HopsFS: Scaling Hierarchical File System Metadata Using NewSQL Databases

Article · June 2016

6 authors, including:

Salman Niazi
KTH Royal Institute of Technology
12 PUBLICATIONS   **119** CITATIONS

SEE PROFILE

Mahmoud Ismail
KTH Royal Institute of Technology
8 PUBLICATIONS   **74** CITATIONS

SEE PROFILE

Mikael Ronström
Oracle Corporation
16 PUBLICATIONS   **141** CITATIONS

SEE PROFILE

Seif Haridi
KTH Royal Institute of Technology
237 PUBLICATIONS   **5,559** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project   Penny - A parallel implementation of AKL View project

Project   distributed system for large scale applications View project

HopsFS: Scaling Hierarchical File System Metadata Using NewSQL Databases

Salman Niazi, Mahmoud Ismail,
Seif Haridi, Jim Dowling
KTH - Royal Institute of Technology
{smkniazi, maism, haridi, jdowling}@kth.se

Steffen Grohsschmiedt
Spotify AB
steffeng@spotify.com

Mikael Ronström
Oracle
mikael.ronstrom@oracle.com

Abstract

Recent improvements in both the performance and scalability of shared-nothing, transactional, in-memory NewSQL databases have reopened the research question of whether distributed metadata for hierarchical file systems can be managed using commodity databases. In this paper, we introduce HopsFS, a next generation distribution of the Hadoop Distributed File System (HDFS) that replaces HDFS' single node in-memory metadata service, with a distributed metadata service built on a NewSQL database. By removing the metadata bottleneck, HopsFS improves capacity and throughput compared to HDFS. HopsFS can store 24 times more metadata than HDFS. We also provide public, fully reproducible experiments based on a workload trace from Spotify that show HopsFS has 2.6 times the throughput of Apache HDFS, lower latency for greater than 400 concurrent clients, and no downtime during failover. Finally, and most significantly, HopsFS allows metadata to be exported to external systems, analyzed or searched online, and easily extended.

1 Introduction

Distributed file systems are an important infrastructure component of many large scale data-parallel processing systems, such as MapReduce [14], Dryad [28], Flink [5] and Spark [77]. By the end of this decade, data centers storing multiple exabytes of data will not be uncommon [13, 48]. For large distributed hierarchical file systems, the metadata management service is the scalability bottleneck [64]. Many existing distributed file systems store their metadata on either a single node or a shared disk architecture, both of which have limited scalability. Well known examples include GFS [18], HDFS [63], QFS [43], Farsite [3], Lustre [60], Ursa Minor [2], GPFS [59], Frangipani [69], GlobalFS [51], and Panasas [73]. Other systems scale out their metadata by statically sharding the namespace and storing the shards on different hosts, such as NFS [46], AFS [38], MapR [66], Locus [50], Coda [58], Sprite [42] and XtreemFS [27]. However, statically sharding the namespace negatively affects file system operations that cross different shards, in particular *move* and *rename* operations. Also, it complicates the management of the file system, as administrators have to map metadata servers to namespace shards that change in size over time.

Recent improvements in both the performance and scalability of shared-nothing, transactional, in-memory NewSQL [44] databases have reopened the possibility of storing distributed file system metadata in a commodity database. To date, the conventional wisdom has been that it is too expensive (in terms of throughput and latency) to store hierarchical file system metadata fully normalized in a distributed database [61, 34].

In this paper we show how to build a high throughput and low operational latency distributed file system using a NewSQL database. We present HopsFS, a new distribution of the Hadoop Distributed File System (HDFS), which decouples file system metadata storage and management services. HopsFS stores all metadata normalized in a highly available distributed relational database called Network Database (NDB) – a NewSQL Storage Engine for MySQL Cluster [39, 55]. HopsFS provides redundant stateless servers (namenodes) that in parallel, read and update metadata stored in the database.

HopsFS encapsulates file system operations in distributed transactions. To improve the performance of file system operations, we leverage both classical database techniques such as *batching* (bulk operations) and *write-ahead* caches within transactions, as well as distribution aware techniques commonly found in NewSQL databases. These distribution aware NewSQL techniques include *application defined partitioning* (we partition the namespace such that the metadata for all immediate descendants of a directory (child files/directories) reside on the same database shard for efficient directory listing and file read operations), and *distribution aware transactions* (we start a transaction on the database shard that stores all/most of the metadata required for the file system operation), and *partition pruned index scans* (scan operations are localized to a single database shard [78]). We also introduce an inode hints cache for faster resolution of file paths. Cache hits when resolving a path of depth N can

reduce the number of database round trips from *N* to *1*.

However, some file system operations on large directory subtrees (such as *move*, *rename*, and *delete*) may be too large to fit in a single database transaction. For example, deleting a folder containing millions of files cannot be performed using a single database transaction due to the limitations imposed by the database management system on the number of operations that can be included in a single transaction. For these *subtree operations*, we introduce a novel protocol that uses application level distributed locking mechanism to isolate large subtrees to perform file system operations. After isolating the subtrees large file system operations are broken down into smaller transactions that execute in parallel for performance. The subtree operations protocol ensures that the consistency of the namespace is not violated when the namenode executing the operation fails.

HopsFS is a drop-in replacement for the Hadoop Distributed File System (HDFS) [63]. HopsFS is running in production, providing Hadoop-as-a-Service at SICS ICE data center in Luleå, Sweden [65]. Our evaluation shows that HopsFS can store 24 times more metadata than Apache HDFS. In experiments, using a real-world workload generated by Hadoop/Spark applications from Spotify, we show that HopsFS delivers 2.6 times higher throughput than HDFS, and HopsFS has no downtime during failover. All the experiments are public and can be fully reproduced with a few mouse clicks on Amazon Web Services (AWS) using Karamel [29]. To the best of our knowledge HopsFS is the first open-source distributed file system that stores fully normalized metadata in a distributed relational database.

2 Background

This section describes HDFS and MySQL Cluster Network Database storage engine.

2.1 Hadoop Distributed File System(HDFS)

HDFS [63] is an open source implementation of the Google File System [18]. In HDFS, the file system metadata is stored on the heap of single Java process called the Active NameNode (ANN), see Figure 1. The files are split into small (typically 128 MB) blocks that are by default triple replicated across the datanodes. For high availability of the metadata management service, the Active namenode logs changes to the metadata to journal servers using quorum based replication. The metadata change log is replicated asynchronously to a Standby NameNode (SbNN), which also performs checkpointing functionality. In HDFS, the ZooKeeper coordination service [26] enables both agreement on which machine is running the active namenode as well as coordinating failover from the active to the standby namenode.

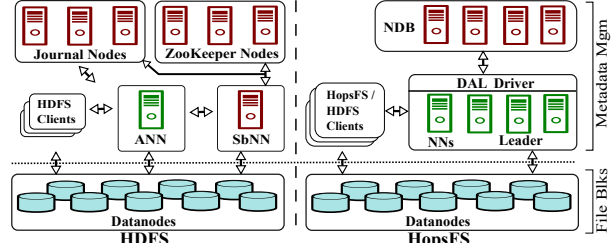The namenode serves requests from potentially thou-

Figure 1: System architecture for HDFS and HopsFS. For high availability, HDFS requires an Active NameNode (ANN), at least one Standby NameNode (SbNN), at least three Journal Nodes for quorum-based replication of the write ahead log of metadata changes, and at least three ZooKeeper instances for quorum based coordination. HopsFS supports multiple stateless namenodes that access the metadata stored in NDB database nodes.

sands of datanodes and clients, and keeps the metadata strongly consistent by executing the file system operations atomically. The namenode implements atomic operations using a single global lock on the entire file system metadata, providing single-writer, multiple-readers concurrency semantics. In order to prevent large file system operations from holding the namespace lock for long durations, which could starve other file system operations, large operations are executed in multiple phases, hence such operation are not atomic. For example, deleting large directories is performed incrementally, that is, in the first phase the inodes are deleted and then the later phases the blocks are deleted incrementally in small batches.

The datanodes are connected to both active and standby namenodes. All the datanodes periodically generate a block report containing information about its own stored blocks. The namenode processes the block report to validate the consistency of the namenode's blocks map with the blocks actually stored at the datanode.

In HDFS the amount of metadata is quite low relative to file data. There is approximately 1 gigabyte of metadata for every petabyte of file system data [64]. Spotify's HDFS cluster has 1600+ nodes, storing 60 petabytes of data, but its metadata fits in 140 gigabytes Java Virtual Machine (JVM) heap. The extra heap space is taken by temporary objects, RPC request queues and secondary metadata required for the maintenance of the file system. However, current trends are towards even larger HDFS clusters (Facebook has HDFS clusters with more than 100 petabytes of data [49]), but current JVM garbage collection technology does not permit very large heap sizes, as the application pauses caused by the JVM garbage collector affects the operations of HDFS [23]. As such, JVM garbage collection technology and the monolithic architecture of the HDFS namenode are now the scalability bottlenecks for Hadoop [64]. Another limitation with this architecture is that data structures are optimized to reduce their memory footprint with the result that metadata is difficult to modify or export to external systems.

2.2 MySQL Cluster

MySQL Cluster is a shared-nothing, replicated, in-memory, auto-sharding, *consistent*, NewSQL relational database [39]. Network DataBase (NDB) is the storage engine for MySQL Cluster. NDB horizontally partitions the tables among storage nodes called NDB datanodes. NDB supports datanode-level failure recovery using transaction redo and undo logs. NDB datanodes also asynchronously snapshot their state to prevent the logs growing too big and to improve datanode recovery time. Cluster-level recovery is supported using a global checkpointing protocol that increments a global epoch counter. On cluster-level recovery, datanodes recover all transactions to the latest epoch-id. NDB also supports application defined partitioning (ADP) for the tables. Transaction coordinators are located at all NDB datanodes, enabling high performance transactions between data shards. Distribution aware transactions (DAT) are possible by providing a *hint*, based on the application defined partitioning scheme, to start a transaction on the NDB datanode containing the data read/updated by the transaction. In particular, single row read operations and partition pruned index scans (scan operations in which a single data shard participates) benefit from distribution aware transactions as they can read all their data locally [78]. Incorrect hints result in additional network traffic being incurred but otherwise correct system operation.

2.2.1 NDB Data Replication and Failure Handling

NDB datanodes are organized into node groups, where the data replication factor, R, determines the number of datanodes in a node group. Given a cluster size N, there are N/R node groups. NDB partitions tables (hash partitioning by default) into a fixed set of partitions distributed across the node groups. New node groups can be added online, and existing data can be rebalanced to the new node group. A partition is a fragment of data stored and replicated by a node group. Each datanode stores a copy (replica) of the partition assigned to the node group of which the node is a member. In NDB the default replication degree is two, which means that each node group can tolerate one NDB datanode failure as the other NDB datanode in the node group contains a full copy of the data assigned to the node group. So, an eight node NDB cluster has four node groups and it can tolerate four NDB datanode failures as long as there is one alive NDB datanode in each of the four node groups. To tolerate multiple failures in each node group the replication degree needs to be increased.

2.2.2 Transaction Isolation

NDB only supports *read-committed* transaction isolation, which does not allow *dirty* reads but *phantom* and *fuzzy* (non-repeatable) reads can happen in a transaction [7]. However, NDB supports row level locks, such as, *exclusive* (write) locks, *shared* (read) locks, and *read-committed* locks that can be used to isolate conflicting transactions.

3 HopsFS Overview

HopsFS is a fork of HDFS v2.0.4. Unlike HDFS, HopsFS provides a scale-out metadata layer by decoupling the metadata storage and manipulation services. HopsFS supports multiple stateless namenodes, written in Java, to handle clients' requests and process the metadata stored in an external distributed database, see Figure 1. Each namenode has a Data Access Layer (DAL) driver that, similar to JDBC, encapsulates all database operations allowing HopsFS to store the metadata in a variety of NewSQL databases. The internal management (housekeeping) operations, such as datanode failure handling, must be coordinated amongst the namenodes. HopsFS solves this problem by electing a leader namenode that is responsible for the housekeeping. We use the database as shared memory to implement a leader election and membership management service. Our leader election protocol defines an alive namenode as one that can write to the database in bounded time, details for which can be found in [57].

In HopsFS the clients can choose between *random*, *round-robin*, and *sticky* policies for selecting a namenode on which to execute file system operations. HopsFS clients periodically refresh the namenode list, enabling new namenodes to join an operational cluster. HDFS v2.x clients are fully compatible with HopsFS, although they do not distribute operations over namenodes, as they assume there is a single active namenode. Like HDFS the datanodes are connected to all the namenodes, however, the datanodes send the block reports to only one namenode. The leader namenode load balances block reports over alive namenodes.

In section 4, we discuss how HopsFS auto sharding scheme enables common file system operations to read the required metadata using few low cost database access queries that are localized to single database shard. Section 5 discusses how the consistency of the file system metadata is maintained by converting file system operations into distributed transactions, and how the latency of the distributed transactions is reduced using per-transaction and namenode level caches. Then, in section 6, a protocol is introduced to handle file system operations that are too large to fit in a single database transaction.

4 HopsFS Distributed Metadata

Metadata for hierarchical distributed file systems typically contains information on inodes, blocks, replicas, quotas, leases and mappings (directories to files, files to blocks, and blocks to replicas). When metadata is distributed, an application defined partitioning scheme is needed to shard the metadata and a distributed consensus pro-

Op Name	Percentage	Op Name	Percentage
append file	0.0%	content summary	0.01%
mkdirs	0.02%	set permissions	0.03% [26.3%*]
set replication	0.14%	set owner	0.32 % [100%*]
delete	0.75% [3.5%*]	create file	1.2%
rename	1.3% [0.03%*]	add blocks	1.5%
list (*listStatus*)	**9% [94.5%*]**	**stat (*fileInfo*)**	**17% [23.3%*]**
read (*getBlkLoc*)	**68.73%**		

Table 1: Relative frequency of file system operations for Spotify's HDFS cluster. List, read, and stat operations account for ≈ 95% of the metadata operations in the cluster.
*Of which, the relative percentage is on directories

Figure 2: (a) Shows the relative cost of different operations in NewSQL database. (b) HopsFS avoids FTS and IS operations as the cost these operation is relatively higher than PPIS, B, and PK operations.

tocol is required to ensure metadata integrity for operations that cross shards. Quorum-based consensus protocols, such as Paxos, provide high performance within a single shard, but are typically combined with transactions, implemented using the two-phase commit protocol, for operations that cross shards, as in Megastore [6] and Spanner [11]. File system operations in HopsFS are implemented primarily using transactions and row-level locks in MySQL Cluster to provide serializability [24] for metadata operations.

The choice of partitioning scheme for the hierarchical namespace is a key design decision for distributed metadata architectures. We base our partitioning scheme on the expected relative frequency of HDFS operations in production deployments and the cost of different database operations that can be used to implement the file system operations. Table 1 shows the relative frequency of selected HDFS operations in a workload generated by Hadoop applications, such as, *Pig*, *Hive*, *HBase*, *MapReduce*, *Tez*, *Spark*, and *Giraph* at Spotify. *List*, *stat* and *file read* operations alone account for ≈ 95% of the operations in the HDFS cluster. These statistics are similar to the published workloads for Hadoop clusters at Yahoo [1], LinkedIn [53], and Facebook [67]. Figure 2a shows the relative cost of different database operations. We can see that the cost of a full table scan or an index scan, in which all database shards participate, is much higher than a partition pruned index scan in which only a single database shard participates. HopsFS metadata design and metadata partitioning enables implementations of common file system operations using only the low cost database operations, that is, primary key operations, batched primary key operations and partition pruned index scans. For example, the read and directory listing operations (see Table 3), are implemented using only (batched) primary key lookups and partition pruned index scans. Index scans and full table scans were avoided, where possible, as they touch all database shards and scale poorly.

4.1 Entity Relation Model

In HopsFS, the file system metadata is stored in tables where a directory inode is represented by a single row in the *Inode* table. File inodes, however, have more associ-
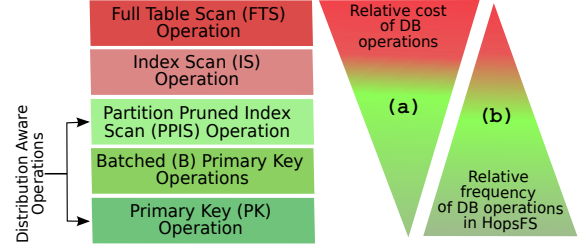
ated metadata, such as a set of blocks, block locations, and checksums that are stored in separate tables, with foreign keys ensuring the integrity of the data.

Figure 3 shows the Entity Relational model depicting key entities in the HopsFS metadata model. Files and directories are represented by the *Inode* entity that contains a reference to its parent inode (parent inode Id) in the file system hierarchy. We store path individual components, not full paths, in inode entries. Each file contains multiple blocks stored in the *Block* entity. The location of each block replica is stored in the *Replica* entity. During its life-cycle a block goes through various phases. Blocks may be under-replicated if a datanode fails and such blocks are stored in the under-replicated blocks table (*URB*). The replication manager, located on the leader namenode, sends commands to datanodes to create more replicas of under-replicated blocks. Blocks undergoing replication are stored in the pending replication blocks table (*PRB*). Similarly, a replica of a block has various states during its life-cycle. When a replica gets corrupted, it is moved to the corrupted replicas (*CR*) table. Whenever a client writes to a new block's replica, this replica is moved to the replica under construction (*RUC*) table. If too many replicas of a block exist (for example, due to recovery of a datanode that contains blocks that were re-replicated), the extra copies are stored in the excess replicas (*ER*) table and replicas that are scheduled for deletion are stored in the invalidation (*Inv*) table. Note that the file inode related entities also contain the inode's foreign key (not shown in figure 3) that is also the partition key, enabling HopsFS to read the file inode related metadata using partition pruned index scans.

4.2 Metadata Partitioning

HopsFS partitions inodes by their *parents' inode Ids*, resulting in inodes with the same parent inode being stored on the same database shard. For common file system operations, such as listing files in a directory, distribution aware transactions enable us to start a transaction on the database shard that holds the directory descendants and use a pruned index scan to retrieve the contents of the directory. File inode related metadata, that is, blocks, replica mappings and checksums, is partitioned using the file's *inode Id*, therefore, all the file related metadata is stored
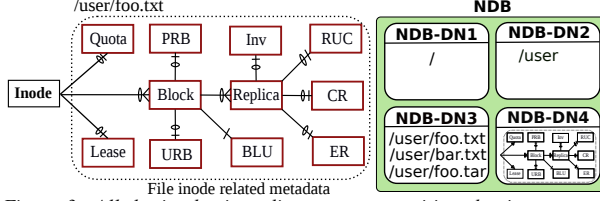
4

Figure 3: All the inodes in a directory are partitioned using a parent inode Id, therefore, all the immediate children of /user dir are stored on NDB-DN-3 for efficient directory listing, for example, ls /user. The file inode related metadata for /user/foo.txt is stored on NDB-DN-4 for efficient file reading operations, for example, cat /user/foo.txt.

in a single database shard for efficient file operations, see Figure 3.

It is not uncommon for big data applications to create millions of files in a single directory [52, 45]. HopsFS metadata partitioning scheme does not lead to hot spots as the database shard that contains the immediate children (files and directories) of a parent directory only stores the names and access attributes while the files inode related metadata, that comprises the majority of all metadata, is spread across all the database shards.

5   HopsFS Transactional Operations

Transactional metadata operations in HopsFS belong to one of the two categories: **Inode** operations that operate on single file, directory or block (for example, create/read file, mkdir, and block state change operations), and **subtree** operations that operate on an unknown number of inodes, potentially millions, (for example, recursive *delete*, *rename*, *chmod*, and *chown* on non-empty directories).

This section describes how HopsFS efficiently encapsulates inode operations in transactions in NDB. The strongest transaction isolation level provided by NDB is *read-committed*, which is not strong enough to provide at least as strong consistency semantics as HDFS which uses single global lock to serialize all HDFS operations. To this end, we use row-level locking to serialize conflicting inode operations. That is, the operations execute in parallel as long as they do not take conflicting locks on the same inodes. However, taking multiple locks in a transaction could lead to extensive deadlocking and transaction timeouts. The reasons are:

Cyclic Deadlocks: In HDFS, not all inode operations follow the same order in locking the metadata which would lead to cyclic deadlocks in our case. To solve this problem, we have reimplemented all inode operations so that they acquire locks on the metadata in the same *total order*, traversing the file system tree from the root down to leave nodes using left-ordered depth-first search.

Lock Upgrades: In HDFS, many inode operations contain read operations followed by write operations on the same metadata. When translated into database operations within the same transaction, this results in deadlocking due to lock upgrades from read to exclusive locks. We

have examined all locks acquired by the inode operations, and re-implemented them so that all data needed in a transaction is read only once at the start of the transaction (see Lock Phase, section 5.2.1) at the strongest lock level that could be needed during the transaction, thus preventing lock upgrades.

5.1   Inode Hint Cache

Resolving paths and checking permissions is by far the most common operation in most HDFS workloads, see Table 1. In HDFS, the full path is recursively resolved into individual components. In HopsFS for a path of depth N, it would require N consecutive database operations to retrieve file path components, resulting in high latency for file system operations.

Similar to AFS [38] and Sprite [42], we use *hints* [31] to speed up the path lookups. Hints are mechanisms to quickly retrieve file path components in parallel (batched operations). In our partitioning scheme, inodes have a composite primary key consisting of the parent inode's Id and the name of the inode (that is, file or directory name), with the parent inode's id acting as the partition key. Each namenode caches only the primary keys of the inodes. Given a pathname and a hit for all path components directories, we can discover the primary keys for all the path components which are used to read the path components in parallel using a single database batch query containing only primary key lookups.

5.1.1   Cache Consistency

Reading all path components using a batch query subsequently validates the cache entires. If a hint is wrong, a primary key read operation fails and path resolution falls back to recursive method for resolving file path components, followed by updating the cache. The cache entries infrequently become stale as rename and move operations, that change the primary key for an inode, are less than 2% of operations in typical Hadoop workloads, see Table 1. Moreover, typical file access patterns follow a heavy-tailed distribution (in Yahoo 3% of files account for 80% of accesses [1]), and using a *sticky* policy for HopsFS clients improves temporal locality and cache hit rates.

5.2   Inode Operations

HopsFS implements a pessimistic control model that supports parallel read and write operations on the namespace, serializing conflicting inode and subtree operations. We chose a pessimistic scheme as, in contrast to optimistic concurrency control, it delivers good performance for medium to high levels of resource utilization [4], and many HDFS clusters, such as Spotify's, run at high load. Inode operations are encapsulated in a single transaction that consists of three distinct phases, which are, **lock**, **execute**, and **update** phases.

```
 1. Get path components hints from the inodes hint cache
 2. Set partition key hint for the transaction
BEGIN TRANSACTION
LOCK PHASE:
 3. Using the inode hints validate the inodes till the
    penultimate inode with a batched operation
 4. If (cache miss | invalid path component) then
      recursively resolve the path & update the cache
 5. Lock and read the last inode
 6. Read Lease, Quota, Blocks, Replica, URB, PRB,
    RUC, CR, ER, Inv using partition pruned index scans
EXECUTE PHASE:
 7. Manipulate the data stored in the transaction cache
UPDATE PHASE:
 8. Transfer the changes to database in batches
COMMIT/ABORT TRANSACTION
```

Figure 4: Transaction template showing different optimizations techniques, for example, setting a partition key hint to start a distribution aware transaction, inode hints to validate the file path components using a batch operation, and partition pruned index scans to read all file inode related metadata.

5.2.1 Lock Phase

In the lock phase, metadata is locked and read from the database with the strongest lock that will be required for the duration of the transaction. Locks are taken in the *total order*, defined earlier. Inode operations are path-based and if they are not read-only operations, they only modify the last component(s) of the path, for example, *rm /etc/conf* and *chmod +x /bin/script*. Thus, only the last component(s) of the file paths are locked for file system operations.

Figure 4 shows a transaction template for HopsFS inode operations. Using the inode hint cache the primary keys for the file path components are discovered, line 1. The transaction is started on the database shard that holds all or most of the desired data, line 2. A batched operation reads all the file path components up to the penultimate path component without locking (*read-committed*) the metadata, line 3. For a path of depth *N*, this removes *N-1* round trips to the database. If the inode hints are invalid then the file path is recursively resolved and the inode hint cache is updated, line 4.

After the path is resolved, either a shared or an exclusive lock is taken on the last inode component in the path, line 5. Shared locks are taken for read-only inode operations, while exclusive locks are taken for inode operations that modify the namespace. Additionally, depending on the operation type and supplied operation parameters, inode related data, such as block, replica, and PRB, etc., are read from the database in a predefined total order using low cost partition pruned scans operations, line 6.

HopsFS uses *hierarchical locking* [20] for inode operations, that is, if data is arranged in tree like hierarchy and all data manipulation operations traverse the hierarchy from top to bottom, then taking a lock on the root of the tree/subtree implicitly locks the children of the tree/subtree. The entity relation diagram for file inode related data, see Figure 3, shows that the entities are arranged in a tree with an inode entity at the root. That is, taking a lock on an inode implicitly locks the tree of file inode related data. As in all operations, inodes are read first, followed by its related metadata. For some operations, such as creating files/directories and listing operations, the parent directory is also locked to prevent *phantom* and *fuzzy* reads for file system operations.

5.2.2 Per-Transaction Cache

All data that is read from the database is stored in a per-transaction cache that withholds the propagation of the updated cache records to the database until the end of the transaction. The cache saves many round trips to the database as the metadata is often read and updated multiple times within the same transaction. Row-level locking of the metadata ensures the consistency of the cache, that is, no other transaction can update the metadata. Moreover, when the locks are released upon the completion of the transaction the cache is cleared.

5.2.3 Execute and Update Phases

The inode operation is performed by processing the metadata in the per-transaction cache. Updated and new metadata generated during the second phase is stored in the cache which is sent to the database in batches in the final *update* phase, after which the transaction is either committed or roll-backed.

6 Handling Large Operations

Recursive operations on large directories, containing millions of inodes, are too large to fit in a single transaction that is, locking millions of rows in a transaction is not supported in online transaction processing systems. These operations include *move*, *rename*, *delete*, *change owner*, *change permissions*, and *set quota* operations. *Move* and *rename* operations change the absolute paths of all the descendant inodes, while *delete* removes all the descendant inodes, and the *set quota* operation affects how all the descendant inodes consume disk space or how many files/directories they can create. Similarly changing the permissions or owner of a directory may invalidate operations executing at the lower subtrees.

6.1 Subtree Operations Protocol

Our solution is a protocol that implements subtree operations incrementally in batches of transactions. Instead of row level database locks, our *subtree operations protocol* uses an application-level distributed locking mechanism to mark and isolate the subtrees. We serialize subtree operations by ensuring that all ongoing inode and subtree operations in a subtree complete before a newly requested subtree operation is executed. We implement this serialization property by enforcing the following invariants: (1) no new operations access the subtree until the operation completes, (2) the subtree is quiesced before the subtree operation starts, (3) no orphaned inodes or inconsistencies

arise if failures occur.

Our subtree operations protocol provides the same consistency semantics as subtree operations in HDFS. For delete subtree operation HopsFS provides even stronger consistency semantics. Failed delete operations in HDFS can result is orphaned blocks that are eventually reclaimed by the block reporting subsystem, which can take hours to detect that the blocks are no longer connected to the namespace. HopsFS improves the semantics of delete operation as failed operations does not cause any metadata inconsistencies, see section 6.2. Subtree operations have the following phases.

Phase 1: In the first phase, an exclusive lock is acquired on the root of the subtree and a *subtree lock* flag (which also contains the Id of the namenode that owns the lock) is set and persisted in the database. The flag is an indication that all the descendants of the subtree are *write* locked.

Before setting the lock it is essential that there are no other *active subtree operations* at any lower level of the subtree. Setting the *subtree lock* could fail active subtree operations executing on a subset of the subtree. We store all active subtree operations in a table and query it to ensure that no subtree operations are executing at lower levels of the subtree. In a typical workload, this table does not grow too large as subtree operations are usually only a tiny fraction of all file system operations. It is important to note that during path resolution, inode and subtree operations that encounter an inode with a subtree lock turned on voluntarily abort the transaction and wait until the *subtree lock* is removed.

Phase 2: To quiesce the subtree we wait for all ongoing inode operations to complete by taking and releasing database write locks on all inodes in the subtree in the same total order used to lock inodes. To do this efficiently, a pool of threads in parallel execute partition pruned index scans that write-lock child inodes. This is repeated down the subtree to the leaves, and, a tree data structure containing the inodes in the subtree is built in memory at the namenode, see Figure 5. The tree is later used by some subtree operations, such as, *rename* and *delete* operations, to process the inodes. We reduce the overhead of reading all inodes in the subtree by using projections to only read the inode Ids.
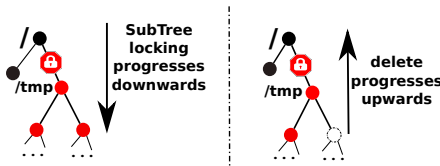

Figure 5: Execution of a delete subtree operation. Parallel transactions progress down (shown left) the subtree waiting for active operations to finish by taking and releasing write locks on all the descendant inodes. In the next phase (shown right), the delete operation is executed in batches using parallel transactions upwards from the leaf nodes.

Phase 3: In the last phase the file system operation is broken down into smaller operations that execute in parallel. For improved performance, large batches of inodes are manipulated in each transaction.

6.2   Handling Failed Subtree Operations

Each namenode maintains a list of the active namenodes in the system provided by the leader election service. If an operation encounters an inode with a subtree lock set and the namenode Id of the *subtree lock* belongs to a dead namenode then the *subtree lock* is reclaimed. However, it is important that when a namenode that is executing a subtree operation fails then it should not leave the subtree in an inconsistent state. The in-memory tree built during the second phase plays an important role in keeping the namespace consistent if the namenode fails. For example, in case of *delete* operations the subtree is deleted incrementally in *post-order* tree traversal manner using transactions. If half way through the operation the namenode fails then the inodes that were not deleted remain connected to the namespace tree. HopsFS clients will transparently resubmit the file system operation to another namenode to delete the remainder of the subtree.

Other subtree operations (*rename*, *move*, *set quota*, *chmod* and *chown*) do not cause any inconsistencies as the actual operation where the metadata is modified is done in the third phase using a single transaction that only updates the root inodes of the subtrees and the inner inodes are left intact. In the case of a failure, the namenode might fail to unset the *subtree lock*, however, this is not a problem as other namenodes can easily reclaim the *subtree lock* when they find out that the *subtree lock* belongs to a dead namenode.

6.3   Inode and Subtree Lock Compatibility

Similar to the inode operation's locking mechanism (see section 5.2.1), subtree operations also implement hierarchical locking, that is, setting a subtree flag on a directory implicitly locks the contents of the directory. Both inode and subtree locking mechanisms are compatible with each other, respecting both of their corresponding locks. That is, a subtree flag cannot be set on a directory locked by an inode operation and an inode operation voluntarily aborts the transaction when it encounters a directory with a subtree lock set.

7   HopsFS Evaluation

As HopsFS addresses how to scale out the metadata layer of HDFS, all our experiments are designed to comparatively test the performance and scalability of the namenode(s) in HDFS and HopsFS in controlled conditions that approximate real-life file system load in big production clusters.

7.1 Experimental Setup

Benchmark: We have extended the benchmarking setup used to test the performance of Quantcast File System (QFS) [43], which is an open source C++ implementation of Google File System. The benchmarking utility is a distributed application that spawns tens of thousands of file system clients, distributed across many machines, which concurrently execute file system (metadata) operations on the namenode(s). The benchmark utility can test the performance of both individual file system operations and file system workloads based on industrial workload traces. HopsFS and the benchmark utility are open source and the experiments described here can be fully reproduced on AWS using Karamel and the cluster definitions found in [22] [25].

HopsFS Setup: All the experiments were run on AWS EC2 using c3.8xlarge instances (32 virtual cores, 60 GB RAM, 2 x 320 GB SSDs) in the EU-West-1 zone. Unless stated otherwise, NDB was deployed on 4 nodes. The NDB datanodes, version 7.4.7, were configured to run using 30 threads and the data replication degree was 2. All other configuration parameters were set to default values.

HDFS Setup: In medium to large Hadoop clusters, 5 to 8 servers are required to provide high availability for HDFS metadata service, see Figure 1 and section section 2. The 5-server setup includes one active namenode, one standby namenode, at least three journal nodes collocated with at least three ZooKeeper nodes. In the 8-server setup, the ZooKeeper nodes are installed on separate servers to prevent multiple services from failing when a server fails. In our experiments Apache HDFS, version 2.4.0 was deployed on five c3.8xlarge instances. Based on Spotify's experience of running HDFS, we configured the HDFS namenodes with 100 client handler threads (*dfs.namenode.handler.count*).

None of the file system clients were co-located with the namenodes or the database nodes. As we are only evaluating metadata performance, all the tests created files of zero length (similar to the NNThroughputBenchmark [64]). Testing with non-empty files requires an order of magnitude of datanodes, and provides no further insight.

7.2 FS Operations' Raw Throughput

In this experiment, for each file system operation, the benchmark utility inundates the namenode(s) with the same file system operation. This test is particularly helpful in determining the maximum throughput and scalability of a particular file system operation. In real deployments, the namenode often receives a deluge of the same file system operations, for example, a big job that reads large amounts of data will generate a huge number of requests to read files and list directories.

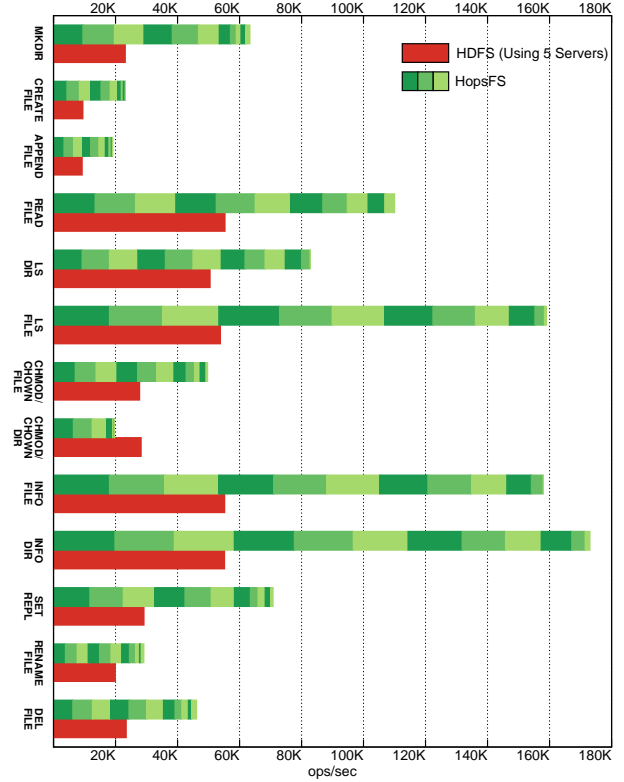Figure 6 shows our results comparing the throughput

Figure 6: HopsFS and HDFS throughput for different operations. For HopsFS each shaded box represents an increase in the throughput of the file system operation when a new namenode is added. For HDFS, the shaded box represents the maximum throughput achieved using the 5-server HDFS namenode setup.

for different file system operations. For each operation, HopsFS' results are displayed as a bar chart of stacked rectangles. Each rectangle represents an increase in the throughput when a new namenode is added. HopsFS outperforms HDFS for almost all file system operations and has significantly better performance than HDFS for the most common file system operations. The only operations where HDFS outperforms HopsFS are *chmod* and *chown*, which change permissions and ownership on a directory subtree, respectively. These operations are slower subtree operations in HopsFS, as the namenode has to make sure that ongoing subtree operations in the subtree are not affected by changing the permission of the root of the subtree (see section 7.4.1 for discussion on the throughput scalability of HopsFS).

In Figure 7, we show the latency for *rename* and *delete* subtree operations on a directory containing a varying number of files, ranging from one quarter to one million files. In HopsFS, large amounts of data are read over the network and the operations are executed in many small transaction batches. The execution time of the rename operation does not increase as rapidly because it does not update the inner nodes or leaves of the subtree. HDFS outperforms HopsFS as all the data is readily available in
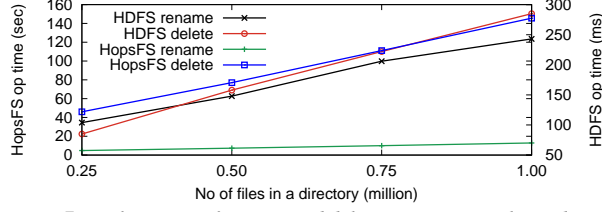
Figure 7: Performance of rename and delete operations on large directories. Note, the different time scales for HDFS and HopsFS. HopsFS is an order of magnitude slower as it reads all the descendant inodes of the subtree from an external database.

the memory. However, due to the low frequency of such operations in typical industrial workloads (see Table 1), we think it is an acceptable trade-off for the higher performance of common file system operations in HopsFS.

7.3 Metadata (Namespace) Scalability

In HDFS, as the entire namespace metadata must fit on the heap of single JVM, the data structures are highly optimized to reduce the memory footprint [62]. In HDFS, a file with two blocks that are replicated three ways requires $448 + L$ bytes of metadata[1] where L represents the filename length. If the file names are 10 characters long, then a 1 GB JVM heap can store 2.3 million files. In reality the JVM heap size has to be significantly larger to accommodate secondary metadata, thousands of concurrent RPC requests, block reports that can each be tens of megabytes in size, as well as other temporary objects.

	Number of Files	
Memory	**HDFS**	**HopsFS**
1 GB	2.3 million	0.44 million
50 GB	115 million	22 million
100 GB	230 million	44 million
200 GB	**460 million**	88 million
500 GB	Does Not Scale	220 million
1 TB	Does Not Scale	440 million
24 TB	Does Not Scale	**10.8 billion**

Table 2: HDFS and HopsFS Metadata Scalability.

Migrating the metadata to a database causes an expansion in the amount of memory required to accommodate indexes, primary/foreign keys and padding. In order to calculate the size of each entity we use a tool called *sizer* [40]. In HopsFS the same file described above takes **2420 bytes** if the metadata is replicated twice. For a highly available deployment with an active and standby namenodes for HDFS, you will need twice the amount of memory, thus, HopsFS requires $\approx$ **2** times more memory than HDFS to store metadata that is highly available.

Table 2 shows the metadata scalability of HDFS and HopsFS. NDB supports up to 48 datanodes, which allows it to scale up to 24 TB of data in a cluster with 512 GB RAM on each NDB datanode. HopsFS can store up to

---

[1]These size estimates are for HDFS version 2.0.4 from which HopsFS was forked. Newer version of HDFS require additional memory for new features such as snapshots and extended attributes.

10.8 billion files using 24 TB of metadata, which is an order of magnitude higher (24 times) than HDFS.

7.4 Industrial Workload Experiments

We benchmarked HopsFS using workloads based on operational traces from Spotify that operates a Hadoop cluster consisting of 1600+ nodes containing 60 petabytes of data. The namespace contains 13 million directories and 218 million files where each file on average contains 1.3 blocks. The Hadoop cluster at Spotify runs on average forty thousand jobs of different applications, such as, *Pig*, *Hive*, *HBase*, *MapReduce*, *Tez*, *Spark*, and *Giraph* every day. The file system workload generated by these application is summarized in Table 1, which shows the relative frequency of HDFS operations. At Spotify the average file path depth is 7 and average inode name length is 34 characters. On average each directory contains 16 files and 2 sub-directories. There are 289 million blocks stored on the datanodes. We use these statistics to generate file system workloads that approximate HDFS usage in production at Spotify.

7.4.1 Scalability Argument

Figure 8 shows that, for our industrial workload, HopsFS delivers 2.6 times the throughput of HDFS with 12 namenodes and 8 NDB nodes. Even for 2-node NDB deployments, HopsFS can outperform HDFS and scale linearly up to 8 namenodes.

As discussed before in medium to large Hadoop clusters **5 to 8 servers** are required to provide high availability for HDFS. With **6 servers**, HopsFS delivers higher throughput than HDFS, increasing further as more namenodes are added to the system.

For higher numbers of namenodes, HopsFS' throughput levels off because NDB becomes overloaded. By increasing the number of NDB datanode instances to 4, HopsFS increases throughput up to 11 namenodes, and by increasing the number of NDB datanode instances to 8, HopsFS scales up to at least 12 namenodes. The 2-node NDB cluster has a similar performance as the 4-node NDB cluster because each NDB datanode in 2-node NDB cluster holds the complete copy of the entire metadata, which improves the performance of transactions as all the
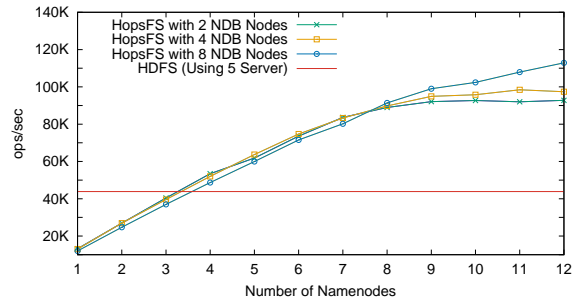

Figure 8: HopsFS and HDFS throughput for our industrial workload.

read operations in the transactions are handled locally. For 1 to 7 namenodes, a 2-node NDB cluster slightly outperforms an 8-node NDB cluster, as we have some index scans that are slower to execute on larger clusters. HopsFS still has further scope for throughput improvements, by testing with even larger clusters and better tuning of NDB. Bare-metal clusters should perform even better, by locking NDB threads to CPU cores and isolating them from handling interrupts [10, 36], which was not possible in our experimental setup due to the virtualized environment.

7.5 Operational Latency

The latency for a single file system operation on an unloaded HDFS namenode will always be lower than in HopsFS, as all the metadata is readily available in main memory for the HDFS namenode, while it is remote for the namenodes in HopsFS. In these experiments, we ran a 12 namenode cluster with 4 NDB datanodes. Figure 9 shows average file system operation latency observed by a client while running the industrial workload. When the number of clients is low, operations in HDFS have lower latency (5ms vs 2.5ms). However, large HDFS deployments may have tens of thousands of clients [63]. For a large number of clients, the end-to-end latency observed by clients increases as the file system operations wait in RPC call queues at the namenode [56]. HopsFS can handle more concurrent clients and operations have lower latency when the number of clients increases beyond 400.
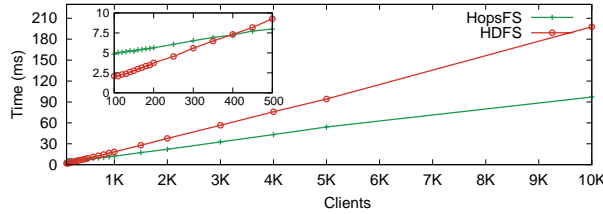
Figure 9: Average operational latency observed by HopsFS and HDFS clients.

Figure 10 shows 99th percentile latencies for different file system operations in a non-overloaded cluster. In this experiment, we ran HopsFS and HDFS under 50% load for the Spotify workload (50 % of the maximum throughput observed in figure 8). We ran 30 clients per namenode. In HopsFS, 99th-percentiles for common file system operations such as touch file, read file, ls dir and stat dir are 76 ms, 12 ms, 13 ms and 9 ms, respectively. In a similar experiment for HDFS, running at 50% load, the 99th-percentile latency for touch file is 6 ms, while other operations had 1 ms 99th-percentile latencies.

A common complaint against the two-phase commit protocol is that it is blocking and failure of a transaction coordinator will cause the system to block. NDB supports very low transaction inactive timeouts. In the experiments, the transaction inactive timeout was set to NDB default of 1.2 seconds, and in the event of a database node
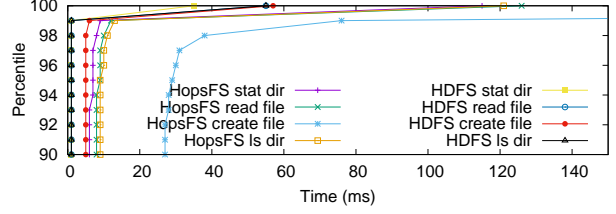
Figure 10: HopsFS and HDFS latency for common file system operations operating at 50% load.

failure, failed transactions are automatically retried by a namenode on a different database node.

7.6 Failure Handling

Now we discuss how the performance of the HDFS and HopsFS is affected when the namenodes, NDB datanodes, and journal nodes fail.

7.6.1 Namenodes failure

Figure 11 shows how the performance of the file system metadata service is affected when a namenode fails. The namenodes failures were simulated by killing and restarting all the file system processes on a namenode. For HDFS, the active namenode was periodically killed while for HopsFS, the namenodes were periodically killed in a round-robin manner. In the Figure 11, vertical lines indicate namenode failures. In HDFS, the standby namenode takes over when it detects that the active namenode has failed. In our experiments we have observed 8 to 10 seconds of downtime during failover in HDFS. During this time no file system metadata operation can be performed. Our failover tests were favorable to Apache HDFS, as the amount of metadata stored by NNs in the experiment is minimal. At Spotify, with 140 gigabytes of metadata and 40 thousand jobs every day, failover takes at least 5 minutes, and often up to 15 minutes. Although we are unsure of the reason why, it may be due to the additional checkpointing role played by the Standby namenode. In contrast, in HopsFS when a namenode fails clients transparently re-execute failed file system operations on one of the remaining namenodes in the system.
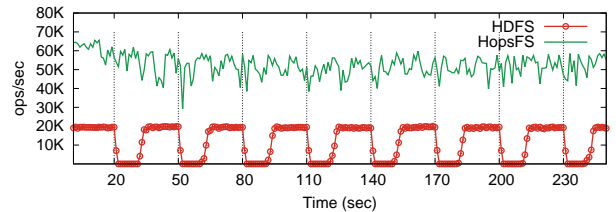
Figure 11: HopsFS and HDFS namenode failover. Vertical lines represent namenodes failures.

7.6.2 Failure of NDB Datanodes or Journal Nodes

For a HDFS cluster with N journal nodes, HDFS can tolerate failure of up to $\lceil N/2 \rceil - 1$ journal nodes. In our tests with a quorum of three journal nodes, HDFS can tolerate only one journal node failure. We have tested HDFS with

3, 5, and 7 journal nodes, and the namenode is not affected when the journal nodes fail provided that the quorum is not lost. When more journal nodes fail and the quorum is lost then the HDFS namenodes shutdown.

The number of NDB nodes failure that HopsFS can tolerate depends on the number of NDB datanodes and the replication degree of NDB. With a default NDB replication degree of 2, a four node NDB cluster can tolerate up to two NDB datanodes failures and an eight node NDB cluster can tolerate up to four NDB datanodes failure. We have tested HopsFS on 2, 4, and 8 node NDB clusters and the performance of HopsFS is not affected when a NDB datanode fails as long as there is at least one remaining NDB datanode alive in each NDB node group. If all the NDB datanodes in a node group fail, then the HopsFS namenodes shutdown.

7.7 Effect of DB Optimization Techniques

In this experiment, we focus on three optimizations; (1) distribution aware transactions (DAT), section 4.2, (2) application defined partitioning (ADP), section 4.2, and (3) the inode hint cache at the namenodes, section 5.1. We ran file creation and file read experiments on a 12 namenode HopsFS cluster while varying the file depth, as shown in Figures 12 and 13.

File system operations in HopsFS are implemented as one or more transactions, where each transaction performs multiple round trips to the database to read and update the metadata. In Table 3, we show the number of round trips needed for each operation in the case of (1) no inode hint cache and (2) inode hint cache hits. For example, if we create an empty file */d1/d2/.../d9/f* at depth $N = 10$, then in the case of no caching, we substitute in the equation from Table 3 and derive the cost of the operation as $17PK_{rc} + 5PK_w + 2PPIS + 2B_1$, that is 26 round trips to the database. In contrast, in the case of cache hits for all inodes hints, we will have 11 round trips to the database, irrespective of the depth of the file in the namespace hierarchy. Cache hits save 15 round trips, giving an expected 58% performance improvement for the file create operation. We validated our analysis of the file system operations by running experiments for both cases, that is, with the cache enabled *(HopsFS-DAT+ADP+Cache)* and with the cache disabled *(HopsFS-DAT+ADP)*, as shown in Figure 13. Our experiments show that at depth $N = 10$ the inode cache improves HopsFS' throughput for file creation by $\approx 58\%$. Similarly, in the case of file read, the cache improves HopsFS by $\approx 68\%$, see Figure 12.

7.8 Block Report Performance

In HDFS, each datanode periodically sends a block report to a namenode, containing Ids of all the blocks stored on the datanode. Block reports serve two purposes: (1) they help to rebuild the block location map when the namenode restarts since HDFS does not persist this informa-
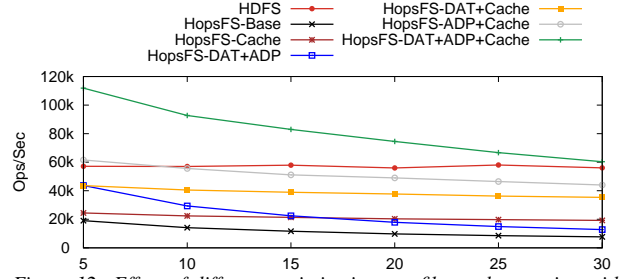

Figure 12: Effect of different optimizations on file read operation with respect to depth of the file in the namespace hierarchy
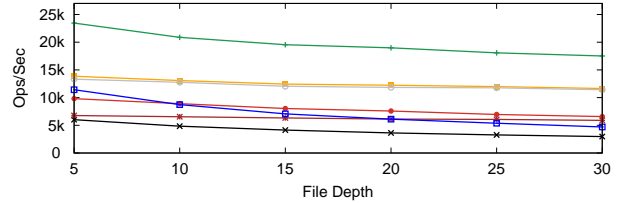

Figure 13: Effect of different optimizations on file create operation with respect to depth of the file in the namespace hierarchy.

tion, (2) they serve as ground truth for available blocks in the system. We reimplemented the HDFS block reporting solution in HopsFS. Although the solution is fully functional it does not deliver as high throughput because a large amount of metadata is read over the network from the database by the namenodes to process a block report.

In an experiment with the same setup, 50 datanodes simultaneously submitted block report containing 100,000 blocks. With 12 namenodes, HopsFS manages to process 10 block reports per second while HDFS managed to process 60 block reports per second. However, full block-reports aren't needed as frequently in HopsFS as in HDFS, as we persist the block location mappings in the database. Even without further optimizations, with a 512 megabyte block size, and datanodes sending block reports every six hours, HopsFS can scale to handle block reporting in an exabyte cluster.

8 Related Work

The InversionFS [41] and Windows Future Storage (WinFS) [74] were some of the first monolithic file systems that stored the metadata in a relational database. Gunawi [21] showed that some file system operations, such as *fsck*, can be more efficient when implemented using a relational database.

Recently, high performance distributed databases such as HBase [17, 9], Cassandra [30], CalvinDB [71] have enabled the development of new distributed metadata management architectures in file systems such as CalvinFS [70], CassandraFS [8] and GiraffaFS [19]. All of these file systems store denormalized metadata, that is, they store the full file path with each inode which affects the subtree operations. GiraffaFS only supports file renames in the same directory. CalvinFS relies on

FS Ops	Without Inode Hint Cache at the NameNodes	Inode Hint Cache Hits
mkdirs(d_N)	$(N-2)PK_{rc} + 2PK_w + B_1$	$2PK_w + (2B)_{N-1}$
addBlk(f_N)	$(N-1)PK_{rc} + PK_w + PK_r + B_1 + (f_s == 0\ ?\ 2PPIS\ :\ 6PPIS)$	$PK_w + PK_r + (2B)_N + (f_s == 0\ ?\ 2PPIS\ :\ 6PPIS)$
create(f_N)	$(2N-3)PK_{rc} + 5PK_w + 2B_1 + (f_s == 0\ ?\ 2PPIS\ :\ 8PPIS) + \sum_{b=b_0}^{b_x} \text{addBlk}(f_N)$	$5PK_w + (4B)_{2N-1} + (f_s == 0\ ?\ 2PPIS\ :\ 8PPIS) + \sum_{b=b_0}^{b_x} \text{addBlk}(f_N)$
read(f_N)	$(N-1)PK_{rc} + PK_r + B_1 + (f_s == 0\ ?\ PPIS\ :\ 5PPIS)$	$PK_r + (2B)_N + (f_s == 0\ ?\ PPIS\ :\ 5PPIS)$
ls(i_N)	$(N-1)PK_{rc} + PK_r + (i\ \text{is a dir}\ ?\ PPIS\ :\ 0)$	$PK_r + B_{N-1} + (i\ \text{is a dir}\ ?\ PPIS\ :\ 0)$
chmod(i_N)	$(2N-2)PK_{rc} + 2PK_w + 2B_1 + (i\ \text{is a dir}\ ?\ IS\ :\ PPIS)$	$2PK_w + (4B)_{2N} + (i\ \text{is a dir}\ ?\ IS\ :\ PPIS)$
info(i_N)	$(N-1)PK_{rc} + PK_r + B_1$	$PK_r + (2B)_N$
del(f_N)	$(2N-2)PK_{rc} + 2PK_w + 2B_1 + (f_s == 0\ ?\ 2PPIS\ :\ 7PPIS)$	$2PK_w + (4B)_{2N} + (f_s == 0\ ?\ 2PPIS\ :\ 7PPIS)$
delTree(i_N)	$(N-2)PK_{rc} + 2PK_w + B_1 + (i_s == 0\ \mid\ i\ \text{is a dir}\ ?\ PPIS\ :\ 7PPIS)$	$2PK_w + (2B)_{N-1} + (i_s == 0\ \mid\ i\ \text{is a dir}\ ?\ PPIS\ :\ 7PPIS)$
del(d_N)	$(3N-1)PK_{rc} + 2PK_w + 3B_1 + IS + \sum_{i=f_1}^{f_x} \text{delTree}(i_N)$	$2PK_w + (3B)_{3N+1} + IS + \sum_{i=f_1}^{f_x} \text{delTree}(i_N)$
mv(f_N,f_M)	$(2M+2N-S-4)PK_{rc} + (3B)_4 + (f_s == 0\ ?\ 2PPIS\ :\ 9PPIS) + (S == M-1\ \&\ M == N\ ?\ 3PK_w\ :\ 4PK_w)$	$(7B)_{2M+2N-S} + (f_s == 0\ ?\ 2PPIS\ :\ 9PPIS) + (S == M-1\ \&\ M == N\ ?\ 3PK_w\ :\ 4PK_w)$
mv(d_N,d_M)	$(2M+3N-S-5)PK_{rc} + 2IS + (4B)_5 + (S == M-1\ \&\ M == N\ ?\ 4PK_w\ :\ 5PK_w)$	$2IS + (8B)_{2M+3N-S} + (S == M-1\ \&\ M == N\ ?\ 4PK_w\ :\ 5PK_w)$

f_N, d_N, i_N :	f=file, d=dir, i=file/dir path at depth N	f_s :	size of a file f
PK_w, PK_{rc} :	a primary key read operation with a <u>w</u>rite or a read-<u>c</u>ommitted lock (i.e., no lock)	$PPIS$:	a partition pruned index scan
xB_N :	x primary key batch operations that read in total N rows from the database	FTS :	Full table scan without using indexes (hits all shards)
S :	number of shared inodes in the paths between (f_N and f_M) or (d_N and d_M)	IS :	Index scan (hits all shards)

Table 3: The cost of file system (metadata) operations in terms of database accesses for (1) no inode hint caching and (2) cache hits.

CalvinDB to perform large transactions. CalvinDB runs large transactions in two phases. In the first phase the lock set is identified, and in the second phase all the locks are acquired and the operation is performed, provided that the lock set has not changed. However, CalvinFS did not experimentally show that this is a viable technique for performing operations on a directory with millions of files. Production-grade online transaction processing systems have an upper bound on the number of operations that can be included in a transaction, where the upper bound is lower than tens of millions.

IndexFS [53] and ShardFS [75] are file systems optimized for metadata workloads with a large number of small files. Their metadata servers handle metadata as well as user data for small files stored in local LevelDB [33] instances, and delegate the management of large files to an underlying distributed file system, for example, HDFS [63], Lustre [68], PanFS [73] and PVFS [32]. IndexFS caches inode information at clients, while ShardFS caches it at metadata servers.

PVFS2 [32], OrangeFS [76], Farsite [15], Lustre [68], Vesta [12], InterMezzo [47], zFS [54], and RAMA [37] shard inodes among multiple metadata servers by either (1) random partitioning or (2) partitioning based on hashed file identifiers or hashed full/partial file paths. This partitioning scheme is typically combined with the caching of metadata at clients, which can cause cache invalidation storms for large subtree operations.

Finally, our architecture supports a pluggable NewSQL storage engine. MemSQL is a candidate, as it supports high throughput cross-partition transactions, application defined partitioning, and partition pruned queries [35]. VoltDB is currently not a candidate as it serializes cross partition transactions [72].

9 External Metadata Implications

Administrators often resort to writing their own tools to analyze the HDFS namespace. HopsFS enables online ad hoc analytics on the metadata. With a NDB backend, HopsFS metadata can be selectively and asynchronously replicated to either a backup cluster or a MySQL slave server, enabling complex analytics without affecting the performance of the active cluster. HopsFS metadata is also easy to export to external systems and it is easy to safely extend the metadata. That is, additional tables can be created that contain a foreign key to the associated inode, thus ensuring the integrity of the extended metadata. Using this approach, we have already added new features to HopsFS, including extended attributes for inodes and erasure coding. In another project, we developed an eventually consistent replication protocol that replicates (extended) HopsFS metadata to Elasticsearch [16] for free-text search. This enables us to search the entire namespace with sub-second latency. We believe that distributed metadata in a commodity database is a significant new enabling technology and a reliable source of ground truth for metadata applications built on top of distributed file systems.

10 Summary

In this paper, we introduced HopsFS, an open-source, highly available file system that scales out in both capacity and throughput by adding new namenodes and database nodes. HopsFS can store 24 times more metadata than HDFS and for a workload from Spotify, HopsFS scales to handle 2.6 times the throughput of HDFS. HopsFS also has lower latency than HDFS for greater than 400 concurrent clients, and no downtime during failover. Our architecture supports a pluggable database storage engine, and other NewSQL databases could be used, providing they have good support for cross-partition transactions and distribution-aware techniques such as partition pruned index scans. Finally, HopsFS makes metadata tinker friendly, opening it up for users and applications to extend and analyze in new and creative ways.

References

[1] C. L. Abad. *Big Data Storage Workload Characterization, Modeling and Synthetic Generation*. PhD thesis, University of Illinois at Urbana-Champaign, 2014.

[2] M. Abd-El-Malek, W. V. Courtright, II, C. Cranor, G. R. Ganger, J. Hendricks, A. J. Klosterman, M. Mesnier, M. Prasad, B. Salmon, R. R. Sambasivan, S. Sinnamohideen, J. D. Strunk, E. Thereska, M. Wachs, and J. J. Wylie. Ursa Minor: Versatile Cluster-based Storage. In *Proceedings of the 4th Conference on USENIX Conference on File and Storage Technologies - Volume 4*, FAST'05, pages 5–5, Berkeley, CA, USA, 2005. USENIX Association.

[3] A. Adya, W. J. Bolosky, M. Castro, G. Cermak, R. Chaiken, J. R. Douceur, J. Howell, J. R. Lorch, M. Theimer, and R. P. Wattenhofer. Farsite: Federated, Available, and Reliable Storage for an Incompletely Trusted Environment. *SIGOPS Oper. Syst. Rev.*, 36(SI):1–14, Dec. 2002.

[4] R. Agrawal, M. J. Carey, and M. Livny. Concurrency control performance modeling: Alternatives and implications. *ACM Trans. Database Syst.*, 12(4):609–654, Nov. 1987.

[5] A. Alexandrov, R. Bergmann, S. Ewen, J.-C. Freytag, F. Hueske, A. Heise, O. Kao, M. Leich, U. Leser, V. Markl, F. Naumann, M. Peters, A. Rheinländer, M. J. Sax, S. Schelter, M. Höger, K. Tzoumas, and D. Warneke. The Stratosphere Platform for Big Data Analytics. *The VLDB Journal*, 23(6):939–964, Dec. 2014.

[6] J. Baker, C. Bond, J. C. Corbett, J. Furman, A. Khorlin, J. Larson, J.-M. Leon, Y. Li, A. Lloyd, and V. Yushprakh. Megastore: Providing scalable, highly available storage for interactive services. In *Proceedings of the Conference on Innovative Data system Research (CIDR)*, pages 223–234, 2011.

[7] H. Berenson, P. Bernstein, J. Gray, J. Melton, E. O'Neil, and P. O'Neil. A critique of ansi sql isolation levels. *SIGMOD Rec.*, 24(2):1–10, May 1995.

[8] Cassandra File System Design. http://www.datastax.com/dev/blog/cassandra-file-system-design. [Online; accessed 1-January-2016].

[9] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber. Bigtable: A Distributed Storage System for Structured Data. *ACM Trans. Comput. Syst.*, 26(2), 2008.

[10] Configuring MySQL Cluster Data Nodes . https://blogs.oracle.com/MySQL/entry/configuring_mysql_cluster_data_nodes. [Online; accessed 1-January-2016].

[11] J. C. Corbett, J. Dean, M. Epstein, A. Fikes, C. Frost, J. J. Furman, S. Ghemawat, A. Gubarev, C. Heiser, P. Hochschild, W. Hsieh, S. Kanthak, E. Kogan, H. Li, A. Lloyd, S. Melnik, D. Mwaura, D. Nagle, S. Quinlan, R. Rao, L. Rolig, Y. Saito, M. Szymaniak, C. Taylor, R. Wang, and D. Woodford. Spanner: Google's Globally-distributed Database. In *Proceedings of the 10th USENIX Conference on Operating Systems Design and Implementation*, OSDI'12, pages 251–264, Berkeley, CA, USA, 2012. USENIX Association.

[12] P. Corbett, D. Feitelson, J.-P. Prost, and S. Baylor. Parallel access to files in the Vesta filesystem. In *Supercomputing '93. Proceedings*, pages 472–481, Nov 1993.

[13] E. Corporation. HADOOP IN THE LIFE SCIENCES:An Introduction. https://www.emc.com/collateral/software/white-papers/h10574-wp-isilon-hadoop-in-lifesci.pdf, 2012. [Online; accessed 30-Aug-2015].

[14] Dean, Jeffrey and Ghemawat, Sanjay. Mapreduce: Simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, Jan. 2008.

[15] J. R. Douceur and J. Howell. Distributed Directory Service in the Farsite File System. In *Proceedings of the 7th Symposium on Operating Systems Design and Implementation*, OSDI '06, pages 321–334, Berkeley, CA, USA, 2006. USENIX Association.

[16] Elasticsearch. https://www.elastic.co/products/elasticsearch. [Online; accessed 1-January-2016].

[17] L. George. *HBase: The Definitive Guide*. Definitive Guide Series. O'Reilly Media, Incorporated, 2011.

[18] S. Ghemawat, H. Gobioff, and S.-T. Leung. The Google File System. *SIGOPS Oper. Syst. Rev.*, 37(5):29–43, Oct. 2003.

[19] GiraffaFS. https://github.com/giraffafs/giraffa. [Online; accessed 1-January-2016].

[20] J. Gray, R. Lorie, G. Putzolu, and I. Traiger. Granularity of Locks and Degrees of Consistency in a Shared Database. In *IFIP Working Conference on Modelling in Data Base Management Systems*, pages 365–394. IFIP, 1976.

[21] H. S. Gunawi, A. Rajimwale, A. C. Arpaci-Dusseau, and R. H. Arpaci-Dusseau. SQCK: A Declarative File System Checker. In *Proc. of OSDI'08*, pages 131–146. USENIX Association, 2008.

[22] Hammer-Bench: Distributed Metadata Benchmark to HDFS. https://github.com/smkniazi/hammer-bench. [Online; accessed 1-January-2016].

[23] Hadoop JIRA: Add thread which detects JVM pauses. https://issues.apache.org/jira/browse/HADOOP-9618. [Online; accessed 1-January-2016].

[24] M. P. Herlihy and J. M. Wing. Linearizability: A correctness condition for concurrent objects. *ACM Trans. Program. Lang. Syst.*, 12(3):463–492, July 1990.

[25] Hadoop Open Platform-as-a-Service (Hops) is a new distribution of Apache Hadoop with scalable, highly available, customizable metadata. https://github.com/smkniazi/hammer-bench. [Online; accessed 1-January-2016].

[26] P. Hunt, M. Konar, F. P. Junqueira, and B. Reed. ZooKeeper: Wait-free Coordination for Internet-scale Systems. In *Proceedings of the 2010 USENIX Conference on USENIX Annual Technical Conference*, USENIXATC'10, pages 11–11, 2010.

[27] F. Hupfeld, T. Cortes, B. Kolbeck, J. Stender, E. Focht, M. Hess, J. Malo, J. Marti, and E. Cesario. The XtreemFS architecture—a case for object-based file systems in Grids. *Concurrency and computation: Practice and experience*, 20(17):2049–2060, 2008.

[28] M. Isard, M. Budiu, Y. Yu, A. Birrell, and D. Fetterly. Dryad: Distributed data-parallel programs from sequential building blocks. In *Proceedings of the 2Nd ACM SIGOPS/EuroSys European Conference on Computer Systems 2007*, EuroSys '07, pages 59–72, New York, NY, USA, 2007. ACM.

[29] Karamel: One Click Installation For Clusters. http://www.karamel.io/this. [Online; accessed 1-January-2016].

[30] A. Lakshman and P. Malik. Cassandra: A Decentralized Structured Storage System. *SIGOPS Oper. Syst. Rev.*, 44(2):35–40, Apr. 2010.

[31] B. W. Lampson. Hints for computer system design. *SIGOPS Oper. Syst. Rev.*, 17(5):33–48, Oct. 1983.

[32] R. Latham, N. Miller, R. B. Ross, and P. H. Carns. A Next-Generation Parallel File System for Linux Clusters. *LinuxWorld Magazine*, January 2004.

[33] LevelDB. http://leveldb.org/. [Online; accessed 1-January-2016].

[34] E. Levy and A. Silberschatz. Distributed file systems: Concepts and examples. *ACM Computing Surveys*, 22:321–374, 1990.

[35] MemSQL, The World's Fastest Database, In Memory Database, Column Store Database. http://www.memsql.com/. [Online; accessed 30-June-2015].

[36] Mikael Ronstrom. Scalability enhancements of MySQL Cluster 7.2. `http://mikaelronstrom.blogspot.se/2012/02/scalability-enhancements-of-mysql.html`. [Online; accessed 1-January-2016].

[37] E. L. Miller and R. Katz. RAMA: An Easy-To-Use, High-Performance Parallel File System. *Parallel Computing*, 23(4):419–446, July 1997.

[38] J. H. Morris, M. Satyanarayanan, M. H. Conner, J. H. Howard, D. S. Rosenthal, and F. D. Smith. Andrew: A Distributed Personal Computing Environment. *Commun. ACM*, 29(3):184–201, Mar. 1986.

[39] MySQL Cluster CGE. `http://www.mysql.com/products/cluster/`. [Online; accessed 30-June-2015].

[40] S. Nines. sizer - capacity planning tool. `http://www.severalnines.com/sizer-capacity-planning-tool`. [Online; accessed 30-Aug-2015].

[41] M. A. Olson and M. A. The Design and Implementation of the Inversion File System, 1993.

[42] J. K. Ousterhout, A. R. Cherenson, F. Douglis, M. N. Nelson, and B. B. Welch. The sprite network operating system. *IEEE Computer*, 21:23–36, 1988.

[43] M. Ovsiannikov, S. Rus, D. Reeves, P. Sutter, S. Rao, and J. Kelly. The Quantcast File System. *Proc. VLDB Endow.*, 6(11):1092–1101, Aug. 2013.

[44] F. Özcan, N. Tatbul, D. J. Abadi, M. Kornacker, C. Mohan, K. Ramasamy, and J. Wiener. Are We Experiencing a Big Data Bubble? In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, pages 1407–1408, 2014.

[45] S. V. Patil, G. A. Gibson, S. Lang, and M. Polte. GIGA+: Scalable Directories for Shared File Systems. In *Proceedings of the 2Nd International Workshop on Petascale Data Storage: Held in Conjunction with Supercomputing '07*, PDSW '07, pages 26–29, New York, NY, USA, 2007. ACM.

[46] B. Pawlowski, C. Juszczak, P. Staubach, C. Smith, D. Lebel, and D. Hitz. NFS Version 3 - Design and Implementation. In *In Proceedings of the Summer USENIX Conference*, pages 137–152, 1994.

[47] Peter Braam Braam and Michael Callahan. The InterMezzo File System. In *In Proceedings of the 3rd of the Perl Conference, O'Reilly Open Source Convention*, Monterey, CA, USA, 1999.

[48] A. J. Peters and L. Janyst. Exabyte Scale Storage at CERN. *Journal of Physics: Conference Series*, 331(5):052015, 2011.

[49] I. Polato, R. Ré, A. Goldman, and F. Kon. A comprehensive view of Hadoop research – A systematic literature review. *Journal of Network and Computer Applications*, 46:1–25, 2014.

[50] G. Popek and B. J. Walker. *The LOCUS distributed system architecture*. MIT Press, 1985.

[51] K. W. Preslan, A. Barry, J. Brassow, R. Cattelan, A. Manthei, E. Nygaard, S. V. Oort, D. Teigland, M. Tilstra, and et al. Implementing Journaling in a Linux Shared Disk File System, 2000.

[52] K. Ren, Y. Kwon, M. Balazinska, and B. Howe. Hadoop's adolescence: an analysis of hadoop usage in scientific workloads. *Proceedings of the VLDB Endowment*, 6(10):853–864, 2013.

[53] K. Ren, Q. Zheng, S. Patil, and G. Gibson. IndexFS: Scaling File System Metadata Performance with Stateless Caching and Bulk Insertion. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '14, pages 237–248, Piscataway, NJ, USA, 2014. IEEE Press.

[54] O. Rodeh and A. Teperman. zFS - a scalable distributed file system using object disks. In *Mass Storage Systems and Technologies, 2003. (MSST 2003). Proceedings. 20th IEEE/11th NASA Goddard Conference on*, pages 207–218, April 2003.

[55] M. Ronström and J. Oreland. Recovery Principles of MySQL Cluster 5.1. In *Proc. of VLDB'05*, pages 1108–1115. VLDB Endowment, 2005.

[56] RPC Congestion Control with FairCallQueue. `https://issues.apache.org/jira/browse/HADOOP-9640`. [Online; accessed 1-January-2016].

[57] G. B. Salman Niazi, Mahmoud Ismail and J. Dowling. Leader Election using NewSQL Systems. In *Proc. of DAIS 2015*, pages 158 –172. Springer, 2015.

[58] M. Satyanarayanan, J. J. Kistler, P. Kumar, M. E. Okasaki, E. H. Siegel, David, and C. Steere. Coda: A Highly available File System for a Distributed Workstation Environment. *IEEE Transactions on Computers*, 39:447–459, 1990.

[59] Schmuck, Frank and Haskin, Roger. GPFS: A Shared-Disk File System for Large Computing Clusters. In *Proceedings of the 1st USENIX Conference on File and Storage Technologies*, FAST '02, Berkeley, CA, USA, 2002. USENIX Association.

[60] P. Schwan. Lustre: Building a File System for 1000-node Clusters. In *Proc. of OLS'03*, 2003.

[61] M. Seltzer and N. Murphy. Hierarchical File Systems Are Dead. In *Proceedings of the 12th Conference on Hot Topics in Operating Systems*, HotOS'09, pages 1–1, Berkeley, CA, USA, 2009. USENIX Association.

[62] K. Shvachko. Name-node memory size estimates and optimization proposal. `https://issues.apache.org/jira/browse/HADOOP-1687`, August 2007. [Online; accessed 11-Nov-2014].

[63] K. Shvachko, H. Kuang, S. Radia, and R. Chansler. The hadoop distributed file system. In *Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, MSST '10, pages 1–10, Washington, DC, USA, 2010. IEEE Computer Society.

[64] K. V. Shvachko. HDFS Scalability: The Limits to Growth. *login: The Magazine of USENIX*, 35(2):6–16, Apr. 2010.

[65] HOPS, Software-As-A-Service from SICS'S new datacenter. `https://www.swedishict.se/hops-software-as-a-service-from-sicss-new-datacenter`. [Online; accessed 23-May-2016].

[66] M. Srivas, P. Ravindra, U. Saradhi, A. Pande, C. Sanapala, L. Renu, S. Kavacheri, A. Hadke, and V. Vellanki. Map-Reduce Ready Distributed File System, 2011. US Patent App. 13/162,439.

[67] The Curse of the Singletons! The Vertical Scalability of Hadoop NameNode. `http://hadoopblog.blogspot.se/2010/04/curse-of-singletons-vertical.html`. [Online; accessed 30-Aug-2015].

[68] The Lustre Storage Architecture. `http://wiki.lustre.org/manual/LustreManual20_HTML/UnderstandingLustre.html`. [Online; accessed 30-Aug-2015].

[69] C. A. Thekkath, T. Mann, and E. K. Lee. Frangipani: A Scalable Distributed File System. In *Proceedings of the Sixteenth ACM Symposium on Operating Systems Principles*, SOSP '97, pages 224–237, New York, NY, USA, 1997. ACM.

[70] A. Thomson and D. J. Abadi. CalvinFS: Consistent WAN Replication and Scalable Metadata Management for Distributed File Systems. In *13th USENIX Conference on File and Storage Technologies (FAST 15)*, pages 1–14, Santa Clara, CA, Feb. 2015. USENIX Association.

[71] A. Thomson, T. Diamond, S.-C. Weng, K. Ren, P. Shao, and D. J. Abadi. Calvin: Fast Distributed Transactions for Partitioned Database Systems. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, pages 1–12, New York, NY, USA, 2012. ACM.

[72] VoltDB Documentation. `http://docs.voltdb.com/ReleaseNotes/`. [Online; accessed 30-June-2015].

[73] Welch, Brent and Unangst, Marc and Abbasi, Zainul and Gibson, Garth and Mueller, Brian and Small, Jason and Zelenka, Jim and Zhou, Bin. Scalable Performance of the Panasas Parallel File System. In *Proceedings of the 6th USENIX Conference on File and Storage Technologies*, FAST'08, Berkeley, CA, USA, 2008. USENIX Association.

[74] WinFS: Windows Future Storage. `https://en.wikipedia.org/wiki/WinFS`. [Online; accessed 30-June-2015].

[75] L. Xiao, K. Ren, Q. Zheng, and G. A. Gibson. ShardFS vs. IndexFS: Replication vs. Caching Strategies for Distributed Metadata Management in Cloud Storage Systems. In *Proceedings of the Sixth ACM Symposium on Cloud Computing*, SoCC '15, pages 236–249, New York, NY, USA, 2015. ACM.

[76] S. Yang, W. B. Ligon III, and E. C. Quarles. Scalable distributed directory implementation on orange file system. *Proc. IEEE Intl. Wrkshp. Storage Network Architecture and Parallel I/Os (SNAPI)*, 2011.

[77] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica. Spark: Cluster computing with working sets. In *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing*, HotCloud'10, pages 10–10, Berkeley, CA, USA, 2010. USENIX Association.

[78] M. Zait and B. Dageville. Method and mechanism for database partitioning, Aug. 16 2005. US Patent 6,931,390.