Improving Data Placement Decisions for Heterogeneous Clustered File Systems

Cyril Allen

A Thesis in the Field of Information Technology

for the Degree of Master of Liberal Arts in Extension Studies

Harvard University

March 2018

Abstract

With the advent of cloud computing, datacenters are using distributed applications more than ever. MapReduce is used to generate over 20 petabytes of data per day by using prodigious numbers of commodity servers (?, ?). Many companies use large scale clusters to perform various computational tasks via the open-source MapReduce implementation, Hadoop (?, ?), or they can possess a virtualized datacenter, allowing them to migrate virtual machines between various machines for high-availability reasons. As economics change for hardware, it is likely that a scalable cloud will have the requirement to mix node types, which will lead to higher performance and higher capacity nodes to be mixed with lower performance, lower capacity nodes. This thesis presents an adaptive data placement method in the Nutanix distributed file system which will remedy some common problems found in many heterogeneous clustered file systems.

Acknowledgements

I doubt that I would be writing this document were it not for my father, Dr. Cyril Allen. Not only did he spark my initial interest in computing when I was a child, but also he convinced me to pursue a graduate degree.

Thanks are also due to my amazing mother, Dr. Hengameh Allen-Schaal, for her unconditional love and support. She has always been there for me when I needed her. I could not have done any of this without her.

I would like to thank my thesis director, Professor Jamie Frankel, for his patience and guidance through multiple drafts of this document. Working with him has been a remarkable learning experience.

None of the work in this thesis would have been possible without my employer, Nutanix. They have helped me tremendously by providing hardware resources and the chance to work on thought-provoking problems.

Contents

List of Figures

List of Tables

1 Introduction

The work in this thesis was performed in the context of a distributed file system designed by Nutanix for enterprise clusters. The file system is hosted on some number of virtual machine hosts (commonly referred to as hypervisors) connected in a full mesh network. Figure **??** shows an example full mesh network topology. This full mesh configuration allows for the clustering of virtual machines (VMs) that provide file system services by allowing each VM to send messages to any other VM.

The basic purpose of a file system is to store and retrieve data. Storage and retrieval of data is performed via write and read operations on the file system. A write operation in the Nutanix file system requires multiple copies, or replicas, of data to be written to disk on multiple physical servers distributed across a network partition. Given this, a write operation is not complete until all replicas of the data are written to disk. A write operation's performance is at the mercy of the slowest disk in the set chosen to host replicas. Variables that affect a disk's performance could include the hosting server's average CPU utilization and the number of operations already in flight on a chosen disk. To mitigate the negative impact these variables could have on a write, the file system should consider how these variables will affect a disk's performance before choosing to target it for a write operation. This results in biasing toward disks that will give better performance overall.

1.1 Nutanix and the Acropolis Base System

1.1.1 Architecture

A Nutanix cluster is facilitated by a clustering of controller virtual machines (CVMs) that reside, one per node, on each server in the cluster, as shown in Figure **??**. CVMs work together to form a distributed system that provides an NFS (for VMware's ESXi