automated action execution by defining custom actions and time-based or environment-based triggers for their execution in the environment.
- □ Level 2 – Built-in features exist for automated action execution by defining custom actions and triggers for their execution in the environment; visualizing these actions and their status during the training is also possible.

EM-4: Support for network attack execution functionality

- □ Level 1 – Built-in features exist for conducting automated network attacks against hosts in the environment.
- □ Level 2 – Built-in features exist for conducting automated network attacks against hosts in the environment; visualizing the scenarios of these attacks or their status during the training is also possible.

EM-5: Support for background traffic generation functionality

- □ Level 1 – Built-in features exist for generating background network traffic.
- □ Level 2 – Built-in features exist for generating background network traffic; visualizing such behaviour during the training is also possible.

3.2. Performance Criteria[2]

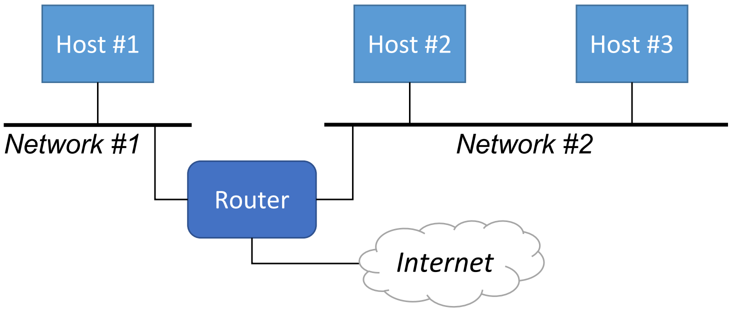

Figure 4: Reference network topology for the performance criteria EM-6 through EM-8.

EM-6: Reliability of network environment creation (N_{failed} is the percentage of failed sandboxes when creating 10 sandboxes with 3 hosts and 2 networks each for a total of ten times)

□ Level 1 – $20\% \leq N_{failed} < 30\%$

□ Level 2 – $10\% \leq N_{failed} < 20\%$

□ Level 3 – $N_{failed} < 10\%$

EM-7: Time efficiency of network environment creation ($T_{creation}$ is the time in minutes needed to create 10 sandboxes with 3 hosts and 2 networks each)

□ Level 1 – $20\ min \leq T_{creation} < 30\ min$

□ Level 2 – $10\ min \leq T_{creation} < 20\ min$

□ Level 3 – $T_{creation} < 10\ min$

EM-8: Resource consumption for a given network environment (R_{total} is the amount of resources including overhead needed to create a sandbox that uses R resources when creating 10 sandboxes with 3 hosts, 1 router, and 2 networks each (R = 40)[3])

□ Level 1 – $1.75R \leq R_{total} < 2R$

□ Level 2 – $1.5R \leq R_{total} < 1.75R$

□ Level 3 – $R_{total} < 1.5R$

3.3. Usability Criteria

EM-9: Type of user interface for network environment creation/deletion

□ Level 1 – Network environments are created/deleted via a CLI and/or REST API.

□ Level 1 – Network environments are created/deleted via a GUI.

□ Level 2 – Network environments can be created/deleted both via a CLI and/or REST API, and via a GUI.

EM-10: Degree of automation for network environment creation

□ Level 1 – Creation of one instance is semi-automated, but users must do several manual actions.

□ Level 2 – Creation of one instance is fully automated, but users must do as many manual actions as the number of instances.

□ Level 3 – Both single- and multiple-instance creation are fully automated.

[2] When assessing performance we advise that the experiments are performed by using the hardware infrastructure (servers, etc.) that is typically used when conducting training activities with a given platform. The sandbox network topology used in the experiments should be similar to that shown in Figure 4. Note that the thresholds used for performance criteria are applicable mainly to platforms that create virtual machine or physical server-based environments, and they may need to be adjusted for container-based ones.

[3] In practice we suggest using the number of VMs/containers as the resources to be evaluated, but other aspects could be assessed as well, such as memory consumption.

EM-11: Ability to monitor the network environment creation process

□ Level 1 – Only the creation result is displayed once the process finishes.

□ Level 2 – The creation process is split into several phases for which the status is displayed.

□ Level 3 – The creation process is split into several phases, and details about their progress are displayed.

EM-12: Degree of automation for network environment deletion

□ Level 1 – Deletion of one instance is semi-automated, but users must do several manual actions.

□ Level 2 – Deletion of one instance is fully automated, but users must do as many manual actions as the number of instances.

□ Level 3 – Both single- and multiple-instance deletion are fully automated.

EM-13: Ability to monitor the network environment deletion process

□ Level 1 – Only the deletion result is displayed once the process finishes.

□ Level 2 – The deletion process is split into several phases for which the status is displayed.

□ Level 3 – The deletion process is split into several phases, and details about their progress are displayed.

4. Training Activity Facilitation Capability Assessment Criteria

4.1. Pre-Training Setup Features

AF-1: Degree of automation for the deployment of the platform itself

□ Level 1 – There are ad-hoc scripts automating the deployment on a host with a specific operating system/hardware.

□ Level 2 – There are configuration files following the declarative Infrastructure as Code (IaC) approach[4] for platform deployment.

AF-2: Support for importing training content into the platform

□ Level 1 – Training content must be registered by a platform administrator who updates the training database.

□ Level 2 – Training content can be imported by the training instructor via a CLI.

□ Level 3 – Training content can be imported by the training instructor via a GUI.

AF-3: Trainee access management features of the training platform

□ Level 1 – Instructors must create user accounts for participants and distribute their credentials manually.

□ Level 2 – The platform supports automatic user registration or the import of user accounts.

□ Level 2 – The platform supports using a third-party identity provider, so that participants can log in using their credentials for another service.

□ Level 3 – The platform supports both automatic user registration or the import of user accounts, and the use of a third-party identity provider.

AF-4: Visibility control features for training sessions

□ Level 1 – Participants need accounts to access the training sessions, but there are no individual visibility controls.

□ Level 2 – Mechanisms exists for instructors to allow only selected participants to access a particular training session.

AF-5: Access time control features for training sessions

□ Level 1 – It is possible to control only the start date and time for accessing a training session.

□ Level 2 – It is possible to control both the start and the end dates and times for accessing a training session.

AF-6: Degree of instructor assistance needed by trainees for participating in training sessions

□ Level 1 – Instructor assistance is required, and without it, participants cannot start their training sessions.

□ Level 2 – Once participants are provided with credentials and step-by-step instructions on how to start the training, they can proceed autonomously, without instructor assistance.

[4] https://www.redhat.com/en/topics/automation/what-is-infrastructure-as-code-iac#declarative-vs-imperative-approach

4.2. Training Execution Features

AF-7: Use of an LMS[5] or ITS[6] during training

□ Level 1 – An LMS/ITS is included, which does not interact directly with the sandbox, but only indirectly via the participant (such as for presenting tasks via a portal, and assessing the submitted answers).

□ Level 2 – An LMS/ITS is included, which can interact with the sandbox to control its content or to process monitoring data collected during the training in order to control the ongoing training or provide feedback during/after the training.

AF-8: Features of the LMS or ITS used during training

□ Level 1 – Defining questions and their correct answers, checking the correctness of the submitted answers.

□ Level 2 – Conditions for serving tasks that are based on participants' proficiency or current performance in the ongoing training, but *without* processing data from the network environment.

□ Level 3 – Setting conditions for serving tasks that are based on participants' proficiency or current performance in the ongoing training, *including* by processing data from the network environment.

AF-9: Types of scaffolding mechanisms available during training

□ Level 1 – On-demand static hints or worked-out solutions for a particular task are available; they are the same for all participants, and do not consider a particular participant's progress.

□ Level 2 – Targeted hints are available based on the analysis of previous actions of an individual participant in an ongoing training session.

AF-10: Cheating prevention or detection features

□ Level 1 – Cheating *prevention* mechanisms[7] are available at training session creation time.

□ Level 1 – Cheating *detection* mechanisms[8] are available for ongoing and/or past training sessions.

□ Level 2 – Both cheating prevention and detection mechanisms are available.

AF-11: Manner of accessing hosts in the created network environment

□ Level 1 – Participants must install platform-specific client software in order to access the hosts.

□ Level 2 – Participants can use standard CLI protocols such as SSH to access the hosts (either directly or via a web browser).

□ Level 3 – Participants can use standard CLI and GUI protocols such as SSH, VNC or Remote Desktop to access the hosts (either directly or via a web browser).

AF-12: Type of user experience when accessing hosts in the created network environment

□ Level 1 – Participants can access the hosts via CLI/GUI, but the user experience is limited (such as copy-and-paste not working).

□ Level 2 – Participants can access the hosts via CLI/GUI, and the user experience is comparable to standard remote/local host access.

AF-13: Support for instructors accessing in-use hosts in sandboxes

□ Level 1 – Instructors can connect to a host that is in use by participants, but they need to use the same credentials with them.

□ Level 2 – Instructors can use special accounts or other mechanisms to access a host without interfering with participant use.

[5] Learning Management System.

[6] Intelligent Tutoring System.

[7] For instance, generating a personalized task assignment and corresponding network environment for each participants so that they cannot submit someone else's correct answer.

[8] For example, analyzing participants' submissions and/or actions performed during the training session with the aim to identify the participants who may have worked together; for details see `https://doi.org/10.1145/3478431.3499420`.

AF-14: Support for assigning one sandbox to multiple participants in the same team

- □ Level 1 – There is no direct support for allowing participants in the same team to access hosts in the same sandbox, but workarounds for achieving this with instructor assistance exist.
- □ Level 2 – The platform supports grouping participants into teams, and can provide team members shared access to all the hosts in a given sandbox.

AF-15: Visualization of the network environment topology

- □ Level 1 – Network environment topology information can be exported for external visualization.
- □ Level 2 – Network environment topology can be visualized within the platform.

AF-16: Support for trainees resuming a training session after exiting

- □ Level 1 – When participants exit a training session, the network environment remains unmodified, but there is not manner of returning in the LMS/ITS to the task/milestone where they exited.
- □ Level 2 – Participants can resume the training and return to their network environment and the task/milestone where they exited the training session.
- □ Level 3 – Participants can resume the training and return to their network environment and the task/milestone where they exited the training session; in addition, it is possible to revert to previous states of the training environment.

AF-17: Situational awareness and learning analytics features for instructors

- □ Level 1 – Instructors can only see who is logged in and is taking part in an ongoing training or has already finished it.
- □ Level 2 – Learning analytics for instructors only include data collected from the LMS/ITS component.
- □ Level 2 – Learning analytics for instructors only include data collected from the network environment.
- □ Level 3 – Instructors are provided with fine-grained data about participant interaction with the LMS/ITS component and the network environment.

AF-18: Learning analytics features for trainees

- □ Level 1 – Learning analytics for trainees only include data collected from the LMS/ITS component.
- □ Level 2 – Learning analytics for trainees include data collected both from the LMS/ITS component and the network environment.

4.3. Post-Training Assessment Features

AF-19: Support for exporting data collected during the training session or participants' score for further processing outside the platform

- □ Level 1 – Only basic statistics from the LMS/ITS component can be exported using CLI/GUI.
- □ Level 2 – Both LMS/ITS statistics and data collected from sandboxes can be exported using CLI/GUI.

AF-20: Support for analyzing the progress or results of participants across different training sessions

- □ Level 1 – Only progress and results for one session are available.
- □ Level 2 – A mechanism is available for creating an aggregated view of participants' progress and results across sessions.

Note

The capability level is indicated for each possible choice, although in some cases the choices can signify the same level of capability. Level 1 represents the minimal requirement for a certain type of functionality. Level 3 is the maximum capability level, although some criteria provide less than three choices. Level 0 is to be assigned if that particular functionality is missing completely (for clarity, we omit the Level 0 label in the assessment tool, and in order to indicate a missing functionality no box should be checked).