Table of Contents

3
Building a Data Parallel Training and Serving Pipeline

4
Bottlenecks and Solutions

Section 2 – Model Parallelism

5

Splitting the Model

6

Pipeline Input and Layer Split

7

Implementing Model Parallel Training and Serving Workflows

8

Achieving Higher Throughput and Lower Latency

Section 3 – Advanced Parallelism Paradigms

9

A Hybrid of Data and Model Parallelism

10
Federated Learning and Edge Devices

11
Elastic Model Training and Serving

12
Advanced Techniques for Further Speed-Ups

Index

Other Books You May Enjoy