Interoperable

Mert Toslali[1], Jonathan Chamberlain[1], Felipe Dale Figeman[1], and Zhengyang Tang[1]

[1]BU EC528 - Cloud Computing

The Open Container Initiative (OCI) has been established to create a open standard for container use regardless of the runtime being used to manage the container. However, the OCI only specifies downloading image then unpacking that image into an OCI Runtime filesystem bundle. It does not standardize lifecycle management of the containers, thus each container implements lifecycle functionality in a different manner. It also does not ensure that consistent standards for the security of containers are present.
In this project, we study the differences in popular runtimes Docker, containerd, and crio. Our focus is on developing an interoperable application focused on ensuring that Center for Internet Security (CIS) Benchmarks are satisfied across runtimes to ensure consistent application of security principles irrespective of container runtime differences. By implementing an application, which can be exposed as a service to validate standard security checks across runtimes, we intend to provide a Proof of Concept that such a common lifecycle management is possible.

Introduction

Virtualization of resources has emerged in recent decades as a means to run multiple OSes on the same hardware. This particularly serves a useful function as this allows multiple applications to coexist on the same server, enabling efficencies in computing such as server consolidation.
Traditional VMs virtualize hardware resources, which results in the VMs taking up more resources. As such, OS-level virtualization, or Containers, have been developed. By sharing OS resources, containers are lightweight and can be spun up quickly while taking up fewer resources. Sharing OS resources such as libraries significantly reduces the need to reproduce the operating system code, and means that a server can run multiple workloads with a single operating system installation. Containers are thus exceptionally light.
Containers are implemented using Linux namespaces and cgroups. **Namespaces** let you virtualize system resources, like the file system or networking, for each container. **Cgroups** provide a way to limit the amount of resources like CPU and memory that each container can use.
Docker, introduced in 2013, is a popular runtime to manage containers as it addresses end-to-end management. However, Docker was initially a monolith with features not inherently dependent on each other being bundled together. As a result, alternative runtimes such as CRI-O and contianerd exist which implement container management at varying levels. [1]
The Open Container Initiative (OCI, https://www.opencontainers.org) has been established to create a open standard for container use regardless of the runtime being used to manage the container. However, the OCI only specifies downloading image then unpacking that image into an OCI Runtime filesystem bundle.
It does not standardize lifecycle management of the containers, thus each container implements lifecycle functionality in a different manner. It also does not ensure that consistent standards for the security of containers are present.
In this project, we study the differences in popular runtimes Docker, containerd, and crio. We propose an interoperable application focused on ensuring that Center for Internet Security (CIS) Benchmarks are satisfied across runtimes to ensure consistent application of security principles irrespective of container runtime differences. By implementing an application, which can be exposed as a service to validate standard security checks across runtimes, we intend to provide a Proof of Concept that such a common lifecycle management is possible. Minimum viable product for this project was to implement 5 CIS benchmarks from chapter 5 over at least two different container runtimes. However, we go way beyond the MVP and implement 23 Benchmark checks for docker, 20 for crio, and 16 for containerd. Detailed implementation for the benchmarks can be found in later sections and appendix.

A. Vision and Goals. Currently if someone wishes to launch an image in a container or perform any other lifecycle management functions on it, they must be sure that the scripts are configured correctly for the target container. For instance, launching images in Docker differs from doing so in crio or containerd. This locks individuals and businesses into whichever container runtime they started with unless they invest the time required to edit the configuration and their scripts which holds the commands for target container.
Our short term goal for this project is to enable the set of Container Runtime tests run in the CIS Docker 1.13.0 Benchmark across any container runtime. These tests are specified in Chapter 5 of the document; example checks include restricting Linux Kernel capabilities within containers, limiting memory usage, and avoiding directly exposing the host devices to the containers. Publishing a minimum viable framework for this purpose will enable users to run their security checks using a single application across the most popular containers.

Users/Personas of the Project. The intended user is a software developer who is developing, testing or a security engineer who is managing applications and ensuring reliability across containers running on different runtimes.
Example Use Case: A software developer would like to launch an image in CRI-O instead of Docker, because he re-

alizes that CRI-O is more adaptable with Kubernetes, and using this capability will provide this application a lot more scalability. Presently, he needs to deal with changing all the continous-integration scripts in order to be able to test and deploy his application on this new container runtime. With our interoperable framework in place, the developer is at least able to run security checks on the new container runtime with our application by specifying the new target container. In this way, the user's workflow is simplified and can apply a standard security checks across runtimes with minimal effort.

Scope. The runtimes in scope for capability for this project will be Docker, and CRI-O. containerd is considered a runtime in scope as a stretch goal.
This project aims to ensure that the framework implements commands that satisfy the CIS Docker 1.13.0 Benchmark related to Container Runtimes across our in-scope runtimes. In doing so, users will be enabled to run their security checks with a single application rather than requiring separate suites for each runtime. The MVP is considered to be implementing at least 5 benchmarks in consultation with our mentor. Implementation of the full suite is a stretch goal.
These benchmarks are specified in pp 126-180 of the Benchmark documentation

Design

Interoperable application exists between developer and container runtimes. Our project aims to implement security checks across runtimes. We implement interoperable application in Python language. Right now, this work is a standalone application which is assumed to have access to container configuration files, sysfs, and procfs in the system. The future work for this project is to enable this application as a service which is deployed as daemon that runs on a Kubernetes cluster and ensures CIS benchmarks on container runtimes within that cluster.

Implications and Discussion. Interoperable application leverages the attributes that are exposed in *sysfs* and *procfs* per process in Linux. The sysfs filesystem is a pseudo-filesystem which provides an interface to kernel data structures. The files under sysfs provide information about devices, kernel modules, filesystems, and other kernel components. The proc filesystem is a pseudo-filesystem which provides an interface to kernel data structures. Interoperable application use procfs and sysfs per container instance to figure out neccesary attributes which are dictated by CIS Benchmarks. The detailed implementation is discussed in later sections. Furthermore, along with the procfs and sysfs, Interoperable application also evaluates the underlying structure of the configuration files per container runtime in order to implement CIS checks.

CIS Benchmarks. CIS Benchmarks are best practices for the secure configuration of a target system. Available for more than 140 technologies, CIS Benchmarks are developed through a unique consensus-based process comprised of cyber-security professionals and subject matter experts around the world. CIS Benchmarks are the only consensus-based, best-practice security configuration guides both developed and accepted by government, business, industry, and academia. In this project, we focus on following chapters from CIS Benchmarks.

CIS Chapter 4. cis chapter 4 [Jonathan]

CIS Chapter 5. The ways in which a container is started governs a lot of security implications. It is possible to provide potentially dangerous run-time parameters that might compromise the host and other containers on the host. Verifying container run-time is thus very important. In this project we implement various recommendations to assess the container run-time security that is provided through CIS Chapter 5.
We mention following benchmarks for the sake of space. Other benchmarks that we implemented can be found from the appendix and their scopes/intentions can be evaluated from (https://www.cisecurity.org/benchmark/docker/):

5.1 Do not disable AppArmor Profile (Scored). AppArmor protects the Linux OS and applications from various threats by enforcing security policy which is also known as AppArmor profile. You can create your own AppArmor profile for containers or use the Docker's default AppArmor profile. This would enforce security policies on the containers as defined in the profile.

5.2 Verify SELinux security options, if applicable (Scored). SELinux provides a Mandatory Access Control (MAC) system that greatly augments the default Discretionary Access Control (DAC) model. You can thus add an extra layer of safety by enabling SELinux on your Linux host, if applicable.

5.3 Restrict Linux Kernel Capabilities within containers (Scored). By default, Containers start with a restricted set of Linux Kernel Capabilities. It means that any process may be granted the required capabilities instead of root access. Using Linux Kernel Capabilities, the processes do not have to run as root for almost all the specific areas where root privileges are usually needed.

- NET_ADMIN
- SYS_ADMIN
- SYS_MODULE

5.6 Do not run ssh within containers (Scored). Running SSH within the container increases the complexity of security management by making it difficult to manage access policies and security compliance for SSH server. Difficult to manage keys and passwords across various containers. Difficult to manage security upgrades for SSH server It is possible to have shell access to a container without using SSH, the needlessly increasing the complexity of security management should be avoided.

5.10 Limit memory usage for container (Scored). By default, container can use all of the memory on the host. You can use memory limit mechanism to prevent a denial of service arising from one container consuming all of the host's resources such that other containers on the same host cannot perform their intended functions. Having no limit on memory can lead to issues where one container can easily make the whole system unstable and as a result unusable.

5.11 Set container CPU priority appropriately (Scored). By default, CPU time is divided between containers equally. If it is desired, to control the CPU time amongst the container instances, you can use CPU sharing feature. CPU sharing allows to prioritize one container over the other and forbids the lower priority container to claim CPU resources more often. This ensures that the high priority containers are served better.

5.24 Confirm cgroup usage (Scored). System administrators typically define cgroups under which containers are supposed to run. At run-time, it is possible to attach to a different cgroup other than the one that was expected to be used. This usage should be monitored and confirmed. By attaching to a different cgroup than the one that is expected, excess permissions and resources might be granted to the container and thus, can prove to be unsafe.

5.28 Use PIDs cgroup limit (Scored). Attackers could launch a fork bomb with a single command inside the container. This fork bomb can crash the entire system and requires a restart of the host to make the system functional again. PIDs cgroup –pids-limit will prevent this kind of attacks by restricting the number of forks that can happen inside a container at a given time.

Interoperable Application

Interoperable application is a Python executable, which accepts two parameters from user. That are, target container runtime and container-id. For instance, to issue CIS Chapter 5 benchmarks over Docker container with id = 0606, user is expected to run command as: `./interoperable_app -docker 0606`.
In this section we demonstrate that we are able to perform more than 10 benchmarks over all container run-times (Docker, Containerd and Cri-o). For the sake of space, we only mention about some of the benchmark implementation that were more challenging. For the other benchmarks, one can refer to tables that are located in appendix.

Docker. Our primary platform is Docker container runtime. Here, we evaluate the how we implement the CIS checks on Docker. First, we find target container's pid from */run/containerd/io.containerd.runtime.v1.linux/moby/<container-id>/init.pid*. In order to issue **5.1 Do not disable AppArmor Profile** and **5.2 Verify SELinux security options** benchmarks over Docker, we use procfs and pid. Apparmor and SELinux are security attributes for a given process. So these values are stored under */proc/<pid>/attr/apparmor/current* and */proc/<pid>/attr/selinux/current* respectively. By exposing these values, we are able to issue **5.1** and **5.2**.
Per container instance, configuration file is created under */run/containerd/io.containerd.runtime.v1.linux/moby/<container-id>/config.json* file. From this file, we get **cgroup** path of a given container. Exposing this value also corresponds to **5.24** benchmark.
Further, by using the cgroup path, we are able to access *sysfs* of a given container. *sysfs* is a pseudo file system provided by the Linux kernel that exports information about various kernel subsystems. From this file, we are able to perform **5.10, 5.11, 5.28**. **5.10 Memory Limit** of a container is found from: */sys/fs/cgroup/memory/<container-id>/memory.limit_in_bytes*. **5.11 CPU Share** of a container is found from: */sys/fs/cgroup/cpu/<container-id>/cpu.shares*. **5.28 PID Limits** of a container is found from: */sys/fs/cgroup/pids/<container-id>/pids.max*.

Containerd. Other platform that we evaluate interoperable application is Containerd run-time. The process for implementing given benchmarks are pretty similar to Docker that is mentioned above. Since different runtimes comes with different configuration files and corresponding structures, we state the implementation details as following. First, we find target container's pid from */run/containerd/io.containerd.runtime.v2.task/default/<container-id>/init.pid*. In order to issue **5.1 Do not disable AppArmor Profile** and **5.2 Verify SELinux security options** benchmarks on Containerd, interoperable application uses procfs and corresponding pid for a container. Apparmor and SELinux are security attributes for a given process. These values are stored under */proc/<pid>/attr/apparmor/current* and */proc/<pid>/attr/selinux/current* respectively. By exposing these values, we are able to issue **5.1** and **5.2**.
Per container instance, configuration file is on containerd created under */run/containerd/io.containerd.runtime.v2.task/default/<container-id>/config.json* file. From this file, we get **cgroup** path of a given container. Exposing this value also corresponds to **5.24** benchmark.
Further, by using the cgroup path, we are able to access *sysfs* and leverage attributes of a container. From this file, we are able to perform **5.10, 5.11, 5.28**. We specifically use followings: **5.10 Memory Limit** of a container is found from: */sys/fs/cgroup/memory/<container-id>/memory.limit_in_bytes*. **5.11 CPU Share** of a container is found from: */sys/fs/cgroup/cpu/<container-id>/cpu.shares*. **5.28 PID Limits** of a container is found from: */sys/fs/cgroup/pids/<container-id>/pids.max*.

Cri-o. Other platform that we evaluate interoperable application is Cri-o run-time. The process for implementing given benchmarks are pretty similar to Docker and Containerd. Since different runtimes comes with different configuration files and corresponding structures, we state the implementation details as following. First, we find target container's

pid from its configuration file (i.e., specifically state.json) that is stored under */var/lib/containers/storage/overlay-containers/ <container-id>/userdata/state.json* In order to issue **5.1 Do not disable AppArmor Profile** and **5.2 Verify SELinux security options** benchmarks on Containerd, interoperable application uses procfs and corresponding pid for a container. Apparmor and SELinux are security attributes for a given process. These values are stored under */proc/<pid>/attr/apparmor/current* and */proc/<pid>/attr/selinux/current* respectively. By exposing these values, we are able to issue **5.1** and **5.2**.
Per container instance, configuration file is on crio created under *//var/lib/containers/storage/overlay-containers/ <container-id>/userdata/config.json* file. From this file, we get **cgroup** path of a given container. Exposing this value also corresponds to **5.24** benchmark.
Further, by using the cgroup path, we are able to access *sysfs* and leverage attributes of a container. From this file, we are able to perform **5.10, 5.11, 5.28**. We specifically use followings: **5.10 Memory Limit** of a container is found from: */sys/fs/cgroup/memory/<container-id>/memory.limit_in_bytes*. **5.11 CPU Share** of a container is found from: */sys/fs/cgroup/cpu/<container-id>/cpu.shares*. **5.28 PID Limits** of a container is found from: */sys/fs/cgroup/pids/<container-id>/pids.max*.

Other benchmarks

For the other benchmarks, interoperable application mainly leverages the fields that exist within the corresponding configuration files per container. For example, to implement **5.4, 5.7, 5.8, 5.9, 5.12, 5.13, 5.14, 5.15, 5.16, 5.17, 5.18, 5.20 and 5.25** on Docker, we use Hostconfig.json file for a given container. This file is created and stored per container in : */var/lib/docker/containers/ <container-id>/hostconfig.json*. To implement **5.3, 5.5, 5.21, 5.24** on docker, our application uses config.json under */run/containerd/io.containerd.runtime.v1.linux/moby/ <container-id>/config.json*.
For crio, interoperable application implements **5.3, 5.4, 5.5, 5.7, 5.8, 5.9, 5.12, 5.13, 5.15, 5.16, 5.17, 5.20, 5.21, 5.24** using the config.json file per container. For crio this file is found from */var/lib/containers/storage/overlay-containers/<container-id>/userdata/config.json*

Containerd....

Bibliography

Supplementary Note 1: Something about something

appendix