FRANKFURT UNIVERSITY OF APPLIED SCIENCES

MASTER'S THESIS

Containerized multi-level deployment for a distributed adaptive microservice application

Author:
Tim WIßMANN

Supervisors:
Prof. Dr. Peter THOMA
Prof. Dr. Eicke GODEHARDT

A thesis submitted in fulfilment of the requirements for the degree of Master of Science

in the course

Allgemeine Informatik Master

March 17th, 2023

Declaration of Authorship

I, Tim Wißmann, declare that this thesis titled, 'Containerized multi-level deployment for a distributed adaptive microservice application' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: ____

Date: ____

"Thanks to my solid academic training, today I can write hundreds of words on virtually any topic without possessing a shred of information, which is how I got a good job in journalism."

Dave Barry

FRANKFURT UNIVERSITY OF APPLIED SCIENCES

Abstract

Faculty 2 - Computer Science and Engineering
Allgemeine Informatik Master

Master of Science

Containerized multi-level deployment for a distributed adaptive microservice application

by Tim Wißmann

The Thesis Abstract is written here (and usually kept to just this page). The page is kept centered vertically so can expand into the blank space above the title too…

Acknowledgements

The acknowledgements and the people to thank go here, don't forget to include your project advisor...

Contents

List of Figures

List of Tables

Abbreviations

LAH	**L**ist **A**bbreviations **H**ere

Chapter 1

Introduction

This thesis is about the deployment of a microservice application.

1.1 Scope

1.2 Intended audience

1.3 Outline

Chapter 2

Background and related work

2.1 Application under study

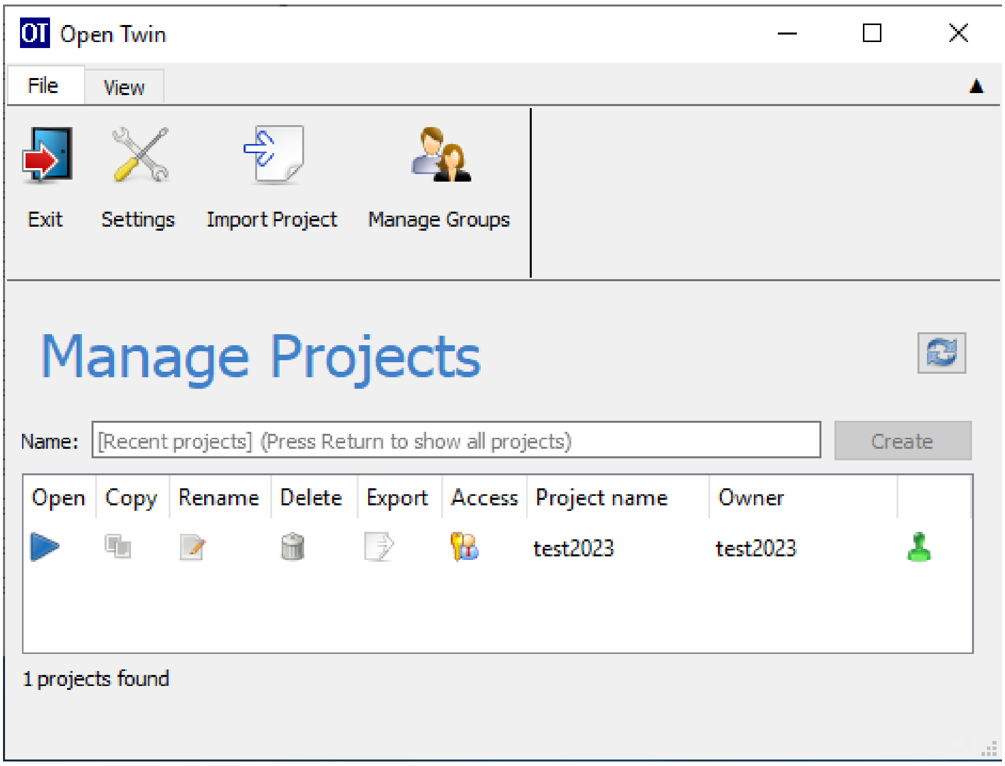
OpenTwin (OT) is a open-source simulation platform developed by the university of Applied Sciences in Frankfurt, Germany. It covers features like computer aided design and meshing and is also a physics simulation (having solvers for Finite Integration and PHREEC). The projects can be administered by a user and group management (see Figure 2.1). Furthermore, all changes on a project are version-controlled. The application is designed in a way, that only a local thin-client needs to run on the users computer. After entering the login credentials (see Figure 2.2), the client securely connects to a centralized service platform where the computation is made. The results and even the User Interface (UI) information is sent back to the client application. This has the benefit, that also weak computers can run the application.

Figure 2.3 shows the application itself with a loaded project and a simple geometric model.

The development team consists of a small core team and several student groups during the semester.

2.2 Baseline architecture

The current system design consists of multiple levels. It is a multi-process application based on the programming languages C++ and Rust. The source code is mainly aligned to be built on Microsoft Microsoft® Windows® (Windows). A port to Unix based systems is currently in



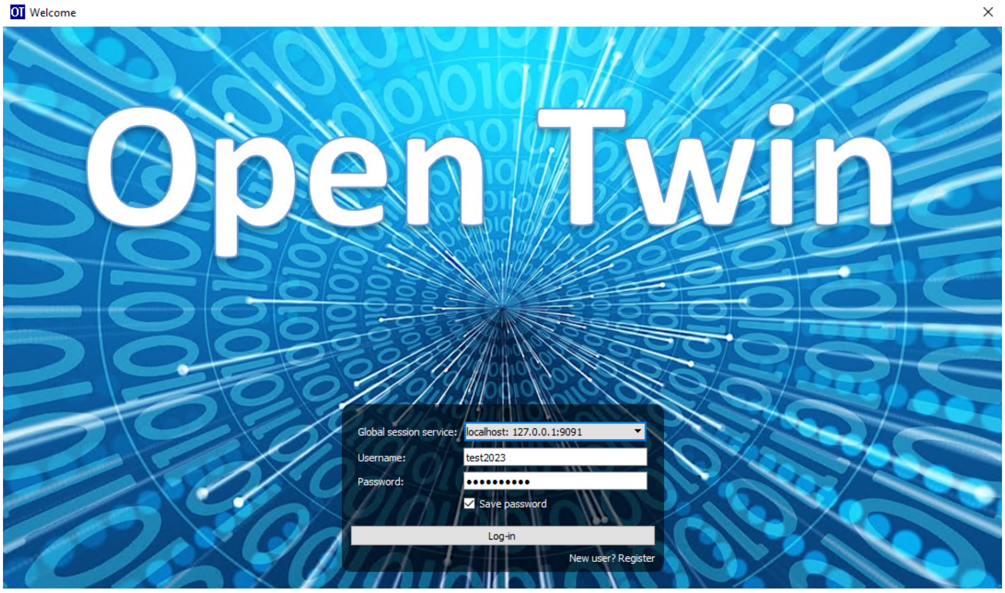
Manage Projects

Name: [Recent projects] (Press Return to show all projects)

Open	Copy	Rename	Delete	Export	Access	Project name	Owner	
▶						test2023	test2023	

1 projects found

Figure 2.1: The OT project overview.



Global session service: localhost: 127.0.0.1:9091

Username: test2023

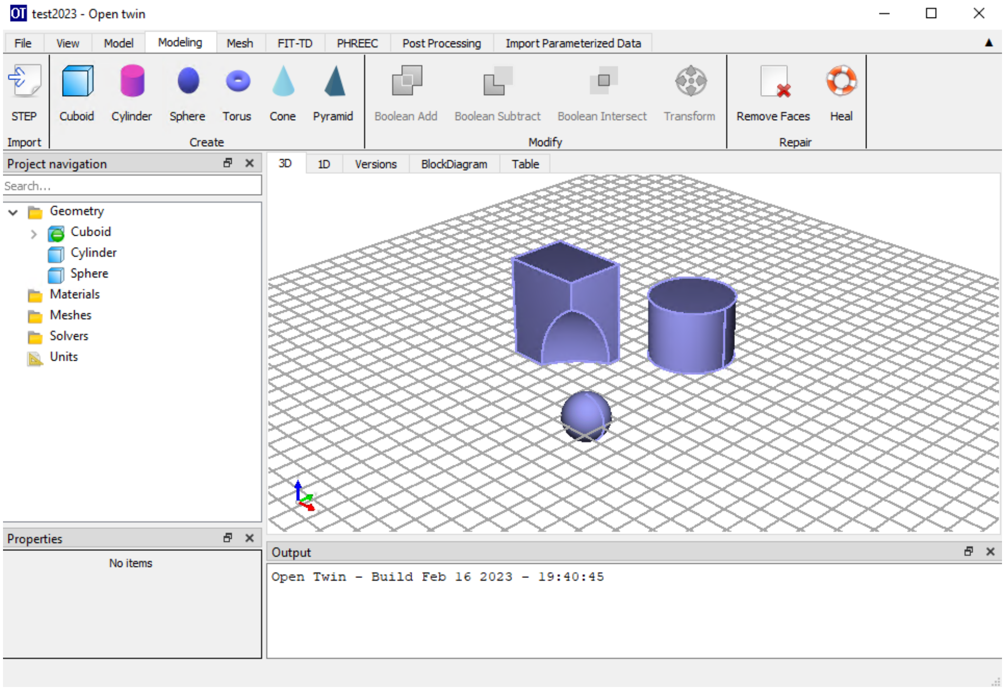
Password: ••••••••••••

☑ Save password

Figure 2.2: The OT login screen.

work. Therefore parts of the code base are aligned for multiple system architectures already, but the application is not yet able to be compiled for Linux.

Each microservice of the application is included dynamically and linked as a Dynamic Link Library (DLL) file. For starting the microservice environment, a central executable ("open_



```
Open Twin - Build Feb 16 2023 - 19:40:45
```

FIGURE 2.3: A opened project inside OT with a few created geometric models and subtracted computation.

twin.exe") is started with the corresponding arguments for the services (like the server's binding address, port numbers, and passwords) (see Listing 2.1) and the path to the DLL file itself. The UI front end, which is started by the user directly, is compiled in its own executable ("uiFrontend.exe").

```
open_twin.exe GlobalSessionService.dll \
  "0" "127.0.0.1:8091" "tls@127.0.0.1:27017" "127.0.0.1:8092"
```

LISTING 2.1: Command line of Open Twin Service start

For conveniently running the services with all their necessary arguments, batch files were provided that read environment variables and convert them into runtime arguments for the service executable. Therefore, if the services are started locally, the user runs a batch file that sets up the environment for the network binding details, path to certificates and encrypted database credentials.

The system consists of the following micro services that are permanently accessible: Global Session Service (GSS), Authorization Service (AUTH) and the database. The database is running on MongoDB[1]. Another Service is the Local Session Service (LSS) that spawns the so called compute services. Those are services for running the actual computation after opening

[1] MongoDB: https://www.mongodb.com/

a project that can dynamically spawn and exit. A partial list of compute services and their corresponding tasks can be found in Table 2.1. Each service runs in its own operating system process.

Name	Task
CartesianMeshService	If demanded, it converts a continuous geometry into a discrete Cartesian mesh.
FITTDService	If demanded, it runs a solver algorithm for transle electromagnetic simulation based on the finite integration technique (FIT).
KrigingService	If demanded, it runs a kriging interpolation of result data.
LoggerService	A background service, that accepts logging messages from other services.
ModelingService	Performs calculations for the creation, modeling and boolean combination of geometric data.
PHREECService	If demanded, it runs simulation based on PHREEC.
TetMeshService	If demanded, it meshes a form with an tetrahedral mesh.
VisualizationService	Runs the graphical calculations for displaying the geometric and data based results on the UI.

TABLE 2.1: List of compute services and their corresponding tasks.

As shown in Figure 2.4, the services can be separated by their network space. Not all services require a public available network address. While GSS, AUTH and database are globally accessible via a fixed network address, the LSS can theoretically run on a dedicated host and is only communicated to other parties after it has registered itself to the GSS. The services, spawned by LSS do not require a public address space either. All communication between the UI front end and the compute services is achieved via a relay service and a web socket communication channel.

The whole process of the LSS registration and connection of the UI front end to the compute services is depicted in Figure 2.5. Once started, the user can login. In order to connect to the database, the following steps are performed:

1. The UI front end requests further service information from the publicly available GSS. The address for this service is provided by the user. The GSS responds with Uniform Resource Locators (URLs) to the database and the AUTH.

2. The UI front end connects to the AUTH using the authentication information provided by the user.

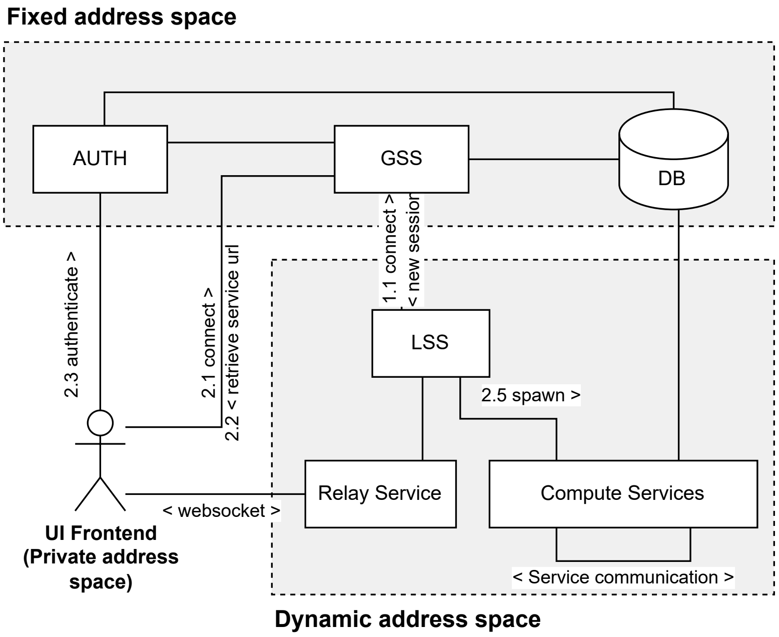


FIGURE 2.4: Communication overview and service organization for OT main services. In 1.1 the LSS registers at GSS. As soon as the UI front end connects to the GSS (2.1), service information is exchanged (2.2) and the user is authenticated (2.3). As a consequence, the GSS creates a new session and tells the LSS to spawn new compute services. From now on the UI front end communicates directly with the Compute services via the Relay Service over a web socket connection.

3. If the AUTH replies with a positive authentication, the UI front end connects to the database and lists the projects.

4. Once a project is opened or created, the UI front end requests a new session from the GSS. The GSS replies with the connection URLs of the LSS. The LSS has been registered to the GSS during its initialization.

5. The UI front end then connects to the LSS and requests a new session. As a result, the LSS spawns new application service processes and replies with the respective service URLs.

6. From now on, the UI front end communicates with the application services via the Relay service over a web socket.

The traffic between services is encrypted using mutual Transfer Layer Security (mTLS) technology. While regular Transfer Layer Security (TLS) ensures the authenticity of the server by using Certificates and the chain of trust, it does not verify the identity of the client. This is the benefit of mTLS. In mTLS, both sides, client and server has to verify their identity by providing a certificate inherited from a common root authority.

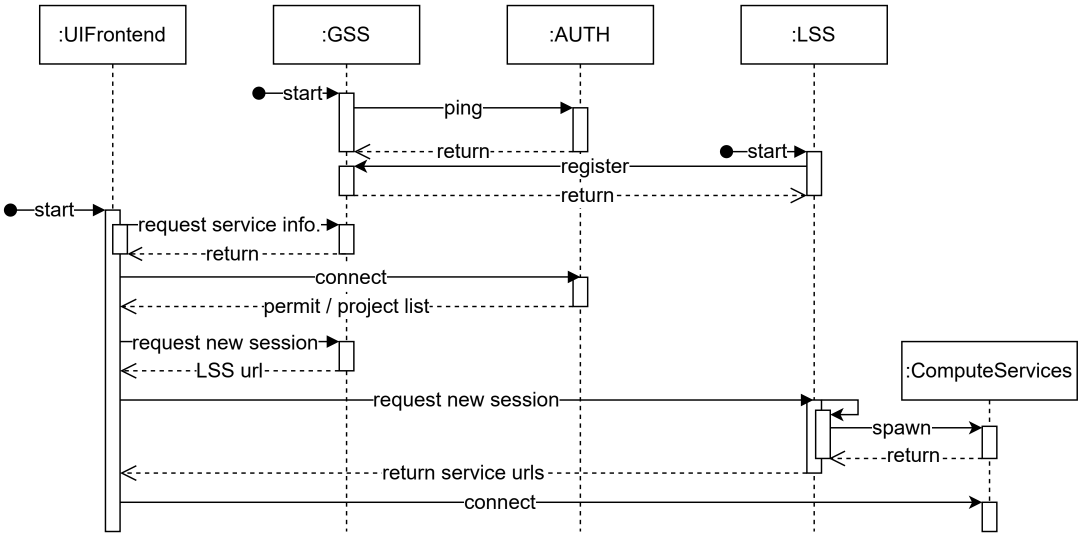


FIGURE 2.5: Service initialization of OT processes. In the beginning, the main services GSS, AUTH and an optional LSS are initialized. While the GSS checks the reachability of AUTH, the LSS registers itself at the GSS. After starting the UI front end, the service information is requested from a GSS and the user is authenticated. Afterwards, the project list for the authenticated user is displayed. After opening a project, the UI front end connects to the LSS and requests a new session. As consequence, the LSS spawns the compute services and connects them to the UI front end via a Relay Service. (Ping messages are omitted.)

2.3 Network traffic encryption

Each service of OT offers two different channels for secure communication between services. One channel supports traditional one-way TLS, while the other uses bidirectional mTLS. The one-way TLS channel is mostly used for checking the general health state of a service, while the two-way mTLS is used for relevant application based communication. In this section both encryption methods are briefly presented.

2.3.1 Transport Layer Security

TLS is an cryptography extension mainly designed for providing security over Hypertext transfer protocol (HTTP). The main goals of cryptography are confidentiality, integrity and authenticity between two communicating parties. This means, the communication on a network is kept secret between the two endpoints (Confidentiality), messages are not subject of manipulation (Integrity), and message exchanges are only allowed between authorized and trusted individuals (Authenticity). TLS ensures the three traits by using certificates.

A simplified handshake of the TLS protocol is depicted in Figure 2.6. After sending the certificate from server to the client, the client validates the certificate based on the chain of trust. This means, it checks if the certificate was issued by a trusted root authority (so called Certificate Authority (CA)). Only if the authenticity of the server is ensured, the encryption key is exchanged in order to start an encrypted communication.

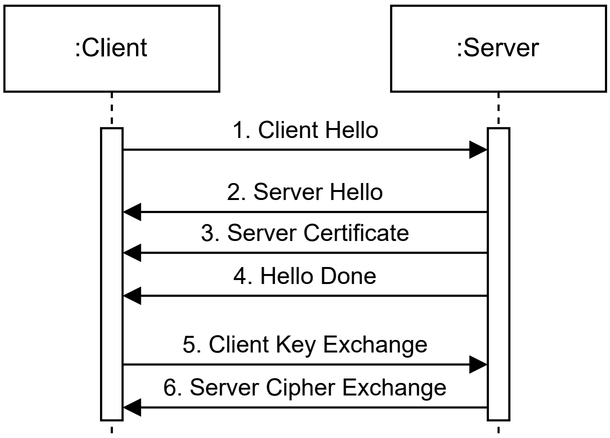


FIGURE 2.6: Simplified visualization of the TLS handshake[1]. After initialization of the handshake (1,2), the server sends its certificate (3) and finishes with a message "Hello Done" (4). The client then validates the certificate and compares it against the chain of trust. Afterwards, the key exchange starts (5,6) to ensure an encrypted communication.

2.3.2 Mutual authentication

The mutual authentication is adding another step to the one-way authentication. The application of it is used in mTLS. Instead of just sending the server's certificate to the client, the client also sends a certificate to the server. Therefore, not only the authenticity of the server is ensured, but also of the client.

As can be seen in Figure 2.7, compared to the TLS handshake, the mTLS handshake involves additional messages 5 and 6 for sending and validating the client's certificate. Unlike with the server certificate, the client certificate is not validated against a publicly available root authority[2]. Instead, the server acts as root authority, creates the client certificate and ships it with the application[2].

2.3.3 Certificate creation

For creation of certificates in the application landscape of OT, CloudFlare's public key infrastructure toolkit "cfssl"[2] is used. For generating certificates with the toolkit, it is fed with a JSON file with subject information for the Certiticate Request (CSR).

[2] cfssl: https://cfssl.org/ or https://github.com/cloudflare/cfssl

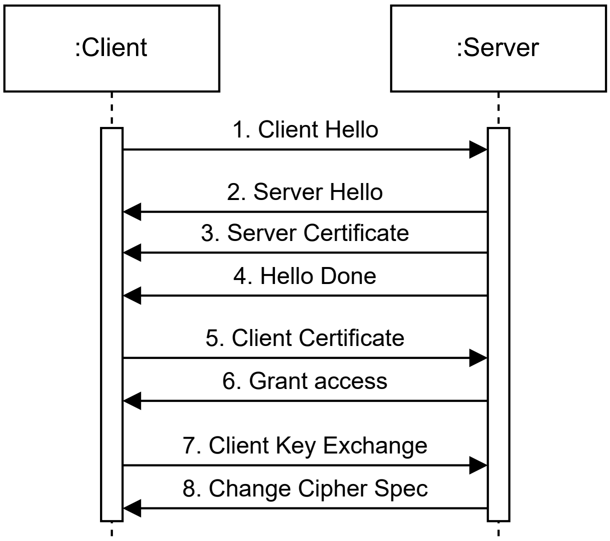


FIGURE 2.7: Simplified visualization of the mTLS handshake[2]. After initialization of the handshake (1,2), the server sends its certificate (3) and finishes with a message “Hello Done” (4). The client then validates the certificate and compares it against the locally stored root authority. Afterwards, the client sends its certificate to the server (5). The server validates the client certificate against its local root authority and grants access to the service (6). Afterwards, the key exchange starts (7,8) to ensure an encrypted communication.

It contains information about the issuer, as well as the name and cryptography algorithm of the certificate. Additionally, the client and server certificate that derive from the root CA contain information for whitelisted hosts in their CSR JSON file. If a host is not mentioned in the resulting certificate, requests from the corresponding host are rejected. Listing 2.2 shows such a configuration with accepted host names in the form of a CSR.

```
{
  "CN": "OpenTwin",
  "hosts": ["localhost", "127.0.0.1"],
  "key": {
    "algo": "rsa",
    "size": 4096},
  "names": [{
    "O": "Frankfurt University of Applied Sciences"
  }]
}
```

LISTING 2.2: Example of meta data in form of CSR configuration. “CN” is the certificate name. “hosts” describes the accepted hostnames, “key” describes information about the cryptography algorithm, “names” contains meta data of the organization

2.4 Problem statement

Even though, the application is clearly based on a microservice architecture and it is able to run on a distributed system, it is not designed for an automated cluster yet. It consists of multiple processes where many of them have to run on the same system and need a full working operating system as baseline. Containerization of the system has never been tested and needs to be introduced. First, the cluster engine needs to be set up for Windows compute nodes to allow Windows containers to run inside the cluster. Additionally, it needs to have full network capabilities as well as inter connectivity between the several services. Next, the application needs to run inside containers. Therefore, container images must be created and provided for the cluster engine. Additionally, the automatic extension of services requires communication between the cluster orchestration management and the applications running on the nodes. A feature that needs to be introduced later on.

Regarding logging, while the front end application does, the microservices currently do not produce log files. Instead, only a few sub processes write the information on its standard output stream. In some cases, the error information given by exceptions is dropped. Furthermore, proper exit codes in error cases are not returned. That is, if the application exits there is currently no way to detect if the process terminated normally or crashed as part of an error.

2.5 Limitations

Due to the limited amount of time, not all code changes are applied. On the one hand, this involves the adaption for automatic extension of services. On the other hand, it implies the changes required to make the application more fault-tolerant. The changes that would be necessary, would be too extensive. Therefore, they are only made to the main processes.

As a first case study, the application is not fully containerized. Since the network connectivity is known to cause troubles in Windows container networks, there is more investigation required later on. As part of this study, only the main services are containerized and the cluster is set up to investigate the behavior in cluster environments. The actual distribution of an full functional cluster network can be part of further studies later on.

2.6 Related Work

Chapter 3

System design

Various applications for realizing the architecture have been compared. In the following sections the different options that were taken into account are presented.

3.1 Orchestration engine

Orchestration engines aggregate the processes and tools that are used to distribute services across multiple machines. Further, multiple replications are provided to maintain reliability. In addition, some solutions offer load balancing of incoming requests and network interconnection. What all of these engines have in common is that a group of virtual machines or containers, known as "nodes", are managed from a central spot. An administrator directs what application is run on the cluster. Based on the application's metadata, the orchestration engine then decides where to run the application by selecting a node inside that cluster.

3.1.1 Hyper-V Replication

Microsoft Windows supports a replication mechanism for virtual machines hosted by Hyper-V. The existing virtual machines are mirrored to secondary virtual machine host servers which highers scalability and reliability. Therefore, by replicating to a secondary Hyper-V host server, enabling process continuity and recovery on outages is ensured. Although there are benefits, like scalability and recovery, Hyper-V is mainly designed for virtual machines. Therefore, the cluster management solution is not applicable for this use case.

3.1.2 Docker Swarm

“Docker Swarm” is a cluster and orchestration engine for the container service “Docker”. The offered extension mode has more features compared to the Hyper-V replication and is specialized for containers. For example, Load Balancing, increased fault tolerance and automatic service discovery. A highlighted feature among Docker Swarm is the decentralized design. That means, manager and application service can both run on any node within the cluster. Since it comes with Docker, no additional installation is required if Docker is already installed on the system. However, since it is bound to the Docker Application Programming Interface (API), using this orchestration technology involves the risk of inflexibility later on (“vendor lock-in”).

3.1.3 Kubernetes

Kubernetes (K8s) is a orchestration engine similar to “Docker Swarm”. Load balancing, autoscaling and automatic service discovery are also offered. However, K8s additionally comes with the ability to rollback to a previous version in a product life cycle and has built-in support for auto-scaling. However, K8s has more sophisticated configuration options which makes it harder to configure in the beginning.

The engine of choice was K8s because of its rich feature set. Also studies showed that K8s outperforms Docker Swarms when it comes to performance. For example, Marathe et. al. [6] compared a simple web server service deployed on a Docker Swarm cluster with a K8s cluster. The results showed better performance for K8s in terms of memory consumption and CPU usage. Another study of Kang et. al. [7] compared the performance of Docker Swarm and K8s in a limited computing environment on Raspberry Pi boards. They also concluded that K8s outperforms Docker Swarm if used with a high amount (=30) of service containers on 3 Pi boards [7]. Since they focused on container distribution and management methods this might get handy in the use case scenario under study.

3.2 Kubernetes

Since K8s is the chosen orchestration engine, the following sections are taking a deeper look inside its architecture.

3.2.1 Entities

There are many entities for objects inside the cluster. For description of those entities the configuration language YAML[1] is used. Some of the most widely used entities are described in the following paragraphs.

Pod A pod represents a set of running containers on a node. Each pod has additional information stored, such as Health state, the cluster internal network Internet Protocol (IP) address or the amount of replications.

Deployment Deployments are used to define declarative states for Pods. This allows to maintain consecutive versions of the pod and upgrade them during runtime.

Daemon set These ensure that multiple (or all) nodes run a certain pod[8]. Common use cases are tasks for all nodes or running the network overlay pod.

Configuration Map A configuration map stores non-confidential data as key-value pair[**?**]. Data stored in a configuration map can be mounted as volume, environment variable or command line argument to make applications more portable[**?**].

User This entity describes a user that can access the K8s cluster and API services. Users can be part of a group and permission roles.

Node A node represents a physical machine inside the cluster. Nodes can run multiple pods.

3.2.2 Services

K8s comes with a set of core services (see Figure 3.1) that ensure the life span of scheduled containers, and the compute services that offer the actual application.

In the following paragraphs, the crucial services are described in detail. Since every service is a pod, they can have multiple replicas. Only the core service have to run on a dedicated Linux

[1] YAML: https://yaml.org/

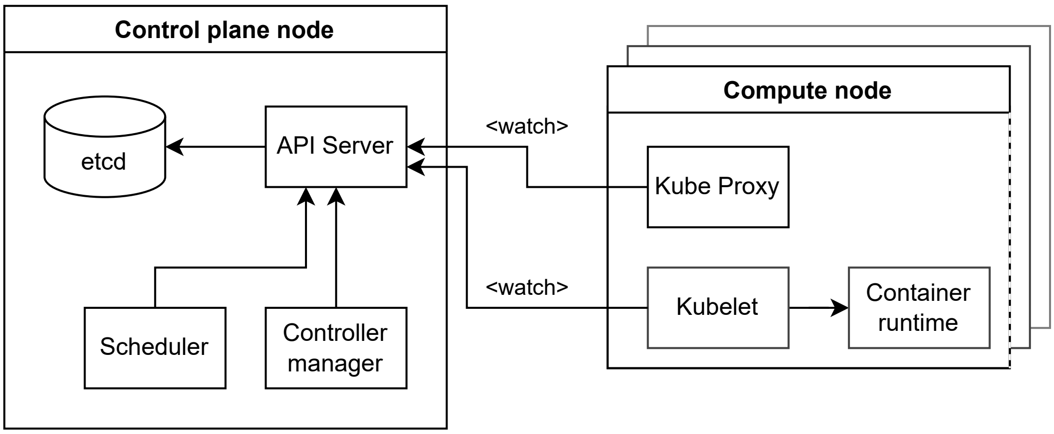


FIGURE 3.1: Core and compute services for Kubernetes[3]

node the so called “control plane node” or “master node”. The underlying operating system must be Linux, because Windows is not supported as control plane. The other services can run on nodes (with any operating system) for executing the applications and perform computations (“compute node” or “worker node”).

etcd The etcd[2] database server is a key-value store designed for distributed systems[3]. That means it could run with multiple replications and would still be able to keep a persistent storage synchronized across multiple instances. It contains the applied configuration of several cluster entities (e.g. User configurations, deployments, pod configurations).

API server This is a RESTful web server that serves the Kubernetes API via HTTP[9]. It is the central joint between the services and establishes communication between users, external components and other core services. It makes the objects stored in etcd accessible over an Open API specification[3, 10] and allows observing changes on the entities. The Command line interface (CLI) tools “kubectl” and “kubeadm” both interact with the API server.

Kubelet Kubelet is the service on the operating system level that maintains the pod life cycle and ensures the runtime of a container inside a pod. Furthermore, it manages the registration of the node to the control plane and reports its health and pod status to the API server.

Kube Proxy The Kube-Proxy runs as a separate pod on every compute node. It maintains the connectivity between the services and pods[3]. For a given IP address and port combination

[2] etcd: https://etcd.io/

it assures the connection to the corresponding pod. If multiple pods can offer a service, the proxy also acts as a load balancer[3].

Scheduler The scheduler is responsible for distributing services on the cluster and determining which node to choose during runtime. It reads conditions for scheduling (e.g. hardware resources, operating system, labels) from the API server and decides which node matches the configuration[3].

Controller manager While the API-Server is responsible for storing data in etcd and announcing changes to the clients, the Controller manager and its parts try to achieve a described target state[3]. The controller manager consists of several controllers for replications, daemon sets, deployments, volumes, and so on.

3.2.3 Pod life cycle

Similar to the underlying application container, Pods in K8s have a ephemeral lifetime[5]. After creation on the cluster, a unique identifier is assigned before a pod gets scheduled to an available node[5]. The pod keeps alive until its termination or deletion[5]. For distinguishing different kind of states of a pod life cycle, K8s defines the pod states as described in Table 3.1.

State	**Description**
Pending	The pod has been set up, the container and pod is currently initialized.
Running	The pod is bound to a node, the container is running.
Succeeded	The container terminated with a zero exit code.
Failed	The container terminated with a non-zero exit code or was terminated by the system.
Unknown	The pod state could not be obtained.

TABLE 3.1: List of K8s states during pod life cycle[5].

A terminated pod automatically gets restarted based on a configured restart policy. As the K8s documentation states, “the kubelet restarts them [the containers of a pod] with an exponential back-off delay (10s, 20s, 40s, ...), that is capped at five minutes”[5]. Furthermore, it is explained that the back-off time gets reset, once a container keeps running for 10 minutes[5].

3.2.4 Cluster networking

Cluster networking is achieved using two components: The network plugin and the Container Network Interface (CNI). Pods receive their own IP address and can communicate with other pods. However, this is not a functionality which is achieved by Kubernetes directly. By using a CNI the automated generation of network addresses and their inclusion is achieved when new containers are create or destroyed. It is crucial that pods share the same subnet across all the nodes in a cluster and Network Address Translation (NAT) is avoided[3].

Network plugins do implement the CNI. They usually come with a manifest for a daemon set that introduces a network agent on all nodes inside the cluster to support the network communication. For setting up the network interface, namespace and its IP address, a dedicated container image is used. This is called the "pause container" image.

Flannel Flannel[3] was originally developed as part from Fedora CoreOS[4][11]. It works with various backends for transferring packets in the internal network. Two possible backends are virtual extensible Local Area Network ("vxlan") and host gateway ("host-gw"). While "host-gw" needs an existing infrastructure and performs routing on the layer 3 network level, VXLAN is more flexible and could also be used in cloud environments[12]. VXLAN is an overlay protocol and encapsulates layer 2 Ethernet frames within datagrams[11]. It is similar to regular VLAN, but offers more than 4,096 network identifiers[11]. Thus, VXLAN is a good choice for highly scalable systems.

Even though, the team behind K8s do not recommend any specific network plugin, there are only a few common network plugins widely used. The amount of available CNI plugins is even more reduced if the support for Windows nodes is taken into account.

Calico Compared to Flannel, Calico[5] is stated to be more performative, flexible and powerful[11, 13]. Calico comes with a sophisticated access control system[13] and more configuration options. However, its advanced configuration makes it hard to maintain long-term.

[3] Flannel: https://github.com/flannel-io/flannel

[4] CoreOS: https://getfedora.org/en/coreos

[5] Tigera's Calico: https://www.tigera.io/project-calico/

For this use case, Flannel is used as network plugin, since it is the described plugin used in the documentations for setting up K8s with Windows containers[14, 15]. Hence, support for this CNI plugin in relation with Windows containers is assumed to be larger than with Calico.

3.3 Container environment

The ecosystem around containerization defines terminology that needs to be looked at before going into details for K8s. First of all, the Container Runtime Interface (CRI) defines the interface between K8s and container runtime. Most of the container runtimes follow the design principles defined by the Open Container Initiative (OCI)[6] for describing images and containers. The actual container runtime runs the isolation layer between the physical host machine and the K8s cluster by using containerization of processes. This is what can be selected when working with K8s.

While K8s used to support Docker as their standard container runtime, they announced it to be deprecated in 2020, and finally removed the support in February 2022[16, 17]. The teams behind K8s decided to drop the hard coded support for Docker and offer ContainerD instead. However, the specification for ContainerD's "Containerfile" has only minor differences compared to Docker's "Dockerfile". Thus, ContainerD files are fully compatible to docker files.

Some of the container runtimes offered by K8s are not available for Windows hosts. For example, Linux containers (LXC)[7] use process groups, control groups (cgroups) and name spaces on the operating system level. The CRI from the Open Container Initiative (CRI-O)[8] is another alternative offered for K8s on Linux systems. Since those are not available in Windows, they are not further considered.

At the current time being, container networking with ContainerD is not well-established on Windows [18–21] even though the docker runtime is already removed in current versions of K8s [16]. However, these are the only two working container backends for Windows containers. Therefore ContainerD as container backend was chosen.

[6] OCI: https://opencontainers.org/

[7] LXC: https://linuxcontainers.org/

[8] CRI-O: https://cri-o.io/

3.3.1 ContainerD

ContainerD is a native version of a container runtime. Newer versions of Docker on Linux, are running ContainerD under the hood for process isolation. On Windows, ContainerD uses slim host process isolation. The process isolation with ContainerD consists of multiple abstraction layers (shown in Figure 3.2). Its back end contacts the containerd-shim which is maintaining an abstraction layer for communication for the underlying layers (depending on Linux and Windows). Below that, Windows offers a custom fork of the CLI *runc*, so called *runhcs*[4]. Using *runc*, new containers can be created by running a simple command[4]. The layer for *runhcs* connects to the Host Compute Service (HCS) which is another abstraction layer of Windows for providing a stable API to the low level functionality of the operating system[22]. ContainerD does not come with any mechanisms for networking. Instead, this is in responsibility of the HCS.

The developers of K8s marked the Docker CRI as deprecated in version v1.20[16]. Since version v1.23 of K8s, Docker was fully removed which lead to ContainerD being the only available CRI for Windows containers.

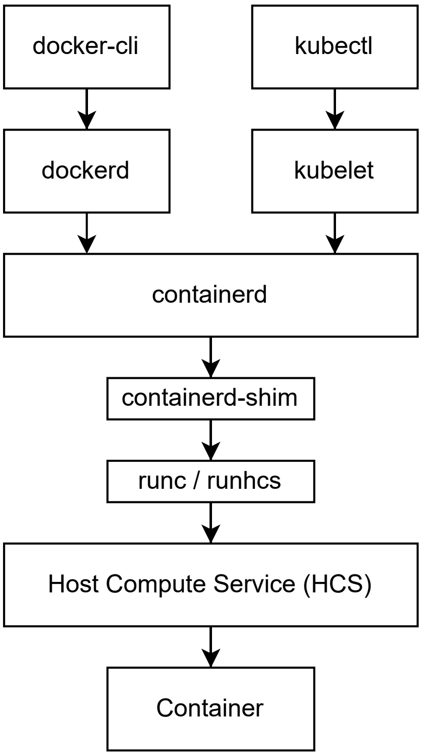


FIGURE 3.2: Abstraction layers for ContainerD on Windows. The image shows the technology stack from the Docker and K8s command line to the Container layer.[4]

3.3.2 Docker Container Runtime Interface

The Docker CRI (so called "Docker shim") is using the internal mechanisms from Docker to run containers. Older versions in Linux were using control group isolation.

For Windows, there are two different modes available. The first option is using the process isolation mode offered by ContainerD. It can be enabled by switching to "Windows Containers". It is the default mode for Windows Server systems. However, older versions of Docker and Docker on Windows 10 and above have the opposite behavior[23]. On client versions of Windows, a dedicated hypervisor-isolated virtualization is the default option[23]. This means, during the installation of Docker on Windows, the underlying Windows container host creates a separate Hyper-V Virtual Machine (VM) to run container images. However, this is not a regular Hyper-V VM. Instead, it is a purpose-built VM, often referred to as utility VM or UVM that can't be managed directly and is fully controlled by the Windows container runtime[23]. For networking, the internal mechanisms from Docker are used. All running containers are deployed inside the dedicated Hyper-V VM. Therefore, this is a mixture of process isolation and full isolation using virtualization.

However, the hypervisor approach still has the disadvantage of using large resources for containers, even though they are running in one virtual machine. In addition, containers running in hypervisor isolation take longer to start up than those running in process isolation[23].

3.4 Container Image

The container image consists of a base part and a part for custom configuration. Both are further explained in the following sections.

Container images are described, using the Containerfile[9] format. There are container images for each process of the system architecture, each of them having their own Containerfile definition with different command line arguments and environment variables.

[9] Containerfile: https://www.mankier.com/5/Containerfile

3.4.1 Base image

The container images to use for running the OpenTwin processes need to run a Windows base image. Beside the full Windows images, Microsoft® offers the more common images "Windows Server Core" and "Windows Nanoserver"[24]. They significantly differ in the download size, their on-disk footprint and the features supported[24]. As Microsoft states, "Nanoserver was built to provide just enough API surface to run apps that have a dependency on .NET core or other modern open source frameworks. PowerShell, Windows Management Instrumentation, and the Windows servicing stack are absent from the Nanoserver image"[24].

The design of containerization of OT envisages the usage of "Server Core" as base image. Even though the "Server Core" image is not the smallest base image, it provides full functionality for the required technologies for the current use case.

3.4.2 Custom image

On top of the base image, customizations and the actual application are applied. The binary files are included in the container image and added during build. The common CRIs only forward the output of processes with process id 1 to the host machine. Furthermore, this is also the only process the CRI is waiting for, to keep the container alive. Thus, instead of using the provided batch files to start the application services, the OpenTwin process is called directly with the appropriate command line arguments as command for the container. Therefore, the environment variables needs to be set up as part of the container file. The root certificate (certificate authority) is passed as file mount into the container later on.

3.5 Target architecture

The application needs to be distributed on multiple systems. Kubernetes supports application rollout only as container images. To be able to distribute the services on a cluster management tool a containerization of the application is necessary.

Chapter 4

Implementation

4.1 Containerization of Services

Container images are provided for running the application inside the CRI. Definition of the container manifest is done in the "Containerfile"[1] format.

4.1.1 Code changes on core application

For containerizing the OT application, several changes were applied to its code base. This improves error handling and error tracing of the application and therefore simplifies development of the cluster. In the following sections, the several code changes are described in detail.

Introduction of exit codes

The microservice DLL files have return codes for different error cases. This is, for instance a code 200 in the AUTH for connection issues to the database. As with process exit codes, a zero return code indicates a successful termination. Even though the main executable "open_twin.exe" retrieves the return code, it did not convert these codes into proper process exit codes. Exit-Codes are a crucial part of the K8s's life cycle management as described in subsection 3.2.3. Therefore, the return codes need to be converted into process exit codes.

The surrounding lines of code in the main executable are shown in Listing 4.1.

[1] Containerfile: https://www.mankier.com/5/Containerfile

```
let _result = initialize(siteid_c_str.as_ptr(), service_c_str.as_ptr(), db_c_str.as_ptr(),
     dir_c_str.as_ptr());
```

Listing 4.1: Former code snippet from main executable for calling the microservice library code in Rust (*/Microservices/OpenTwin/src/main.rs*)

A new condition for non-zero exit codes was added as shown in Listing 4.2. The variable "`_result`" contains the returned exit code of the library DLL, whereas "`lib_path`" contains the path to the library DLL. As can be seen, it is used to describe the affected service name in a error message (line 87).

```
86  if _result != 0 {
87    eprintln!("Library/Service initialization ({}:init()) failed with exit code {}",
        lib_path, _result);
88    std::process::exit(_result);
89  }
```

Listing 4.2: Code changes in Rust main executable for additional treatment of exit codes (*/Microservices/OpenTwin/src/main.rs*)

Debug verbosity in launcher

By default Rust programs show a window for console output no matter if it was built in release mode or with debug parameters. However, the main executable uses conditional compilation to set configuration attributes about the windows subsystem. Rust provides a compilation option "`debug_assertions`" that is set to "true" for compilations without code optimization[25]. Therefore it is set to "true" if the application runs in debug mode. As shown in Listing 4.3 with conditional compilation, this build option is checked and console output is only shown for debug builds.

```
#![cfg_attr(not(debug_assertions), windows_subsystem = "windows")]
```

Listing 4.3: Conditional compilation for disabling console output in non-debug builds (*/Microservices/OpenTwin/src/main.rs*)

Even if output is shown on the console window for debug builds only, the error messages are not able to be read conveniently in cases where the application crashes. This aggravates debugging of regular application errors outside of a containerized environment. To avoid this behavior, the launcher batch files were adapted. For this, a new parameter was invented to the batch file for running the batch file in verbose mode.

```
18 IF "%~1"=="/V" (
19   REM OT is opening console windows in debug build, so we want to pause them at the end
20   SET pause_prefix=cmd.exe /S /C "
21   SET pause_suffix=" ^& pause
22   ECHO ON
23 )
```

Listing 4.4: Additional command argument for preventing close of window after termination (*/Microservices/Launcher/OpenTwin_session.bat*)

As first step, a new command line argument has been introduced. If "/V" is appended to the start of the launcher batch file it will run with higher verbosity. As can be seen in Listing 4.4, with "/V" appended, two new variables "`pause_prefix`" and "`pause_suffix`" are set (line 20-21). Furthermore, command output is enabled to debug the launcher file itself (line 22).

```
START "AUTHORIZATION SERVICE" %pause_prefix%open_twin.exe AuthorisationService.dll
 "%OPEN_TWIN_SERVICES_ADDRESS%:%OPEN_TWIN_AUTH_PORT%" "%OPEN_TWIN_MONGODB_ADDRESS%"
 "%OPEN_TWIN_MONGODB_PWD%"%pause_suffix%
```

Listing 4.5: Additional command extension for preventing close of window after termination (*/Microservices/Launcher/OpenTwin_session.bat*)

As can be seen in Listing 4.5, the newly defined variables "`pause_prefix`" and "`pause_suffix`" are appended to the command of starting a service. This ensures that a service process is started in a new window and the command is followed by the Windows "pause" command to stop and wait for user interaction.

Enhanced logging and error tracing

For improving the error tracing, the overall log amount has been increased. This involves enabling the logging of the central logging functions inside the library "OpenTwinCommunication". The logging mechanism did not use logging to standard output. Instead, calling the respective functions for logs were just ignored. This was changed and logging to standard output has been introduced. Additionally, more calls to the logging mechanism, including caller information, have been added. For instance, the AUTH now shows error messages for caught exceptions and errors during database initialization. Also, the general error tracing has been improved. In the main services GSS, LSS and AUTH, exception handling has been added, where it was not present before. Furthermore, the formatting of exception messages was improved and more information has been added.

For the current time being, there is an ongoing work to replace the logging to standard output by forwarding the log lines to a central logging library.

Certificate changes

Since OT is using mTLS for communication between services, the CSR file needs to contain proper host specifications. There is a script for the substitution of placeholders in a template CSR file. However, this process does not work if performed for a containerized application. If the replacement takes place during the creation of the container, the final host name or IP address does not yet exist. If, on the other hand, the automatic replacement is done later, it cannot be carried out from outside the container, as the script automatically replaces only the local host name.

Listening on all interfaces

Container images have a certain network image for communicating to external hosts. Processes inside the container have to bind to this network interface. However, the IP address of this network interface is unpredictable during compile time. The server processes that run inside a container therefore have to bind to all available network interfaces to be accessible to the outer network. Furthermore, the services exchange service information with other services as described in section 2.2. However, the binding address must differ from this published address in a containerized environment, since the binding address is mostly not accessible from the public.

Instead of binding to all network interfaces, the main executable in OT was only able to bind to a given IP address (based on the published service address) only. Also, it was not possible to configure a different binding address than the one published to other services.

Listing 4.6 shows the affected lines of code in the main executable. The passed argument for the service URL is forwarded to the server listener class. Afterwards, the address is printed on the console. The service URL, here passed as variable, consists of the service address and the port of the service.

```
let listener = net::TcpListener::bind(&service_url).await?;
println!("Starting server at {:?}", service_url);
```

Listing 4.6: Listener binding before the applied changes (*Microservices/OpenTwin/src/main.rs*)

To not affect the outer interface, the changes work with the data already provided. This means, the passed arguments to the main executable do not require a change.

```
150 let service_port = Url::parse(&format!("https://{}", service_url))
151 .expect(&format!("Invalid service url. Unable to parse service url: {}", service_url))
152 .port();
153 if service_port.is_none() {
154   panic!("Invalid service url. Service url is lacking port defintion: {}", service_url);
155 }
156 let binding_address = format!("0.0.0.0:{}", service_port.unwrap().to_string());
157 let listener = net::TcpListener::bind(&binding_address).await?;
158 println!("Server listening on {:?} (publishing {:?})", binding_address, service_url);
```

Listing 4.7: Listener binding after the applied changes. The service url is parsed, based on its port. Binding is done on all interfaces. (*Microservices/OpenTwin/src/main.rs*)

The binding address is separated from the published address, since the address in the argument still gets passed to the microservice DLL. Afterwards, the service URL is processed as shown in Listing 4.7. First, the given service URL is parsed and the port number is extracted (lines 150-152). If no port is found or the parsing failed, the application fails and shows an error (lines 153-155). Afterwards, the port number is concatenated with the binding address “0.0.0.0” and therefore the server binds to all addresses (lines 156-157). The last line shows the new output containing the port number and the address published to other services.

4.1.2 Container definition

There are currently three container images prepared for containerization of OT. These cover the main services for GSS, LSS and AUTH. The structure is the same for each of the container files. The first part of the container file is shown in Listing 4.8.

```
ARG BASE_IMAGE_TAG=ltsc2022
FROM mcr.microsoft.com/windows/servercore:$BASE_IMAGE_TAG
```

Listing 4.8: Containerfile for the GSS. Description of the base image and variable tagging using a build argument. (*Distribution/Container/globalsession.Containerfile*)

The first line introduces a build argument for defining the target image tag from the command line without altering the container file. It is used afterwards to pass the image tag to the base container image. The subsequent lines define labels for the resulting image.

The container files differ in the runtime specification. They run the same entry point, but have a different process running as command. Besides, the exposed port depends on the service inside the container. After defining the command and copying the files into the container image, the certificates are built as part of the image file system. The discussed section of the container file can be seen in Listing 4.9.

```
ENTRYPOINT ["cmd", "/C"]
CMD ["open_twin.exe", "GlobalSessionService.dll", "0", \
"%OPEN_TWIN_GSS_SERVICE_ADDRESS%:%OPEN_TWIN_GSS_PORT%", \
"%OPEN_TWIN_MONGODB_ADDRESS%", \
"%OPEN_TWIN_AUTH_SERVICE_ADDRESS%:%OPEN_TWIN_AUTH_PORT%"]
EXPOSE 8091
WORKDIR C:/app/Deployment/Certificates
COPY ./ ../
RUN createCertificate.bat && certutil -addstore root %OPEN_TWIN_CA_CERT%
WORKDIR C:/app/Deployment
```

Listing 4.9: Containerfile for the GSS. Description of the command line and certificate creation. (*Distribution/Container/globalsession.Containerfile*)

4.2 Cluster Setup

The following section describes the setup of different nodes in the cluster. While the master node refers to the K8s Control-Plane node which is responsible for distribution of the workers, the worker nodes are the actual machines that are executing the applications. The used version of K8s is “v1.25.3”. The version of ContainerD is “v1.6.8”, on both sides, the worker nodes and the master node.

4.2.1 Setting up the master node

For setting up the master node on Linux, a system based on Debian Bullseye 11.5 is used. Basis for the performed steps is the K8s documentation for setting up a container runtime[?] and the online tutorial of Basappa[?].

4.2.1.1 Installing prerequisites and ContainerD

After installing and setting up the operating system, the swap mechanism needs to be permanently turned off. This is done by editing the file system table (“`fstab`”), in file “`/etc/fstab`” respectively, by commenting out the swap partitions and masking the systemd swap units as shown in Listing 4.10.

```
$ sed -i '/ swap / s/^\(.*\)$/#\1/g' /etc/fstab
$ systemctl mask dev-sda3.swap
```

Listing 4.10: Bash commands for disabling Swap mechanism and masking.

The container runtime requires certain kernel features to be enabled. On Debian 11, the required kernel modules for virtual networking facilities (“`br_netfilter`”) and overlay filesystems (“`overlay`”) are not enabled by default. Thus, they have to be manually enabled. To permanently enable them and keep them enabled across reboots, the setting is stored in a file in the directory for loaded modules (*/etc/modules-load.d/k8s.conf*).

For functional networking on the master node, K8s also requires internal network packets being forwarded to the pods. Additionally, the tracking table for tracing network packets and their assigned connection is increased. This is a prerequisite for the applied NAT which is used in container networking. New files with respective settings is created in the directory for kernel settings as shown in Listing 4.11.

```
$ cat << EOF | tee /etc/sysctl.d/k8s.conf
net.bridge.bridge-nf-call-iptables  = 1
net.bridge.bridge-nf-call-ip6tables = 1
net.ipv4.ip_forward                 = 1
net.netfilter.nf_conntrack_max      = 524288
EOF
$ echo 1 > /proc/sys/net/ipv4/ip_forward
```

Listing 4.11: Bash commands for enabling IP forwarding and network filtering.

After installing ContainerD with all its prerequisites from the package registry, a configuration file (“`config.toml`”) is created. Subsequently, the SystemD cgroup is added to the runtime options of ContainerD and its service is restarted. For this, the commands from Listing 4.12 are applied.

```
$ containerd config default |
  sed 's/\[plugins.\"io.containerd.grpc.v1.cri\".containerd.runtimes.runc.options\]/&\n
    SystemdCgroup = true/' | tee /etc/containerd/config.toml >/dev/null
$ service containerd restart
```

LISTING 4.12: Bash command for setting up containerd config

To finally start the cluster, the packages for the Kubelet service, “kubeadm” and “kubectl” need to be installed. Then, the cluster can be initialized by running the command line tool as shown in Listing 4.13. Since the configuration parameters can also be passed as YAML file, this is the preferred method of passing arguments and allows version control of the required configuration.

```
$ kubeadm init --config kubeadm_config.yaml
```

LISTING 4.13: Bash command for setting up the cluster

Since Flannel is used as network overlay, the IP address range provided for pods must be “10.244.0.0/16”[14]. Additionally, a value is provided that sets the CRI socket address for newly registered nodes to the one of ContainerD. Furthermore, the cgroup driver for the Kubelet service on worker nodes is set to SystemD. Thus, the cgroup driver on Linux worker nodes should be aligned to the master node. The information passed as configuration is presented in Listing 4.14.

```
kind: ClusterConfiguration
apiVersion: kubeadm.k8s.io/v1beta3
kubernetesVersion: v1.25.3
networking:
  podSubnet: "10.244.0.0/16" # pod-network-cidr
---
kind: InitConfiguration
apiVersion: kubeadm.k8s.io/v1beta3
nodeRegistration:
  criSocket: unix:///var/run/containerd/containerd.sock
---
kind: KubeletConfiguration
apiVersion: kubelet.config.k8s.io/v1beta1
cgroupDriver: systemd
```

LISTING 4.14: YAML configuration for providing configuration for the cluster and newly registered ndoes. (*Distribution/Controlplane/kubeadm_config.yaml*)

4.2.1.2 Applying a Container Network Interface

After successfully initializing the cluster, the overlay network for Flannel must be set up. For this, the respective entity description can be directly downloaded from the vendor[2]. Since networking has to blend with Flannel on Windows, the Virtual Network Identifier (VNI) (4096) and port (4789) for Flannel on Windows must be set as part of the configuration. For this, manual editing of the description must be performed. The respective values are added to the configuration map section where the network configuration file is described. This file is automatically created on newly registered nodes. The changed entity description is applied on the cluster respectively.

4.2.1.3 Adding the proxy daemon sets

The usage of Windows worker nodes in a K8s cluster requires additional configuration mappings and daemon sets to be applied on the cluster. Those are used for setting up a proxy for Flannel. The Special Interest Group (SIG) “Windows Tools” of K8s provides the additional objects[14]. Before applying them a replacement of the K8s version and the Flannel version is performed. This procedure is achieved by running the two commands shown in Listing 4.15.

```
$ curl -L https://raw.githubusercontent.com/kubernetes-sigs/sig-windows-tools/master \
  /hostprocess/flannel/kube-proxy/kube-proxy.yml |
  sed 's/KUBE_PROXY_VERSION/v1.25.3/g' |
  kubectl apply -f -
$ curl -L https://raw.githubusercontent.com/kubernetes-sigs/sig-windows-tools/master \
  /hostprocess/flannel/flanneld/flannel-overlay.yml |
  sed 's/FLANNEL_VERSION/v0.17.0/g' |
  kubectl apply -f -
```

LISTING 4.15: Bash command for adding the flannel overlay configuration[14]

4.2.2 Setting up the worker node

The initialization of the worker node is mainly oriented on the guide for adding Windows nodes that is provided from the SIG “Windows Tools” on GitHub[14]. Although there is an original guide from K8s[15], it was not used as basis because it is outdated.

[2] Flannel entity description: https://raw.githubusercontent.com/flannel-io/flannel/master/Documentation/kube-flannel.yml

First, the node preparation scripts of the K8s SIG “Windows Tools” are retrieved and executed as shown in Listing 4.16. The first script retrieved (*Install-Containerd.ps1*) installs ContainerD and its prerequisites on Windows. Installation of the prerequisites involves the Windows features “Hyper-V”, “Hyper-V Tools”, “Hyper-V for PowerShell” and “Containers”. Once the features are installed and the system has successfully rebooted, the script proceeds by downloading the binary of ContainerD and adding its location to the “PATH” environment variable. ContainerD’s configuration file is changed to align the CNI file locations with K8s. The script ends by registering ContainerD as a service.

The second script (*PrepareNode.ps1*) prepares the node for joining the cluster. It first downloads the Kubelet executable and creates a script file that is called by the Kubelet service. The created script file contains runs the Kubelet in consideration of version dependent command arguments. Afterwards, the Non-Sucking Service Manager (NSSM)[3] is downloaded and the Kubelet script is registered as a service. Finally, the *PrepareNode* script also sets up firewall rules for Kubelet.

Furthermore the command line tools “kubectl”, “kubeadm”, “crictl” and “nerdctl” are installed. While installing “crictl”, it is added to the “PATH” environment variable.

```
> curl.exe -LO https://raw.githubusercontent.com/kubernetes-sigs/sig-windows-tools/master/
    kubeadm/scripts/Install-Containerd.ps1
> .\Install-Containerd.ps1
> curl.exe -LO https://raw.githubusercontent.com/kubernetes-sigs/sig-windows-tools/master/
    kubeadm/scripts/PrepareNode.ps1
> .\PrepareNode.ps1 -KubernetesVersion v1.25.3
```

Listing 4.16: Retrieval of node preparation scripts[14]

After successfully running the preparation script the node should be ready to join the cluster. For joining the cluster a token is generated on the master node. The token is copied to the preapared node and used as part of a “join” command. Listing 4.17 shows the command for joining the cluster, whereas the token is substituted by “TOKEN”.

```
> kubeadm.exe join 1.2.3.4:6443 --token TOKEN --discovery-token-ca-cert-hash sha256:HASH \
  --cri-socket "npipe:////./pipe/containerd-containerd"
```

Listing 4.17: Command for joining new nodes on Windows

[3] NSSM: https://nssm.cc/

Besides containing the token and the master node's IP address, the command also has the hash value of the CA of the master node. For K8s versions 1.25 and below the command also requires the explicit declaration of ContainerD as container runtime. Thus, the additional option "`cri-socket`" is defined.

The successful join of the node is verified by running "`kubectl get nodes`" on the master node. This prints the newly added node.

4.2.3 Deployment of the application

As a first draft, only the LSS of the application is deployed. Deployment of the application is done by applying the manifest as shown in Listing 4.18. The manifest defines a pod using the container image of the LSS (lines 2-4). Since the images are not yet provided via an image registry, the images are stored locally. Thus, the K8s is directed to never pull the image and retrieve it locally instead. The passed environment variables (lines 5-13) are defined by the section "`env`" section, whereas the variable "`OPEN_TWIN_LSS_SERVICE_ADDRESS`" is set with the IP address of the pod. The external port is defined as 8093, which is equivalent to the port of the LSS. The node selector property (lines 16-17) defines a requirement for the node scheduler to select only Windows nodes. The manifest is applied on the cluster using "`kubectl apply`".

```
 1  containers:
 2  - name: opentwin-session
 3    image: local.dev/opentwin-session:latest
 4    imagePullPolicy: Never
 5    env:
 6      - name: OPEN_TWIN_LSS_SERVICE_ADDRESS
 7        valueFrom:
 8          fieldRef:
 9            fieldPath: status.podIP
10      - name: OPEN_TWIN_MONGODB_ADDRESS
11        value: 1.2.3.4:27017
12      - name: OPEN_TWIN_GSS_SERVICE_ADDRESS
13        value: 5.6.7.8
14    ports:
15    - containerPort: 8093
16  nodeSelector:
17    kubernetes.io/os: windows
```

LISTING 4.18: Partial section of configuration for cluster deployment of the LSS (*Distribution/Kubernetes/open_twin.yaml*)

4.3 Automatic setup

Even though the scripts provided from the SIG "Windows Tools" cover most of the preparation steps, some manual steps are still required as described in the respective guide[14]. To automate the manual steps where possible, and to also provide tools for debugging error cases, further steps are required. Thus, a custom script (*Distribution/Container/setup-node.ps1*) was created to provide this aid.

The script automates the process of rebooting and rerunning the preparation after the required Windows features are installed, and provides instructions for steps that cannot be automated. It also checks the prerequisites before performing any action. This means, it is checked whether the operating system is "Windows Server" and if a certain minimum version of Windows is used. The check for prerequisites is shown in Listing 4.19.

```
if (-not ((Get-ComputerInfo).WindowsProductName | Select-String "Server")) {
  throw "Prerequisites not met. Windows Server is required as operating system."
}
if (-not (Get-Hotfix -ErrorAction Ignore KB4489899)) {
  if (-not ([System.Environment]::OSVersion.Version -gt [System.Version]"10.0.17763.0")) {
    throw "Prerequisites not met. You either need KB4489899 installed, or a Windows
    Version higher than 10.0.17763"
  }
}
```

Listing 4.19: Powershell commands in the automated setup script. Checks for prerequisites.

Besides checking for system requirements, the custom script also installs tools for debugging. This includes the Windows version of "kubeadm" and "kubectl", as well as the command line tools for debugging containers on Windows, namely "crictl" and "nerdctl". Furthermore, the custom script also excludes the ContainerD process from the Windows Defender firewall as shown in Listing 4.20. This increases the pull and general runtime performance of containers. The installation of the command line tools also involves setting up PowerShell auto completion and include to the "PATH" environment variable.

```
Add-MpPreference -ExclusionProcess "$Env:ProgramFiles\containerd\containerd.exe"
```

Listing 4.20: Powershell command in the automated setup script. Exclusion of all actions performed by containerd.exe from Windows Defender.

Chapter 5

Results

The development of the K8s cluster and the containerization of OT hides several pitfalls. The results are discussed in this chapter.

5.1 Containerization

Issues occurred during the containerization of the application. This involves pitfalls with the development of the container images and the special cases with the new Windows host-process container technology. The required considerations are discussed in this section.

5.1.1 Container manifest

Even though the format of the container manifests ”Containerfile” is compatible to the proprietary ”Dockerfile” format from Docker, the CRIs do not follow the specification everywhere[26] This was an issue while writing a Containerfile for the ContainerD CRI. Especially in cases where line breaks in the Containerfile were necessary to shorten long lines and increase readability. ContainerD is treating line breaks paths in string notation different compared to paths in JSON array notation and is not following the specification[27].

```
ENTRYPOINT open_twin.exe \
Service.dll
```

Listing 5.1: Containerfile entrypoint specification across multiple lines in text format.

Listing 5.1 shows an example of the problem. While the entry point in Docker is interpreted as *open_twin.exe Service.dll*, the interpreter in ContainerD only reads the first line as entry point and ignores the line break character ”äs in *open_twin.exe*. This causes troubles during build and execution of a container image. The container image is built in Docker, whereas it does not run in ContainerD. Since, ContainerD on Windows cannot build container images, it does also not work the other way around. To overcome this flaw, the *ENTRYPOINT* definition has to be written in JSON notation. If defined as in Listing 5.2 the interpreter

```
ENTRYPOINT ["open_twin.exe",
"Service.dll"]
```

Listing 5.2: Containerfile entrypoint specification across multiple lines in JSON format.

5.1.2 Windows Base Image

While it is normal to have a requirement for the same operating system kernel when running container images, it is a uncommon requirement to have the exact same build version.

- OpenTwin needs to be compiled on the system it is running. Developers need to fix this issue.

X Containerfile support differs from Dockerfile: Path of ENTRYPOINT is not treated the same way.

- Container namespace: If imported manually, image has to be shifted from default ns to k8s.io ns - Container cannot be built on Windows in ContainerD /nerdctl - need to be build on docker first and then transferred

5.1.3 Local images namespace

If images are locally imported, they have to be in the ContainerD namespace ”k8s.io” to get recognized by K8s. However, if images are built with Docker, they are not recognized by ContainerD and additionally get the default namespace assigned. Altough, images pulled with K8s from a image registry are getting imported correctly. The faulty namespace behavior cannot be changed during the image build. Therefore, the change of the namespace is required for locally imported images on the one hand, and it requires a transferation of the image from the Docker environment to the ContainerD environemnt on the other hand. Thus, the only solution to use local images without uploading them to a image registry, is by performing the following steps for each image:

1. Building the image in Docker (using `docker build`)
2. Exporting the image to an archive file (using `docker save`)
3. Loading the image to ContainerD to correct namespace (using `ctr --namespace k8s.io image import`)

To automate this process for a batch of container files, a powershell script was created.

- K8s Documentation redirects to a different page when searching for tutorial for adding windows nodes - Error messages of Windows hcs shim are not explanatory

- Windows needs to have certain Update installed to run flannel container (KB4489899)

- Windows base image has to have same build number than host computer (not the same in linux containers)

- Scripts and docs are maintained by a relatively small community (SIG windows tools)

- Cluster networking throws weird error messages when performed inside hyper-v vm ("Directory not found")

- Docker support was removed, containerd is not fully supported yet (without bugs) in Windows

- OpenTwin needs to be compiled on the system it is running. Development team needs to fix this issue.

- MongoDB server container image cannot be initialized easily on Windows

Chapter 6

Discussion

6.1 Analysis

Chapter 7

Conclusion and future work

7.1 Future work

- Linux port and cluster based on linux - Automated image building using Packer.io (for multiple platforms) - Ranger for streamlining kubernetes deployment - Images verkleinern - nur die Dateien ins Image bundlen, die für den entsprechenden Service notwendig sind. - Put certificate files into container via file mount - Design?

- Falls OT sowieso in Container gebaut werden muss, dann kann dieser Schritt auch automatisiert werden (clone aus git, bauen mit Build Tools) - Dann ist auch nicht mehr nötig, Deployment Verzeichnis manuell auf jeden Node zu kopieren.

- K8s checkliste, Blockzitat zur korrekten Funktionsweise des Netwerks -> Funktioniert nicht

7.2 Conclusion

Appendix A

Overview about submitted code

The created code and configuration file was uploaded to the project GitHub repository. The repository has private access only. Access to the repository can be granted by request.

The GitHub repository is located at: https://github.com/pth68/SimulationPlatform

The location of the additional files that are relevant for creation of the cluster are located in the **"Distribution"** sub directory of the repository. Table A.1 describes the submitted files (relative from the "Distribution" directory) and their intention.

Path	**Filename**	**Intention**
Container/	build-image.ps1	Automation script for building, saving, and re-importing container images for ContainerD.
	setup-node.ps1	Script for automated installation of Kubernetes node prerequisites and setup by treating the flaws of a manual installation.
	authorisation.Containerfile	Containerfile for OpenTwin's Authorisation service.
	globalsession.Containerfile	Containerfile for OpenTwin's global session service.
	mongodb.Containerfile	Containerfile for the containerized MongoDB server.
	session.Containerfile	Containerfile for OpenTwin's local session service.
	compose.yaml	A container compose file for testing the setup of containerized services.
Container/mongodb/	0-init.js	A initialization script for setting up the admin user and roles of the OpenTwin database.
	mongodb.conf	The MongoDB Server configuration file.
Controlplane/	kubeadm_config.yaml	The configuration which should be passed during initialization of the cluster using kubeadm.
	open_twin.yaml	The K8s configuration file for deploying the cluster.
Kubernetes/	open_twin.yaml	The Kubernetes configuration file for deploying the cluster.

TABLE A.1: List of files in the code repository and their usage description.

Acronyms

OT	OpenTwin	2
Windows	Microsoft® Windows®	2
cgroup	control group	18
SIG	Special Interest Group	30
DLL	Dynamic Link Library	3
URL	Uniform Resource Locator	5
UI	User Interface	2
TLS	Transfer Layer Security	6
mTLS	mutual Transfer Layer Security	6
CA	Certificate Authority	8
CSR	Certiticate Request	8
GSS	Global Session Service	4
AUTH	Authorization Service	4
LSS	Local Session Service	4
HTTP	Hypertext transfer protocol	7
API	Application Programming Interface	13
CLI	Command line interface	15
K8s	Kubernetes	13
IP	Internet Protocol	14
NAT	Network Address Translation	17
VNI	Virtual Network Identifier	30
CNI	Container Network Interface	17
CRI	Container Runtime Interface	18
NSSM	Non-Sucking Service Manager	31
VM	Virtual Machine	20
HCS	Host Compute Service	19

Bibliography

[1] *The Transport Layer Security (TLS) Protocol Version 1.3*, Request for Comments Std. 8446, 2018. [Online]. Available: https://www.rfc-editor.org/rfc/rfc8446.html

[2] “Mutual tls - what is mtls?” Cloudflare. [Online]. Available: https://www.cloudflare.com/learning/access-management/what-is-mutual-tls/ (Visited on 2023-03-07)

[3] M. Lukša, *Kubernetes in action: Anwendungen in Kubernetes-Clustern bereitstellen und verwalten*, ser. Hanser eLibrary. München: Hanser, 2018. ISBN: 9783446456020. [Online]. Available: https://www.hanser-elibrary.com/doi/book/10.3139/9783446456020

[4] Scooley, “Windows container platform: Microsoft learn,” 2022. [Online]. Available: https://learn.microsoft.com/en-us/virtualization/windowscontainers/deploy-containers/containerd (Visited on 2023-02-28)

[5] “Pod lifecycle: Documentation,” Kubernetes, 2023-02-17. [Online]. Available: https://kubernetes.io/docs/concepts/workloads/pods/pod-lifecycle/ (Visited on 2023-03-04)

[6] N. Marathe, A. Gandhi, and J. M. Shah, “Docker swarm and kubernetes in cloud computing environment,” in *Proceedings of the International Conference on Trends in Electronics and Informatics (ICOEI 2019)*. Piscataway, NJ: IEEE, 2019, pp. 179–184.

[7] B. Kang, J. Jeong, and H. Choo, “Docker swarm and kubernetes containers for smart home gateway,” *IT Professional*, vol. 23, no. 4, pp. 75–80, 2021.

[8] “Daemonset: Kubernetes documentation,” Kubernetes, 2022-08-31. [Online]. Available: https://kubernetes.io/docs/concepts/workloads/controllers/daemonset/ (Visited on 2023-02-12)

[9] “The kubernetes api: Documentation,” Kubernetes, 2022-10-24. [Online]. Available: https://kubernetes.io/docs/concepts/overview/kubernetes-api/ (Visited on 2023-02-10)

[10] "Version 3.1.0 of the openapi-specification: Github - oai/openapi-specification," OpenAPI Initiative, 2023-02-10. [Online]. Available: https://github.com/OAI/OpenAPI-Specification/blob/main/versions/3.1.0.md (Visited on 2023-02-10)

[11] Suse Rancher Community, "Comparing kubernetes cni providers: Flannel, calico, canal, and weave," 2023-02-12. [Online]. Available: https://www.suse.com/c/rancher_blog/comparing-kubernetes-cni-providers-flannel-calico-canal-and-weave/ (Visited on 2023-02-12)

[12] "Flannel backends documentation," 2023-02-12. [Online]. Available: https://github.com/flannel-io/flannel/blob/master/Documentation/backends.md (Visited on 2023-02-12)

[13] "Project calico," Tigera, 2023-02-10. [Online]. Available: https://www.tigera.io/project-calico/ (Visited on 2023-02-13)

[14] "Guide for adding windows node: Documentation," Kubernetes SIG Windows Tools, 2023-02-13. [Online]. Available: https://github.com/kubernetes-sigs/sig-windows-tools/blob/727707fa7d83d401956b467a5ce41700cce7d9a3/guides/guide-for-adding-windows-node.md (Visited on 2023-02-13)

[15] "Adding windows nodes: Kubernetes - documentation (archived / v1.23)," Kubernetes, 2022-04-19. [Online]. Available: https://v1-23.docs.kubernetes.io/docs/tasks/administer-cluster/kubeadm/adding-windows-nodes/ (Visited on 2023-02-12)

[16] "Don't panic: Kubernetes and docker," Kubernetes, 2020. [Online]. Available: https://kubernetes.io/blog/2020/12/02/dont-panic-kubernetes-and-docker/ (Visited on 2023-02-02)

[17] "Kubernetes is moving on from dockershim: Commitments and next steps," Kubernetes, 2022. [Online]. Available: https://kubernetes.io/blog/2022/01/07/kubernetes-is-moving-on-from-dockershim/ (Visited on 2023-02-13)

[18] "Guide for adding windows node: Rbac config not found: Issue #261 - kubernetes-sigs/sig-windows-tools," Kubernetes SIG Windows Tools, 2023-02-02. [Online]. Available: https://github.com/kubernetes-sigs/sig-windows-tools/issues/261 (Visited on 2023-02-02)

[19] "Windows node with containerd can't run flannel and kubeproxy daemonsets: Issue #128 - kubernetes-sigs/sig-windows-tools," Kubernetes SIG Windows Tools, 2023-02-02.

[Online]. Available: https://github.com/kubernetes-sigs/sig-windows-tools/issues/128 (Visited on 2023-02-02)

[20] amaltinsky, “Windows containers issue: Windows server 2022 container on windows 10 (10.0.19044) host.” Microsoft, 2022. [Online]. Available: https://github.com/microsoft/Windows-Containers/issues/258 (Visited on 2022-10-18)

[21] “kube-flannel for linux fails to (re)start after rbac for kube-flannel for windows is applied: Issue #277 - kubernetes-sigs/sig-windows-tools,” Kubernetes SIG Windows Tools, 2023-02-02. [Online]. Available: https://github.com/kubernetes-sigs/sig-windows-tools/issues/277 (Visited on 2023-02-02)

[22] “Introducing the host compute service (hcs): Blog post, archived technet article,” Microsoft, 2017. [Online]. Available: https://techcommunity.microsoft.com/t5/containers/introducing-the-host-compute-service-hcs/ba-p/382332 (Visited on 2023-02-27)

[23] V. Ramos Apolinario, *Windows Containers for IT Pros: Transitioning Existing Applications to Containers for on-Premises, Cloud, or Hybrid*. Berkeley, CA: Apress L. P, 2021. ISBN: 9781484266861. [Online]. Available: https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=6508410

[24] Mattbriggs, “Windows container base images: Microsoft documentation,” 2023-02-14. [Online]. Available: https://learn.microsoft.com/en-us/virtualization/windowscontainers/manage-containers/container-base-images (Visited on 2023-02-14)

[25] “Conditional compilation: The rust reference,” Rust, 2023-02-09. [Online]. Available: https://doc.rust-lang.org/reference/conditional-compilation.html (Visited on 2023-03-05)

[26] S. C. Gadde, “failed to create containerd task: hcsshim::createcomputesystem kube-proxy: The directory name is invalid.” [Online]. Available: https://stackoverflow.com/questions/74799620/kubernetes-windows-worker-node-addition-failed-to-create-containerd-task-hcss

[27] ManKier, “Containerfile format: automate the steps of creating a container image.” [Online]. Available: https://www.mankier.com/5/Containerfile (Visited on 2023-02-23)