


Article

Leadership hijacking in docker swarm and its consequences*

Adi Farshteindiker¹ and Rami Puzis^{1,*} ¹ Cyber@BGU, Software and Information Systems Engineering at Ben-Gurion University of the Negev

* Correspondence: puzis@bgu.ac.il; Tel.: +972-54-4764010

Abstract: With the advent of microservice-based software architectures, an increasing number of modern cloud environments and enterprises use operating system level virtualization, often referred to as containers. Docker Swarm is one of the most popular container orchestration infrastructures, providing high availability and fault tolerance. Occasionally discovered container escape vulnerabilities allow adversaries to execute code on the host operating system and operate within the cloud infrastructure. We show that docker swarm is currently not secured against misbehaving manager nodes and allows a high impact, high probability privilege escalation attack that we refer to as leadership hijacking. Cloud lateral movement and defense evasion payloads allow an adversary to leverage the docker swarm functionality to control each and every host in the underlying cluster. We demonstrate an end-to-end attack, in which an adversary with access to an application running on the cluster achieves full control of the cluster. To reduce the probability of a successful high impact attack, container orchestration infrastructures must reduce the trust level of participating nodes and in particular, incorporate adversary immune leader election algorithms.

Keywords: docker swarm; leader election; privilege escalation; defense evasion; cloud

1. Introduction

As Docker gained popularity among cloud service providers, attackers began to develop various techniques to attack Docker-based applications. One class of techniques used by attackers exploits classical application vulnerabilities, such as SQL injection, buffer overflow, command injection, etc. Using such vulnerabilities, an attacker can control the victim's container and data inside it. Container escape exploits are another technique class; in this case, after successful container exploitation, the attacker exploits a vulnerability allowing the attacker to escape from the container to the underlying host. Access to the underlying host grants an attacker access to data and other containers that run on the compromised host.

There are many products and protocols that try to mitigate the abovementioned techniques. First, Docker offers built in protections,¹ such as, protecting the Docker daemon socket, using data encryption between Docker daemon and public registries etc. These protections harden Docker hosts with a "security in depth" approach. In addition, software such as SE-Linux and App-Armor can help harden container isolation and minimize the attack surface between containers and the host. Furthermore, Docker offers an image scanning service,² which can detect vulnerabilities in Docker images.

Although a lot of attention has been dedicated to securing Docker hosts from the technique classes mentioned above, little to no solution exists for securing against privilege escalation among different hosts in a Docker cluster. In this work, we show how an attacker with access to a manager host inside a Docker cluster can escalate his/her privileges in the cluster. The research scope is presented in Figure 1.

There are many algorithms inside a Docker cluster which can be exploited for that purpose. For example, Raft, a consensus algorithm used to manage a replicated log [1], is used in Docker Swarm to synchronize the cluster's state between all managers of the cluster. See Section 2.2 for details. The logs are replicated using a strong leader, which is elected

*This research was partially supported by the Cyber Security Research Center at Ben-Gurion University of the Negev.

¹ <https://docs.docker.com/engine/security/security/>

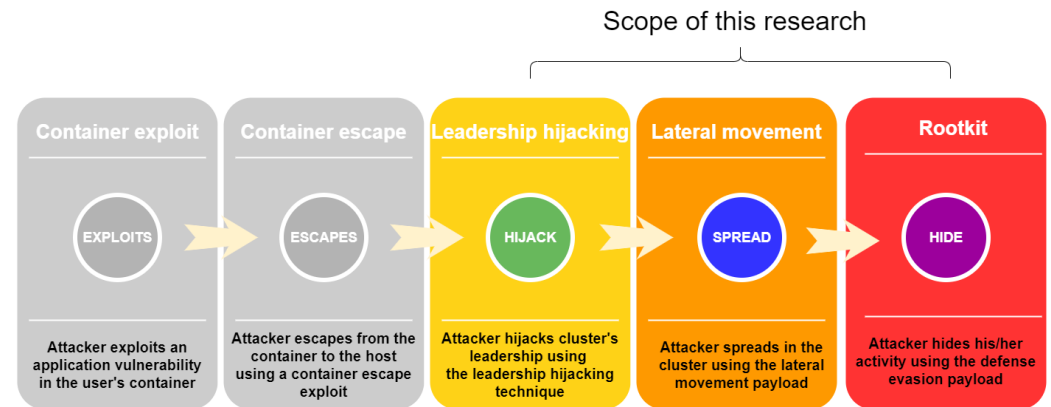


Figure 1. High-level description of the end-to-end scenario

in the leader election phase in the algorithm. In case of a leader failure (a crash, network issues, etc.), the rest of the managers choose a new leader using the Raft algorithm. Despite its many advantages, Raft is a non-Byzantine algorithm allowing a malicious insider to become a leader.

1.1. Scope and purpose

In this paper, we highlight a new privilege escalation technique called **leadership hijacking** (see Section 3). An attacker with access to a manager node in Docker Swarm can use this technique to escalate his/her cluster privileges and become the cluster leader. By doing so, the attacker can control all messages and decisions within the cluster.

In addition, we demonstrate two possible malicious payloads expected to be executed by a typical attacker: a **lateral movement** payload and a **defense evasion** payload. The former utilizes cluster leader privileges and allows the attacker to execute code on every host in the cluster. The latter is used by an attacker in order to hide his/her malicious activity from infrastructure management tools.

In order to illustrate the feasibility and impact of the leadership hijacking technique and the malicious payloads, we developed an end-to-end attack scenario that shows how an external attacker can chain exploits seen in the wild with our technique and payloads, in order to obtain full control of a cluster. For this purpose, we created a test-bed that, on one hand, mimics a cloud environment with a Docker Swarm and multiple client's services; and, on the other hand, includes a typical attackers' tool set. A Detailed description of the attack scenario is discussed in Section 5. We executed a full exploit chain which, when combined with our leadership hijacking technique, ultimately gave the attacker cluster leader privileges. Later, we showed how our malicious payloads can be used to completely compromise our simulated cloud environment. A high-level overview of the end-to-end scenario can be seen in Figure 1.

As shown in the figure, the end-to-end attack consists of five major steps:

1. Exploitation of an application vulnerability inside a container, in which an attacker gains a foothold within the user's container
2. Container escape exploitation, in which an attacker obtains access to the container's underlying host
3. Leadership hijacking, in which an attacker executes the privilege escalation technique presented in Section 3 and obtains cluster leader privileges
4. Lateral movement, in which an attacker executes the lateral movement payload described in Subsection 4.1 and gains privileged access to all hosts in the cluster

² <https://docs.docker.com/ee/dtr/user/manage-images/scan-images-for-vulnerabilities/>

5. Defense evasion, in which an attacker uses the defense evasion payload described in Subsection 4.2 in order to hide his/her lateral movement payload from management tools

A detailed description of this scenario is provided in Section 5. Steps 1 and 2 are implemented in order to demonstrate the feasibility of our work, but they are not elaborated upon, since they are out of the scope of our research.

The rest of this paper is structured as follows: Section 2 reviews technical background. In Section 3, we introduce the novel privilege escalation technique, called leadership hijacking. Next, in Section 4 we investigate malicious payloads that can be executed after the privilege escalation. In Section 5, we demonstrate an end-to-end attack scenario that illustrates the potential security risk and the impact of the investigated attack. Finally, in Section 6 we discuss possible mitigation and propose countermeasures. Final remarks can be found in Section 7.

2. Background and related work

2.1. Docker Swarm

An increasing number of organizations are moving their digital systems to the cloud. The benefits of cloud servers are easy deployment, high availability, continuous maintenance, system security, and more. From online websites to internal servers and databases, cloud servers store a lot of sensitive information, making them an attractive target for attackers. As the cost of hardware has decreased, software has become the main performance bottleneck. In order to fully utilize the available hardware, cloud service providers use virtualization technology to run different applications on the same hardware.

Until recently, the most advanced solution was virtual machine (VM) technology. VM technology allows one physical server to run many different virtual servers, all of them running different operating systems. From a security point of view, a VM is a good solution, since breaking out of a VM is a relatively complex task[2]. On the other hand, VMs suffer from significant performance overhead[3]. The main reason for the reduced performance is the overhead added by the hypervisor to each hardware operation emulated to the VM.

Today, many cloud service providers use operating system level virtualization, which employs isolated user space instances called containers. In contrast to a VM which includes its own operating system, containers run under the host's operating system and communicate with it directly. During runtime, container communicates through a regular system calls interface with the host OS, without any intermediate software. The architectural difference is illustrated in Figure 2.

At the time of this writing, Docker is one of the leading OS virtualization solutions.³ Docker is implemented in the Go programming language and enables the creation, deployment, and management of containers on a host computer. A Docker container is a lightweight software unit that bundles its own tools and libraries. Typically, one container includes one instance of an application or service, e.g., a Web server, database, scientific software package, etc.

Docker is a rich ecosystem. One of the main components of this ecosystem is the Docker daemon. The Docker daemon is software that runs on the host and is responsible for the creation of images and containers. The Docker daemon can run containers and create their runtime environment; it can also create a container's networking interfaces, mount points, and can trigger actions and execute commands inside a running container. The Docker daemon implements Docker's main logic and many of its features.

When deploying an application in a production environment, it's important to ensure that when a container fails, a new container will start and replace the faulty container. In addition, it is highly recommended to run several instances of a container for high availability and load balancing. To address these issues, Docker introduced a feature called Swarm.

³ <https://resources.flexera.com/web/media/documents/rightscale-2019-state-of-the-cloud-report-from-flexera.pdf>

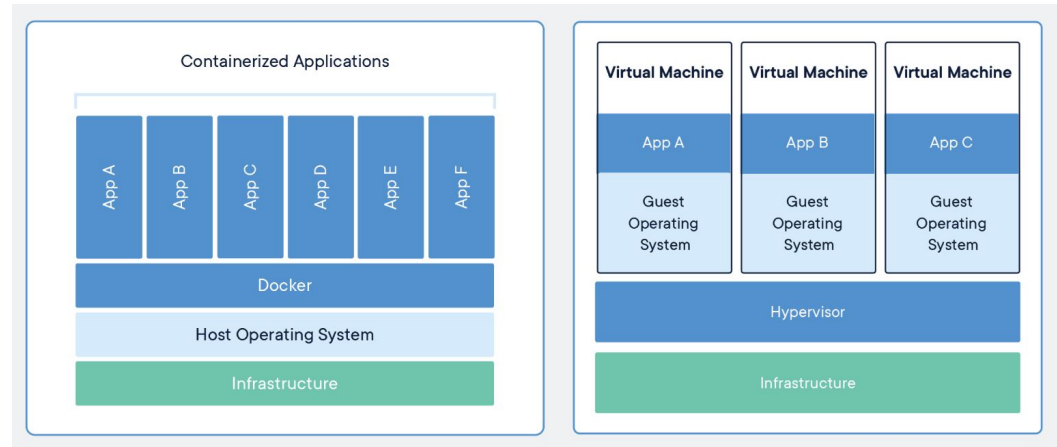


Figure 2. Container vs VM architecture [4]

Docker Swarm abstracts many Docker hosts to one virtual Docker host. Each host that participates in the swarm cluster is called a **node**. Each node can have two roles: **manager** or **worker**. A manager's job is to keep a replicated state of the cluster. One manager node is also a **leader**. The cluster's leader is responsible for scheduling new containers and services for the cluster. A worker's job is to get container tasks from the leader and to actually run the container. The weakest point in the design of Docker Swarm exploited in this research is the Raft leader election algorithm.

2.2. Leader election

Raft [1] is a consensus algorithm used to manage a replicated log. Raft was designed with the aim of producing an efficient and understandable algorithm which, unlike Paxos [18] [20] [19], would be easy to learn and use in practical systems. Raft was chosen in Docker Swarm due to its important features:

- Strong leader – Raft uses a stronger form of leadership than other consensus algorithms. For example, log entries only flow from the leader to other servers.
- Leader election – Raft uses randomized timers to elect leaders. This adds only a small number of mechanisms to the already existing heartbeat mechanism and facilitates simpler conflict resolution.
- Membership changes – Raft's mechanism for changing the set of servers in the cluster uses a new joint consensus approach, which allows the cluster to continue operating normally during configuration changes.

Yet, Raft assumes that all nodes are honest and is not tolerant to malicious (Byzantine) nodes participating in the leader election process.

Byzantine fault tolerant (BFT) leader election algorithms have existed for a long time. These algorithms provide the ability to overcome failures in networks where some nodes are Byzantine. For example, Castro et al.[21,22] described a state machine replication algorithm able to tolerate Byzantine faults. The algorithm guarantees safety, i.e., each replicated log is agreed on by all non-faulty nodes.

Bessani et al. [31] introduced open-source Java library implementing robust BFT state machine replication. Key features of their implementation are reliability, modularity and flexible application programming interface (API). Moreover, their implementation achieves good performance and can tolerate real world faults.

Castro et al. [21] implemented a BFT library, that can be used to build highly available systems that tolerate Byzantine faults. Moreover, Castro et al. used the library to implement a Byzantine-fault-tolerant NFS file system. They showed that the replicated library can be even more efficient than the non-replicated version of NFS.

2.3. Related work on Docker security

In this section, we overview previous work on cloud security related to Docker.

Singh et al.[5] demonstrated primary techniques used by attackers to attack cloud services. There are many potential attack vectors that attackers can use, including: DoS and DDoS attacks [6] [7], malware injection, and side-channel attacks [8–11]. In their study, Jensen et al.[12] demonstrated an attack on the software of the cloud itself and outlined the threat of flooding attacks on cloud systems. The authors suggested improving the cloud's security by first improving the security of frameworks used in the cloud.

In [13], Liu et al. provide an overview of the latest technologies in cloud computing and discuss how Docker is integrated into it. According to Liu et al., the major difference between classic VM and containers is that a VM contains not only the application and its dependencies but also the entire guest operating system. The authors list rapid application deployment, portability across machines, lightweight footprint, and minimal overhead as the main advantages of Docker over traditional VM-based virtualization software.

Xavier et al.[14] performed numerous experiments in order to evaluate the performance of container-based cloud environments compared to VM-based cloud environments, as well as the trade-off between performance and isolation. They found that the cloud environment would benefit from container-based solutions, due to the fact that container-based solutions achieve near-native performance.

Other research [15] suggests a new attack surface in the Docker environment: namely indirect adversaries. Unlike a direct adversary who exploits vulnerabilities in the cluster directly, an indirect adversary exploits third party appliances (e.g., Docker Hub) in order to attack Docker's environment.

In his master's work, Kabbe [16] compared the security model of containers to hypervisor-based systems and virtual machines. He compared the outcome of known attacks (DirtyCow,⁴ Heartbleed,⁵ and Shellshock⁶) in a containerized environment, with the outcome of the same attacks performed in hypervisor/virtual machine environments. He found that containers offer at least the same amount of security as hypervisor/virtual machine environments.

In his master thesis[17], Seather reviews the underlying security of the Docker Swarm infrastructure. Namely, Seather tests many adversarial scenarios, including: flooding the orchestrator with invalid/corrupted requests, sniffing the network from within the cluster, impersonating as a cluster member, performing man-in-the-middle attacks between containers within Docker's internal network, and more. The conclusions of his thesis are that Docker's infrastructure is secure, Docker Swarm's design is good (from a security point of view), the technology stack used by Docker is immune to known attacks, and the development community responds quickly to security incidents.

3. Leadership hijacking

In this section we introduce an adversarial technique named leadership hijacking. A precondition to employing this technique is code execution access to a manager node. In Section 5, we show how this precondition can be achieved in a production environment. From now on, we will refer to the manager host compromised by the attacker as the attacker's manager. The main idea of our technique is to repeatedly trigger a leader election phase, until the attacker's manager becomes the cluster leader. The technique's pseudocode is shown in algorithm 1.

As shown in algorithm 1, the first step of the technique is to identify the current cluster leader. If the current leader is the attacker's manager, the technique's code will exit. Otherwise, the technique starts a loop.

⁴ <https://nvd.nist.gov/vuln/detail/CVE-2016-5195>

⁵ <https://nvd.nist.gov/vuln/detail/CVE-2014-0160>

⁶ <https://nvd.nist.gov/vuln/detail/CVE-2014-6271>

Algorithm 1 Attack’s pseudo code

```
1: procedure GET-LEADERSHIP
2:   if Current manager is cluster leader then
3:     Exit
4:   while Current manager is not leader do
5:     leader_id  $\leftarrow$  find out current leader ID
6:     demote node with id leader_id
7:     wait constant time
8:     promote node with id leader_id to be manager
9:                                      $\triangleright$  At the end of the loop, current manager is the leader
```

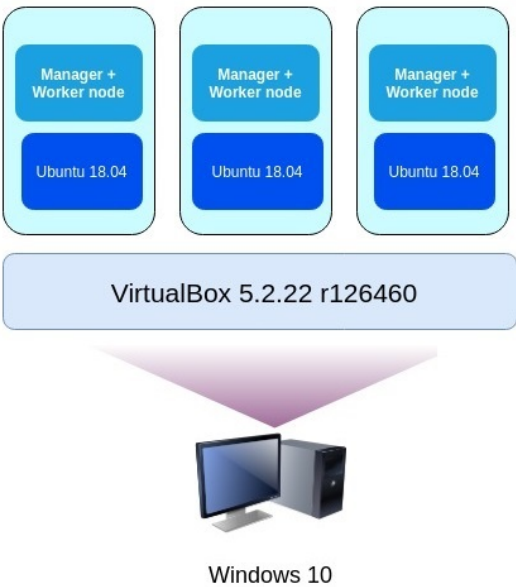


Figure 3. Overview of lab architecture

In each loop iteration, the technique demotes the current cluster leader. This will cause the cluster to initiate a leader election algorithm and elect a new leader. The first manager that reaches timeout proposes itself as the cluster leader. Afterwards, each manager votes in favor of one manager, and the manager that receives the majority of the votes becomes the new cluster leader.

In the final step of the iteration, the current cluster leader is identified again. If the attacker’s manager is the leader, the technique exits. Otherwise, it will continue the loop until the attacker’s manager becomes the cluster leader.

In order to prove that the technique works in practice, we implemented the pseudocode shown in algorithm 1. We set up a lab to test the implementation, and its architecture is illustrated in Figure 3. Running our technique’s implementation in the lab was successful: the attacker was able to escalate privileges in order to become the new cluster leader.

3.1. Analysis

3.1.1. Probability of success

In each iteration, the technique code demotes the leader. According to the Docker Swarm documentation, a manager that doesn’t receive the heartbeat from the leader during the predefined time window assumes that the leader is unavailable and proposes itself to

be the new cluster leader. Since the leader has been demoted, none of the managers receive the heartbeat from the leader, and hence a new leader election phase will start when the first manager reaches its timeout.

According to the Docker Swarm source code, the manager's heartbeat timer is randomly selected. We can refer to that timeout as a uniformly distributed number.

Since the technique is activated at a random time, each of the manager hosts has the same probability of reaching its timeout first and proposing itself as the new cluster leader. We would like to calculate the probability that the attacker's manager will reach its timeout first and hence propose itself as the leader.

In our scenario, each iteration may have two outcomes: success or failure. Success means the attacker's host reached its timeout first and hence proposed itself as the new cluster leader, and failure otherwise. Since timeouts are random in each host, the probability that a host will reach timeout first is uniformly distributed. Since we demoted the leader to be a worker, there are $n - 1$ remaining managers that may propose themselves as leader. Hence

$$P(h \text{ reach timeout first}) = \frac{1}{n-1}, \forall h \in \text{manager's hosts}$$

In particular,

$$\begin{aligned} P(\text{success}) &= P(\text{attacker's manager reaches timeout first}) \\ &= \frac{1}{n-1} \end{aligned}$$

In addition, in each round the probability for success remains the same. According to this, the number of failed iterations before the first success distributes in a geometric distribution, given by the formula:

$$P(X = k) = (1 - p)^{k-1} * p$$

and the cumulative distribution function is given by the formula:

$$P(X \leq k) = F(k) = 1 - (1 - p)^k \quad (1)$$

where X is a random variable that represents the number of iterations and $p = P(\text{success})$.

For a given k , we'll calculate the probability that our attack succeeds before k iterations. By substituting Formula 1,

$$P(X \leq k) = 1 - (1 - \frac{1}{n})^k = 1 - (\frac{n-1}{n})^k$$

For a large k ,

$$(\frac{n-1}{n})^k \rightarrow 0 \Rightarrow 1 - (\frac{n-1}{n})^k \rightarrow 1 \Rightarrow P(X \leq k) \rightarrow 1$$

Thus for a large k , our technique will succeed before k iterations in probability $P \rightarrow 1$.

3.1.2. Advantages

The first advantage of the technique is its simple implementation. In order to prove its feasibility, we decided to implement the technique in the most simple way possible. After reviewing the Docker Swarm API, we realized that our technique could be implemented with repeated calls to the demote and promote API. This simple implementation makes our technique stable and reliable.

The second advantage of our technique is its stealthiness. Typical attacker would like to stay undetected as long as possible while in an engagement. Our technique can be implemented in many ways, but some are rather loud, which will increase the chance to get caught by the system administrators. For example, attacker can demote all other

managers of the cluster and become the only manager, hence the cluster leader. The obvious issue of this implementation is that the system administrators will quickly notice that the cluster state has changed. On the other hand, our implementation's changes to the cluster state are minimal, which make it harder to detect the technique.

3.1.3. Limitations

The main limitation of our technique is that it is probabilistic. Although we showed that our technique completes successfully with probability $P \rightarrow 1$, the number of iterations in each execution may differ. An unknown number of iterations is particularly problematic in a real-world scenario.

4. Malicious payloads

In order to illustrate the impact of the leadership hijacking technique, we developed malicious payloads that use cluster leader privileges, and use them to perform some malicious operations.

Typically, an attacker who has access to one host inside a cluster would like to spread and obtain a wider foothold in the cluster. Ideally, the attacker would like to have access to all hosts in the cluster, with high privileges in each host. Moreover, once the attacker controls a cluster he/she would like to remain undetectable by the users/system administrators for as long as possible.

To achieve the above goals, the attacker has to find a way to spread inside the cluster and hide his/her malicious activity from the users and monitoring tools. In this work, we introduce and develop two types of malicious payloads: a lateral movement payload and a defense evasion payload. These payloads utilize leader privileges and allow an attacker to execute high privileged code on every node in the cluster and hide from monitoring tools.

4.1. Lateral movement

Typically, an attacker would like to establish a wide foothold in a cluster, preferably with high privileges. In this work, we create a payload that enables lateral movement in the cloud. Using this payload, we demonstrate how an attacker with leader privileges in a Docker Swarm cluster can execute high privileged code on each host in the cluster.

Due to the fact that after successful execution of leadership hijacking, the attacker gained leader privileges, the attacker can control all messages that come out of the leader node. By hooking the leader's function responsible for sending messages between the leader and other nodes, the attacker can change these messages and alter their content.

In order to execute code on other nodes in the cluster, an attacker who is in control of a leader host can send the victim node a task to run. The attacker instructs the worker to run a container task with an image controlled by the attacker. As we show in Section 5, the victim node will execute the container. The container's image will be a malicious image.

However, the malicious container runs in an isolated environment in the host. As discussed in Section 2.3, containers run in a separate namespace from the host. Thus, for example, a process inside a container can't sniff the host's network.

There are many ways to overcome this limitation. In addition to controlling what image the container will run on each host, the attacker also controls the creation flags of the container. Thus, for example, the attacker can mount the main file system of the host to the container. Then, from inside the container, the attacker can alter the host's executable files with a malicious code. In order to obtain high privileged code execution, the attacker has to alter a file that is executed by a high privileged user on the host. When the user executes the file, the attacker's malicious code will get executed as well, resulting in high privileged code execution on the host.

4.2. Defense evasion

With the above lateral movement payload, the attacker can spread and move laterally by deploying service with malicious image to every host in the cluster. In this subsection,

we show how attacker can stay undetected in the cluster and hide malicious activity from the cloud's management tools. We introduce the cloud defense evasion payload, which offers rootkit-like functionality in the cloud.

In this subsection we assume that the attacker is the cluster leader and has a malicious service in the cluster which he/she wishes to hide from system administrators, e.g., a malicious cryptocurrency mining service.

Default Docker Swarm command line offers a rich variety of commands for cluster administration. In particular, Swarm offers the `docker service`⁷ command for viewing and updating services that run on the cluster. In order to view services that run on the cluster, the system administrator can issue the `docker service ls`⁸ command and view its output. The output includes the service's name, image, number of replicas, exposed ports, etc.

In order to obtain this information, the Docker daemon of the host that issued the command queries the leader of the cluster and retrieves the information from the leader.

However, the attacker is in control of the leader host. Hence, the attacker can hook the function that returns this information on the leader's Docker daemon and spoof the answers. In this way, the attacker can change malicious service's name, image, ports, or even the service itself (i.e., the attacker can trick the user into thinking that there is no such service at all, by removing any information related to the malicious service).

In a similar manner, the system administrator can view what containers are running for each service. Using `docker service ps`⁹ command, the system administrator can obtain information about a container's image, name, state, etc. In a similar way to the `docker service ls` command, the issuing host queries the leader host and retrieves that information. The attacker has access to the leader host, and thus he/she can alter that information as well. By doing so, the attacker can trick the system administrator and show him/her that a container is running a different image than the real image, for example.

In that way, the attacker can hide malicious activity from Docker's default tools, which query the cluster leader to obtain information about objects (running services, containers, etc.) in the cluster.

5. End-to-end attack demonstration

In order to prove that our leadership hijacking technique and malicious payloads are feasible, we created a combined scenario that demonstrates the impact of our technique and the payloads. We show the importance of our technique and payloads, as well as that the initial assumption regarding the attack is reasonable. We show how an external attacker can leverage practical exploitation which has been seen in the wild together with our leadership hijacking technique and malicious payloads, in order to ultimately control the entire cluster.

5.1. Lab setup

We set up a lab that simulates a cloud environment. Cloud nodes are simulated using virtual machines which run the Ubuntu guest OS. We set up a Docker Swarm cluster in which all hosts are both manager and worker hosts.

In addition, an external laptop is used as the attacking machine. The lab's architecture is shown in Figure 4.

5.2. Scenario overview

In our end-to-end attack scenario, the attacker starts on an external laptop with network access to a Docker container that runs inside a Docker cluster. Ultimately, the

⁷ <https://docs.docker.com/engine/reference/commandline/service/>

⁸ https://docs.docker.com/engine/reference/commandline/service_ls/

⁹ https://docs.docker.com/engine/reference/commandline/service_ps/

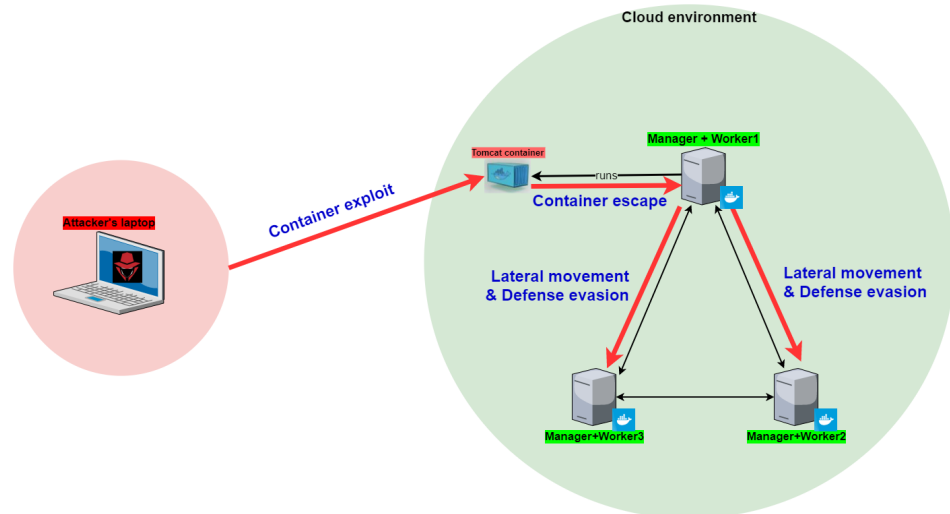


Figure 4. Diagram of attack steps

attacker will obtain high privileged code execution on each host in the cluster. The scenario contains five major steps:

1. Container exploitation
2. Container escape exploitation
3. Leadership hijacking
4. Lateral movement
5. Defense evasion

In each step, the attacker expands his/her foothold in the cluster. An illustration of the entire scenario and its steps can be seen in Figure 4.

The next subsections explain these steps in greater detail.

5.3. Container exploitation

First, the attacker needs to have some initial foothold in the cluster. He/she has network access to an application that runs on a container in the cluster. In order to obtain an initial foothold, the attacker exploits a vulnerability in the application.

In this case, the application running inside the container is the Apache Tomcat Web server, version 8.5.19. The attacker finds a one-day exploit for that Web server in the Metasploit framework; after successful exploit completion, the attacker has shell access to the application's container.

5.4. Container escape

After the attacker has successfully exploited the application, the attacker has a shell in the restricted Docker environment. In order to execute our privilege escalation technique, the attacker needs to escape from the restricted environment and retrieve a shell on the underlying host of the container.

The attacker then exploits a vulnerability in the host's RunC component.¹⁰ RunC is a container runtime that was originally developed as part of Docker, which is responsible for running and managing new container environments.

A vulnerability resides in RunC version < 1.0-rc6 (which is used by Docker < 18.09.2), allowing the attacker to overwrite the host's RunC binary and thus achieve code execution with root privileges on the host.

¹⁰ <https://www.cvedetails.com/cve/CVE-2019-5736/>

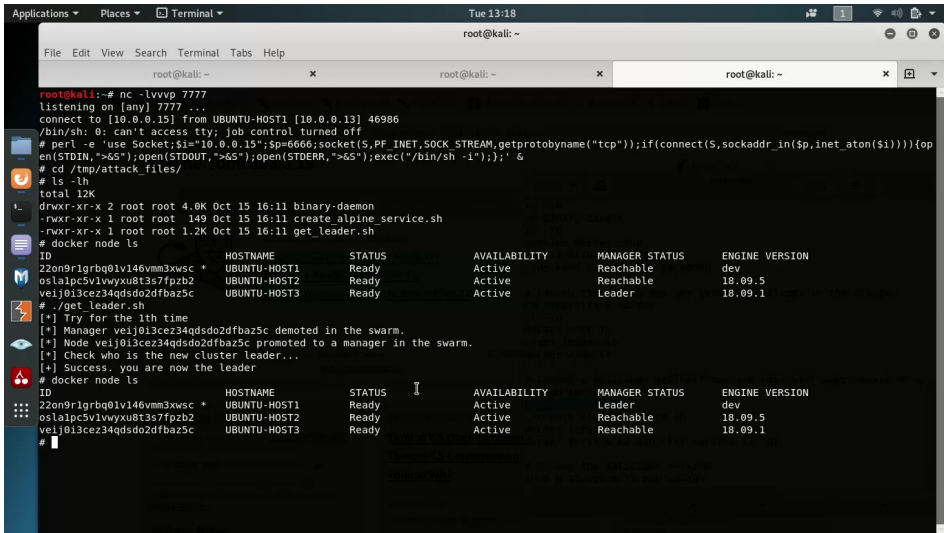


Figure 5. Successful attack attempt

5.5. Cloud privilege escalation

Once the attacker has achieved code execution on Docker’s manager host, he/she can execute the leadership hijacking technique and escalate his/her privileges in order to become the cluster leader (see Section 3 for a description of the leadership hijacking technique).

After the leadership hijacking technique’s successful execution, the attacker obtains leader privileges in the cluster and thus will be able to control all messages that flow between the leader and other hosts in the cluster.

The result of the technique’s successful execution can be seen in Figure 5. In this figure, we can see that before the attack, UBUNTU-HOST3 was the cluster leader, and after the technique was successfully executed, UBUNTU-HOST1 (which is the attacker’s manager) obtained the leadership role in the cluster.

5.6. Lateral movement and defense evasion

Armed with leader privileges, the attacker can now control all messages that flow between the leader and other hosts in the cluster. As described in Subsections 4.1 and 4.2, the attacker can execute a malicious container on each host in the cluster and hide these actions from various management tools.

To effectively demonstrate the attack and its potential impact, in our scenario, the attacker will run a WebShell service which will run a WebShell container on every host in the cluster.

The malicious WebShell container provides a root privileged command execution environment on the underlying host. The host’s file system is mounted in the container’s /tmp directory. This allows the attacker to view, modify, and delete the host’s files. Effectively, the attacker runs a root WebShell on all hosts in the cluster.

The output of the WebShell can be seen in Figure 6. In addition, the figure shows that the WebShell is executed with high privileges (root).

Moreover, the attacker uses the defense evasion functionality described in Subsection 4.2, hooking the leader’s Docker daemon function which is responsible for listing services and containers of services. By doing so, any service listing request that is made to the cluster leader will be monitored by the attacker. In cases in which the attacker’s malicious service is running, the attacker will spoof the answer of the listing and hide his/her malicious service image with a benign Alpine image. As seen in Figure 7, docker service ls command reveals a single running service, with image "alpine:latest." In addition, it seems that there are no listening ports, but in actuality, container on each host is listening on port 80.

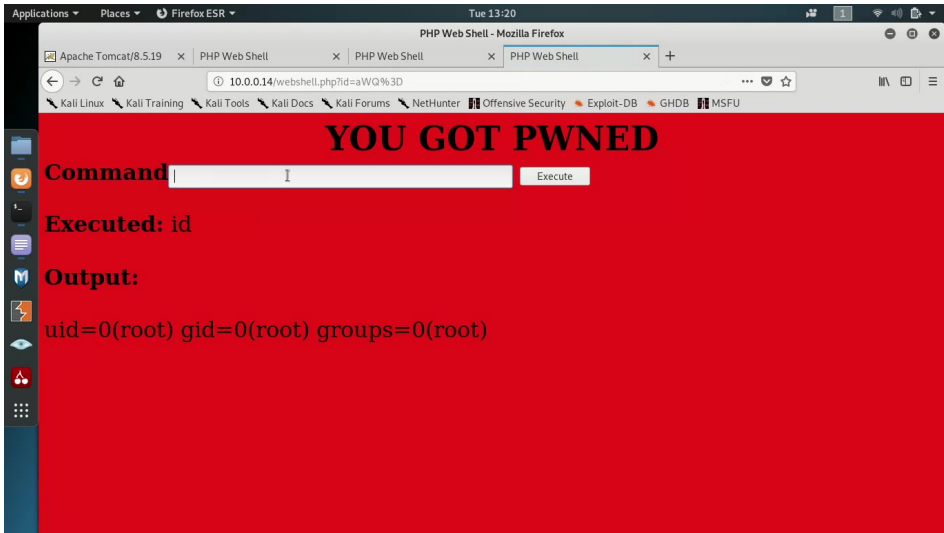


Figure 6. Output of malicious WebShell

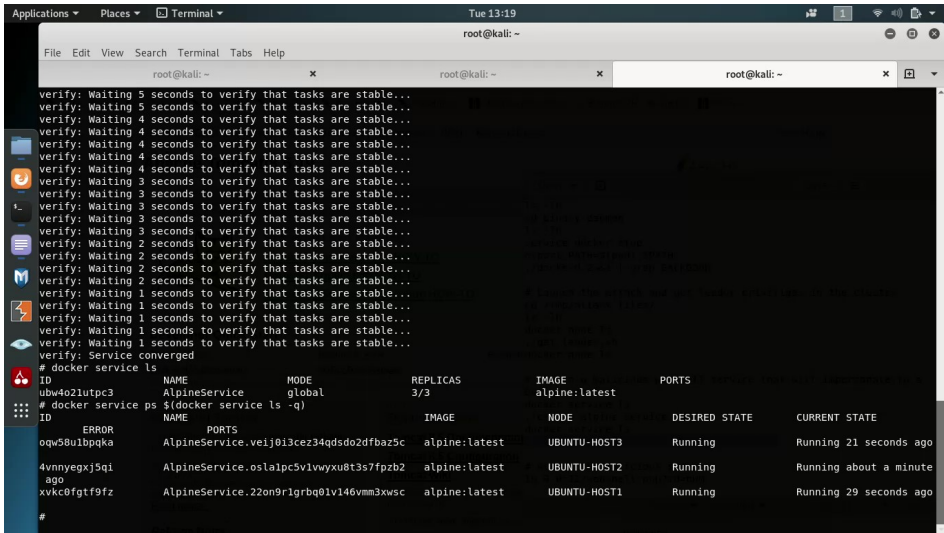


Figure 7. Docker’s default tools used for viewing information about malicious services

Furthermore, the attacker also hooks the function responsible for listing container of each service; thus, the output of `docker service ps $(docker service ls -q)` doesn't reveal the real image that each container is actually running.

According to Docker’s default tools, it looks like the service running is a benign alpine service but accessing each host in port 80 reveals the true "face" of the service.

6. Discussion

The main advantage of our technique is that unlike many techniques seen in the wild, our technique doesn't exploit any software bugs. A software bug is usually a mistake in a program's code, which can lead to an undefined behavior of the program. In most cases, software bugs are easily fixed. However, our technique doesn't exploit any programming errors, but rather exploits a design flaw. Unlike programming bugs, logical bugs are much harder to fix, since in many scenarios a large amount of code should be changed, which can be costly and time-consuming for software developers.

As shown in Section 3, our technique exploits the fact that the Raft algorithm is used to replicate logs in the Docker Swarm environment, but it is a non-adversarial algorithm. Raft is a key component of Docker Swarm's management infrastructure, and it is integrated

into the core logic of Docker Swarm. Replacing the Raft algorithm in Docker Swarm is a mandatory step to mitigate our proposed technique, since exploits used to escape from container to host (as shown in Section 5) is very common and relatively easy to find. Since it's a design bug, replacing Raft requires a significant amount of work.

First, Docker's developers should choose and implement byzantine fault tolerant algorithm [31] [21] in Go, or find such an implementation as a Go package. The implementation should be high quality, since it will be deployed to every manager in the cluster. Next, the developers should modify Docker Swarm's source code. In Docker swarm, Raft's implementation is encapsulated with a wrapper object. The developers of Docker Swarm should change the entire wrapper object to encapsulate the new package instead of Raft. Then, series of tests should be ran to ensure that the new package meets Docker's efficiency requirements: both local and network. The new package should not consume significant amount of host's resources, as well as be efficient in terms of network activity between hosts in Docker swarm. Moreover, the tests should ensure that the new package works as expected on every operating system supported by Docker Swarm. Since managers are the most valuable servers in the cluster, any bug in a manager can be fatal. The tests should ensure as much as possible that the new package is bug free, and that it has no unwanted side effects. In any case, replace Raft implementation holds a major risk and may cause a service degradation.

There are some best practices which may block our attack, The most common is to separate manager nodes from worker nodes. In such case, even if attacker compromised a worker node, he will not be able to escalate his privileges in the way we suggested in this article, since attacker's node is not part of the managers group. However, although considered a best practice, this is not the default behaviour of Docker Swarm. We believe that Docker's developers chose to make manager node a worker too by default in order to not waste expensive computing power. If a node is just a manager, it won't receive client container to execute, hence cluster's computing capacity decrease. Regardless, in this article we chose to research and exploit systems in their default state, and not delve into best practices.

We offer two strategies in order to effectively mitigate our technique. In the short term, the technique can be mitigated by detecting and blocking container escape exploits. As discussed in Section 3, the leadership hijacking technique should be executed from a manager host. We showed in Subsection 5 that an attacker can gain such access using a container escape exploit. In case that the container escape exploit fails, attacker can't launch the technique and therefore can't escalate his privileges in the cluster. In order to reduce the amount of container escape exploits, Docker can start a bug bounty program. We believe that it will help Docker patch container escape vulnerabilities before they can be exploited by real attackers in the wild.

In the long term, we offer to replace Raft algorithm with a byzantine fault tolerant algorithm. As discussed earlier, Raft is a non adversarial algorithm, hence attacker who is in control of a Raft's participant can forge and spoof messages. In that way, the attacker can trick other participants to vote for him in the leader election phase, and become cluster's leader. In case that a BFT algorithm is used, other participants would not vote for the attacker since the algorithm can tolerate byzantine participants. In that way, attacker would not be able to escalate his privileges to cluster leader. Furthermore, in order to support future changes, developers of Docker should divide Docker's infrastructure from the leader election algorithm. The architecture of Docker Swarm should be "plug and play", such that the leader election algorithm is chosen as a configuration option instead of a source code modification.

7. Conclusion

In this work, we suggest a new attack vector on the Docker Swarm orchestrator. Our technique demonstrates a new concept in offensive security, in which a cluster is treated as a single unit of processing, and an attacker is able to escalate his/her privileges in that unit

and thereafter perform malicious activity on every component of that unit separately (i.e., every host in the cluster).

Future research can explore other ways in which attackers can obtain leader privileges in cloud environments, including environments that are orchestrated by other solutions besides Docker Swarm, e.g., Kubernetes.

References

1. Diego Ongaro and John K Ousterhout. In search of an understandable consensus algorithm. In *USENIX Annual Technical Conference*, pages 305–319, 2014.
2. Jenni Susan Reuben. A survey on virtual machine security. *Helsinki University of Technology*, 2(36), 2007.
3. Kim-Thomas Moeller. *Virtual machine benchmarking*. PhD thesis, Citeseer, 2007.
4. Figure, container vs vm arch. <https://www.docker.com/resources/what-container>.
5. Ajey Singh and Maneesh Shrivastava. Overview of attacks on cloud computing. *International Journal of Engineering and Innovative Technology (IJEIT)*, 1(4), 2012.
6. Meiko Jensen, Nils Gruschka, and Norbert Luttenberger. The impact of flooding attacks on network-based services. In *2008 Third International Conference on Availability, Reliability and Security*, pages 509–513. IEEE, 2008.
7. Marwan Darwish, Abdelkader Ouda, and Luiz Fernando Capretz. Cloud-based ddos attacks and defenses. In *International Conference on Information Society (i-Society 2013)*, pages 67–71. IEEE, 2013.
8. Daniel Gruss, Clémentine Maurice, Klaus Wagner, and Stefan Mangard. Flush+ flush: a fast and stealthy cache attack. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*, pages 279–299. Springer, 2016.
9. Yuval Yarom and Katrina Falkner. Flush+ reload: A high resolution, low noise, l3 cache side-channel attack. In *USENIX Security Symposium*, volume 1, pages 22–25, 2014.
10. Fangfei Liu, Yuval Yarom, Qian Ge, Gernot Heiser, and Ruby B Lee. Last-level cache side-channel attacks are practical. In *2015 IEEE Symposium on Security and Privacy*, pages 605–622. IEEE, 2015.
11. Michael Weiß, Benedikt Heinz, and Frederic Stumpf. A cache timing attack on aes in virtualization environments. In *International Conference on Financial Cryptography and Data Security*, pages 314–328. Springer, 2012.
12. Meiko Jensen, Jörg Schwenk, Nils Gruschka, and Luigi Lo Iacono. On technical security issues in cloud computing. In *2009 IEEE International Conference on Cloud Computing*, pages 109–116. Ieee, 2009.
13. D. Liu and L. Zhao. The research and implementation of cloud computing platform based on docker. In *2014 11th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages 475–478, Dec 2014.
14. Miguel G Xavier, Marcelo V Neves, Fabio D Rossi, Tiago C Ferreto, Timoteo Lange, and Cesar AF De Rose. Performance evaluation of container-based virtualization for high performance computing environments. In *2013 21st Euromicro International Conference on Parallel, Distributed, and Network-Based Processing*, pages 233–240. IEEE, 2013.
15. T. Combe, A. Martin, and R. Di Pietro. To docker or not to docker: A security perspective. *IEEE Cloud Computing*, 3(5):54–62, Sept 2016.
16. Jon-Anders Kabbe. Security analysis of docker containers in a production environment. June 2017.
17. Didrik Sæther. Security in docker swarm: orchestration service for distributed software systems. Master’s thesis, The University of Bergen, 2018.
18. Leslie Lamport. The part-time parliament. *ACM Transactions on Computer Systems (TOCS)*, 16(2):133–169, 1998.
19. Butler W Lamport. How to build a highly available system using consensus. In *International Workshop on Distributed Algorithms*, pages 1–17. Springer, 1996.
20. Leslie Lamport. Leaderless byzantine paxos. In *DISC 2011*. Citeseer, 2011.
21. Miguel Castro and Barbara Liskov. Practical byzantine fault tolerance and proactive recovery. *ACM Transactions on Computer Systems (TOCS)*, 20(4):398–461, 2002.
22. Miguel Castro, Barbara Liskov, et al. Practical byzantine fault tolerance. In *OSDI*, volume 99, pages 173–186, 1999.
23. Iulian Moraru, David G Andersen, and Michael Kaminsky. There is more consensus in egalitarian parliaments. In *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles*, pages 358–372. ACM, 2013.
24. Leslie Lamport et al. Paxos made simple. *ACM Sigact News*, 32(4):18–25, 2001.
25. Docker node demote api. https://docs.docker.com/engine/reference/commandline/node_demote/.
26. Docker node promote api. https://docs.docker.com/engine/reference/commandline/node_promote/.
27. Docker api. <https://docs.docker.com/engine/api/v1.24>.
28. Giuliana Santos Veronese, Miguel Correia, Alysson Neves Bessani, Lau Cheuk Lung, and Paulo Verissimo. Efficient byzantine fault-tolerance. *IEEE Transactions on Computers*, 62(1):16–30, 2011.
29. James Cowling, Daniel Myers, Barbara Liskov, Rodrigo Rodrigues, and Liuba Shrira. Hq replication: A hybrid quorum protocol for byzantine fault tolerance. In *Proceedings of the 7th symposium on Operating systems design and implementation*, pages 177–190. USENIX Association, 2006.
30. Ramakrishna Kotla, Lorenzo Alvisi, Mike Dahlin, Allen Clement, and Edmund Wong. Zyzzyva: speculative byzantine fault tolerance. In *ACM SIGOPS Operating Systems Review*, volume 41, pages 45–58. ACM, 2007.

-
31. Alysson Bessani, João Sousa, and Eduardo EP Alchieri. State machine replication for the masses with bft-smart. In *2014 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, pages 355–362. IEEE, 2014.