In *Proceedings of the 9th USENIX Conference on File and Storage Technologies (FAST '11), San Jose CA, February 2011.*

Scale and Concurrency of GIGA+: File System Directories with Millions of Files

Swapnil Patil and Garth Gibson
Carnegie Mellon University
{firstname.lastname @ cs.cmu.edu}

Abstract – We examine the problem of scalable file system directories, motivated by data-intensive applications requiring millions to billions of small files to be ingested in a single directory at rates of hundreds of thousands of file creates every second. We introduce a POSIX-compliant scalable directory design, GIGA+, that distributes directory entries over a cluster of server nodes. For scalability, each server makes only local, independent decisions about migration for load balancing. GIGA+ uses two internal implementation tenets, asynchrony and eventual consistency, to: (1) partition an index among all servers without synchronization or serialization, and (2) gracefully tolerate stale index state at the clients. Applications, however, are provided traditional strong synchronous consistency semantics. We have built and demonstrated that the GIGA+ approach scales better than existing distributed directory implementations, delivers a sustained throughput of more than 98,000 file creates per second on a 32-server cluster, and balances load more efficiently than consistent hashing.

1 Introduction

Modern file systems deliver scalable performance for large files, but not for large numbers of files [Wheeler 2010; Fikes 2010]. In particular, they lack scalable support for ingesting millions to billions of small files in a single directory - a growing use case for data-intensive applications [Fikes 2010; Ross 2006; Newman 2008]. We present a file system directory service, GIGA+, that uses highly concurrent and decentralized hash-based indexing, and that scales to store at least millions of files in a single POSIX-compliant directory and sustain hundreds of thousands of create insertions per second.

The key feature of the GIGA+ approach is to enable higher concurrency for index mutations (particularly creates) by eliminating system-wide serialization and synchronization. GIGA+ realizes this principle by aggressively distributing large, mutating directories over a cluster of server nodes, by disabling directory entry caching in clients, and by allowing each node to migrate, without notification or synchronization, portions of the directory for load balancing. Like traditional hash-based distributed indices [Litwin 1996; Fagin 1979; Schmuck 2002], GIGA+ incrementally hashes a directory into a growing number of partitions. However, GIGA+ tries harder to eliminate synchronization and prohibits migration if load balancing is unlikely to be improved.

Clients do not cache directory entries; they cache only the directory index. This cached index can have stale pointers to servers that no longer manage specific ranges in the space of the hashed directory entries (filenames). Clients using stale index values to target an incorrect server have their cached index corrected by the incorrectly targeted server. Stale client indices are aggressively improved by transmitting the history of splits of all partitions known to a server. Even the addition of new servers is supported with minimal migration of directory entries and delayed notification to clients. In addition, because 99.99% of the directories have less than 8,000 entries [Dayal 2008; Agrawal 2007], GIGA+ represents small directories in one partition so most directories will be essentially like traditional directories.

Since modern cluster file systems have support for data striping and failure recovery, our goal is not to compete with all features of these systems, but to offer additional technology to support high rates of mutation of many small files.[1] We have built a skeleton cluster file system with GIGA+ directories that layers on existing lower layer file systems using FUSE [FUSE]. Unlike the current trend of using special purpose storage systems with custom interfaces and semantics [Shvachko 2010; Ghemawat 2003; Beaver 2010], GIGA+ directories use the traditional UNIX VFS interface and provide POSIX-like semantics to support unmodified applications.

Our evaluation demonstrates that GIGA+ directories scale linearly on a cluster of 32 servers and deliver a throughput of more than 98,000 file creates per second – outscaling the Ceph file system [Weil 2006] and the HBase distributed key-value store [HBase], and exceeding peta-scale scalability requirements [Newman 2008]. GIGA+ indexing also achieves effective load balancing with one to two orders of magnitude less re-partitioning

[1] OrangeFS is currently integrating a GIGA+ based distributed directory implementation into a system based on PVFS [Ligon 2010; OrangeFS].

than if it was based on consistent hashing [Stoica 2001; Karger 1997].

In the rest of the paper, we present the motivating use cases and related work in Section 2, the GIGA+ indexing design and implementation in Sections 3-4, the evaluation results in Section 5, and conclusion in Section 6.

2 Motivation and Background

Over the last two decades, research in large file systems was driven by application workloads that emphasized access to very *large files*. Most cluster file systems provide scalable file I/O bandwidth by enabling parallel access using techniques such as data striping [Hartman 1993; Gibson 1998; Ghemawat 2003], object-based architectures [Gibson 1998; Lustre; Weil 2006; Welch 2008] and distributed locking [Thekkath 1997; Schmuck 2002; Weil 2006]. Few file systems scale metadata performance by using a coarse-grained distribution of metadata over multiple servers [PVFS2; Schmuck 2002; Douceur 2006; Weil 2006]. But most file systems cannot scale access to a *large number of files*, much less efficiently support concurrent creation of millions to billions of files in a single directory. This section summarizes the technology trends calling for scalable directories and how current file systems are ill-suited to satisfy this call.

2.1 Motivation

In today's supercomputers, the most important I/O workload is checkpoint-restart, where many parallel applications running on, for instance, ORNL's CrayXT5 cluster (with 18,688 nodes of twelve processors each) periodically write application state into a file per process, all stored in one directory [Top500 2010; Bent 2009]. Applications that do this per-process checkpointing are sensitive to long file creation delays because of the generally slow file creation rate, especially in one directory, in today's file systems [Bent 2009]. Today's requirement for 40,000 file creates per second in a single directory [Newman 2008] will become much bigger in the impending Exascale-era, when applications may run on clusters with up to billions of CPU cores [Kogge 2008].

Supercomputing checkpoint-restart, although important, might not be a sufficient reason for overhauling the current file system directory implementations. Yet there are diverse applications, such as gene sequencing, image processing [Tweed 2008], phone logs for accounting and billing, and photo storage [Beaver 2010], that essentially want to store an unbounded number of files that are logically part of one directory. Although these applications are often using the file system as a fast, lightweight "key-value store", replacing the underlying file system with a database is an oft-rejected option because it is undesirable to port existing code to use a new API (like SQL) and because traditional databases do not provide the scalability of cluster file systems running on thousands of nodes [Seltzer 2008; Agrawal 2008; Stonebraker 2005; Abouzeid 2009].

Authors of applications seeking lightweight stores for lots of small data can either rewrite applications to avoid large directories or rely on underlying file systems to improve support for large directories. Numerous applications, including browsers and web caches, use the former approach where the application manages a large logical directory by creating many small, intermediate sub-directories with files hashed into one of these sub-directories. This paper chose the latter approach because users prefer this solution. Separating large directory management from applications has two advantages. First, developers do not need to re-implement large directory management for every application (and can avoid writing and debugging complex code). Second, an application-agnostic large directory subsystem can make more informed decisions about dynamic aspects of a large directory implementation, such as load-adaptive partitioning and growth rate specific migration scheduling.

Unfortunately most file system directories do not currently provide the desired scalability: popular local file systems are still being designed to handle little more than tens of thousands of files in each directory [ZFS-discuss 2009; NetApp-Community-Form 2010; StackOverflow 2009] and even distributed file systems that run on the largest clusters, including HDFS [Shvachko 2010], GoogleFS [Ghemawat 2003], PanFS [Welch 2008] and PVFS [PVFS2], are limited by the speed of the single metadata server that manages an entire directory. In fact, because GoogleFS scaled up to only about 50 million files, the next version, ColossusFS, will use BigTable [Chang 2006] to provide a distributed file system metadata service [Fikes 2010].

Although there are file systems that distribute the directory tree over different servers, such as Farsite [Douceur 2006] and PVFS [PVFS2], to our knowledge, only three file systems now (or soon will) distribute single large directories: IBM's GPFS [Schmuck 2002], Oracle's Lustre [Lustre 2010], and UCSC's Ceph [Weil 2006].

2.2 Related work

GIGA+ has been influenced by the scalability and concurrency limitations of several distributed indices and their implementations.

GPFS: GPFS is a shared-disk file system that uses a distributed implementation of Fagin's extendible hashing for its directories [Fagin 1979; Schmuck 2002]. Fagin's extendible hashing dynamically doubles the size of the hash-table pointing pairs of links to the original bucket and expanding only the overflowing bucket (by restricting implementations to a specific family of hash functions)

[Fagin 1979]. It has a two-level hierarchy: buckets (to store the directory entries) and a table of pointers (to the buckets). GPFS represents each bucket as a disk block and the pointer table as the block pointers in the directory's i-node. When the directory grows in size, GPFS allocates new blocks, moves some of the directory entries from the overflowing block into the new block and updates the block pointers in the i-node.

GPFS employs its client cache consistency and distributed locking mechanism to enable concurrent access to a shared directory [Schmuck 2002]. Concurrent readers can cache the directory blocks using shared reader locks, which enables high performance for read-intensive workloads. Concurrent writers, however, need to acquire write locks from the lock manager before updating the directory blocks stored on the shared disk storage. When releasing (or acquiring) locks, GPFS versions before 3.2.1 force the directory block to be flushed to disk (or read back from disk) inducing high I/O overhead. Newer releases of GPFS have modified the cache consistency protocol to send the directory insert requests directly to the current lock holder, instead of getting the block through the shared disk subsystem [Schmuck 2010; Hedges 2010; GPFS 2008]. Still GPFS continues to synchronously write the directory's i-node (i.e., the mapping state) invalidating client caches to provide strong consistency guarantees [Schmuck 2010]. In contrast, GIGA+ allows the mapping state to be stale at the client and never be shared between servers, thus seeking even more scalability.

Lustre and Ceph: Lustre's proposed clustered metadata service splits a directory using a hash of the directory entries only once over all available metadata servers when it exceeds a threshold size [Lustre 2010, 2009]. The effectiveness of this "split once and for all" scheme depends on the eventual directory size and does not respond to dynamic increases in the number of servers. Ceph is another object-based cluster file system that uses dynamic sub-tree partitioning of the namespace and hashes individual directories when they get too big or experience too many accesses [Weil 2006, 2004]. Compared to Lustre and Ceph, GIGA+ splits a directory incrementally as a function of size, i.e., a small directory may be distributed over fewer servers than a larger one. Furthermore, GIGA+ facilitates dynamic server addition achieving balanced server load with minimal migration.

Linear hashing and LH:* Linear hashing grows a hash table by splitting its hash buckets in a linear order using a pointer to the *next* bucket to split [Litwin 1980]. Its distributed variant, called LH* [Litwin 1993], stores buckets on multiple servers and uses a central split coordinator that advances permission to split a partition to the next server. An attractive property of LH* is that it does not update a client's mapping state synchronously after every new split.

GIGA+ differs from LH* in several ways. To maintain consistency of the split pointer (at the coordinator), LH* splits only one bucket at a time [Litwin 1993, 1996]; GIGA+ allows any server to split a bucket at any time without any coordination. LH* offers a complex partition pre-split optimization for higher concurrency [Litwin 1996], but it causes LH* clients to continuously incur some addressing errors even after the index stops growing; GIGA+ chose to minimize (and stop) addressing errors at the cost of more client state.

Consistent hashing: Consistent hashing divides the hash-space into randomly sized ranges distributed over server nodes [Stoica 2001; Karger 1997]. Consistent hashing is efficient at managing membership changes because server changes split or join hash-ranges of adjacent servers only, making it popular for wide-area peer-to-peer storage systems that have high rates of membership churn [Dabek 2001; Rowstron 2001; Muthitacharoen 2002; Rhea 2003]. Cluster systems, even though they have much lower churn than Internet-wide systems, have also used consistent hashing for data partitioning [DeCandia 2007; Lakshman 2009], but have faced interesting challenges.

As observed in Amazon's Dynamo, consistent hashing's data distribution has a high load variance, even after using "virtual servers" to map multiple randomly sized hash-ranges to each node [DeCandia 2007]. GIGA+ uses threshold-based binary splitting that provides better load distribution even for small clusters. Furthermore, consistent hashing systems assume that every data-set needs to be distributed over many nodes to begin with, i.e., they do not have support for incrementally growing data-sets that are mostly small – an important property of file system directories.

Other work: DDS [Gribble 2000] and Boxwood [MacCormick 2004] also used scalable data-structures for storage infrastructure. While both GIGA+ and DDS use hash tables, GIGA+'s focus is on directories, unlike DDS's general cluster abstractions, with an emphasis on indexing that uses inconsistency at the clients; a non-goal for DDS [Gribble 2000]. Boxwood proposed primitives to simplify storage system development, and used B-link trees for storage layouts [MacCormick 2004].

3 GIGA+ Indexing Design

3.1 Assumptions

GIGA+ is intended to be integrated into a modern cluster file system like PVFS, PanFS, GoogleFS, HDFS etc. All these scalable file systems have good fault tolerance usually including a consensus protocol for node membership and global configuration [Burrows 2006; Hunt 2010; Welch 2007]. GIGA+ is not designed to replace membership or fault tolerance; it avoids this where possible and employs them where needed.

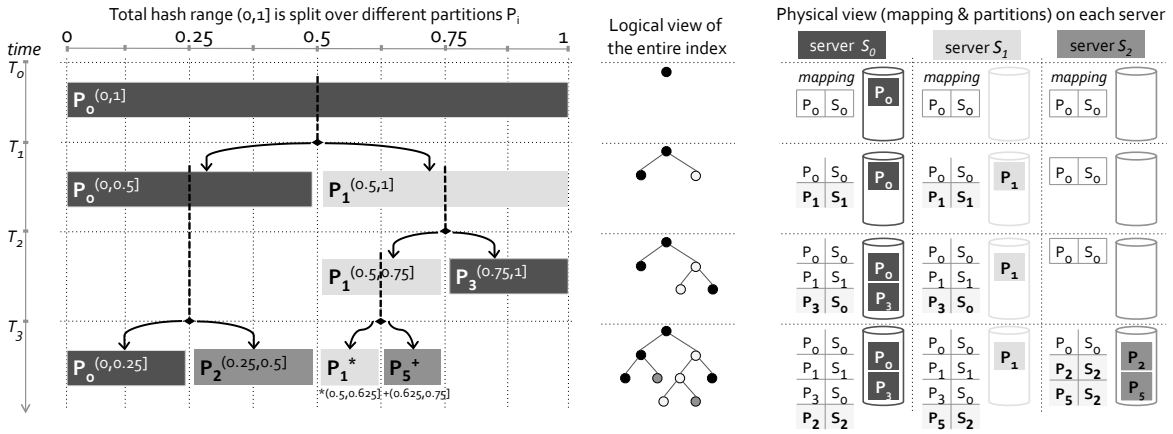

Figure 1 – Concurrent and unsynchronized data partitioning in GIGA+. *The hash-space* $(0,1]$ *is divided into multiple partitions* (P_i) *that are distributed over many servers (different shades of gray). Each server has a local, partial view of the entire index and can independently split its partitions without global co-ordination. In addition to enabling highly concurrent growth, an index starts small (on one server) and scales out incrementally.*

GIGA+ design is also guided by several assumptions about its use cases. First, most file system directories start small and remain small; studies of large file systems have found that 99.99% of the directories contain fewer than 8,000 files [Dayal 2008; Agrawal 2007]. Since only a few directories grow to really large sizes, GIGA+ is designed for incremental growth, that is, an empty or a small directory is initially stored on one server and is partitioned over an increasing number of servers as it grows in size. Perhaps most beneficially, incremental growth in GIGA+ handles adding servers gracefully. This allows GIGA+ to avoid degrading small directory performance; striping small directories across multiple servers will lead to inefficient resource utilization, particularly for directory scans (using `readdir()`) that will incur disk-seek latency on all servers only to read tiny partitions.

Second, because GIGA+ is targeting concurrently shared directories with up to billions of files, caching such directories at each client is impractical: the directories are too large and the rate of change too high. GIGA+ clients do not cache directories and send all directory operations to a server. Directory caching only for small rarely changing directories is an obvious extension employed, for example, by PanFS [Welch 2008], that we have not yet implemented.

Finally, our goal in this research is to complement existing cluster file systems and support unmodified applications. So GIGA+ directories provide the strong consistency for directory entries and files that most POSIX-like file systems provide, i.e., once a client creates a file in a directory all other clients can access the file. This strong consistency API differentiates GIGA+ from "relaxed" consistency provided by newer storage systems including NoSQL systems like Cassandra [Lakshman 2009] and Dynamo [DeCandia 2007].

3.2 Unsynchronized data partitioning

GIGA+ uses hash-based indexing to incrementally divide each directory into multiple partitions that are distributed over multiple servers. Each filename (contained in a directory entry) is hashed and then mapped to a partition using an index. Our implementation uses the cryptographic MD5 hash function but is not specific to it. GIGA+ relies only on one property of the selected hash function: for any distribution of unique filenames, the hash values of these filenames must be uniformly distributed in the hash space [Rivest 1992]. This is the core mechanism that GIGA+ uses for load balancing.

Figure 1 shows how GIGA+ indexing grows incrementally. In this example, a directory is to be spread over three servers $\{S_0, S_1, S_2\}$ in three shades of gray color. $P_i^{(x,y]}$ denotes the hash-space range $(x,y]$ held by a partition with the unique identifier i.[2] GIGA+ uses the identifier i to map P_i to an appropriate server S_i using a round-robin mapping, i.e., server S_i is i `modulo` *num_servers*. The color of each partition indicates the (color of the) server it resides on. Initially, at time T_0, the directory is small and stored on a single partition $P_0^{(0,1]}$ on server S_0. As the directory grows and the partition size exceeds a threshold number of directory entries, provided this server knows of an underutilized server, S_0 splits $P_0^{(0,1]}$ into two by moving the greater half of its hash-space range to a new partition $P_1^{(0.5,1]}$ on S_1. As the directory expands, servers continue to split partitions onto more servers until all have about the same fraction of the hash-space to manage (analyzed in Section 5.2 and 5.3). GIGA+ computes a split's target partition identifier using well-known radix-based

[2] For simplicity, we disallow the hash value zero from being used.

techniques.[3]

The key goal for GIGA+ is for each server to split independently, without system-wide serialization or synchronization. Accordingly, servers make local decisions to split a partition. The side-effect of uncoordinated growth is that GIGA+ servers do not have a global view of the partition-to-server mapping on any one server; each server only has a partial view of the entire index (the mapping tables in Figure 1). Other than the partitions that a server manages, a server knows only the identity of the server that knows more about each "child" partition resulting from a prior split by this server. In Figure 1, at time T_3, server S_1 manages partition P_1 at tree depth $r = 3$, and knows that it previously split P_1 to create children partitions, P_3 and P_5, on servers S_0 and S_2 respectively. Servers are mostly unaware about partition splits that happen on other servers (and did not target them); for instance, at time T_3, server S_0 is unaware of partition P_5 and server S_1 is unaware of partition P_2.

Specifically, each server knows only the split history of its partitions. The full GIGA+ index is a complete history of the directory partitioning, which is the transitive closure over the local mappings on each server. This full index is also not maintained synchronously by any client. GIGA+ clients can enumerate the partitions of a directory by traversing its split histories starting with the zeroth partition P_0. However, such a full index constructed and cached by a client may be stale at any time, particularly for rapidly mutating directories.

3.3 Tolerating inconsistent mapping at clients

Clients seeking a specific filename find the appropriate partition by probing servers, possibly incorrectly, based on their cached index. To construct this index, a client must have resolved the directory's parent directory entry which contains a cluster-wide i-node identifying the server and partition for the zeroth partition P_0. Partition P_0 may be the appropriate partition for the sought filename, or it may not because of a previous partition split that the client has not yet learned about. An "incorrectly" addressed server detects the addressing error by recomputing the partition identifier by re-hashing the filename. If this hashed filename does not belong in the partition it has, this server sends a split history update to the client. The client updates its cached version of the global index and retries the original request.

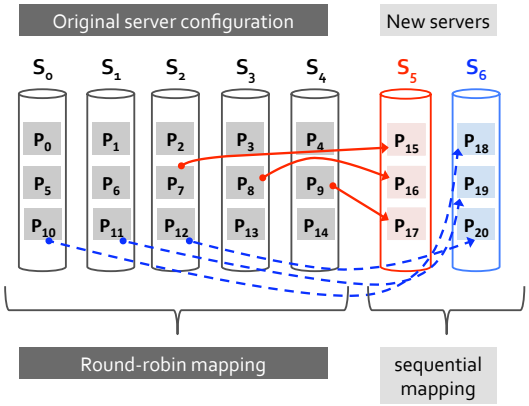

Figure 2 – Server additions in GIGA+. *To minimize the amount of data migrated, indicated by the arrows that show splits,* GIGA+ *changes the partition-to-server mapping from round-robin on the original server set to sequential on the newly added servers.*

The drawback of allowing inconsistent indices is that clients may need additional probes before addressing requests to the correct server. The required number of incorrect probes depends on the client request rate and the directory mutation rate (rate of splitting partitions). It is conceivable that a client with an empty index may send $O(log(N_p))$ incorrect probes, where N_p is the number of partitions, but GIGA+'s split history updates makes this many incorrect probes unlikely (described in Section 5.4). Each update sends the split histories of all partitions that reside on a given server, filling all gaps in the client index known to this server and causing client indices to catch up quickly. Moreover, after a directory stops splitting partitions, clients soon after will no longer incur any addressing errors. GIGA+'s eventual consistency for cached indices is different from LH*'s eventual consistency because the latter's idea of independent splitting (called pre-splitting in their paper) suffers addressing errors even when the index stops mutating [Litwin 1996].

3.4 Handling server additions

This section describes how GIGA+ adapts to the addition of servers in a running directory service.[4]

When new servers are added to an existing configuration, the system is immediately no longer load balanced, and it should re-balance itself by migrating a minimal number of directory entries from all existing servers equally. Using the round-robin partition-to-server mapping, shown in Figure 1, a naive server addition scheme would require re-mapping almost all directory entries whenever a new server is added.

[3] GIGA+ calculates the identifier of partition i using the depth of the tree, r, which is derived from the number of splits of the zeroth partition P_0. Specifically, if a partition has an identifier i and is at tree depth r, then in the next split P_i will move half of its filenames, from the larger half of its hash-range, to a new partition with identifier $i + 2^r$. After a split completes, both partitions will be at depth $r + 1$ in the tree. In Figure 1, for example, partition $P_1^{(0.5,0.75]}$, with identifier $i = 1$, is at tree depth $r = 2$. A split causes P_1 to move the larger half of its hash-space $(0.625, 0.75]$ to the newly created partition P_5, and both partitions are then at tree depth of $r = 3$.

[4] Server removal (i.e., decommissioned, not failed and later replaced) is not as important for high performance systems so we leave it to be done by user-level data copy tools.

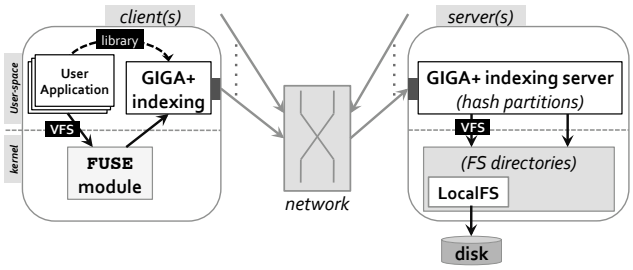

Figure 3 – GIGA+ experimental prototype.

GIGA+ avoids re-mapping all directory entries on addition of servers by differentiating the partition-to-server mapping for initial directory growth from the mapping for additional servers. For additional servers, GIGA+ does not use the round-robin partition-to-server map (shown in Figure 1) and instead maps all future partitions to the new servers in a "sequential manner". The benefit of round-robin mapping is faster exploitation of parallelism when a directory is small and growing, while a sequential mapping for the tail set of partitions does not disturb previously mapped partitions more than is mandatory for load balancing.

Figure 2 shows an example where the original configuration has 5 servers with 3 partitions each, and partitions P_0 to P_{14} use a round-robin rule (for P_i, server is i `mod` N, where N is number of servers). After the addition of two servers, the six new partitions P_{15}-P_{20} will be mapped to servers using the new mapping rule: i `div` M, where M is the number of partitions per server (e.g., 3 partitions/server).

In GIGA+ even the number of servers can be stale at servers and clients. The arrival of a new server and its order in the global server list is declared by the cluster file system's configuration management protocol, such as Zookeeper for HDFS [Hunt 2010], leading to each existing server eventually noticing the new server. Once it knows about new servers, an existing server can inspect its partitions for those that have sufficient directory entries to warrant splitting and would split to a newly added server. The normal GIGA+ splitting mechanism kicks in to migrate only directory entries that belong on the new servers. The order in which an existing server inspects partitions can be entirely driven by client references to partitions, biasing migration in favor of active directories. Or based on an administrator control, it can also be driven by a background traversal of a list of partitions whose size exceeds the splitting threshold.

4 GIGA+ Implementation

GIGA+ indexing mechanism is primarily concerned with distributing the contents and work of large file system directories over many servers, and client interactions with these servers. It is not about the representation of directory entries on disk, and follows the convention of reusing mature local file systems like ext3 or ReiserFS (in Linux) for disk management found as is done by many modern cluster file systems [Shvachko 2010; Welch 2008; Lustre; Weil 2006; PVFS2].

The most natural implementation strategy for GIGA+ is as an extension of the directory functions of a cluster file system. GIGA+ is not about striping the data of huge files, server failure detection and failover mechanism, or RAID/replication of data for disk fault tolerance. These functions are present and, for GIGA+ purposes, adequate in most cluster file systems. Authors of a new version of PVFS, called OrangeFS, and doing just this by integrating GIGA+ into OrangeFS [OrangeFS; Ligon 2010]. Our goal is not to compete with most features of these systems, but to offer technology for advancing their support of high rates of mutation of large collections of small files.

For the purposes of evaluating GIGA+ on file system directory workloads, we have built a skeleton cluster file system; that is, we have not implemented data striping, fault detection or RAID in our experimental framework. Figure 3 shows our user-level GIGA+ directory prototypes built using the FUSE API [FUSE]. Both client and server processes run in user-space, and communicate over TCP using SUN RPC [Srinivasan 1995]. The prototype has three layers: unmodified applications running on clients, the GIGA+ indexing modules (of the skeletal cluster file system on clients and servers) and a backend persistent store at the server. Applications interact with a GIGA+ client using the VFS API (e.g., `open()`, `creat()` and `close()` syscalls). The FUSE kernel module intercepts and redirects these VFS calls the client-side GIGA+ indexing module which implements the logic described in the previous section.

4.1 Server implementation

The GIGA+ server module's primary purpose is to synchronize and serialize interactions between all clients and a specific partition. It need not "store" the partitions, but it owns them by performing all accesses to them. Our server-side prototype is currently layered on lower level file systems, ext3 and ReiserFS. This decouples GIGA+ indexing mechanisms from on-disk representation.

Servers map logical GIGA+ partitions to directory objects within the backend file system. For a given (huge) directory, its entry in its parent directory names the "zeroth partition", $P_0^{(0,1]}$, which is a directory in a server's underlying file system. Most directories are not huge and will be represented by just this one zeroth partition.

GIGA+ stores some information as extended attributes on the directory holding a partition: a GIGA+ directory ID (unique across servers), the the partition identifier P_i and its range $(x,y]$. The range implies the leaf in the directory's logical tree view of the huge directory associated

with this partition (the center column of Figure 1) and that determines the prior splits that had to have occurred to cause this partition to exist (that is, the split history).

To associate an entry in a cached index (a partition) with a specific server, we need the list of servers over which partitions are round robin allocated and the list of servers over which partitions are sequentially allocated. The set of servers that are known to the cluster file system at the time of splitting the zeroth partition is the set of servers that are round robin allocated for this directory and the set of servers that are added after a zeroth partition is split are the set of servers that are sequentially allocated.[5]

Because the current list of servers will always be available in a cluster file system, only the list of servers at the time of splitting the zeroth server needs to be also stored in a partition's extended attributes. Each split propagates the directory ID and set of servers at the time of the zeroth partition split to the new partition, and sets the new partition's identifier P_i and range $(x,y]$ as well as providing the entries from the parent partition that hash into this range $(x,y]$.

Each partition split is handled by the GIGA+ server by locally locking the particular directory partition, scanning its entries to build two sub-partitions, and then transactionally migrating ownership of one partition to another server before releasing the local lock on the partition [Sinnamohideen 2010]. In our prototype layered on local file systems, there is no transactional migration service available, so we move the directory entries and copy file data between servers. Our experimental splits are therefore more expensive than they should be in a production cluster file system.

4.2 Client implementation

The GIGA+ client maintains cached information, some potentially stale, global to all directories. It caches the current server list (which we assume only grows over time) and the number of partitions per server (which is fixed) obtained from whichever server GIGA+ was mounted on. For each active directory GIGA+ clients cache the cluster-wide i-node of the zeroth partition, the directory ID, and the number of servers at the time when the zeroth partition first split. The latter two are available as extended attributes of the zeroth partition. Most importantly, the client maintains a bitmap of the global index built according to Section 3, and a maximum tree-depth, $r = \lceil log(i) \rceil$, of any partition P_i present in the global index.

Searching for a file name with a specific hash value, H, is done by inspecting the index bitmap at the offset j determined by the r lower-order bits of H. If this is set to '1' then H is in partition P_j. If not, decrease r by one and repeat until $r = 0$ which refers to the always known zeroth partition P_0. Identifying the server for partition P_j is done by lookup in the current server list. It is either $j \bmod N$, where N is the number of servers at the time the zeroth partition split), or $j \operatorname{div} M$, where M is the number of partitions per server, with the latter used if j exceeds the product of the number of servers at the time of zeroth partition split and the number of partitions per server.

Most VFS operations depend on lookups; `readdir()` however can be done by walking the bitmaps, enumerating the partitions and scanning the directories in the underlying file system used to store partitions.

4.3 Handling failures

Modern cluster file systems scale to sizes that make fault tolerance mandatory and sophisticated [Ghemawat 2003; Welch 2007; Braam 2007]. With GIGA+ integrated in a cluster file system, fault tolerance for data and services is already present, and GIGA+ does not add major challenges. In fact, handling network partitions and client-side reboots are relatively easy to handle because GIGA+ tolerates stale entries in a client's cached index of the directory partition-to-server mapping and because GIGA+ does not cache directory entries in client or server processes (changes are written through to the underlying file system). Directory-specific client state can be reconstructed by contacting the zeroth partition named in a parent directory entry, re-fetching the current server list and rebuilding bitmaps through incorrect addressing of server partitions during normal operations.

Other issues, such as on-disk representation and disk failure tolerance, are a property of the existing cluster file system's directory service, which is likely to be based on replication even when large data files are RAID encoded [Welch 2008]. Moreover, if partition splits are done under a lock over the entire partition, which is how our experiments are done, the implementation can use a non-overwrite strategy with a simple atomic update of which copy is live. As a result, recovery becomes garbage collection of spurious copies triggered by the failover service when it launches a new server process or promotes a passive backup to be the active server [Burrows 2006; Hunt 2010; Welch 2007].

While our architecture presumes GIGA+ is integrated into a full featured cluster file system, it is possible to layer GIGA+ as an interposition layer over and independent of a cluster file system, which itself is usually layered over multiple independent local file systems [Ghemawat 2003; Shvachko 2010; Welch 2008; PVFS2]. Such a layered GIGA+ would not be able to reuse the fault tolerance services of the underlying cluster file system, leading to an extra layer of fault tolerance. The primary function

[5] The contents of a server list are logical server IDs (or names) that are converted to IP addresses dynamically by a directory service integrated with the cluster file system. Server failover (and replacement) will bind a different address to the same server ID so the list does not change during normal failure handling.

	File System	File creates/second in one directory
GIGA+ *(layered on Reiser)*	Library API	17,902
	VFS/FUSE API	5,977
Local file systems	Linux ext3	16,470
	Linux ReiserFS	20,705
Networked file systems	NFSv3 filer	521
	HadoopFS	4,290
	PVFS	1,064

Table 1 – File create rate in a single directory on a single server. *An average of five runs (with 1% standard deviation).*

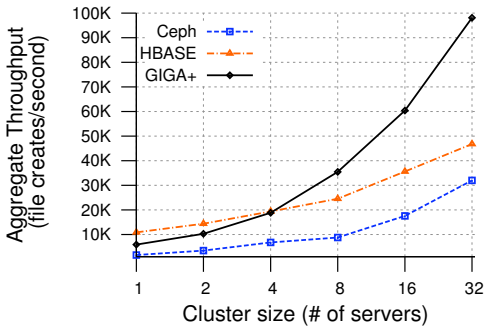

Figure 4 – Scalability of GIGA+ FS directories. GIGA+ *directories deliver a peak throughput of roughly 98,000 file creates per second. The behavior of underlying local file system (ReiserFS) limits* GIGA+*'s ability to match the ideal linear scalability.*

of this additional layer of fault tolerance is replication of the GIGA+ server's write-ahead logging for changes it is making in the underlying cluster file system, detection of server failure, election and promotion of backup server processes to be primaries, and reprocessing of the replicated write-ahead log. Even the replication of the write-ahead log may be unnecessary if the log is stored in the underlying cluster file system, although such logs are often stored outside of cluster file systems to improve the atomicity properties writing to them [Chang 2006; HBase]. To ensure load balancing during server failure recovery, the layered GIGA+ server processes could employ the well-known chained-declustering replication mechanism to shift work among server processes [Hsaio 1990], which has been used in other distributed storage systems [Lee 1996; Thekkath 1997].

5 Experimental Evaluation

Our experimental evaluation answers two questions: (1) How does GIGA+ scale? and (2) What are the tradeoffs of GIGA+'s design choices involving incremental growth, weak index consistency and selection of the underlying local file system for out-of-core indexing (when partitions are very large)?

All experiments were performed on a cluster of 64 machines, each with dual quad-core 2.83GHz Intel Xeon processors, 16GB memory and a 10GigE NIC, and Arista 10 GigE switches. All nodes were running the Linux 2.6.32-js6 kernel (Ubuntu release) and GIGA+ stores partitions as directories in a local file system on one 7200rpm SATA disk (a different disk is used for all non-GIGA+ storage). We assigned 32 nodes as servers and the remaining 32 nodes as load generating clients. The threshold for splitting a partition is always 8,000 entries.

We used the synthetic `mdtest` benchmark [MDTEST] (used by parallel file system vendors and users) to insert zero-byte files in to a directory [Hedges 2010; Weil 2006]. We generated three types of workloads. First, a *concurrent create* workload that creates a large number of files concurrently in a single directory. Our configuration uses eight processes per client to simultaneously create files in a common directory, and the number of files created is proportional to the number of servers: a single server manages 400,000 files, a 800,000 file directory is created on 2 servers, a 1.6 million file directory on 4 servers, up to a 12.8 million file directory on 32 servers. Second, we use a *lookup workload* that performs a `stat()` on random files in the directory. And finally, we use a mixed workload where clients issue create and lookup requests in a pre-configured ratio.

5.1 Scale and performance

We begin with a baseline for the performance of various file systems running the `mdtest` benchmark. First we compare `mdtest` running locally on Linux ext3 and ReiserFS local file systems to `mdtest` running on a separate client and single server instance of the PVFS cluster file system (using ext3) [PVFS2], Hadoop's HDFS (using ext3) [Shvachko 2010] and a mature commercial NFSv3 filer. In this experiment GIGA+ always uses one partition per server. Table 1 shows the baseline performance.

For GIGA+ we use two machines with ReiserFS on the server and two ways to bind `mdtest` to GIGA+: direct library linking (non-POSIX) and VFS/FUSE linkage (POSIX). The library approach allows `mdtest` to use custom object creation calls (such as `giga_creat()`) avoiding system call and FUSE overhead in order to compare to `mdtest` directly in the local file system. Among the local file systems, with local `mdtest` threads generating file creates, both ReiserFS and Linux ext3 deliver high directory insert rates.[6] Both file systems were configured with `-noatime` and `-nodiratime` option; Linux ext3 used write-back journaling and the `dir_index` option to enable hashed-tree indexing, and ReiserFS was configured with the `-notail` option, a small-file optimization

[6] We tried XFS too, but it was extremely slow during the create-intensive workload and do not report those numbers in this paper.

that packs the data inside an i-node for high performance [Reiser 2004]. GIGA+ with `mdtest` workload generating threads on a different machine, when using the library interface (sending only one RPC per create) and ReiserFS as the backend file system, creates at better than 80% of the rate of ReiserFS with local load generating threads. This comparison shows that remote RPC is not a huge penalty for GIGA+. We tested this library version only to gauge GIGA+ efficiency compared to local file systems and do not use this setup for any remaining experiments.

To compare with the network file systems, GIGA+ uses the VFS/POSIX interface. In this case each VFS file `creat()` results in three RPC calls to the server: `getattr()` to check if a file exists, the actual `creat()` and another `getattr()` after creation to load the created file's attributes. For a more enlightening comparison, cluster file systems were configured to be functionally equivalent to the GIGA+ prototype; specifically, we disabled HDFS's write-ahead log and replication, and we used default PVFS which has no redundancy unless a RAID controller is added. For the NFSv3 filer, because it was in production use, we could not disable its RAID redundancy and it is correspondingly slower than it might otherwise be. GIGA+ directories using the VFS/FUSE interface also outperforms all three networked file systems, probably because the GIGA+ experimental prototype is skeletal while others are complex production systems.

Figure 4 plots aggregate operation throughput, in file creates per second, averaged over the complete *concurrent create* benchmark run as a function of the number of servers (on a log-scale X-axis). GIGA+ with partitions stored as directories in ReiserFS scales linearly up to the size of our 32-server configuration, and can sustain 98,000 file creates per second - this exceeds today's most rigorous scalability demands [Newman 2008].

Figure 4 also compares GIGA+ with the scalability of the Ceph file system and the HBase distributed key-value store. For Ceph, Figure 4 reuses numbers from experiments performed on a different cluster from the original paper [Weil 2006]. That cluster used dual-core 2.4GHz machines with IDE drives, with equal numbered separate nodes as workload generating clients, metadata servers and disk servers with object stores layered on Linux ext3. HBase is used to emulate Google's Colossus file system which plans to store file system metadata in BigTable instead of internally on single master node[Fikes 2010]. We setup HBase on a 32-node HDFS configuration with a single copy (no replication) and disabled two parameters: blocking while the HBase servers are doing compactions and write-ahead logging for inserts (a common practice to speed up inserting data in HBase). This configuration allowed HBase to deliver better performance than GIGA+ for the single server configuration because the HBase tables are striped over all 32-nodes in the HDFS cluster.

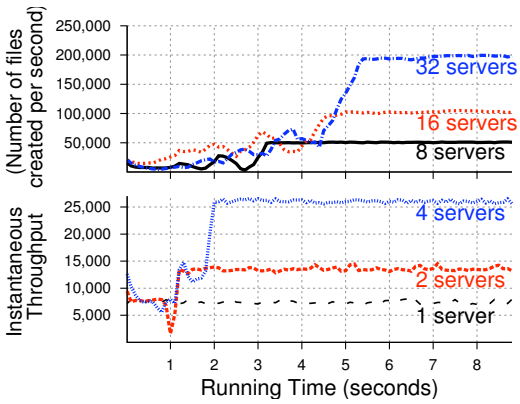

Figure 5 – Incremental scale-out growth. GIGA+ *achieves linear scalability after distributing one partition on each available server. During scale-out, periodic drops in aggregate create rate correspond to concurrent splitting on all servers.*

But configurations with many HBase servers scale poorly.

GIGA+ also demonstrated scalable performance for the *concurrent lookup* workload delivering a throughput of more than 600,000 file lookups per second for our 32-server configuration (not shown). Good lookup performance is expected because the index is not mutating and load is well-distributed among all servers; the first few lookups fetch the directory partitions from disk into the buffer cache and the disk is not used after that. Section 5.4 gives insight on addressing errors during mutations.

5.2 Incremental scaling properties

In this section, we analyze the scaling behavior of the GIGA+ index independent of the disk and the on-disk directory layout (explored later in Section 5.5). To eliminate performance issues in the disk subsystem, we use Linux's in-memory file system, `tmpfs`, to store directory partitions. Note that we use `tmpfs` only in this section, all other analysis uses on-disk file systems.

We run the *concurrent create* benchmark to create a large number of files in an empty directory and measure the aggregate throughput (file creates per second) continuously throughout the benchmark. We ask two questions about scale-out heuristics: (1) what is the effect of splitting during incremental scale-out growth? and (2) how many partitions per server do we keep?

Figure 5 shows the first 8 seconds of the *concurrent create* workload when the number of partitions per server is one. The primary result in this figure is the near linear create rate seen after the initial seconds. But the initial few seconds are more complex. In the single server case, as expected, the throughput remains flat at roughly 7,500 file creates per second due to the absence of any other server. In the 2-server case, the directory starts on a single server and splits when it has more than 8,000 entries in the partition. When the servers are busy splitting, at the

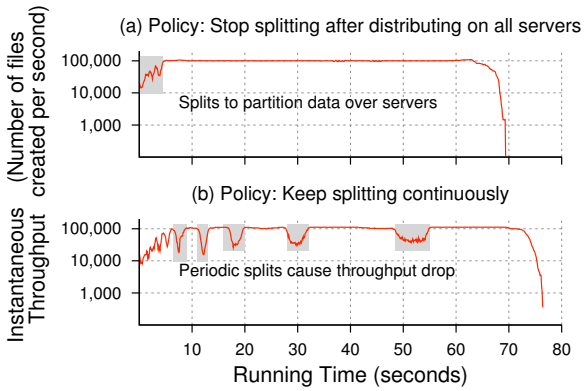

Figure 6 – Effect of splitting heuristics. GIGA+ *shows that splitting to create at most one partition on each of the 16 servers delivers scalable performance. Continuous splitting, as in classic database indices, is detrimental in a distributed scenario.*

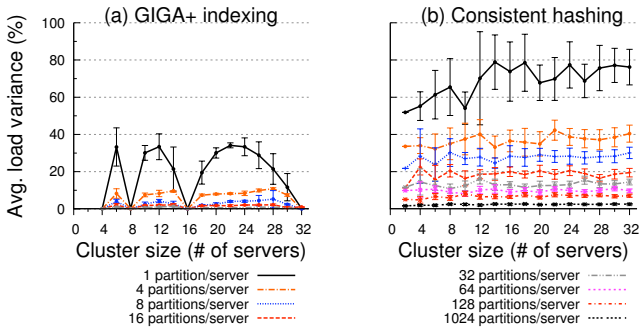

Figure 7 – Load-balancing in GIGA+. *These graphs show the quality of load balancing measured as the mean load deviation across the entire cluster (with 95% confident interval bars). Like virtual servers in consistent hashing,* GIGA+ *also benefits from using multiple hash partitions per server.* GIGA+ *needs one to two orders of magnitude fewer partitions per server to achieve comparable load distribution relative to consistent hashing.*

0.8-second mark, throughput drops to half for a short time.

Throughput degrades even more during the scale-out phase as the number of directory servers goes up. For instance, in the 8-server case, the aggregate throughput drops from roughly 25,000 file creates/second at the 3-second mark to as low as couple of hundred creates/second before growing to the desired 50,000 creates/second. This happens because all servers are busy splitting, i.e., partitions overflow at about the same time which causes all servers (where these partitions reside) to split without any co-ordination at the same time. And after the split spreads the directory partitions on twice the number of servers, the aggregate throughput achieves the desired linear scale.

In the context of the second question about how many partitions per server, classic hash indices, such as extendible and linear hashing [Fagin 1979; Litwin 1980], were developed for out-of-core indexing in single-node databases. An out-of-core table keeps splitting partitions whenever they overflow because the partitions correspond to disk allocation blocks [Gray 1992]. This implies an unbounded number of partitions per server as the table grows. However, the splits in GIGA+ are designed to parallelize access to a directory by distributing the directory load over all servers. Thus GIGA+ can stop splitting after each server has a share of work, i.e., at least one partition. When GIGA+ limits the number of partitions per server, the size of partitions continue to grow and GIGA+ lets the local file system on each server handle physical allocation and out-of-core memory management.

Figure 6 compares the effect of different policies for the number of partitions per server on the system throughput (using a log-scale Y-axis) during a test in which a large directory is created over 16 servers. Graph (a) shows a split policy that stops when every server has one partition, causing partitions to ultimately get much bigger than 8,000 entries. Graph (b) shows the continuous splitting policy used by classic database indices where a split happens whenever a partition has more than 8,000 directory entries. With continuous splitting the system experiences periodic throughput drops that last longer as the number of partitions increases. This happens because repeated splitting maps multiple partitions to each server, and since uniform hashing will tend to overflow all partitions at about the same time, multiple partitions will split on all the servers at about the same time.

Lesson #1: To avoid the overhead of continuous splitting in a distributed scenario, GIGA+ stops splitting a directory after all servers have a fixed number of partitions and lets a server's local file system deal with out-of-core management of large partitions.

5.3 Load balancing efficiency

The previous section showed only configurations where the number of servers is a power-of-two. This is a special case because it is naturally load-balanced with only a single partition per server: the partition on each server is responsible for a hash-range of size 2^r-th part of the total hash-range $(0, 1]$. When the number of servers is not a power-of-two, however, there is load imbalance. Figure 7 shows the load imbalance measured as the average fractional deviation from even load for all numbers of servers from 1 to 32 using Monte Carlo model of load distribution. In a cluster of 10 servers, for example, each server is expected to handle 10% of the total load; however, if two servers are experiencing 16% and 6% of the load, then they have 60% and 40% variance from the average load respectively. For different cluster sizes, we measure the variance of each server, and use the average (and 95% confidence interval error bars) over all the servers.

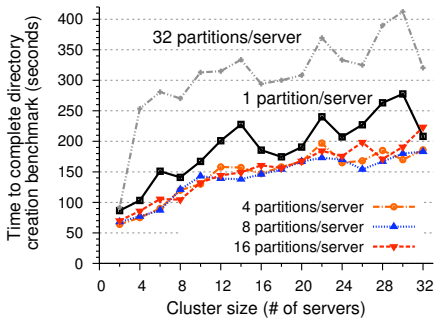

Figure 8 – Cost of splitting partitions. *Using 4, 8, or 16 partitions per server improves the performance of* GIGA+ *directories layered on Linux ext3 relative to 1 partition per server (better load-balancing) or 32 partitions per server (when the cost of more splitting dominates the benefit of load-balancing).*

We compute load imbalance for GIGA+ in Figure 7(a) as follows: when the number of servers N is not a power-of-two, $2^r < N < 2^{r+1}$, then a random set of $N - 2^r$ partitions from tree depth r, each with range size $1/2^r$, will have split into $2(N-2^r)$ partitions with range size $1/2^{r+1}$. Figure 7(a) shows the results of five random selections of $N - 2^r$ partitions that split on to the $r+1$ level. Figure 7(a) shows the expected periodic pattern where the system is perfectly load-balanced when the number of servers is a power-of-two. With more than one partition per server, each partition will manage a smaller portion of the hash-range and the sum of these smaller partitions will be less variable than a single large partition as shown in Figure 7(a). Therefore, more splitting to create more than one partition per server significantly improves load balance when the number of servers is not a power-of-two.

Multiple partitions per server is also used by Amazon's Dynamo key-value store to alleviate the load imbalance in consistent hashing [DeCandia 2007]. Consistent hashing associates each partition with a random point in the hash-space $(0,1]$ and assigns it the range from this point up to the next larger point and wrapping around, if necessary. Figure 7(b) shows the load imbalance from Monte Carlo simulation of using multiple partitions (virtual servers) in consistent hashing by using five samples of a random assignment for each partition and how the sum, for each server, of partition ranges selected this way varies across servers. Because consistent hashing's partitions have more randomness in each partition's hash-range, it has a higher load variance than GIGA+ – almost two times worse. Increasing the number of hash-range partitions significantly improves load distribution, but consistent hashing needs more than 128 partitions per server to match the load variance that GIGA+ achieves with 8 partitions per server – an order of magnitude more partitions.

More partitions is particularly bad because it takes longer for the system to stop splitting, and Figure 8 shows

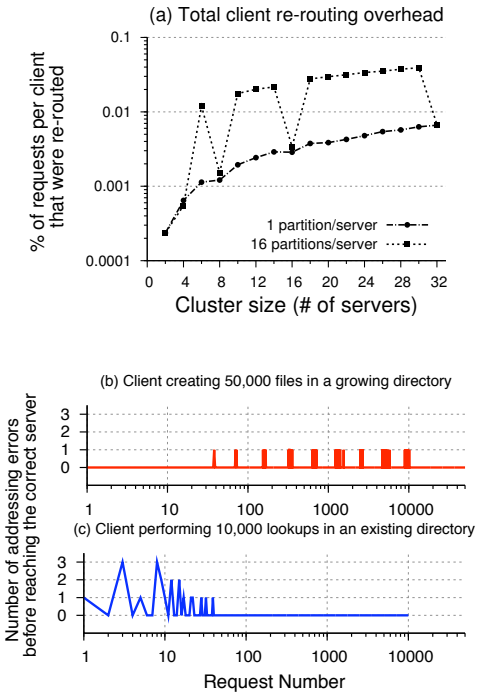

Figure 9 – Cost of using inconsistent mapping at the clients. *Using weak consistency for mapping state at the clients has a very negligible overhead on client performance (a). And this overhead – extra message re-addressing hops – occurs for initial requests until the client learns about all the servers (b and c).*

how this can impact overall performance. Consistent hashing theory has alternate strategies for reducing imbalance but these often rely on extra accesses to servers all of the time and global system state, both of which will cause impractical degradation in our system [Byers 2003].

Since having more partitions per server always improves load-balancing, at least a little, how many partitions should GIGA+ use? Figure 8 shows the *concurrent create* benchmark time for GIGA+ as a function of the number of servers for 1, 4, 8, 16 and 32 partitions per server. We observe that with 32 partitions per server GIGA+ is roughly 50% slower than with 4, 8 and 16 partitions per server. Recall from Figure 7(a) that the load-balancing efficiency from using 32 partitions per server is only about 1% better than using 16 partitions per server; the high cost of splitting to create twice as many partitions outweighs the minor load-balancing improvement.

Lesson #2: Splitting to create more than one partition per server significantly improves GIGA+ load balancing for non power-of-two numbers of servers, but because of the performance penalty during extra splitting the overall performance is best with only a few partitions per server.

5.4 Cost of weak mapping consistency

Figure 9(a) shows the overhead incurred by clients when their cached indices become stale. We measure the per-

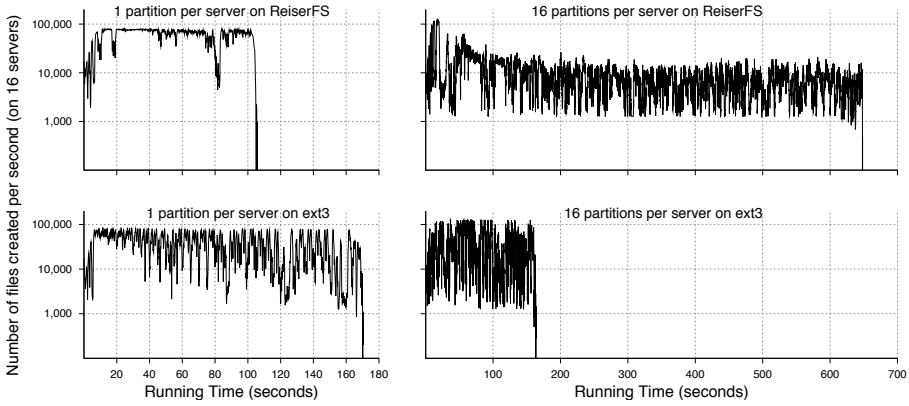

Figure 10 – Effect of underlying file systems. *This graph shows the concurrent create benchmark behavior when the* GIGA+ *directory service is distributed on 16 servers with two local file systems, Linux ext3 and ReiserFS. For each file system, we show two different numbers of partitions per server, 1 and 16.*

centage of all client requests that were re-routed when running the *concurrent create* benchmark on different cluster sizes. This figure shows that, in absolute terms, fewer than 0.05% of the requests are addressed incorrectly; this is only about 200 requests per client because each client is doing 400,000 file creates. The number of addressing errors increases proportionally with the number of partitions per server because it takes longer to create all partitions. In the case when the number of servers is a power-of-two, after each server has at least one partition, subsequent splits yield two smaller partitions on the same server, which will not lead to any additional addressing errors.

We study further the worst case in Figure 9(a), 30 servers with 16 partitions per server, to learn when addressing errors occur. Figure 9(b) shows the number of errors encountered by each request generated by one client thread (i.e., one of the eight workload generating threads per client) as it creates 50,000 files in this benchmark. Figure 9(b) suggests three observations. First, the index update that this thread gets from an incorrectly addressed server is always sufficient to find the correct server on the second probe. Second, that addressing errors are bursty, one burst for each level of the index tree needed to create 16 partitions on each of 30 servers, or 480 partitions ($2^8 < 480 < 2^9$). And finally, that the last 80% of the work is done after the last burst of splitting without any addressing errors.

To further emphasize how little incorrect server addressing clients generate, Figure 9(c) shows the addressing experience of a new client issuing 10,000 lookups after the current create benchmark has completed on 30 servers with 16 partitions per server.[7] This client makes no more than 3 addressing errors for a specific request, and no more than 30 addressing errors total and makes no more addressing errors after the 40th request.

Lesson #3: GIGA+ clients incur neglible overhead (in terms of incorrect addressing errors) due to stale cached indices, and no overhead shortly after the servers stop splitting partitions. Although not a large effect, fewer partitions per server lowers client addressing errors.

5.5 Interaction with backend file systems

Because some cluster file systems represent directories with equivalent directories in a local file system [Lustre] and because our GIGA+ experimental prototype represents partitions as directories in a local file system, we study how the design and implementation of Linux ext3 and ReiserFS local file systems affects GIGA+ partition splits. Although different local file system implementations can be expected to have different performance, especially for emerging workloads like ours, we were surprised by the size of the differences.

Figure 10 shows GIGA+ file create rates when there are 16 servers for four different configurations: Linux ext3 or ReiserFS storing partitions as directories, and 1 or 16 partitions per server. Linux ext3 directories use h-trees [Cao 2007] and ReiserFS uses balanced B-trees [Reiser 2004]. We observed two interesting phenomenon: first, the benchmark running time varies from about 100 seconds to over 600 seconds, a factor of 6, and second, the backend file system yielding the faster performance is different when there are 16 partitions on each server than with only one.

Comparing a single partition per server in GIGA+ over ReiserFS and over ext3 (left column in Figure 10), we observe that the benchmark completion time increases from about 100 seconds using ReiserFS to nearly 170 seconds using ext3. For comparison, the same bench-

[7] Figure 9 predicts the addressing errors of a client doing only lookups on a mutating directory because both `create(filename)` and `lookup(filename)` do the same addressing.

mark completed in 70 seconds when the backend was the in-memory `tmpfs` file system. Looking more closely at Linux ext3, as a directory grows, ext3's journal also grows and periodically triggers ext3's `kjournald` daemon to flush a part of the journal to disk. Because directories are growing on all servers at roughly the same rate, multiple servers flush their journal to disk at about the same time leading to troughs in the aggregate file create rate. We observe this behavior for all three journaling modes supported by ext3. We confirmed this hypothesis by creating an ext3 configuration with the journal mounted on a second disk in each server, and this eliminated most of the throughput variability observed in ext3, completing the benchmark almost as fast as with ReiserFS. For ReiserFS, however, placing the journal on a different disk had little impact.

The second phenomenon we observe, in the right column of Figure 10, is that for GIGA+ with 16 partitions per server, ext3 (which is insensitive to the number of partitions per server) completes the create benchmark more than four times faster than ReiserFS. We suspect that this results from the on-disk directory representation. ReiserFS uses a balanced B-tree for *all objects* in the file system, which re-balances as the file system grows and changes over time [Reiser 2004]. When partitions are split more often, as in case of 16 partitions per server, the backend file system structure changes more, which triggers more re-balancing in ReiserFS and slows the create rate.

Lesson #4: Design decisions of the backend file system have subtle but large side-effects on the performance of a distributed directory service. Perhaps the representation of a partition should not be left to the vagaries of whatever local file system is available.

6 Conclusion

In this paper we address the emerging requirement for POSIX file system directories that store massive number of files and sustain hundreds of thousands of concurrent mutations per second. The central principle of GIGA+ is to use asynchrony and eventual consistency in the distributed directory's internal metadata to push the limits of scalability and concurrency of file system directories. We used these principles to prototype a distributed directory implementation that scales linearly to best-in-class performance on a 32-node configuration. Our analysis also shows that GIGA+ achieves better load balancing than consistent hashing and incurs a neglible overhead from allowing stale lookup state at its clients.

Acknowledgements. This work is based on research supported in part by the Department of Energy, under award number DE-FC02-06ER25767, by the Los Alamos National Laboratory, under contract number 54515-001-07, by the Betty and Gordon Moore Foundation, by the National Science Foundation under awards CCF-1019104 and SCI-0430781, and by Google and Yahoo! research awards. We thank Cristiana Amza, our shepherd, and the anonymous reviewers for their thoughtful reviews of our paper. John Bent and Gary Grider from LANL provided valuable feedback from the early stages of this work; Han Liu assisted with HBase experimental setup; and Vijay Vasudevan, Wolfgang Richter, Jiri Simsa, Julio Lopez and Varun Gupta helped with early drafts of the paper. We also thank the member companies of the PDL Consortium (including APC, DataDomain, EMC, Facebook, Google, Hewlett-Packard, Hitachi, IBM, Intel, LSI, Microsoft, NEC, NetApp, Oracle, Seagate, Sun, Symantec, and VMware) for their interest, insights, feedback, and support.

References

[Abouzeid 2009] Azza Abouzeid, Kamil Bajda-Pawlikowski, Daniel J. Abadi, Avi Silberschatz, and Alex Rasin. HadoopDB: An Architectural Hybrid of MapReduce and DBMS Technologies for Analytical Workloads. In *Proceedings of the 35th International Conference on Very Large Data Bases (VLDB '09)*, Lyon, France, August 2009.

[Agrawal 2007] Nitin Agrawal, William J. Bolosky, John R. Douceur, and Jacob R. Lorch. A Five-Year Study of File-System Metadata. In *Proceedings of the 5th USENIX Conference on File and Storage Technologies (FAST '07)*, San Jose CA, February 2007.

[Agrawal 2008] Rakesh Agrawal, Anastasia Ailamaki, Philip A. Bernstein, Eric A. Brewer, Michael J. Carey, Surajit Chaudhuri, AnHai Doan, Daniela Florescu, Michael J. Franklin, Hector Garcia-Molina, Johannes Gehrke, Le Gruenwald, Laura M. Haas, Alon Y. Halevy, Joseph M. Hellerstein, Yannis E. Ioannidis, Henry F. Korth, Donald Kossmann, Samuel Madden, Roger Magoulas, Beng Chin Ooi, Tim O'Reilly, Raghu Ramakrishnan, Sunita Sarawagi, Michael Stonebraker, Alexander S. Szalay, and Gerhard Weikum. The Claremont report on database research. *ACM SIGMOD Record*, 37(3), September 2008.

[Beaver 2010] Doug Beaver, Sanjeev Kumar, Harry C. Li, Jason Sobel, and Peter Vajgel. Finding a Needle in Haystack: Facebook's Photo Storage. In *Proceedings of the 9th USENIX Symposium on Operating Systems Design and Implementation (OSDI '10)*, Vancouver, Canada, October 2010.

[Bent 2009] John Bent, Garth Gibson, Gary Grider, Ben McClelland, Paul Nowoczynski, James Nunez, Milo Polte, and Meghan Wingate. PLFS: A Checkpoint Filesystem for Parallel Applications. In *Proceedings of the ACM/IEEE Transactions on Computing Conference on High Performance Networking and Computing (SC '09)*, Portland OR, November 2009.

[Braam 2007] Peter Braam and Byron Neitzel. Scalable Locking and Recovery for Network File Systems. In *Proceedings of the 2nd International Petascale Data Storage Workshop (PDSW '07)*, Reno NV, November 2007.

[Burrows 2006] Mike Burrows. The Chubby lock service for loosely-coupled distributed systems. In *Proceedings of the 7th USENIX Symposium on Operating Systems Design and Implementation (OSDI '06)*, Seattle WA, November 2006.

[Byers 2003] John Byers, Jeffrey Considine, and Michael Mitzenmacher. Simple Load Balancing for Distributed Hash Tables. In *Proceedings of the 2nd International Workshop on Peer-to-Peer Systems (IPTPS '03)*, Berkeley CA, February 2003.

[Cao 2007] Mingming Cao, Theodore Y. Ts'o, Badari Pulavarty, Suparna Bhattacharya, Andreas Dilger, and Alex Tomas. State of the Art: Where we are with the ext3 filesystem. In *Proceedings of the Ottawa Linux Symposium (OLS '07)*, Ottawa, Canada, June 2007.

[Chang 2006] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Michael Burrows, Tushar Chandra, Andrew Fikes, and Robert Gruber. Bigtable: A Distributed Storage System for Structured Data. In *Proceedings of the 7th USENIX Symposium on Operating Systems Design and Implementation (OSDI '06)*, Seattle WA, November 2006.

[Dabek 2001] Frank Dabek, M. Frans Kaashoek, David Karger, Robert Morris, and Ion Stoica. Wide-area cooperative storage with CFS. In *Proceedings of the 18th ACM Symposium on Operating Systems Principles (SOSP '01)*, Banff, Canada, October 2001.

[Dayal 2008] Shobhit Dayal. Characterizing HEC Storage Systems at Rest. Technical Report CMU-PDL-08-109, Carnegie Mellon University, July 2008.

[DeCandia 2007] Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, Gunavardhan Kakulapati, Avinash Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Vosshall, and Werner Vogels. Dynamo: Amazon's Highly Available Key-Value Store. In *Proceedings of the 21st ACM Symposium on Operating Systems Principles (SOSP '07)*, Stevenson WA, October 2007.

[Douceur 2006] John R. Douceur and Jon Howell. Distributed Directory Service in the Farsite File System. In *Proceedings of the 7th USENIX Symposium on Operating Systems Design and Implementation (OSDI '06)*, Seattle WA, November 2006.

[Fagin 1979] Ronald Fagin, Jurg Nievergelt, Nicholas Pippenger, and H. Raymond Strong. Extendible Hashing – A Fast Access Method for Dynamic Files. *ACM Transactions on Database Systems*, 4(3), September 1979.

[Fikes 2010] Andrew Fikes. Storage Architecture and Challenges. Presentation at the 2010 Google Faculty Summit. Talk slides at `http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en/us/university/relations/facultysummit2010/storage_architecture_and_challenges.pdf`, June 2010.

[FUSE] FUSE. Filesystem in Userspace. `http://fuse.sf.net/`.

[Ghemawat 2003] Sanjay Ghemawat, Howard Gobioff, and Shuan-Tek Lueng. Google file system. In *Proceedings of the 19th ACM Symposium on Operating Systems Principles (SOSP '03)*, Bolton Landing NY, October 2003.

[Gibson 1998] Garth A. Gibson, Dave F. Nagle, Khalil Amiri, Jeff Butler, Fay W. Chang, Howard Gobioff, Charles Hardin, Erik Riedel, David Rochberg, and Jim Zelenka. A Cost-Effective, High-Bandwidth Storage Architecture. In *Proceedings of the 8th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS '98)*, San Jose CA, October 1998.

[GPFS 2008] GPFS. An Introduction to GPFS Version 3.2.1. `http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp`, November 2008.

[Gray 1992] Jim Gray and Andreas Reuter. *Transaction Processing: Concepts and Techniques*. Morgan Kaufmann Publishers, 1992.

[Gribble 2000] Steve Gribble, Eric Brewer, Joe Hellerstein, and David Culler. Scalable Distributed Data Structures for Internet Service Construction. In *Proceedings of the 4th USENIX Symposium on Operating Systems Design and Implementation (OSDI '00)*, San Diego CA, October 2000.

[Hartman 1993] John H. Hartman and John K. Ousterhout. The Zebra Striped Network File System. In *Proceedings of the 14th ACM Symposium on Operating Systems Principles (SOSP '93)*, Asheville NC, December 1993.

[HBase] HBase. The Hadoop Database. `http://hadoop.apache.org/hbase/`.

[Hedges 2010] Richard Hedges, Keith Fitzgerald, Mark Gary, and D. Marc Stearman. Comparison of leading parallel NAS file systems on commodity hardware. Poster at the Petascale Data Storage Workshop 2010. `http://www.pdsi-scidac.org/events/PDSW10/resources/posters/parallelNASFSs.pdf`, November 2010.

[Hsaio 1990] Hui-I Hsaio and David J. DeWitt. Chained Declustering: A New Availability Strategy for Multiprocessor Database Machines. In *Proceedings of the 6th International Conference on Data Engineering (ICDE '90)*, Washington D.C., February 1990.

[Hunt 2010] Patrick Hunt, Mahadev Konar, Flavio P. Junqueira, and Benjamin Reed. ZooKeeper: Wait-free Coordination for Internet-scale Systems. In *Proceedings of the USENIX Annual Technical Conference (USENIX ATC '10)*, Boston MA, June 2010.

[Karger 1997] David Karger, Eric Lehman, Tom Leighton, Matthew Levine, Daniel Lewin, and Rina Panigrahy. Consistent Hashing and Random Trees: Distributed Caching Pro-

tocols for Relieving Hot Spots on the World Wide Web. In *Proceedings of the ACM Symposium on Theory of Computing (STOC '97)*, El Paso TX, May 1997.

[Kogge 2008] Paul Kogge. ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems. DARPA IPTO Report at http://www.er.doe.gov/ascr/Research/CS/DARPAexascale-hardware(2008).pdf, September 2008.

[Lakshman 2009] Avinash Lakshman and Prashant Malik. Cassandra - A Decentralized Structured Storage System. In *Proceedings of the Workshop on Large-Scale Distribued Systems and Middleware (LADIS '09)*, Big Sky MT, October 2009.

[Lee 1996] Edward K. Lee and Chandramohan A. Thekkath. Petal: Distributed virtual disks. In *Proceedings of the 7th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS '96)*, Cambridge MA, October 1996.

[Ligon 2010] Walt Ligon. Private Communication with Walt Ligon, OrangeFS (http://orangefs.net), November 2010.

[Litwin 1980] Witold Litwin. Linear Hashing: A New Tool for File and Table Addressing. In *Proceedings of the 6th International Conference on Very Large Data Bases (VLDB '80)*, Montreal, Canada, October 1980.

[Litwin 1993] Witold Litwin, Marie-Anne Neimat, and Donovon A. Schneider. LH* - Linear Hashing for Distributed Files. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data (SIGMOD '93)*, Washington D.C., June 1993.

[Litwin 1996] Witold Litwin, Marie-Anne Neimat, and Donovon A. Schneider. LH* - A Scalable, Distributed Data Structure. *ACM Transactions on Database Systems*, 21(4), December 1996.

[Lustre] Lustre. Lustre File System. http://www.lustre.org.

[Lustre 2009] Lustre. Clustered Metadata Design. http://wiki.lustre.org/images/d/db/HPCS_CMD_06_15_09.pdf, September 2009.

[Lustre 2010] Lustre. Clustered Metadata. http://wiki.lustre.org/index.php/Clustered_Metadata, September 2010.

[MacCormick 2004] John MacCormick, Nick Murphy, Marc Najork, Chandramohan A. Thekkath, and Lidong Zhou. Boxwood: Abstractions as the Foundation for Storage Infrastructure. In *Proceedings of the 6th USENIX Symposium on Operating Systems Design and Implementation (OSDI '04)*, San Francisco CA, November 2004.

[MDTEST] MDTEST. mdtest: HPC benchmark for metadata performance. http://sourceforge.net/projects/mdtest/.

[Muthitacharoen 2002] Athicha Muthitacharoen, Robert Morris, Thomer Gil, and Benjie Chen. Ivy: A Read/Write Peer-to-peer File System. In *Proceedings of the 5th USENIX Symposium on Operating Systems Design and Implementation (OSDI '02)*, Boston MA, November 2002.

[NetApp-Community-Form 2010] NetApp-Community-Form. Millions of files in a single directory. Discussion at http://communities.netapp.com/thread/7190?tstart=0, February 2010.

[Newman 2008] Henry Newman. HPCS Mission Partner File I/O Scenarios, Revision 3. http://wiki.lustre.org/images/5/5a/Newman_May_Lustre_Workshop.pdf, November 2008.

[OrangeFS] OrangeFS. Distributed Directories in OrangeFS v2.8.3-EXP (November, 2010). http://orangefs.net/trac/orangefs/wiki/Distributeddirectories.

[PVFS2] PVFS2. Parallel Virtual File System, Version 2. http://www.pvfs2.org.

[Reiser 2004] Hans Reiser. ReiserFS. http://www.namesys.com/, 2004.

[Rhea 2003] Sean Rhea, Patrick Eaton, Dennis Geels, Hakim Weatherspoon, Ben Zhao, and John Kubiatowicz. Pond: the Oceanstore Prototype. In *Proceedings of the 2nd USENIX Conference on File and Storage Technologies (FAST '03)*, San Francisco CA, March 2003.

[Rivest 1992] Ronald A. Rivest. The MD5 Message Digest Algorithm. Internet RFC 1321, April 1992.

[Ross 2006] Rob Ross, Evan Felix, Bill Loewe, Lee Ward, James Nunez, John Bent, Ellen Salmon, and Gary Grider. High End Computing Revitalization Task Force (HECRTF), Inter Agency Working Group (HECIWG) File Systems and I/O Research Guidance Workshop 2006. http://institutes.lanl.gov/hec-fsio/docs/HECIWG-FSIO-FY06-Workshop-Document-FINAL6.pdf, 2006.

[Rowstron 2001] Anthony Rowstron and Peter Druschel. Storage Management and Caching in PAST, A Large-scale, Persistent Peer-to-peer Storage Utility. In *Proceedings of the 18th ACM Symposium on Operating Systems Principles (SOSP '01)*, Banff, Canada, October 2001.

[Schmuck 2010] Frank Schmuck. Private Communication with Frank Schmuck, IBM, February 2010.

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

[Seltzer 2008] Margo Seltzer. Beyond Relational Databases. *Communications of the ACM*, 51(7), July 2008.

[Shvachko 2010] Konstantin Shvachko, Hairong Huang, Sanjay Radia, and Robert Chansler. The Hadoop Distributed File System. In *Proceedings of the 26th IEEE Transactions on Computing Symposium on Mass Storage Systems and Technologies (MSST '10)*, Lake Tahoe NV, May 2010.

[Sinnamohideen 2010] Shafeeq Sinnamohideen, Raja R. Sambasivan, James Hendricks, Likun Liu, and Gregory R. Ganger. A Transparently-Scalable Metadata Service for the Ursa Minor Storage System. In *Proceedings of the USENIX Annual Technical Conference (USENIX ATC '10)*, Boston MA, June 2010.

[Srinivasan 1995] R. Srinivasan. RPC: Remote Procedure Call Protocol Specification Version 2. Internet RFC 1831, August 1995.

[StackOverflow 2009] StackOverflow. Millions of small graphics files and how to overcome slow file system access on XP. Discussion at `http://stackoverflow.com/questions/1638219/`, October 2009.

[Stoica 2001] Ion Stoica, Robert Morris, David Karger, Frans Kaashoek, and Hari Balakrishnan. Chord: A Scalable Peer-To-Peer Lookup Service for Internet Applications. In *Proceedings of the ACM SIGCOMM 2001 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM '01)*, San Diego CA, August 2001.

[Stonebraker 2005] Michael Stonebraker, Daniel J. Abadi, Adam Batkin, Xuedong Chen, Mitch Cherniack, Miguel Ferreira, Edmond Lau, Amerson Lin, Samuel R. Madden, Elizabeth J. O'Neil, Patrick E. O'Neil, Alexander Rasin, Nga Tran, and Stan B. Zdonik. C-Store: A Column-Oriented DBMS. In *Proceedings of the 31st International Conference on Very Large Data Bases (VLDB '05)*, Trondheim, Norway, September 2005.

[Thekkath 1997] Chandramohan A. Thekkath, Timothy Mann, and Edward K. Lee. Frangipani: A Scalable Distributed File System. In *Proceedings of the 16th ACM Symposium on Operating Systems Principles (SOSP '97)*, Saint-Malo, France, October 1997.

[Top500 2010] Top500. Top 500 Supercomputer Sites. `http://www.top500.org`, December 2010.

[Tweed 2008] David Tweed. One usage of up to a million files/directory. Email thread at `http://leaf.dragonflybsd.org/mailarchive/kernel/2008-11/msg00070.html`, November 2008.

[Weil 2006] Sage A. Weil, Scott A. Brandt, Ethan L. Miller, Darrell D. E. Long, and Carlos Maltzahn. Ceph: A Scalable, High-Performance Distributed File System. In *Proceedings of the 7th USENIX Symposium on Operating Systems Design and Implementation (OSDI '06)*, Seattle WA, November 2006.

[Weil 2004] Sage A. Weil, Kristal Pollack, Scott A. Brandt, and Ethan L. Miller. Dynamic Metadata Management for Petabyte-Scale File Systems. In *Proceedings of the ACM/IEEE Transactions on Computing Conference on High Performance Networking and Computing (SC '04)*, Pittsburgh PA, November 2004.

[Welch 2007] Brent Welch. Integrated System Models for Reliable Petascale Storage Systems. In *Proceedings of the 2nd International Petascale Data Storage Workshop (PDSW '07)*, Reno NV, November 2007.

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

[Wheeler 2010] Ric Wheeler. One Billion Files: Scalability Limits in Linux File Systems. Presentation at LinuxCon '10. Talk Slides at `http://events.linuxfoundation.org/slides/2010/linuxcon2010_wheeler.pdf`, August 2010.

[ZFS-discuss 2009] ZFS-discuss. Million files in a single directory. Email thread at `http://mail.opensolaris.org/pipermail/zfs-discuss/2009-October/032540.html`, October 2009.