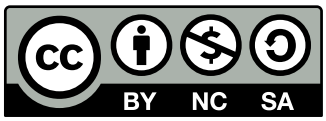
Aalto University
School of Science

Master's programme in Computer, Communication and Information Sciences

Kubernetes inter-pod container isolation

Aarni Halinen

Master's Thesis
2023



Aalto University
School of Science

Author Aarni Halinen

Title Kubernetes inter-pod container isolation

Degree programme Computer, Communication and Information Sciences

Major Computer Science

Supervisor Prof. Mario Di Francesco

Advisors M.Sc. (Tech.) José Luis Martin Navarro, M.Sc. (Tech.) Jacopo Bufalino

Date 1 June 2023 **Number of pages** 33+2 **Language** English

Abstract

The abstract is a short description of the essential contents of the thesis: what was studied and how and what were the main findings. For a Finnish thesis, the abstract should be written in both Finnish and English; for a Swedish thesis, in Swedish and English. The abstracts for English theses written by Finnish or Swedish speakers should be written in English and either in Finnish or in Swedish, depending on the student's language of basic education. Students educated in languages other than Finnish or Swedish write the abstract only in English. Students may include a second or third abstract in their native language, if they wish. The abstract text of this thesis is written on the readable abstract page as well as into the pdf file's metadata via the thesisabstract macro (see the comment in the TeX file). Write here the text that goes into the metadata. The metadata cannot contain special characters, linebreak or paragraph break characters, so these must not be used here. If your abstract does not contain special characters and it does not require paragraphs, you may take advantage of the abstracttext macro (see the comment in the TeX file below). Otherwise, the metadata abstract text must be identical to the text on the abstract page.

Keywords Kubernetes, Container, Docker, Security

Aalto-yliopisto
Perustieteiden
korkeakoulu

Tekijä Aarni Halinen

Työn nimi Opinnäyteen otsikko

Koulutusohjelma Computer, Communication and Information Sciences

Pääaine Computer Science

Työn valvoja Prof. Mario Di Francesco

Työn ohjaajat DI José Luis Martin Navarro, DI Jacopo Bufalino

Päivämäärä 1.6.2023 **Sivumäärä** 33+2 **Kieli** englanti

Tiivistelmä

Tiivistelmä on lyhyt kuvaus työn keskeisestä sisällöstä: mitä tutkittiin ja miten sekä mitkä olivat tärkeimmät tulokset. Suomenkielisen opinnäytteen tiivistelmä kirjoitetaan suomeksi ja englanniksi ja ruotsinkielisen vastaavasti ruotsiksi ja englanniksi. Suomen- tai ruotsinkielisten opiskelijoiden, joiden opinnäytteen kieli on englanti, tulee kirjoittaa tiivistelmänsä englanniksi ja koulusivistyskielellään. Muiden kuin koulusivistyskieleltään suomen- tai ruotsinkielisten tulee kirjoittaa tiivistelmänsä vain englanniksi. Opiskelija voi halutessaan lisätä opinnäytteeseensä toisen tai kolmannen tiivistelmän omalla äidinkielellään. Tämän opinnäytteen tiivistelmäteksti kirjoitetaan opinnäytteen luettavan osan lomakkeen lisäksi myös pdf-tiedoston metadataan. Kirjoita tähän metadataan kirjoitettavaa teksti. Metadatatekstissa ei saa olla erikoismerkkejä, rivinvaiho- tai kappaleenjakomerkkiä, joten näitä merkkeja ei saa käyttää tässä. Jos tiivistelmäsi ei sisällä erikoimerkkejä eikä kaipaa kappaleenjakoa, voit hyödynttää makroa abstracttext luodessasi lomakkeen tiivistelmää (katso kommentti tässä TeX-tiedostossa alla). Metadatatiivistelmatekstin on muuten oltava sama kuin lomakkeessa oleva teksti.

Avainsanat Vastus, resistanssi, lämpötila

Preface

I want to thank Professor Pirjo Professor and my instructors Dr Alan Advisor and Ms Elsa Expert for their guidance.

I also want to thank my partner for keeping me sane and alive.

Otaniemi, 9 February 2023

Aarni O. Halinen

Contents

Symbols and abbreviations

Symbols

$\uparrow$	electron spin direction up
$\downarrow$	electron spin direction down

Operators

$\nabla \times \mathbf{A}$	curl of vectorin $\mathbf{A}$

Abbreviations

K8s	Kubernetes
STRIDE	an object-oriented analog circuit simulator and design tool

1 Introduction

1.1 Problem Statement

While the sidecar pattern makes it easier to add peripheral tasks to applications, it opens up questions about application security. In Kubernetes, there is limited amount of security features available on container-level. Most of the security related policies and capabilities are defined for the Pod, which essentially means that any capability required by the main application is inherited in the sidecar. Any privilege or network policy granted for the main application can be used by the sidecar for escalation and lateral movement.

Most often, developers rely on containers by third parties for the sidecar tasks. The source code of the sidecar containers, even if it were open, can be hard or even impossible to verify for known vulnerabilities. This, combined with the limited security features for sidecars, makes any exploitable security issue in the sidecar an optimal launchpad for attack against the whole cluster. Furthermore, malicious actors can use supply chain attacks and typosquatting to trick victims into installing malicious sidecars to their clusters.

This thesis proposes a solution for limiting capabilities of sidecar without limiting those of the main container, thus extending the principle of least privilege to within the pod.

1.2 Thesis outline

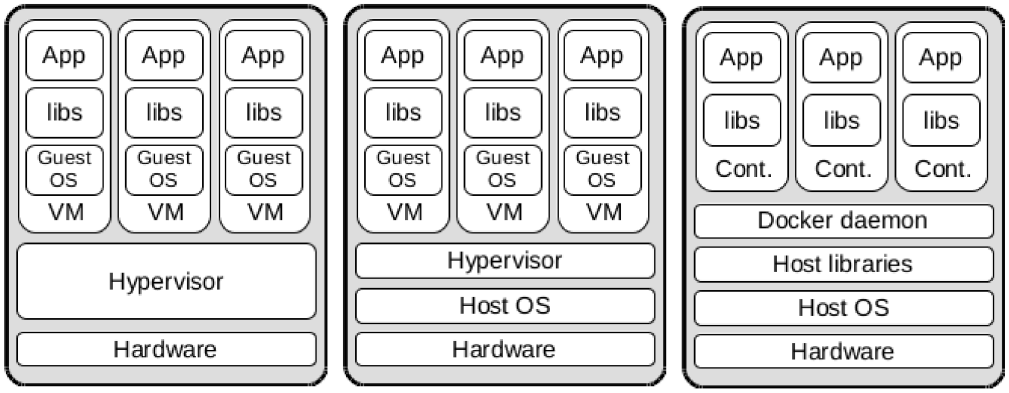
The following chapter 2 gives background about containers, Kubernetes and explains their common attack vectors. It also discusses Kubernetes networking and container network interface plugins. Chapter 5 proposes ideas for isolating sidecars from main application container. The chapter discusses both container and networking security in the context of Kubernetes Pod. Chapter 6 introduces an implementation based on the findings of the previous chapter. The pros and cons of the solution are discussed in Chapter 7. Finally, Chapter 8 discusses future research and concludes the thesis.

2 Background

2.1 Principle of least privilege, Zero trust...

2.2 Microservices architecture

2.3 Containerization and Docker



(a) Type1 hypervisor (b) Type2 hypervisor (c) Container

Figure 1: Virtualization models [20]

Figure 1 illustrates common virtualization models. Whereas traditional virtualization techniques virtualize workloads on top of a hypervisor which shares hardware resources between the virtual machines, containerization is a technique where virtualization happens on a operating system level [45]. Processes executing in containers run on the host machine kernel. However, each container is isolated to its own network, process namespace and so on; two containers on the same host OS do not know that they share resources. Furthermore, containers are similarly isolated from accessing host OS resources.

BSD jails and *chroot* can be considered early forms of containerization technology, so the idea of containers is not new [20]. Recent Linux container solutions rely on two main implementations: Linux Containers (LXC) -based solution that relies on kernel features such as control groups (cgroups) and namespaces, and a custom kernel and Linux distribution called Open Virtuozzo (OpenVZ). Docker [32] is a hugely popular container runtime that is based on LXC and provides an easy-to-use API and tooling for creating and managing containers. Docker also provides containerization for other OSes as well. However, in this thesis we focus only on the Linux implementation.

2.3.1 Linux containers

The Linux containers technology implements container isolation and containement using Linux kernel feature called namespaces [43]. Namespaces [29] are a construct that wraps a global system resource in an abstraction which makes it appear to the processes in the namespace that they have their own, isolated, instance of the global resource. There are total of eight namespaces: i) Cgroup which is used for resource management, ii) Inter-process communication (IPC) which isolates POSIX message queues etc., iii) Network which isolates network devices, stack ports etc., iv) Mount for file system isolation, v) Process ID (PID), vi) Time, vii) User for isolating user and group identifiers and viii) UTS which isolates hostnames and NIS domain names. For example, network namespace provides each container their own loopback device and even `iptables` rules. In another example, mount namespace makes sure that container has no visibility nor access to the host's or other container's file system. Compared to other namespaces that concern isolation of kernel data, cgroups focuses on limiting available system resources per namespace [43]. Each namespace can be setup with their own limits on CPU and memory usage and available devices. Using Docker as an example, setting *–cpu*, *–memory* and *–devices* options will limit available resources for the container.

Since all containers and the host machine run on same kernel, any container that manages to breakout from isolation may compromise other containers, the host and the whole kernel. To combat this container breakout, several Linux kernel security mechanisms are adopted to constrain the capabilities of containers [43]. The mechanisms include Discretionary Access Control (DAC) mechanisms like Capability [28] and Secure computing mode (Seccomp) [30], and Mandatory Access Control (MAC) mechanisms like Security-Enhanced Linux (SELinux) and AppArmor [1]. With Capability, the superuser (i.e. the root user) privilege is divided into distinct units, each of which represent a permission to process some specific kernel resources. The feature turns the binary "root/non-root" security mechanism into fine-grained access control system, which makes it easier to follow the principle of least privilege. For example, processes like web servers that just need to bind on a Internet domain privileged port (numbers below 1024) do not need to run as root; they can just be granted with `CAP_NET_BIND_SERVICE` capability instead [33]. The Seccomp mechanism constrains which system calls a process can invoke. The available system calls are defined for a container through Seccomp profile which is defined as a JSON file. The Docker default Seccomp profile [31] includes over 300 system calls. SELinux is integrated to CentOS/RHEL/Fedora distributions and utilizes a label-based enforcement model, while AppArmor is available in Debian and Ubuntu distros and adopts a path-based enforcement model [43].

2.3.2 Docker

Docker is an open-source container technology written in Go and launched in 2013 [3, text]. The platform consists of Docker Engine packaging tool, Docker image registries like the public image repository Docker Hub and Docker desktop application [2]. In

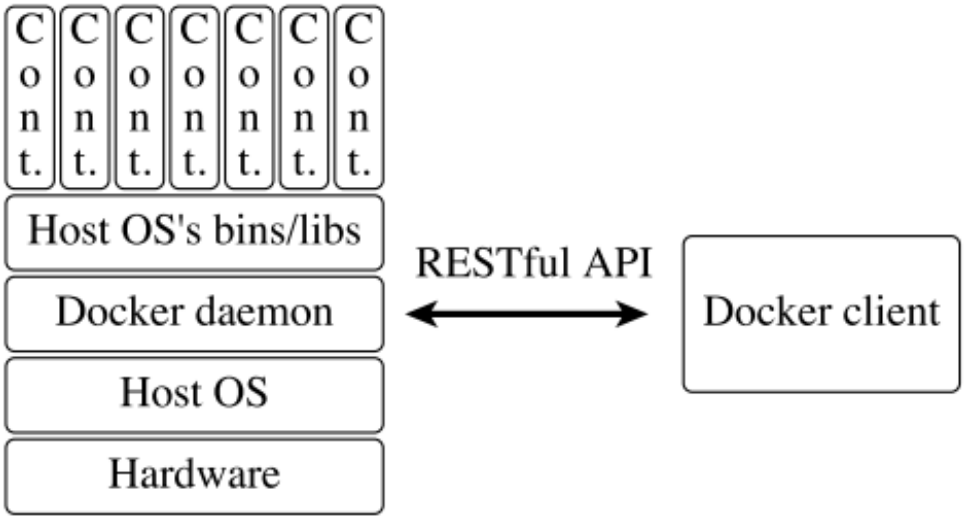


Figure 2: Architecture of Docker engine [16]

general, the engine architecture is similar to container-based virtualization, as visible in figure 2 [16]. The containers run on top of Docker daemon which manages and executes all the containers. The daemon is exposed to Docker clients via RESTful HTTP API. The Docker client is a command line tool which provides user interface for commanding the daemon and thus containers. By exposing the API outside the host machine, the architecture enables remote control of daemon with the client. For security reasons, remote communication should be secured with TLS.

Docker image is a read-only template with instructions for creating a Docker container [2]. The images are often based on another image, such as OS images `ubuntu` and `alpine`, with some additional customizations like installation of web server binaries. The customizations are added to the image as series of data layers so that each new command creates a new layer. This process makes the distribution of image more efficient since only the changes between layers need to be distributed [16]. The layering is achieved with special filesystem inspired by UnionFS which allows files and directories in different file systems to be combined into a single consistent file system.

Docker users can share their custom images publicly or privately in Docker Hub, or even host their own image registry platform. Most cloud providers also offer container registry services so even propietary software can be published in a private registry and used by other cloud services, like Kubernetes clusters. Whenever the image is not found locally, the client automatically tries to search and pull image from connected registries.

2.4 Kubernetes

Kubernetes (K8s) [8] is an open-source container orchestrator, i.e. a system for automating deployment, scaling and management of containerized applications. It allows creation of a cluster which consists of a set of servers, called Nodes, on which application containers are scheduled by the system. The automation provides resilience and efficient resource utilization for workloads in the cluster: if a container or node dies, the system tries to restart and re-schedule containers so that the desired cluster state is maintained. K8s is hosted by the Cloud Nativce Computing Foundation (CNCF), but it's origins are at Google where it was created as an open-source option for Google's propietary Borg and Omega orchestrators [18]. K8s was open-sourced in 2014.

2.4.1 Kubernetes objects

Pods are the basic atomic scheduling unit in K8s. Pods consists of one or more tightly-coupled containers with shared storage volume and networking [12]. Containers in a pod are always co-located and co-scheduled and run in a shared context, i.e. a set of Linux namespaces. Network, UTS and IPC namespaces are shared by default, and process namespace can be shared with `v1.PodSpec.shareProcessNamespace`. The common network namespace means that containers in a pod can communicate with each other via localhost, have common IP address and cannot re-use same port numbers. In addition to normal application container, Pods can include special `initContainers` that are only run on Pod startup. These pods are used for modifying Pod context before the actual workload starts. Multiple `initContainers` are run sequentially and a failing container blocks the execution of the following initialization and normal workloads. All Pods accross the cluster share same subnet and can access each other via IP address. However, connecting to a Pod with IP address is sub-optimal since Pods are ephermal and restarting a dead pod may receive a new IP address. Furthermore, horizontally scaled Pods with multiple replicas have as many IP addresses, thus making load balancing difficult. Kubernetes concept called Services solves these issues.

Instead of creating Pods directly, **Deployment** workload resources are used for creating Pods in a cluster, even with singleton Pods [12]. With Deployments, user describes the desired state in a declarative manner. The Kubernetes control loop then creates **ReplicaSet** based on the Deployment resource, which in turn guarantees the availibility of desired amount of Pods [7]. **DaemonSet** on the other hand is a workload resource that ensures all or some Nodes run a copy of a Pod. Typical usecases for daemons are running Node monitoring and logging, and network plugins which we discuss in depth in section 2.5.1.

Services are an object for exposing groups of Pods over an network [13]. The object defines a set of enpoints, i.e. the targeted pods, along with a policy about how to make the pods accessible. The targeted pods are determined with a `selector` field in the object specification. Meanwhile, the `type` field determines how the Service is exposed. There are four different `ServiceTypes` levels: i) the default `ClusterIP`

which exposes Service inside the cluster with its own IP address, ii) `NodePort` which exposes service in each Node's IP address on static port (by default within a range of 30000-32767), iii) `LoadBalancer` which exposes the Service externally using cloud provider's load balancer and iv) `ExternalName` which is used to map Service to DNS name instead of a group of Pods. The field is designed as a nested functionality; each `ServiceType` level adds up to the previous one. Ingress object can also be used for exposing Services to outside of cluster. The Ingress object requires installation of an Ingress Controller to the cluster. Cloud providers often have their own controllers and all the examples in this thesis are executed on a local cluster where no external access is needed. Thus, the controllers are left as an exercise for the reader.

Namespaces provide isolation for cluster objects and allow grouping of objects under a single name. New K8s cluster starts with four namespaces: `default`, `kube-node-lease`, `kube-public` and `kube-system`. Namespaced objects like Deployments, Services and Pods are always deployed under a namespace which is `default` if not explicitly defined. `kube-system` is the namespace for all objects created by the K8s system which we focus more on the next section 2.4.2. Namespaces also provide a scope for naming; names of resources need to be unique within a namespace, but not across namespaces. Namespaces are also used to enforce resource quotas, access control, and isolation for cluster users, for example in multi-tenancy setups. Pod Security Standards [11], which are used by Pod Security admission controller, are also defined at namespace level. Admission controllers are discussed in section 2.4.3.

2.4.2 Kubernetes components

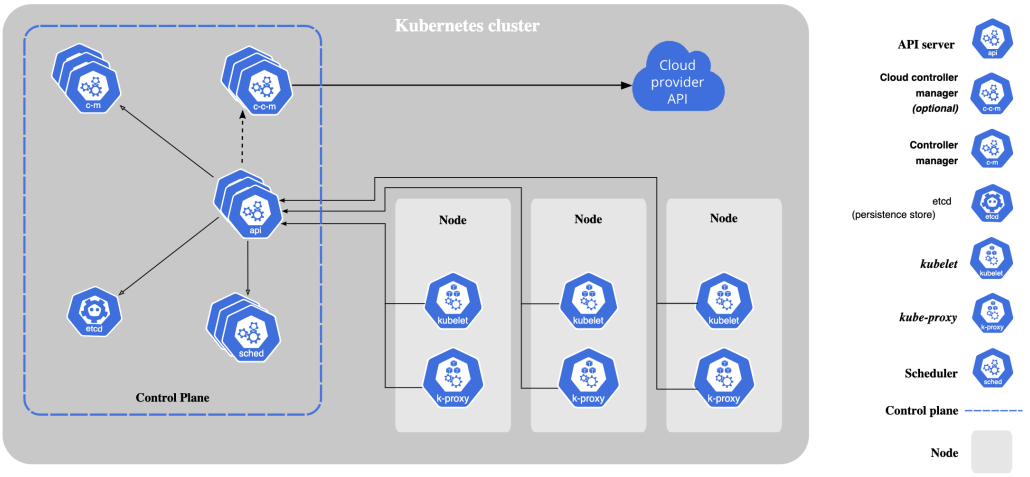


Figure 3: Kubernetes cluster architecture [9]

The figure 3 describes Kubernetes cluster with control plane and three worker nodes. The control plane consists of components that control, monitor and store the state of the cluster; essentially these are the components that are needed for complete and working Kubernetes cluster [9]. The control plane components can be run on any

worker node. However, clusters often have specialized master node for control plane components, which does not run any other containers. For fault-tolerance and high availability, control plane components should run on multiple Nodes in production environments. The control plane consists of these main components:

API server is a frontend component of the control plane. It is a stateless HTTP server which validates and authenticates commands given to the cluster. For valid commands, the server then forwards these to other control plane components. For example the `kubectl` CLI tool sends commands to the API server with HTTP. The main implementation of the server is `kube-apiserver`. The server can be horizontally scaled by running several instances on multiple Nodes and load balancing traffic between the instances.

Etcd [4] is a strongly consistent, distributed key-value store. It is the stateful component of the control plane: all of the cluster data is stored in etcd. Thus, the stability of the component is critical for the whole cluster. To tolerate failures, etcd implements a leader-based architecture. Multiple etcd clients automaitcally elect a leader instance as the source of truth. Other instances periodically update their state from the leader instance, so that the state stays eventually conistent across all the instances. On leader failure, the other instances automatically elect new leader to keep the system functioning.

Scheduler watches for newly created Pods that have no assigned worker node, and selects one of the active Nodes for them to run on [9]. The scheduling takes into account resource availibility on Nodes, Pod resource requirements, object specification affinity rules and hardware, software and policy constraints, among others.

Controller manager is a control plane component that runs all the controller loop processes [9]. Controller loops, like the Deployment controller, continously watch the current and desired cluster state. When the states differ, they send commands via the API server so that the cluster moves towards the desired state. All the built-in controllers are compiled into a single binary, even though the controllers are logically different processes.

Each Node also has components that are essential for Kubernetes to work properly. **Kubelet** is an agent that makes sure containers are running in a Pod [9]. It receives a set of pod specification from the API server and ensures that containers are running on the Node, follow the pod specifications and are healthy. Any container which is not created by Kubernetes is not managed by `kubelet`. **Kube-proxy** maintains network rules on Nodes. Part of the Serivce objects' networking is implemented by **kube-proxy**; the proxy writes `iptables` rules that route the traffic [24].

2.4.3 Admission controllers

Admission controllers are a feature of Kubernetes API server, made for validating and modifying requests made to the server [5]. The controllers execute prior to persistence of the object, but after the request is authenticated and authorized by the server. Several important features of Kubernetes are implemented with admission controllers, and these should be enabled in a properly configured API server. In addition to the built-in controllers, Kubernetes provides `MutatingAdmissionWebhook` and

`ValidatingAdmissionWebhook` controllers for building own admission logic.

Admission controllers can be validating, mutating, or both [5]. Mutating controllers may modify related objects to the requests they admit, while validating controllers either approve or reject the request. The control process executes mutating controllers first so that no mutations happen after the validation. If a controller in either phase rejects the request, the request is not processed further and error is returned to the end-user.

The important Admission controller for the scope of this thesis is the `PodSecurity` controller. The admission controller validates Pods before they are admitted, making sure that the requested Pod security context and other restrictions are permitted in the namespace that the Pod is assigned to [5]. The controller is enabled by default, and can be taken into use just by configuring Pod Security Admission labels for Namespace objects.

The labels use `pod-security.kubernetes.io/<MODE>: <LEVEL>` format, where MODE defines the action to be taken when the security level is violated and the LEVEL is a pre-defined Pod Security Standard level. The three available levels are `privileged`, `baseline` and `restricted` [11].

The actions available are i) `enforce`, which will reject the pod on violation, ii) `audit`, which triggers an event about the violation in the audit log, and iii) `warn`, which triggers user-facing warning about the violation [10]. A namespace can configure any or all three of the available modes, and even set a different level for the modes. For example, it is possible to warn the user about security policy violation without blocking the request by setting the `warn` mode more restrictive than the `enforce`.

2.4.4 Sidecar pattern

As mentioned before, Pods are the basic scheduling abstraction in Kubernetes and they support management and co-scheduling of multiple containers as an atomic unit. This co-scheduling and management of multiple symbiotic containers as a single unit enables multi-container application design patterns to emerge [17]. Sidecar pattern is the most common of these design patterns. As an example of this pattern, the main application container can be a simple web server pairted with a container that collects server's log file and streams them to a centralized log management system. Another example of this pattern is Istio service mesh [27] and its Envoy proxy sidecar which routes all traffic thorugh the Istio control plane for management, observability and security reasons.

In the pattern, peripheral tasks such as logging, configuration and observability are isolated from the main application into helper containers. These containers, sidecars, are tightly-coupled to the parent application container and should share the lifecycle of the parent. Even though the functionality of the sidecars could be build into the main container, there are benefits for using separate containers [17]. The isolation allows tweaking of containers' cgroups so that CPU cycles can be prioritized for the main container. The isolation also provides failure containement boundary between the main and sidecar processes. Since container is also the unit of deployment, the sidecar containers could be developed, tested and deployed independently from one another.

Sidecar containers can also be developed with different tools and dependencies, and in a way that they are re-usable with other application containers. However, this multiplies the amount of moving parts of the overall system, which increases the size of test matrix considering all of the container version combinations that might be seen in the production environment.

2.5 Kubernetes network model

Integral part of Kubernetes cluster is how nodes and resources are networked together. Specifically, the networking model needs to address four different type of networking problems: i) intra-Pod (ie. container-to-container within same Pod) communication, ii) inter-Pod communication between Pods, iii) Service-to-Pod communication and iv) communication from external sources to Services [6]. The model also requires that each Pod is IP addressable and can communicate with other Pods without network address translation (NAT), even when Pods are scheduled on different hosts [48]. All agents on a host should also be able to communicate with Pods on the same host. The implementation of this model is not part of Kubernetes, but is handed to special plugins that implement Container Network Interface (CNI) specification.

2.5.1 Container Network Interface

The Container Network Interface (CNI) [14] is a networking specification, which has become de facto industry standard for container networking. It is backed by CNCF [48]. CNI was first developed for the container runtime `rkt`, but it is supported by all container runtimes and there is a large number of implementations to choose from [41]. Most of the container orchestrators have adopted the specification as their networking solution. The biggest outlier is Docker Swarm, which instead implements `libnetwork` [34].

The CNI specification has five distinct definitions: i) a format for network configuration, ii) a execution protocol between the container runtimes and the plugin binary, iii) a procedure for the runtime to interpret the configuration and execute the plugins, iv) a procedure for deletegating functionality between the plugins and v) data types for plugins to return their results to the runtime [14]. The network configuration is defined as a JSON file and it includes a list of plugins and their configuration. The container runtime interprets the configuration file at plugin execution time and transforms it into a form to be passed to the plugins. The execution protocol defines a set of operations (ADD, DEL, CHECK, VERSION) for adding and removing containers from the network. The operation command, similarly to other protocol parameters, are passed to the plugins via OS environment variables. The configuration file is supplied to the plugin via stdin. On successful execution, the plugin returns the result via stdout with a return code of 0. On errors, the plugin returns a specific JSON structure error message to stderr and a non-zero return code. When the runtime mutates a container network, it results in a series of ADD, DELETE or CHECK executions. These are then executed in same order as defined in the `plugins` list, or reversed order for DELETE executions. Each plugin then returns either `Success` or

`Error` JSON object. The execution of a series of operations ends when it encounters the first Error response, or when all the operations have been performed.

The CNI plugin must provide at least connectivity and reachability for the containers [40]. For connectivity, each Pod must have a NIC for communication outside its networking namespace. The NIC must have IP address reachable from the host Node, so that cluster processes like Kubelet health and readiness checks can reach the Pod. Reachability means that all Pods can be reached from other Nodes directly without NAT. Thus, each Pod receives an unique IP address from the `PodCIDR` range configured on the Node by the Kubelet bootstrapping phase. The end-to-end reachability between different Node `PodCIDR`s is established by encapsulating in the overlay network (for example with VXLAN) or orchestrating on the underlay network, e.g. with Border Gateway Protocol (BGP).

Since Kubernetes does not provide networking between the Pods, it has no capabilities to enforce network isolation between workloads. Thus, another key feature for CNI plugins is enforcing network traffic rules. Kubernetes provides a common object called `NetworkPolicy` for CNI plugins to consume. The NetworkPolicy specification consists of a `podSelector` that specifies pods that are subject to the policy and `policyTypes` to specify Ingress and Egress rules for the traffic [15] to the target Pod. Each rule includes `to` or `from` field for selecting Pod, Namespace or IP address block in CIDR notation on the other side of the connection, and `ports` field for explicitly specifying which ports and protocols are part of the rule. The policies are additive; when multiple rules are defined for a Pod, the traffic is restricted to what is allowed by the union of the policies. Many CNI plugins also introduce Custom Resource Definitions for their own, more granular, network policy rules.

While all CNI plugins meet the requirements listed above, they may differ in architecture significantly. The plugins can be classified based on which OSI model network layers they operate on, which Linux kernel features they use for packet filtering and which encapsulation and routing model they support for inter-host and intra-host communication between Pods. In this thesis, we focus on three different CNI plugins: Calico, Cilium and Multus.

2.5.2 Calico

Calico [49] is an open-source CNI plugin with modular architecture that supports wide range of deployment options. Each Pod created to the Calico network receives one end of a virtual ethernet device link as its default `eth0` network interface, while other end is left dangling on the host Node [38]. The Pod end of the link receives IP address from Pod CIDR, but the Node end does not. Instead, a `proxy_arp` flag is set on the on the host side of the interface while containers have a route to link-local address `169.254.1.1`, thus making the host behave like a gateway router. For routing packets between Nodes, Calico creates a VXLAN overlay network. Optionally, Calico supports IP-in-IP overlay or non-overlay network with BGP protocol.

On each Node, a `calico-node` daemon setups CNI plugin, IPAM and possible eBPF programs. The daemon subscribes to Kubernetes API for Pod events and manages both container and host networking namespaces. Calico also deploys a single-container

`calico-kube-controllers` Pod into the Kubernetes control plane. The container executes a binary that consists of controller loops for Namespace, NetworkPolicy, Node, Pod and ServiceAccount Kubernetes objects. The Calico project also introduces own CLI tool, called `calicoctl` [51], for managing Calico's custom resources. The tool provides extra validation for the resources which is not possible with `kubectl`.

Calico supports Kubernetes NetworkPolicies as well as its own namespaced `projectcalico.org/v3.NetworkPolicy` Custom Resource Definition (CRD). Both of the policies work on OSI layers L3 (identity, e.g. IP address) and L4 (ports). Compared to the built-in policy, the Calico policy includes features such as policy ordering, log action in rules, more flexible matching criteria (e.g., mathcing on ServiceAccounts) [50]. The policy can also match on other Calico CRDs such as **HostEndpoints** and **NetworkSets**, which allows implementing rules on host interfaces and non-Kubernetes resources. If Calico is installed along Istio service mesh, the Calico Network Policy can enforce L7 (e.g. HTTP methods and URL paths) policies on the Envoy proxy. For policies that are not tied to a Kubernetes namespace, Calico provides a `GlobalNetworkPolicy` CRD.

2.5.3 Cilium

Cilium [21] is one of the most advanced and powerful CNI plugins for Kubernetes. Similarly to Calico, it creates virtual ethernet device for each Pod and sets one side of the link into Pod's network namespace [39] as the default interface. Cilium then attaches extended Berkeley Packet Filter (eBPF) programs to ingress traffic control (`tc`) hooks of these virtual ethernet devices for intercepting all incoming packets from the Pod. The packets are intercepted and processed before the network stack and thus `iptables`, reducing latency 20%-30% and even doubling the throughput of packets in some scenarios [15]. The network between Pods running on different hosts is handled by default with VXLAN overlay, but there is support for Geneve interfaces and native-routing with BGP protocol as well [21].

The Cilium system consists of an agent (`cilium-agent`) daemon running on each Node, one or more operator (`cilium-operator`) Pods and a CLI client (`cilium`) [22]. The agent daemons subscribe to events from Kubernetes API and manage containers' networking and eBPF programs. The CLI tool, which is installed on each agent, interacts with the REST API of the agent and allows inspecting the state and status of the local agent. The tool should not be confused with Cilium management CLI tool, also incidentally named `cilium`, which is typically installed remote from the cluster. The operator is responsible for all management operations which should be handled once for the entire cluster, rather than once for each Node. This includes for example registering of CRDs.

While default Kubernetes Network Policy provides security on OSI layers L3 and L4, Cilium provides CRDs that also support for L7 policies [23]. If L7 policies exist, the traffic is directed to Envoy instance bundled into the agent Pod which filters the traffic. Unlike on layers 3 and 4, policy violation does not result in dropped packet but an application protocol specific denied message. For example, HTTP traffic is denied with `HTTP 403 Forbidden` and DNS requests with `DNS REFUSED`. Cilium provides

`CiliumNetworkPolicy` CRD that supports all L3, L4 and L7 policies. Cilium also provides `CiliumClusterwideNetworkPolicy` custom resource which is used to apply network rules to every namespace in the cluster or even to nodes when using `nodeSelector`.

As even more advanced features, Cilium also includes natively `kube-proxy` replacement, encryption for Cilium-managed traffic and Service Mesh, among others. By default, `kube-proxy` uses `iptables` to route the Service traffic [24]. With `kubeProxyReplacement` installation option, Cilium implements Service load-balancing as XDP and TC programs on Node network stack. For encryption, Cilium supports both IPsec and WireGuard implementations [25]. The Service mesh performs variety of features directly in eBPF, thus functioning without sidecar containers or proxying requests through the agent Pod's Envoy [35]. Since all features are not available as eBPF programs or on all kernel versions, Cilium automatically probes the underlying kernel and automatically reverts to Envoy proxy when needed. For capabilities beyong the built-in mesh, Cilium also provides an integration with Istio.

2.5.4 Multus

Traditionally CNI plugins only provide a single network interface for a Pod, apart from loopback device. Multus [36] is a CNI plugin that allows attaching multiple network interfaces for a Pod. It does not provide any connectivity or reachability for the containers like other plugins. Instead, it is installed as the first plugin in the CNI plugin chain. When executed, the plugin delegates interface creation to other installed plugins. Since Multus does not provide any networking and thus does not independently, it is often called "meta plugin" to distinguish it from common CNI plugins like the previous Calico and Cilium.

Multus system includes a binary, a CNI configuration file and a namespaced `NetworkAttachmentDefinition` CRD that is used to define network interfaces used in Pods. The binary and the configuration file are often installed to cluster Nodes via a DaemonSet. The daemon consists of an *initContainer* that copies the binary into the `/opt/cni/bin` directory, and a daemon container that setups the configuration file and optionally spawns a HTTP server for additional features such as metrics [36]. The configuration file satisfies the CNI specification with few extra attributes of which the combination of `clusterNetwork` and `defaultNetworks` or `delegates` are imperative for the CNI plugin to function [37]. The `clusterNetwork` specifies the main network of the cluster, which implements the `eth0` interface and Pod IP address. The `defaultNetworks` is an optional array of networks that should be added for any Pods by default. The values can be names of the `NetworkAttachmentDefinition` objects or paths to CNI plugin's JSON configuration files. Optionally, the `delegates` attribute can be used; it supports similar format of values. In this scenario, the first element of the array functions as `clusterNetwork` and the rest are infered as `defaultNetworks`.

Attaching additional interfaces to workloads is most often configured by adding a special annotations field `k8s.v1.cni.cncf.io/networks` to workload resource definitions. In the simplest configuration, the field takes a comma-separated list of

`NetworkAttachmentDefinition` names as input. The network interface identifiers can be modified by giving the attachment input in *name@interface-identifier* format. Otherwise, Multus names the interfaces *net0*, *net1* and so on. If extra configuration for the networks is needed, the annotation also supports a JSON array format.

2.5.5 Extended Berkeley Packet Filter

Berkeley Packet Filter (BPF, or nowadays often cBPF) was originally developed in early 1990s as a high-performance tool for user-space packet captures [44]. BPF works by deploying the filtering part of the application, `packet filter`, in the kernel-space as an agent. The `packet filter` is provided with a program (often denoted as BPF program) consisting of BPF instructions, which works as a set of rules for selecting which packets are of interest in the user-space application and should be copied from kernel-space to user-space. The instuctions are executed in a register-based pseudo machine. Since network monitors are often interested only in subset of network traffic, this limits the number of expensive copy operations across the kernel/user-space protection boundary only to packets that are of interest in the user-space application. A notable usecase for BPF is *libpcap* library, which is used by network monitoring tool called `tcpdump`.

Later in the 2010s the Linux community realized that BPF and it's ability to instrument the kernel could benefit other areas than packet filtering as well [52]. This reworked version of BPF was first merged in to Linux kernel in 2014 and is publicly called extended Berkeley Packet Filter (eBPF) to distinguish it from the original cBPF. The kernel development community continues to call the newer version BPF, but instead of the original acronym consider it a name of a technology. Similarly to the kernel community, the term BPF always refers to the eBPF in this thesis.

The eBPF programs are compiled to bytecode and loaded to kernel with `bpf()` system call [46]. Most often programs are written in restricted C and compiled with LLVM Clang compiler to bytecode. It also possible to use eBPF assembly instructions and `bpf_asm` utility for converting instructions to bytecode. eBPF programs follow a event-driven architecture: a loaded eBPF program is hooked to a particular type of event and each occurence of the event triggers the program execution.

For networking purposes, there are two eBPF hooks available for intercepting and mangling, forwarding or dropping network packets: eXpress Data Path (XDP) and Traffic Control (TC) [46]. In Cloudflares DDoS testing benchmark [19], XDP program was capable to drop 10 million and TC program 2 million packets per second, while common `iptables` INPUT rule was able to drop less than one million packets per second.

XDP programs are attached to a network interface controller (NIC) and can handle only incoming packets [42]. The programs are called directly by the NIC driver if it has XDP support, thus executing before packets enters the network stack. This skips expensive packet parsing and memory allocation operations, and allow XDP programs to run at very high throughput. Thus, even the main networking buffer *skbuff* is not populated. Some SmartNICs even support offloading the program to the NIC's own processor from host CPU, improving host machine performance even further [26]. If

the driver does not support XDP, generic XDP is used and the programs run after the packet has been parsed by the network stack.

XDP programs can read and modify contents of the packets [52]. Since the packets are not parsed the network stack, the programs have to work with raw packets and implement own parsing functionality. The program's return value determines how the packet should be processed further. With `XDP_DROP` and `XDP_PASS` return values, the packet can be dropped or passed further to the networking stack respectively. The packet can also be bounced back to the same NIC it arrived on with `XDP_TX`, usually after modifying the packet contents. `XDP_REDIRECT` is used for redirecting the packet to a different NIC, CPU or even to another socket.

TC programs are executed when both incoming and outgoing packets reach kernel traffic control function within the Linux network stack [52]. The ingress hook executes after the packet is parsed to *skbuff* but before most of the network stack. On egress the stack is traversed in reverse, thus the hook executes after most of the network stack. TC programs can read and write directly to packet in memory. Similarly to XDP programs, the return value of the program determines further processing of the packet. The packet can be passed furhter in the stack with `TC_ACT_OK`, dropped with `TC_ACT_SHOT`, or the modified packet can be redirected back to the start of the classification with `TC_ACT_RECLASSIFY`, among others.

3 Pod Privilege Escalation

This chapter discusses securing K8s Pods from privilege escalation. We start by introducing possible attack scenarios and then propose a solution for mitigating these issues. The solution itself sets some constraints on the network solution we discuss in next chapter. As a baseline requirement, the networking solution should not enable sidecar container breakouts. Finally, we discuss some noted limitations of the solution.

3.1 Example attack scenarios

All of the example attack scenarios start by attacker getting shell access to a container running in a Pod, usually through a remote code exectuion (RCE) flaw on the container application and then executing reverse shell inside the Pod. The scope of the examples is not in the initial attack, but in the Pod template misconfiugrations that then provide some path for the attacker to escalate the attack and even take over the whole cluster in the end.

TODO: add file

The first attack scenario includes a privileged container, as described in file 3.1. Privileged containers have all the capabilities of the host machine, so practically they can be perform almost any action available on the host. This includes but is not restricted to for example "pipeing" (TODO: better word for undock.sh and other scripts) commands as root to the Node's shell and mounting the host's disks. If combined with `hostPID: true`, the attacker can see all the processes on the host, and use `nsenter` to execute commands in the other processes' namespaces.

TODO: add file

The second example file 3.1 does not have similar privileges for executing commands as the first, but has unlimited access for mounting the whole host's filesystem, with both read and write access. Thus, the attacker can try to find any credentials stored on the host machine and use these to escalate the attack. Important credentials include `kubeconfig` files which store access token to K8s API server, ServiceAccount tokens that may have been mounted on any Pod on the host, SSH keys and hashed user passwords in `/etc/shadows`.

TODO: some extra issues (finding .kubeconfig files, running on control-plane -> etcd and secrets within, cloud metadata)

3.2 Hardening containers

For Kubernetes versions bigger than 1.25, the easiest way to enforce secure Pod configurations is to use Pod Security Admission controller, which we discussed in chapter 2.4.3. The `restricted` Pod Security Standard aims on Pod hardening best practices [11], so we will use it in the solution. The table 1 describes all the fields affected by the standard.

Table 1: Restricted security standard enforcement

Field name	Usage	Allowed values
hostPID, hostIPC, hostNetwork	Controls whether container uses host's PID, IPC and network namespace.	false
privileged	Controls whether Pod can run privileged containers.	false
capabilities.add	Defines Linux capabilities for the container.	NET_BIND_SERVICE
capabilities.drop	Defines Linux capabilities for the container.	ALL
volumes[*]	All volume types are not allowed. For example, hostPath, that maps host directories, are not allowed.	`volumes[*].configMap`, `volumes[*].csi`, `volumes[*].downwardAPI`, `volumes[*].emptyDir`, `volumes[*].ephemeral`, `volumes[*].persistentVolumeClaim`, `volumes[*].projected`, `volumes[*].secret`
hostPort	Expose container via host's network port.	undefined
container.apparmor.security.beta.kubernetes.io/* annotation	Sets the AppArmor profile used by containers. On supported hosts, the runtime/default AppArmor profile is applied by default.	runtime/default, localhost/*
seLinuxOptions	Sets the SELinux context of the container.	Set if supported by environment.
procMount	The default /proc masks are set up to reduce attack surface, and should be required.	Default
seccompProfile.type	Sets the seccomp profile used to sandbox containers.	RuntimeDefault or Localhost
sysctls[*].name	Sysctls can disable security mechanisms or affect all containers on a host, and should be disallowed except for an allowed "safe" subset.	kernel.shm_rmid_forced, net.ipv4.ip_local_port_range, net.ipv4.ip_unprivileged_port_start, net.ipv4.tcp_syncookies, net.ipv4.ping_group_range
allowPrivilegeEscalation	Restricts escalation to root privileges.	false
runAsNonRoot	Controls whether container can run as root user.	true
runAsUser, runAsGroup	Controls the user and	Set both to non-zero

[47]

- Privileged container
- CAP_SYS_ADMIN, mounting /proc and chroot
- CAP_SYS_PTRACE, shellcode injection to running program, nc 172.17.0.1 on port running shell
- Mounted docker socket, creating privileged containers

3.3 Pod Security Admission

3.4 Solution

4 Network Isolation

5 Research material and methods

5.1 IPTables

- If executed from containers in Pod (init, lifecycle, wrapper container), it breaks Security admission rules (root user and NET_ADMIN)
- Can be executed from Node itself, using DaemonSet (sort of a CNI plugin), but a bit hacky.
- Use owner module for catching egress packets (userId, groupId, processId)

5.2 eBPF program firewall

- XDP works only for ingress
- TC needs some way to catch egress from sidecar

5.3 Own pod for sidecar

- Guaranteed to work, since own network namespaces. What type of issues arise from this?
- Need to force pods to same node, for common volumes (if using host as storage)
- Implementation by hand, or Admission controller that catches sidecars?
- Loopback is not the same anymore. DNAT that changes localhost to external? => impossible with IPTables!

5.4 Multus

- Requires breaking sidecar pattern with multiple Pods
- Allows use of custom IP addresses
- Hard to implement between nodes, affinity rules
- Loopback not easy to hijack for forwarding to new IPs

6 Solution Evaluation

7 Discussion

8 Conclusion

References

[1] AppArmor. *AppArmor*. 2022. URL: https://apparmor.net/ (visited on 03/31/2023).

[2] Docker Authors. *Docker overview*. 2023. URL: https://docs.docker.com/get-started/overview/ (visited on 04/06/2023).

[3] Docker Authors. *Use containers to Build, Share and Run your applications*. 2023. URL: https://www.docker.com/resources/what-container/ (visited on 04/06/2023).

[4] etcd Authors. *etcd*. 2023. URL: https://etcd.io/ (visited on 04/05/2023).

[5] Kubernetes Authors. *Admission Controllers Reference*. 2023. URL: https://kubernetes.io/docs/reference/access-authn-authz/admission-controllers/ (visited on 04/06/2023).

[6] Kubernetes Authors. *Cluster Networking*. 2022. URL: https://kubernetes.io/docs/concepts/cluster-administration/networking/ (visited on 03/09/2023).

[7] Kubernetes Authors. *Deployments*. 2023. URL: https://kubernetes.io/docs/concepts/workloads/controllers/deployment/ (visited on 04/04/2023).

[8] Kubernetes Authors. *Kubernetes*. 2023. URL: https://kubernetes.io/ (visited on 03/31/2023).

[9] Kubernetes Authors. *Kubernetes components*. 2023. URL: https://kubernetes.io/docs/concepts/overview/components/ (visited on 04/05/2023).

[10] Kubernetes Authors. *Pod Security Admission*. 2023. URL: https://kubernetes.io/docs/concepts/security/pod-security-admission/ (visited on 04/26/2023).

[11] Kubernetes Authors. *Pod Security Standards*. 2023. URL: https://kubernetes.io/docs/concepts/security/pod-security-standards/ (visited on 04/04/2023).

[12] Kubernetes Authors. *Pods*. 2023. URL: https://kubernetes.io/concepts/workloads/pods/ (visited on 04/04/2023).

[13] Kubernetes Authors. *Services*. 2023. URL: https://kubernetes.io/docs/concepts/services-networking/service/ (visited on 04/04/2023).

[14] The CNI Authors. *Container Network Interface (CNI) Specification*. 2023. URL: https://www.cni.dev/docs/spec/ (visited on 04/20/2023).

[15] Gerald Budigiri et al. "Network policies in kubernetes: Performance evaluation and security analysis". In: *2021 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*. IEEE. 2021, pp. 407–412.

[16] Thanh Bui. "Analysis of docker security". In: *arXiv preprint arXiv:1501.02967* (2015).

[17] Brendan Burns and David Oppenheimer. "Design patterns for container-based distributed systems". In: *8th {USENIX} workshop on hot topics in cloud computing (HotCloud 16)*. 2016.

[18] Brendan Burns et al. "Borg, Omega, and Kubernetes". In: *Communications of the ACM* 59.5 (2016), pp. 50–57.

[19] Cloudflare. *How to drop 10 million packets per second*. 2018. URL: https://blog.cloudflare.com/how-to-drop-10-million-packets/ (visited on 03/15/2023).

[20] Theo Combe, Antony Martin, and Roberto Di Pietro. "To docker or not to docker: A security perspective". In: *IEEE Cloud Computing* 3.5 (2016), pp. 54–62.

[21] Cilium Developers. *Cilium*. 2023. URL: https://cilium.io/ (visited on 03/14/2023).

[22] Cilium Developers. *Component overview*. 2023. URL: https://docs.cilium.io/en/v1.13/overview/component-overview/ (visited on 04/12/2023).

[23] Cilium Developers. *Component overview*. 2023. URL: https://docs.cilium.io/en/v1.13/security/policy/language/ (visited on 04/12/2023).

[24] Cilium Developers. *Kubernetes Without kube-proxy*. 2023. URL: https://docs.cilium.io/en/v1.13/network/kubernetes/kubeproxy-free (visited on 04/12/2023).

[25] Cilium Developers. *Transparent Encryption*. 2023. URL: https://docs.cilium.io/en/v1.13/security/network/encryption/ (visited on 04/12/2023).

[26] Cilium developers. *Program types*. 2018. URL: https://docs.cilium.io/en/latest/bpf/progtypes/ (visited on 03/16/2023).

[27] Istio developers. *The Istio service mesh*. 2023. URL: https://istio.io/latest/about/service-mesh/ (visited on 04/06/2023).

[28] Linux Developers. *capabilities(7) — Linux manual page*. 2023. URL: https://man7.org/linux/man-pages/man7/capabilities.7.html (visited on 03/31/2023).

[29] Linux Developers. *namespaces(7) — Linux manual page*. 2023. URL: https://man7.org/linux/man-pages/man7/namespaces.7.html (visited on 03/31/2023).

[30] Linux Developers. *seccomp(2) — Linux manual page*. 2023. URL: https://man7.org/linux/man-pages/man2/seccomp.2.html (visited on 03/31/2023).

[31] Docker. *Docker default Seccomp profile*. 2022. URL: https://github.com/moby/moby/blob/23.0/profiles/seccomp/default.json (visited on 03/31/2023).

[32] Docker. *Docker overview*. 2018. URL: https://docs.docker.com/get-started/overview/ (visited on 03/16/2023).

[33] Docker. *Docker security*. 2023. URL: https://docs.docker.com/engine/security/ (visited on 03/31/2023).

[34] Docker. *libnetwork*. 2023. URL: https://github.com/moby/moby/tree/master/libnetwork (visited on 03/10/2023).

[35] Thomas Graf. *Cilium Service Mesh – Everything You Need to Know*. 2023. URL: https://isovalent.com/blog/post/cilium-service-mesh/ (visited on 04/12/2023).

[36] Network Plumbing Working Group. *Multus-CNI*. 2023. URL: https://github.com/k8snetworkplumbingwg/multus-cni (visited on 04/25/2023).

[37] Network Plumbing Working Group. *Multus-CNI Configuration Reference*. 2023. URL: https://github.com/k8snetworkplumbingwg/multus-cni/blob/master/docs/configuration.md (visited on 04/25/2023).

[38] The Kubernetes Networking Guide. *Calico*. 2023. URL: https://www.tkng.io/cni/calico/ (visited on 03/14/2023).

[39] The Kubernetes Networking Guide. *Cilium*. 2023. URL: https://www.tkng.io/cni/cilium/ (visited on 03/14/2023).

[40] The Kubernetes Networking Guide. *CNI*. 2023. URL: https://www.tkng.io/cni/ (visited on 04/20/2023).

[41] Michael Hausenblas. *Container Networking*. O'Reilly Media, Incorporated, 2018.

[42] Toke Høiland-Jørgensen et al. "The express data path: Fast programmable packet processing in the operating system kernel". In: *Proceedings of the 14th international conference on emerging networking experiments and technologies*. 2018, pp. 54–66.

[43] Xin Lin et al. "A measurement study on linux container security: Attacks and countermeasures". In: *Proceedings of the 34th Annual Computer Security Applications Conference*. 2018, pp. 418–429.

[44] Steven McCanne and Van Jacobson. "The BSD Packet Filter: A New Architecture for User-level Packet Capture." In: *USENIX winter*. Vol. 46. 1993.

[45] Dirk Merkel et al. "Docker: lightweight linux containers for consistent development and deployment". In: *Linux j* 239.2 (2014), p. 2.

[46] Sebastiano Miano et al. "A framework for eBPF-based network functions in an era of microservices". In: *IEEE Transactions on Network and Service Management* 18.1 (2021), pp. 133–151.

[47] National Security Agency (NSA) and Cybersecurity and Infrastructure Security Agency (CISA). *Kubernetes Hardening Guide*. 2022. URL: https://media.defense.gov/2022/Aug/29/2003066362/-1/-1/0/CTR_KUBERNETES_HARDENING_GUIDANCE_1.2_20220829.PDF (visited on 05/05/2023).

[48] Shixiong Qi, Sameer G Kulkarni, and KK Ramakrishnan. "Assessing container network interface plugins: Functionality, performance, and scalability". In: *IEEE Transactions on Network and Service Management* 18.1 (2020), pp. 656–671.

[49] Tigera. *About Calico*. 2023. URL: https://docs.tigera.io/calico/3.25/about (visited on 04/13/2023).

[50] Tigera. *About Network Policy*. 2023. URL: https://docs.tigera.io/calico/latest/about/about-network-policy (visited on 04/13/2023).

[51] Tigera. *Install calicoctl*. 2023. URL: https://docs.tigera.io/calico/3.25/operations/calicoctl/install (visited on 04/13/2023).

[52] Marcos AM Vieira et al. "Fast packet processing with ebpf and xdp: Concepts, code, challenges, and applications". In: *ACM Computing Surveys (CSUR)* 53.1 (2020), pp. 1–36.

A Contents of an appendix

Appendices are not essential in a thesis, and so you must plan the content of your thesis as if it does not contain an appendix. The appendix cannot be used as a dumping ground for text and ideas from an overgrown thesis.

An appendix is an independent entity, even though it complements the thesis. So, the appendix is not, say, just a list or image or table, but contains explanatory text as well that indicates the purpose of its content. It can contain code listings, like the one below for a simplified list of commands to create an appendix.

```
\clearpage
\appendix
\addcontentsline{toc}{section}Contents of an appendix}
\thispagestyle{empty}
\section*{Contents of an appendix}
...
text
```

Equation numbering in the appendix forms a separate, complete entity. Here are a couple of examples how equations in an appendix are numbered: