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 making use of distributed applications more than ever. MapReduce is used to generate over 20 petabytes of data per day using very large numbers of commodity servers [3]. Many companies use large scale clusters to perform various computational tasks via the the open-source MapReduce implementation, Hadoop [4], 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 attempt to remedy the common problems found in many heterogeneous clustered file systems.

Acknowledgements

First off, I would like to thank my father, Dr. Cyril Allen, for sparking my initial interest in computing as a child and convincing me to pursue a graduate degree.

I also want to thank my mother, Dr. Hengameh Allen-Schaal, for her unconditional love and support. She has always been there for me when I needed her, even when we began living in different states. I could not have done any of this without her.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

A number of scenarios arise in heterogeneous Nutanix clusters that can degrade performance for an entire cluster. The currently replica disk selection logic in Stargate uses does not take into account a number of variables such as disparities in tier size, CPU power, workloads, and disk health among other things.

Given that a write is not complete until all replicas are written, the write's performance is at the mercy of the slowest disk and node. There are several scenarios, both pathological and daily occurences, where a more robust replica placement heuristic is required. For the work in this thesis, I will focus on two orthogonal cases described in the next section.

1.1.1 Interfering Workloads

An example of interfering workloads can take the form of a 3-node homogeneous cluster with only 2 nodes hosting active workloads as shown in Figure 1. In the current random selection scheme in use by Nutanix clusters, writes are equally likely to place their replica on the other node with an active workload as they would be to place it on the idle node. This can impact performance on both the local and remote workloads as secondary writes will be slower on nodes whose resources are being utilized by their primary workloads. An adaptive replica placement scheme is needed to avoid the busy node and bias secondary replica placement on an idle node.

1.1.2 Nodes with Tier Size Disparities

A cluster containing nodes with a tier size disparity are susceptible to a skew in node fullness, even if the workload on each node is identical. This can be illustrated via

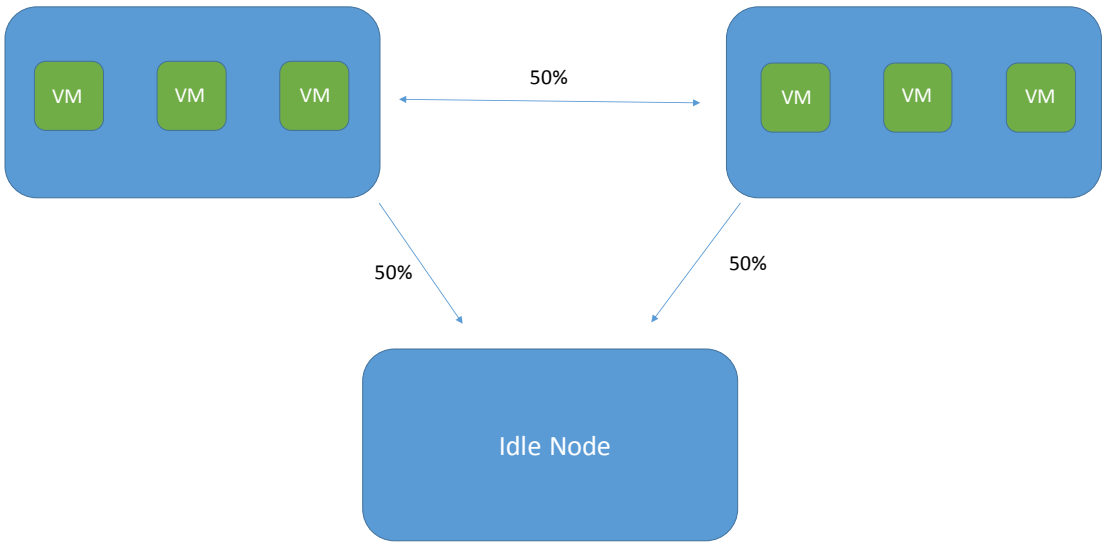

Figure 1: A cluster with identical nodes running a heterogeneous workload.

Figure 2 where we have a 3-node heterogeneous cluster with 2 high-end nodes and a single weak node. Suppose these high-end nodes have 500GB of SSD tier and 6TB of HDD tier and the single weak node has only 128GB of SSD tier and 1TB of HDD tier. If 3 simultaneous workloads were to generate data such that the working sets of the workloads are 50% of the local SSD tier, the weaker node is at a significant disadvantage. Given the replica selection algorithm in Nutanix cluster versions prior to 5.0, we can expect 500GB of replica traffic to flood the weak node and fill up its SSD tier well before the workload is finished. This results in an inability for the workload on the smaller node to place its primary replicas locally and forces the workload to rely on remote CVMs, increasing latency. An adaptive replica placement heuristic would mitigate this issue by taking disk usages into consideration during the placement of secondary replicas and biasing placement of secondary replicas on the nodes with more free capacity.

To create a context for the work in this thesis, it's necessary to delve deeper into

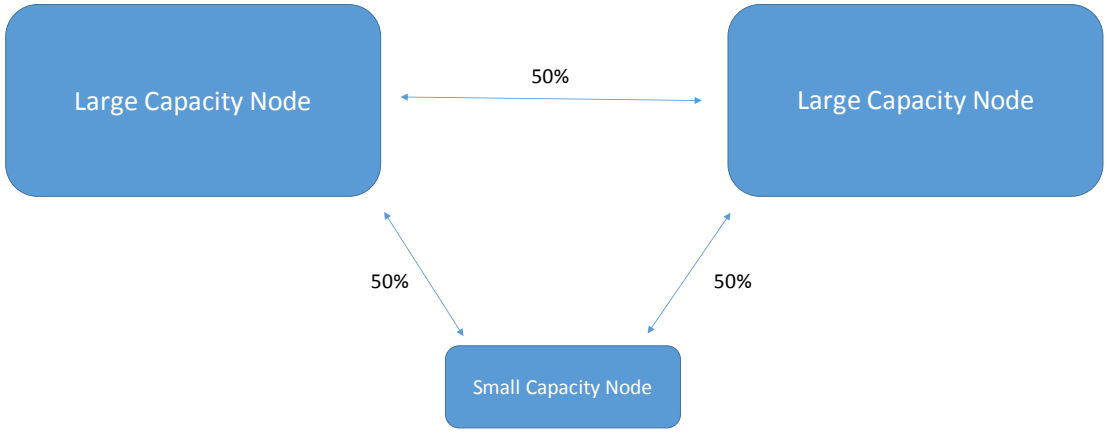

Figure 2: A cluster with nodes of varying resource capacity.

what a Nutanix cluster is and give a brief overview of the I/O path. Section 1.2 will be devoted to this overview.

1.2 Nutanix and the Acropolis Base System

The work in this thesis was performed in the context of a distributed file system designed by Nutanix for enterprise clusters. In this section I'll provide an overview of the key features provided by the storage system to provide the necessary background for understanding the changes performed to the data replica placement.

1.2.1 Architecture

A Nutanix cluster is facilitated by a clustering of controller virtual machines (CVMs) which reside, one per node, on each server in the cluster. CVMs work together to form a distributed system that provides an NFS (for VMWare's ESXi [14]), SMB (for Microsoft's Hyper-V [17]), or iSCSI (for Nutanix's AHV [1]) interface to each

hypervisor that they reside on. For example, the interface provided by the CVMs to VMware's ESXi hypervisor will be interfaced with as a datastore. The virtual machines' virtual disk files will reside on the Nutanix datastore and be accessed via NFS through the CVM sharing a host with the user VM.

Users of a Nutanix cluster will typically have virtual machines (VMs) hosted by one of the hypervisors previously mentioned. These VMs will have their storage requests forwarded to the local CVM hosted on the same node, or another CVM in the cluster in the event that the local CVM is down. This fault tolerance is one of the advantages to using a clustered file system such as Nutanix as opposed to a monolithic storage array.

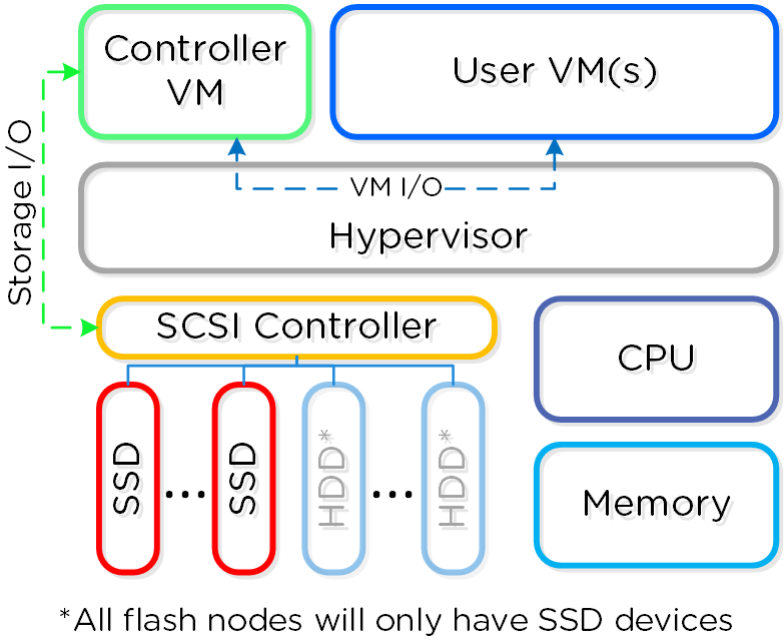

Figure 3: Nutanix node architecture diagram from the Nutanix Bible[1].

The CVMs expose some number of block devices, known as vDisks, that are used by the VMs hosted by the hypervisor. Within each CVM exists an ecosystem of processes responsible for the services provided by the Nutanix cluster. The work in this thesis is scoped specifically to the I/O manager in the CVM ecosystem, Stargate,

which is discussed in the following section.

1.2.2 Stargate

The Stargate process is responsible for all data management and I/O operations. The messages sent via NFS/SMB/iSCSI from the hypervisor to the CVM are consumed and acted upon by Stargate. All file allocations and data replica placement decisions are made by this process.

As the Stargate process facilitates reads and writes to physical disks, it gathers statistics for each disk such as the number of operations currently in flight on the disk (queue length), how much data in bytes currently resides on the disk, and average time to complete an operation on the disk. These statistics are only gathered on the local disks; however, they are then stored in a distributed database provided by another service on the CVM, known internally as Arithmos, along with the statistics gathered by every other Stargate in the cluster. These disk statistics stored in the database and are pulled periodically and are then used to make decisions on data placement when performing writes.

1.2.3 Cassandra

Cassandra stores and manages all cluster metadata in a distributed manner. The version of Cassandra running in NDFS is a heavily modified Apache Cassandra [2]. One of the main differences between Nutanix Cassandra and Apache Cassandra is that Nutanix has implemented the Paxos [18] protocol to enforce strict consistency.

1.2.4 The Extent Store

The Extent Store is a sub-component of Stargate that serves as the persistent bulk storage. Data stored within the Extent Store are referred to as extent groups (also

referred to as egroups), or 4MB pieces of physically contiguous data. These extent groups are replicated a number of times dependent on the cluster replication factor and each replica is placed on a different CVM except for a single replica that resides on the local node. The current replica selection algorithm will be explained in further detail in the "Replica Selection" section.

1.2.5 Storage Tiering and Data Replication

Data in the system is physically stored on disks of varying type. Within a given cluster, one can find NVMe, PCIe SSD, SATA SSD, and HDD drive types. These disks are separated out into groupings of similar drive types referred to as storage tiers. Extent groups and their replicas are created on a particular storage tier, but may be migrated between tiers depending on data access patterns.

Data replicas for each extent group are written on separate fault domains on the same tier to ensure that in the event a failure occurs and a replica becomes unavailable, that piece of data can still be accessed. An example of a fault domain can be a node (single CVM) or a block (a rackable unit of hardware containing multiple nodes).

1.2.6 Replica Selection

Stargate is inherently limited in its choices of disk to service reads since it must select a disk that contains the desired data. If the data is on a local disk, this disk is always selected due to the unnecessary overhead of data traversal over the network when selecting a remote disk. However, when selecting disks to service writes, Stargate must decide whether to perform an in-place overwrite for writes over existing ranges of data or whether to generate new data. In-place overwrites are similar to reads in that the choice of target disk is limited by where the relevant data resides. In-place overwrites are ruled out in the case of deduplicated data, since multiple ranges within

a vDisk point to the same physical data. In the case of deduplicated data and all other scenarios where data is being created for the first time, Stargate has the freedom to choose any disk in the cluster.

When a Stargate write operation wants to select disks for data placement upon entrance into the Extent Store, it is done on a per-tier basis. The interface for a replica selection function call, DetermineNewReplicas, takes the following arguments:

Table 1: Argument descriptions for DetermineNewReplicas.

Argument Name	**Description**
`replica_vec`	A std::vector that is populated with the selected disk IDs. The number of selections made is the size of this vector.
`storage_tier`	The name of the storage tier (set of disks) to select for data placement.
`storage_pool_vec`	The collection of all disks in the cluster.
`preferred_node_vec`	Nodes we prefer to choose disks from. This means we will consider disks from these nodes first and only consider other disks if the ones belonging to these nodes are not suitable.
`predicate_func`	A function that accepts a disk ID and returns a boolean. If this function evaluates to false for any disk, it is not considered for selection.
`exclude_replica_vec`	Disk IDs that will be excluded from selection.
`exclude_node_vec`	Node IDs whose disks will be excluded from selection.
`exclude_rackable_unit_vec`	Rack IDs whose disks will be excluded from selection.

Upon calling DetermineNewReplicas, we first attempt to select replicas from the preferred nodes. This involves finding the set of disks that belong to each node on the specified tier, shuffling the disks, and sequentially evaluating each disk until a suitable one is found (the evaluation step). If a suitable disk is not found on a preferred node, all other disks belonging to the specified tier are shuffled and considered sequentially until enough disks are found to satisfy the requirement set by replica_vec.

A suitable disk is one that:

1. Contains enough space to accept new data. This is less than 95% by default for Nutanix clusters.
2. Returns "true" when evaluated by the predicate_func.
3. Is not included in the exclude_replic_vec, exclude_node_vec, and exclude_rackable_unit_vec.

If a disk does not meet the criteria above, we simply continue searching the shuffled set of disks belonging to the specified tier. If a disk is found to be suitable, we must add the disk to the replic_vec and add the node that the disk belongs to in the exclude_nod_vec. This prevents us from considering other disks on that node to maintain a node fault tolerance guaranteed by the cluster's replication factor.

2 Prior Work

2.1 Clustered Storage Systems

Among the contemporary distributed storage systems, Apache's Hadoop Distributed File System (HDFS) [4] is one of the most well known. HDFS is an open source distributed file system written as a userspace library in Java that is used by Hadoop clusters for storing large volumes of data. An HDFS cluster is comprised of a single NameNode and multiple DataNodes which respectively manage cluster metadata and store the data. HDFS stores files as a sequence of same-size 128MB blocks that are replicated across DataNodes to provide fault tolerance in the event of node failures. Decisions regarding the replication and placement of blocks of data are made by the NameNode. Since large HDFS clusters will span multiple racks, replica placement within HDFS is made in a rack-aware manner to improve data reliability and network utilization.

While HDFS is an open source distributed file system that is part of the greater Apache Hadoop ecosystem, the Google File System (GFS) [25] is a proprietary distributed file system developed by Google that inspired HDFS. A GFS cluster is comprised of a single Master node and multiple Chunkservers that store 64MB "chunks," or blocks, of data. Each file in the GFS is divided into the 64MB chunks and each chunk is replicated by the Chunkservers multiple times throughout the GFS cluster. Master nodes store the metadata associated with the chunks such as chunk location or which processes are operating on a chunk.

2.2 Data Placement Schemes in Clustered Storage Systems

Xie et. al. [5] showed that data placement schemes based on the computing capacities of nodes in the HDFS significantly improved workload performance. These computing

capacities are determined for each node in the cluster by profiling a target application leveraging the HDFS. Their MapReduced wordcount and grep results showed up to a 33.1% reduction in response time. Similarly, Perez et. al. [8] applied adaptive data placement features to the Expand parallel file system based on available free space. Though effective in their given contexts, the main drawback to this work is that it assumes the specific application is working without interference and does not account for other workloads on the system.

One adaptive data placement approach that can account for other workloads on the system was introduced by Jin et. al. [7] in their work on ADAPT . The work predicts how failure-prone a node in a MapReduce cluster is and advises their availability-aware data placement algorithm to avoid those nodes. This proves useful for performance by avoiding faulty nodes that could fail mid-task and cause data transfers and re-calculation of data.

Work by Suresh et. al. [16] approaches adaptive replica selection in much the same way proposed in this paper, though their work was mainly focused on decreasing tail latencies for Cassandra [2] reads. Their load balancing algorithm, C3, incorporates the concept of a value calculated from request feedback from their servers that allows for decisions to be made on server selection. In addition to the ranking function in C3, they implemented a distributed rate control mechanism to prevent scenarios where many individual clients can bombard a single highly desirable server with requests. Many of the same problems that the work in this proposal seeks to remedy are also addressed by the C3 algorithm; however, given the Nutanix file system's architecture, some C3 solutions are not feasible.

The C3 algorithm takes into account the request queue length of certain servers similar to the way I will use disk queue lengths. In addition Suresh et. al. factor in the service latencies of each server so that they may consider a different ideal

queue length for each server. With this approach, longer service times will warrant a lower queue length and vice versa. This is beneficial for scenarios where there are multiple underlying storage technologies such as NVMe drives, SSDs, and HDDs under consideration, but the Nutanix file system's architecture does not allow for multiple replicas to span storage tiers. This forces the ideal queue lengths for each selection pool to be the same. Therefore, my work does not incorporate service latencies in the fitness value calculations.

Herding of requests to a single highly suitable server is a problem that arises in any replica ranking algorithm. C3 mitigates this issue by rate limiting client requests to each server via a decentralized calculation using a configured time interval and local sending rate information. I've opted to use a simpler probabilistic spreading of client requests via selecting remote nodes using a weighted random selection tied to the calculated fitness values.

2.3 Non-clustered Data Replication via RAID

So far, I have only discussed distributed file systems and data replica placement schemes in a clustered/distributed context. It is worth noting that not all storage replication occurs across a network partition. For example, it is common to see Redundant Arrays of Inexpensive Disks (RAID) [27] [28] used on servers to increase performance or to protect data written to the disks in the event of a drive failure.

Table 2: Common RAID levels and their descriptions.

RAID Level	Description
RAID 0	Disk striping with no replication.
RAID 1	Disk mirroring.
RAID 5	Block-level striping with distributed parity.
RAID 6	Block-level striping with doubly distributed parity.

RAID combines multiple physical disks into a single logical disk that is then presented to the operating system as a single device. The exact data replication scheme that dictates the way data is distributed across the drives is known as the RAID level. Table 2 shows some of the standard RAID levels.

3 Implementation

3.1 Stargate Disk Stats Collection

Prior to this work, Stargate's periodic disk stats collection was limited to caching solely disk usage stats for all disks in the cluster. This has been expanded to now include disk performance stats for use in disk fitness values.

Stargate maintains a mapping, henceforth referred to as the *disk_map_*, from cluster disk ID to a disk state structure. The *disk_map_*'s state structure contains disk usage and performance information that has been published to Arithmos by other Stargates in the cluster. Upon gathering fresh stats from Arithmos, the information is used to create a disk fitness value for each disk in the cluster.

3.2 Fitness Values and Functions

To calculate a value to represent the desirability of a disk for replica placement, we'll use a function, $f_{fitness}$ that takes as its argument disk stats and returns a positive number we will call a fitness value. A low fitness value indicates a poor placement candidacy for a disk and a high fitness value will indicate a highly desirable disk for replica placement. In this thesis, I evaluate two fitness functions that are comprised of terms that utilize a disk's average queue length over some stretch of time, t_q, and a disk's percentage utilization, t_u.

$$t_q = 1 - \frac{q}{q_{ceil}} \tag{1}$$

$q_{ceiling}$ is defined as the maximum observed queue length such that beyond this value, t_q does not contribute to the fitness value. This ensures that as the queue length grows, t_q approaches zero.

$$t_u = \frac{1}{a^u} \tag{2}$$

u is the disk utilization percentage and a is an aggression variable used to control the exponential decay of t_u. The larger a is, the more aggressively t_u will decay as u increases. An aggressive decay results in more preference given to less utilized disks when compared with disks that are slightly more utilized.

These terms are used in two different fitness functions evaluated in this thesis:

$$f_{add} = t_u + t_q \tag{3}$$

$$f_{mult} = t_u t_q \tag{4}$$

3.3 Weighted Random Selection Algorithms

After a weight is calculated for a disk in the cluster that will store a replica, the WeightedVector class' Sample() calls will perform a weighted random selection on the set of potential candidate disks. To determine the best method of weighted random selection for Stargate's WeightedVector class, an exploration of various weighted random selection algorithms was necessary. Since the file system only supports replication factors of 2 or 3, the investigation was limited to algorithms that allow for a weighted N choose Y, where Y is the data replication factor.

This section provides an overview of the algorithms investigated via simulations to compare each algorithm at different orders of magnitude.

3.3.1 Scalability Simulation Methodology

To test the scalability of the weighted random selection algorithms evaluated in the next section, a single-threaded Python [23] script was written to to evaluate the change in run time as sample sets increase. Each algorithm's time to select is calculated for each of a fixed number of iterations for multiple sample sets. pseudo-code for the simulations can be written as follows:

```
for each selection_algorithm in algorithm_list:
    run_time_values = empty_list()
    for each sample_set_size in all_sample_set_sizes:
        sample_set = generate_sample_set(sample_set_size)
        all_elapsed_times = empty_list()
        for each iteration:
            start_time = time.now()
            selection_algorithm(sample_set)
            elapsed_time = time.now() - start_time
            all_elapsed_times.append(elapsed_time)
        calculate_avg_elapsed_time(all_elapsed_times)
        calculate_std_error(all_elapsed_times)
```

Object weights are constant throughout the simulation, so selection schemes that require some amount of preprocessing (such as the top T% calculation for truncation selection) are performing their preprocessing steps for each selection. This gives information about the worst-case behavior for each algorithm in comparison with others'.

3.3.2 Herding Behavior Evaluation Methodology

Herding behaviors can be seen in some weighted random selection algorithms when the weights of a subset of objects in the sampling pool cause a disproportionate amount of selections to target those objects. In the case of replica disk selections in a Nutanix cluster, this can cause too many operations to target an especially suitable

disk, resulting in poor performance. We can simulate an exaggerated scenario in which the susceptibility to herding behavior can be observed by having a single object with a weight that is multiple orders of magnitude heavier than the next highest object in the sampling set. It is also necessary to observe any herding behavior for a sampling set with low weight skew. This section describes the simulation methodology for each.

High-skew sampling sets of 11 objects were generated such that the array index of the first 10 objects was assigned as the object weight, and the last object was given a weight of 1000. This creates an extremely large skew in weights and makes the high-weight object a target for herding behavior. Given this sampling set, weighted random selections were performed and a histogram was kept that tracked the number of selections for each object. 10^3, 10^4, and 10^5 iterations were performed to observe any changes in herding behavior at larger time scales. In addition to the high-skew sampling sets, low-skew sets of 100 objects were also simulated. These low-skew sets were identical to the high-skew sets, except there was not a single object with an exaggerated weight of 1000. All element weights were their array indices.

3.3.3 Stochastic Universal Sampling (SUS) Simulations

SUS is another sampling technique first introduced by Baker in 1987 [10]. The algorithm can be understood as follows: On a standard roulette wheel there's a single pointer that indicates the winner. The roulette wheel's "bins" can all be the same size which would indicate a uniform probability of selecting any bin and could also be unevenly sized which would indicate a weighted probability. SUS uses this same concept except allows for N evenly spaced pointers corresponding to the selection of N items. Key things to note are that the set, or "bins" in my roulette analogy, must be shuffled prior to selection. Also, there is a minimum spacing allowed for the pointers to prevent selection of the same bin.

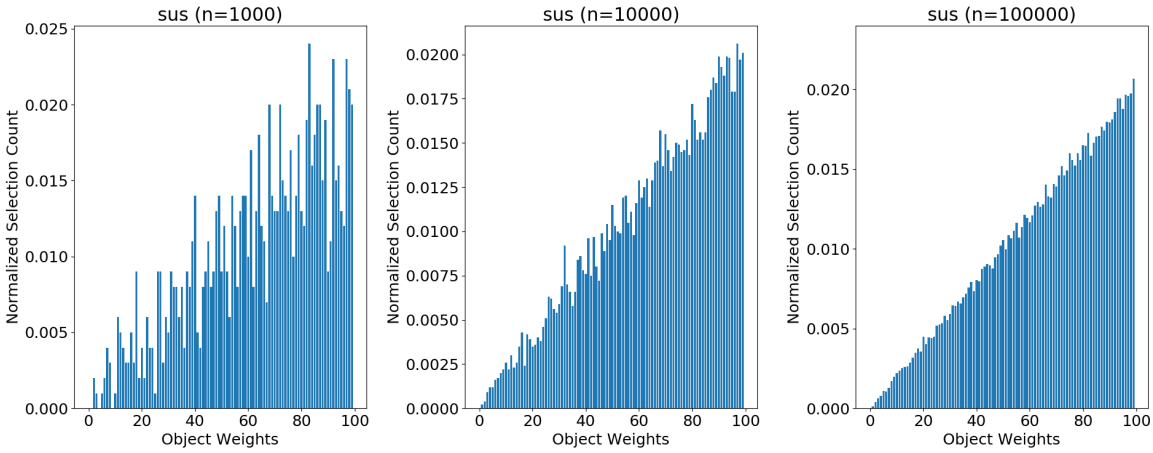

Figure 4: Histograms generated by simulation of Stochastic Universal Sampling to illustrate selection distributions.

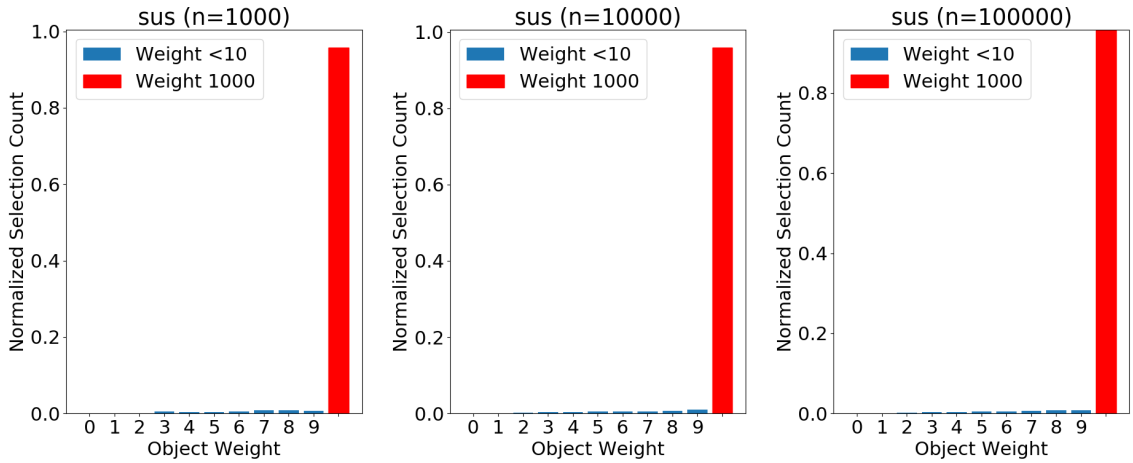

Figure 5: Histograms generated by simulation of Stochastic Universal Sampling to illustrate herding behavior.

Figure 4 shows the evolution of the distribution of selection frequencies for SUS as the number of samples increases. We can see that the selection frequency is proportional to the object weight. This can prove problematic for outlier objects with weights that are much larger than the other objects in the set as shown in Figure 5. We can see that the high weight object's selection frequency eclipses all other objects in the selection pool which can lead to extreme herding behaviors.

3.3.4 Tructation Selection Simulations

Truncation selection [20] does not consider any objects for selection below some threshold, T. In figure 6, only the top 10% of objects ranked by weight are considered for selection. Within this subset, weight has no meaning so there is a uniform distribution of selections. However, this may not be suitable for use cases where all objects must be candidates for selection. Depending on the threshold chosen, this algorithm can be resistant to herding behavior as shown in Figure 7.

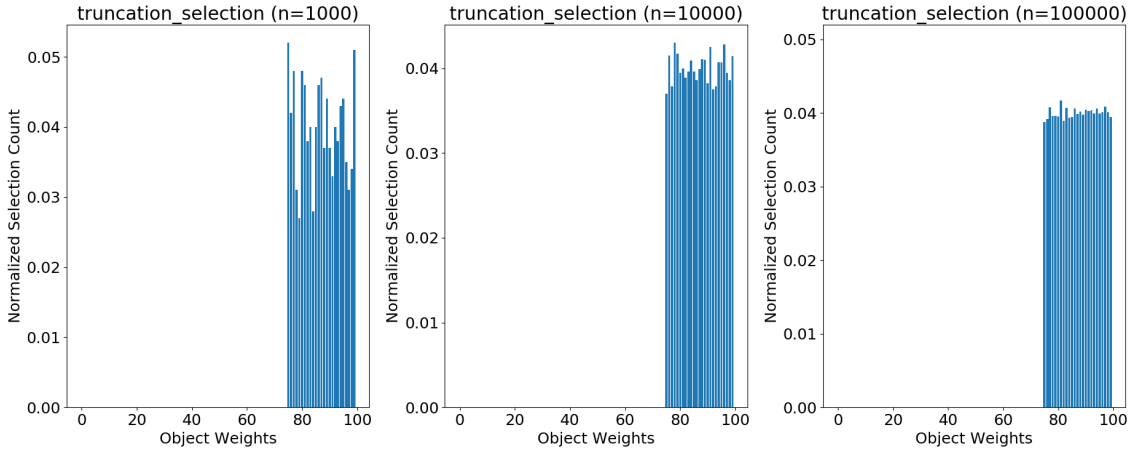

Figure 6: Histograms generated by simulation of truncation selection to illustrate selection distributions.

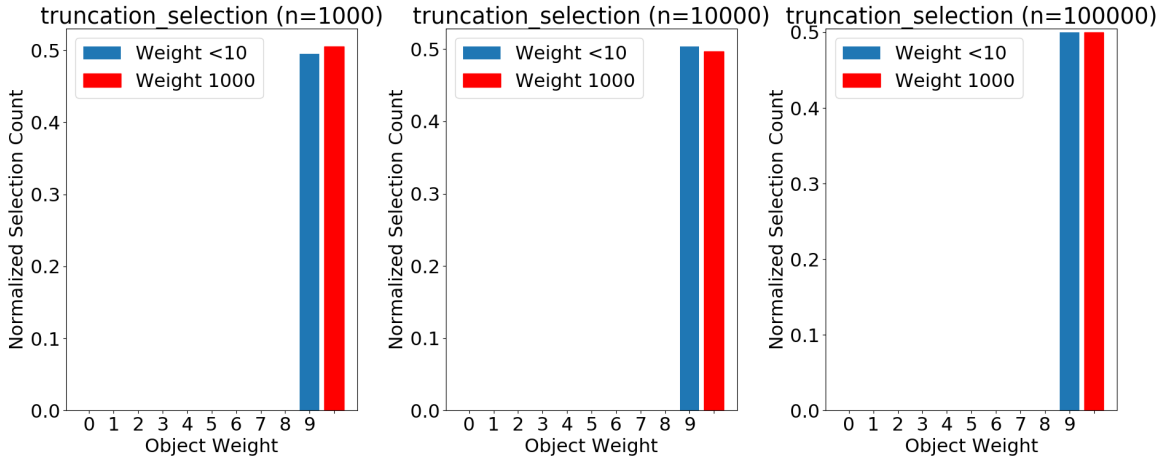

Figure 7: Histograms generated by simulation of truncation selection to illustrate herding behavior.

3.3.5 Two-choice Sampling Simulations

Two-choice sampling [21], has proven to be extremely resilient to herding behavior as shown in Figure 9 and selection frequencies for all objects are influenced by object weights in a way similar to SUS. While an object with a higher weight is more likely to be selected, it is not selected with a probability proportional to its fitness value. This makes the algorithm resistant to any herding behaviors, but will not be a good candidate if we desire the WeightedVector to select objects with probability proportional to its weight.

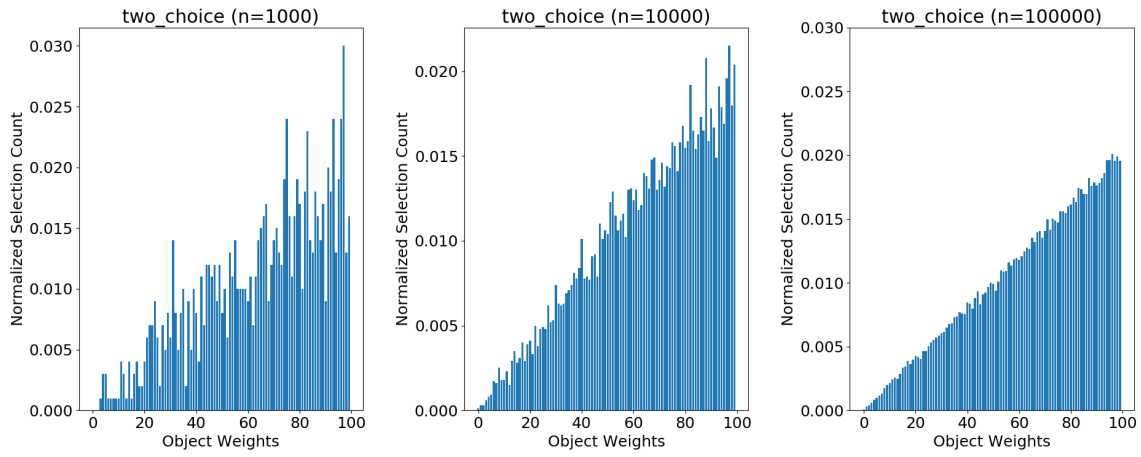

Figure 8: Histograms generated by simulation of two-choice selection to illustrate selection distributions.

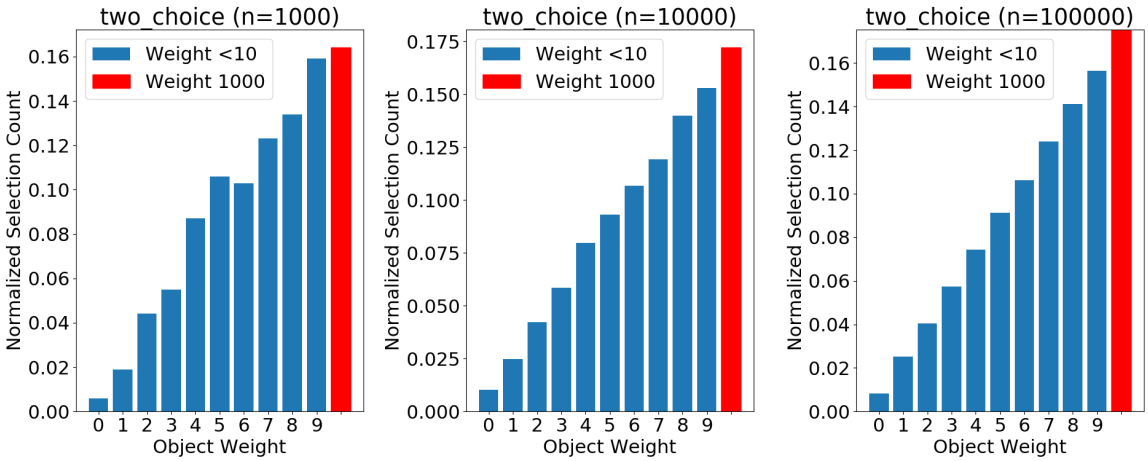

Figure 9: Histograms generated by simulation of two-choice selection to illustrate herding behavior.

3.3.6 Uniform Random Selection Simulations

Uniform random selection is completely indifferent to object weights. Therefore, it exhibits no herding behavior and no usefulness for the WeightedVector's sampling, but it is important to include its analysis for comparison with other selection algorithms.

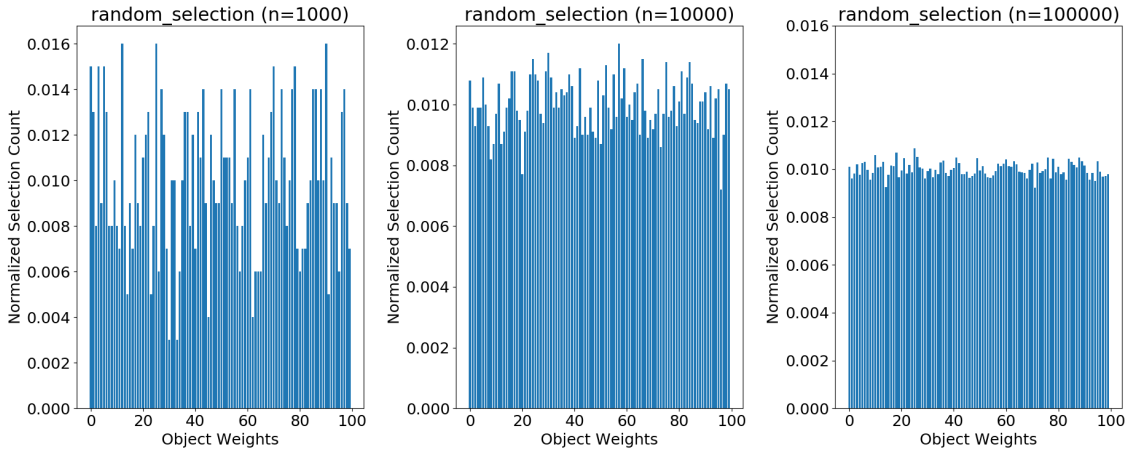

Figure 10: Histograms generated by simulation of uniform random selection to illustrate selection distributions.

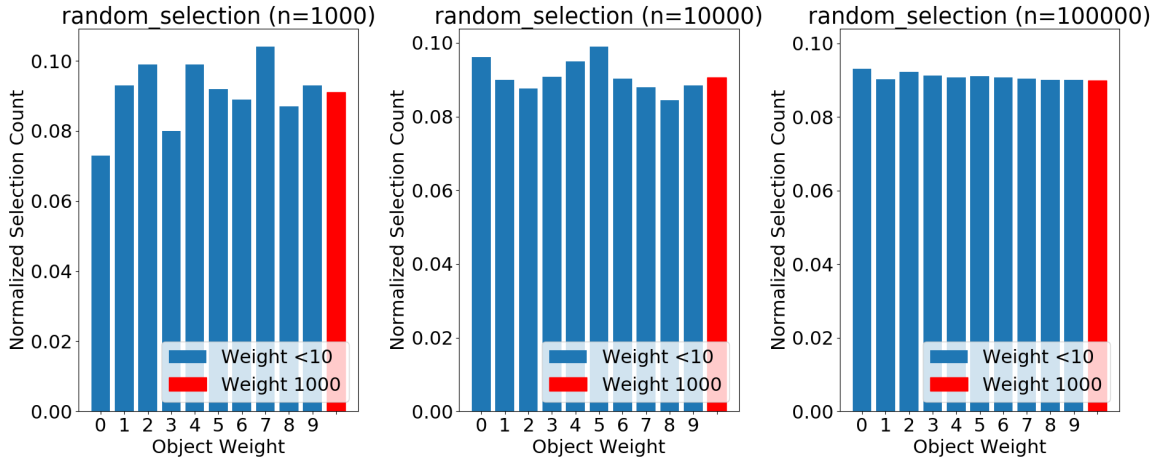

Figure 11: Histograms generated by simulation of uniform random selection to illustrate herding behavior.

Figure 7 shows that we can expect uniform selection probabilities and Figure 11

confirms that we can expect no herding behavior with a uniform random selection scheme.

3.3.7 Weighted Random Algorithm Scalability Simulations

Even though both SUS and truncation selection have $O(N^2)$ time complexity, truncation selection is slower than SUS by an order of magnitude. This is mainly due to the need for the truncation selection algorithm to calculate the top T% of the set for every selection performed. As expected, two-choice and random selections are observed to be constant-time algorithms. Two-choice is slightly slower than random selection due to the second selection and comparison operation that must occur.

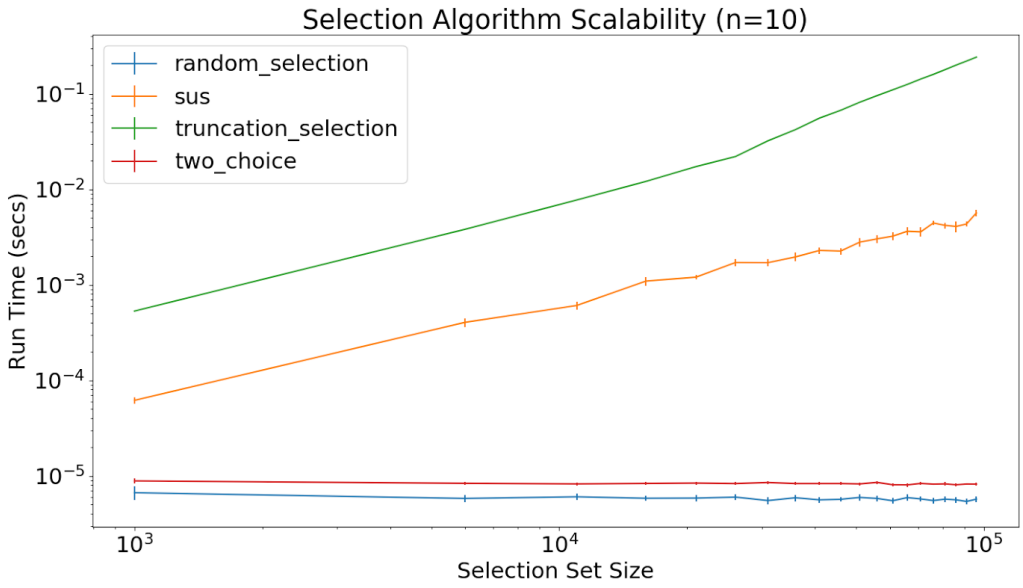

Figure 12: Running times of various weight random selection algorithms.

3.4 WeightedVector Class

It's necessary to perform repeated weighted sampling of disk IDs, easily, and given their fitness values. An object container class is needed similar to a C++ STL con-

tainer [24] that can also access elements of arbitrary types via arbitrary weighted random sampling methods. The WeightedVector class was implemented to fulfill this requirement.

Upon instantiation of the WeightedVector class, a fitness function is provided to store the fitness values of all objects internally. The WeightedVector class has implemented insertion, removal, and weighted random sampling of objects of some arbitrary type, T, based on the object's fitness value as a weight.

The set of objects and their fitness values are stored in $std :: vector$ objects , one for the object references stored in the WeightedVector and one for their corresponding fitness values. The objects and their corresponding fitness values share an index into their respective internal vectors.

Table **??** shows the WeightedVector public interface.

3.4.1 Weighted Vector Internals

Internally, the WeightedVector class keeps 3 different std:vector objects to keep track of sampling state and the object-to-fitness value mapping. There is the inner_vector_ variable which stores the objects inserted via EmplaceBack(), the weights_ variable which stores the fitness values of all objects in the inner_vector_, and the sampling_weights_ variable which is identical to weights_ except that objects weights are set to 0 when sampled. This allows us to exclude objects from being sampled multiple times with time complexity $O(1)$, since a weight of 0 gives a probability of 0 for sampling using Stochastic Universal Sampling as shown in the example below.

Before sampling, we have identical weights_ and sampling_weights_ accompanying an inner_vector_ of objects:

Suppose we then call Sample() on the WeightedVector and it returns D. The probability of this event is 0.5, so to maintain sampling state within the WeightedVector,

Table 3: WeightedVector public interface.

`WeightedVector(const std::function<double(const T)>)`
The WeightedVector class is instantiated by providing a fitness function that accepts a reference to an object of type T and returns a fitness value that is of type *double*. The provided fitness function is used to create a mapping between inserted objects and their corresponding fitness values for sampling. The WeightedVector class asserts that the fitness function returns positive values.
`void EmplaceBack(const T& element)`
The EmplaceBack function constructs and inserts an object of type T at the end of the vector. This increases the WeightedVector's size by 1.
`bool Empty()`
Returns true if the WeightedVector is of size 0.
`void Clear()`
Resets all state in the WeightedVector object, sets the size to 0, and clears all internal vectors.
`T& Sample()`
Returns a reference to an object stored inside of the WeightedVector. Objects are sampled via weighted random selection and subsequent calls to Sample are guaranteed not to return the same object unless Reset is called before sampling again. The WeightedVector class asserts that Sample is not called more times than the size of the WeightedVector.
`void Reset()`
Resets the sampling state of the WeightedVector, allowing sampled objects to be eligible for selection in subsequent Sample calls.
`size_t size()`
Returns the size of the WeightedVector.

Table 4: Inner vector example part 1.

`inner_vector_`	A	B	C	D
`weights_`	1	3	3	7
`sampling_weights_`	1	3	3	7

the weight associated with object D is set to zero:

Table 5: Inner vector example part 2.

inner_vector_	A	B	C	D
weights_	1	3	3	7
sampling_weights_	1	3	3	0

If we call Sample() again, it is not possible to return D since its sampling weight is now zero. The probabilities of returning A, B, or C in subsequent calls to Sample() are $\frac{1}{7}$, $\frac{3}{7}$, and $\frac{3}{7}$ respectively. Suppose two more calls to Sample() lead to the following state:

Table 6: Inner vector example part 3.

inner_vector_	A	B	C	D
weights_	1	3	3	7
sampling_weights_	0	3	0	0

We may only call Sample() one more time without triggering an assertion failure and the WeightedVector is obligated to return D. The only way to restore sampling state is to call Reset(). A Reset call simply copies weights_ into sampling_weights_ and all objects are eligible for sampling again:

Table 7: Inner vector example part 3.

inner_vector_	A	B	C	D
weights_	1	3	3	7
sampling_weights_	1	3	3	7

3.4.2 Weighted Vector Unit Testing

The WeightedVector class' unit test has four phases:

1. Test Average() functionality
2. Test there are no duplicate samples
3. Test sampling with uniform probabilities
4. Test sampling with non-uniform probabilities

The testing of the Average() function's behavior includes simply adding all zeros, all ones, and monotonically increasing integers in the range $[0, N]$. For each phase, we verify that the reported average is 0, 1, and $\frac{N(N-1)}{2N}$ respectively.

Verification that there are no duplicate samples involves adding monotonically increasing integers in the range $[0, N]$, inserting all sampled elements into a hash set, and verifying that the size of the hash set is equal to the size of the WeightedVector.

To test sampling of objects with uniform and non-uniform weights, I add monotonically increasing integers in the range $[0, N]$ to the WeightedVector. The fitness function provided for the uniform test simply returns a weight of 1 for all objects inserted into the WeightedVector. Elements are then sampled and Reset() is called for a number of times several orders of magnitude larger than the number of elements. Each sampled element's number of times being selected is tracked in a test-local hash map. We then calculate the actual selection probability of each element in the WeightedVector with the expected value and verify the difference is within an acceptable tolerance. For the uniform test, we expect all integers to be sampled roughly the same amount and for the non-uniform test, we expect larger integers to be sampled an amount of times proportional to their value.

3.5 Replica Selection Changes

Stargate was modified to store disk IDs in a WeightedVector rather than a `std::vector`. When considering a disk for candidacy, rather than shuffling the `std::vector` and iterating through each disk until enough replica targets are found, we call `WeightedVector::Sample()` until enough replica targets are found. All of Stargate's replica placement logic is untouched and I've simply modified the order in which candidate disks are considered.

4 Evaluation and Results

The experiments below seek to measure the effect of the additive and multiplicative term fitness functions on both the tier utilization of each node and the queue lengths of disks residing on those nodes compared with uniform random selection.

4.1 Experimental Setup

The replica selection schemes were evaluated using a NX-1350 for evaluating the disk fullness and a NX-3-node cluster). Each node contains a single 300GB SSD and 4 HDDs 1TB in size. When evaluating the new replica disk selection framework, two heterogeneous workload scenarios are tested:

1. Two worker VMs on separate nodes running a workload with low outstanding ops.

2. Two worker VMs on separate nodes with running a workload with high outstanding ops.

4.2 Fio and Write Patterns

When generating I/O in these experiments, Fio is used on the worker VMs. Fio, short for Flexible IO, is an I/O workload generator that can take configuration files to specify the parameters of a test. On each worker VM, fio is used to generate a sequential write workload that completely fills the cluster's hot-tier. I choose to use exclusively sequential writes for all tests because they are the default for fio tests and the purpose of these experiments is to generate new replicas in a consistent manner. For the purposes of replica placement, the Nutanix file system does not distinguish targets based on the write pattern that generatred an extent group.

4.3 Tier Utilization Experiments

The tier utilization experiments define the hot-tier deviation, $d_{hottier}$, as the average SSD utilization percentage of the nodes running a workload, u_w, subtracted by the SSD utilization percentage of the node without a workload, u_o:

$$d_{hot\ tier} = \frac{u_{w1} + u_{w2}}{u_o} \tag{5}$$

Ideally, the idle node would absorb the majority of secondary replicas from the running workloads. However, uniform random selection causes only 50% of secondary replicas to go to the idle node even though it can potentially handle more work due to the fact that it does not have to service a local workload. In a uniform random replica selection scheme, we expect the nodes running a workload have to bear 100% of their own primary replicas and 50% of secondary replicas from the other worker node. This causes total SSD utilization to be skewed towards the worker nodes and for this skew to grow as the tests run. This is indicated by higher $d_{hot\ tier}$ values. We expect a more sophisticated replica selection scheme to minimize $d_{hot\ tier}$ by biasing secondary replicas towards the idle node and limiting the skew.

4.3.1 Low Outstanding Operation Results

Figure 13 shows the results of a workload with only a single outstanding operation. This causes the queue length reported by Stargate to be at most 1, resulting in the fitness function's queue length term to be roughly constant and approximately 1.

We can see that the additive fitness function does not minimize the hot-tier deviation as well as the multiplicative. This is because an additive fitness function's behavior varies depending on the weight chosen for the queue length term. By default, the linear fitness function gives equal weight to both the disk fullness and queue

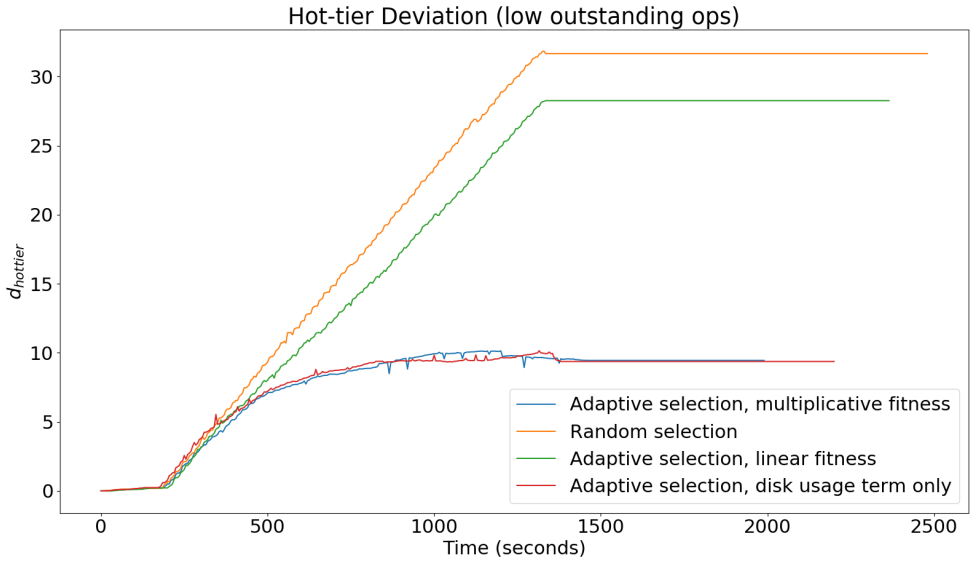

Figure 13: $d_{hot\ tier}$ values over time for low outstanding I/O operations.

length terms; however, for one run of this experiment the queue length term was given no weight. Linear fitness that gives equal weight to both terms does not reduce skew by very much while giving no weight to the queue length term keeps all nodes' usages within 10% of each other. This is because the queue length term is contributing the maximum amount possible to the fitness value due to the consistently low queue length values. We can illustrate this by defining a disk selection bias, b_r, as the probability some disk, d, will be selected when compared with another disk, d' whose utilization is 10% higher and queue length value is identical. Given a fitness function, f, we can calculate b_r as follows:

$$b_r = \frac{f(d)}{f(d) + f(d')} \tag{6}$$

Figures 14 and 15 show the disk selection biases for very large and very small queue lengths respectively.

We can see that for an additive fitness function, the bias towards a less utilized

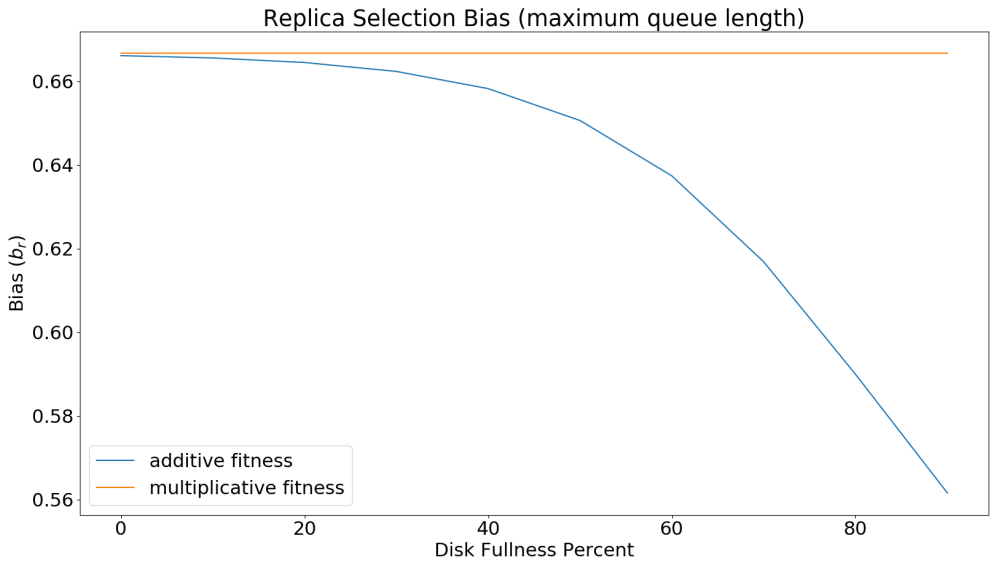

Figure 14: b_r values with static queue lengths at the fitness function ceiling values.

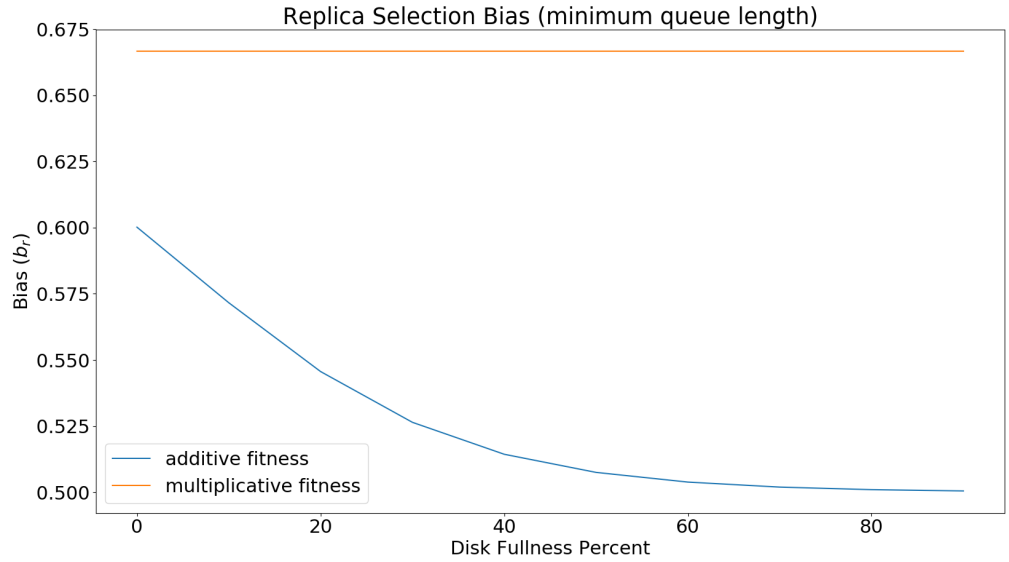

Figure 15: b_r values with static queue lengths at 1.

disk decreases as the disk fullness percentages for d and d' increase, even though they still only differ by 10% in the figure above. This is because the entire linear fitness function does not scale with each term, so we can conclude that the multiplicative fitness function is superior.

4.3.2 High Outstanding Operation Results

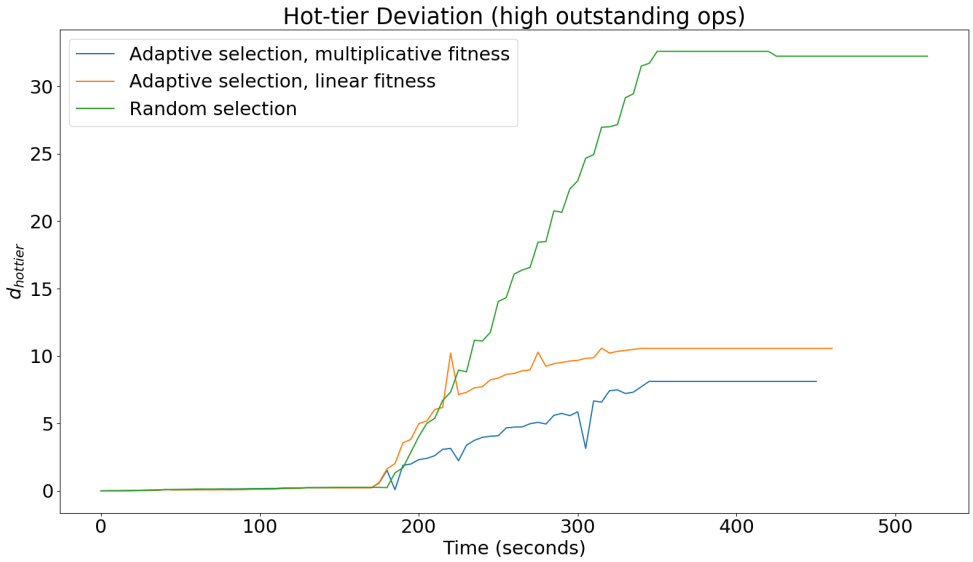

Figure 16: $d_{hot\ tier}$ values over time for low outstanding I/O operations.

In Figure 16, we can see that both additive and multiplicative fitness functions reduce the disk fullness skew from 30% to less than 10%. Multiplicative fitness performs slightly better at minimizing $d_{hot\ tier}$ than additive fitness, possibly due to scaling the fitness value by the value of both the fullness and queue length terms, rather than weights.

4.4 Disk Queue Length Experiments

Since a low outstanding operation workload would not give useful information for measuring the effects of fitness-based replica selection on disk queue lengths, the high outstanding I/O operation experiment in the previous section was re-run for all fitness function types and for fitness function queue length term ceilings of 200 and 100. Figures 17 and 18 show a reduction in queue length quartiles when fitness-based selection is used for disks on nodes that host local workloads. Lower queue length ceilings are observed to provide better results in reducing the queue lengths for the worker nodes.

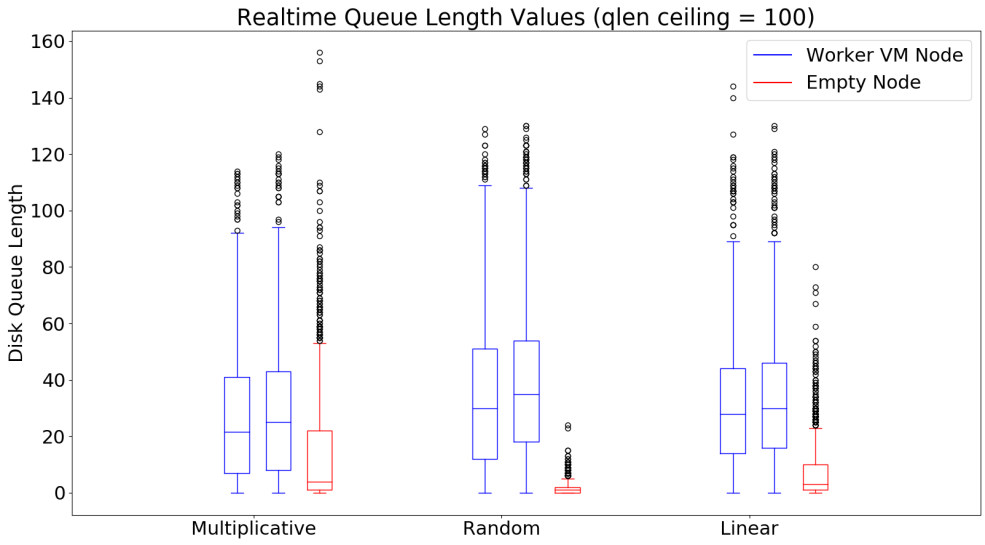

Figure 17: Queue lengths for all SSDs on the specified nodes sampled every 1 second.

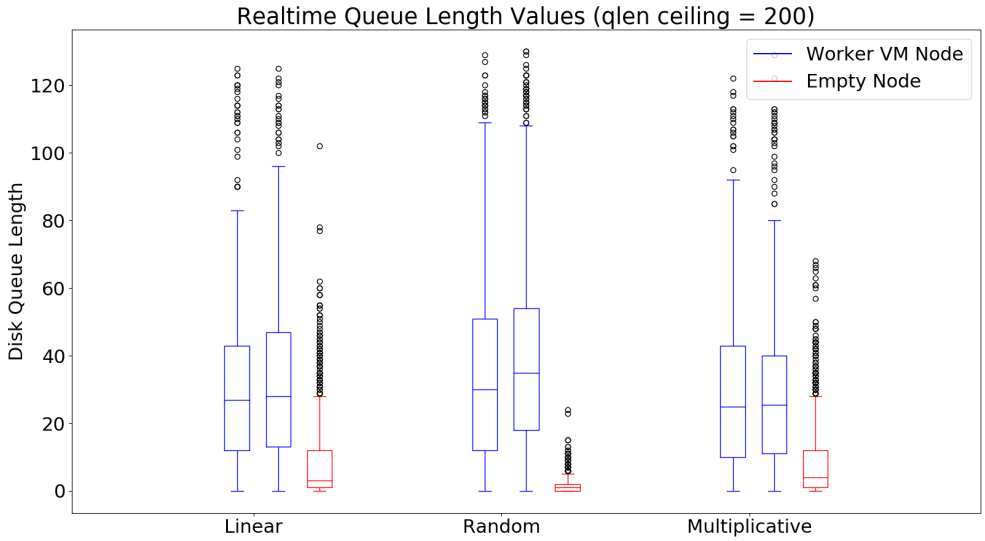

Figure 18: Queue lengths for all SSDs on the specified nodes sampled every 1 second.

5 User Guide

5.1 Scraping Data From the Nutanix Cluster

Stargate exposes a webserver on the CVM, listening on port 2009, that exposes various real-time stats such as the number of operations in flight, disk throughput, disk queue length, and many other pieces of information. I wrote a Python script to parse and derive the following information for each node:

1. SSD tier usage
2. SSD tier availability
3. SSD tier max capacity
4. Read/Write counts for each disk
5. Queue lengths for each disk

The script will filter out all disks in the HDD tier and only pull the statistics for the SSD tier. Upon each invocation of the script, information is appended to a file for each node that is created if it is non-existent. The data is layed out in a CSV format for easy plotting via the Matplotlib library [22]. For each experiment, the script was called every 5 seconds via the command `watch -n 5`.

5.2 Re-creating Experiments

An NX-1350 with VMware ESXi 5.5u2 hypervisor and a single 300GB SSD on each node was used for all experiments in this thesis. This means that for an RF2 cluster, there is less than 150GB of useable hot tier available since a small amount of space is reserved on each disk for other services on the CVM.

A single CentOS 6.5 worker VM was manually created on a single node with 4GB RAM, single 150GB disk, and installation of fio. The 150GB disk size ensures that the entire hot tier of each node will be fully utilized when the disk is filled via workload generation on each node. Each other node then clones the worker VM so that all nodes in the cluster have an identical VM and tier utilization.

The fio script used to generate a sequential write workload on each node is as follows:

```
[global]
direct=1
ioengine=libaio
bs=32k
iodepth=128
randrepeat=0
group_reporting
filesize=150G

[job1]
rw=write
filename=/dev/sdb
name=sequential-write
```

The `iodepth` variable is modified depending on the number of outstanding I/Os needed.

When resetting the cluster for a new test, on the CVM I run `cluster stop ; cluster destroy` and swap out the Stargate binary in the `/home/nutanix/bin/` directory with the binary required for the next test. When ready, I recreate the Nutanix cluster via `cluster -s <cvm IPs> create`. The process is then repeated for each test.

6 Future Work

6.1 Real-time Fitness Feedback

The replica placement scheme shown in this work relies on periodic stats updates, so the system is always working with older stats for data placement decisions. One change that can be made to this system is to track the time to completion of various ops that place data on each disk and bias the replica placements towards the faster disks. This would require Stargate to keep historical latency data and would remove the dependence on a centralized stats repository.

6.2 Read Replica Selection

When choosing which replicas to read from, we always select local disks form the SSD tier and sometimes select remote HDDs at random. This process does not take the new disk fitness information that is available and would benefit from adapting read replica selection decisions based on the disk fitness values.

6.3 More Fitness Function Variables

The disk fitness values do not need to be limited to derivation from disk queue length and fullness percentage. The cluster tracks many other variables such as average node CPU utilization, number of Stargate failures in a specified time window, number of active user VMs, and data access patterns among other pieces of information. More investigation should occur to see how this data can be used to make better data placement decisions.

A Appendix

A.1 Herding Behavior Due to Implementation Bug

During the 128 outstanding op experiments, herding behavior was observed unexpectedly after implementing fitness based replica placement as shown in Figure 19. By default, disk usage and performance stats are supposed to be refreshed every 10 seconds. This is frequent enough to avoid herding behavior, but the 128 outstanding op experiments exhibited herding.

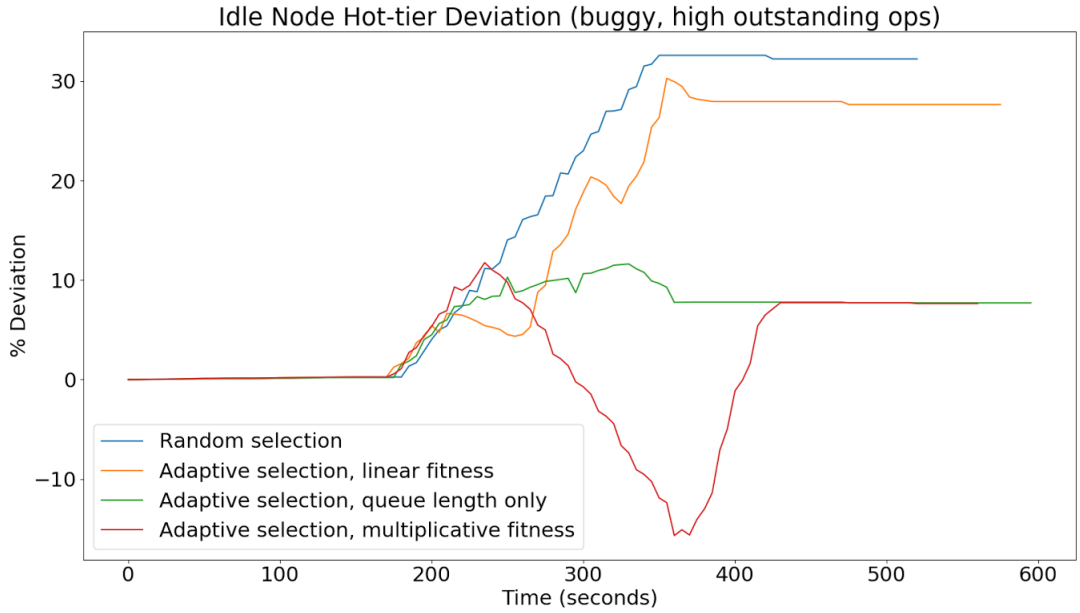

Figure 19: $d_{hot\ tier}$ values over time for low outstanding I/O operations. This set of experiments contains the stats update bug.

The experiment with additive fitness seemed to only exhibit mild herding behavior, whereas the test with multiplicative fitness showcased much more dramatic shifts in SSD usage skews. In conjunction with the complete absence of this behavior in the single outstanding op experiments, it was thought to be highly likely that this herding behavior was caused by a bug in the queue length term of the fitness functions. An

additive fitness function is less affected by this bug due to the attenuated effect of each term in the fitness function.

Within Stargate, we keep a mapping from disk ID to a DiskState object (called `disk_map_`) containing information and cached statistics related to the disk. The disk performance and disk usage stats are two separate elements within the DiskState structure.

Every 10 seconds, an alarm handler will execute and iterate through each active disk in the cluster and asynchronously query disk stats and bind a callback to each query to be executed when a response is received. Disk usage and performance lookups each have their own callback functions:

Table 8: Disk usage and performance lookup callback functions.

Function Name	**Description**
`UsageStatLookupCallback`	Decrement the outstanding stats lookup counter, acquire lock and populate performance stats in `disk_map_`, and leave performance stats untouched.
`PerformanceStatLookupCallback`	Decrement outstanding stats lookup counter, lock and populate performance stats in `disk_map_`, and leave usage stats untouched.

The two callbacks introduce a race condition regarding `disk_map_` even though the structure is locked. Any time PerformanceStatLookupCallback returns before the callback for usage stats, all performance stats will be cleared and cause the fitness function to assume worst-case values for the queue length term.

This problem was fixed by simply serializing our usage and performance stats lookups.

References

[1] Poitras, S. (n.d.) The Nutanix Bible. Retrieved August 09, 2017, from http://www.nutanixbible.com/

[2] Lakshman, A., and Malik, P. (2008, August 25). Cassandra A structured storage system on a P2P Network. Retrieved February 15, 2016, from https://www.facebook.com/notes/facebook-engineering/cassandra-a-structured-storage-system-on-a-p2p-network/24413138919/

[3] Dean, J., and Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. Communications of the ACM, 51(1), 107-113.

[4] Hadoop, A. (2009). Hadoop. 2009-03-06]. http://hadoop.apache.org.

[5] Xie, J., Yin, S., Ruan, X., Ding, Z., Tian, Y., Majors, J., ... and Qin, X. (2010, April). Improving mapreduce performance through data placement in heterogeneous hadoop clusters. In Parallel and Distributed Processing, Workshops and Phd Forum (IPDPSW), 2010 IEEE International Symposium on (pp. 1-9). IEEE.

[6] Zaharia, M., Konwinski, A., Joseph, A. D., Katz, R. H., and Stoica, I. (2008, December). Improving MapReduce Performance in Heterogeneous Environments. In OSDI (Vol. 8, No. 4, p. 7).

[7] Jin, H., Yang, X., Sun, X. H., and Raicu, I. (2012, June). Adapt: Availability-aware mapreduce data placement for non-dedicated distributed computing. In Distributed Computing Systems (ICDCS), 2012 IEEE 32nd International Conference on (pp. 516-525). IEEE.

[8] Perez, J. M., Garcia, F., Carretero, J., Calderon, A., and Sanchez, L. M. (2003, May). Data allocation and load balancing for heterogeneous cluster storage systems. In Cluster Computing and the Grid, 2003. Proceedings. CCGrid 2003. 3rd IEEE/ACM International Symposium on (pp. 718-723). IEEE.

[9] Schlierkamp-Voosen, D., and Mhlenbein, H. (1993). Predictive models for the breeder genetic algorithm. Evolutionary Computation, 1(1), 25-49.

[10] Baker, J. E. (1987, July). Reducing bias and inefficiency in the selection algorithm. In Proceedings of the second international conference on genetic algorithms (pp. 14-21).

[11] Vitter, J. S. (1985). Random sampling with a reservoir. ACM Transactions on Mathematical Software (TOMS), 11(1), 37-57.

[12] Efraimidis, P. S., and Spirakis, P. G. (2006). Weighted random sampling with a reservoir. Information Processing Letters, 97(5), 181-185.

[13] Chao, M. T. (1982). A general purpose unequal probability sampling plan. Biometrika, 69(3), 653-656.

[14] Chaubal, C. (2008). The architecture of vmware esxi. VMware White Paper, 1(7).

[15] Borthakur, D. (2008). HDFS architecture guide. HADOOP APACHE PROJECT http://hadoop. apache. org/common/docs/current/hdfs design. pdf, 39.

[16] Suresh, L., Canini, M., Schmid, S., and Feldmann, A. (2015). C3: Cutting tail latency in cloud data stores via adaptive replica selection. In 12th USENIX Symposium on Networked Systems Design and Implementation (NSDI 15) (pp. 513-527).

[17] Velte, A., and Velte, T. (2009). Microsoft virtualization with Hyper-V. McGraw-Hill, Inc..

[18] Lamport, L. (2005). Generalized consensus and Paxos. Technical Report MSR-TR-2005-33, Microsoft Research.s

[19] Y. Azar, A. Broder, A. Karlin, and E. Upfal. Balanced allocations. In Proceedings of the 26th ACM Symposium on the Theory of Computing, pages 593{602, 1994.

[20] Crow, J. F., and Kimura, M. (1979). Efficiency of truncation selection. Proceedings of the National Academy of Sciences of the United States of America, 76(1), 396399.

[21] Mitzenmacher, M. (2001). The power of two choices in randomized load balancing. IEEE Transactions on Parallel and Distributed Systems, 12(10), 1094-1104.

[22] Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing In Science and Engineering, 9(3), 90-95.

[23] Van Rossum, G., and Drake, F. L. (2003). Python language reference manual (p. 144). Network Theory.

[24] Plauger, P. J., Lee, M., Musser, D., and Stepanov, A. A. (2000). C++ Standard Template Library. Prentice Hall PTR.

[25] Ghemawat, S., Gobioff, H., and Leung, S. T. (2003, October). The Google file system. In ACM SIGOPS operating systems review (Vol. 37, No. 5, pp. 29-43). ACM.

[26] Weil, S. A., Brandt, S. A., Miller, E. L., Long, D. D., and Maltzahn, C. (2006, November). Ceph: A scalable, high-performance distributed file system. In Pro-

ceedings of the 7th symposium on Operating systems design and implementation (pp. 307-320). USENIX Association.

[27] Patterson, D. A., Gibson, G., and Katz, R. H. (1988). A case for redundant arrays of inexpensive disks (RAID) (Vol. 17, No. 3, pp. 109-116). ACM.

[28] Chen, P. M., Lee, E. K., Gibson, G. A., Katz, R. H., and Patterson, D. A. (1994). RAID: High-performance, reliable secondary storage. ACM Computing Surveys (CSUR), 26(2), 145-185.