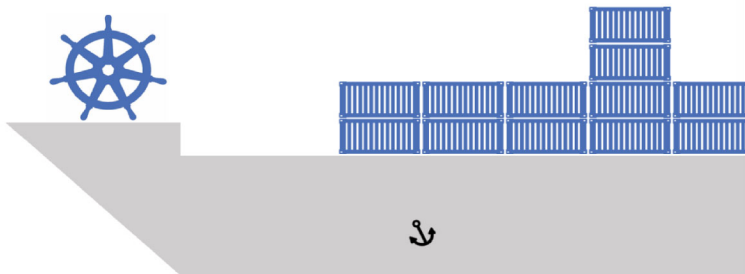


Mar 2019

The Kubernetes Book



Nigel Poulton
& Pushkar Joglekar

The Kubernetes Book

Nigel Poulton

This book is for sale at <http://leanpub.com/thekubernetesbook>

This version was published on 2019-03-09



Leanpub

This is a [Leanpub](#) book. Leanpub empowers authors and publishers with the Lean Publishing process. [Lean Publishing](#) is the act of publishing an in-progress ebook using lightweight tools and many iterations to get reader feedback, pivot until you have the right book and build traction once you do.

© 2017 - 2019 Nigel Poulton

Education is about inspiring and creating opportunities. I hope this book, and my video training courses, inspire you and create some great new opportunities!

A huge thanks to my family for putting up with me. I'm a geek who thinks he's software running on midrange biological hardware. I know it's not easy living with me.

Thanks as well to everyone who watches my Pluralsight and A Cloud Guru training videos. I love connecting with you and appreciate all the feedback I've had over the years. That is what inspired me to write this book. I think you'll love it, and I hope it helps drive your career forward.

@nigelpoulton

Contents

0: About the book	1
Paperback	1
Audio book	1
eBook and Kindle editions	1
Feedback	1
Why should anyone read this book or care about Kubernetes?	2
Should I buy the book if I've already watched your video training courses?	2
Free updates to the book	2
Versions of the book	3
1: Kubernetes primer	4
Kubernetes background	4
Where did Kubernetes come from	4
A data center OS	7
Chapter summary	9
2: Kubernetes principles of operation	10
Kubernetes from 40K feet	10
Masters and nodes	12
Packaging apps	17
The declarative model and desired state	18
Pods	20
Deployments	23
Services	23
Chapter summary	26
3: Installing Kubernetes	27
Play with Kubernetes	28
Docker Desktop	31
Minikube	32
Google Kubernetes Engine (GKE)	37
Installing Kubernetes on AWS with kops	40
Installing Kubernetes with kubeadm	45

CONTENTS

kubectl	50
Chapter summary	52
4: Working with Pods	53
Pod theory	53
Hands-on with Pods	59
Chapter Summary	66
5: Kubernetes Deployments	67
Deployment theory	67
How to create a Deployment	73
Performing a rolling update	77
How to perform a rollback	79
Chapter summary	81
6: Kubernetes Services	82
Setting the scene	82
Theory	82
Hands-on with Services	89
Real world example	96
Chapter Summary	99
7: Kubernetes storage	100
The big picture	100
Storage Providers	102
The Container Storage Interface (CSI)	102
The Kubernetes persistent volume subsystem	102
Storage Classes and Dynamic Provisioning	109
Demo	112
Chapter Summary	117
8: Other important Kubernetes stuff	118
DaemonSets	118
StatefulSets	119
Jobs and CronJobs	120
Autoscaling	120
Role-based access control (RBAC)	121
Helm	122
Chapter Summary	122
9: Threat modeling Kubernetes	123
Threat model	123
Spoofing	123
Tampering	125

CONTENTS

Repudiation	127
Information Disclosure	129
Denial of Service	130
Elevation of privilege	133
Pod Security Policies	139
Summary	141
10: Real-world Kubernetes security	142
CI/CD pipeline	142
Infrastructure and networking	148
Identity and access management (IAM)	154
Auditing and security monitoring	155
Real world example	157
Summary	157
11: What next	158
Practice makes perfect	158
More books	158
Video training	158
Events and meetups	159
Feedback	160

0: About the book

This is an *up-to-date* book about Kubernetes. It's relatively short, and it's straight-to-the-point.

Let me be clear about this, as I don't want to mislead people... **This is not a deep dive, and it does not attempt to cover everything.** This is an easy-to-read book that covers the fundamental and important parts of Kubernetes.

Paperback

A paperback version is available in selected Amazon markets. I have no control over which markets Amazon makes the paperback available in – if it was my choice, I'd make it available everywhere.

I've opted for a **high-quality, full-color paperback** that I think you'll love. That means no cheap paper, and no black-and-white diagrams from the 1990's.

Audio book

I plan to make an audio version of the book available via Audible in April 2019. There will be minor tweaks to the examples and labs so that they are easier to follow in an audio book.

eBook and Kindle editions

The easiest place to get an electronic copy is leanpub.com. It's a slick platform and updates are free.

You can also get a Kindle edition from Amazon, which also gets free updates. However, Kindle is notoriously bad at delivering updates. If you have problems getting updates to your Kindle, contact Kindle Support and they will resolve the issue.

Feedback

I'd love it if you'd give the book a review on Amazon. Writing technology books is lonely work. I literally spent months making this book as great as possible, so a couple of minutes of your time for a review would be magic. No pressure though, I won't hunt you down at the next KubeCon if you don't.

Why should anyone read this book or care about Kubernetes?

Kubernetes is white-hot, and Kubernetes skills are in high demand. So, if you want to push ahead with your career and work with a technology that's shaping the future, you need to read this book. If you don't care about your career and are fine being left behind, don't read it. It's the truth.

Should I buy the book if I've already watched your video training courses?

Kubernetes is Kubernetes. So yes, there's obviously some similar content between my books and video courses. But reading books and watching videos are totally different experiences. In my opinion, videos are more fun, but books are easier to write notes in and flick through when you're trying to find something.

If I was you, I'd watch the videos *and* get the book. They complement each other, and learning via multiple methods is a proven strategy.

Final word: Take a look at the reviews my videos and books have. That should reassure you they'll be great investments.

Some of my Video courses:

- Getting Started with Kubernetes (pluralsight.com)
- Kubernetes Deep Dive (acloud.guru)
- Docker Deep Dive (pluralsight.com)

Free updates to the book

I've done everything I can to make sure your investment in this book is as safe as possible!

All Kindle and Leanpub customers receive all updates at no extra cost. Updates work well on Leanpub, but it's a different story on Kindle. Many readers complain that their Kindle devices don't get access to updates. This is a common issue, and one that is easily resolved by *contacting Kindle Support*.

If you buy the paperback version from **Amazon.com**, you can get the Kindle version at the discounted price of \$2.99. This is done via the *Kindle Matchbook* program. Unfortunately, Kindle Matchbook is only available in the US, and it's buggy — sometimes the Kindle Matchbook icon doesn't appear on the book's Amazon selling page. Contact Kindle Support if you have issues like this and they'll sort things out.

That's the best I can do!

Things will be different if you buy the book through other channels, as I have no control over them. I'm a techie, not a book publisher ;-)

Versions of the book

Kubernetes is developing fast! As a result, the value of a book like this is inversely proportional to how old it is. In other words, the older any Kubernetes book is, the less valuable it is. With this in mind, **I'm committed to updating the book at least once per year.** And when I say “update”, I mean real updates — every word and concept is reviewed, and every example is tested and updated. **I'm 100% committed to making this book the best Kubernetes book in the world**

If at least one update a year seems like a lot... welcome to the new normal.

We no longer live in a world where a 2-year-old technology book is valuable. In fact, I question the value of a 1-year-old book on a topic that's developing as fast as Kubernetes. As an author I'd love to write a book that was useful for 5 years. But that's not the world we live in. Again... welcome to the new normal.

- **Version 4** March 2019. All content updated and all examples tested on the latest versions of Kubernetes. Added new Storage Chapter. Added new real-world security section with two new chapters.
- **Version 3** November 2018. Re-ordered some chapters for better flow. Removed the *ReplicaSets* chapter and shifted that content to an improved *Deployments* chapter. Added new chapter giving overview of other major concepts not covered in dedicated chapters.
- **Version 2.2** January 2018. Fixed a few typos, added several clarifications, and added a couple of new diagrams.
- **Version 2.1** December 2017. Fixed a few typos and updated Figures 6.11 and 6.12 to include missing labels.
- **Version 2.** October 2017. Added new chapter on *ReplicaSets*. Added significant changes to *Pods* chapter. Fixed typos and made a few other minor updates to existing chapters.
- **Version 1.** Initial version.

1: Kubernetes primer

This chapter is split into two main sections.

- Kubernetes background – where it came from etc.
- The idea of Kubernetes as a data center OS

Kubernetes background

Kubernetes is an orchestrator. For the most part, it orchestrates containerized cloud-native apps. However, there are projects that enable it to orchestrate things like virtual machines and functions (serverless workloads). All of this is adding up to Kubernetes being the de-facto orchestrator for *cloud-native applications*.

That's great, but what do we mean when use terms like *orchestrator* and *cloud-native*?

An *orchestrator* is a back-end system that deploys and manages applications. This means it helps you deploy your application, scale it up and down, perform updates and rollbacks, and more. If it's a good orchestrator, it does this without you having to supervise.

A *cloud-native application* is a business application that is made from a set of small independent services that communicate and form into a useful application. As the name suggests, this design allows it to cope with cloud-like demands and run natively on cloud platforms. As an example, cloud-native applications are designed and written so that they can easily be scaled up and down as demand rises and falls. It's also simple to update them and perform rollbacks. They can also self-heal.

More on these concepts throughout the book.

Note: Despite the name, *cloud-native* apps can also run on-premises. In fact, an attribute of a cloud-native app might be the ability to run anywhere – any cloud, or any on-prem datacenter.

Where did Kubernetes come from

Let's start from the beginning...

Kubernetes came out of Google. It is the product of Google's many years orchestrating containers at extreme scale. It was open-sourced in the summer of 2014 and handed over to the Cloud Native Computing Foundation (CNCF).



Figure 1.1

Since then, it's become the most important cloud-native technology on the planet.

Like many of the modern cloud-native projects, it's written in Go (Golang), it lives on Github at `kubernetes/kubernetes`, it's actively discussed on the IRC channels, you can follow it on Twitter (`@kubernetesio`), and `slack.k8s.io` is a pretty good slack channel. There are also regular meetups and conferences all over the planet.

Kubernetes and Docker

Kubernetes and Docker are complementary technologies. For example, it's common to develop your applications with Docker and use Kubernetes to orchestrate them.

In this model, you write your code in your favourite languages, and then use Docker to package it, test it, and ship it. But the final step of running it in test or production is handled by Kubernetes.

At a high-level, you might have a Kubernetes cluster with 10 nodes to run your production applications. Behind the scenes though, each node is running Docker as its container runtime. This means that Docker is the low-level technology that starts and stops containers etc., and Kubernetes is the higher-level technology that looks after the bigger picture things like; deciding which nodes to run containers on, deciding when to scale up or down, and executing updates.

Figure 1.2 shows a simple Kubernetes cluster with some nodes using Docker as the container runtime.

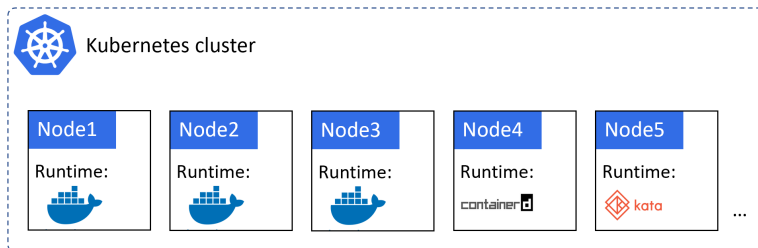


Figure 1.2

As can be seen in Figure 1.2, Docker isn't the only container runtime that Kubernetes supports. In fact, Kubernetes has a couple of features that abstract the container runtime:

1. The *Container Runtime Interface (CRI)* is an abstraction layer that standardizes the way 3rd-party container runtimes interface with Kubernetes. It allows the container runtime code to exist outside of Kubernetes, but interface with it in a supported and standardized way.

2. *Runtime Classes* is a new feature that was introduced in Kubernetes 1.12. The feature is currently in alpha and allows for different classes of runtimes. For example, the *gVisor* or *Kata Containers* runtimes might provide better isolation than Docker and containerd.

At the time of writing, *containerd* is catching up to Docker as the most commonly used container runtime in Kubernetes. It is a stripped-down version of Docker with just the stuff that Kubernetes needs.

While all of this is interesting, it's low-level stuff that should not impact your experience as a Kubernetes user. Whichever container runtime you use, the regular Kubernetes commands and patterns will continue to work as normal.

What about Kubernetes vs Docker Swarm

In 2016 and 2017 we had the *orchestrator wars* where Docker Swarm, Mesosphere DCOS, and Kubernetes fought over which would be the de-facto container orchestrator. To cut a long story short, Kubernetes won.

It's true that Docker Swarm and other container orchestrators still exist, but their development and market-share are small compared to Kubernetes.

Kubernetes and Borg: Resistance is futile!

There's a pretty good chance you'll hear people talk about how Kubernetes relates to Google's *Borg* and *Omega* systems.

It's no secret that Google has been running many of its systems on containers for years. Legendary stories of them crunching through *billions of containers a week* are common. So yes, for a very long time – even before Docker came along – Google has been running things like *search*, *Gmail*, and *GFS* on **lots** of containers.

Pulling the strings, and keeping those billions of containers in check, are a couple of in-house Google technologies and frameworks called *Borg* and *Omega*. So, it's not a huge stretch to make the connection with Kubernetes – they're all in the game of orchestrating containers at scale, and they're all related to Google.

This has occasionally led to people thinking Kubernetes is an open-sourced version of either *Borg* or *Omega*. But it's not. It's more like Kubernetes shares its DNA and family history with Borg and Omega. A bit like this... In the beginning was Borg, and Borg begat Omega. Omega *knew* the open-source community and begat her Kubernetes. Or something like that ;-)

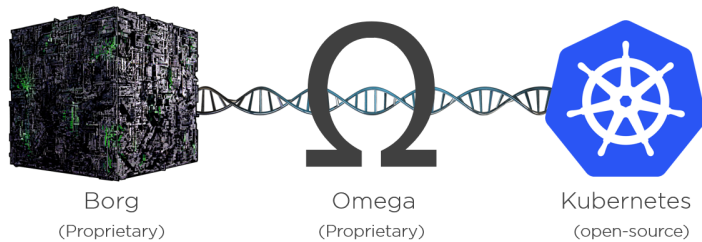


Figure 1.3 - Shared DNA

The point is, all three are separate, but all three are related. In fact, a lot of the people involved with building Borg and Omega were also involved in building Kubernetes. So, although Kubernetes was built from scratch, it leverages much of what was learned at Google with Borg and Omega.

As things stand, Kubernetes is an open-source project under the CNCF, licensed under the Apache 2.0 license, and version 1.0 shipped way back in July 2015.

Kubernetes – what’s in the name

The name **Kubernetes** (koo-ber-net-eez) comes from the Greek word meaning *Helmsman* – the person who steers a ship. This theme is reflected in the logo.



Figure 1.4 - The Kubernetes logo

Apparently, Kubernetes was originally going to be called *Seven of Nine*. If you know your Star Trek, you’ll know that *Seven of Nine* is a female **Borg** rescued by the crew of the USS Voyager under the command of Captain Kathryn Janeway. The seven spokes on the logo are also a reference to Seven of Nine. However, copyright laws prevented this.

One last thing about the name before moving on. You’ll often see Kubernetes shortened to **K8s**. The number 8 replacing the 8 characters between the K and the S – great for tweets and lazy typists like me ;-)

A data center OS

Generally speaking, containers make our previous scalability challenges look easy – we’ve just said that Google goes through billions of containers per week!

That's great, but not everybody is the size of Google. What about the rest of us?

Well... as a general rule, if your legacy apps have hundreds of VMs, there's a good chance your containerized cloud-native apps will have thousands of containers. With this in mind, we desperately need a way to manage them.

Say hello to Kubernetes.

Also, we live in a business and technology world that is increasingly fragmented and in a constant state of disruption. With this in mind, we desperately need a framework and platform that is ubiquitous and hides the complexity.

Again, say hello to Kubernetes.

When getting your head around something like Kubernetes it's important to understand modern data center architectures. For example, we're moving away from the traditional view of the data center as collection of computers. Instead, we're viewing it as a *single large computer*.

OK... but what does that even mean?

A typical computer is a collection of CPU, RAM, storage, and networking. But we've done a great job of building operating systems (OS) that abstract most of that away. For example, it's rare for a developer to care which CPU core or exact memory address their application uses – we let the OS decide all of that. And it's a good thing, the world of application development is a far friendlier place for it.

So, it's natural to take this to the next level and apply those same abstractions to data center resources – to view the data center as just a pool of compute, network and storage, and have a data center operating system that abstracts it. This means we no longer need to care about which server or storage volume our containers are running on – just leave this up to the data center OS.

In some ways, Kubernetes is a data center OS. Others do exist, but they're all in what we often call *the cattle business*. In this model, you forget about naming your servers, mapping volumes in a spreadsheet, and otherwise treating your data center assets like *pets*. Systems like Kubernetes don't care. Gone are the days of taking your app and saying “*Run this part of the app on this specific named-node, with this IP, on this specific volume...*”. In the cloud-native Kubernetes world, we're more about saying “*Hey Kubernetes, I've got this app and it consists of these parts... just run it for me please*”. Kubernetes then goes off and does all the hard work of scheduling and orchestrating.

Note: No offence is intended to anyone, or any animals, when using the terms *pets* and *cattle*.

Let's look at a quick analogy...

Think about the process of sending goods via a courier service. You package the goods in the courier's standard packaging, put a label on it, and hand it over to the courier. The courier takes care of everything else – all the complex logistics of which planes and trucks it goes on, which highways to use, and who the driver should be etc. They also provide services that let you do things like track

your package and make delivery changes. The point is, the only thing that the courier requires is that the goods are packaged and labelled according to their requirements.

The same goes for apps in Kubernetes. Package them as containers, give them a declarative manifest, and let Kubernetes take care of running them and keeping them running. You also get a rich set of tools and APIs that let you introspect (observe and examine) your app . It's a beautiful thing.

While all of this sounds great, don't take this *data center OS* analogy too far. It's not a point-and click install, you don't end up with a shell prompt to control your entire data center, and you definitely don't get a free solitaire card game.

We're in the early stages, but Kubernetes is leading the way and I think you'll love it.

Chapter summary

Kubernetes is the leading orchestrator of cloud-native apps. At the highest-level, it's all about providing an industry-standard API in front of a datacenter. It pools data center resources such as, CPU, RAM and storage, and fronts them with an extensive API. We then give Kubernetes an application, along with a description of how it should run, and let Kubernetes make it happen.

It came out of Google, it's open-sourced under the Apache 2.0 license, and lives within the Cloud Native Computing Foundation (CNCF).

Tip!

Kubernetes is a fast-moving project under active development. This means things are changing fast. But don't let that put you off – embrace it. **Change is the new normal!**

As well as reading this book, I suggest you follow @kubernetesio on Twitter, hit the various k8s slack channels, and attend your local meetups. These will all help to keep you up-to-date with the latest and greatest in the Kubernetes world. I'll also be updating the book regularly and producing more video training courses.

Keep an eye on pluralsight.com and acloud.guru for my latest video courses.

2: Kubernetes principles of operation

In this chapter, we'll learn about the major components needed to build a Kubernetes cluster and deploy an app. The aim is to give you an overview of the major concepts. But don't worry if you don't understand everything straight away, we'll cover most things again as we progress through the book.

We'll divide the chapter as follows:

- Kubernetes from 40K feet
- Masters and nodes
- Packaging apps
- Declarative configuration and desired state
- Pods
- Deployments
- Services

Kubernetes from 40K feet

At the highest level, Kubernetes is two things:

- A cluster for running applications
- An orchestrator of cloud-native microservices apps.

On the *cluster* front, Kubernetes is like any other cluster – a bunch of nodes and a control plane. The control plane exposes an API, has a scheduler for assigning work to nodes, and state is recorded in a persistent store. Nodes are where application services run.

Kubernetes is API-driven and uses standard HTTP RESTful verbs to view and update the cluster.

On the *orchestrator* front, “orchestrator” is just a fancy name for an application that's made from lots of small independent services that work together to form a useful app.

Let's look at a quick analogy.

In the real world, a football (soccer) team is made of individuals. No two are the same, and each has a different role to play in the team – some defend, some attack, some are great at passing, some tackle, some shoot... Along comes the coach, and he or she gives everyone a position and organizes them into a team with a purpose. We go from Figure 2.1 to Figure 2.2.

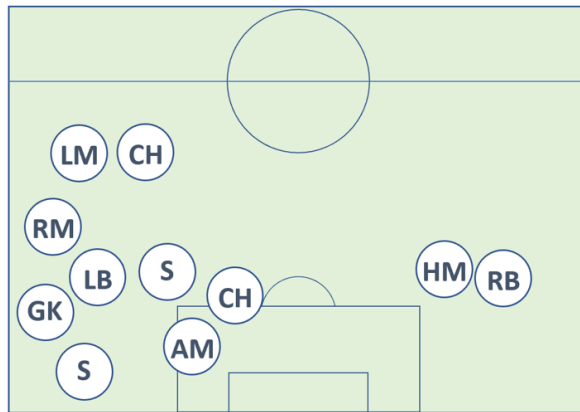


Figure 2.1

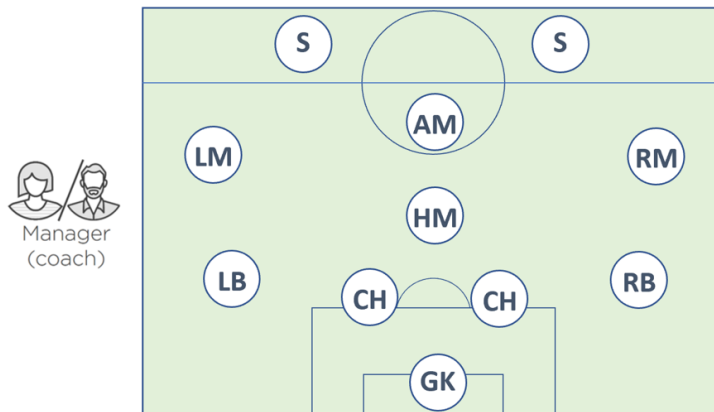


Figure 2.2

The coach also makes sure that the team maintains its formation, sticks to the game-plan, and deals with any injuries and other changes in circumstances. Well guess what... microservices apps in the Kubernetes world are the same.

Stick with me on this...

We start out with lots of individual specialised services. Some serve web pages, some do authentication, some perform searches, others persist data. Kubernetes comes along – a bit like the coach in the football analogy -- organizes everything into a useful app and keeps things running smoothly. It even responds to changes in circumstances.

In the sports world we call this *coaching*. In the application world we call it *orchestration*. Kubernetes is an *orchestrator*.

To make this happen, we start out with an app, package it up and give it to the cluster (Kubernetes).

The cluster is made up of one or more *masters* and a bunch of *nodes*.

The masters, sometimes called *heads* or *head nodes*, are in-charge of the cluster. This means they make the scheduling decisions, perform monitoring, implement changes, respond to events, and more. For these reasons, we often refer to the masters as the *control plane*.

The nodes are where application services run, and we sometimes call them the *data plane*. They have a reporting line back to the masters, and constantly watch for new work assignments.

To run applications on a Kubernetes cluster we follow this simple pattern:

1. Write the application as small independent services in our favourite languages.
2. Package each service in its own container.
3. Wrap each container in its own Pod.
4. Deploy Pods to the cluster via higher-level controllers such as; *Deployments*, *DaemonSets*, *StatefulSets*, *CronJobs* etc.

We're still near the beginning of the book and you're not expected to know what all of this means yet. However, at a high-level, *Deployments* offer scalability and rolling updates, *DaemonSets* run one instance of a Pod on every node in the cluster, *StatefulSets* are for stateful application components, and *CronJobs* are for work that needs to run at set times. There are more than these, but these will do for now.

Kubernetes likes to manage applications *declaratively*. This is a pattern where we describe how we want our application to look and feel in a set of YAML files, POST these files to Kubernetes, then sit back while Kubernetes makes it all happen.

But it doesn't stop there. Because the declarative pattern defines how we want an application to look, Kubernetes can watch it and make sure things are how they should be. If something isn't as it should be, Kubernetes tries to fix it.

That's the big picture. Let's dig a bit deeper.

Masters and nodes

A Kubernetes cluster is made of masters and nodes. These are Linux hosts that can be VMs, bare metal servers in your data center, or instances in a private or public cloud.

Masters (control plane)

A Kubernetes master is a collection of system services that make up the control plane of the cluster.

The simplest setups run all the master *services* on a single host. However, this is only suitable for labs and test environments. For production environments, multi-master high availability (HA) is a **must**

have. This is why the major cloud providers implement HA masters as part of their Kubernetes-as-a-Service platforms such as Azure Kubernetes Service (AKS), AWS Elastic Kubernetes Service (EKS), and Google Kubernetes Engine (GKE).

Generally speaking, running 3 or 5 replicated masters in an HA configuration is recommended.

It's also considered a good practice not to run user applications on masters. This allows masters to concentrate entirely on managing the cluster.

Let's take a quick look at the different master services that make up the control plane.

The API server

The API server is the Grand Central Station of Kubernetes. All communication, between all components, goes through the API server. We'll get into the detail later in the book, but it's important to understand that internal system components, as well as external user components, all communicate via the same API.

It exposes a RESTful API that we POST YAML configuration files to over HTTPS. These YAML files, which we sometimes call *manifests*, contain the desired state of our application. This includes things like; which container image to use, which ports to expose, and how many Pod replicas to run.

All requests to the API Server are subject to authentication and authorization checks, but once these are done, the config in the YAML file is validated, persisted to the cluster store, and deployed to the cluster.

You can think of the API server as the brains of the cluster – where the smarts are implemented.

The cluster store

If the API server is the brains of the cluster, the *cluster store* is its heart. It's the only stateful part of the control plane, and it persistently stores the entire configuration and state of the cluster. As such, it's a vital component of the cluster – no cluster store, no cluster.

The cluster store is currently based on **etcd**, a popular distributed database. As it's the *single source of truth* for the cluster, you should run between 3-5 etcd replicas for high-availability, and you should provide adequate ways to recover when things go wrong.

On the topic of *availability*, etcd prefers consistency over availability. This means that it will not tolerate a split-brain situation and will halt updates to the cluster in order to maintain consistency. However, if etcd becomes unavailable, applications running on the cluster should continue to work, it's just updates to the cluster configuration that will be halted.

As with all distributed databases, consistency of writes to the database is important. For example, multiple writes to the same value originating from different nodes needs to be handled. etcd uses the popular RAFT consensus algorithm to accomplish this.

The controller manager

The controller manager is a *controller of controllers* and is shipped as a single monolithic binary. However, despite it running as a single process, it implements multiple independent control loops that watch the cluster and respond to events.

Some of the control loops include; the node controller, the endpoints controller, and the replicaset controller. Each one runs as a background watch-loop that is constantly watching the API Server for changes – the aim of the game is to ensure the *current state* of the cluster matches the *desired state* (more on this shortly).

The logic implemented by each control loop is effectively this:

1. Obtain desired state
2. Observe current state
3. Determine differences
4. Reconcile differences

This logic is at the heart of Kubernetes and declarative design patterns.

Each control loop is also extremely specialized and only interested in its own little corner of the Kubernetes world. No attempt is made to over-complicate things by implementing awareness of other parts of the system – each takes care of its own task and leaves other components alone. This is key to the distributed design of Kubernetes and adheres to the Unix philosophy of building complex systems from small specialized parts.

Note: Throughout the book we'll use terms like *control loop*, *watch loop*, and *reconciliation loop* to mean the same thing.

The scheduler

At a high level, the scheduler watches for new work tasks and assigns them to appropriate healthy nodes. Behind the scenes, it implements complex logic that filters out nodes incapable of running the Pod and then ranks the nodes that are capable. The ranking system itself is complex, but the node with the highest ranking points is eventually selected to run the Pod.

When identifying nodes that are capable of running the Pod, the scheduler performs various predicate checks. These include; is the node tainted, are there any affinity or anti-affinity rules, is the Pod's network port available on the node, does the node have sufficient free resources etc. Any node incapable of running the Pod is ignored, and the remaining Pods are ranked according to things such as; does the node already have the required image, how much free resource does the node have, how many Pods is the node already running. Each criteria is worth points, and the node with the most points is selected to run the Pod.

If the scheduler cannot find a suitable node, the Pod cannot be scheduled and goes into pending.

It's not the job of the scheduler to perform the mechanics of *running* Pods, it just picks the nodes they will be *scheduled* on.

The cloud controller manager

If you're running your cluster on a supported public cloud platform, such as AWS, Azure, or GCP, your control plane will be running a *cloud controller manager*. Its job is to manage integrations with underlying cloud technologies and services such as, instances, load-balancers, and storage.

Control Plane summary

Kubernetes masters run all of the cluster's control plane services. Think of it as brains of the cluster where all the control and scheduling decisions are made. Behind the scenes, a master is made up of lots of small specialized control loops and services. These include the API server, the cluster store, the controller manager, and the scheduler.

The API Server is the front-end into the control plane and the only component in the control plane that we interact with directly. By default, it exposes a RESTful endpoint on port 443.

Figure 2.3 shows a high-level view of a Kubernetes master (control plane).

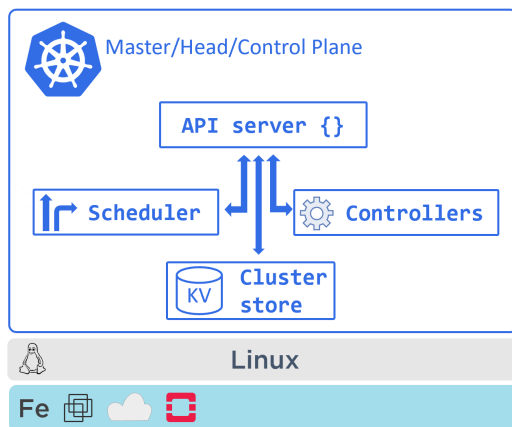


Figure 2.3 - Kubernetes Master

Nodes

Nodes are the workers of a Kubernetes cluster. At a high-level they do three things:

1. Watch the API Server for new work assignments
2. Execute new work assignments
3. Report back to the control plane

As we can see from Figure 2.4, they're a bit simpler than *masters*. Let's look at the three major components of a node.

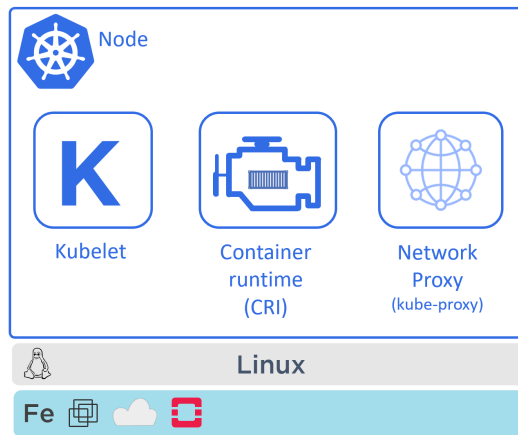


Figure 2.4 - Kubernetes Node (formerly Minion)

Kubelet

The Kubelet is the star of the show on every Node. It's the main Kubernetes agent, and it runs on every node in the cluster. In fact, it's common to use the terms *node* and *kubelet* interchangeably.

When you join a new node to a cluster, the process involves installation of the kubelet which is then responsible for the node registration process. This effectively pools the node's CPU, RAM, and storage into the wider cluster pool. Think back to the previous chapter where we talked about Kubernetes being a data center OS and abstracting data center resources into a single usable pool.

One of the main jobs of the kubelet is to watch the API server for new work assignments. Any time it sees one, it executes the task and maintains a reporting channel back to the control plane. It also keeps an eye on local static Pod definitions.

If a kubelet can't run a particular task, it reports back to the master and lets the control plane decide what actions to take. For example, if a Pod fails to start on a node, the kubelet is **not** responsible for finding another node to run it on. It simply reports back to the control plane and the control plane decides what to do.

Container runtime

The Kubelet needs a container runtime to perform container-related tasks – things like pulling images and starting and stopping containers.

In the early days, Kubernetes had native support for a few container runtimes such as Docker. More recently, it has moved to a plugin model called the Container Runtime Interface (CRI). This is an abstraction layer for external (3rd-party) container runtimes to plug in to. At a high-level, the CRI masks the internal machinery of Kubernetes and exposes a clean documented interface for 3rd-party container runtimes to plug in to.

The CRI is the supported method for integrating runtimes into Kubernetes.

There are lots of container runtimes available for Kubernetes. One popular example is `cri-containerd`. This is a community-based open-source project porting the CNCF `containerd` runtime to the CRI interface. It has a lot of support and is replacing Docker as the preferred container runtime used in Kubernetes.

Note: `containerd` (pronounced “container-dee”) is the container supervisor and runtime logic stripped out from the Docker Engine. It was donated to the CNCF by Docker, Inc. and has a lot of community support. Other CRI-compliant container runtimes exist.

Kube-proxy

The last piece of the *node* puzzle is the kube-proxy. This runs on every node in the cluster and is responsible for local networking. For example, it makes sure each node gets its own unique IP address, and implements local IPTables or IPVS rules to handle routing and load-balancing of traffic on the Pod network.

Kubernetes DNS

As well as the various control plane and node components, every Kubernetes cluster has an internal DNS service that is vital to operations.

The cluster’s DNS service has a static IP address that is hard-coded into every Pod on the cluster, meaning all containers and Pods know how to find it. Every new service is automatically registered with the cluster’s DNS so that all components in the cluster can find every Service by name. Some other components that are registered with the cluster DNS are StatefulSets and the individual Pods that a StatefulSet manages.

Cluster DNS is based on CoreDNS (<https://coredns.io/>).

Now that we understand the fundamentals of masters and nodes, let’s switch gears and look at how we package applications to run on Kubernetes.

Packaging apps

For an application to run on a Kubernetes cluster it needs to tick a few boxes. These include:

1. Packaged as a container
2. Wrapped in a Pod
3. Deployed via a declarative manifest file

It goes like this... We write an application service in a language of our choice. We then build it into a container image and store it in a registry. At this point, the application service is *containerized*.

Next, we define a Kubernetes Pod to run the containerized service in. At the kind of high level we're at, a Pod is just a wrapper that allows containers to run on a Kubernetes cluster. Once we've defined a Pod for the container, we're ready to deploy it on the cluster.

Kubernetes offers several objects for deploying and managing Pods. The most common is the *Deployment*, which offers scalability, self-healing, and rolling updates. We define them in a YAML file that specifies things like which image to use and how many replicas to deploy.

Figure 2.5 shows application code packaged as a *container*, running inside a *Pod*, managed by a *Deployment*.

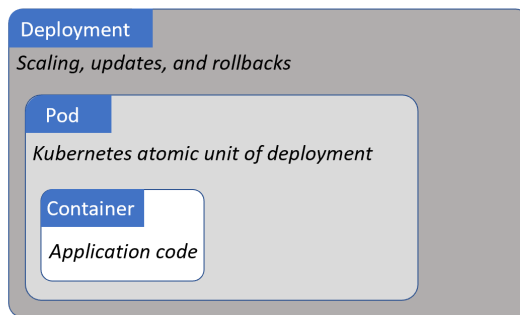


Figure 2.5

Once everything is defined in the *Deployment* YAML file, we POST it to the cluster as the *desired state* of the application and let Kubernetes implement it.

Speaking of desired state...

The declarative model and desired state

The *declarative model* and the concept of *desired state* are at the very heart of Kubernetes. Take them away and Kubernetes crumbles.

In Kubernetes, the declarative model works like this:

1. Declare the desired state of the application (microservice) in a manifest file
2. POST it to the Kubernetes API server
3. Kubernetes stores this in the cluster store as the application's *desired state*
4. Kubernetes implements the desired state on the cluster
5. Kubernetes implements watch loops to make sure the *current state* of the application doesn't vary from the *desired state*

Let's look at each step in a bit more detail.

Manifest files are written in simple YAML, and they tell Kubernetes how we want an application to look. We call this the *desired state*. It includes things such as; which image to use, how many replicas to have, which network ports to listen on, and how to perform updates.

Once we've created the manifest, we POST it to the API server. The most common way of doing this is with the `kubectl` command-line utility. This POSTs the manifest as a request to the control plane, usually on port 443.

Once the request is authenticated and authorized, Kubernetes inspects the manifest, identifies which controller to send it to (e.g. the *Deployments controller*), and records the config in the cluster store as part of the cluster's overall *desired state*. Once this is done, the work gets scheduled on the cluster. This includes the hard work of pulling images, starting containers, building networks, and starting the application's processes.

Finally, Kubernetes utilizes background reconciliation loops that constantly monitor the state of the cluster. If the *current state* of the cluster varies from the *desired state*, Kubernetes will perform whatever tasks are necessary to reconcile the issue.

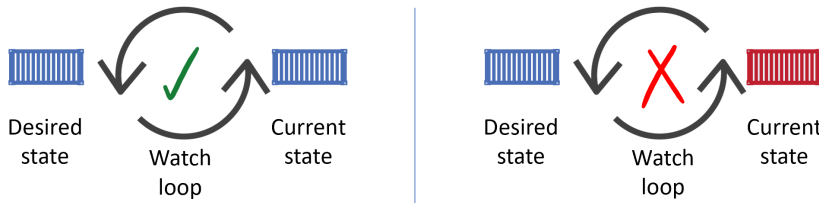


Figure 2.6

It's important to understand that what we've described is the opposite of the traditional *imperative model*. The imperative model is where we issue long lists of platform-specific commands to build things.

Not only is the declarative model a lot simpler than long lists of imperative commands, it also enables self-healing, scaling, and lends itself to version control and self-documentation. It does this by telling the cluster *how things should look*. If they stop looking like this, the cluster notices the discrepancy and does all of the hard work to reconcile the situation.

But the declarative story doesn't end there – things go wrong, and things change. When they do, the **current state** of the cluster no longer matches the **desired state**. As soon as this happens, Kubernetes kicks into action and attempts to bring the two back into harmony.

Let's consider an example.

Assume we have an app with a desired state that includes 10 replicas of a web front-end Pod. If a node that was running two replicas fails, the *current state* will be reduced to 8 replicas, but the *desired state* will still be 10. This will be observed by a reconciliation loop and Kubernetes will schedule two new replicas on other nodes in the cluster.

The same thing will happen if we intentionally scale the desired number of replicas up or down. We could even change the image we want to use. For example, if the app is currently using `v2.00` of an image, and we update the desired state to use `v2.01`, Kubernetes will notice the discrepancy and go through the process of updating all replicas so that they are using the new image version specified in the new *desired state*.

To be clear. Instead of writing a long list of commands to go through the process of updating every replica to the new version, we simply tell Kubernetes we want the new version, and Kubernetes does the hard work for us.

Despite how simple this might seem, it's extremely powerful. It's also at the very heart of how Kubernetes operates. We give Kubernetes a declarative manifest that describes how we want an application to look. This forms the basis of the application's desired state. The Kubernetes control plane records it, implements it, and runs background reconciliation loops that constantly check what is running is what we've asked for. When current state matches desired state, the world is a happy place. When it doesn't, Kubernetes gets busy fixing it.

Pods

In the VMware world, the atomic unit of scheduling is the virtual machine (VM). In the Docker world, it's the container. Well... in the Kubernetes world, it's the *Pod*.

It's true that Kubernetes runs containerized apps. However, you cannot run a container directly on a Kubernetes cluster – containers must **always** run inside of Pods.

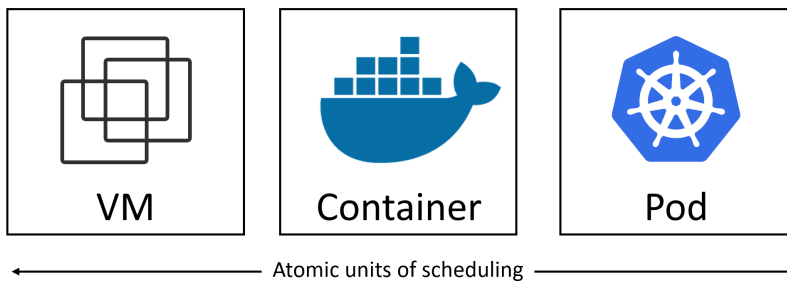


Figure 2.7

Pods and containers

The very first thing to understand is that the term *Pod* comes from a *pod of whales* – in the English language we call a group of whales a *pod of whales*. As the Docker logo is a whale, it makes sense that we call a group of containers a *Pod*.

The simplest model is to run a single container per Pod. However, there are advanced use-cases that run multiple containers inside a single Pod. These *multi-container Pods* are beyond the scope of what we're discussing here, but powerful examples include:

- Service meshes
- Web containers supported by a *helper* container that pulls the latest content
- Containers with a tightly coupled log scraper

The point is, a Kubernetes Pod is a construct for running one or more containers. Figure 2.8 shows a multi-container Pod.

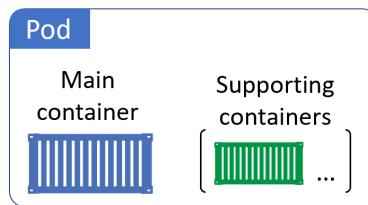


Figure 2.8

Pod anatomy

At the highest-level, a *Pod* is a ring-fenced environment to run containers. The Pod itself doesn't actually run anything, it's just a sandbox for hosting containers. Keeping it high level, you ring-fence an area of the host OS, build a network stack, create a bunch of kernel namespaces, and run one or more containers in it. That's a Pod.

If you're running multiple containers in a Pod, they all share the **same environment**. This includes things like the IPC namespace, shared memory, volumes, network stack and more. As an example, this means that all containers in the same Pod will share the same IP address (the Pod's IP). This is shown in Figure 2.9.

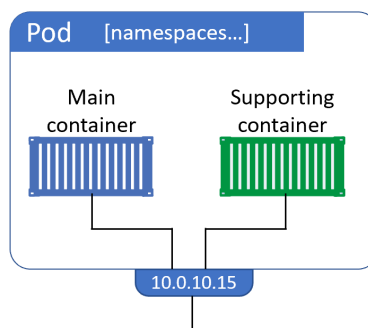


Figure 2.9

If two containers in the same Pod need to talk to each other (container-to-container within the Pod) they can use ports on the Pod's `localhost` interface as shown in Figure 2.10.

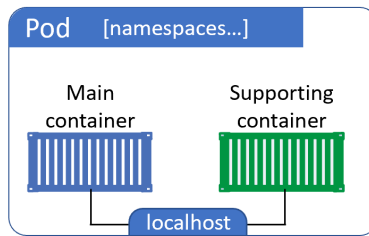


Figure 2.10

Multi-container Pods are ideal when you have requirements for tightly coupled containers that may need to share memory and storage. However, if you don't **need** to tightly couple your containers, you should put them in their own Pods and loosely couple them over the network. This keeps things clean by having each Pod dedicated to a single task.

Pods as the unit of scaling

Pods are also the minimum unit of scheduling in Kubernetes. If you need to scale your app, you add or remove Pods. You **do not** scale by adding more containers to an existing Pod. Multi-container Pods are only for situations where two different, but complimentary, containers need to share resources. Figure 2.11 shows how to scale the `nginx` front-end of an app using multiple Pods as the unit of scaling.

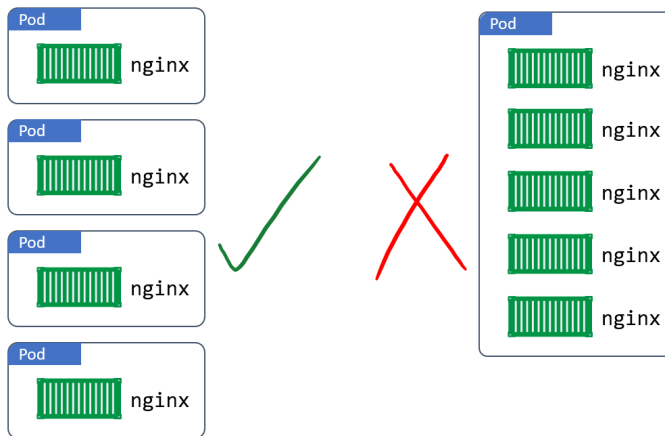


Figure 2.11 - Scaling with Pods

Pods - atomic operations

The deployment of a Pod is an atomic operation. This means that a Pod is either entirely deployed, or not deployed at all. There is never a situation where a partially deployed Pod will be servicing requests. The entire Pod either comes up and is put into service, or it doesn't, and it fails.

A single Pod can only be scheduled to a single node. This is also true of multi-container Pods – all containers in the same Pod will run on the same node.

Pod lifecycle

Pods are mortal. They're created, they live, and they die. If they die unexpectedly, we don't bring them back to life. Instead, Kubernetes starts a new one in its place. However, even though the new Pod looks, smells, and feels like the old one, it isn't. It's a shiny new Pod with a shiny new ID and IP address.

This has implications on how we should design our applications. Don't design them so they are tightly coupled to a particular instance of a Pod. Instead, design them so that when Pods fail, a totally new one (with a new ID and IP address) can pop up somewhere else in the cluster and seamlessly take its place.

Deployments

We normally deploy Pods indirectly as part of something bigger. Examples include; *Deployments*, *DaemonSets*, and *StatefulSets*.

For example, a Deployment is a higher-level Kubernetes object that wraps around a particular Pod and adds features such as scaling, zero-downtime updates, and versioned rollbacks.

Behind the scenes, they implement a controller and a watch loop that is constantly observing the cluster making sure that current state matches desired state.

Deployments have existed in Kubernetes since version 1.2 and were promoted to GA (stable) in 1.9. You'll see them a lot.

Services

We've just learned that Pods are mortal and can die. However, if they're managed via Deployments or DaemonSets, they get replaced when they fail. But replacements come with totally different IPs. This also happens when we perform scaling operations – scaling up adds new Pods with new IP addresses, whereas scaling down takes existing Pods away. Events like these cause a lot of *IP churn*.

The point we're making is that **Pods are unreliable**, which poses a challenge... Assume we've got a microservices app with a bunch of Pods performing video rendering. How will this work if other parts of the app that need to use the rendering service cannot rely on the rendering Pods being there when they need them?

This is where *Services* come in to play. **Services provide reliable networking for a set of Pods.**

Figure 2.12 shows the uploader microservice talking to the renderer microservice via a Kubernetes Service. The Kubernetes Service is providing a reliable name and IP, and is load-balancing requests to the two renderer Pods behind it.

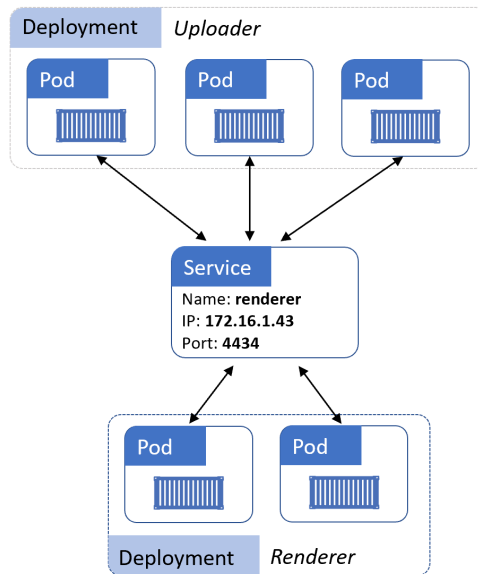


Figure 2.12

Digging in to a bit more detail. Services are fully-fledged objects in the Kubernetes API – just like Pods and Deployments. They have a front-end that consists of a stable DNS name, IP address, and port. On the back-end, they load-balance across a dynamic set of Pods. Pods come and go, the Service observes this, automatically updates itself, and continues to provide that stable networking endpoint.

The same applies if we scale the number of Pods up or down. New Pods are seamlessly added to the Service, whereas terminated Pods are seamlessly removed.

That’s the job of a Service – it’s a stable network abstraction point that provides TCP and UDP load-balancing across a dynamic set of Pods.

As they operate at the TCP and UDP layer, Services do not possess application intelligence and cannot provide application-layer routing. For that, you need an Ingress, which understands HTTP and provides host and path-based routing.

Connecting Pods to Services

Services use *labels* and a *label selector* to know which set of Pods to load-balance traffic to. The Service has a *label selector* that is a list of all the *labels* a Pod must possess in order for it to receive traffic from the Service.

Figure 2.13 shows a Service configured to send traffic to all Pods on the cluster possessing the following three labels:

- zone=prod

- env=be
- ver=1.3

Both Pods in the diagram have all three labels, so the Service will load-balance traffic to them both.

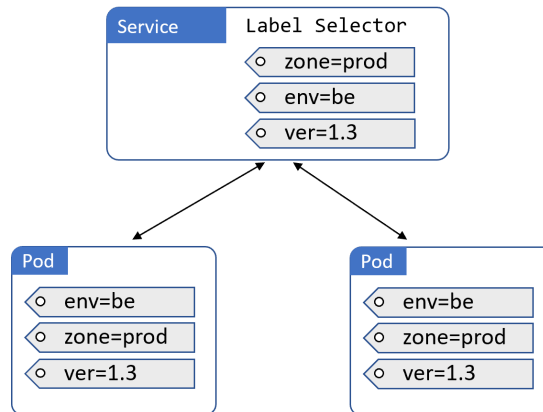


Figure 2.13

Figure 2.14 shows a similar setup. However, an additional Pod, on the right, does not match the set of labels configured in the Service's label selector. This means the Service will not load balance requests to it.

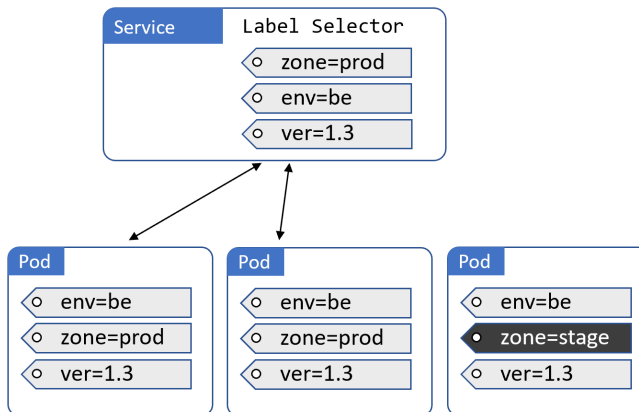


Figure 2.14

One final thing about Services. They only send traffic to **healthy Pods**. This means a Pod that is failing health-checks will not receive traffic from the Service.

That's the basics. Services bring stable IP addresses and DNS names to the unstable world of Pods.

Chapter summary

In this chapter, we introduced some of the major components of a Kubernetes cluster.

The masters are where the control plane components run. Under-the-hood, there's a combination of several system-services, including the API Server that exposes the public REST interface. Masters make all of the deployment and scheduling decisions, and multi-master HA is important for production-grade environments.

Nodes are where user applications run. Each node runs a service called the `kubelet` that registers the node with the cluster and communicates with the control plane. This includes receiving new work tasks and maintaining a reporting channel. Nodes also have a container runtime and the `kube-proxy` service. The container runtime, such as Docker or containerd, is responsible for all container-related operations. The `kube-proxy` is responsible for networking on the node.

We also talked about some of the major Kubernetes API objects such as Pods, Deployments, and Services. The Pod is the basic building-block. Deployments add self-healing, scaling and updates. Services add stable networking and load-balancing.

Now that we know the basics, we're going to start getting into the detail.

3: Installing Kubernetes

In this chapter, we'll look at a few different ways to install Kubernetes.

Things have changed a lot since I wrote the first edition of the book in July 2017. Back then, installing Kubernetes was hard. These days it's a lot easier. In fact, we're approaching the point where we can just *ask for a Kubernetes cluster and get one* – a zero-effort Kubernetes cluster. This is especially true with *hosted Kubernetes services* like Azure Kubernetes Service (AKS) and Google Kubernetes Engine (GKE).

On the topic of *hosted Kubernetes services*... More and more companies are choosing to use hosted Kubernetes services, and with things like *GKE On-Prem* (<https://cloud.google.com/gke-on-prem/>), it's looking increasingly likely that a large number of Kubernetes clusters will be built and managed via the major cloud providers.

With all of this in mind, ask yourself the following question before building your own Kubernetes cluster: *Is building and managing your own Kubernetes cluster the best use of your time and other resources?* If the answer isn't **"Hell yes"**, I strongly suggest you consider a hosted service.

OK, we'll look at the following ways to get a Kubernetes cluster:

- Play with Kubernetes (PWK)
- Docker Desktop: local development cluster on your laptop
- Minikube: local development cluster on your laptop
- Google Kubernetes Engine (GKE): production-grade hosted cluster
- Kops: Install your own cluster on AWS
- Kubeadm: Manual installation with kubeadm
- kubectl: The Kubernetes command-line tool

A couple of quick things to point out before diving in...

Firstly, there are a lot of other ways to install Kubernetes. The ones we're covering here are the ones I think will be most helpful.

Secondly, a *hosted Kubernetes service* is one where the control plane (masters) is managed by a 3rd-party such as your cloud platform. For example, AKS, EKS, and GKE are all hosted Kubernetes services where management of the control plane is taken care of by the cloud platform (you aren't responsible). 2018 saw a huge increase in use of hosted Kubernetes platforms and the trend is continuing in 2019.

Play with Kubernetes

Play with Kubernetes (PWK) is free and is a great way to get your hands on a Kubernetes cluster without having to install anything. All you need is a computer, an internet connection, and an account on Docker Hub or GitHub. In my opinion, it's the fastest and easiest way to get your hands on Kubernetes.

However, it has limitations. For starters, it's a time-limited playground – you get a lab that lasts for 4 hours. It also lacks some integrations with external services such as cloud-based load-balancers. However, limitations aside, it's a great tool and I use it all the time.

Let's see what it looks like.

1. Point your browser at <http://play-with-k8s.com>
2. Login with your GitHub or Docker Hub account and click **Start**
3. Click + **ADD NEW INSTANCE** from the navigation pane on the left of your browser

You will be presented with a terminal window in the right of your browser. This is a Kubernetes node (node1).

4. Run a few commands to see some of the components pre-installed on the node.

```
$ docker version
Docker version 18.09.0-ce...

$ kubectl version --output=yaml
clientVersion:
...
  major: "1"
  minor: "11"
```

As the output shows, the node already has Docker and `kubectl` (the Kubernetes client) pre-installed. Other tools, including `kubeadm`, are also pre-installed.

It's also worth noting that although the command prompt is a `$`, we're actually running as `root`. We can confirm this by running `whoami` or `id`.

5. Use the `kubeadm` command to initialize a new cluster

When you added a new instance in step 3, PWK gave you a short list of commands to initialize a new Kubernetes cluster. One of these was `kubeadm init...` The following command will initialize a new cluster and configure the API server to listen on the correct IP interface.

You may be able to specify the version of Kubernetes to install by adding the `--kubernetes-version` flag to the command. The latest versions can be seen at <https://github.com/kubernetes/kubernetes/releases>. Not all versions work with PWK.

```
$ kubectl init --apiserver-advertise-address $(hostname -i)
[kubeadm] WARNING: kubeadm is in beta, do not use it for prod...
[init] Using Kubernetes version: v1.11.1
[init] Using Authorization modes: [Node RBAC]
<Snip>
Your Kubernetes master has initialized successfully!
<Snip>
```

Congratulations! You have a brand new single-node Kubernetes cluster. The node that we executed the command from (node1) is initialized as the *master*.

The output of the `kubeadm init` gives you a short list of commands it wants you to run. These will copy the Kubernetes config file and set permissions. You can ignore these, as PWK has already configured them for you. Feel free to poke around inside of `$HOME/.kube`.

6. Verify the cluster with the following `kubectl` command.

```
$ kubectl get nodes
NAME      STATUS    AGE       VERSION
node1     NotReady  1m        v1.11.2
```

The output shows a single-node Kubernetes cluster. However, the status of the node is `NotReady`. This is because we haven't configured the *Pod network* yet. When you first logged on to the PWK node, you were given a list of three commands to configure the cluster. So far, we've only executed the first one (`kubeadm init...`).

7. Initialize the Pod network (cluster networking).

Copy the second command from the list of three commands that were printed on the screen when you first created node1 (this will be a `kubectl apply` command). Paste it onto a new line in the terminal. In the book, the command may wrap over multiple lines and insert backslashes (`\`). You should remove any backslashes that occur at the right-edge of the page.

```
$ kubectl apply -n kube-system -f \
  "https://cloud.weave.works/k8s/net?k8s-version=$(kubectl version | base64 | tr
  -d '\n')"
```

```
serviceaccount "weave-net" created
clusterrole "weave-net" created
clusterrolebinding "weave-net" created
role "weave-net" created
rolebinding "weave-net" created
daemonset "weave-net" created
```

8. Verify the cluster again to see if node1 has changed to `Ready`.

```
$ kubectl get nodes
NAME          STATUS    AGE          VERSION
node1         Ready     2m           v1.11.2
```

Now that the *Pod network* has been initialized and the control plane is *Ready*, you're ready to add some worker nodes.

9. Copy the `kubeadm join` command from the output of the `kubeadm init`.

When you initialized the new cluster with `kubeadm init`, the final output of the command listed a `kubeadm join` command to use when adding nodes. This command includes the cluster join-token, the IP socket that the API server is listening on, and other bits required to join a new node to the cluster. Copy this command and be ready to paste it into the terminal of a new node (`node2`).

10. Click the + ADD NEW INSTANCE button in the left pane of the PWK window.

You will be given a new node called `node2`.

1. Paste the `kubeadm join` command into the terminal of `node2`.

The join-token and IP address will be different in your environment.

```
$ kubeadm join --token 948f32.79bd6c8e951cf122 10.0.29.3:6443...
Initializing machine ID from random generator.
[preflight] Skipping pre-flight checks
<Snip>
Node join complete:
* Certificate signing request sent to master and response received.
* Kubelet informed of new secure connection details.
```

1. Switch back to `node1` and run another `kubectl get nodes`

```
$ kubectl get nodes
NAME          STATUS    AGE          VERSION
node1         Ready     5m           v1.11.2
node2         Ready     1m           v1.11.2
```

Your Kubernetes cluster now has two nodes – one master and one worker node.

Feel free to add more nodes.

Congratulations! You have a fully working Kubernetes cluster that you can use as a test lab.

It's worth pointing out that `node1` was initialized as the Kubernetes *master* and additional nodes will join the cluster as *nodes*. PWK usually puts a blue icon next to *masters* and a transparent one next to *nodes*. This helps you identify which is which.

Finally, PWK sessions only last for 4 hours and are obviously not intended for production use.

Have fun!

Docker Desktop

In my opinion, *Docker Desktop* is the best way to get a local development cluster on your Mac or Windows laptop. With a few easy steps, you get a single-node Kubernetes cluster that you can develop and test with. I use it nearly every day.

It works by creating a virtual machine (VM) on your laptop and starting a single-node Kubernetes cluster inside that VM. It also configures your `kubectl` client with a context that allows it to talk to the cluster. Finally, you get a simple GUI that lets you to perform basic operations such as switching between all of your `kubectl` contexts.

Note: A `kubectl` context is a bunch of settings that the `kubectl` command uses to know which cluster to issue commands to.

1. Point your web browser to `www.docker.com` and choose **Products > Docker Desktop**.
2. Click the download button for either Mac or Windows.

You may need to login to the Docker Store. Accounts are free, and so is the product.

3. Open the installer and follow the simple installation instructions.

Once the installer is complete, you'll get a whale icon on the Windows task bar, or the menu bar on a Mac.

4. Click the whale icon (you may need to right-click it), go to **Settings** and enable Kubernetes from the **Kubernetes** tab.

You can open a terminal window and see your cluster:

```
$ kubectl get nodes
```

NAME	STATUS	ROLES	AGE	VERSION
docker-for-desktop	Ready	master	28d	v1.13.0

Congratulations, you now have a local development cluster.

Minikube

Minikube is another option if you're a developer and need a local Kubernetes development environment on your laptop. Like *Docker Desktop*, you get a local VM running a single-node Kubernetes cluster for development. It's not for production!

Note: I've had mixed results with Minikube. It's great when it works, but sometimes it's hard to get working. For this reason, I prefer Docker Desktop for Mac and Windows.

You can get Minikube for Mac, Windows, and Linux. We'll take a quick look at Mac and Windows, as this is what most people run on their laptops.

Note: Minikube requires virtualization extensions enabled in your system's BIOS.

Installing Minikube on Mac

It's probably a good idea to install `kubectl` (the Kubernetes client) before you install Minikube. You'll use this later to issue commands to the Minikube cluster.

1. Use Brew to install `kubectl`

```
$ brew install kubernetes-cli
Updating Homebrew...
```

This puts the `kubectl` binary in `/usr/local/bin` and makes it executable.

2. Verify that the install worked.

```
$ kubectl version --client
Client Version: version.Info{Major:"1", Minor:"13"...
```

Now that we've installed the `kubectl` client, let's install Minikube.

1. Use Brew to install Minikube.

```
$ brew cask install minikube
==> Downloading https://storage.googleapis.com/minikube...
```

Provide your password if prompted.

2. Use Brew to install the **hyperkit** lightweight hypervisor for Mac.

Other Hypervisor options are available – VirtualBox and VMware Fusion – but we're only showing hyperkit.

```
$ brew install hyperkit
==> Downloading https://homebrew.bintray...
```

3. Start Minikube with the following command.

```
$ minikube start --vm-driver=hyperkit
Starting local Kubernetes cluster...
Starting VM...
```

`minikube start` is the simplest way to start Minikube. Specifying the `--vm-driver=hyperkit` flag will force it to use the **hyperkit** hypervisor instead of VirtualBox.

You now have a Minikube instance up and running on your Mac.

Installing Minikube on Windows 10

In this section we'll show you how to use Minikube on Windows using Hyper-V as the virtual machine manager. Other options exist, but we're not showing them here. We'll also be using a PowerShell terminal opened with Administrative privileges.

Before installing Minikube, let's install the `kubectl` client. There are a couple of ways to do this:

1. Using the Chocolatey package manager
2. Downloading via your web browser

If you are using Chocolatey, you can install it with the following command.

```
> choco install kubernetes-cli
```

If you are not using Chocolatey, you can install `kubectl` using your web browser.

Point your web browser to <https://kubernetes.io/docs/tasks/tools/install-kubectl/> and click the **Install kubectl binary using curl** option. Click the **Windows** tab. Copy and paste the URL into your web browser – this will download the `kubectl` binary. Be sure that you just copy and paste the URL and not the full `curl` command.

Once the download is complete, copy the `kubectl.exe` file to a folder in your system's `%PATH%`.

Verify the installation with a `kubectl version` command.

```
> kubectl version --client=true --output=yaml
clientVersion:
  ...
  gitVersion: v1.13.0
  ...
  major: "1"
  minor: "13"
  platform: windows/amd64
```

Now that you have `kubectl`, you can proceed to install Minikube for Windows.

1. Open a web browser to the Minikube Releases page on GitHub – <https://github.com/kubernetes/minikube/releases>
2. Click `minikube-installer.exe` from beneath the latest version of Minikube. This will download the 64-bit Windows installer.
3. Start the installer and click through the wizard accepting the default options.
4. Make sure Hyper-V has an external vSwitch .

Open Hyper-V Manager (`virtmgmt.msc`) and go to `Virtual Switch Manager...` If there is no Virtual Switch configured with the following two options, create a new one:

- Connection type = External network
- Allow management operating system to share this network adapter

For the remainder of this section we'll assume that you have Hyper-V configured with an external vSwitch called `external`. If yours has a different name, you will need to substitute the name of yours in the following commands.

5. Verify the Minikube version with the following command.

```
> minikube version
minikube version: v0.35.0
```

6. Use the following command to start a local Minikube instance running Kubernetes version 1.12.1.

The command assumes a Hyper-V vSwitch called `external` and uses backticks “`” to allow the command to span multiple lines for readability.

It can take a while to download and start the cluster the first time.


```
> minikube start `
--vm-driver=hyperv `
--hyperv-virtual-switch="external" `
--kubernetes-version="v1.13.0" `
--memory=4096

Starting local Kubernetes v1.13.0 cluster...
Starting VM...
139.09 MB / 139.09 MB [=====] 100.00% 0s
<Snip>
Starting cluster components...
Kubectl is now configured to use the cluster.
```

7. Verify the installation by checking the version of the Kubernetes master.

```
> kubectl version -o yaml
clientVersion:
<Snip>
serverVersion:
  buildDate: 2018-10-05T16:36:14Z
  compiler: gc
  gitCommit: 4ed3216f3ec431b140b1d899130a69fc671678f4
  gitTreeState: clean
  gitVersion: v1.13.0
  goVersion: go1.10.4
  major: "1"
  minor: "13"
  platform: linux/amd64
```

If the target machine actively refuses the network connection with an `Unable to connect to the server: dial tcp... error`, this is most likely a network-related error. Make sure that your vSwitch is configured correctly and that you specified it correctly with the `--hyperv-virtual-switch` flag. `kubectl` talks to Kubernetes inside the `minikube` Hyper-V VM over port 8443.

Congratulations! You've got a fully working Minikube cluster up and running on your Windows 10 PC.

You can now type `minikube` on the command line to see a full list of minikube sub-commands. A good one to try might be `minikube ip` which will give you the IP address that the Minikube cluster is operating on.

Use `kubectl` to verify the Minikube install

The `minikube start` operation configures a *kubectl context* so that you can use `kubectl` against your new Minikube environment. Test this by running the following `kubectl` command from the same shell that you ran `minikube start` from.

```
$ kubectl config current-context
minikube
```

Great, your `kubectl` context is set to Minikube. This means `kubectl` commands will be sent to the Minikube cluster.

It's worth pointing out that `kubectl` can be configured to talk to any Kubernetes cluster by setting different contexts – you just need to switch between contexts to send commands to different clusters.

Use the `kubectl get nodes` command to list the nodes in the cluster.

```
$ kubectl get nodes
NAME          STATUS    AGE   VERSION
minikube      Ready    1m    v1.13.0
```

That's a single-node Minikube cluster ready to use.

You can use the `minikube ip` command to get the IP address of your cluster.

Deleting a Minikube cluster

We spun up the Minikube cluster with a single `minikube start` command. We can stop it with a `minikube stop` command.

```
$ minikube stop
Stopping local Kubernetes cluster...
Machine stopped
```

Stopping a Minikube keeps all the config on disk. This makes it easy to start it up again and pick things up from where you left off.

To blow it away completely – leaving no trace – use the `minikube delete` command.

```
$ minikube delete
Deleting local Kubernetes cluster...
Machine deleted
```

Running a particular version of Kubernetes inside of Minikube

Minikube lets you use the `--kubernetes-version` flag specify the version of Kubernetes you want to run. This is useful if you need to match the version of Kubernetes used in your production environment.

The following command will start a Minikube cluster running Kubernetes version 1.10.7.

```
$ minikube start \
  --kubernetes-version=v1.10.7

Starting local Kubernetes cluster...
Starting VM...
```

Run another `kubectl get nodes` command to verify the version.

```
$ kubectl get nodes
NAME          STATUS    AGE      VERSION
minikube      Ready     1m       v1.10.7
```

That's Minikube. A great way to spin up a simple Kubernetes cluster on your Mac or PC. But it's not for production.

Google Kubernetes Engine (GKE)

Google Kubernetes Engine is a *hosted Kubernetes* service that runs on the Google Cloud Platform (GCP). Like most *hosted Kubernetes* services, it provides:

- A fast and easy way to get a production-grade Kubernetes cluster
- A managed control plane (you do not manage the *masters*)
- Itemized billing

Warning: GKE and other hosted Kubernetes services are not free. Some services might provide a *free tier* or an initial amount of *free credit*. However, generally speaking, you have to pay to use them.

Configuring GKE

To work with GKE you'll need an account on the Google Cloud with billing configured and a blank project. These are all simple to setup, so we won't spend time explaining them here – for the remainder of this section we'll be assuming you have these.

The following steps will walk you through configuring GKE via a web browser. Some of the details might change in the future, but the overall flow will be the same.

1. From within the Console of your Google Cloud Platform (GCP) project, open the navigation pane on the left-hand side and select **Kubernetes Engine > Clusters**. You may have to click the three horizontal bars at the top-left of the Console to make the navigation pane visible.
2. Click the **Create cluster** button.

This will start the wizard to create a new Kubernetes cluster.

3. The wizard currently offers a few templated options. This may change in the future, but the overall flow will be the same. Choose a template (**Your first cluster** or **Standard cluster** will probably be good options to choose from).
4. Give the cluster a meaningful name and description.
5. Choose whether you want a **Regional** or **Zonal** cluster. Regional is newer and potentially more resilient – your masters and nodes will be distributed across multiple zones but still accessible via a single highly-available endpoint.
6. Choose the **Region** or **Zone** for your cluster.
7. Select the **Cluster Version**. This is the version of Kubernetes that will run on your master and nodes. You are limited to the versions available in the drop-down list. Choose an up-to-date version.
8. You can select the number and size of your worker nodes under the **Node pools** section. This allows you to choose the size and configuration of your worker nodes, as well as how many. Larger and faster nodes incur higher costs.

If you are building a Regional cluster, the number you specify will be the number of nodes **in each zone**, not the total number.

9. Leave all other options with default values and click **Create**.

You can also click the **More** link to see a long list of other options you can customize. It's worth looking at them, but we won't be discussing them in this book.

Your cluster will now be created.

Exploring GKE

Now that you have a cluster, it's time to have a quick look at it.

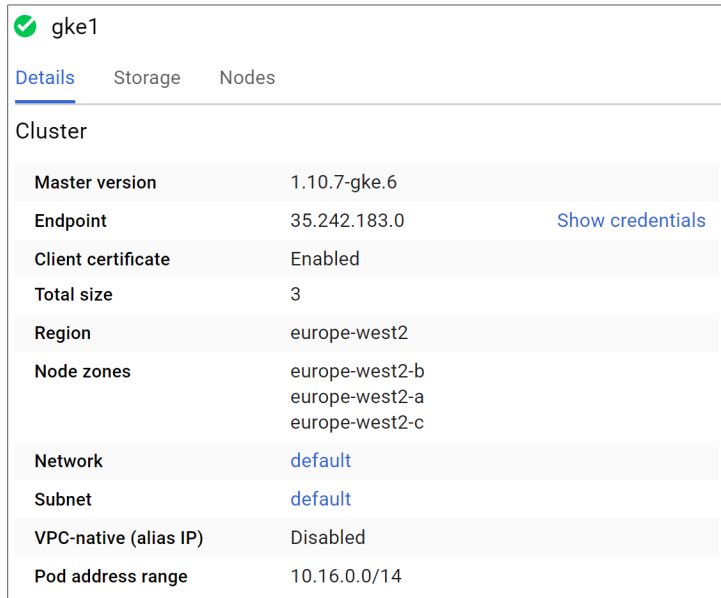
Make sure you're logged on to the GCP Console and are viewing **Clusters** under **Kubernetes Engine**.

The clusters page shows a high-level overview of the Kubernetes clusters you have in your project. Figure 3.1 shows a single 3-node cluster called `gke1`.

Kubernetes clusters							+ CREATE CLUSTER	+ DEPLOY	↻ REFRESH	🗑 DELETE
<input type="checkbox"/>	Name ^	Location	Cluster size	Total cores	Total memory	Notifications	Labels			
<input type="checkbox"/>	✓ gke1	europe-west2	3	3 vCPUs	5.10 GB			Connect	✎	🗑

Figure 3.1

Click the cluster name to drill in to more detail. Figure 3.2 shows a screenshot of some of the detail you can view.



The screenshot shows the 'gke1' cluster details page in the Google Cloud console. The 'Details' tab is selected, showing various cluster configuration parameters.

Cluster	
Master version	1.10.7-gke.6
Endpoint	35.242.183.0 Show credentials
Client certificate	Enabled
Total size	3
Region	europe-west2
Node zones	europe-west2-b europe-west2-a europe-west2-c
Network	default
Subnet	default
VPC-native (alias IP)	Disabled
Pod address range	10.16.0.0/14

Figure 3.2

Clicking the > CONNECT icon towards the top of the web UI (not shown in Figure 3.2) gives you a command you can run on your laptop to configure your local `gcloud` and `kubectl` tools to talk to your cluster. Copy this command to your clipboard.

For the following step to work, you will need to download and install the Google Cloud SDK from <https://cloud.google.com/sdk/>. This will download several utilities, including the `gcloud` and `kubectl` command-line utilities.

Open a terminal and paste the long `gcloud` command into it. This will configure your `kubectl` client to talk to your new GKE cluster.

Run a `kubectl get nodes` command to list the nodes in the cluster.

```
$ kubectl get nodes
NAME                STATUS    AGE     VERSION
gke-cluster-1-pool Ready    5m      v1.12.5-gke.5
gke-cluster-1-pool Ready    6m      v1.12.5-gke.5
gke-cluster-1-pool Ready    6m      v1.12.5-gke.5
```

Congratulations! You know how to create a production-grade Kubernetes cluster using Google Kubernetes Engine (GKE). You also know how to inspect it and connect to it.

Warning! Be sure to delete your GKE cluster as soon as you are finished using it. GKE, and other hosted K8s platforms may incur costs even when they are not in use.

Installing Kubernetes on AWS with kops

kops is short for Kubernetes Operations. It's a *highly-opinionated* cluster bootstrapping tool that makes installing Kubernetes on supported platforms *relatively simple*.

By *highly opinionated*, we mean it's limited in how much you can customize the installation. If you need a highly-customized cluster you should look at kubeadm.

By *relatively simple*, we mean it's easier than compiling the binaries yourself :-D There are still parts that can be complicated if you're not already experienced in those areas. For example, kops is extremely opinionated when it comes to DNS configuration – if you get the DNS wrong you'll be in a world of pain. Fortunately it supports gossip-based installations that don't use DNS. This is intended for development use-cases where the additional hassle of configuring DNS isn't needed.

Kops currently supports bootstrapping a cluster on AWS and GCE. Other platforms might be supported in the future.

At the time of writing, the kops command-line tool is only available on Mac and Linux.

You'll need all of the following to bootstrap a cluster with kops:

- An AWS account and a decent understanding of AWS fundamentals
- `kubectl`
- The latest version of the kops binary for your OS (Mac or Linux)
- The `awscli` tool
- The credentials of an AWS account with the following permissions:
 - `AmazonEC2FullAccess`
 - `AmazonRoute53FullAccess`
 - `AmazonS3FullAccess`
 - `IAMFullAccess`
 - `AmazonVPCFullAccess`

The following examples are from a Linux machine, but it works the same on a Mac (and possibly Windows in the future).

The following examples show both installation options:

1. DNS
2. Gossip

The gossip-based installation is the simplest and is ideal for situations where private DNS domains are not available. It's also ideal for AWS locations, such as China, where Route53 isn't available.

The DNS installation is more involved and requires a top-level-domain, as well as a sub-domain delegated to AWS Route53. The DNS examples in this chapter use a domain called `tf1.com` that is hosted with a 3rd party provider such as GoDaddy. It has a subdomain called `k8s` that is delegated to Amazon Route53. If you're following along with the DNS examples you will need your own working domains.

Download and install kubectl

For Mac, the download and installation is a simple `brew install kubernetes-cli`.

The following procedure is for a Linux machine.

1. Use the following command to download the latest kubectl binary to your home directory.

```
$ curl -LO https://storage.googleapis.com/kubernetes-release/release/$(curl -s
https://storage.googleapis.com/kubernetes-release/release/stable.txt)/bin/linux
/amd64/kubectl
```

The command is a single command, but is quite long and will wrap over multiple lines in the book. This process might introduce backslashes at the edge of the printed page that are not part of the command and will need to be removed.

2. Make the downloaded binary executable and move it to a directory in your PATH.

```
$ chmod +x ./kubectl
$ mv ./kubectl /usr/local/bin/kubectl
```

Run a kubectl command to make sure it's installed and working.

Download and install kops

For Mac, you just need to run `brew install kops`.

Use the following procedure for Linux.

1. Download the kops binary with the following `curl` command.

The command should be issued on one line and have no backslashes \ in it. It also has the version of the kops tool embedded in the URL, you can change this. See <https://github.com/kubernetes/kops/releases> for the latest versions.

```
$ curl -LO https://github.com/kubernetes/kops/releases/download/1.11.1/kops-linux-amd64
```

2. Make the downloaded binary executable and move it to a directory in your system's PATH.

```
$ chmod +x kops-linux-amd64
$ mv kops-linux-amd64 /usr/local/bin/kops
```

Run a `kops version` command to verify the installation.

```
$ kops version
Version 1.11.1
```

Install and configure the AWS CLI

You can install the AWS CLI tool on Mac OS using the `brew install awscli`.

The following example shows how to install the AWS CLI from the default app repos used by Ubuntu 18.04. The installation method will obviously be different if you're using a different Linux distro.

1. Run the following command to install the AWS CLI

```
$ sudo apt-get install awscli -y
```

2. Run the `aws configure` command to configure your instance of the AWS CLI

You will need the credentials of an AWS IAM account with *AmazonEC2FullAccess*, *AmazonRoute53FullAccess*, *AmazonS3FullAccess*, *IAMFullAccess*, and *AmazonVPCFullAccess* to complete this step.

```
$ aws configure
AWS Access Key ID [None]: *****
AWS Secret Access Key [None]: *****
Default region name [None]: enter-your-region-here
Default output format [None]:
```

3. Create a new S3 bucket for kops to store configuration and state information.

Kops requires cluster names to be valid DNS names. We'll use the name `cluster1.k8s.tf1.com` in these examples. You will have to use a different name in your environment. Let's quickly break-down how it works. The example assumes I own a domain called `tf1.com`, and that I've delegated a sub-domain called `k8s` to AWS Route53. Within that sub-domain I can create clusters with whatever names I like. In the example, we'll create a cluster called `cluster1`. This will make the fully-qualified domain name for the cluster `cluster1.k8s.tf1.com`.

I've created NS records in the parent `tf1.com` domain to point to the `k8s` hosted domain in Route53. `tf1.com` is fictional and only being used in these examples to keep the command-line arguments short.

If you plan on creating a gossip-based cluster, you will need to use a cluster name that ends with `.k8s.local`.

```
$ aws s3 mb s3://cluster1.k8s.tf1.com
make_bucket: cluster1.k8s.tf1.com
```

4. List your S3 buckets and `grep` for the name of the bucket you created. This will prove that the bucket created successfully.

```
$ aws s3 ls | grep k8s
2018-10-10 13:09:11 cluster1.k8s.tf1.com
```

5. Tell **kops** where to find its config and state – this will be the S3 bucket created in the previous step.

```
$ export KOPS_STATE_STORE=s3://cluster1.k8s.tf1.com
```

6. Create a new cluster with one of the following `kops create cluster` commands.

The first command creates the cluster using gossip instead of DNS. To work with gossip the cluster name **must** end with `.k8s.local`.

The second command creates the cluster with DNS and assumes a working DNS configuration as previously explained.

You will need a copy of your AWS public key for the command to work. In the examples, the key is called `np-k8s.pub` and is in the current working directory.

```
$ kops create cluster \
  --cloud-aws \
  --zones=eu-west-1b \
  --name=mycluster.k8s.local \
  --ssh-public-key ~/np-k8s.pub \
  --yes
```

```
$ kops create cluster \
  --cloud=aws \
  --zones=eu-west-1b \
  --dns-zone=k8s.tf1.com \
  --name cluster1.k8s.tf1.com \
  --ssh-public-key ~/np-k8s.pub \
  --yes
```

The command is broken down as follows. `kops create cluster` tells **kops** to create a new cluster. `--cloud=aws` tells it to create the cluster in AWS using the AWS provider. `--zones=eu-west-1b` tells **kops** to create the cluster in the eu-west-1b zone. If creating the cluster with DNS, the `--dns-zone` flag tells it to use the delegated zone. We name the cluster with the `--name` flag – remember to end your cluster name with “k8s.local” if creating with gossip. `--ssh-public-key` tells it which key to use. Finally, the `--yes` flag tells **kops** to go ahead and deploy the cluster. If you omit the `--yes` flag, a cluster config will be created but it will not be deployed.

It may take a few minutes for the cluster to deploy. This is because **kops** is creating the AWS resources required to build the cluster. This includes things like a VPC, EC2 instances, launch configs, auto scaling groups, security groups etc. After it has built the AWS infrastructure, it also has to build the Kubernetes cluster.

7. Once the cluster is deployed you can validate it with the `kops validate cluster` command. It may take a while for the cluster to completely come up, so be patient.

```
$ kops validate cluster
Using cluster from kubectl context: cluster1.k8s.tf1.com
```

INSTANCE GROUPS

NAME	ROLE	MACHINETYPE	MIN	MAX	SUBNETS
master..	Master	m3.medium	1	1	eu-west-1b
nodes	Node	t2.medium	2	2	eu-west-1b

NODE STATUS

NAME	ROLE	READY
ip-172-20-38..	node	True
ip-172-20-58..	master	True
ip-172-20-59..	node	True

Your cluster cluster1.k8s.tf1.com is ready

Congratulations! You now know how to create a Kubernetes cluster in AWS using the `kops` tool.

Now that your cluster is up and running you can issue `kubectl` commands against it. It might also be worth having a poke around in the AWS console to see some of the resources that `kops` created.

Warning! Be sure to delete your cluster when you're finished using it. Clusters running on cloud platforms may incur costs even when they are not actively being used.

Deleting a Kubernetes cluster in AWS with `kops`

You can use the `kops delete cluster` command to delete the cluster you just created. This will also delete all of the AWS resources that were created to support the cluster.

The following command will delete the cluster created in the previous steps.

```
$ kops delete cluster --name=cluster1.k8s.tf1.com --yes
```

Installing Kubernetes with `kubeadm`

In this section, we'll see how to install Kubernetes using `kubeadm`.

One of the best things about `kubeadm` is that you can use it to install Kubernetes nearly anywhere – laptop, bare metal in your data center, even on public clouds. It also does a lot more than just install Kubernetes. You can upgrade, manage, and query your clusters too. It's often said that `kubeadm` is `kubectl` for clusters – a great tool for building **and** managing Kubernetes clusters. Anyway, `kubeadm` is a core Kubernetes project and has a promising future.

The examples in this section are based on Ubuntu 18.04. Some of the commands in the pre-reqs section will be different if you're using a different Linux distro. However, the procedure we're showing can be used to install Kubernetes on your laptop, in your data center, or even in the cloud.

We'll be walking through a simple example using three Ubuntu 18.04 machines configured as one master and two nodes as shown in Figure 3.3.

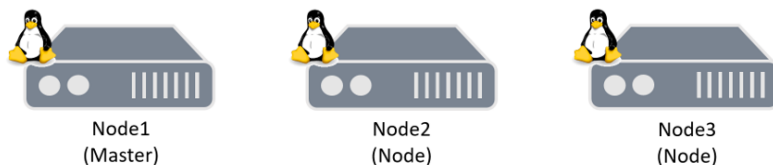


Figure 3.3

The high-level plan will be as follows:

1. Install the pre-requisites

2. Initialize a new cluster with **node1** as the master
3. Create the Pod network
4. Add **node2** and **node3** as worker nodes.

All three nodes will get the following:

- Docker
- kubeadm
- kubelet
- kubectl

Docker is the container runtime. Other runtimes exist, but we'll go with Docker. **kubeadm** is the tool we'll use to build the cluster, **kubelet** is the Kubernetes node agent, and **kubectl** is the Kubernetes command-line utility.

Pre-requisites

The following commands are specific to Ubuntu 18.04 and need to be ran on **all three nodes**. They set things up so that we can install the right packages from the right repos. Equivalent commands and packages exist for other flavors of Linux.

Use the following two commands to get the latest versions of a few packages that will be required in later steps.

```
$ sudo apt-get update  
<Snip>
```

```
$ sudo apt-get install -y \  
  apt-transport-https \  
  ca-certificates \  
  curl \  
  software-properties-common
```

Download and install the following two repository keys. One of the repositories has the Kubernetes tools and the other has Docker. We'll need these keys in later steps.

```
$ curl -s \
https://packages.cloud.google.com/apt/doc/apt-key.gpg | sudo apt-key add -
OK
```

```
$ curl -fsSL \
https://download.docker.com/linux/ubuntu/gpg | sudo apt-key add -
OK
```

Create, or edit, the following file and add the repo required to install the Kubernetes packages.

```
$ sudo vim /etc/apt/sources.list.d/kubernetes.list
```

Add the following line.

```
deb https://apt.kubernetes.io/ kubernetes-xenial main
```

The next step is to install kubectl, kubeadm, and the kubelet. Use the following two commands.

```
$ sudo apt-get update
<Snip>
```

```
$ sudo apt-get install -y kubelet kubeadm kubectl
<Snip>
```

If you run the `apt-get install` command again you can see the versions that were installed.

Now let's install Docker..

Add the required fingerprint.

```
$ sudo apt-key fingerprint 0EBFCD88
pub  rsa4096 2017-02-22 [SCEA]
      9DC8 5822 9FC7 DD38 854A  E2D8 8D81 803C 0EBF CD88
uid          [ unknown] Docker Release (CE deb) <docker@docker.com>
sub  rsa4096 2017-02-22 [S]
```

Now add the *stable* Docker repository. This is a single command with backslashes used to spread it over multiple lines.

```
$ sudo add-apt-repository \
    "deb [arch=amd64] https://download.docker.com/linux/ubuntu \
    $(lsb_release -cs) \
    stable"
```

Install Docker.

```
$ sudo apt-get update
<Snip>
```

```
$ sudo apt-get install docker-ce
<Snip>
```

That's the pre-reqs done.

Initialize a new cluster

Initializing a new Kubernetes cluster with `kubeadm` is as simple as typing `kubeadm init`.

```
$ sudo kubeadm init
<SNIP>
Your Kubernetes master has initialized successfully!
```

To start using your cluster, you need to run (as a regular user):

```
mkdir -p $HOME/.kube
sudo cp -i /etc/kubernetes/admin.conf $HOME/.kube/config
sudo chown $(id -u):$(id -g) $HOME/.kube/config
```

```
<SNIP>
You can join any number of machines by running the following...
```

```
kubeadm join --token b90685.bd53aca93b758efc 172.31.32.74:6443
```

The command pulls all required images and builds the cluster. When the process is complete, it spits out a few short commands that enable you to manage the cluster as a regular user. It also gives you the `kubeadm join` command that will let you add additional nodes to the cluster.

Congratulations! That's a brand-new single-master Kubernetes cluster.

Complete the process by running the commands listed in the output of the `kubeadm init`.

```
$ mkdir -p $HOME/.kube
$ sudo cp -i /etc/kubernetes/admin.conf $HOME/.kube/config
$ sudo chown $(id -u):$(id -g) $HOME/.kube/config
```

These commands may be different, or even no longer required in the future. However, they copy the Kubernetes config file from `/etc/kubernetes` into your home directory and change the ownership to you.

Use `kubectl` to verify that the cluster initialized successfully.

```
$ kubectl get nodes
NAME      STATUS      ROLES    AGE      VERSION
node1     NotReady    master   2m43s    v1.13.4
```

Run the following `kubectl` command to find the reason why the cluster STATUS is showing as `NotReady`.

```
$ kubectl get pods --all-namespaces
NAMESPACE   NAME             READY   STATUS              RESTARTS   AGE
kube-system  coredns-...vt    0/1     ContainerCreating   0           8m33s
kube-system  coredns-...xw    0/1     ContainerCreating   0           8m33s
kube-system  etcd-...         1/1     Running             0           7m46s
kube-system  kube-api-...     1/1     Running             0           7m36s
...
```

This command shows all Pods in all namespaces, including system Pods in the system (`kube-system`) namespace.

As we can see, none of the `coredns` Pods are running. This is preventing the cluster from entering the Ready state and is happening because we haven't created the Pod network yet.

Create the Pod network. The following example creates a multi-host overlay network provided by Weaveworks. Other options exist, and you do not have to go with the example shown here.

The command may wrap over multiple lines in the book. Any backslashes (`\`) at the edge of the printed page should be removed.

```
$ kubectl apply -n kube-system -f \
  "https://cloud.weave.works/k8s/net?k8s-version=$(kubectl version | base64 | \
  tr -d '\n')"
```

Check if the status of the master has changed from `NotReady` to `Ready`.

```
$ kubectl get nodes
NAME      STATUS    ROLES    AGE      VERSION
node1     Ready     master   3m51s    v1.13.4
```

Great, the cluster is ready and the DNS Pods will now be running.

Now that the cluster is up and running, it's time to add some nodes.

Adding worker nodes requires the cluster's join token. You might remember that this was provided as part of the output when the cluster was first initialized. Scroll back up to that output, copy the `kubeadm join` command to the clipboard and then run it on **node2** and **node3**.

Note: The following must be performed on **node2** and **node3** and you must have already installed the pre-reqs (Docker, kubeadm, kubectl, and the kubelet) on these nodes.

```
node2$ kubeadm join 172.31.32.74:6443 --token b90...
<SNIP>
Node join complete:
* Certificate signing request sent to master and response received
* Kubelet informed of new secure connection details.
```

Repeat the command on **node3**.

Make sure that both nodes successfully registered by running another `kubectl get nodes` on the master.

```
$ kubectl get nodes
NAME      STATUS    ROLES    AGE      VERSION
node1     Ready     master   10m      v1.13.4
node2     Ready     master   55s      v1.13.4
node3     Ready     <none>   34s      v1.13.4
```

Congratulations! You've manually built a 3-node cluster using kubeadm. But remember that it's running a single master without H/A.

Feel free to poke around the cluster with kubeadm. You should also investigate ways kubeadm can install clusters with H/A managers.

kubectl

`kubectl` is the main Kubernetes command-line tool and is what you should use for your Kubernetes management activities. In fact, it's useful to think of `kubectl` as *SSH for Kubernetes*. It's available for Linux, Mac and Windows.

As it's the main command-line tool, it's important that you use a version that is no more than one minor version higher or lower than your cluster. For example, if your cluster is running Kubernetes 1.13.x, your `kubectl` should be between 1.12.x and 1.14.x.

At a high-level, `kubectl` converts user-friendly commands into the JSON payload required by the API server. It uses a configuration file to know which cluster and API server endpoint to POST to.

By default, the `kubectl` configuration file is called `config` and lives in `$HOME/.kube`. It contains definitions for:

- Clusters
- Users
- Contexts

Clusters lets define multiple clusters and is ideal if you plan on using a single workstation to manage multiple clusters. Each cluster definition has a name, certificate info, and API server endpoint.

Users let you define different users that might have different level of permissions on each cluster. For example, you might have a *dev* user and an *ops* user, each with different permissions. Each *user* definition has a friendly name, a username, and a set of credentials.

Contexts bring together clusters and users under a friendly name. For example, you might have a context called `deploy-prod` that combines the `deploy` user credentials with the `prod` cluster definition. If you use `kubectl` with this context you will be POSTing commands to the API server of the `prod` cluster as the `deploy` user.

The following is a simple `kubectl` config file with a single cluster called `minikube`, a single user called `minikube`, and a single context called `minikube`. The `minikube` context combines the `minikube` user and cluster, and is also set as the default context.

```
apiVersion: v1
clusters:
- cluster:
    certificate-authority: C:\Users\nigel\.minikube\ca.crt
    server: https://192.168.1.77:8443
  name: minikube
contexts:
- context:
    cluster: minikube
    user: minikube
  name: minikube
current-context: minikube
kind: Config
preferences: {}
```

```
users:
- name: minikube
  user:
    client-certificate: C:\Users\nigel\.minikube\client.crt
    client-key: C:\Users\nigel\.minikube\client.key
```

You can view your `kubectl` config using the `kubectl config view` command. Sensitive data will be redacted from the output.

You can use `kubectl config current-context` to see your current context. The following example shows a system where `kubectl` is configured to issue commands to a cluster that is called `eks-k8sbook` in `$HOME/.kube/config`.

```
$ kubectl config current-context
eks_k8sbook
```

You can change the current/active context with `kubectl config use-context`. The following command will set the current context to `docker-desktop` so that future commands will be sent to the cluster defined in the `docker-desktop` context. It obviously requires that a context called `docker-desktop` exists in the `kubectl` config file.

```
$ kubectl config use-context docker-desktop
Switched to context "docker-desktop".
```

```
$ kubectl config current-context
docker-desktop
```

Chapter summary

In this chapter, we learned how to install Kubernetes in a few different ways on a few different platforms.

We saw how fast and simple it is to setup a Kubernetes cluster on Play with Kubernetes (PWK). We got a 4-hour playground without having to install anything on our laptop or in our own cloud.

We setup Docker Desktop and Minikube for a great developer experience on our laptops.

We learned how to spin up a managed/hosted Kubernetes cluster in the Google Cloud using Google Kubernetes Engine (GKE).

Then we looked at how to use the `kops` tool to spin up a cluster in AWS using the AWS provider.

We finished the chapter seeing how to perform a manual install using the `kubeadm` tool.

There are other ways and places we can install Kubernetes. But the chapter is already long enough and I've pulled out way too much of my hair already :-D

4: Working with Pods

We'll split this chapter in to two main parts:

- Theory
- Hands-on

Let's crack on with the theory.

Pod theory

The atomic unit of scheduling in the virtualization world is the Virtual Machine (VM). This means **deploying applications** in the virtualization world means scheduling them on VMs.

In the Docker world, the atomic unit is the container. This means **deploying applications** on Docker means deploying them inside of containers.

In the Kubernetes world, the atomic unit is the *Pod*. Ergo, **deploying applications** on Kubernetes means stamping them out in Pods.

This is fundamental to understanding Kubernetes, so be sure to tag it in your brain as important >> Virtualization does VMs, Docker does containers, and **Kubernetes does Pods**.

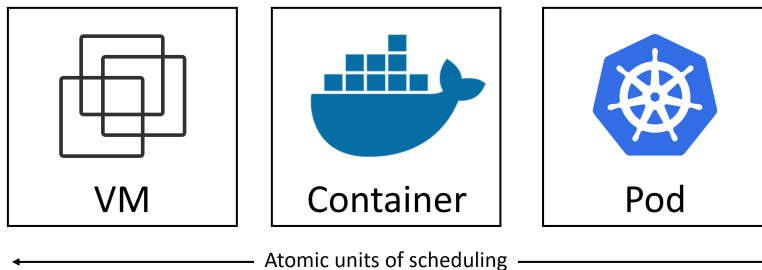


Figure 4.1

As Pods are the fundamental unit of deployment in Kubernetes, it's vital we understand how they work.

Note: We're going to talk a lot about Pods in this chapter. However, it's important to remember that Pods are just a vehicle for **deploying applications**.

Pods vs containers

In a previous chapter we said that a Pod hosts one or more containers. From a footprint perspective, this puts Pods somewhere in between containers and VMs – they’re bigger than a container, but a lot smaller than a VM.

Digging a bit deeper, a Pod is a shared execution environment for one or more containers. Quite often it’s one container per Pod, but multi-container Pods are gaining in popularity. One use-case for multi-container Pods is co-scheduling tightly-coupled workloads. For example, two containers that share memory wouldn’t work if they were scheduled on different nodes in the cluster. Other increasingly common use-cases include logging and service meshes.

Pods: the canonical example

A common example for comparing single-container and multi-container Pods is a web server that utilizes a file synchronizer.

In this example we have two clear *concerns*:

1. Serving the web page
2. Making sure the content is up-to-date

In our context, a *concern* is a requirement or a task. Generally speaking, microservices design patterns dictate that we *separate concerns*. This means one concern per container. Assuming the previous example, that would require one container for the web service, and another container for the file-sync service.

This model of separating concerns has a lot of advantages.

Instead of building a monolithic container – where a single container runs the web service *and* file-sync service – we build two containers. One container does the web serving, the other does the file synchronizing. Some of the advantages this brings include:

- Different teams can be responsible for each of the two concerns
- Each can be scaled independently
- Each can be developed and iterated independently
- Each can have its own release cadence
- If one fails, the other keeps running

Despite the benefits of separating concerns, there are situations where it makes sense to co-schedule multiple containers in a single Pod. Use-cases include; two containers that need to share memory or share a volume (see Figure 4.2). An increasingly popular use-case is a service mesh where a second container is inserted into every Pod to provide things like network proxying.

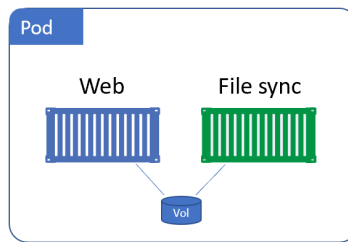


Figure 4.2

Figure 4.2 depicts a scenario where a single Pod is responsible for two concerns. The simplest way to implement a shared volume is to schedule the containers that will be sharing the volume on the same node. By running the web service container and the file-sync container in the same Pod, we ensure they will always be deployed to the same node. We also give them a shared operating environment where both can access the same shared memory and shared volumes etc. More on this later.

In summary, the general rule is to separate concerns by designing containers do a single job, and then scheduling a single container per Pod. However, there are use-cases where breaking this rule has advantages.

How do we deploy Pods

To deploy a Pod to a Kubernetes cluster we define it in a *manifest file* and POST that manifest file to the API server. The control plane examines it, writes it to the cluster store as a record of intent, and the scheduler deploys it to a healthy node with enough available resources. This process is identical for single-container Pods and multi-container Pods.

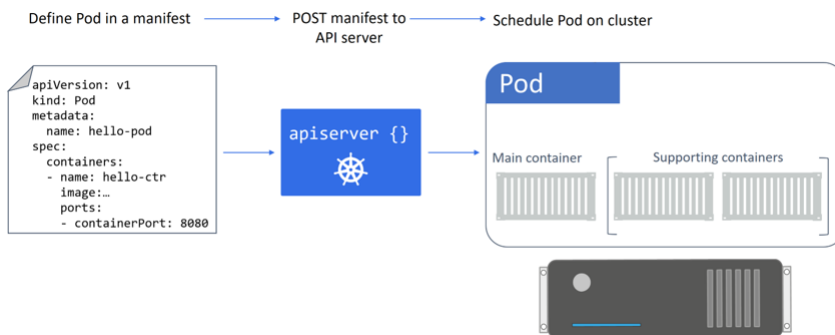


Figure 4.3

Let's dig a bit deeper...

The anatomy of a Pod

At the highest level, a Pod is a shared execution environment for one or more containers. *Shared execution environment* means that the Pod has a set of resources that are shared by every container that is part of the Pod. These resources include; IP addresses, ports, hostname, sockets, memory, volumes, and more...

If you're using Docker as the container runtime, a Pod is actually a special type of container called a **pause container**. That's right, a Pod is just a fancy name for a special container. This means containers running inside of Pods are really containers running inside of containers. For more information, watch "Inception" by Christopher Nolan, starring Leonardo DiCaprio ;-)

Seriously though, the Pod (pause container) is just a collection of system resources that containers running inside of it will inherit and share. These system resources are kernel namespaces and include:

- **Network namespace:** IP address, port range, routing table...
- **UTS namespace:** Hostname
- **IPC namespace:** Unix domain sockets...

As we just mentioned, this means that all containers in a Pod share a hostname, IP address, memory address space, and volumes.

Let's look at how this affects networking.

Each Pod creates its own network namespace. This includes; a single IP address, a single range of TCP and UDP ports, and a single routing table. This is true even if the Pod is a multi-container Pod – each container in a Pod shares the Pod's; IP, range of ports, and routing table.

Figure 4.4 shows two Pods, each with its own IP. Even though one of them is a multi-container Pod, it still only gets a single IP.

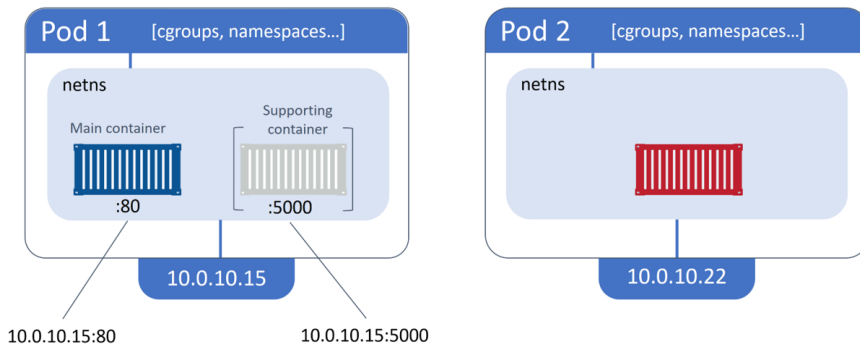


Figure 4.4

In the example shown in Figure 4.4, external access to the containers in Pod 1 is achieved via the IP address of the Pod coupled with the port of the container you wish to reach. For example,

`10.0.10.15:80` would get you to the main container. Container-to-container communication works via the Pod's localhost adapter. For example, the main container can reach the supporting container via `localhost:5000`.

One last time (apologies if it feels like I'm over-repeating myself)... Each container in a Pod shares the Pod's entire network namespace – IP, localhost adapter, port range, routing table, and more.

However, as we've already said, it's more than just networking. All containers in a Pod have access to the same volumes, the same memory, the same IPC sockets, and more. Technically speaking, the Pod (pause container) holds all the namespaces, any containers that are part of the Pod inherit them and share them.

This networking model makes *inter-Pod* communication really simple. Every Pod in the cluster has its own IP addresses that's fully routable on the *Pod network*. If you read the chapter on installing Kubernetes, you'll have seen how we created a Pod network at the end of the *Play with Kubernetes*, and *kubeadm* sections. Because every Pod gets its own routable IP, every Pod on the Pod network can talk directly to every other Pod without messing around with things like nasty port mappings.



Figure 4.5 Inter-Pod communication

As previously mentioned, *intra-Pod* communication – where two containers in the same Pod need to communicate – can happen via the Pod's localhost interface.

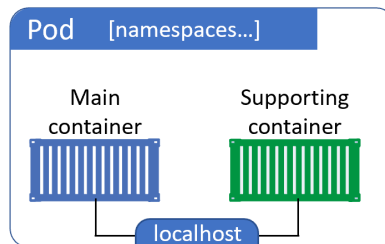


Figure 4.6 Intra-Pod communication

If you need to make multiple containers in the same Pod available to the outside world, you can expose them on individual ports. Each container needs its own port, and two containers in the same Pod cannot use the same port.

In summary. It's all about the **Pod**! The **Pod** gets deployed, the **Pod** gets the IP, the **Pod** owns all of the namespaces... The **Pod** is at the center of the Kuberverse.

Pods and cgroups

At a high level, Control Groups (cgroups) prevent individual containers from consuming all of the available CPU, RAM and IOPS on a node. We could say that cgroups actively *police* resource usage.

Individual containers have their own cgroup limits.

This means it's possible for two containers in the same Pod to have their own set of cgroup limits. This is a powerful and flexible model. If we assume the canonical multi-container Pod example from earlier in the chapter, we could set a cgroup limit on the file sync container so that it has access to less resources than the web service container. This might reduce the risk of it starving the web service container of CPU and memory.

Atomic deployment of Pods

Deploying a Pod is an *atomic operation*. This means it's an all-or-nothing operation – there's no such thing as a partially deployed Pod that can service requests. It also means that all containers in a Pod will be scheduled on the same node.

Once all Pod resources are ready, the Pod becomes available.

Pod lifecycle

The lifecycle of a typical Pod goes something like this. You define it in a YAML manifest file and POST the manifest to the API server. Once there, the contents of the manifest are persisted to the cluster store as a record of intent (desired state), and the Pod is scheduled to a healthy node with enough resources. Once it's scheduled to a node, it enters the *pending* state while the node downloads images and starts any containers. The Pod remains in this *pending* state until **all of its resources** are up and ready. Once everything's up and ready, the Pod enters the *running* state. Once it has completed all of its tasks, it gets terminated and enters the *succeeded* state.

When a Pod can't start, it can remain in the *pending* state or go to the *failed* state. This is all shown in Figure 4.7.

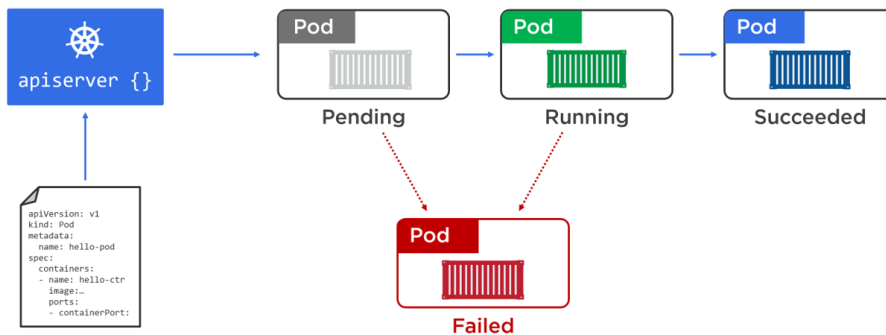


Figure 4.7 Pod lifecycle

Pods that are deployed via Pod manifest files are *singletons* – they are not replicated and have no self-healing capabilities. For this reason, we almost always deploy Pods via higher-level objects like *Deployments* and *DaemonSets*, as these can reschedule Pods when they fail.

On that topic, it's important to think of Pods as *mortal*. When they die, they're gone. There's no bringing them back from the dead. This follows the *pets vs cattle* analogy, and Pods should be treated as *cattle*. When they die, you replace them with another. There's no tears and no funeral. The old one is gone, and a shiny new one -- with the same config, but a different ID and IP -- magically appears and takes its place.

This is one of the main reasons you should code your applications so that they don't store *state* in Pods. It's also why we shouldn't rely on individual Pod IPs. Singleton Pods are not reliable!

Note: No offense is intended to any person or any animal when referring to the *pets* and *cattle*.

Pod theory summary

1. Pods are the atomic unit of scheduling in Kubernetes
2. You can have more than one container in a Pod. Single-container Pods are the simplest, but multi-container Pods are ideal for containers that need to be tightly coupled. They're also great for logging and service meshes
3. Pods get scheduled on nodes – you can't schedule a single Pod instance to span multiple nodes
4. Pods are defined declaratively in a manifest file that is POSTed to the API server and assigned to nodes by the scheduler
5. We almost always deploy Pods via higher-level objects

Hands-on with Pods

It's time to see Pods in action.

For the examples in the rest of this chapter we'll use the 3-node cluster shown in Figure 4.8.

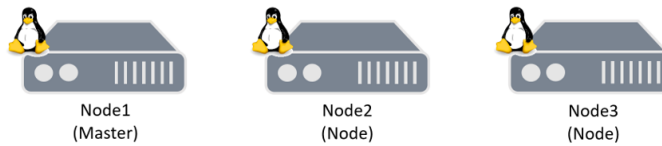


Figure 4.8

It doesn't matter where this cluster is or how it was deployed. All that matters is that you have three Linux hosts configured into a Kubernetes cluster with at least one master and two nodes. You'll also need `kubectl` installed and configured to talk to the cluster.

If you do not have a cluster but would like to follow along, go to <http://play-with-k8s.com> and build a quick cluster. It's free and easy.

Following the Kubernetes mantra of *composable infrastructure*, we define Pods in manifest files, POST these to the API server, and let the scheduler instantiate them on the cluster.

Pod manifest files

For the examples in this chapter we're going to use the following Pod manifest. It's available in the book's GitHub repo under the `pods` folder called `pod.yml`:

```
apiVersion: v1
kind: Pod
metadata:
  name: hello-pod
  labels:
    zone: prod
    version: v1
spec:
  containers:
  - name: hello-ctr
    image: nigelpoulton/k8sbook:latest
    ports:
    - containerPort: 8080
```

Let's step through what the YAML file is describing.

Straight away we can see four top-level resources:

- `apiVersion`

- `kind`
- `metadata`
- `spec`

The `.apiVersion` field tells us two things – the *API group* and the *API version* that will be used to create the object. Normally the format is `<api-group>/<version>`. However, Pods are defined in a special API group called the *core* group which omits the *api-group* part. For example, `StorageClass` objects are defined in `v1` of the `storage.k8s.io` API group and are described in YAML files as `storage.k8s.io/v1`. However, Pods are in the *core* API group which is special, as it omits the API group name, so we describe them in YAML files as just `v1`.

It's possible for a resource to be defined in multiple versions of an API group. For example, `some-api-group/v1` and `some-api-group/v2`. In this case, the definition in the newer group would probably include additional features and fields that extend the capabilities of the resource. Think of the *version* field as defining the schema – newer is usually better. Interestingly, there may be occasions where you deploy an object via one version in the YAML file, but when you introspect it, the return values show it as another version. For example, you may deploy an object by specifying `v1` in the YAML file, but when you run commands against it the returns might show it as `v1beta1`. This is normal behavior.

Anyway, Pods are currently defined at the `v1` path.

The `.kind` field tells Kubernetes the type of object being deployed.

So far, we know we're deploying a Pod object as defined in `v1` of the *core API group*.

The `.metadata` section is where we attach a name and labels. These help us identify the object in the cluster, and labels help us create loose couplings. We can also define the `namespace` that an object should be deployed to. Keeping things brief, namespaces allow us to logically divide clusters for management purposes. In the real world, it's highly recommended to use namespaces, however, you should not think of them as strong security boundaries.

The `.metadata` section of this Pod manifest is naming the Pod “hello-pod” and assigning it two labels. Labels are simple key-value pairs, but they're insanely powerful. We'll talk more about labels later as we build our knowledge.

As the `.metadata` section does not specify a namespace, the `default` namespace is assumed. It's not good practice to use the default namespace in the real world.

The `.spec` section is where we define any containers that will run in the Pod. Our example is deploying a Pod with a single container based on the `nigelpoulton/k8sbook:latest` image. It's calling the container `hello-ctr` and exposing it on port `8080`.

If this was a multi-container Pod, we'd define additional containers in the `.spec` section.

Manifest files: Empathy as Code

Quick side-step.

Configuration files, like Kubernetes manifest files, are excellent sources of documentation. As such, they have some secondary benefits. Two of these include:

- Speeding-up the on-boarding process for new team members
- Bridging the gap between developers and operations

For example, if you need a new team member to understand the basic functions and requirements of an application, get them to read the application’s Kubernetes manifest files.

Also, if your operations teams complain that developers don’t give accurate application requirements and documentation, make your developers use Kubernetes. Kubernetes forces developers to describe their applications through Kubernetes manifests, which can then be used by operations staff to understand how the application works and what it requires from the environment.

These kinds of benefits were described as a form of *empathy as code* by Nirmal Mehta in his 2017 DockerCon talk entitled “A Strong Belief, Loosely Held: Bringing Empathy to IT”.

I understand that describing YAML files like these as “*empathy as code*” sounds a bit extreme. However, there is merit to the concept – they definitely help.

Back to business...

Deploying Pods from a manifest file

If you’re following along with the examples, save the manifest file as `pod.yml` in your current directory and then use the following `kubectl` command to POST the manifest to the API server.

```
$ kubectl apply -f pod.yml
pod/hello-pod created
```

Although the Pod is showing as created, it might not be fully deployed and available yet. This is because it takes time to pull the image.

Run a `kubectl get pods` command to check the status.

```
$ kubectl get pods
```

NAME	READY	STATUS	RESTARTS	AGE
hello-pod	0/1	ContainerCreating	0	9s

We can see that the container is still being created – no doubt waiting for the image to be pulled from Docker Hub.

You can add the `--watch` flag to the `kubectl get pods` command so that you can monitor it and see when the status changes to running.

Congratulations! Your Pod has been scheduled to a healthy node in the cluster and is being monitored by the local `kubelet` process. The `kubelet` process is the Kubernetes agent running on the node.

In future chapters, we’ll see how to connect to the web server running in the Pod.

Introspecting running Pods

As good as the `kubectl get pods` command is, it's a bit light on detail. Not to worry though, there's plenty of options for deeper introspection.

First up, the `kubectl get` command offers a couple of really simple flags that give you more information:

The `-o wide` flag gives a couple more columns but is still a single line of output.

The `-o yaml` flag takes things to the next level. It returns a full copy of the Pod manifest from the cluster store. The output is broadly divided into two parts:

- desired state (`.spec`)
- current observed state (`.status`)

The following command shows a snipped version of a `kubectl get pods -o yaml` command.

```
$ kubectl get pods hello-pod -o yaml
apiVersion: v1
kind: Pod
metadata:
  annotations:
    kubectl.kubernetes.io/last-applied-configuration: |
      ...
  name: hello-pod
  namespace: default
spec: #Desired state
  containers:
  - image: nigelpoulton/k8sbook:latest
    imagePullPolicy: Always
    name: hello-ctr
    ports:
status: #Observed state
  conditions:
  - lastProbeTime: null
    lastTransitionTime: 2018-10-19T15:24:24Z
  state:
    running:
      startedAt: 2018-10-19T15:26:04Z
  ...
```

Notice how the output contains more values than we initially set in our 13-line YAML file. Where does this extra information come from?

Two main sources:

- The Kubernetes Pod object has far more than we defined in the manifest. Those that we don't set explicitly are automatically expanded with default values by Kubernetes.
- When you run a `kubectl get pods` with `-o yaml` you get the Pods *current observed state* as well as its *desired state*. This observed state is listed in the `.status` section.

Another great Kubernetes introspection command is `kubectl describe`. This provides a nicely formatted multi-line overview of an object. It even includes some important object lifecycle events. The following command describes the state of the `hello-pod` Pod.

```
$ kubectl describe pods hello-pod
Name:          hello-pod
Namespace:     default
Node:          docker-for-desktop/192.168.65.3
Start Time:    Fri, 19 Oct 2018 16:24:24 +0100
Labels:        version=v1
               zone=prod
Status:        Running
IP:            10.1.0.21
Containers:
  hello-ctr:
    Image:      nigelpoulton/k8sbook:latest
    Port:       8080/TCP
    Host Port:  0/TCP
    State:      Running
Conditions:
  Type           Status
  Initialized    True
  Ready          True
  PodScheduled   True
...
Events:
  Type    Reason      Age    Message
  ----    -
  Normal  Scheduled   2m     Successfully assigned...
  Normal  Pulling     2m     pulling image "nigelpoulton/k8sbook:latest"
  Normal  Pulled      2m     Successfully pulled image
```

Normal	Created	2m	Created container
Normal	Started	2m	Started container

The output has been snipped to help it fit the book.

Another way to introspect a running Pod is to log into it or execute commands in it. We can do both of these with the `kubectl exec` command. The following example shows how to execute a `ps aux` command in the first container in the `hello-pod` Pod.

```
$ kubectl exec hello-pod ps aux
PID    USER      TIME    COMMAND
   1    root       0:00    node ./app.js
  40    root       0:00    ps aux
```

You can also log-in to containers running in Pods using `kubectl exec`. When you do this, your terminal prompt will change to indicate your session is now running inside of a container in the Pod, and you'll be able to execute commands from there (as long as the command binaries are installed in the container).

The following `kubectl exec` command will log-in to the first container in the `hello-container` Pod. Once inside the container, install the `curl` utility and run a `curl` command to transfer data from the process listening on port 8080.

```
$ kubectl exec -it hello-pod sh
```

```
# apk add curl
```

```
<Snip>
```

```
# curl localhost:8080
```

```
<html><head><title>Pluralsight Rocks</title><link rel="stylesheet" href="http://\
/netdna.bootstrapcdn.com/bootstrap/3.1.1/css/bootstrap.min.css"/></head><body><\
div class="container"><div class="jumbotron"><h1>Yo Pluralsighters!!!</h1><p>Cl\
ick the button below to head over to my podcast...</p><p> <a href="http://intec\
hwetrustpodcast.com" class="btn btn-primary">Podcast</a></p><p></p></div></div>\
</body></html>
```

The `-it` flags make the `exec` session interactive and connects STDIN and STDOUT on your terminal to STDIN and STDOUT inside the first container in the Pod. When the command completes, your shell prompt will change to indicate your shell is now connected to the container.

If you are running multi-container Pods, you will need to pass the `kubectl exec` command the `--container` flag and give it the name of the container that you want to create the `exec` session with. If you do not specify this flag, the command will execute against the first container in the Pod.

You can see the ordering and names of containers in a Pod with the `kubectl describe pods <pod>` command.

One other useful command for introspecting Pods is the `kubectl logs` command. Like other Pod-related commands, if you don't use `--container` to specify a container by name, it will execute against the first container in the Pod. The format of the command is `kubectl logs <pod>`.

There's obviously a lot more to Pods than what we've covered. However, we've learned enough to get started.

Clean-up up the lab by typing `exit` to quit your shell session inside the container, then run `kubectl delete` to delete the Pod.

```
# exit
$ kubectl delete -f pod.yml
pod "hello-pod" deleted
```

Chapter Summary

In this chapter, we learned that the atomic unit of deployment in the Kubernetes world is the *Pod*. Each Pod consists of one or more containers and gets deployed to a single node in the cluster. The deployment operation is an all-or-nothing *atomic transaction*.

Pods are deployed declaratively using a YAML manifest file, and it's normal to deploy them via higher-level controllers such as Deployments. We use the `kubectl` command to `POST` the manifest to the API server, it gets stored in the cluster store and converted into a `PodSpec` that gets scheduled to a healthy cluster node with enough available resources.

The process on the worker node that accepts the `PodSpec` is the `kubelet`. This is the main Kubernetes agent running on every node in the cluster. It takes the `PodSpec` and is responsible for pulling all images and starting all containers in the Pod.

If a singleton Pod fails, it is not automatically rescheduled. Because of this, we usually deploy Pods via higher-level objects like Deployments and DaemonSets. These add capabilities such as self-healing and roll-backs which are at the heart of what makes Kubernetes so powerful.

5: Kubernetes Deployments

In this chapter, we'll see how *Deployments* bring self-healing, scalability, rolling updates, and versioned rollbacks to Kubernetes.

We'll divide the chapter as follows:

- Deployment theory
- How to create a Deployment
- How to perform a rolling update
- How to perform a rollback

Deployment theory

At a high level, we start with application code. That gets packaged as a container and wrapped in a Pod so it can run on Kubernetes. However, Pods don't self-heal, they don't scale, and they don't allow for easy updates or rollbacks. Deployments do all of these. As a result, we almost always run Pods via Deployments.

Figure 5.1 shows some Pods being managed by a Deployment.

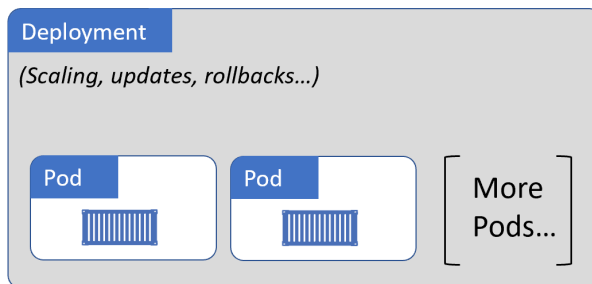


Figure 5.1

It's important to know that a single Deployment can only manage a single type of Pod. For example, if you have an application with a Pod for the web front-end and another Pod for the catalog service, you'll need two Deployments. However, as we saw in Figure 5.1, a Deployment can manage multiple replicas of the same Pod. For example, Figure 5.1 could be a deployment that currently manages two replicated web server Pods.

The next thing to know is that Deployments are fully-fledged objects in the Kubernetes API. This means we define them in manifest files that we POST to the API server.

The last thing to note, is that behind-the-scenes, Deployments leverage another object called a ReplicaSet. While it's best-practice that we don't directly manage ReplicaSets, it's important to understand the role they play as it will help explain the mechanics of some of the operations we're about to describe.

Keeping it high-level, Deployments use ReplicaSets to provide self-healing and scalability.

Figure 5.2. shows the same Pods managed by the same Deployment. However, this time we've added a ReplicaSet object into the relationship and shown how which object is responsible for which feature.

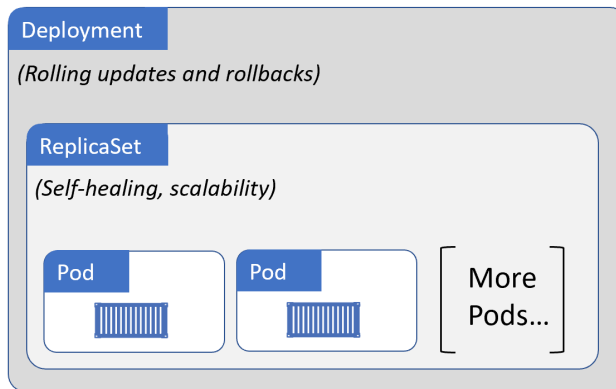


Figure 5.2

In summary, think of *Deployments* as managing *ReplicaSets*, and *ReplicaSets* as managing *Pods*. Put them all together, and we've got a great way to deploy and manage applications on Kubernetes.

Self-healing and scalability

Pods are great. They augment containers by allowing co-location of containers, sharing of volumes, sharing of memory, simplified networking, and a lot more. But they offer nothing in the way of self-healing and scalability – if a Pod fails, or the node it's running on fails, the Pod will not be restarted.

Enter Deployments.

Deployments augment Pods by adding things like self-healing and scalability. This means:

- If a Pod being managed by a Deployment fails, it will be replaced – *self-healing*.
- If a Pod being managed by a Deployment sees increased load, more of the same Pod can be added to deal with the load – *scaling*.

Remember though, behind-the-scenes, Deployments use an object called a ReplicaSet to accomplish self-healing and scalability. However, ReplicaSets operate in the background and we should always manage them via Deployments. For this reason, we'll focus on Deployments.

It's all about the *state*

Before going any further, it's critical to understand three concepts that are fundamental to everything about Kubernetes:

- Desired state
- Current state (sometimes called *actual state* or *observed state*)
- Declarative model

Desired state is what you **want**. *Current state* is what you **have**. If the two match, everybody's happy.

A *declarative model* is a way of telling Kubernetes what our *desired state* is, without having to get into the detail of *how* to implement it.

The declarative model

There are two competing models. The *declarative model* and the *imperative model*.

The declarative model is all about describing the end-goal – telling Kubernetes what you want. The imperative model is all about lists of things to do to achieve an end-goal – telling Kubernetes **how** to do something.

The following is an extremely simple analogy that might help:

- **Declarative:** I need a chocolate cake that will feed 10 people.
- **Imperative:** Drive to the store. Buy, eggs, milk, flour, cocoa powder... Drive home. Turn on oven. Mix ingredients. Place in baking tray. Place tray in oven for 30 minutes. Remove from oven and turn oven off. Add icing. Leave to stand.

The declarative model is stating what you want (chocolate cake for 10). The imperative model is a long list of steps required to make a chocolate cake for 10.

Let's look at a more concrete example.

Assume you've got an application with two services – front-end and back-end. You've built container images so that you can have a Pod for the front-end service, and a separate Pod for the back-end service. To meet expected demand, you always need 5 instances of the front-end Pod, and 2 instances of the back-end Pod.

Taking the declarative approach, you write a configuration file that tells Kubernetes what you want your application to look like. For example, *I want 5 replicas of the front-end Pod all listening externally on port 80 please. And I also want 2 back-end Pods listening internally on port 27017.* That's the desired state. Obviously, the YAML format of the config file will be different, but you get the picture.

Once you've described the desired state, you give the config file to Kubernetes and sit back while Kubernetes does the hard work of implementing it.

But things don't stop there... Kubernetes implements watch loops that are constantly checking that you've got what you asked for – does current state match desired state.

Believe me when I tell you, it's great!

The opposite of the declarative model is the imperative model. In the imperative model, there's no concept of what you actually want. At least there's no *record* of what you want, all you get is a list of instructions.

To make things worse, imperative instructions might have multiple variations. For example, the commands to start `containerd` containers are different from the commands to start `rkt` containers. This ends up being more work, prone to more errors, and because it's not declaring a desired state, there's no self-healing.

Believe me when I tell you, this isn't so great.

Kubernetes supports both models, but strongly prefers the declarative model.

Reconciliation loops

Fundamental to desired state is the concept of background reconciliation loops.

For example, ReplicaSets implement a background reconciliation loop that is constantly checking whether the right number of Pod replicas are present on the cluster. If there aren't enough, it adds more. If there are too many, it terminates some.

To be crystal clear, **Kubernetes is constantly making sure that *current state* matches *desired state*.**

If they don't match – may be desired state is 10 replicas, but only 8 are running – Kubernetes switches to red-alert, orders the control plane to battle-stations, and brings up two more replicas. And the best part... it does all of this without calling you at 4:20 in the morning!

But it's not just failure scenarios. These very-same reconciliation loops enable scaling.

For example, if you POST an updated config that changes replica count from 3 to 5. The new value of 5 will be registered as the application's new *desired state*, and the next time the ReplicaSet reconciliation loop runs, it will notice the discrepancy and follow the same process – sounding the claxon horn for red alert and spinning up two more replicas.

It's a beautiful thing.

Rolling updates with Deployments

As well as self-healing and scaling, Deployments give us zero-downtime rolling-updates.

As previously mentioned, Deployments use ReplicaSets for some of the background legwork. In fact, every time we create a Deployment, we automatically get a ReplicaSet that manages the Deployment's Pods.

Note: Best practice states that you should not manage ReplicaSets directly. You should perform all actions against the Deployment object and leave the Deployment to manage its own ReplicaSets.

It works like this. We design applications with each discrete service as a Pod. For convenience – self-healing, scaling, rolling updates and more – we wrap Pods in Deployments. This means creating a YAML configuration file describing all of the following:

- How many Pod replicas
- What image to use for the Pod’s container(s)
- What network ports to use
- Details about how to perform rolling updates

You POST the YAML file to the API server, and Kubernetes does the rest.

Once everything is up and running, Kubernetes sets up watch loops to make sure observed state matches desired state.

All good so far.

Now, assume you’ve experienced a bug, and you need to deploy an updated image that implements a fix. To do that, you update the **same Deployment YAML file** with the new image version, and re-POST it to the API server. This registers a new desired state on the cluster requesting the same number of Pods, but all running the new version of the image. To make this happen, Kubernetes creates a new ReplicaSet for the Pods with the new image. We now have two ReplicaSets – the original one for the Pods with the old version of the image, and another for the Pods with the new version. As Kubernetes increases the number of Pods in the new ReplicaSet (with the new version of the image) it decreases the number of Pods in the old ReplicaSet (with the old version of the image). Net result, we get a smooth rolling update with zero downtime. And we can rinse and repeat the process for future updates – just keep updating that manifest file (which should be stored in a version control system).

Brilliant.

Figure 5.3 shows a Deployment that has been updated once. The initial deployment created the ReplicaSet on the left, and the update created the ReplicaSet on the right. We can see that the ReplicaSet for the initial deployment has been wound down and no longer has any Pods. The ReplicaSet associated with the update is active and owns all of the Pods.

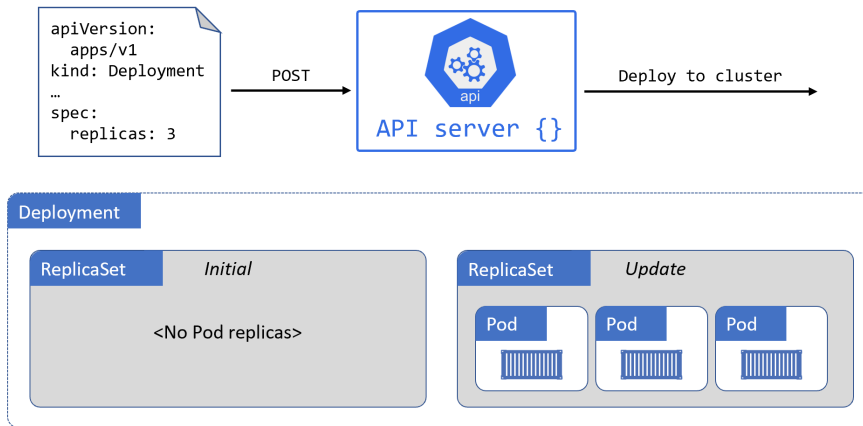


Figure 5.3

It's important to understand that the old ReplicaSet still has its entire configuration, including the older version of the image it used. This will be important in the next section.

Rollbacks

As we've seen in Figure 5.3, older ReplicaSets are wound down and no longer manage any Pods. However, they still exist with their full configuration. This makes them a great option for reverting to previous versions.

The process of rolling back is essentially the opposite of a rolling update – wind one of the old ReplicaSets up, and wind the current one down. Simple.

Figure 5.4 shows the same app rolled back to the initial revision.

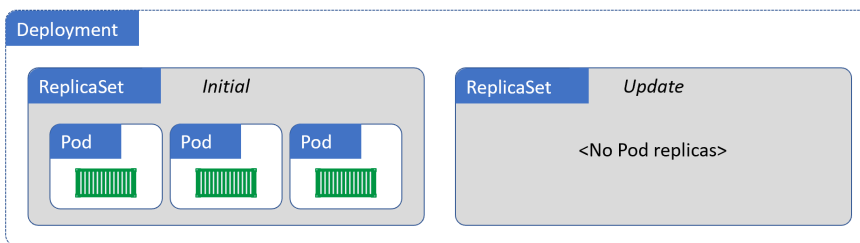


Figure 5.4

That's not the end though. There's built-in intelligence that lets us say things like *"wait X number of seconds after each Pod comes up before we mark it as healthy..."*. There's also liveness checks, readiness checks, and more. All-in-all, Deployments are excellent for performing rolling updates and versioned rollbacks.

With all that in mind, let's get our hands dirty and create a Deployment.

How to create a Deployment

In this section, we'll create a brand-new Kubernetes Deployment from a YAML file. We can do the same thing imperatively using the `kubectl run` command, but we shouldn't. The right way is the declarative way!

The following code snippet is the Deployment manifest file that we'll use. It's available in the book's GitHub repo in the "deployments" folder and is called `deploy.yaml`.

The examples assume it's in your system's PATH and is called `deploy.yaml`.

```
apiVersion: apps/v1 #Older versions of k8s use apps/v1beta1
kind: Deployment
metadata:
  name: hello-deploy
spec:
  replicas: 10
  selector:
    matchLabels:
      app: hello-world
  minReadySeconds: 10
  strategy:
    type: RollingUpdate
    rollingUpdate:
      maxUnavailable: 1
      maxSurge: 1
  template:
    metadata:
      labels:
        app: hello-world
    spec:
      containers:
        - name: hello-pod
          image: nigelpoulton/k8sbook:latest
          ports:
            - containerPort: 8080
```

Warning: The images used in this book are not maintained and may contain vulnerabilities and other security issues. Use with caution.

Let's step through the config and explain some of the important parts.

Right at the very top we specify the API version to use. Assuming that you're using an up-to-date version of Kubernetes, Deployment objects are in the `apps/v1` API group. If you're using an older version, you can try `apps/v1beta1` or `extensions/v1beta2`.

Next, the `.kind` field tells Kubernetes we're defining a Deployment.

The `.metadata` section is where we give the Deployment a name and labels.

The `.spec` section is where most of the action happens. Anything directly below `.spec` relates to the Pod. Anything nested below `.spec.template.spec` relates to the containers that will be part of the Pod.

At the Pod-level, `.spec.replicas` tells Kubernetes how many Pod replicas for this Deployment. `spec.selector` is a list of labels that Pods must have in order for the Deployment to manage them. And `.spec.strategy` tells Kubernetes how to perform updates to the Deployment.

Use `kubectl apply` to implement it on the cluster.

Note: `kubectl apply` POSTs the YAML file to the Kubernetes API server.

```
$ kubectl apply -f deploy.yml
deployment.apps/hello-deploy created
```

The Deployment is now instantiated on the cluster.

Inspecting Deployments

We can use the usual `kubectl get` and `kubectl describe` commands to see details of the Deployment.

```
$ kubectl get deploy hello-deploy
NAME          DESIRED   CURRENT   UP-TO-DATE   AVAILABLE   AGE
hello-deploy  10        10        10           10          24s

$ kubectl describe deploy hello-deploy
Name:          hello-deploy
Namespace:     default
Selector:      app=hello-world
Replicas:      10 desired | 10 updated | 10 total ...
StrategyType:  RollingUpdate
MinReadySeconds: 10
RollingUpdateStrategy: 1 max unavailable, 1 max surge
Pod Template:
```



```

Labels:                app=hello-world
Containers:
  hello-pod:
    Image:              nigelpoulton/k8sbook:latest
    Port:               8080/TCP

```

<SNIP>

The command outputs have been trimmed for readability. Your outputs will show more information.

As we mentioned earlier, Deployments automatically create associated ReplicaSets. Use the following `kubectl` command to confirm this.

```

$ kubectl get rs
NAME                                DESIRED   CURRENT   READY   AGE
hello-deploy-7bbd...               10        10        10      1m

```

Right now, we only have one ReplicaSet. This is because we’ve only performed the initial rollout of the Deployment. We can also see that the name of the ReplicaSet matches the name of the Deployment with a hash on the end. This hash is a hash of the Pod template section (anything below `.spec`) of the YAML manifest file.

We can get more detailed information about the ReplicaSet with the usual `kubectl describe` command.

Accessing the app

In order to access the application from a stable name or IP address, or even from outside the cluster, we need a Kubernetes Service object (more on these in the next chapter). For now, all you need to know is that Kubernetes Services provide a stable DNS name and IP address for a set of Pods.

The following YAML defines a Service that will work with the Pods previously deployed. The YAML is included in the “deployments” folder of the book’s GitHub repo called `svc.yml`.

```

apiVersion: v1
kind: Service
metadata:
  name: hello-svc
  labels:
    app: hello-world
spec:
  type: NodePort
  ports:

```

```
- port: 8080
  nodePort: 30001
  protocol: TCP
selector:
  app: hello-world
```

Deploy it with the following command (the command assumes the manifest file is called `svc.yml` and is in your system's PATH).

```
$ kubectl apply -f svc.yml
service/hello-svc created
```

Now that the Service is deployed, you can access it:

1. From inside the cluster using the DNS name `hello-svc` on port `8080`
2. From outside the cluster by hitting any of the cluster nodes on port `30001`

Figure 5.5 shows the Service being accessed from outside of the cluster via a node called `node1` on port `30001`. It assumes that `node1` is resolvable, and that port `30001` is allowed by any intervening firewalls.

If you are using Minikube, you should append port `30001` to the end of the Minikube IP address. Use the `minikube ip` command to get the IP address of your Minikube.

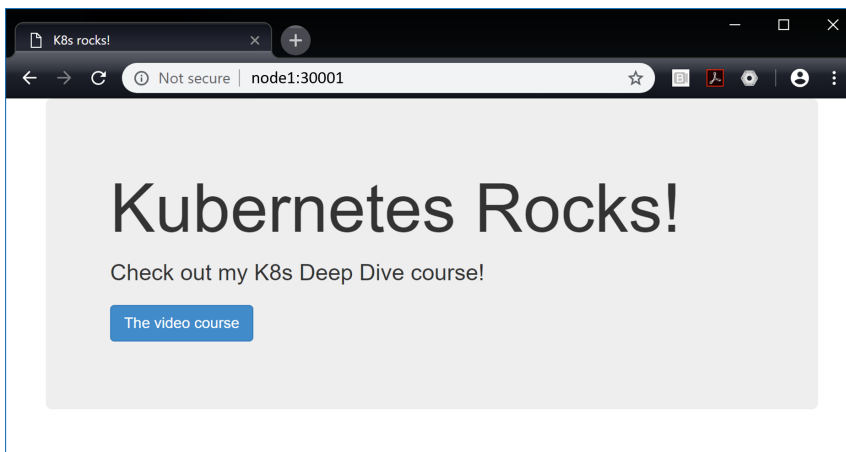


Figure 5.5

Performing a rolling update

In this section, we'll see how to perform a rolling update on the app we've just deployed. We'll assume the new version of the app has already been created and containerized as a Docker image with the `edge` tag. All that is left to do is use Kubernetes to push the update to production. For this example, we're ignoring real-world CI/CD workflows and version control tools.

The first thing we need to do is update the image tag used in the Deployment's manifest file. The initial version of the app used an image tagged as `nigelpoulton/k8sbook:latest`. We'll update the `.spec.template.spec.containers` section of the Deployment manifest to reference the new `nigelpoulton/k8sbook:edge` image. This will ensure that next time the manifest is POSTed to the API server, all Pods in the Deployment will be updated to run the new `edge` image.

The following is the updated `deploy.yml` manifest file – the only change is to `.spec.template.spec.containers` indicated by the commented line.

```
apiVersion: apps/v1
kind: Deployment
metadata:
  name: hello-deploy
spec:
  replicas: 10
  selector:
    matchLabels:
      app: hello-world
  minReadySeconds: 10
  strategy:
    type: RollingUpdate
    rollingUpdate:
      maxUnavailable: 1
      maxSurge: 1
  template:
    metadata:
      labels:
        app: hello-world
    spec:
      containers:
        - name: hello-pod
          image: nigelpoulton/k8sbook:edge    # This line changed
          ports:
            - containerPort: 8080
```

Before POSTing the updated configuration to Kubernetes, let's look at the settings that govern how the update will proceed.

The `.spec` section of the manifest contains all of the settings relating to how updates will be performed. The first value of interest is `.spec.minReadySeconds`. This is set to `10`, telling Kubernetes to wait for 10 seconds between each individual Pod being updated. This is useful for throttling the rate at which updates occur – longer waits give us a chance to spot problems and avoid situations where we update all Pods to a faulty configuration.

We also have a nested `.spec.strategy` map that tells Kubernetes we want this Deployment to:

- Update using the `RollingUpdate` strategy
- Never go more than one Pod below desired state (`maxUnavailable: 1`)
- Never go more than one Pod above desired state (`maxSurge: 1`)

As the desired state of the app demands 10 replicas, `maxSurge: 1` means we will never have more than 11 Pods during the update process, and `maxUnavailable: 1` means we'll never have less than 9.

With the updated manifest ready, we can initiate the update by re-POSTing the updated YAML file to the API server.

```
$ kubectl apply -f deploy.yml --record
deployment.apps/hello-deploy configured
```

The update may take some time to complete. This is because it will iterate one Pod at a time, pull down the new image on each node, start the new Pod, and then wait 10 seconds before moving on to the next Pod.

We can monitor the progress of the update with `kubectl rollout status`.

```
$ kubectl rollout status deployment hello-deploy
Waiting for rollout to finish: 4 out of 10 new replicas...
Waiting for rollout to finish: 4 out of 10 new replicas...
Waiting for rollout to finish: 5 out of 10 new replicas...
^C
```

If you press `Ctrl+C` to stop watching the progress of the update, you can run `kubectl get deploy` commands while the update is in process. This lets us see the effect of some of the update-related settings in the manifest. For example, the following command shows that 5 of the replicas have been updated and we currently have 11. 11 is 1 more than the desired state of 10. This is a result of the `maxSurge=1` value in the manifest.

```
$ kubectl get deploy
```

NAME	DESIRED	CURRENT	UP-TO-DATE	AVAILABLE	AGE
hello-deploy	10	11	5	9	28m

Once the update is complete, we can verify with `kubectl get deploy`.

```
$ kubectl get deploy hello-deploy
```

NAME	DESIRED	CURRENT	UP-TO-DATE	AVAILABLE	AGE
hello-deploy	10	10	10	10	39m

The output shows the update as complete – 10 Pods are up-to-date.

You can get more detailed information about the state of the Deployment with the `kubectl describe deploy` command. This will include the new version of the image in the Pod Template section of the output.

If you’ve been following along with the examples, you’ll be able to hit refresh in your browser and see the updated app (Figure 5.6).

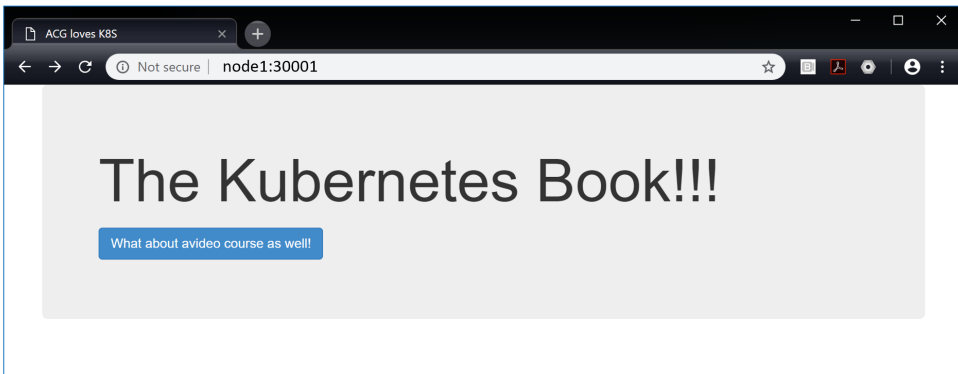


Figure 5.6

The old version of the app displayed “Kubernetes Rocks!”, the new version displays “The Kubernetes Book!!!”.

How to perform a rollback

A moment ago, we used `kubectl apply` to perform a rolling update on a Deployment. We used the `--record` flag so that Kubernetes would maintain a documented revision history of the Deployment. The following `kubectl rollout history` command shows the Deployment with two revisions.

```
$ kubectl rollout history deployment hello-deploy
deployment.apps/hello-deploy
REVISION  CHANGE-CAUSE
1          <none>
2          kubectl apply --filename-deploy.yml --record=true
```

Revision 1 was the initial deployment that used the latest image tag. Revision 2 is the rolling update we just performed, and we can see that the command we used to invoke the update has been recorded in the object's history. This is only there because we used the `--record` flag as part of the command to invoke the update. This might be a good reason for you to use the `--record` flag.

Earlier in the chapter we said that updating a Deployment creates a new ReplicaSet, and that any previous ReplicaSets are not deleted. We can verify this with a `kubectl get rs`.

```
$ kubectl get rs
```

NAME	DESIRED	CURRENT	READY	AGE
hello-deploy-6bc8...	10	10	10	10m
hello-deploy-7bbd...	0	0	0	52m

The output shows that the ReplicaSet for the initial revision still exists (`hello-deploy-7bbd...`) but that it has been wound down and is not managing any replicas. The `hello-deploy-6bc8...` ReplicaSet is the one from the latest revision and is active with 10 replicas under management. However, the fact that the previous version still exists makes rollbacks extremely simple.

If you're following along, it's worth running a `kubectl describe rs` against the old ReplicaSet to prove that its configuration still exists.

The following example uses the `kubectl rollout` command to roll the application back to revision 1. This is an imperative operation and not recommended. However, it can be convenient for quick rollbacks, just remember to update your source YAML files to reflect the imperative changes you make to the cluster.

```
$ kubectl rollout undo deployment hello-deploy --to-revision=1
deployment.apps "hello-deploy" rolled back
```

Although it might look like the rollback operation is instantaneous, it's not. Rollbacks follow the same rules set out in the Deployment manifest – `minReadySeconds: 10`, `maxUnavailable: 1`, and `maxSurge: 1`. You can verify this and track the progress with the following `kubectl get deploy` and `kubectl rollout` commands.

```
$ kubectl get deploy hello-deploy
NAME          DESIRED  CURRNET  UP-TO-DATE  AVAILABE  AGE
hello-deploy  10       11       4            9          45m

$ kubectl rollout status deployment hello-deploy
Waiting for rollout to finish: 6 out of 10 new replicas have been updated...
Waiting for rollout to finish: 7 out of 10 new replicas have been updated...
Waiting for rollout to finish: 8 out of 10 new replicas have been updated...
Waiting for rollout to finish: 1 old replicas are pending termination...
Waiting for rollout to finish: 9 of 10 updated replicas are available...
^C
```

Congratulations. You’ve performed a rolling update and a successful rollback.

Use `kubectl delete -f deploy.yml` and `kubectl delete -f svc.yml` to delete the Deployment and Service used in the examples.

Just a quick reminder. The rollback operation we just initiated was an imperative operation. This means that the current state of the cluster will not match your source YAML files – the latest version of the YAML file lists the `edge` image, but we’ve rolled the cluster back to the `latest` image. This is a problem with the imperative approach. In the real world, following a rollback operation like this, you should manually update your source YAML files to reflect the changes incurred by the rollback.

Chapter summary

In this chapter, we learned that *Deployments* are a great way to manage Kubernetes apps. They build on top of Pods by adding self-healing, scalability, rolling updates, and rollbacks. Behind-the-scenes, they leverage ReplicaSets for the self-healing and scalability parts.

Like Pods, Deployments are objects in the Kubernetes API, and we should work with them declaratively.

When we perform updates with the `kubectl apply` command, older versions of ReplicaSets get wound down, but they stick around making it easy to perform rollbacks.

6: Kubernetes Services

In the previous chapters, we've launched Pods and used Deployments to give applications self-healing, scalability, and rolling updates. However, despite all of this, **we still cannot rely on Pod IPs!** In this chapter, we'll see how Kubernetes *Services* give us networking that we **can** rely on.

We'll divide the chapter as follows:

- Setting the scene
- Theory
- Hands-on
- Real world example

Setting the scene

Before diving in, we need to remind ourselves that Pod IPs are unreliable. When Pods fail, they get replaced with new Pods that have new IPs. Scaling-up a Deployment introduces new Pods with new IP addresses. Scaling-down a Deployment removes Pods. All of this creates a large amount of *IP churn*, and creates a situation where Pod IPs cannot be relied on.

We also need to know 3 fundamental things about Kubernetes Services.

First, we need to clear up some terminology. When talking about a *Service* in this chapter we're talking about the Service REST object in the Kubernetes API. Just like a *Pod*, *ReplicaSet*, or *Deployment*, a Kubernetes **Service** is an object in the API that we define in a manifest and POST to the API server.

Second, we need to know that every Service gets its own **stable IP address**, its own **stable DNS name**, and its own **stable port**.

Third, we need to know that Services use labels to dynamically select the Pods in the cluster they will send traffic to.

The last two points are what allow Services to provide stable networking to a dynamic set of Pods.

Theory

Figure 6.1 shows a simple Pod-based application deployed via a Kubernetes Deployment. It shows a client (which could be another component of the app) that does not have a reliable network endpoint for accessing the Pods.

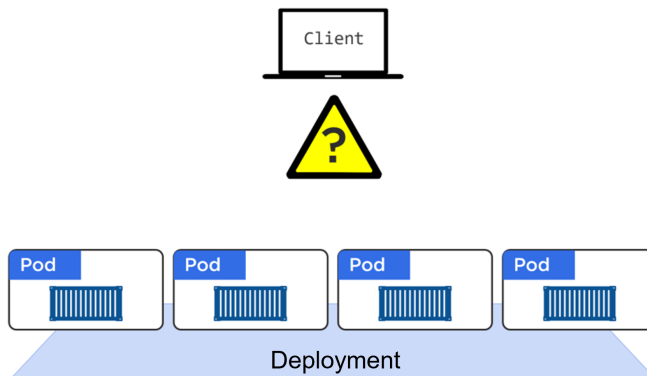


Figure 6.1

Figure 6.2 shows the same application with a Service added into the mix. The Service is associated with the Pods and fronts them with a stable IP, DNS, and port. It also load-balances requests across the Pods.

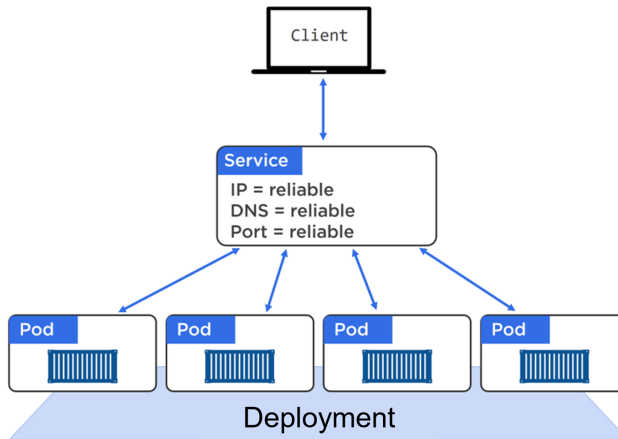


Figure 6.2

With a Service in front of a set of Pods, the Pods can scale up and down, they can fail, and they can be updated... As these events occur, the Service in front of them observes the changes and updates its knowledge of the Pods. But it never changes the stable IP, DNS and port that it exposes.

Think of Services as having a static front-end and a dynamic back-end. The front-end consists of the IP, DNS name, and port, and never changes. The back-end consists of the Pods, which are fluid and can be constantly changing.

Labels and loose coupling

Pods and Services are loosely coupled via *labels* and *label selectors*. This is the same technology that loosely couples Deployments to Pods and is key to the flexibility provided by Kubernetes. Figure 6.3 shows an example where 3 Pods are labelled as `zone=prod` and `version=v1`, and the Service has a *label selector* that matches.

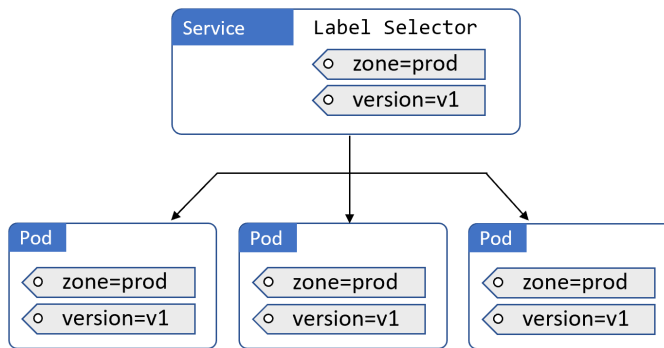


Figure 6.3

In Figure 6.3, the Service is providing stable networking to all three Pods – you can send requests to the Service and it will proxy them on to the Pods. It also provides simple load-balancing.

For a Service to match a set of Pods, and therefore send traffic to them, it's label selector only needs to match *some* of the labels on a Pod. However, for a Pod to match a Service, it must have all of the labels the Service is looking for. If that sounds confusing, the examples in Figures 6.4 and 6.5 should help.

Figure 6.4 shows an example where the Service does not match any of the Pods. This is because the Service is looking for Pods that have two labels, but the Pods only possess one of them. The logic behind the selection process is a Boolean AND.

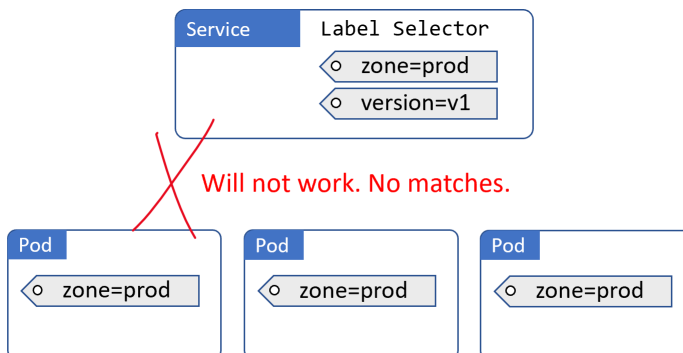


Figure 6.4

Figure 6.5 shows an example that does work. It works because the Service is looking for two labels, and the Pods in the diagram possess both. It doesn't matter that the Pods possess additional labels that the Service isn't looking for. The Service is looking for Pods with two labels, it finds them, and ignores the fact that the Pods have additional labels – all that is important is that the Pods possess the labels the Service is looking for.

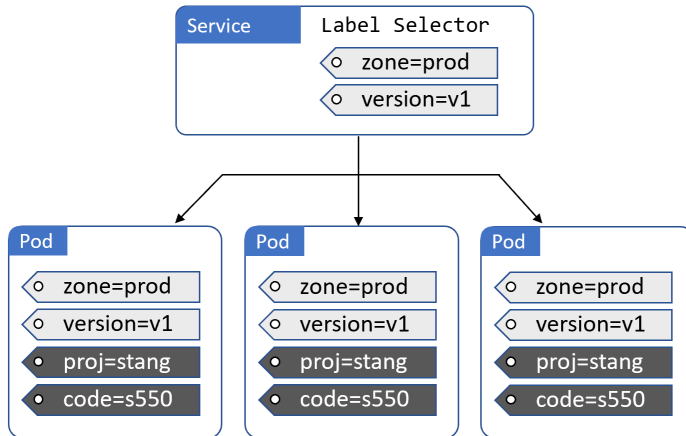


Figure 6.5

The following excerpts, from a Service YAML and Deployment YAML, show how *selectors* and *labels* are implemented. We've added comments to the lines we're interested in.

svc.yml

```

apiVersion: v1
kind: Service
metadata:
  name: hello-svc
  labels:
    app: hello-world
spec:
  type: NodePort
  ports:
    - port: 8080
      nodePort: 30001
      protocol: TCP
  selector:
    app: hello-world    # Label selector
    # Service is looking for Pods with the label `app=hello-world`
  
```

deploy.yml

```

apiVersion: apps/v1
kind: Deployment
metadata:
  name: hello-deploy
spec:
  replicas: 10
  selector:
    matchLabels:
      app: hello-world
  template:
    metadata:
      labels:
        app: hello-world    # Pod labels
        # The label above matches what the svc.yml is looking for
    spec:
      containers:
        - name: hello-ctr
          image: nigelpoulton/k8sbook:latest
          ports:
            - containerPort: 8080

```

In the example files, the Service has a label selector (`.spec.selector`) with a single value `app=hello-world`. This is the label that the Service is looking for when it queries the cluster for matching Pods. The Deployment specifies a Pod template with the same `app=hello-world` label (`.spec.template.metadata.labels`). This means that any Pods it deploys will have the `app=hello-world` label. It is these two attributes that loosely couple the Service to the Deployments Pods.

When the Deployment and the Service are deployed, the Service will select all 10 Pod replicas and provide them with a stable networking endpoint and load-balance traffic to them.

Services and Endpoint objects

As Pods come-and-go (scaling up and down, failures, rolling updates etc.), the Service dynamically updates its list of healthy matching Pods. It does this through a combination of the label selector and a construct called an *Endpoint* object.

Each Service that is created, automatically gets an associated *Endpoint* object. All this Endpoint object is, is a dynamic list of all of the healthy Pods on the cluster that match the Service's label selector.

It works like this...

Kubernetes is constantly evaluating the Service's label selector against the current list of healthy Pods on the cluster. Any new Pods that match the selector get added to the Endpoint object, and any Pods that disappear get removed. This means the Endpoint is always up-to-date. Then, when a Service is sending traffic to Pods, it queries its Endpoint object for the latest list of healthy matching Pods.

The Endpoint object has its own API endpoint that Kubernetes-native apps can query for the latest list of matching Pods (like the Service object does). These apps can then send traffic directly to Pods. Non-native Kubernetes apps, that cannot query the Endpoints object, send traffic to the Service's stable IP (VIP).

Now that we know the fundamentals of how Services work, let's look at some use-cases.

Accessing Services from inside the cluster

Kubernetes supports several *types* of Service. The default type is **ClusterIP**.

A ClusterIP Service has a stable IP address and port that is only accessible from inside the cluster. We call this IP its *ClusterIP* and it's programmed into the network fabric and guaranteed to be stable for the life of the Service.

The ClusterIP gets registered against the name of the Service on the cluster's native DNS service. All Pods in the cluster are pre-programmed to know about the cluster's DNS service, meaning all Pods are able to resolve Service names.

Let's look at a simple example.

Creating a new Service called "hellcat-svc" will trigger the following. Kubernetes will register the name "hellcat-svc", along with the ClusterIP and port, with the cluster's DNS service. The name, ClusterIP, and port are guaranteed to be long-lived and stable, and all Pods in the cluster will be able to resolve "hellcat-svc" to the ClusterIP. IPTables or IPVS rules are distributed across the cluster that ensure traffic sent to the ClusterIP gets routed to Pods on the backend.

Net net... as long as a Pod (application microservice) knows the name of a Service, it can resolve that to its ClusterIP address and connect to the desired Pods.

This only works for Pods and other objects on the cluster, as it requires access to the cluster's DNS service. It does not work outside of the cluster.

Accessing Services from outside the cluster

Kubernetes has another type of Service called a **NodePort Service**. This builds on top of ClusterIP and enables access from outside of the cluster.

We already know that the default Service type is ClusterIP, and that it registers a DNS name, virtual IP, and port with the cluster DNS. A different type of Service, called a NodePort Service builds on this by adding another port that can be used to reach the Service from outside the cluster. This additional port is called the *NodePort*.

The following example represents a NodePort Service:

- **Name:** hellcat-svc
- **ClusterIP:** 172.12.5.17
- **port:** 8080
- **NodePort:** 30050

This Service can be accessed directly from inside the cluster via any of the first three values (Name, ClusterIP, port). It can also be accessed from outside of the cluster by hitting any cluster node on port 30050.

At the bottom of the stack we have *nodes* in the cluster hosting Pods. Then we create a Service and use labels to associate it with Pods. The Service object has a reliable NodePort mapped to every node in the cluster -- the NodePort value is the same on every node. This means that traffic from outside of the cluster can hit any node in the cluster on the NodePort and get through to the application (Pods).

Figure 6.6 shows a NodePort Service where 3 Pods are exposed externally on port 30050 on every node in the cluster. In step 1, an external client hits **Node2** on port 30050. In step 2 it is redirected to the Service object (this happens even though **Node2** isn't running a Pod from the Service). Step 3 shows that the Service has an associated Endpoint object with an always-up-to-date list of Pods matching the label selector. Step 4 shows the client being directed to **pod1** on **Node1**.

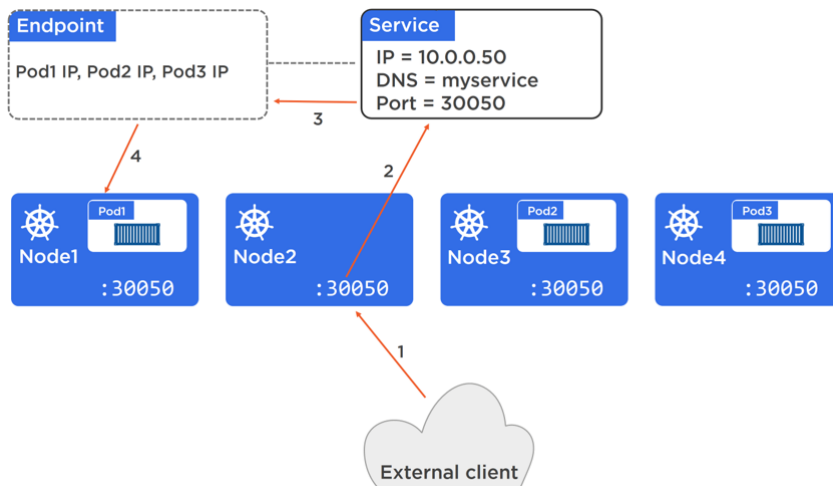


Figure 6.6

The Service could just as easily have directed the client to pod2 or pod3. In fact, future requests may go to other Pods as the Service performs basic load-balancing.

There are other types of Services, such as LoadBalancer Services. These integrate with load-balancers from your cloud provider such as AWS, Azure, and GCP. They build on top of NodePort Services (which in turn build on top of ClusterIP Services) and allow clients on the internet to reach your Pods via one of your cloud's load-balancers.

LoadBalancer Services are extremely easy to setup. However, they only work if you're running your Kubernetes cluster on a supported cloud platform. E.g. you cannot leverage an ELB load-balancer on AWS if your Kubernetes cluster is running on Microsoft Azure.

Service discovery

Kubernetes implements Service discovery in a couple of ways:

- DNS (preferred)
- Environment variables (definitely not preferred)

DNS-based Service discovery requires the DNS *cluster-add-on* – this is just a fancy name for the native Kubernetes DNS service. If you followed the installation methods from the “Installing Kubernetes” chapter, you'll already have this. It implements a Pod-based DNS service in the cluster and configures all kubelets (nodes) to use it for DNS.

The DNS add-on constantly watches the API server for new Services and automatically registers them in DNS. This means every Service gets a DNS name that is resolvable across the entire cluster.

The alternative form of service discovery is through environment variables. Every Pod gets a set of environment variables that resolve every Service currently on the cluster. However, this is a fall-back in case you're not using DNS in your cluster.

The problem with environment variables, is that they're only inserted into Pods when the Pod is initially created. This means that Pods have no way of learning about new Services that are added to the cluster after the Pod itself is created. This is far from ideal, and a major reason DNS is the preferred method.

Summary of Service theory

Services are all about providing stable networking for Pods. They also provide load-balancing and ways to be accessed from outside of the cluster.

The front-end of a Service provides a stable IP, DNS name and port that is guaranteed not to change for the entire life of the Service. The back-end of a Service uses labels to load-balance traffic across a potentially dynamic set of application Pods.

Hands-on with Services

We're about to get hands-on and put the theory to the test.

We'll augment a simple single-Pod app with a Kubernetes Service. And we'll show how to do it in two ways:

- The imperative way (only use in emergencies)
- The declarative way

The imperative way

Warning! The imperative way is **not** the Kubernetes way. It introduces the risk that imperative changes never make back to declarative manifests, rendering the manifests stale. This introduces the risk that stale manifests are used to update the cluster at a later date, unintentionally overwriting important changes that were made imperatively.

Use `kubectl` to declaratively deploy the following Deployment (later steps will be done imperatively).

The YAML file is called `deploy.yml` and can be found in the `services` folder in the book's GitHub repo.

```
apiVersion: apps/v1
kind: Deployment
metadata:
  name: web-deploy
spec:
  replicas: 10
  selector:
    matchLabels:
      app: hello-world
  template:
    metadata:
      labels:
        app: hello-world
    spec:
      containers:
        - name: hello-ctr
          image: nigelpoulton/k8sbook:latest
          ports:
            - containerPort: 8080
```

```
$ kubectl apply -f deploy.yml
deployment.apps/hello-deploy created
```

Now that the Deployment is running, it's time to imperatively deploy a Service for it.

Use the following `kubectl` command to create a new Service that will provide networking and load-balancing for the Pods deployed in the previous step.


```
$ kubectl expose deployment web-deploy \
  --name=hello-svc \
  --target-port=8080 \
  --type=NodePort
```

service/hello-svc exposed

Let's explain what the command is doing. `kubectl expose` is the imperative way to create a new *Service* object. `deployment web-deploy` is telling Kubernetes to expose the `web-deploy` Deployment that we created in the previous step. `--name=hello-svc` tells Kubernetes that we want to call this Service "hello-svc", and `--target-port=8080` tells it which port the app is listening on (this is **not** the cluster-wide NodePort that we'll access the Service on). Finally, `--type=NodePort` tells Kubernetes we want a cluster-wide port for the Service.

Once the Service is created, you can inspect it with the `kubectl describe svc hello-svc` command.

```
$ kubectl describe svc hello-svc
Name:                hello-svc
Namespace:           default
Labels:              app=hello-world
Annotations:         <none>
Selector:            app=hello-world
Type:               NodePort
IP:                 100.70.80.47
Port:               <unset> 8080/TCP
NodePort:           <unset> 30175/TCP
Endpoints:          100.96.1.10:8080, 100.96.1.11:8080, + more...
Session Affinity:   None
Events:             <none>
```

Some interesting values in the output include:

- **Selector** is the list of labels that Pods must have in order for the Service to send traffic to them
- **IP** is the permanent ClusterIP (VIP) of the Service
- **Port** is the port that the app and Service listens on
- **NodePort** is the cluster-wide port that can be used to access it from outside the cluster
- **Endpoints** is the dynamic list of healthy Pods that currently match the Service's label selector.

Now that we know the cluster-wide port that the Service is accessible on, we can open a web browser and access the app. In order to do this, you will need to know the IP address of at least one of the

nodes in your cluster, and you will need to be able to reach it from your browser - e.g. a publicly routable IP if you're accessing via the internet.

Figure 6.7 shows a web browser accessing a cluster node with an IP address of 54.246.255.52 on the cluster-wide NodePort 30175.

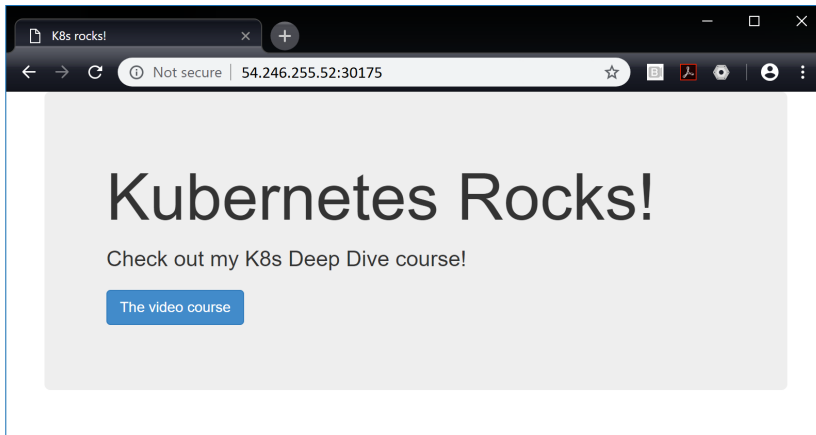


Figure 6.7

The app we've deployed is a simple web app. It's built to listen on port 8080, and we've configured a Kubernetes *Service* to map port 30175 on every cluster node back to port 8080 on the app. By default, cluster-wide ports (NodePort values) are between 30,000 - 32,767.

Coming up next we're going to see how to do the same thing the proper way – the declarative way. To do that, we need to clean up by deleting the Service we just created. We can do this with the `kubectl delete svc` command

```
$ kubectl delete svc hello-svc
service "hello-svc" deleted
```

The declarative way

Time to do things the proper way... the Kubernetes way!

A Service manifest file

We'll use the following Service manifest file to deploy the same *Service* that we deployed in the previous section. However, this time we'll specify a value for the cluster-wide port.

```
apiVersion: v1
kind: Service
metadata:
  name: hello-svc
  labels:
    app: hello-world
spec:
  type: NodePort
  ports:
  - port: 8080
    nodePort: 30001
    protocol: TCP
  selector:
    app: hello-world
```

Let's step through some of the lines.

Services are mature objects and are fully defined in the `v1` core API group (`.apiVersion`).

The `.kind` field tells Kubernetes we're defining a Service.

The `.metadata` section defines a name and a label for the Service. The label here is a label for the Service itself. It is not the label that the Service uses to select Pods.

The `.spec` section is where we actually define the Service. In this example, we're telling Kubernetes to deploy a `NodePort` Service and to map port `8080` from the app to port `30001` on each node in the cluster. Then we're explicitly telling it to use `TCP` (default).

Finally, `.spec.selector` tells the Service to send traffic to all Pods in the cluster that have the `app=hello-world` label. This means it will provide stable networking and load-balancing across all Pods with that label.

Before deploying and testing the Service, let's remind ourselves of the major Service types.

Common Service types

The three common *ServiceTypes* are:

- **ClusterIP**. This is the default option and gives the *Service* a stable IP address internally within the cluster. It will not make the Service available outside of the cluster.
- **NodePort**. This builds on top of **ClusterIP** and adds a cluster-wide TCP or UDP port. It makes the Service available outside of the cluster on this port.
- **LoadBalancer**. This builds on top of **NodePort** and integrates with cloud-based load-balancers.

The manifest needs POSTing to the API server. The simplest way to do this is with `kubectl apply`.

The YAML file is called `svc.yml` and can be found in the `services` folder of book's GitHub repo.

```
$ kubectl apply -f svc.yml
service/hello-svc created
```

This command tells Kubernetes to deploy a new object from a file called `svc.yml`. The `.kind` field tells Kubernetes that we are deploying a new Service object.

Introspecting Services

Now that the Service is deployed, we can inspect it with the usual `kubectl get` and `kubectl describe` commands.

```
$ kubectl get svc hello-svc
```

NAME	TYPE	CLUSTER-IP	EXTERNAL-IP	PORT(S)	AGE
hello-svc	NodePort	100.70.40.2	<nodes>	8080:30001/TCP	8s

```

$ kubectl describe svc hello-svc
Name:                hello-svc
Namespace:           default
Labels:              app=hello-world
Annotations:         <none>
Selector:            app=hello-world
Type:               NodePort
IP:                 100.70.40.2
Port:               <unset> 8080/TCP
NodePort:           <unset> 30001/TCP
Endpoints:          100.96.1.10:8080, 100.96.1.11:8080, + more...
Session Affinity:   None
Events:             <none>
```

In the previous example, we've exposed the Service as a `NodePort` on port 30001 across the entire cluster. This means we can point a web browser to that port on any node and reach the Service and the Pods it's proxying. You will need to use the IP address of a node you can reach, and you will need to make sure that any firewall and security rules allow the traffic to flow.

Figure 6.8 shows a web browser accessing the app via a cluster node with an IP address of 54.246.255.52 on the cluster-wide port 30001.

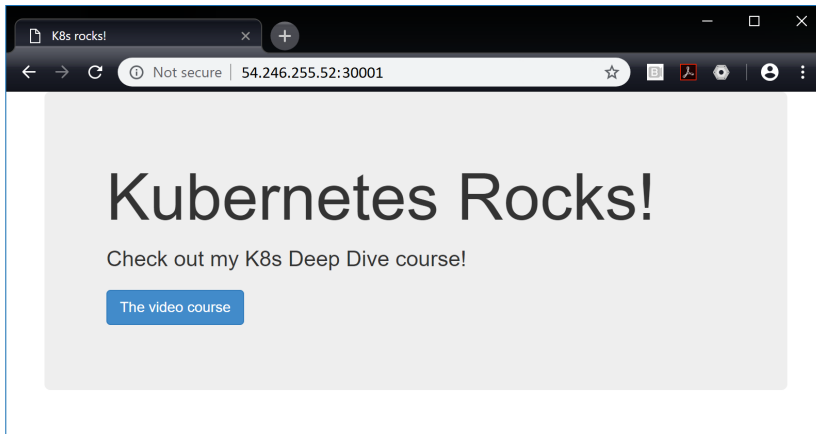


Figure 6.8

Endpoint objects

Earlier in the chapter, we said that every Service gets its own Endpoint object with the same name as the Service. This object holds a list of all the Pods the Service matches and is dynamically updated as Pods come and go. We can see Endpoints with the normal `kubectl` commands.

In the following command, we use the Endpoint controller's `ep` shorthand.

```
$ kubectl get ep hello-svc
NAME           ENDPOINTS                                     AGE
hello-svc      100.96.1.10:8080, 100.96.1.11:8080 + 8 more... 1m
kubernetes     172.20.32.78:443
```

```
$ kubectl describe ep hello-svc
Name:          hello-svc
Namespace:     default
Labels:        app=hello-world
Annotations:   <none>
Subsets:
  Addresses:    100.96.1.10,100.96.1.11,100.96.1.12...
  NotReadyAddresses: <none>
  Ports:
    Name      Port      Protocol
    ----      -
    <unset>    8080      TCP
Events: <none>
```

Summary of deploying Services

As with all Kubernetes objects, the preferred way of deploying and managing Services is the declarative way. Labels allow them to send traffic to a dynamic set of Pods. This means you can deploy new Services that will work with Pods and Deployments that are already running on the cluster and already in-use. Each Service gets its own Endpoint object that maintains an up-to-date list of matching Pods.

Real world example

Although everything we've learned so far is cool and interesting, the important questions are: *How does it bring value?* and *How does it keep businesses running and make them more agile and resilient?*

Let's take a minute to run through a common real-world example – making updates to applications.

We all know that updating applications is a fact of life – bug fixes, new features etc.

Figure 6.9 shows a simple application deployed on a Kubernetes cluster as a bunch of Pods managed by a Deployment. As part of it, we've got a Service selecting on Pods with labels that match `app=biz1` and `zone=prod` (notice how the Pods have both of the labels listed in the label selector). The application is up and running.

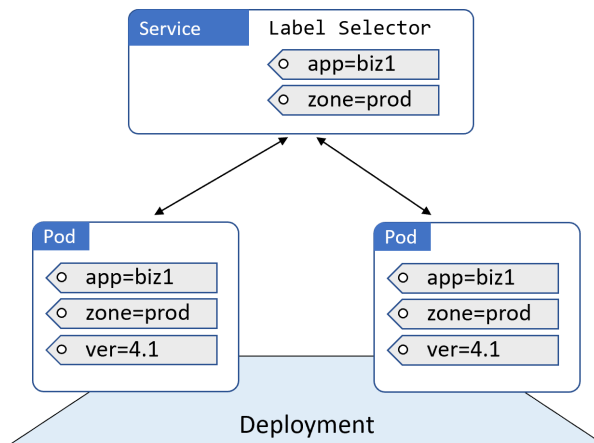


Figure 6.9

Let's assume we need to push a new version. But we need to do it without incurring downtime.

To do this, we can add Pods running the new version of the app as shown in Figure 6.10.

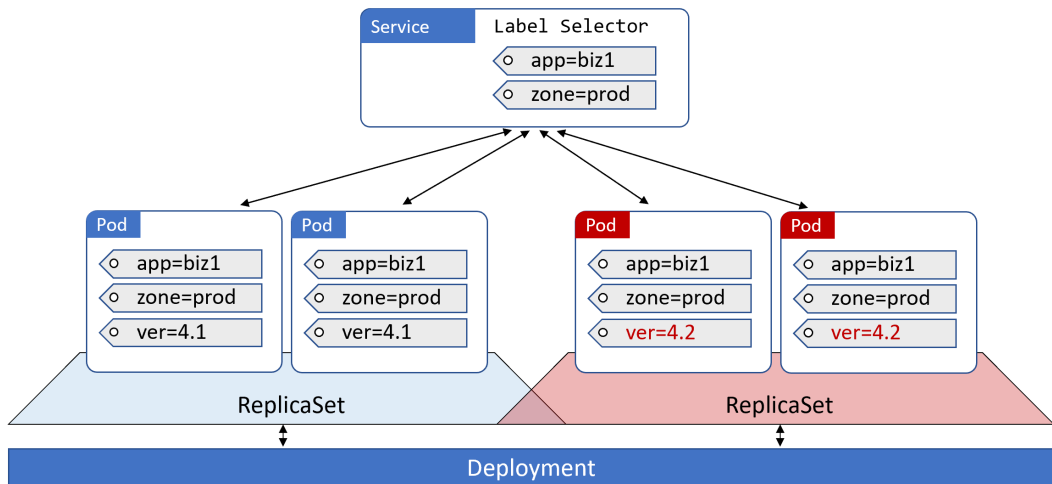


Figure 6.10

Behind the scenes, the updated *Pods* are labelled so that they match the existing label selector. The Service is now load-balancing requests across **both versions of the app** (`version=4.1` and `version=4.2`).

This happens because the Service's label selector is being constantly evaluated, and its Endpoint object and ClusterIP are constantly being updated with new matching Pods.

Once you're happy with the updated version, forcing all traffic to use it is as simple as updating the Service's label selector to include the label `version=4.2`. Suddenly the older Pods no longer match, and the Service is only forwarding traffic to the new version (Figure 6.11).

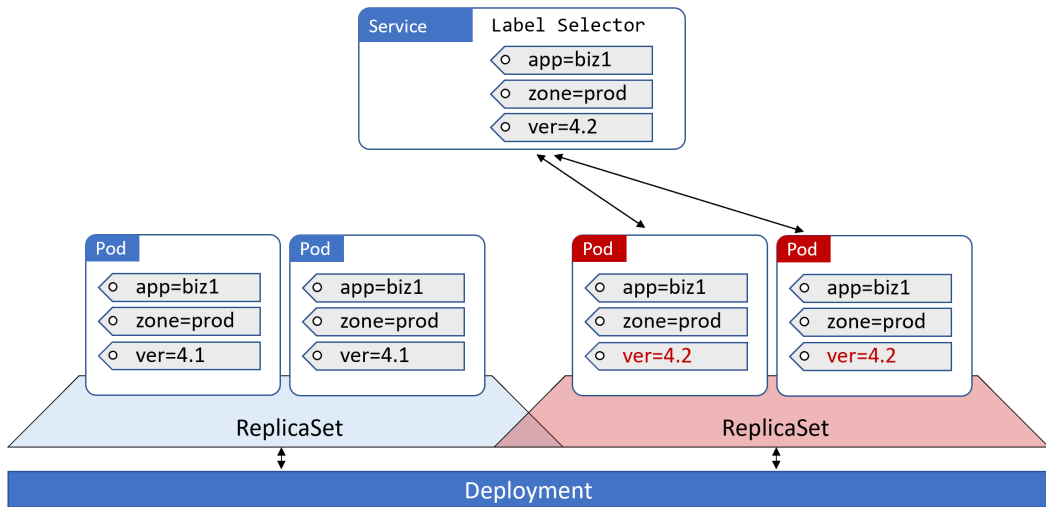


Figure 6.11

However, the old version still exists, we're just not sending traffic to it any more. This means that if we experience an issue with the new version, we can switch back to the previous version by simply changing the label selector on the Service to select on `version=4.1` instead of `version=4.2`. See Figure 6.12.

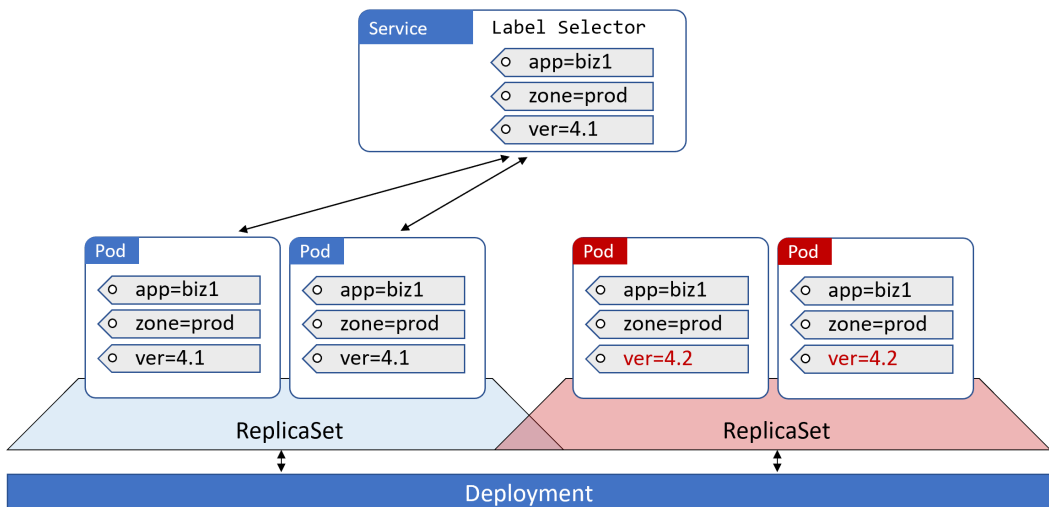


Figure 6.12

Now everybody's getting the old version.

This functionality can be used for all kinds of things – blue-greens, canaries, you name it. So simple,

yet so powerful.

Clean-up the lab with the following commands. These will delete the Deployment and Service used in the examples.

```
$ kubectl delete -f deploy.yml
$ kubectl delete -f svc.yml
```

Chapter Summary

In this chapter, we learned that *Services* bring stable and reliable networking to apps deployed on Kubernetes. They also perform load-balancing and allow us to expose elements of our application to the outside world (outside of the Kubernetes cluster).

The front-end of a Service is fixed, providing stable networking for the Pods behind it. The back-end of a Service is dynamic, allowing Pods to come and go without impacting the ability of the Service to provide load-balancing.

Services are first-class objects in the Kubernetes API and can be defined in the standard YAML manifest files. They use label selectors to dynamically match Pods, and the best way to work with them is declaratively.

7: Kubernetes storage

Storage is critical to most real-world production applications. Fortunately, Kubernetes has a mature and feature-rich storage subsystem called the *persistent volume subsystem*.

We'll divide this chapter as follows:

- The big picture
- Storage provisioners
- The Container Storage Interface (CSI)
- The Kubernetes persistent volume subsystem
- Storage Classes and Dynamic Provisioning
- Demo

The big picture

First things first, Kubernetes supports lots of types of storage from lots of different places. For example, iSCSI, SMB, NFS, and object storage blobs, all from a variety of external storage systems that can be in the cloud or in your on-premises datacenter. However, no matter what type of storage you have, or where it comes from, when it's surfaced on your Kubernetes cluster it's called a ***volume***. For example, Azure File resources surfaced in Kubernetes are called *volumes*, as are block devices from AWS Elastic Block Store. All storage on a Kubernetes cluster is called a *volume*.

Figure 7.1 shows the high-level architecture.

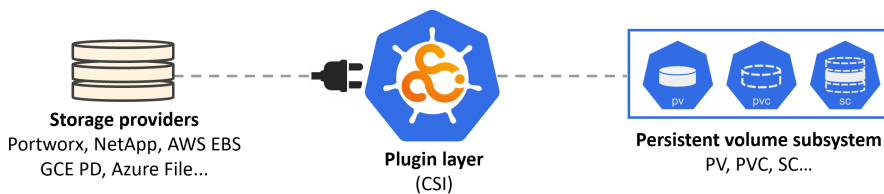


Figure 7.1

On the left, we've got storage providers. They can be your traditional enterprise storage arrays from vendors like EMC and NetApp, or they can be cloud storage services such as AWS Elastic Block Store (EBS) and GCE Persistent Disks (PD). All you need, is a plugin that allows their storage to be surfaced as volumes in Kubernetes.

In the middle of the diagram is the plugin layer. In the simplest terms, this is the glue that connects external storage with Kubernetes. Going forward, plugins will be based on the Container Storage Interface (CSI) which is an open-standard aimed at providing a clean interface for plugins. If you're a developer writing storage plugins, the CSI abstracts the internal Kubernetes storage detail and lets you develop *out-of-tree*.

Note: Prior to the CSI, all storage plugins were implemented as part of the main Kubernetes code tree (*in-tree*). This meant they all had to be open-source, and all updates and bug-fixes were tied to the main Kubernetes release-cycle. This was a nightmare for plugin developers as well as the Kubernetes maintainers. However, now that we have the CSI, storage vendors no longer need to open-source their code, and they can release updates and bug-fixes against their own timeframes.

On the right of Figure 7.1 is the Kubernetes persistent volume subsystem. This is a set of API objects that allow applications to consume storage. At a high-level, Persistent Volumes (PV) are how we map external storage onto the cluster, and Persistent Volume Claims (PVC) are like tickets that authorize applications (Pods) to use a PV.

Let's assume the quick example shown in Figure 7.2.

A Kubernetes cluster is running on AWS, and the AWS administrator has created a 25GB EBS volume called "ebs-vol". The Kubernetes administrator creates a PV called "k8s-vol" that links back to the "ebs-vol" via the `kubernetes.io/aws-ebs` plugin. While that might sound complicated, it's not. The PV is simply a way of representing the external storage on the Kubernetes cluster. Finally, the Pod uses a PVC to claim the PV and start using it.

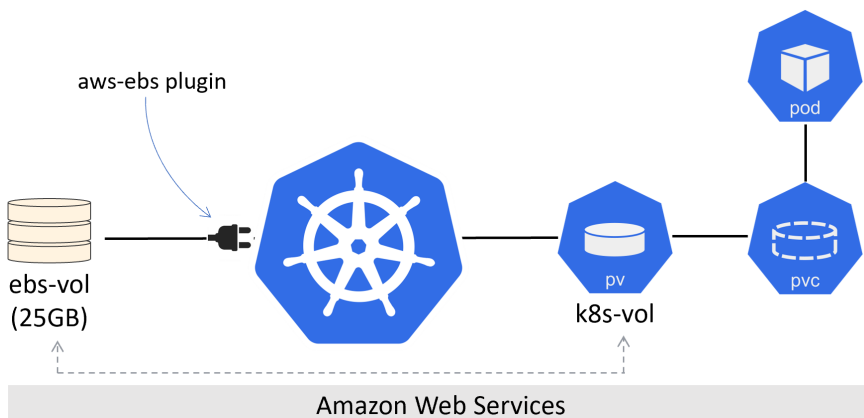


Figure 7.2

A couple of points worth noting.

1. There are rules safeguarding access to a single volume from multiple Pods (more on this later).

2. A single external storage volume can only be used by a single PV. For example, you cannot have a 50GB external volume that has two 25GB Kubernetes PVs each using half of it.

Now that we have an idea of the fundamentals, let's dig a bit deeper.

Storage Providers

Kubernetes can use storage from wide range external systems. These will often be native cloud services such as `AWSElasticBlockStore` or `AzureDisk`, but they can also be traditional on-premises storage arrays providing `iSCSI` or `NFS` volumes. Other options exist, but the take-home point is that Kubernetes gets its storage from a wide range of external systems.

Some obvious restrictions apply. For example, you cannot use the `AWSElasticBlockStore` provisioner if your Kubernetes cluster is running in Microsoft Azure.

The Container Storage Interface (CSI)

The CSI is an important piece of the Kubernetes storage jigsaw. However, unless you're writing storage plugins, you're unlikely to interact with it very often.

It's an open-source project that defines a standards-based interface so that storage can be leveraged in a uniform way across multiple container orchestrators. In other words, a storage vendor *should* be able to write a single CSI plugin that works across multiple orchestrators like Kubernetes and Docker Swarm.

In the Kubernetes world, the CSI is the preferred way to write drivers and means that plugin code no longer needs to exist in the main Kubernetes code tree. It also provides a clean and simple interface that abstracts all the complex internal Kubernetes storage machinery.

From a day-to-day management perspective, your only real interaction with the CSI will be referencing the appropriate plugin in your YAML manifest files. Also, it may take a while for existing in-tree plugins to be replaced by CSI plugins.

Sometimes we call plugins "*provisioners*", especially when we talk about Storage Classes later in the chapter.

The Kubernetes persistent volume subsystem

From a day-to-day perspective, this is where you'll spend most of your time configuring and interacting with Kubernetes storage.

You start out with raw storage on the left of Figure 7.3. This *plugs in* to Kubernetes via a CSI plugin. You then use the resources provided by the persistent volume subsystem to leverage the storage in your apps.

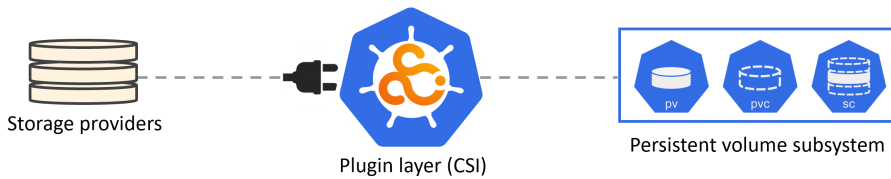


Figure 7.3

The three main resources in the persistent volume subsystem are:

- Persistent Volumes (PV)
- Persistent Volume Claims (PVC)
- Storage Classes (SC)

At a high level, **PVs** are how we represent storage in Kubernetes. **PVCs** are like tickets that let a Pod use a PV. **SCs** make it all dynamic.

Let's walk through a quick example.

Assume that you have a Kubernetes cluster and an external storage system. The storage vendor provides a CSI plugin so that you can leverage its storage assets inside of your Kubernetes cluster. You provision 3 x 10GB volumes on the storage array and create 3 Kubernetes PV objects to make them available to the cluster. Each PV references one of the volumes on the storage array via the CSI plugin. At this point, the three volumes are visible and available for use on the Kubernetes cluster.

Now assume you're about to deploy an application that requires 10GB of storage. That's great, you already have three 10GB PVs. In order for the app to use one of them, it needs a PVC. As previously mentioned, a PVC is like a ticket that lets a Pod (application) use a PV. Once the app has the PVC, it can mount the respective PV into its Pod as a volume. Refer back to Figure 7.2 if you need a visual representation.

That was a high-level example. Let's do it.

This example is for a Kubernetes cluster running on the Google Cloud. I'm using a cloud option as they're the easiest to follow along with and you *may* be able to use the cloud's free tier/initial free credit. It's also possible to follow along on other clouds by changing a few values.

The example assumes 10GB SSD volume called "uber-disk" has been pre-created in the same Google Cloud Region or Zone as the cluster. The Kubernetes steps will be:

1. Create the PV
2. Create the PVC
3. Mount the volume into a Pod

The following YAML file creates a PV object that maps back to the pre-created Google Persistent Disk called "uber-disk". The YAML file is available in the `storage` folder of the book's GitHub repo called `gke-pv.yml`.

```

apiVersion: v1
kind: PersistentVolume
metadata:
  name: pv1
spec:
  accessModes:
    - ReadWriteOnce
  storageClassName: test
  capacity:
    storage: 10Gi
  persistentVolumeReclaimPolicy: Retain
  gcePersistentDisk:
    pdName: uber-disk

```

Let's step through the file.

PersistentVolume resources are defined in `v1` of the *core* API group. We're naming this PV "pv1", setting its access mode to `ReadWriteOnce`, and making it part of a class of storage called "test". We're defining it as a 10GB volume, setting a reclaim policy, and mapping it back to a pre-created GCE persistent disk called "uber-disk".

The following command will create the PV. It assumes the YAML file is in your `PATH` and is called `gke-pv.yml`. The operation will fail if you have not pre-created "uber-disk" on the back-end storage system (in this example the back-end storage is provided by Google Compute Engine).

```

$ kubectl apply -f gke-pv.yml
persistentvolume/pv1 created

```

Check the PV exists.

```

$ kubectl get pv pv1

```

NAME	CAPACITY	MODES	RECLAIM POLICY	STATUS	STORAGECLASS ...
pv1	10Gi	RWO	Retain	Available	test

If you want, you can see more detailed information with `kubectl describe pv pv1`, but at the moment we have what is shown in Figure 7.4.

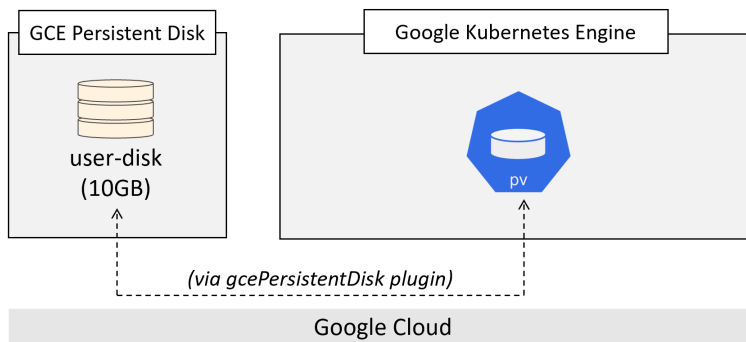


Figure 7.4

Let's quickly explain some of the PV properties set out in the YAML file.

`.spec.accessModes` defines how the PV can be mounted. Three options exist:

- `ReadWriteOnce (RWO)`
- `ReadWriteMany (RWM)`
- `ReadOnlyMany (ROM)`

`ReadWriteOnce` defines a PV that can only be mounted/bound as R/W by a single PVC. Attempts to bind it via multiple PVCs will fail.

`ReadWriteMany` defines a PV that can be bound as R/W by multiple PVCs. This mode is usually only supported by file and object storage, block storage normally only supports `RWO`.

`ReadOnlyMany` defines a PV that can be bound by multiple PVCs as R/O.

A couple of things are worth noting. First up, a PV can only be opened in one mode - it is not possible for a PV to have a PVC bound to it in ROM mode, and another PVC bound to it in RWM mode. Second up, Pods do not act directly on PVs, they always act on the PVC object that is bound to the PV.

`.spec.storageClassName` tells Kubernetes to group this PV in a storage class called "test". We'll learn more about storage classes later in the chapter, but we need this here to make sure the PV will correctly bind with a PVC in a later step.

Another property is `spec.persistentVolumeReclaimPolicy`. This tells Kubernetes what to do with a PV when its PVC has been released. Two policies currently exist:

- `Delete`
- `Retain`

`Delete` is the most dangerous, and is the default for PVs that are created dynamically via *storage classes* (more on these later). This policy deletes the PV **and associated storage resource on the external storage system**, so will result in data loss! You should obviously use this policy with caution.

Retain will keep the associated PV object on the cluster as well as any data stored on the associated external asset. However, it will prevent another PVC from using the PV.

If you want to re-use a *retained* PV, you need to perform the following three steps:

1. Manually delete the PV on Kubernetes
2. Re-format the associated storage asset on the external storage system to wipe any data
3. Recreate the PV

Tip: If you are experimenting in a lab and re-using PVs, it's easy to forget that you will have to perform the previous three steps.

`.spec.capacity` tells Kubernetes how big the PV should be. This value can be less than the actual physical storage asset but cannot be more. For example, you cannot create a 100GB PV that maps back to a 50GB device on the external storage system.

Finally, the last line of the YAML file links the PV to the name of the pre-created device on the back-end.

You can also specify vendor-specific attributes using the `.parameters` section of a PV YAML. We'll see more of this later when we look at *storage classes*, but for now, if your storage system supports pink fluffy NVMe devices, this is where you'd specify them.

Now that we've got a PV, let's create a PVC so that a Pod can claim access to the storage.

The following YAML defines a PVC that can be used by a Pod to gain access to the PV we've just created. The file is available in the `storage` folder in the book's GitHub repo called `gke-pvc.yaml`.

```
apiVersion: v1
kind: PersistentVolumeClaim
metadata:
  name: pvc1
spec:
  accessModes:
    - ReadWriteOnce
  storageClassName: test
  resources:
    requests:
      storage: 10Gi
```

As with the PV, PVCs are a stable v1 resource in the *core* API group.

The most important thing to note about a PVC object is that the values in the `.spec` section must match with the PV you intend to bind it with. In our example, *access modes*, *storage class*, and *capacity* must match.

Note: It is possible for a PV to have more capacity than a PVC. For example, a 10GB PVC can be bound to a 15GB PV (obviously this will waste 5GB of the PV). However, a 15GB PVC cannot be bound to a 10GB PV.

Figure 7.5 shows a side-by-side comparison of our example PV and PVC YAML files and highlights the properties that need to match.

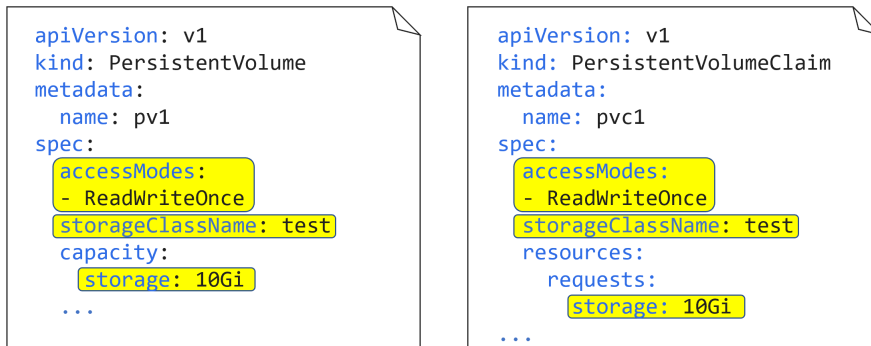


Figure 7.5

Deploy the PVC with the following command. It assumes the YAML file is called “gke-pvc.yml” and exists in your PATH.

```
$ kubectl apply -f gke-pvc.yml
persistentvolumeclaim/pvc1 created
```

Check that the PVC is created and bound to the PV.

```
$ kubectl get pvc pvc1
```

NAME	STATUS	VOLUME	CAPACITY	ACCESS MODES	STORAGECLASS
pvc1	Bound	pv1	10Gi	RWO	test

OK, we’ve got a PV representing 10GB of external storage on our Kubernetes cluster and we’ve bound a PVC to it. Let’s find out how a Pod can leverage that PVC and mount the actual storage.

More often than not, you’ll deploy your applications via higher-level controllers like *Deployments* and *StatefulSets*, but to keep the example simple, we’ll deploy a single *Pod*. Pods deployed like this are often referred to as “*singletons*” and are not recommended for production as they do not provide high availability and cannot self-heal.

The following YAML defines a single-container Pod with a volume called “data” that leverages the PVC and PV objects we already created. The file is available in the `storage` folder of the book’s GitHub repo called `volpod.yml`.

```
apiVersion: v1
kind: Pod
metadata:
  name: volpod
spec:
  volumes:
    - name: data
      persistentVolumeClaim:
        claimName: pvc1
  containers:
    - name: ubuntu-ctr
      image: ubuntu:latest
      command:
        - /bin/bash
        - "-c"
        - "sleep 60m"
      volumeMounts:
        - mountPath: /data
          name: data
```

We can see that the first reference to storage is `.spec.volumes`. This defines a volume called “data” that leverages our previously created PVC called “pvc1”.

You can run the `kubectl get pv` and `kubectl get pvc` commands to show that we’ve already created a PVC called “pvc1” that is bound to a PV called “pv1”. `kubectl describe pv pv1` will also prove that pv1 relates to a 10GB GCE persistent disk called “uber-disk”.

Deploy the Pod with the following command.

```
$ kubectl apply -f volpod.yml
pod/volpod created
```

You can run a `kubectl describe pod volpod` command to see that the Pod is successfully using the data volume and the pvc1 claim.

Time for a quick summary before we look at how we can make all of this dynamic with *storage classes*.

We start out with a storage assets on an external storage system. We use a CSI plugin to make the external storage system work with Kubernetes, and we use Persistent Volume (PV) objects to make the external systems assets accessible and usable. Each PV is an object on the Kubernetes cluster that maps back to a specific storage asset (LUN, share, blob...) on the external storage system. Finally, for a Pod to use a PV, it needs a Persistent Volume Claim (PVC). This is like a ticket that allows a Pod to use

a PV. Once the PV and PVC objects are created and bound, the PVC can be referenced in a PodSpec to mount the associated PV as a volume in a container.

Don't worry if this seems complicated, we'll pull it all together in a demo at the end of the chapter.

Storage Classes and Dynamic Provisioning

Everything we've seen so far is correct and fundamental to Kubernetes storage. But it doesn't scale – there's no way somebody managing a large Kubernetes environment can manually create and maintain large numbers of PVs and PVCs. We need something more dynamic. Enter *storage classes*...

As the name suggests, storage classes allow us to define different *classes*, or tiers, of storage. How you define your classes is arbitrary but will depend on the types of storage you have access to. For example, you might define a *fast* class, a *slow* class, and an *encrypted* class.

As far as Kubernetes goes, storage classes are defined as resources in the `storage.k8s.io/v1` API group. The resource type is `StorageClass`, and we define them in regular YAML files that we POST to the API server for deployment. You can use the `sc` shorthand to refer to `StorageClass` objects when using `kubectl`.

Note: You can see a full list of API resources, and their shortnames, using the `kubectl api-resources` command. The output of the command shows; the API group that each resource belongs to (an empty string indicates the *core* API group), if the resource is namespaced, and what its equivalent `kind` is when writing YAML files.

A StorageClass YAML

The following is a simple example of a `StorageClass` YAML file. It defines a class of storage called “fast”, that is based on AWS solid state drives (`io1`) in the Ireland Region (`eu-west-1a`). It also requests a performance level of 10 IOPs per gigabyte.

```
kind: StorageClass
apiVersion: storage.k8s.io/v1
metadata:
  name: fast
provisioner: kubernetes.io/aws-efs
parameters:
  type: io1
  zones: eu-west-1a
  iopsPerGB: "10"
```

As with all Kubernetes YAML, `kind` tells the API server what type of object is being defined, and `apiVersion` tells it which version of the schema to apply to the resource. `metadata.name` is an

arbitrary string value that lets you give the object a friendly name – in this example we’re defining a class called “fast”. `provisioner` tells Kubernetes which plugin to use, and the `parameters` field lets us finely tune the type of storage to leverage from the back-end.

A few quick things worth noting:

1. StorageClass objects are immutable, meaning you cannot modify them once deployed
2. `metadata.name` should be meaningful as it’s how other objects will refer to the class
3. We use the terms *provisioner* and *plugin* interchangeably
4. The `parameters` section is for plugin-specific values, and each plugin is free to support its own set of values. Configuring this section requires knowledge of the storage plugin and associated storage back-end.

Multiple StorageClasses

You can configure as many StorageClass objects as you need. However, each one relates to a single provisioner. For example, if you have a Kubernetes cluster with StorageOS and Portworx storage back-ends, you will need at least two StorageClass objects. That said, each back-end can offer multiple classes/tiers of storage, each of which can have its own StorageClass. For example, you could have the following two StorageClass objects for different classes of storage **from the same back-end**:

1. “db-secure” for encrypted database volumes
2. “db-basic” for unencrypted database volumes

An example of a StorageClass defining an encrypted volume on a Portworx back-end might look like the following. It will only work if you have a Portworx.

```
kind: StorageClass
apiVersion: storage.k8s.io/v1
metadata:
  name: portworx-db-secure
provisioner: kubernetes.io/portworx-volume
parameters:
  fs: "xfs"
  block_size: "32"
  repl: "2"
  snap_interval: "30"
  io-priority: "medium"
  secure: "true"
```

As we can see, the `.parameters` section requires knowledge of the plugin and what is supported on the storage back-end. Consult your storage plugin documentation for details.

Implementing StorageClasses

The basic workflow for deploying *and using* a StorageClass on your cluster is as follows:

1. Create your Kubernetes cluster with a storage back-end
2. Ensure the plugin for the storage back-end is available
3. Create a StorageClass object
4. Create a PVC object that references the StorageClass by name
5. Deploy a Pod that uses volume based on the PVC

Notice that the workflow does include creating a PV. This is because storage classes create PVs dynamically.

The following YAML snippet contains the definitions for a StorageClass, a PersistentVolumeClaim, and a Pod. All three objects can be defined in a single YAML file by separating each one with three dashes.

Notice how the PodSpec references the PVC by name, and in turn, the PVC references the SC by name.

```
kind: StorageClass
apiVersion: storage.k8s.io/v1
metadata:
  name: fast # Referenced by the PVC
provisioner: kubernetes.io/gce-pd
parameters:
  type: pd-ssd
---
apiVersion: v1
kind: PersistentVolumeClaim
metadata:
  name: mypvc # Referenced by the PodSpec
  namespace: mynamespace
spec:
  accessModes:
    - ReadWriteOnce
  resources:
    requests:
      storage: 50Gi
  storageClassName: fast # Matches name of the SC
---
apiVersion: v1
kind: Pod
```

```
metadata:
  name: mypod
spec:
  volumes:
    - name: data
      persistentVolumeClaim:
        claimName: mypvc # Matches PVC name
  containers: ...
<SNIP>
```

The previous YAML is truncated and does not include a full PodSpec.

So far, we've seen a few SC definitions. However, each one has been slightly different as each one has related to a different provisioner (storage plugin/back-end). You will need to refer to the documentation of your storage plugin to know which options your provisioners support.

Let's quickly summarize what we've learned about storage classes before walking through a demo.

StorageClasses make it so that we don't have to create PVs manually. We create the StorageClass object and use a plugin to tie it to a particular type of storage on a particular storage back-end. For example, high-performance AWS SSD storage in the AWS Mumbai Region. It needs a name and is defined in a YAML file that we deploy using `kubectl`. Once deployed, the StorageClass watches the API server for new PVC objects that reference its name. When matching PVCs appear, the StorageClass dynamically creates the required PV.

There's always more detail, such as *mount options* and *volume binding modes*, but what we've learned so far is enough to get you more than started.

Let's bring everything together with a demo.

Demo

In this section, we'll walk through a demo that uses a StorageClass. The basic steps of the demo will be:

1. Create a StorageClass
2. Create a PVC
3. Create a Pod

The Pod will mount a volume using the PVC, which in turn will trigger the SC to dynamically create a PV. The demo will be on the Google Cloud Platform and assumes you have a working cluster with `kubectl` correctly configured.

Clean-up

If you've been following along, you'll have a Pod, a PVC, and PV already created. Let's delete these before we go ahead with the demo.

```
$ kubectl delete pods volpod
pod "volpod" deleted
```

```
$ kubectl delete pvc pvc1
persistentvolumeclaim "pvc1" deleted
```

```
$ kubectl delete pv pv1
persistentvolume "pv1" deleted
```

Create a StorageClass

We'll use the following YAML to create a StorageClass called "slow" based on Google GCE standard persistent disks. We won't get into the details of the storage back-end, but suffice to say it's a slow tier of disk. The YAML also sets the reclaim policy so that data will not be lost when PVC bindings are released. Finally, we use an annotation to attempt to set this as the default storage class on the cluster.

Here's the YAML file, it's available in the storage folder of the book's GitHub repo called `google-sc.yml`.

```
kind: StorageClass
apiVersion: storage.k8s.io/v1
metadata:
  name: slow
  annotations:
    storageclass.kubernetes.io/is-default-class: "true"
provisioner: kubernetes.io/gce-pd
parameters:
  type: pd-standard
reclaimPolicy: Retain
```

Two things to note before we deploy the SC:

1. This lab was tested on GKE using Kubernetes 1.12.5. Even though Kubernetes 1.13 is available, 1.12.5 was the latest version available on GKE.

2. Setting default storage classes on the tested version of Kubernetes is done via annotations. However, this will likely change as the feature matures. You may have limited success with this setting depending on your version of Kubernetes.

Deploy the SC with the following command:

```
$ kubectl apply -f .\google-sc.yml
storageclass.storage.k8s.io/slow created
```

You can check and inspect it with `kubectl get sc slow` and `kubectl describe sc slow`. For example:

```
$ kubectl get sc slow
NAME                PROVISIONER             AGE
slow (default)      kubernetes.io/gce-pd    32s
```

Create a PVC

We'll use the following YAML to create a PVC object that references the `slow` StorageClass created in the previous step. The YAML is available in the `storage` folder of the book's GH repo called `google-pvc.yml`.

```
apiVersion: v1
kind: PersistentVolumeClaim
metadata:
  name: pv-ticket
spec:
  accessModes:
    - ReadWriteOnce
  storageClassName: slow
  resources:
    requests:
      storage: 25Gi
```

The important things to note are that the PVC is called `pv-ticket`, it's linked to the `slow` class, and it's for a 25GB volume.

Let's deploy it.


```
$ kubectl apply -f google-pvc.yml
persistentvolumeclaim/pv-ticket created
```

Verify the operation with a `kubectl get pvc`.

```
$ kubectl get pvc pv-ticket
```

NAME	STATUS	VOLUME	CAPACITY	ACCESS MODES	STORAGECLASS
pv-ticket	Bound	pvc-881a23...	25Gi	RWO	slow

Notice that the PVC is already bound to the `pvc-881a23...` volume – we didn’t have to manually create a PV. The mechanics behind the binding are as follows:

1. We created the `slow` StorageClass
2. A loop was created to watch the API server for new PVCs referencing the `slow` StorageClass
3. We created the `pv-ticket` PVC that requested binding to a 25GB volume from the **slow** StorageClass
4. The StorageClass noticed this PVC and dynamically created the requested PV

Use the following command to verify the presence of the automatically created PV on the cluster.

```
$ kubectl get pv
```

NAME	CAPACITY	Mode	STATUS	CLAIM	STORAGECLASS
pvc-881...	25Gi	RWO	Bound	pv-ticket	slow

Some of the columns have been trimmed from the output to better fit the book.

The following YAML defines a single-container Pod with a volume called `data`, mounted at `/data` based on the `pv-ticket` PVC. The YAML file is in the storage folder of the book’s GitHub repo called `google-pod.yml`.

```
apiVersion: v1
kind: Pod
metadata:
  name: class-pod
spec:
  volumes:
    - name: data
      persistentVolumeClaim:
        claimName: pv-ticket
  containers:
    - name: ubuntu-ctr
```

```
image: ubuntu:latest
command:
- /bin/bash
- "-c"
- "sleep 60m"
volumeMounts:
- mountPath: /data
  name: data
```

Deploy the Pod with `kubectl apply -f google-pod.yml`.

Congratulations. You’ve deployed a new default StorageClass and used a PVC to dynamically create a PV. You also have a Pod that has mounted the PVC as a volume in a container.

Clean-up

If you’ve followed along with the demo, you’ll have a Pod called “class-pod” with a volume using the “pv-ticket” PVC that was dynamically created via the “slow” SC. The following commands will delete all of these objects.

```
$ kubectl delete pod class-pod
pod "class-pod" deleted
```

```
$ kubectl delete pvc pv-ticket
persistentvolumeclaim "pv-ticket" deleted
```

```
$ kubectl delete sc slow
storageclass.storage.k8s.io "slow" deleted
```

Using the default StorageClass

One last thing...

If your cluster has a *default storage class*, you can deploy a Pod using just a PodSpec and a PVC. You do not need to manually create a StorageClass. However, real-world production clusters will usually have multiple StorageClasses, so it’s best practice to create and manage StorageClasses that suit your business and application needs. The default StorageClass is normally only useful in development environments and times when you do not have specific storage requirements.

Chapter Summary

In this chapter, we've learned that Kubernetes has a powerful storage subsystem that allows it to leverage storage from a wide variety of external storage back-ends.

Each back-end requires a plugin so that its storage assets can be used on the cluster, and the preferred type of plugin is a CSI plugin. Once a plugin is enabled, Persistent Volumes (PV) are used to represent external storage resources within the cluster, and Persistent Volume Claims (PVC) are used to give Pods access to PV storage.

Storage Classes take things to the next level by allowing applications to dynamically request storage. You create a Storage Class object that references a class, or tier, of storage from a storage back-end. Once created, the Storage Class watches the API server for new Persistent Volume Claims that reference it by name. When a matching PVC arrives, the SC dynamically creates the storage and makes it available as a PV that can be mounted as a volume into a Pod (container).

8: Other important Kubernetes stuff

There's a lot more to Kubernetes than can fit in a book. If we tried, we'd fill volumes!

In this chapter, we'll briefly mention some important Kubernetes features and projects that will take you to the next level:

- DaemonSets
- StatefulSets
- Jobs
- CronJobs
- Autoscaling
- RBAC
- Helm

Many of these will become their own chapters in future editions of the book. For now, let's have a taste of each.

DaemonSets

DaemonSets manage Pods and are a resource in the `apps` API group. They're useful when you need a replica of a particular Pod running on every node in the cluster. Some examples include; *monitoring Pods* and *logging Pods* that you need to run on every node in the cluster.

As you'd expect, it implements a controller and a watch loop. This means that you can dynamically add and remove nodes from the cluster, and the DaemonSet will ensure you always have one Pod replica on each of them.

The following command shows two DaemonSets in the `kube-system` namespace that exist on a newly installed 3-node cluster.

The output is trimmed so that it fits the page.

```
kubect1 get ds -n kube-system
```

NAME	DESIRED	CURRENT	READY	NODE SELECTOR
kube-proxy	3	3	3	beta.kubernetes.io/arch=amd64
weave-net	3	3	3	<none>

Notice that *desired state* for each DaemonSet is 3 replicas. You do not need to specify this in the DaemonSet YAML file as it is automatically implied based on the number of nodes in the cluster.

As well as the default behaviour of running one Pod replica on every cluster node, you can also run DaemonSets against subsets of nodes.

DaemonSets are stable in the `apps/v1` API group and can be managed with the usual `kubectl get`, and `kubectl describe` commands. If you already understand Pods and Deployments, you will find DaemonSets really simple.

StatefulSets

StatefulSets are a stable resource in the `apps/v1` API group. Their use-case is stateful components of your application, such as Pods that are not intended to be ephemeral and need more order than is provided by something like a Deployment.

Stateful components of a microservices application are usually the hardest to implement, and platforms like Kubernetes have been somewhat slow to implement features to handle them. StatefulSets are step towards improving this.

In many ways, StatefulSets are like Deployments. For example, we define them in a YAML file that we POST to the API server as desired state. A controller implements the work on the cluster and a background watch loop makes sure current state matches desired state. However, there are several significant differences. These include:

- StatefulSets give Pods deterministic meaningful names. Deployments do not.
- StatefulSets always start and delete Pods in a specific order. Deployments do not.
- Pods deployed via StatefulSet are not interchangeable. Pods deployed by Deployments are interchangeable.

Let's quickly look at each point a bit closer.

When a Pod is created by a Deployment, its name is a combination of the name of the Deployment plus a hash. When a Pod is created by a StatefulSet, its name is a combo of the name of the StatefulSet plus an integer. The first Pod deployed by a StatefulSet gets integer 1, the second gets integer 2 and so on. This effectively names Pods according to the order they were created. Scaling up a StatefulSet will cause the new Pod to get the next integer in the list, and scaling down a StatefulSet will start by deleting the highest numbered Pod. Finally, when a Pod managed by a StatefulSet fails, it is replaced by another Pod with the same name, ID, and IP address.

Potential use-cases for StatefulSets are any services in your application that maintain state. These can include:

- Pods that require access to specific named volumes
- Pods that require a persistent network identity

- Pods that must come online in a particular order

A StatefulSet guarantees all of these will be maintained across Pod failures and subsequent rescheduling operations.

Due to the more complex nature of stateful applications, StatefulSets can be complex to configure.

In summary, StatefulSets ensure a deterministic order for Pod creation and deletion based on the meaningful name of each managed Pod.

Jobs and CronJobs

Jobs, a.k.a. batch jobs, are stable resources in the `batch/v1` API group. They are useful when you need to run a specific number of a particular Pod, and you need guarantees that they'll all successfully complete.

A couple of subtleties worth noting:

1. Jobs don't have the concept of desired state
2. Pods that are part of a Job are short-lived

These two concepts separate *Jobs* from other objects like Deployments, DaemonSets, and StatefulSets. Whereas those objects keep a specified number of a certain Pod running indefinitely, *Jobs* manage a specified number of a certain Pod and make sure they complete and exit successfully.

The *Job* object implements the usual controller and watch loop. If a Pod that the *Job* object spawns fails, the *Job* will create another in its place. Once all the Pods managed by a Job complete, the Job itself completes.

Use-cases include typical batch-type workloads.

Interestingly, *Jobs* can be useful even if you only need to run a single Pod through to completion. Basically, any time you need to run one or more short-lived Pods, and you need to guarantee they complete successfully, the *Job* object is your friend!

CronJobs are just *Jobs* that run against a time-based schedule.

Autoscaling

The Deployments chapter showed us how to manually scale the number of Pod replicas. However, manually scaling a set of Pods does not scale (excuse the pun). As an example, if demand on your application spikes at 4:20 a.m. it's far from ideal if you need to page an operator who will then log-on to the cluster and manually increase the number of replicas. The same applies if you need to scale the number of *nodes* to your cluster.

With these challenges in mind, Kubernetes offers several auto-scaling technologies.

The **Horizontal Pod Autoscaler (HPA)** dynamically increases and decreases the number of *Pods* in a Deployment based on demand.

The **Cluster Autoscaler (CA)** dynamically increases and decreases the number of *nodes* in your cluster based on demand.

The **Vertical Pod Autoscaler (VPA)** attempts to right-size your Pods, but it's currently an *alpha* product.

Horizontal Pod Autoscaler (HPA)

HPA's are stable resources in the `autoscaling/v1` API group and their job is to scale the number of replicas in a Deployment based on observed CPU metrics. At the time of writing, the `autoscaling/v2` API is being worked on and will allow scaling based on more than just CPU.

It works like this... You define a Deployment that makes use of Pod resource requests – where each container in the Pod requests an amount of CPU. You deploy this to the cluster. You also create an HPA object that targets that Deployment and has a rule that says something like: *if any Pod in this Deployment uses more than 60% of its requested CPU, spin up an additional Pod.*

Once the Deployment and HPA are deployed to the cluster, scaling operations become automatic.

One thing worth noting is that HPAs update the `.spec.replicas` field of the targeted Deployment. While this update is recorded against the Deployment object in the cluster store, it can lead to situations where the copy of the Deployment YAML file in your external version control system gets out of sync with what is currently observed on the cluster.

Cluster Autoscaler (CA)

CAs are all about right-sizing your Kubernetes cluster. At a high-level, they increase and decrease the number of nodes in your cluster based on demand.

Getting under the covers a little... CAs periodically check Kubernetes for any Pods that are in the *pending* state due to lack of node resources. If it finds any, it adds nodes to the cluster so that the pending Pod(s) can be scheduled.

This requires integrations with your cluster's underlying infrastructure platform via a public API that allows Kubernetes to add and remove nodes (cloud instances). The major cloud platforms implement Cluster Autoscaler with varying levels of support. Check your cloud provider documentation for the latest support info.

Role-based access control (RBAC)

Kubernetes implements a least-privilege RBAC subsystem. When enabled, it locks down a cluster and allows you to grant permissions based on specific users and groups.

The model is based on three major components:

- Subjects
- Operations
- Resources

Subjects are users and groups, and these must be managed outside of Kubernetes. *Operations* are what the subject is allowed to do (create, list, delete etc.). *Resources* are objects on the cluster such as Pods. Put the three together, and you have an RBAC rule. For example, **Abi** (subject) is allowed to **create** (operation) **Pods** (resource).

RBAC has been stable (v1) since Kubernetes 1.8 and leverages two objects that are defined in the `authorization.rbac.k8s.io` API group. The two objects are `Roles` and `RoleBindings`. The `Role` is where you define the resource and the operation that you want to allow, and the `RoleBinding` connects it with a subject.

Helm

Helm is the de facto Kubernetes package manager and greatly simplifies installation and management of Kubernetes applications.

Helm was accepted into the Cloud Native Computing Foundation (CNCF) in 2018 as an official top-level project. As such, it sits alongside Kubernetes, Prometheus, gRPC, and others.

As a *package manager*, it's like `apt` for Ubuntu and `brew` for Mac. In the case of Helm, it hides all the complexities of things like Deployments, Pods, and Persistent Volumes in a construct called a *chart*. Think of a chart as the Helm equivalent of a YUM, DEB, or homebrew package. As such, you can install, update, and delete Kubernetes applications via the application's Helm chart.

You can also share your charts with the community as well as re-use existing charts.

Chapter Summary

The idea of this chapter was to make you aware of *some* of the other important Kubernetes technologies so that you have an idea of where you might want to go next. However, Kubernetes is huge, and we haven't covered everything.

The plan going forward is to make some of these topics their own chapters in future versions of the book. If you can't wait for that, I already cover *Autoscaling*, *RBAC*, and *Storage* in my **Kubernetes Deep Dive** video course on acloud.guru.

<https://acloud.guru/learn/kubernetes-deep-dive>

9: Threat modeling Kubernetes

Security is more important than ever before, and Kubernetes is no exception. Fortunately, there are a lot of things that can be done to secure Kubernetes, and we'll cover some of them in the next chapter. However, before we do this, it's worth taking a moment to model some of the common threats.

Threat model

Threat modeling is the process of identifying vulnerabilities so that we can put measures in place to prevent and mitigate them. In this chapter, we'll look at the popular **STRIDE** model and see how it can be applied to Kubernetes.

STRIDE defines six categories of potential threat:

- Spoofing
- Tampering
- Repudiation
- Information disclosure
- Denial of service
- Elevation of privilege

While the model is good, it's important to keep in mind that no threat model guarantees to cover all possible threats. However, models like this are useful in giving us a structured way to look at an entire system.

For the rest of this chapter, we'll look at each of the six threat categories in turn. For each one, we'll give a quick description, and then look at some of the ways it applies to Kubernetes and how we can prevent and mitigate.

This chapter doesn't attempt to cover everything. It's intended to give you ideas and get you started.

Spoofing

Spoofing is pretending to be something, or somebody, you are not. In the context of information security, it's pretending to be a different user or entity, with the aim of gaining extra privileges on a system.

Let's look at how Kubernetes authenticates users to prevent spoofing.

Securing communications with the API server

Kubernetes is comprised of lots of small components that work together. These include control plane components such as the API server, controller manager, scheduler, cluster store, and others. It also includes node components such as the kubelet and container runtime. Each of these has its own set of privileges that allow it to interact with, and even modify the cluster. Even though Kubernetes implements a least-privilege model, spoofing the identity of any of these components can have unforeseen and potentially disastrous consequences.

Fortunately, Kubernetes implements a security model that requires components to authenticate via mutual TLS (mTLS). This requires both parties (the sender and the receiver) to authenticate each other via cryptographically signed certificates. This is good, and Kubernetes makes things easy by auto-rotating certificates etc. However, it's vital that you consider the following:

1. A typical Kubernetes installation will auto-generate a self-signed certificate authority (CA) during the bootstrap process. This is the CA that will issue certificates to all cluster components. It's better than nothing, but on its own it probably isn't enough for your production environment.
2. Mutual TLS is only as secure as the CA that issued the certificates. Compromising the CA can render the entire mTLS layer ineffective. So, keep the CA secure!

A good practice is to ensure that certificates issued by the internal Kubernetes CA are only used and trusted *within* the Kubernetes cluster. This requires careful approval of certificate signing requests, as well as making sure the Kubernetes CA is not added as a trusted CA for any components outside of Kubernetes.

As mentioned in previous chapters, all interaction with Kubernetes is via the API server and subject to authentication and authorization checks. This is true for internal and external components. As a result, the API server needs a way to authenticate (trust) internal and external sources. A good way to do this is to have two trusted key pairs – one for authenticating internal components and the other for authenticating external components. To accomplish this, Kubernetes leverages an internal self-signed CA to issue keys to internal components, as well as one or more trusted 3rd-party CAs to issue keys to external components (Kubernetes obviously needs configuring the trust 3rd-party CAs). This configuration ensures the API server trusts internal components possessing a certificate issued by the cluster's self-signed CA, as well as external components possessing a certificate signed by the 3rd-party CA.

Securing Pod communications

As well as spoofing access to the *cluster*, there is also the threat of spoofing an application for app-to-app communications. This is when one Pod spoofs another. Fortunately, we can leverage Kubernetes *Secrets* to mount certificates into Pods that can then be used to authenticate Pod identity.

While on the topic of Pods, every Pod has an associated `ServiceAccount` that is used to provide an identity for the Pod within the cluster. This is achieved by automatically mounting a service account token into every Pod as a *Secret*. Two points to note:

1. The service account token allows access to the API server
2. Most Pods probably don't need to access the API server

With these two points in mind, it is recommended to set `automountServiceAccountToken` to `false` for Pods that do not need to communicate with the API server. The following Pod manifest shows how to do this.

```
apiVersion: v1
kind: Pod
metadata:
  name: service-account-example-pod
spec:
  serviceAccountName: some-service-account
  automountServiceAccountToken: false
<Snip>
```

Tampering

Tampering is the act of changing something in a malicious way. In relation to information security, the goal of tampering is usually to cause one of the following:

- Denial of service. Tampering with the resource to make it unusable.
- Elevation of privilege. Tampering with a resource to gain additional privileges.

Tampering can be hard to avoid, so a common counter-measure is to make it obvious when something has been tampered with. A common example, outside of information security, is drug packaging. Most over-the-counter drugs are packaged with tamper-proof seals. These make it obvious to the consumer if the product has been tampered with because the tamper-proof seal has been broken.

Let's first look at some of the cluster components that can be tampered with.

Tampering with Kubernetes components

All of the following Kubernetes components, if tampered with, can cause harm:

- etcd
- Configuration files for the API server, controller-manager, scheduler, etcd, and kubelet
- Container runtime binaries
- Container images
- Kubernetes binaries

Generally speaking, tampering happens either *in transit* or *at rest*. In transit refers to data while it is being transmitted over the network, whereas at rest refers to data stored in memory or on disk.

TLS is a great tool for protecting against *in transit* tampering as it provides built-in integrity guarantees – You’ll be warned if the data has been tampered with.

The following recommendations can also help prevent tampering with data when it is *at rest* in a Kubernetes cluster:

- Restrict access to the servers that are running Kubernetes components – especially control plane components.
- Restrict access to repositories that store Kubernetes configuration files.
- Only perform remote bootstrapping over SSH (remember to safely guard your SSH keys).
- Always perform SHA-2 checksums on downloaded binaries.
- Restrict access to your image registry and associated repositories.

This isn’t an exhaustive list, but if you implement it, you will greatly reduce the chances of having your data tampered with while at rest.

As well as the items listed, it’s good production hygiene to configure auditing and alerting for important binaries and configuration files. If configured and monitored correctly, these can help detect potential tampering attacks.

The following example uses a common Linux audit daemon to audit access to the docker binary. It also audits attempts to change the binary’s file attributes.

```
auditctl -w /var/lib/docker -p rwx -k audit-docker
```

We’ll refer to this example later in the chapter.

Tampering with applications running on Kubernetes

As well as infrastructure components, application components are also potential tampering targets.

A good way to prevent a live Pod from being tampered with, is setting its filesystems to read-only. This guarantees filesystem immutability and can be accomplished through a Pod Security Policy or the securityContext section of a Pod’s manifest file.

Note: PodSecurityPolicy objects are a relatively new feature that allow us to force security settings on all Pods in a cluster, or targeted sub-sets of Pods. They’re a great way to enforce standards without developers and operations staff having to remember to do it for every individual Pod.

You can make a container's root filesystem read-only by setting the `readOnlyRootFilesystem` property to `true`. As previously mentioned, this can be set via a `PodSecurityPolicy` object, or in Pod manifest files. The same can be done for other filesystems that are mounted into containers via the `allowedHostPaths` property.

The following example shows how to use both settings in a Pod manifest. The `allowedHostPaths` section makes sure anything mounted beneath `/test` will be read-only.

```
apiVersion: v1
kind: Pod
metadata:
  name: readonly-test
spec:
  securityContext:
    readOnlyRootFilesystem: true
  allowedHostPaths:
    - pathPrefix: "/test"
      readOnly: true
```

The same can be implemented in a `PodSecurityPolicy` object as follows:

```
apiVersion: policy/v1beta1 # Will change in future versions
kind: PodSecurityPolicy
metadata:
  name: tampering-example
spec:
  readOnlyRootFilesystem: true
  allowedHostPaths:
    - pathPrefix: "/test"
      readOnly: true
```

Repudiation

At a very high level, *repudiation* is casting doubt on something. *Non-repudiation* is providing proof about something. In the context of information security, non-repudiation is **proving** certain actions were carried out by certain individuals.

Digging a little deeper, non-repudiation includes the ability to prove:

- What happened

- When it happened
- Who made it happen
- Where it happened
- Why it happened
- How it happened

Answering the last two usually requires the correlation of several events over a period of time.

Fortunately, auditing of Kubernetes API server events can usually help answer these questions. The following is an example of an API server audit event (you may need to manually enable auditing on your API server).

```
{
  "kind": "Event",
  "apiVersion": "audit.k8s.io/v1",
  "metadata": { "creationTimestamp": "2019-03-03T10:10:00Z" },
  "level": "Metadata",
  "timestamp": "2019-03-03T10:10:00Z",
  "auditID": "7e0cbccf-8d8a-4f5f-aefb-60b8af2d2ad5",
  "stage": "RequestReceived",
  "requestURI": "/api/v1/namespaces/default/persistentvolumeclaims",
  "verb": "list",
  "user": {
    "username": "fname.lname@example.com",
    "groups": [ "system:authenticated" ]
  },
  "sourceIPs": [ "123.45.67.123" ],
  "objectRef": {
    "resource": "persistentvolumeclaims",
    "namespace": "default",
    "apiVersion": "v1"
  },
  "requestReceivedTimestamp": "2019-03-03T10:10:00.123456Z",
  "stageTimestamp": "2019-03-03T10:10:00.123456Z"
}
```

Although the API server is central to most things in Kubernetes, it's not the only component that requires auditing for non-repudiation. At a minimum, you should also collect audit logs from container runtimes, kubelets, and the applications running on your cluster. This is without even mentioning network firewalls and the likes.

Once you start auditing multiple components, you quickly need a centralised location to store and correlate events. A common way to do this is deploying an agent to all nodes via a DaemonSet. The agent collects logs (runtime, kubelet, application...) and ships them to a secure central location.

If you do this, it's vital that the centralised log store is secure. If the security of the central log store is compromised, you can no longer trust the logs, and their contents can be repudiated.

To provide non-repudiation relative to tampering with binaries and configuration files, it might be useful to use an audit daemon that watches for write actions on certain files and directories on your Kubernetes masters and nodes. For example, earlier in the chapter we showed an example that enabled auditing of changes to the `docker` binary. With this enabled, starting a new container with the `docker run` command will generate an event like this:

```
type=SYSCALL msg=audit(1234567890.123:12345): arch=abc123 syscall=59 success=yes
exit=0 a0=12345678abc a1=0 a2=abc12345678 a3=a items=1 ppid=1234 pid=1234 au
id=0 uid=0 gid=0 euid=0 suid=0 fsuid=0 egid=0 sgid=0 fsgid=0 tty=pts0 ses=1 comm\
="docker" exe="/var/lib/docker" subj=system_u:object_r:container_runtime_exec_t\
:s0 key="audit-docker"
type=CWD msg=audit(1234567890.123:12345): cwd="/home/firstname"
type=PATH msg=audit(1234567890.123:12345): item=0 name="/var/lib/docker" inode=\
123456 dev=fd:00 mode=0100600 ouid=0 ogid=0 rdev=00:00 obj=system_u:object_r:co
ntainer_runtime_exec_t:s0
```

Audit logs like this, when combined and correlated with Kubernetes' audit features, create a comprehensive and trustworthy picture that cannot be repudiated.

Information Disclosure

Information disclosure is when sensitive data is leaked. There are lots of ways it can happen, from leaving an insecure USB drive on a plane, all the way to data stores being hacked and APIs that unintentionally expose sensitive data.

Protecting cluster data

In the Kubernetes world, the entire configuration of the cluster is stored in the cluster store (currently etcd). This includes network and storage configuration, as well as passwords and other sensitive data stored in Secrets. For obvious reasons, this makes the cluster store a prime target for information disclosure attacks.

As a minimum, you should limit *and* audit access to the nodes hosting the cluster store. As will be seen in the next paragraph, gaining access to a cluster node can allow the logged-on user to bypass some of the security layers.

Kubernetes 1.7 introduced encryption of Secrets but doesn't enable it by default. Even when this becomes default, the data encryption key (DEK) is stored on the same node as the Secret! This means that gaining access to a node allows you to bypass encryption. This is especially worrying on nodes that host the cluster store (etcd nodes).

Fortunately, Kubernetes 1.11 enabled a beta feature that lets you store *key encryption keys (KEK)* outside of the Kubernetes cluster. These types of key are used to encrypt and decrypt data encryption keys and should be safely guarded. You should seriously consider Hardware Security Modules (HSM) or cloud-based Key Management Stores (KMS) for storing your key encryption keys.

Keep an eye on upcoming versions of Kubernetes for further improvements to encryption of Secrets.

Protecting data in Pods

As previously mentioned, Kubernetes has an API resource called a Secret that is the preferred way to store and share sensitive data such as passwords. For example, a front-end container accessing an encrypted back-end database can have the key to decrypt the database mounted as a Secret. This is a far better solution than storing the decryption key in a plain-text file or environment variable.

It is also common to store data and configuration information outside of Pods and containers in Persistent Volumes and ConfigMaps. If the data on these is encrypted, keys for decrypting them should also be stored in Secrets.

With all of this, it's vital that you consider the caveats outlined in the previous section relative to Secrets and how their encryption keys are stored. You don't want to do the hard work of locking the house but leaving the keys in the door.

Denial of Service

Denial of Service (DoS) is all about making something unavailable. There are many types of DoS attack, but a well-known variation is overloading a system to the point it can no longer service requests. In the Kubernetes world, a potential attack might be to overload the API server so that cluster operations grind to a halt (even essential system services have to communicate via the API server).

Let's take a look at some potential Kubernetes systems that might be targets of DoS attacks, and some ways to protect and mitigate.

Protecting cluster resources against DoS attacks

It's a time-honored best practice to replicate essential control plane services on multiple nodes for high availability (HA). Kubernetes is no different, and you should run multiple master nodes in an HA configuration for your production environments. Doing this will prevent a single master from becoming a single point of failure. In relation to certain types of DoS attacks, an attacker would potentially need to attack more than one master to have a meaningful impact.

You should also consider replicating control plane nodes across availability zones. This may prevent a DoS attack on the *network* of a particular availability zone from taking down your entire control plane.

The same principle applies to worker nodes. Having multiple worker nodes allows the scheduler to spread your application over multiple nodes and availability zones. Not only might this allow the scheduler to run your application on a different node if the one it's currently running on is subject to a DoS attack. It also means that replicated parts of your application can be distributed over multiple nodes and zones, potentially rendering a DoS attack on any single node or zone ineffective (or less effective).

You should also configure appropriate limits for the following:

- Memory
- CPU
- Storage
- Kubernetes objects

Limiting Kubernetes object includes things like; limiting the number of ReplicaSets, Pods, Services, Secrets, and ConfigMaps in a particular namespace.

Placing limits on things can help prevent important system resources from being starved, therefore preventing potential DoS.

Here's an example manifest that limits the number of Pod objects in the `skippy` namespace to 100.

```
apiVersion: v1
kind: ResourceQuota
metadata:
  name: pod-quota
spec:
  hard:
    pods: "100"
```

Use the following command to apply it to the `skippy` namespace. The command assumes the manifest file is called `quota.yml`.

```
$ kubectl apply -f quota.yml --namespace=skippy
```

Protecting the API Server against DoS attacks

All communication in Kubernetes goes through the API server. The API server exposes a RESTful interface over a TCP socket, making it susceptible to botnet-based DoS attacks.

The following may be helpful in either preventing or mitigating such attacks.

- Highly available masters. Having multiple API server replicas running on multiple nodes across multiple availability zones.
- Monitoring and alerting of API server requests (based on sane thresholds)
- Not exposing the API server to the internet (firewall rules etc.)

As well as botnet DoS attacks, an attacker may also attempt to spoof a user or other control plane service in an attempt to cause an overload. Fortunately, Kubernetes has robust authentication and authorization controls in place to help prevent spoofing. However, even with a robust RBAC model, it is vital that you safeguard access to accounts with high privileges.

Protecting the cluster store against DoS attacks

Cluster configuration is stored in etcd, making it vital that etcd be available and secure. The following recommendations will help accomplish this:

- Configure an HA etcd cluster with either 3 or 5 nodes
- Configure monitoring and alerting of requests to etcd
- Isolate etcd at the network level so that only members of the control plane can interact with it

A default installation of Kubernetes will install etcd on the same servers as the rest of the control plane. This is usually fine for development and testing; however, large production clusters should seriously consider a dedicated etcd cluster. This will provide better performance and greater resilience.

On the performance front, etcd is probably the most common choking point for large Kubernetes clusters. With this in mind, you should perform testing to ensure the infrastructure it runs on is capable of sustaining performance at scale – a poorly performing etcd can be as bad as an etcd cluster under a sustained DoS attack. Operating a dedicated etcd cluster also provides additional resilience by protecting it from other parts of the control plane that might be compromised.

Monitoring and alerting of etcd should be based on sane thresholds, and a good place to start is by monitoring etcd log entries.

Protecting application components against DoS attacks

Most Pods expose their main service on the network, and without additional controls in place, anyone with access to the network can perform a DoS attack on the Pod. Fortunately, Kubernetes provides Pod resource request limits to prevent such attacks from exhausting Pod and node resources. As well as these, the following will be helpful:

- Define Kubernetes Network Policies that restrict Pod-to-Pod and Pod-to-external communications
- Utilize mutual TLS and API token-based authentication for application-level authentication (reject any unauthenticated requests)

For defence in depth, you should also implement application-layer authorization policies that implement least privilege.

Figure 9.1 shows how all of these can be combined to make it hard for an attacker to successfully DoS an application.

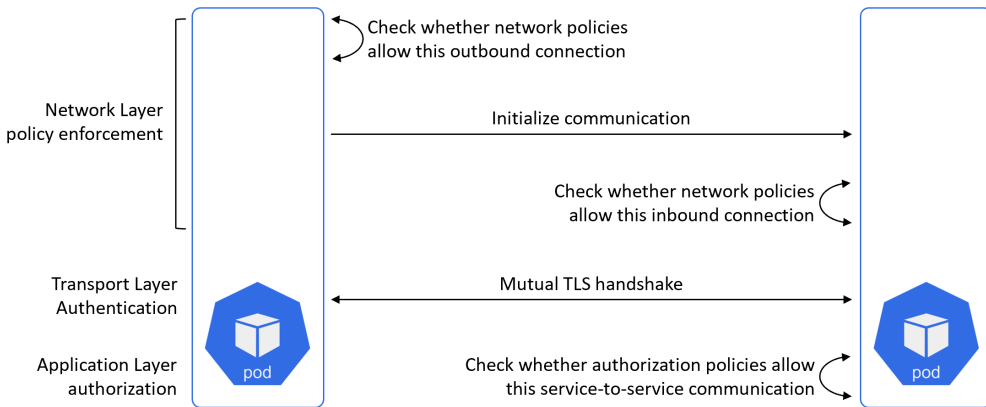


Figure 9.1

Elevation of privilege

Elevation of privilege, a.k.a. a privilege escalation, is gaining higher access than what is granted, usually in order to cause damage or gain unauthorized access.

Let's look at a few ways to prevent this in a Kubernetes environment.

Protecting the API server

Kubernetes offers several authorization modes that help safeguard access to the API server. These include:

- Role-based Access Control (RBAC)
- Webhook
- Node

You should run multiple authorizers at the same time. For example, a common best practice is to always have *RBAC* and *node* enabled.

RBAC mode lets us restrict API operations to sub-sets of users. These *users* can be regular user accounts as well as system services. The idea is that all requests to the API server must be authenticated **and** authorized. Authentication ensures that requests are coming from a validated user – the user performing the request is who they claim to be. Authorization ensures the validated user is allowed to perform the requested operation on the targeted cluster resource. For example, can *Lily* *create* *Pods*? In this example, *Lily* is the user, *create* is the operation, and *Pods* is the resource. Authentication makes sure that it really is Lily making the request, and authorization determines if she's allowed to create Pods.

Webhook mode lets you offload authorization to an external REST-based policy engine. However, it requires additional effort to build and maintain the external engine. It also makes the external engine a potential single-point-of-failure for every request to the API server. For example, if the external webhook system becomes unavailable, you may not be able to make any requests to the API server. With this in mind, you should be rigorous in vetting and implementing any webhook authorization service.

Node authorization is all about authorizing API requests made by kubelets (cluster nodes). The types of requests made to the API server by nodes is obviously different to those generally made by regular users, and the node authorizer is designed to help with this.

RBAC and node are two recommended authorization modes. RBAC mode is extremely configurable, and you should use it to implement a least privilege model for users accessing the API server. When implemented, it is a deny-by-default system that requires you to specifically grant individual permissions. If implemented well, it does an excellent job of ensuring users and Service Accounts do not have more access than required.

Protecting Pods

The next few sections will look at a few of the technologies that help reduce the risk of elevation of privilege attacks against Pods and containers. We'll look at the following:

- Preventing processes from running as `root`
- Dropping capabilities
- Filtering syscalls
- Preventing privilege escalation

As we proceed through the following sections, it's important to remember that a Pod is just an execution environment for one or more containers – application code runs in containers, which in turn, run inside of Pods. Some of the terminology used will refer to Pods and containers interchangeably, but usually we will mean container.

Do not run processes as root

The `root` user is the most powerful user on a Linux system and is always User ID 0 (UID 0). Therefore, running application processes as `root` is almost always a bad idea as it grants the application process full access to the container. This is made even worse by the fact that the `root` user of container often has unrestricted `root` access on the node as well. If that doesn't make you afraid, it should!

Fortunately, Kubernetes lets us force container processes to run as unprivileged non-root users.

The following Pod manifest configures all containers that are part of this Pod to run processes as UID 1000. If the Pod has multiple containers, all processes in all containers will run as UID 1000

```
apiVersion: v1
kind: Pod
metadata:
  name: demo
spec:
  securityContext: # Applies to all containers in this Pod
    runAsUser: 1000 # Non-root user
  containers:
  - name: demo
    image: example.io/simple:1.0
```

`runAsUser` is one of many settings that can be configured as part of what we refer to as a `PodSecurityContext` (`.spec.securityContext`).

It is possible for two or more Pods to be configured with the same `runAsUser` UID. When this happens, the containers from both Pods will run with the same security context and potentially have access to the same resources. This *might* be fine if they are replicas of the same Pod or container. However, there is a high chance that this will cause problems if they are different containers. For example, two different containers with R/W access to the same host directory can cause data corruption (both writing to the same dataset without co-ordinating write operations). Shared security contexts also increase the possibility of a compromised container tampering with a dataset it should not have access to.

With this in mind, it is possible to use the `securityContext.runAsUser` property at the container level instead of at the Pod level:

```
apiVersion: v1
kind: Pod
metadata:
  name: demo
spec:
  securityContext: # Applies to all containers in this Pod
    runAsUser: 1000 # Non-root user
  containers:
  - name: demo
    image: example.io/simple:1.0
    securityContext:
      runAsUser: 2000 # Overrides the Pod setting
```

This example sets the UID to 1000 at the Pod level but overrides it at the container level so that processes in one particular container run as UID 2000. Unless otherwise specified, all other containers in the Pod will use UID 1000.

A couple of other things that might help get around the issue of multiple Pods and containers using the same UID include:

- Enabling *user namespaces*
- Maintaining a map of UID usage

User namespaces is a Linux kernel technology that allows a process to run as root within a container but run as a different user outside of the container. For example, UID 0 (the root user) in the container gets mapped to UID 1000 on the host. This can be a good solution for processes that *need* to run as root inside the container, but you should check whether it has full support from your version of Kubernetes and your container runtime.

Maintaining a map of UID usage is a clunky way to prevent multiple different Pods and containers using overlapping UIDs. It's a bit of a hack and requires strict adherence to a gated release process for releasing Pods into production.

Note: A strict gated release process is a good thing for production environments. The *hacky* part of the previous section is the UID map itself, as well as the fact that you're introducing an external dependency and complicating releases and troubleshooting.

Drop capabilities

While *user namespaces* allow container processes to run as root inside the container but not on the host machine, it remains a fact that most processes do not need all of the privileges that the root has. However, it is equally true that many processes do require more privileges than a typical non-root

user has. What we need, is a way to grant the exact set of privileges a process requires in order to run. Enter *capabilities*.

Time for a quick bit of background...

We've already said that the `root` user is the most powerful user on a Linux system. However, its power is a combination of lots of small privileges that we call *capabilities*. For example, the `SYS_TIME` capability allows a user to set the system clock, whereas the `NET_ADMIN` capability allows a user to perform network-related operations such as modifying the local routing table and configuring local interfaces. The root user holds every *capability* and is therefore extremely powerful.

Having a modular set of *capabilities* like this allows us to be extremely granular when granting permissions. Instead of an all or nothing (root or non-root) approach, we can grant a process the exact set of privileges it requires to run.

There are currently over 30 capabilities and choosing the right ones can be daunting. With this in mind, an out-of-the-box Docker runtime drops over half of them by default. This is a *sensible-default* that is designed to allow most processes to run, without *leaving the keys in the front door*. While sensible defaults like these are better than nothing, they will usually not be enough for a lot of production environments.

A common way to find the absolute minimum set of capabilities an application requires, is to run it in a test environment with all capabilities dropped. This will cause the application to fail and log messages about the missing permissions. You map those permissions to *capabilities*, add them to the application's Pod spec, and run the application again. You rinse and repeat this process until the application runs properly with the minimum set of capabilities.

As good as this is, there are a few things to consider.

Firstly, you **must** perform extensive testing of your application. The last thing you want is a production edge case that you hadn't accounted for in your test environment. Such occurrences can crash your application in production!

Secondly, every fix and change to your application requires the exact same extensive testing against the capability set.

With these considerations in mind, it is vital that you have testing procedures and production release processes that can handle all of this.

By default, Kubernetes implements the default set of *capabilities* implemented by your chosen container runtime (E.g. containerd or Docker). However, you can override this in a Pod Security Policy, or as part of a container's `securityContext` field.

The following Pod manifest shows how to add the `NET_ADMIN` and `CHOWN` capabilities to a container.

```
apiVersion: v1
kind: Pod
metadata:
  name: capability-test
spec:
  containers:
  - name: demo
    image: example.io/simple:1.0
    securityContext:
      capabilities:
        add: ["NET_ADMIN", "CHOWN"]
```

Filter syscalls

seccomp is similar in concept to *capabilities* but works on syscalls rather than capabilities.

The way that Linux processes ask the kernel to perform an operation is by issuing a syscall to the kernel. *seccomp* lets us configure which syscalls a particular container can make to the host kernel. As with capabilities, a least privilege model is preferred where the only syscalls a container is allowed to make are the ones it needs to in order to run.

Be careful though, Linux has over 300 syscalls, and at the time of writing *seccomp* is an alpha feature in Kubernetes. You should also check support from your container runtime.

Prevent privilege escalation by containers

The only way to create a new process in Linux is for one process to clone itself and then load new instructions on to the new process. We're obviously over-simplifying, but the original process is called the *parent* process, and the copy is called the *child*.

By default, Linux allows a *child* process to claim more privileges than its *parent*. This is usually a bad idea. In fact, you will often want a child process to have the same, or less privileges than its parent. This is especially true for containers, as their security configurations are defined against their initial configuration, and not against potentially escalated privileges.

Fortunately, it's possible to prevent privilege escalation through a Pod Security Policy or the *securityContext* property of an individual container.

The following Pod manifest shows how to prevent privilege escalation for an individual container.


```
apiVersion: v1
kind: Pod
metadata:
  name: demo
spec:
  containers:
  - name: demo
    image: example.io/simple:1.0
    securityContext:
      allowPrivilegeEscalation: false
```

Pod Security Policies

As we've seen throughout the chapter, we can enable security settings on a per-Pod basis by setting security context attributes in individual Pod YAML files. However, this approach doesn't scale, requires developers and operators to remember to do this for every Pod, and is prone to errors. *Pod Security Policies* offer a better way.

Pod Security Policies are a relatively new feature that allow us to define security settings at the cluster level. We can then apply these to targeted sets of Pods as part of the deployment process. As such, this solution scales better, requires less work from developers and admins, and is less prone to error. It also lends itself to situations where you have a team dedicated to securing apps in production.

Pod Security Policies are implemented as an *admission controller*, and in order to use them, a Pod's serviceAccount must be authorized to use it. Once this is done, their policies are applied to new requests to create Pods as they pass through the API admission chain.

Pod Security Policy example

Let's finish the chapter with a quick look at an example of a Pod Security Policy that covers many of the points discussed in this chapter, as well as some other known secure defaults.

The example is based on an example from the [official Kubernetes docs](https://kubernetes.io/docs/concepts/policy/pod-security-policy/#example-policies)¹:

¹<https://kubernetes.io/docs/concepts/policy/pod-security-policy/#example-policies>

```

apiVersion: policy/v1beta1
kind: PodSecurityPolicy
metadata:
  name: restricted
  annotations:
    seccomp.security.alpha.kubernetes.io/allowedProfileNames: 'docker/default'
    apparmor.security.beta.kubernetes.io/allowedProfileNames: 'runtime/default'
    seccomp.security.alpha.kubernetes.io/defaultProfileName: 'docker/default'
    apparmor.security.beta.kubernetes.io/defaultProfileName: 'runtime/default'
spec:
  privileged: false
  allowPrivilegeEscalation: false # Prevent privilege escalation
  requiredDropCapabilities:
    - ALL # Drops all root capabilities (non-privileged user)
  # Allow core volume types.
  volumes:
    - 'configMap'
    - 'emptyDir'
    - 'projected'
    - 'secret'
    - 'downwardAPI'
    # Assume that PVs set up by the cluster admin are safe to use.
    - 'persistentVolumeClaim'
  hostNetwork: false # Prevent access to the host network namespace
  hostIPC: false # Prevent access to the host IPC namespace
  hostPID: false # Prevent access to the host PID namespace
  runAsUser:
    rule: 'MustRunAsNonRoot' # Prevent from running as root
  selinux:
    rule: 'RunAsAny' # Any SELinux options can be used
  supplementalGroups:
    rule: 'MustRunAs' # Allow all except root (UID 0)
    ranges:
      - min: 1
        max: 65535
  fsGroup:
    rule: 'MustRunAs' # Sets range for groups that own Pod volumes
    ranges:
      - min: 1
        max: 65535

```

```
readOnlyRootFilesystem: true # Force root filesystem to be R/O
```

There's no denying that configuring effective security policies is both important and challenging. A common practice is to start with a restrictive policy like the one just shown, then tweak it to fit your requirements. A lot of experimenting will be required.

It may also be a good idea to configure several Pod Security Policies that vary in how restrictive they are, then allow development teams to work with cluster administrators to choose the one that best fits the application.

Summary

In this chapter, we used STRIDE to threat model Kubernetes. We stepped through the six categories of threat and looked at some ways to prevent and mitigate them.

We saw that one threat can often lead to another, and that there are multiple ways to mitigate a single threat. As always, defence in depth is a key tactic.

We finished the chapter by discussing how Pod Security Policies provide a flexible and scalable way to implement Pod security defaults.

In the next chapter, we'll see some best practices and lessons learned from running Kubernetes in production

10: Real-world Kubernetes security

In the previous chapter, we threat modeled Kubernetes using STRIDE. In this chapter, we'll cover some common security-related challenges that you're likely to encounter when implementing Kubernetes in the real world.

While we accept that every Kubernetes deployment is different, there are many similarities. As a result, the examples that we cover will affect most Kubernetes deployments, large and small.

Now then, we won't be offering *cookbook style* solutions. Instead, we'll be looking at things from a high-level view, similar to what a *security architect* has.

We'll divide the chapter into the following four sections:

- CI/CD pipeline
- Infrastructure and networking
- Identity and access management
- Security monitoring and auditing

CI/CD pipeline

Containers are a revolutionary application *packaging* and *runtime* technology.

On the packaging front, we conveniently bundle application code and dependencies into an *image*. As well as code and dependencies, the image contains the commands required to run the application. This has allowed containers to hugely simplify the process of building, shipping, and running applications. It has also overcome the infamous “*it worked on my laptop*” issue.

However, containers also make running dangerous code easier than ever before.

With this in mind, let's look at some ways we can secure the flow of application code from a developer's laptop to production servers.

Image Repositories

We store images in registries, and registries are either public or private.

Note: Each registry is divided into one or more repositories, and we actually store images in repositories.

Public registries are on the internet and are the easiest way to download images and run containers. However, it's important to understand that they host a mixture of *official images* and *community images*. Official images are usually provided by product vendors and have undergone a vetting process to ensure certain levels of quality. Typically, official images will; implement best practices, be scanned for known vulnerabilities, contain up-to-date code, and be supported by the product vendor. *Community images* are none of that. Yes, there are some excellent community images, but you should practice extreme caution when using them.

With all of this in mind, it's important that you implement a standard way for developers to obtain and consume images in your environments. It's also vital that any such process be as frictionless as possible for developers – if there's too much friction, your developers will look for ways to bypass them.

Let's discuss a few things that might help.

Use approved base images

Images are made up of multiple layers that build on top of each other to form a useful image. But all images start with a base layer.

Figure 10.1 shows a simple example of an image comprising three layers. The base layer contains the core OS and filesystem components that applications need in order to run. The middle layer contains the application library dependencies. The top layer contains the code that your developers have written. We call the combination of these layers an *image*, and it contains everything needed to run the application.

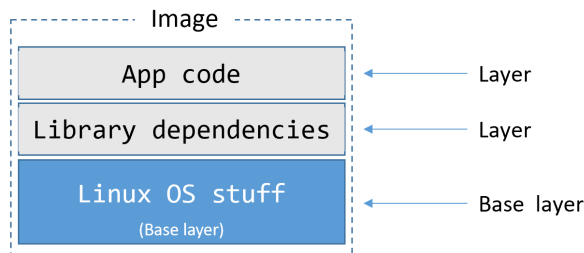


Figure 10.1

As all images have a base layer containing the required operating system (OS) and filesystem constructs for applications to build on, it's a common practice for organizations to have a small number of *approved base images*. It's also common, but not essential, for these base images to be derived from *official images*. For example, if you develop your applications on CentOS Linux, your base images *may* be based on the official CentOS image – you take the official CentOS base image and tweak it for your requirements.

In this model, all of your applications will build on top of a common approved base image like shown in Figure 10.2.

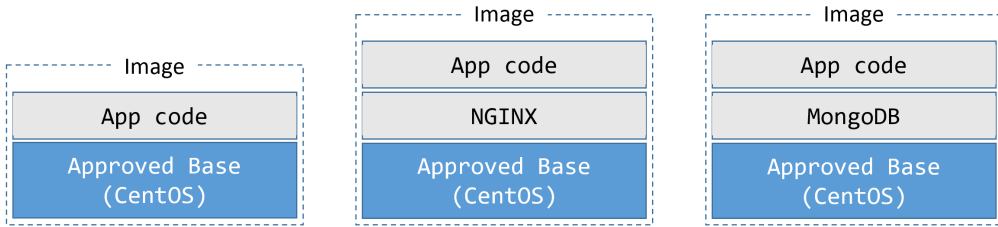


Figure 10.2

While there is some up-front effort required to create and implement base images, the long-term security benefits are worth it.

From a developer perspective, they can focus their entire efforts on the application and its dependencies without having to worry about maintaining OS components – don't worry about patching, drivers, audit settings, and more.

From an operations perspective, base images reduce software sprawl. This makes testing easier, as you will always be testing on a known base image. It makes pushing updates easier, you only need to update a small number of approved base images and have these easily rolled out to all developers. It also makes troubleshooting easier, as you have a small number of well-known base images providing your building blocks. It may also reduce the number of base image configurations that need tying into support contracts.

Non-standard base images

As good as it is to have a small number of approved base images, there may still be occasions when an application needs something different. This means you will need processes in place that:

- Identify why an existing approved base image cannot be used
- Determine whether an existing approved base image can be updated to meet requirements (including if it is worth the effort)
- Determine the support implications of bringing an entirely new image into the environment

Generally speaking, updating an existing base image – such as adding a device driver for GPU computing – should be preferred over introducing an entirely new image.

Control access to images

Various options exist that allow you to protect your organization's container images. The most secure practical option is to host your own private registry within your own firewall. This allows your organization to manage how the registry is deployed, how it is replicated, and how it is patched.

It also integrates permissions with existing identity management providers, such as Active Directory, and allows you to create repositories that fit your organizational structure.

If you do not have the means for a dedicated private registry, you can host your images in private repositories on public registries such as Docker Hub. However, this is not as secure as hosting your own private registry within your own firewalled network.

Whichever solution you choose, you should only host images that are approved to be used within your organization. Normally, these will be from a *trusted* source and vetted by your information security team. You should place access controls on the repositories that store these images, so that only approved users can push and pull them.

Away from the registry itself, you should also:

- Restrict which cluster nodes have internet access, keeping in mind that your image registry may be on the internet
- Configure access controls that only allow authorized users/nodes can push to repositories

Expanding on the list above...

If you are using a public registry, you will probably need to grant your worker nodes access to the internet so they can pull images. In this situation, a best practice is to limit internet access to the addresses and ports of any registries you use. You should also implement strong RBAC rules so that you can maintain control over who is pushing and pulling images from which repositories. For example, developers should probably be able to push and pull from non-production repositories, but not production. Whereas operations teams should probably be able to pull from non-production, as well as push and pull to production repos.

Finally, you may only want a sub-set of nodes (*build nodes*) to be able to push images. You may even wish to lock things down so that only your automated build systems can push to certain repositories

Moving images from non-production to production

Many organizations have separate environments for development, testing, and production.

Generally speaking, development environments have less rules and are commonly used as places where developers can experiment. Such experimenting often involves using non-standard images that your developers eventually want to use in production.

The following sub-sections will outline some measures you can take to ensure only safe images get approved into production.

Vulnerability scanning

Top of the list for vetting images before allowing them into production should be *vulnerability scanning*. This is a process where your images are scanned at a binary level and their contents checked against a database of known security vulnerabilities.

If your organization has an automated CI/CD build pipeline, you should definitely integrate vulnerability scanning. As part of this, you should consider defining policies that automatically fail builds and quarantine images containing certain categories of vulnerabilities. For example, you might implement a build phase that scans images and automatically fails anything using images with known *critical* vulnerabilities.

Two things to keep in mind if you do this...

Firstly, scanning engines are only as good as the vulnerability databases they use.

Secondly, scanning engines might not implement intelligence. For example, a method in Python that performs TLS verification might be vulnerable to Denial of Service attacks when the Common Name contains a lot of wildcards. However, if you never use Python in this way, the vulnerability *might* not be relevant and you *might* want to consider it a false positive. With this in mind, you may want to implement a solution that provides the ability to mark certain vulnerabilities as *not applicable*.

Configuration as code

Scanning application code for vulnerabilities is a widely adopted production best practice. However, reviewing application configurations, such as Dockerfiles and Kubernetes YAML files, is less widely adopted.

The *build once, run anywhere* mantra of containers means that a single container or Pod configuration can have hundreds or thousands of running instances. If a single one of these configurations pulls in vulnerable code, you can easily end up running hundreds or thousands of instances of vulnerable code. With this in mind, if you are not already reviewing your Dockerfiles and Kubernetes YAML files for security issues, you should start now!

A well-publicised example of not reviewing configurations was when an IBM data science experiment embedded private TLS keys in its container images. This made it possible for an attacker to pull the image and gain root access to the nodes that were hosting the containers. This would not have happened if a security review had been performed against the application's Dockerfiles.

There continue to be advancements in automating these types of checks with tools that implement *policy as code* rules.

Sign container images

Trust is a big deal in today's world, and cryptographically signing content at every stage in the software delivery pipeline is becoming a *must have*. Fortunately, Kubernetes, and many container runtimes, support the ability to cryptographically sign and verify images.

In this model, developers cryptographically sign their images, and consumers cryptographically verify them when they pull and run them. This process gives the consumer confidence that the image they are working with is the image they asked for and has not been tampered with.

Figure 10.3 shows the high-level image signing and verification process.

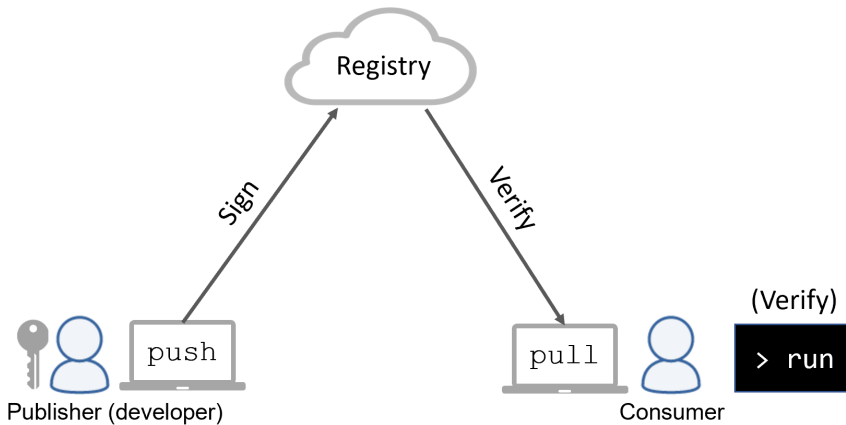


Figure 10.3

Image signing, and verification of signatures is usually implemented by the container runtime and Kubernetes does not get actively involved.

As well as signing images like this, higher-level tools, such as Docker Universal Control Plane, allow you to implement enterprise-wide policies that require certain teams to sign images before allowing them to be used.

Image promotion workflow

With everything that we've covered so far, a CI/CD pipeline for promoting an image to production should include as many of the following security-related steps as possible:

1. Configure environment to only `pull` and `run` signed images
2. Configure network rules to restrict which nodes can `push` and `pull` images
3. Configure repositories with RBAC rules
4. Developers build images using approved base images
5. Developers sign images and push to approved repos
6. Images are scanned for known vulnerabilities
 - Policies dictate whether images are promoted or quarantined based on scan results
7. Security team:
 - Reviews source code and scan results
 - Updates vulnerability rating as appropriate
 - Reviews container and Pod configuration files
8. Security team signs the image
9. All image pull and container run operations verify image signatures

These steps are examples and not intended to represent an exact workflow.

Let's switch our focus away from images and CI/CD pipelines.

Infrastructure and networking

In this section, we'll look at some of the ways we can isolate workloads.

We'll start at the cluster level, switch to the runtime level, and then look outside of the cluster at supporting infrastructure such as network firewalls.

Cluster-level workload isolation

Cutting straight to the chase, **Kubernetes does not support secure multi-tenant clusters. The only cluster-level security boundary in Kubernetes is the cluster itself.**

Let's look a bit closer...

The only way to divide a Kubernetes cluster is by creating *namespaces*. A Kubernetes namespace is not the same as a Linux kernel namespace, it is a logical partition of a single Kubernetes cluster. In fact, it's little more than a way of grouping resources and applying things like:

- Limits
- Quotas
- RBAC rules
- More...

The take-home point is that Kubernetes namespaces cannot guarantee a Pod in one namespace will not impact a Pod in another namespace. As a result, you should not run potentially hostile production workloads on the same physical cluster. The only way to run potentially hostile workloads, and guarantee true isolation, is to run them on separate clusters.

Despite this, Kubernetes namespaces are useful, and you *should* use them – just don't use them as security boundaries.

Let's look at how namespaces relate to *soft multi-tenancy* and *hard multi-tenancy*.

Namespaces and soft multi-tenancy

For our purposes, we'll define *soft multi-tenancy* as hosting multiple trusted workloads on shared infrastructure. By *trusted*, we mean workloads that do not require absolute guarantees that one Pod/container cannot impact another.

An example of two trusted workloads might be an e-commerce application with a web front-end service and a back-end recommendation service. Both services are part of the same e-commerce application, so are not hostile, but they might benefit from:

- Isolating the teams responsible for the different services
- Having different resource limits and quotas for each service

In this situation, a single cluster with one namespace for the front-end service and another for the back-end service might be a good solution. However, exploiting a vulnerability in one service might give the attacker access to Pods in the other service.

Namespaces and hard multi-tenancy

Let's define *hard multi-tenancy* as hosting untrusted and potentially hostile workloads on shared infrastructure. Only... as we said before, this isn't *currently* possible with Kubernetes.

This means that truly hostile workloads – workloads that require a strong security boundary – need to run on separate Kubernetes clusters! Examples include:

- Isolating production and non-production workloads on dedicated clusters
- Isolating different customers on dedicated clusters
- Isolating sensitive projects and business functions on separate clusters

Other examples exist, but you get the picture. If you have workloads that require strong separation, put them on their own clusters.

Note: The Kubernetes project has a dedicated *Multitenancy Working Group* that is actively working on the multitenancy models that Kubernetes supports. This means that future releases of Kubernetes might support hard multitenancy.

Node isolation

There are times when individual applications require non-standard privileges such as running as root or executing non-standard syscalls. Isolating these on their own clusters might be overkill, but the increased risk of collateral damage would probably justify running them on a ring-fenced subset of worker nodes. In this case, if one Pod is compromised it can only impact other Pods on the same node.

You should also apply *defence in depth* principles by enabling stricter audit logging and tighter runtime defence options on nodes running workloads with non-standard privileges.

Kubernetes offers several technologies, such as labels, affinity and anti-affinity rules, and taints, to help target workloads to sub-sets of nodes.

Runtime isolation

So far, we've looked at cluster-level isolation and node-level isolation. Now let's turn our attention to the various types of runtime isolation.

Containers versus virtual machines can be a polarizing topic. However, when it comes to workload isolation there is only one winner... the virtual machine!

The typical container model has multiple containers sharing a single kernel, and isolation is provided by kernel constructs that were never designed as *strong* security boundaries. We often call these *namespaced containers*.

In the hypervisor model, every virtual machine gets its own dedicated kernel and is strongly isolated from other virtual machines using hardware enforcement.

From a workload isolation perspective, virtual machines win.

However, it is becoming easier and more common to augment containers with additional kernel-level isolation technologies such as apparmor and SELinux, seccomp, capabilities, and user namespaces. Unfortunately, these can add significantly to the complexity of the setup and are still considered less secure than a virtual machine.

Another thing to consider is different classes of container runtime. Two prominent examples are **gVisor** and **Kata Containers**, both of which are re-writing the rules and providing stronger levels of workload isolation. Integrating runtimes like these with Kubernetes is made simple thanks to Kubernetes supporting the Container Runtime Interface (CRI) and Runtime Classes.

There are also projects that enable Kubernetes to orchestrate other workloads such as virtual machines and functions.

While all of this might feel overwhelming, everything discussed here needs to be considered when deciding what levels of isolation your workloads require.

To summarize, the following workload isolation options exist:

1. **Virtual Machines:** Every workload gets its own virtual machine and kernel. This provides excellent isolation but is relatively slow and heavy-weight.
2. **Traditional namespaced containers:** Every workload gets its own container but shares a common kernel. Not the best isolation, but fast and light-weight.
3. **Run every container in its own virtual machine:** This option attempts to combine the versatility of containers with the security of VMs by running every container in its own dedicated VM. Despite using specialized lightweight VMs this loses some of the appeal of containers and is not a popular solution.
4. **Use appropriate runtime classes:** This is extremely new but has a lot of potential. All workloads can run as containers, but workloads requiring stronger isolation are targeted to a class of container runtime that provides appropriate isolation (gVisor, Kata Containers etc.). Runtime classes is currently an alpha feature in Kubernetes.

A couple of other security-related things to consider...

Running lots of virtual machines can complicate things when it's time to patch operating systems. Also, running a mix of containers and virtual machines can increase network complexity.

Network isolation

On the topic of networking, firewalls are an integral part of any layered information security system. At a high level, they implement a set of rules that either *allow* or *deny* system-to-system communication.

As the names suggest, *allow rules* permit traffic to flow, whereas *deny rules* stop traffic flowing. The overall intent is to lock things down so that only authorized communications occur.

In Kubernetes, Pods communicate with each other over a special internal network called the *Pod network*. However, Kubernetes does not implement this *Pod network*, instead, it implements a plugin model called the Container Network Interface (CNI). Vendors and the community are responsible for writing the CNI plugins that actually provide the *Pod network*. Fortunately, there are lots of plugins available, and the networking options they support fall into the following two categories:

- Overlay
- BGP

Each of these is different, and each has a different impact on firewall implementation. Let's take a quick look at each.

Kubernetes and overlay networking

The most common way to build the *Pod network* is as an overlay network. In the Kubernetes world, overlay networking allows us to build a simple flat Pod network that hides any complexity that might exist between the nodes in the cluster. For example, you might have your cluster deployed across two different networks but have all Pods on a single flat Pod network. In this scenario, the Pods only know about the flat overlay Pod network and have no knowledge of the networks that the nodes are on. Figure 10.4 shows four nodes on two different networks, with Pods connected to a single overlay Pod network.

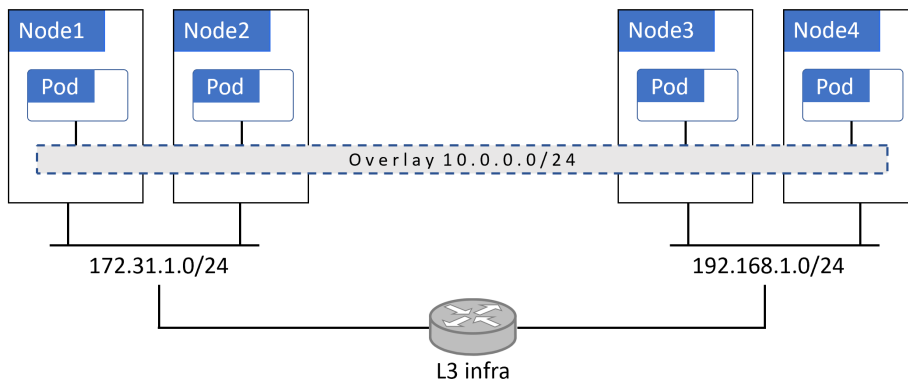


Figure 10.4

Generally speaking, overlay networks encapsulate packets for transmission over VXLAN tunnels. In this model, the overlay network is a virtual Layer 2 network operating on top of existing Layer 3 infrastructure. Traffic is encapsulated in order to pass between Pods on different nodes. This simplifies implementation, but encapsulation poses challenges for some firewalls. See Figure 10.5

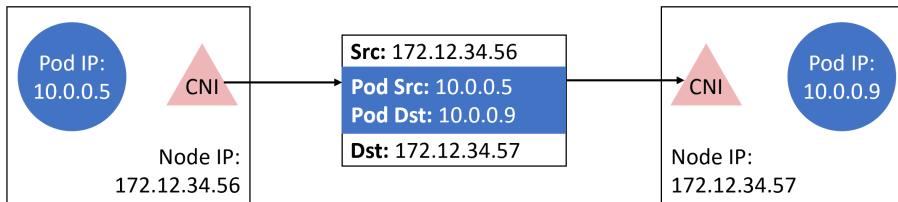


Figure 10.5

Kubernetes and BGP

BGP is the protocol that powers the internet. However, at its core it's a simple and scalable protocol that creates peer relationships that are used to share routes and perform routing.

The following analogy might help if you're new to BGP. Imagine you want to send a birthday card to a friend who you lost contact with and no longer have their address. However, your child has a friend at school whose parents are still in touch with your old friend. In this situation, you give the card to your child and ask them to give it to their friend at school. This friend gives it to their parents who deliver it to your friend.

This is similar to BGP. BGP Routing happens through a network of *peers* that help each other find a route for packets to go from one Pod to another.

BGP does not encapsulate packets, making life easier for firewalls. See Figure 10.6.

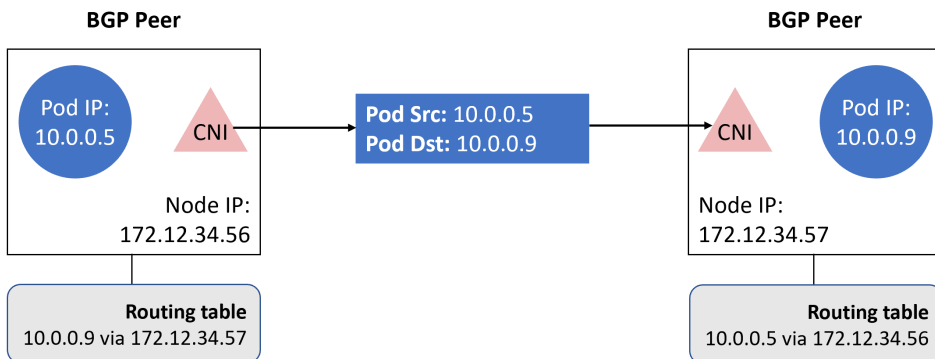


Figure 10.6

How this impacts firewalls

We've already defined a firewall as a network entity that allows or disallows traffic-flow based on source and destination addresses. For example:

- Allow traffic from the 10.0.0.0/24 network
- Disallow traffic from the 192.168.0.0/24 network

If your Pod network is an overlay network, source and destination Pod IP addresses are encapsulated so they can traverse the underlay network. This means firewalls that do not crack open packets and inspect their contents will not be able to filter based on Pod source and Pod destination IPs. You should consider this when choosing your Pod network and your firewall solutions.

With this in mind, if your Pod-to-Pod traffic has to traverse existing firewalls that do not perform deep packet inspection, it might be a better idea to choose a BGP-based Pod network. This is because BGP does not obscure Pod source and destination addresses.

You should also consider whether to deploy *physical firewalls*, *host-based firewalls*, or a combination of both.

Physical firewalls are dedicated network hardware devices that are usually managed by central team. Host-based firewalls are operating systems (OS) features and are usually managed by the team that manages your OS. For example, the Linux sysadmins. Both solutions have their pros and cons, and a combination of the two is probably the most secure. However, you should consider things such as; whether your organization has a long and protracted procedure for implementing changes to physical firewalls. If it does, it might not suit the nature of your Kubernetes deployment and a different firewall solution might be preferable.

Packet capture

On the topic of networking and IP addresses, not only are Pod/container IP addresses sometimes obscured by encapsulation, they are also dynamic.

Pods and containers are designed to be ephemeral, meaning they are not long lived. Scaling part of an application up adds more Pods and more IP addresses, whereas scaling it down removes Pods and IP addresses. IP addresses can even be recycled and re-used by different Pods and containers. This causes a lot of *IP churn* and reduces how useful IP addresses are in identifying systems and workloads. With this in mind, the ability to associate IP addresses with Kubernetes-specific identifiers such as; Pod IDs, Service aliases, and container IDs when performing things like packet capturing is extremely useful.

Let's switch tack and look at some ways of controlling user access to Kubernetes.

Identity and access management (IAM)

Controlling user access to Kubernetes is important in any production environment. Fortunately, Kubernetes has a robust RBAC subsystem that integrates with existing IAM providers such as Active Directory and other LDAP systems.

Most organizations already have a centralized IAM provider, such as Active Directory, that is integrated with company HR systems to simplify employee lifecycle management.

Fortunately, Kubernetes leverages existing IAM providers instead of implementing its own. For example, a new employee joining the company will automatically get an identity in Active Directory, which integrates with Kubernetes RBAC to automatically grant that user certain access to Kubernetes. Likewise, an employee leaving the company will automatically have his or her Active Directory identity removed or disabled, resulting in their access to Kubernetes being revoked.

RBAC went GA in Kubernetes 1.8 and it is highly recommended that you leverage its full capabilities.

Managing Remote SSH access to cluster nodes

Almost all Kubernetes administration is done via the API server, meaning it should be rare for a user to require remote SSH access to Kubernetes cluster nodes. In fact, remote SSH access to cluster nodes should only be for the following types of reason:

- Performing *node management* activities that cannot be performed via the Kubernetes API
- *Break the Glass* activities such as when the API server is down
- Deep troubleshooting

You should probably have tighter controls over who has remote access to the control plane nodes.

Multi-factor authentication (MFA)

With great power comes great responsibility...

Accounts with administrator access to the API server, and root access to cluster nodes, are extremely powerful and are prime targets for attackers and disgruntled employees. As such, their use should be protected by multi-factor authentication (MFA) where possible. This is where a user has to input a username and password followed by a second stage of authentication. For example:

- Stage 1: Tests *knowledge* of a username and password
- Stage 2: Tests *possession* of something like a one-time password device

Or...

- Stage 1: Tests *knowledge* of a username and password

- Stage 2: Tests something *about* the user, such as fingerprint or facial recognition

An easy, and important, place to implement multi-factor authentication is remote SSH access to cluster nodes. You should also consider it for access to workstations and user profiles that have `kubectrl` installed.

Auditing and security monitoring

No system is 100% secure, and you should plan for the eventuality that your systems will be breached. When breaches happen, it is vital that you can do at least two things:

1. Recognize that a breach has occurred
2. Build a detailed timeline of events that cannot be repudiated

Auditing is key to both of these requirements, and the ability to build a reliable timeline helps answer the following post-event questions; *what happened, how did it happen, when did it happen* and *who did it...* In extreme circumstances, information like this can even be called upon in court.

Good auditing and monitoring solutions also help to identify vulnerabilities in your security systems.

With these points in mind, you should ensure that reliable auditing and monitoring is high on your list of priorities, and you should not go live in production without them.

Secure Configuration

There are various tools and checks that can be useful in ensuring your Kubernetes environment is provisioned according to best practices and in-line with company policies.

The Center for Information Security (CIS) has published an industry standard benchmark for Kubernetes security, and Aqua Security (aquasec.com) has written an easy-to-use tool called `kube-bench` to implement the CIS tests. In its most basic form, you run `kube-bench` against each node in your cluster and get a report outlining which tests passed and which failed.

Many organizations consider it a best practice to run `kube-bench` on all production nodes as part of the node provisioning process. Then, depending on your risk appetite, you can pass or fail provisioning tasks based on the results.

`kube-bench` reports can also serve as a valuable baseline in the aftermath of an incident. In situations like this, you run an additional `kube-bench` after a breach and compare the results with the initial baseline to determine if and where the configuration has changed.

Container and Pod lifecycle Events

As previously mentioned, Pods and containers are ephemeral in nature, meaning they don't live for long – certainly not as long as VMs and physical servers. This means you will see a lot of events announcing new Pods and containers, as well as a lot of events announcing terminated Pods and containers. It also means you may need a solution that stores container logs in an external system and keeps them around for a while after their Pods and containers have terminated. If you don't, you *may* find it frustrating that you do not have logs for old terminated containers available for inspection.

Logs entries relating to container lifecycle events may also be available from your container runtime (engine) logs.

Application logs

In some situation there is not a lot Kubernetes can do to protect the applications it is running. For example, Kubernetes cannot prevent an application from running vulnerable code. This means it is important to capture and analyse application logs as a way to identify potential security-related issues.

Fortunately, most containerized applications will log messages to standard out (stdout) and standard error (stderr), which are then directed to the container's logs. However, some applications send log messages to other locations such as proprietary log files, so be sure to check your application's documentation.

Actions performed by users

Most of your Kubernetes configuration will be done via the API server where all requests should be logged. However, it is also possible to gain remote SSH access to control plane nodes and directly manipulate Kubernetes objects. This may include local unauthenticated access to the API, as well as directly modifying control plane systems such as etcd.

We've already spoken about limiting who has remote SSH access to nodes and bolstering security via things like multi-factor authentication. However, logging all activities performed via SSH sessions and shipping those logs to a secure log aggregator is highly recommended. As is the practice of always having a second pair of eyes involved in remote access sessions.

Managing log data

A key advantage of containers is application density – we can run a lot more applications on our servers and in our data centers. While this is great, it has the side-effect of generating massive amounts of logging and audit-related data that can easily become too much to analyse using traditional tools. At the time of writing, there is a lot of work being done to resolve this, including areas such as machine learning, but there is currently no easy solution.

On the negative side, such vast amounts of log-related data makes proactive analysis difficult – too much data to analyse. However, on the positive side, we have a lot of valuable data that can be used by security first-responders as well as for post-event reactive analysis.

Real world example

A great example of a container-related vulnerability, that can be prevented by implementing some of the best practices we've discussed, occurred in February 2019. CVE-2019-5736 allows a container process running as `root` to escape its container and gain root access on the host **and** all containers running on that host.

As dangerous as the vulnerability is, the following things that we covered in this chapter would've prevented the issue:

- Vulnerability scanning
- Not running processes as root
- Enabling SELinux

As the vulnerability has a CVE number, security scanning tools would've found it and alerted on it. Also, organizations that did not allow container processes to run as root will have been protected, as the issue only affects processes running as root. Finally, common SELinux policies, such as those that ship with RHEL and CentOS, prevented the issue.

All in all, a great real-world example of the benefits of defence-in-depth and other security-related best practices.

Summary

The purpose of this chapter was to give you an idea of some of the real-world security considerations effecting many Kubernetes clusters.

We started out by looking at ways to secure the software delivery pipeline by discussing some image-related best practices. These included; how to secure your image registries, scanning images for vulnerabilities, and cryptographically signing images. Then we looked at some of the workload isolation options that exist at different layers of the infrastructure stack. In particular, we looked at cluster-level isolation, node-level isolation, and some of the different runtime isolation options. We talked about identity and access management, including places where additional security measures might be useful. We then talked about auditing, and finished up with a real-world issue that could be easily avoided by implementing some of the best practices already covered.

Hopefully you now have enough understanding to go away and start securing your own Kubernetes clusters.

11: What next

There are lots of ways of taking your Kubernetes journey to the next level, and fortunately most of them are easy. We'll list a few here.

Practice makes perfect

I know I'm being *Captain Obvious* with this one, but there's no substitute for hands-on. Fortunately, it's never been easier to spin-up a Kubernetes playground where you can play around until you're a world authority!

I'm a huge fan of Play with Kubernetes (<https://labs.play-with-k8s.com/>). I love Docker Desktop, and I spin up test clusters on Google's GKE all the time.

Other options exist, and all of them are a lot simpler than how things used to be. I remember studying for my MSCE in Windows NT, and spending countless hours rebuilding NT domains from CD installs every time I trashed my lab. Things are so much simpler these days, and there really is no excuse for not getting our hands dirty.

More books

I've got a book on Docker that's had stellar reviews. Docker and containers are integral to Kubernetes, so if you need to know Docker, go check it out - it's called Docker Deep Dive, and you can get it on Amazon and Leanpub.

Video training

If you liked this book, you'll **love** my video courses!

I've got a **Getting Started with Kubernetes** course on Pluralsight, and a **Kubernetes Deep Dive** on A Cloud Guru. If you've read all the book, you probably want to go straight to the deep dive course.

<https://acloud.guru/learn/kubernetes-deep-dive>

I've also got a ton of Docker courses on Pluralsight.

If you're not a member of Pluralsight or A Cloud Guru, I recommend becoming one! Yes, they cost money, but they could be the best investments you ever make into your career! A monthly subscription on each platform gets you access to **every course in that platform's library** - everything from developer to IT ops. And if you're unsure about spending your money, there's usually a free trial where you can get free access for a limited time.

Events and meetups

You should hit events like KubeCon and your local Kubernetes meetups. They're full of great people and are great places to learn!

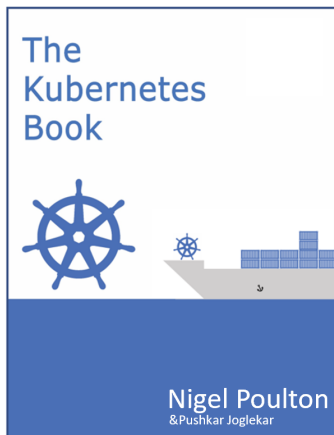
That's it for now. Keep learning!

Feedback

Thanks for reading my book. I hope you loved it.

Time for me to ask a favour (that's how we spell it in the UK)...

It takes a **ridiculous** amount of time and dedication to write a book. So, I'd really appreciate a quick review on Amazon. It'll take a couple of minutes, and you can even post an Amazon review if you bought the book from somewhere else. All types of reviews welcome!



★★★★★ ▾ 117 customer reviews

The Kubernetes Book

by Nigel Poulton ▾ (Author)

Also.... feel free to hit me on [Twitter](https://twitter.com/nigelpoulton)².



That's it. Live long and prosper...

²<https://twitter.com/nigelpoulton>