Big Book of MLOps

Jamie Ralph

2024-09-24

Table of contents

Preface

MLOps is a big field. I work in MLOps and spend a lot of time learning things. Unfortunately my brain can only store so much information, and it's easy to forget stuff.

This book is a repository of the knowledge I've learned since starting in MLOps. The content can roughly be divided into these categories:

- Short summaries of topics with links to articles that explain things nicely
- Notes I've taken from video courses (I find these the most time consuming and attention hungry. Taking notes mean I only sit through them once)
- Useful hints from technical books I've read

Networking in Kubernetes

Networking is a fundamental component of a Kubernetes cluster. There are many networking requirements within a cluster: containers communicate with each other in a pod, pods communicate with each other (both on the same node and across different nodes), and traffic from outside the cluster must reach the correct pod. Pods are ephemeral (they are destroyed and recreated) and have changing IP addresses, so how do we handle this? The resources here give a good overview of how Kubernetes networking is implemented.

Resources on Kubernetes networking

- Services, Load Balancing, Networking from the official Kubernetes docs gives a nice overview of K8s networking API and requirements.

- The official docs on Services goes into more detail on how services enable reliable communication with groups of pods whose IP addresses change as they're destroyed and recreated.
- This article on different service types goes into detail about the different service types available in Kubernetes
- Tracing the path of network traffic in Kubernetes by Kristijan Mitevski goes into much more detail about the mechanisms underlying pod to pod communication. The author covers key concepts like network namespaces, interfaces, network switches, and container network interfaces (CNI).
- This video on Kubernetes ingress by TechWorld with Nana gives a really nice overview of how Kubernetes services are exposed securely to the outside world. The video includes details on routing, ingress controllers, enabling HTTPS, and how ingress looks different in cloud platforms vs bare metal.

Resources on proxies and reverse proxies

- What is a reverse proxy? by CloudFlare provides a good overview of why we need proxies and reverse proxies for network communication
- NGINX is a popular reverse proxy - this introduction to NGINX by Sanjeev Sharma provides a hands-on demo of setting up NGINX and testing its capabilities