Iris: A Scalable Cloud File System with Efficient Integrity Checks

Emil Stefanov*
UC Berkeley
emil@cs.berkeley.edu

Marten van Dijk
RSA Laboratories
mvandijk@rsa.com

Alina Oprea
RSA Laboratories
aoprea@rsa.com

Ari Juels
RSA Laboratories
ajuels@rsa.com

Abstract

We present Iris, a practical, authenticated file system designed to support workloads from large enterprises storing data in the cloud and be resilient against potentially untrustworthy service providers. As a transparent layer enforcing strong integrity guarantees, Iris lets an enterprise tenant maintain a large file system in the cloud. In Iris, tenants obtain strong assurance not just on data integrity, but also on data freshness, as well as data retrievability in case of accidental or adversarial cloud failures.

Iris offers an architecture scalable to many clients (on the order of hundreds or even thousands) issuing operations on the file system in parallel. Iris includes new optimization and enterprise-side caching techniques specifically designed to overcome the high network latency typically experienced when accessing cloud storage. Iris also includes novel erasure coding techniques for efficient support of dynamic Proofs of Retrievability (PoR) protocols over the file system.

We describe our architecture and experimental results on a prototype version of Iris. Iris achieves end-to-end throughput of up to 260MB per second for 100 clients issuing simultaneous requests on the file system. (This limit is dictated by the available network bandwidth and maximum hard drive throughput.) We demonstrate that strong integrity protection in the cloud can be achieved with minimal performance degradation.

1 Introduction

Organizations that embrace cloud computing outsource massive amounts of data, as well as workloads to external cloud providers. Cost savings, lower management overhead, and rapid elasticity are just some of the attractions of the cloud.

But cloud computing entails a sacrifice of control. Tenants give up configuration and management oversight of the infrastructure that contains their data and computing resources. In cloud storage systems today, for example, tenants can only discover corruption or loss of their data (particularly infrequently

*This research was performed while visiting RSA Laboratories.

accessed data) if their service providers faithfully report failures or security lapses—or when a system failure occurs. This integrity-measurement gap creates business risk and complicates compliance with regulatory requirements.

We propose a cloud-oriented *authenticated file system* called Iris that gives tenants efficient, comprehensive, and real-time data-integrity verification. The Iris system enables an enterprise tenant—or an auditor acting on the tenant's behalf—to verify the integrity and freshness of any data retrieved from the file system while performing typical file system operations. Data integrity ensures that data has not been accidentally modified or corrupted, while freshness ensures that the latest version of the data is always retrieved (and thus prevents rollback attacks reverting the file system state to a previous version). Moreover, tenants in Iris can efficiently audit the cloud provider on a regular basis and obtain continuous guarantees about the correctness and availability of the entire file system.

Motivating scenario We envision a scenario in which a large enterprise migrates its internal distributed file system to a cloud storage service. An important requirement for our system is that enterprise users (called herein *clients*) perform the same file system operations as they typically do (e.g., file read, write, update, and delete operations, creation and removal of directories) without modifying applications running on user machines. The slowdown in operation latency should be small enough to be unnoticed by users even when a large number of clients (on the order of hundreds and even thousands) issue operations on the file system in parallel.

Design goals in Iris Iris aims to support outsourcing of enterprise-class file-systems to the cloud seamlessly and with minor performance degradation. Thus the design goals of Iris stem from the most common needs of enterprise-class tenants:

- *Efficiency:* Cloud file systems need to achieve throughputs close to those offered by local file systems under thousand of operations issued concurrently by many clients. Individual file system operation latency overhead should also be minimal.

- *Scalability:* A cloud file system should be scalable to large enterprise file systems under a variety of workloads with potentially very sensitive performance requirements. The system should also be scalable to multiple clients issuing operations on the file system in parallel.

- *Transparency:* Transparency and backwards compatibility with existing file system interfaces is important to facilitate migration to the cloud seamlessly.

- *Strong integrity protection:* Data and file system meta-data retrieved from the cloud need to be both *authentic* and *fresh*. Tenants' ability to verify continuously the integrity and availability of their data with minimal bandwidth and computation is a desirable feature, as well.

Contributions of Iris

In more detail, the key technical contributions and novel elements in Iris are:

- **Authenticated file system design:** The first contribution of Iris is to provide data integrity and freshness for an enterprise-class file system in an efficient way. To that end, we design a balanced Merkle-tree data structure that authenticates both file-system data and meta-data blocks. The distinctive features of our data structure design compared to other authenticated file systems is that it efficiently supports updates from multiple clients *in parallel* (without blocking) and it handles *all existing* file system operations (including delete, move and truncate) with minimal overhead. Iris further implements many optimizations for typical file system workloads (e.g., those involving sequential file accesses).

In addition, Iris is designed to overcome the main economic barrier in migrating storage to the cloud: the impact of high network latency. Iris implements novel caching techniques locally, within the enterprise trust boundary. A lightweight (possibly distributed) trusted entity called *the portal* mediates file-system operations passing between the enterprise clients and cloud and caches most recently accessed blocks. We develop techniques to cache the authentication information (nodes of the Merkle tree), handle dependencies among nodes, and preserve Merkle tree consistency when multiple clients simultaneously access nodes from the (partially cached) data structure.

- **Continuous auditing of file system correctness:** Iris enables an enterprise tenant to continuously monitor the operation of the cloud storage service and obtain strong guarantees about the correctness and availability of the *entire file system*. The auditing protocol is a kind of Proof of Retrievabiliity (PoR) [17]. With a PoR, a tenant can verify the correctness and availability of large data collection stored in the cloud with low computation and bandwidth cost. While previous PoR protocols are designed for static data (e.g., archival files), our protocol is the first to efficiently support *dynamic PoR* protocols over the entire file system. One of the key innovations in Iris is the design of a sparse randomized erasure code over the file-system data and metadata. The new erasure code is specifically crafted to hide the code parity structure (typically revealed by other codes during file updates) and be resilient against a potentially adversarial cloud. It enables recovery when corruptions are detected through auditing.

- **End-to-end design and implementation:** One of our main contributions is the end-to-end design and full implementation of Iris consisting of 25,000 lines of code. We show through our performance evaluation that the caching mechanism in Iris is effective in achieving low latency for file system operations (similar to LAN latencies). Moreover, Iris achieves high throughput (up to 260MB for 100 clients issuing simultaneous requests on the file system in our local testbed), with the bottleneck given by the available network bandwidth and hard drive throughput. Finally, we demonstrate that the overall cost of adding strong integrity protection to Iris is minimal.

Organization

In Section 2, we review related work. We give an overview of Iris and describe its architecture in Section 3, discussing the specifics of its integrity layer in Section 4 and the auditing protocol in Section 5. In Section 6, we describe our implementation; we report on our experimental evaluation in Section 7. We conclude in Section 8. In the Appendix, we give an analysis and show some parameterizations of our new sparse erasure code construction needed in the design of the dynamic PoR protocol.

2 Related Work

File systems with integrity support: Early cryptographic file systems were designed to protect data confidentiality [5] and the integrity of data [30] in local storage. Later cryptographic networked file systems provided different integrity guarantees. TCFS [7] and SNAD [22] provide data integrity by storing a hash for each file data block. A number of systems construct a Merkle tree over files in order to authenticate file blocks more efficiently (e.g., [13, 12, 18, 3, 23, 24]).

Many cryptographic file systems to date provide data integrity, but do not authenticate the file system directory structure (or meta-data), e.g., [18, 23, 24]. Others, while authenticating both file system data and meta-data, do not provide strong freshness guarantees. SiRiUS [15] does not ensure data freshness, but only partial meta-data freshness by periodically requiring clients to sign meta-data entries. SUNDR [20] implements a property called "fork consistency" that detects freshness violations only when clients communicate

out of band. More recently, SPORC [11] supports the building of collaborative cloud applications, enabling clients to recover from malicious forks performed by untrusted cloud servers. Depot [21] reconciles malicious forks even in the presence of faulty clients.

To the best of our knowledge, few cryptographic file systems provide freshness of both file system data and meta-data. SFSRO [13] and Cepheus [12] build a Merkle tree over the file system directory tree. While this approach efficiently supports file-system operations like moving or deletion of entire directories, it results in an unbalanced authentication data structure and thus has a high authentication cost for directories with many entries. Athos [16] constructs a balanced data structure that maps the directory tree of the file system in a set of node relations represented as a skip list. Athos abstracts away the hierarchical structure of the directory tree, however, and doesn't provide efficient support for some existing file-system operations, e.g., garbage collection. Moreover, its primary, prototyped design handles only a single client. FARSITE [3] is a peer-to-peer storage system that uses a distributed directory group to maintain meta-data information. Meta-data freshness is guaranteed when more than two thirds of the directory group members are correct. Data freshness is provided by storing hashes of file Merkle trees in the directory group.

Other systems provide data integrity guarantees for key-value stores. Venus [28] implements strong consistency semantics for a key-value store with malicious storage in the back-end. CloudProof [25] provides a mechanism for clients to verify the integrity and freshness, as well as other properties of cloud-stored data.

PoRs/PDPs: A *Proof of Retrievability* (PoR) [17] is a challenge-response protocol that enables a cloud provider to demonstrate to a client that a file is retrievable, i.e., recoverable without any loss or corruption. *Proofs of data possession* (PDP) [4] are related protocols that only detect a large amount of corruption in outsourced data. Most existing PDP [4] and PoR [17, 26, 6, 9] protocols are designed for static data, i.e., infrequently modified data.

Dynamic PDP protocols have been proposed by Erway et al. [10], but they were not designed to handle typical file system operations. For instance, Erway et al. [10] support operations like insertion in the middle of a file, but do not efficiently support moving and deleting entire files or directories. The CS2 system [19] designs and implements an efficient dynamic PDP protocol, as well as techniques for searching over encrypted data.

Several papers ([32] and [33]) claim to construct dynamic PoRs, but in fact only provide dynamic PDP schemes. To the best of our knowledge, designing efficient dynamic PoR protocols is extremely challenging and has stood as an open problem in the community.

3 System model and overview

Iris is designed as an enterprise file system using back-end cloud storage. Clients in Iris (enterprise users) issue file system operations intermediated by Iris and relayed to the public cloud. An important design consideration is that heavy caching on the enterprise side is strictly necessary. There are several reasons for this. First, if local caching is not performed, the cost of network transfer to and from the cloud will far outweigh any storage costs savings ([8] points to the extremely high cost of network transfer). Second, without local caching individual operation latency will be prohibitive for the system to be usable.

Existing network file systems are not designed with similar requirements in mind. For instance, NFS is not optimized for high network latency scenarios [14]. Moreover, most cloud storage systems available today (e.g., Amazon S3) export a key-value store interface and employ a flat namespace. Our system is unique in providing a file system interface to enterprise clients (for compatibility with existing applications), and at the same time ensuring low operation latency. In addition, our main goal is to support integrity protection of both file system data and meta-data and continuous verification of full file system correctness

and availability with minimum overhead.

We describe here Iris's architecture, threat model, and give an overview of our solution and technical challenges.

3.1 System architecture

In our architecture (shown in Figure 1), a trusted portal residing within the enterprise trust boundary intermediates all communication between enterprise clients and the cloud. The portal caches data and meta-data blocks recently accessed by enterprise clients. Cached blocks are evicted once the cache is full and they are not utilized by a pending operation. The portal is also responsible for checking data integrity and freshness for all file system operations (with the *integrity layer* component). Data integrity ensures that data retrieved from the cloud has been written by authorized clients and has not been accidentally modified or corrupted at the cloud side. A stronger property, data freshness, ensures that data accessed by a client during a file system operation is always the latest version written to the cloud by any client.

The portal offers a *portal service* to clients issuing file system operations, and communicates to the cloud through the *storage interface* component. The auditing component issues challenges to the cloud periodically to verify the correctness and availability of the entire file system. The portal plays a central role in recovering from data corruptions: The portal caches error-correcting information (or more concisely, *parities*) for the full file system. When corruption is detected through the auditing protocol, these parities enable recovery of lost or corrupted data. Parities are backed up to the cloud on a regular basis (e.g., once a day or once a week).

To scale to large organizations with tens of thousands of clients, the portal needs to be distributed internally using a tool to ensure consistency of distributed caches (e.g., memcached [2]). For purposes of our prototype detailed in Section 6, we have instantiated the portal on a single server machine and show that it can scale up to 100 clients executing sequential workloads in parallel on the file system.

The cloud maintains the distributed file system, consisting of all files and directories belonging to enterprise users. Iris is designed to use any existing cloud storage system transparently in the back end without modification. In addition, the cloud also stores the MACs and Merkle tree necessary for authenticating data, as well as the checkpointed parity information needed to recover from potential corruptions. As an additional resilience measure, the parity information could be stored on a different cloud or replicated internally within the enterprise.

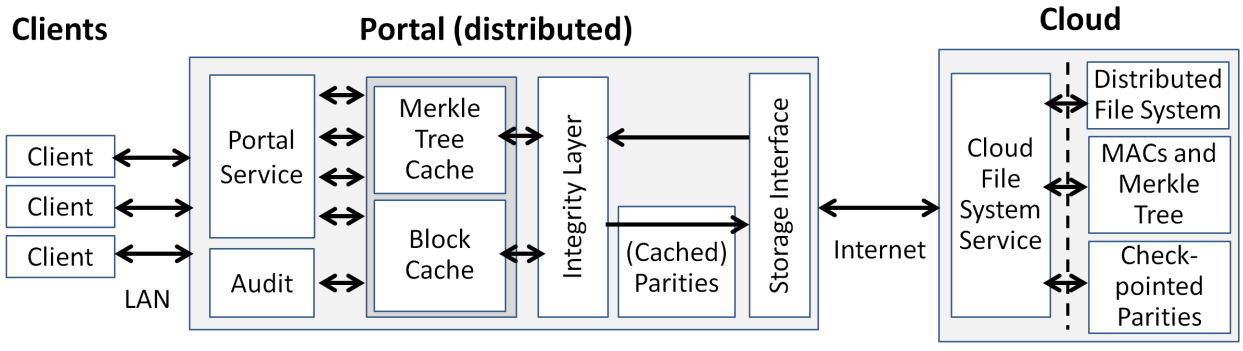

Figure 1: System architecture.

3.2 Threat model

Iris treats the portal, which is controlled by the enterprise, as a trusted component, in the sense that it executes client file-system operations faithfully. No trust assumption is required on clients: They may act arbitrarily within the parameters of the file system. (The file-system may enforce access-control policies on clients through the portal, but such issues lie outside the scope of Iris.)

The cloud, on the other hand, is presumed to be potentially untrustworthy. It may corrupt the file-system in a fully Byzantine manner. The cloud may alter or drop file-system operations transmitted by the portal; it may corrupt or erase files and/or metadata; it may also attempt to present the portal with stale, incorrect, and/or inconsistent views of file-system data. The objective of the portal in Iris is to detect the presentation of *any invalid data by the cloud*, i.e., immediately identify any cloud output that reflects a file-system state different from that produced by a correct execution of the operations emitted by the portal.

3.3 Solution overview and challenges

Iris consists of two major components:

Authenticated file system: As already described, the first challenge we address in building an authenticated enterprise-class file system is the high cost of network latency and bandwidth between the enterprise and cloud. Another challenge is efficient management and caching of the authenticating information. Integrity and freshness verification should be extremely efficient for existing file system operations and induce minimal latency.

Iris employs a two-layer authentication scheme. In its lower layer, it stores on every file block a message-authentication code (MAC)—generated by the portal when a client writes to the file system. These MACs ensure data integrity. To ensure freshness, it is necessary to authenticate not just data blocks, but also their *versions*. Each block has an associated version counter that is incremented every time the block is modified. This version number is bound to the file-block's MAC: To protect against cloud replay of stale file-blocks (rollback attacks), the counters themselves must be authenticated.

The upper layer of the authenticated data structure in Iris is a balanced Merkle-tree-based structure that protects the integrity of the file-block version counters. This data structure embeds the file-system directory tree, and balances each directory for optimization. Attached to each node representing a file is a sub-tree containing file-block version counters. The root of the Merkle tree stored at the portal guarantees the integrity and freshness of both data and meta-data in the file system.

This Merkle-tree-based structure has two distinctive features compared to other authenticated file systems: (1) *Support for existing file system operations:* Iris maintains a balanced binary tree over the file system directory structure to efficiently support existing file system calls; and (2) *Support for concurrent operations:* The Merkle tree supports efficient updates from multiple clients operating on the file system in parallel. Iris also optimizes for the common case of sequential file-block accesses: Sequences of identical version counters are compacted into a single leaf. We detail the data structure in Section 4, and the Merkle tree caching mechanism in Section 6.

Auditing protocol: Iris enables the enterprise tenant to continuously monitor and assess the correctness and availability of the entire file system through the auditing protocol. The auditing protocol in Iris is an instantiation of a PoR protocol and, in fact, the first one that supports data updates. Previous PoR protocols have been designed for static data (files that do not undergo modifications). In any PoR, the tenant samples and checks the correctness of random data blocks retrieved from the cloud to detect any large-scale data corruption. To recover from small-scale damage, parity information computed with an erasure code needs to be maintained over the data.

The main challenge in designing a PoR protocol is that the erasure code structure, i.e., mapping of data blocks to parity blocks, must be randomized to prevent an adversarial server from introducing targeted, undetectable file corruptions. File updates are most problematic as they partially reveal the code structure (in particular the parity blocks corresponding to updated file blocks). At the same time, file updates should be efficient and involve only a small fraction of parity blocks.

We overcome this challenge with two techniques. First, we design Iris to cache parity information locally at the portal (and only checkpoint it to the cloud at fixed time intervals). As the cloud does not perceive individual file updates, but only parity modifications aggregated over a long time interval, the cloud cannot easily infer the mapping from file blocks to parity blocks. Second, we design a new sparse, binary code structure that combines randomly chosen blocks from the file system into a codeword. The code supports updates to the file system very efficiently through binary XOR operations. Its sparse structure supports very large file systems. This novel code construction is carefully parameterized to optimize local storage at the portal side, update cost, and bandwidth and computation in the auditing protocol. We describe the auditing protocol and the erasure code construction in Section 5.

4 Authentication in Iris

We describe in this section how Iris provides strong data protection, including integrity and freshness, for both file system data and meta-data. The authentication scheme in Iris is based on Merkle trees, and designed to support existing file-system operations. In addition, random access to files for both read and write operations is a desirable feature (offered by existing file systems like NFS) that we also choose to implement. The tenant needs to maintain at all times the root of the Merkle trees for checking the integrity and freshness of data retrieved from the cloud. For reducing operation latency, recently accessed nodes in the tree are also cached at the portal (the caching mechanism is described in Section 6).

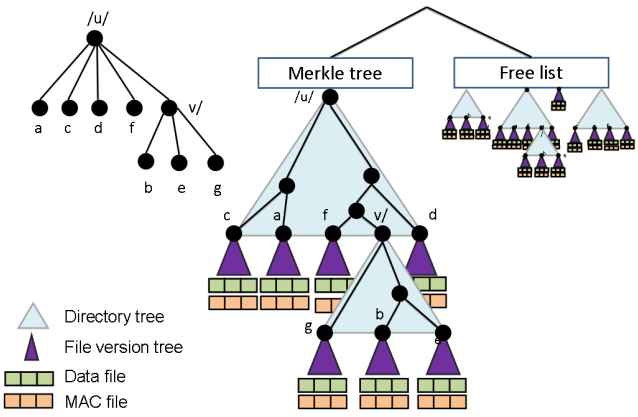

Figure 2: Authenticated tree. A file system directory on the left and its mapping to the Merkle tree on the right.

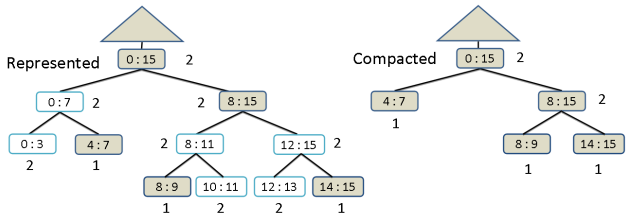

Figure 3: File version tree for a file with 16 blocks. Blocks 0-3 and 10-13 have been written twice, all other blocks have been written once. White nodes on the left are removed in the compacted version on the right. Version numbers are adjacent to nodes.

Figure 2 depicts the main components of our tree-based structure used for authentication:

Block-level MACs: To provide file-block integrity, we store a MAC for each file block, and combine block MACs from the same file in a *MAC file*. We choose to store MACs for each file block (instead of a single

MAC for each file) to support random accesses to files. Block MACs are computed by the portal when a client writes to the file system. For providing freshness, we need to bind a unique version number to each file block every time it's updated and include the version number in the block MAC. To protect against rollback attacks (in which clients are presented with an old state of the file system), version numbers will have to be authenticated as well.

File version trees: We construct a *file version tree* per file that authenticates version numbers for all file blocks in a compressed form. Briefly, the file version tree compresses the versions of a consecutive range of blocks into a single node, storing the index range of the blocks and their common version number. File version trees are optimized for sequential access to files. For instance, if a file is always written sequentially then its file version tree consists of only one root node. An example of a compacted tree is shown in Figure 3.

Directory trees: To authenticate file-system meta-data (or the directory structure of the file system), the file-system directory tree is transformed into a Merkle tree in which every directory is mapped to a *directory subtree*. We have chosen to map our authenticated data structure onto the existing file-system tree in order to efficiently support file-system operations like delete or move of entire directories. To support directories with large number of files efficiently, we create a balanced binary tree for each directory that contains file and subdirectory nodes in the leaves, and includes intermediate, empty internal nodes for balancing. Nodes in a directory tree have unique identifers assigned to them, chosen as random strings of fixed length. A leaf for each file and subdirectory is inserted into the directory tree in a position given by a keyed hash applied to its name and its parent's identifier (to ensure tree balancing).

At the leaves of the directory tree, we insert the file version trees in compacted form, as described above. Internal nodes in the Merkle tree contain hash values computed over their children, as well as some additional information, e.g., node identifiers, their rank (defined as the size of the subtree rooted at the node), file and directory names.

Our Merkle tree supports the following operations. Clients can insert or delete file-system object nodes (files or directories) at certain positions in the tree. Those operations trigger updates of the hashes stored on the path from the inserted/deleted nodes up to the root of the tree. Deleted subtrees are added to the free list, as explained below. Clients can verify a file block version number, by retrieving all siblings on the path from the leaf corresponding to that file block up to the root of the tree. Searches of files or directories in the tree can also be performed, given absolute path names.

We also implement an operation randompath-dir-tree for directory trees. This feature is needed to execute the challenge-response protocols of the auditing component in Iris. A (pseudo)-random path in the tree is returned by traversing the tree from the root, and selecting at each node a child uniformly at random, weighted by rank. In addition, the authentication information for the random path is returned, so the tenant can verify that the path has been chosen pseudo-randomly.

With this Merkle tree construction, we authenticate both file system meta-data, as well as file block version numbers. Together with the file block MACs, this mechanism ensures data integrity and freshness, assuming that the portal always stores the root of the Merkle tree.

Free list: As an optimization, we also maintain in the data structure a *free list* containing pointers of nodes deleted from the data structure, i.e., subtrees removed as part of delete or truncate operations. The aim of the free list is to defer garbage collection of deleted nodes and support remove and truncate file-system operations efficiently. We omit further details due to space limitations.

5 Auditing protocol

The authentication mechanism in Iris presented in the previous section can be used to verify the correctness of all blocks retrieved from the file system during the course of normal operations issued by clients. A challenging question that we address in this section is how can the enterprise verify infrequently accessed blocks and detect even small amounts of corruptions spread throughout the file system. We are particularly interested in offering strong assurances to the enterprise about the correctness and availability of *the entire file system*. An important requirement is that auditing of correctness should be performed with minimal bandwidth and computation. For instance, downloading a substantial fraction of the file system to verify its correctness would not be an acceptable solution. In addition, a recovery mechanism is needed to reconstruct the original data once corruptions are detected.

Several different protocols that address to some extent this question have been proposed in the literature. PoR protocols provide strong assurances about availability of data outsourced to the cloud, and a recovery mechanism, but they have only been designed for static data (files that do not undergo modifications). PDP protocols, while supporting updates to data, ensure only *detection* of a certain amount of data corruption, but do not implement a recovery mechanism. To the best of our knowledge, our solution here is the first *dynamic PoR protocol* over an entire file system, supporting updates and providing an efficient recovery mechanism in case corruption is detected.

We start by presenting at a high level how existing PoR protocols work, and then describe the challenges of adapting these ideas to a dynamic setting. We then discuss our main insights and contributions in constructing a dynamic PoR protocol.

5.1 Static PoR protocols

In a PoR protocol, the tenant encodes a single file with an error-correcting code (ECC) and stores the encoded file in the cloud. The encoded file contains the original file and some *parity blocks*, redundant blocks computed with the ECC that are needed in recovering from corruption. To ensure correctness and availability of the data, the tenant periodically challenges the cloud for a few randomly selected file blocks, and verifies their correctness. Through this auditing protocol, the tenant can detect large-scale corruption to the file (exceeding a certain fixed threshold). Small corruptions, while not detectable through sampling, can be recovered from the redundancy embedded in the encoded file.

An important parameter in a PoR is the recovery-failure probability ρ. This is the probability, assuming that the cloud replies correctly to all challenges during an audit, that the tenant can't recover the file from the cloud's storage. The size and frequency of challenges in a PoR may be calibrated to achieve a target parameter ρ given the file size, and error-correcting code parameters.

5.2 Challenges for dynamic PoRs

The main challenge in adapting a static PoR protocol to a dynamic setting is the construction of an error-correcting code with several required properties. As a reminder, the error-correcting code is used to recover from corruptions once the auditing protocol detects missing or corrupted data at the cloud. An additional requirement our system has compared to previous PoR protocols is that it needs to recover from corruptions of both data and meta-data in the entire file system (while previous PoR protocols have been designed for single files).

Our first observation is that we can use in our system an erasure code instead of a more expensive error-correcting code. The reason is that Iris's main service is authentication of file system blocks, and, therefore,

the portal can verify the correctness of file blocks and Merkle tree nodes during recovery and determine the positions of corrupted blocks. We present the remaining challenges in achieving an efficient dynamic PoR protocol:

Challenge 1: Update efficiency The erasure code has to support updates to the file system efficiently. In particular a modification to a file block or Merkle-tree node should require the update of only a small number of parity blocks. Additionally, it would be desirable to avoid expensive Galois field arithmetic (as employed by Reed-Solomon codes, for instance) in the parity computation. Instead, efficient, binary operations (e.g., XORs) are preferable.

This requirement excludes upfront the use of maximum-distance separable (MDS) codes. While such codes are attractive for their correction capability, a parity block in an MDS codes depends on all message blocks, and therefore updates to the codeword are quite impractical.

Thus we must use a non-MDS code, with a lower error-correction capability. For instance, we might *stripe* the file system, that is, partition it into a number of smaller components, called stripes, and apply an erasure code individually to each stripe (striping is a common technique employed in most storage systems today). With this approach, updates would be more efficient as an update to a file block or Merkle tree node would involve updating only parity blocks within a single stripe.

Challenge 2: Hiding code structure Nevertheless, striping introduces a problem. When a client updates a block of the file system along with the corresponding stripe parities, it *reveals code-structure information* to the cloud, namely the correspondence between the file blocks and the parity blocks. A malicious cloud can then create a targeted corruption against the file system, e.g., it can corrupt a single stripe and its corresponding parity blocks. Such corruption, being focused, will be hard to detect by sampling (since sampling detects only a large amount of corruption).

We overcome this challenge with two key techniques:

1. *Cache parities at the portal* We cache the parity information at the enterprise side and only transmit parities to the cloud at regular time intervals for back up (e.g., at the end of the week). With this approach, the cloud does not perceive individual updates to the file system, but only the aggregate parity structure over a large number of updates and can not infer the exact code structure. Moreover, updates are extremely efficient if parities are stored in main memory at the portal.

2. *Randomize code structure* Even when parities are stored at the portal, it is important that the stripe structure is not revealed to the cloud to avoid targeted corruptions. To enforce this, we randomize the assignment of file blocks to stripes.

If these two design principles are employed, it might seem that after caching the parities locally and randomizing the assignment of file blocks and tree nodes to stripes, any erasure code could be used for computing the parity blocks within a stripe. But our system has to overcome another subtle challenge:

Challenge 3: Variable-length encoding Typically, the code parameters for an erasure code, including the message size, and the size of parity information are fixed and known in advance (before encoding is performed). However in Iris we need to compute parity blocks over an entire file system data and meta-data blocks without knowing in advance the total size of the file system. At the same time, we have to enforce a randomization of the mapping of file system blocks to parity blocks at any given time. Therefore, approaches in which new parity blocks are created as more data is added to the file system in a streaming fashion (e.g., LDPC codes) would not be applicable here.

New sparse randomized erasure code construction Our solution is to set an upper bound on the total size of the file system, and design a novel erasure code construction that is *sparse* in the sense that it supports

incremental updates to the codeword very efficiently, even when only a fraction of the maximum size is used by the file system. The construction randomizes the mapping of file system blocks to parity blocks, and uses binary XOR operations. The size of the parity information is also constrained to fit into the main memory of typical servers today (an important consideration for efficient updates). Lastly, we are able to prove for this construction an exponentially small bound for the recovery-failure probability.

5.3 Our erasure code

Parameter overview: We first set an upper bound for the entire file system size, denoted n. In our example parameterization, n is the number of 4KB blocks needed for a file system of maximum size 1PB. Our erasure code construction is scalable up to that size, but once the file system exceeds the upper bound, the code parameters need to be changed and the file system has to be re-encoded.

To correct a fraction α of erasures, the storage for parities must be at least $s \geq \alpha n$ blocks—a coding-theoretic lower bound. Here s is limited by the sizes of current memories to about $s = O(\sqrt{n})$ for practical file system sizes and thus $\alpha = O(1/\sqrt{n})$. (To obtain a probabilistic guarantee that at most an α-fraction of all stored file blocks is missing or corrupted, the tenant must challenge $c = O(1/\alpha) = O(\sqrt{n})$ randomly selected file blocks.)

To support updates efficiently we split the huge codeword into $m \approx \alpha n$ stripes; each stripe being a codeword itself with p parities. With high probability, given an α-fraction of erasures, each stripe is affected by only $O(\log n)$ erasures. Thus to correct and recover stripes, we need $p = O(\log n)$ parity blocks per stripe, leading to $s = O(\alpha n \log n) = O(\sqrt{n} \log n)$ memory. Each write only involves updating $u = O(\log n)$ parities within the corresponding stripe. By using a sparse parity structure, though, we are able to reduce u to $O(\log \log n)$.

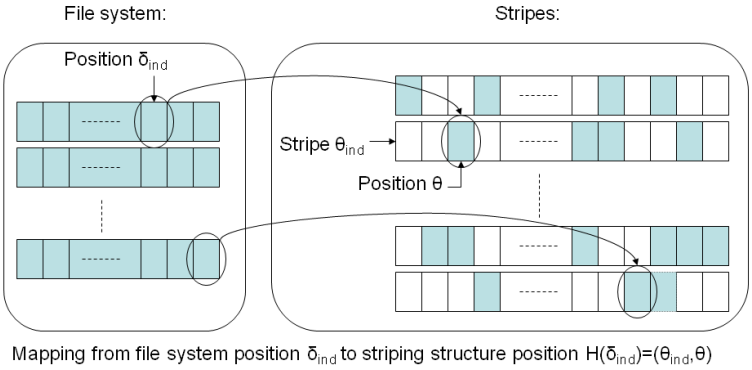

Figure 4: Randomized mapping of blocks in the file system to erasure code stripes.

Details on our erasure-code construction: Our erasure code is a sparse one based on efficient XOR operations. Although the new construction is probabilistic in that successful erasure decoding is not guaranteed for any number of erasures, its main advantage is that it is a binary efficient code scalable to large codeword lengths.

The portal computes parities over both file blocks and Merkle tree nodes when block values are updated by a client operation. For the purpose of erasure coding, we view data blocks or tree nodes as identifier-value pairs $\delta = (\delta_{id}; \delta_{val})$, where δ_{id} is a unique identifier (a unique block ID in the file system) and $\delta_{val} = (\delta_1, \ldots, \delta_b)$ is a sequence of b bits denoting the change in block value. (We assume all blocks are initialized

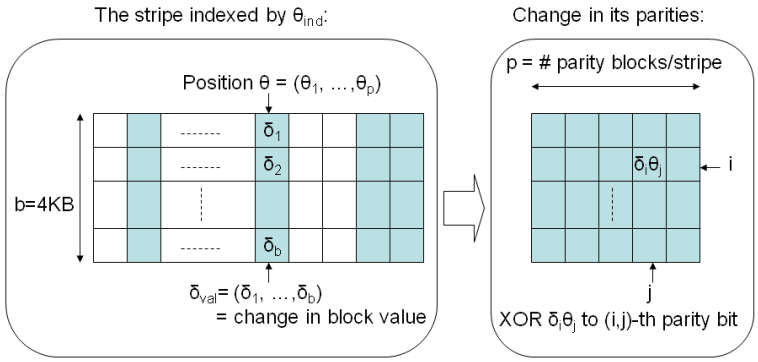

Figure 5: New erasure code construction.

with 0.) To randomize the mapping from data blocks to parity blocks, we use a keyed hash function $H_k(.)$ that maps an identifier δ_{id} to a pair (θ_{ind}, θ), where θ_{ind} is a random stripe index and $\theta = (\theta_1, \ldots, \theta_p)$ is a binary vector of p bits. The randomization is graphically depicted in Figure 4.

The 1s in vector θ indicate the parity bits that need to be updated. Each update modifies at most u of the p parities of the stripe to which δ belongs. That is, $H_k(\delta_{id})$ is designed to produce a binary random vector θ of length p with at most u entries equal to 1. For $u = O(\log p) = O(\log \log n)$ this leads to a sparse erasure code that still permits decoding, but entails relatively few parity updates.

Encoding: We maintain a parity matrix $P[i]$ for each stripe i, $1 \leq i \leq m$. To change the value of block δ_{id} with δ_{val}, the portal computes $H_k(\delta_{id}) = (\theta_{ind}; \theta)$; constructs $A = \delta_{val} \otimes \theta = \{\delta_i \theta_j\}_{i \in [1,b], j \in [1,p]}$; and updates $P[\theta_{ind}] \leftarrow P[\theta_{ind}] \oplus A$. The change in parity structure is shown graphically in Figure 5.

Since vector θ has at most u non-zero positions, the number of XOR operations for updating a block is u. The total storage for all parities is $s = bpm$ bits.

Decoding: Erasure decoding of the multi-striped structure involves decoding each stripe separately. Gaussian elimination is performed m times, each time computing the right inverse of a $(\leq p) \times p$ matrix–at a cost of at most $p^2 = O((\log n)^2)$ XOR operations. As an additional benefit of our construction, decoding can be done in place, and thus within memory at the portal.

Analysis: Let n denote the maximum number of blocks in the file system. On the assumption of 4KB-sized blocks, each file block stores $b = 2^{15}$ bits and the file system's total possible *storage* equals nb bits ("storage" denotes the total file-system size).

The sparse erasure code has m stripes, each stripe has p parity blocks, and thus the total amount of *memory* needed at the portal in order to store all parity blocks is equal to $s = mpb$. Here, "memory" refers to the size of main memory at the portal devoted to parity storage.

Suppose that the portal issues $c \leq n$ random challenge-response pairs. (We require the technical condition that the challenges contain c/m blocks within each of the m stripes.) Each response includes a Merkle-tree path as well as a data block. Merkle-tree paths can vary in length, but are no more than 1KB for reasonable file-system parameterizations. For this reason we assume an upper bound of 5KB communication per response, i.e., $c \cdot 5$KB *verification bandwidth*.

Our goal is to compute the *recovery failure probability* ρ, defined as the probability that, given c random challenge-response pairs erasure decoding fails despite all challenge-response pairs verifying correctly. The following theorem provides a tight upper bound on ρ. We will use this bound to compute an upper bound on the verification bandwidth in the recovery failure probability ρ, the block size b, the file system's storage

Table 1: Examples for recovery failure probability $\rho \leq 0.0074$, i.e., $\leq 0.74\%$, block size 4KB and 5KB communication per response.

	Theorem 1(i)		Theorem 1(iii)	
Memory/Storage	p	c	p	c
4GB/1TB	175*	$2^{14.2} \rightarrow 94$MB	1667	$2^{12.3} \rightarrow 24$MB
4GB/10TB	169	$2^{17.5} \rightarrow 0.9$GB	1781	$2^{15.6} \rightarrow 0.23$GB
16GB/100TB	186	$2^{19.0} \rightarrow 2.5$GB	1965	$2^{17.0} \rightarrow 0.63$GB
16GB/1000TB	196	$2^{22.3} \rightarrow 25$GB	2014	$2^{20.2} \rightarrow 5.8$GB

nb and memory at the portal mpb (Table 1).

We remind the reader that in the sparse erasure code, when updating a file block, at most u out of the p parity blocks of the stripe to which the block belongs need to be updated. The sparse code is completely defined by the number of stripes m, the number of parities per stripe p and parameter u (indicating the "sparsity" of the code).

As a technical preliminary, for integers p and u and $0 \leq \beta \leq 1$, let $R(p, u, \beta)$ be the probability that a binary $\beta p \times p$ rectangular sparse matrix in which each entry is chosen independently and at random to be 1 with probability u/p does not have full rank. In the purely random case $u = p/2$ and $R(p, u = p/2, \beta) \leq 2^{-(1-\beta)p}$. Based on extensive simulation and literature, [31] states the conjecture that for $u > 2 \ln p$ and β sufficiently close to 1, $R(p, u, \beta) \approx R(p, p/2, \beta) \leq 2^{-(1-\beta)p}$. One of our bounds is based on this conjecture.

Theorem 1 *In order to achieve a recovery failure probability $\rho \leq 3 \cdot e^{-l}$ with $e = 2.718$ for some $l \geq 1$ we can use the following parameter settings:*

(i) $u = p/2$, $s/b \leq 2.0 \cdot \sqrt{np}$ and

$$p \geq 4.6 \cdot (l + \ln(1.24 \cdot n) + \ln(s/(pb))),$$
$$c = 5.1 \cdot (nb/s) \cdot (l + \ln(s/(bp))).$$

(ii) $u = p/2$ and

$$p \geq 4.6 \cdot (l + \ln(1.27 \cdot \frac{n(n-c)}{\sqrt{c(3n+c)}}) + \ln(s/(2pb))),$$
$$c = 5.1 \cdot (nb/s) \cdot (l + \ln(s/(bp))).$$

(iii) $u > 2 \ln p$ with $R(p, u, 0.972) \leq 2^{-(1-0.972)p}$ (e.g., $u = p/2$), $s/b \leq 2.0 \cdot \sqrt{np}$ and

$$p \geq 51.45 \cdot (l + \ln(1.71 \cdot n) + \ln(s/(pb))),$$
$$c = 1.54 \cdot (nb/s) \cdot (l + \ln(s/(bp))).$$

Example parameters: For a selection of four example system parameters, Table 1 lists parities-per-stripe p and number of PoR challenges c (together with their corresponding verification bandwidth). This example is parameterized under a recovery failure upper bound of 0.74% (corresponding to technical parameter $l = 6$ in Theorem 1). Values in the left column are based on Theorem 1(i) and values in the right column are based on Theorem 1(iii). For the entry labeled $*$, Theorem 1(i) yields $p = 159$, which does not satisfy the condition $s/b \leq 2.0 \cdot \sqrt{np}$. In this case we need to use Theorem 1(ii).

If we assume $R(p, u, 0.972) \leq 2^{-(1-0.972)p}$ for $u > 2 \ln p = 2 \ln 2000 \approx 15.2$, then for the right column the values also hold if each block affects an expected $u = 16$ parity blocks out of the p parity blocks of the stripe to which it belongs. Thus it suffices that each file-block update induces only an expected $u = 16$ operations (XORs) over the parity structure. Compared to the left column where $u = p/2$, this is a factor 5 improvement.

Remark: Notice that p and c are relatively independent of l; e.g., in the left column slightly increasing c with $5.1 \cdot nb/s$ (increasing the verification bandwidth by at most several percent) and by adding ≈ 5 more parities to p, decreases ρ by a factor e.

5.4 Deployment of erasure-coding in the Iris auditing protocol

We now explain how our erasure code functions in Iris.

PoR encoding and update: During encoding, the portal constructs two parity structures: the *data parity structure* constructed over the file-system data blocks (including the data blocks in the free list) and the *meta-data parity structure* over the meta-data blocks (internal nodes in the data structure comprising the Merkle tree and free list).

The challenge-response protocol: The portal challenges the cloud to return a set of c (again $c = O(\sqrt{n})$ randomly selected file-system data blocks, including data blocks from the free list. These blocks are all leaf nodes in the authenticated data structure containing the Merkle tree and free list. As an optimization, the portal sends a seed from which the challenge set is pseudo-randomly derived.

The c selected random blocks together with the authenticating paths from the authenticated data structure are transmitted back to the portal. The portal verifies the correctness of the responses by performing two checks. First, it verifies the integrity and freshness of the selected blocks, checking the block MACs and the path to the root in the authenticated data structure. Second, it verifies that the blocks have been correctly indexed by the challenges according to the node ranks/weights. (This proves that the file-system data blocks are selected with uniform probability.) As a byproduct of these checks the challenge-response protocol also verifies the integrity and freshness of the meta-data blocks (internal nodes in the authenticated data structure). We can immediately infer that if a fraction α of file-system data blocks don't verify correctly, then at most a fraction α of internal nodes in the Merkle tree and free list are either missing or corrupted.

Recovery: If the portal ever receives an incorrect response to a PoR challenge (presumably a very rare event), it can trigger recovery of the file system in a streaming fashion. Assuming that the portal does not have enough storage to download the full file system, it can download and recover parts of the file system and stream the corrected version to a new provider. For simplicity of description, we assume full recovery by the portal itself.

The recovery proceeds in two steps. The portal first needs to decode the Merkle tree and free list structure containing information about the directory structure of the file system. The portal iterates through the tree in pre-order, and verifies the authenticity of each tree node. Validated nodes are subtracted from the meta-data parity structure. While iterating through the tree, the portal creates an authenticated list of the identities of correctly verified nodes. This list represents a connected subtree rooted at the root of the authenticated data structure. (A corrupted internal node can not have descendants in its subtree that verify correctly.)

Omissions from the list correspond to erasures. In order to recover these erasures, the portal sorts the list according to node identities. Since the portal stores a counter indicating the range of all node identities that have been used, the portal can retrieve, by reading out the sorted list, the identities of all erasures and store these in its memory. Using the decoding algorithm of the erasure code, the erased tree nodes can

be recovered from the parity blocks over the authenticated data structure. The recovered tree nodes are streamed to the new provider.

Second, after the complete tree structure has been recovered, the portal verifies the integrity and freshness of file-system data blocks. The portal requests blocks in a standard traversal order, and marks as erasures the blocks whose integrity is not verified. Each correct block is subtracted from the parity structure computed over file-system data blocks. Block identifiers of missing or corrupted file blocks can be retrieved from the file-version tree within the recovered authenticated data structure. The portal stores the block identifiers of missing or corrupted file blocks in its memory. Using the decoding algorithm of the erasure code, the portal recovers all the corrupted file blocks from the parity blocks computed over the file-system data blocks.

6 Implementation

Our implementation of Iris is a 25,000-line end-to-end system with all integrity checking in place. The system is fully asynchronous and never holds a lock while waiting for network or disk I/O operations. The code runs in user space as a transparent layer that can take advantage of any existing storage system at the cloud provider. Our implementation uses the open-source .NET framework Mono, which is advantageously platform-independent: Iris can run on Linux, Windows, and MAC OS.

Our implementation includes the Portal, a simple Cloud storage server, and clients that run traces and benchmarks, as depicted in the detailed system architecture in Figure 1.

6.1 Cloud

The cloud stores not only regular file system data, but also authenticating meta-data, including MAC files and our Merkle tree authenticated data structure and checkpointed parities needed for recovery. The repositories for these data types are shown at the top of Figure 1.

The portal performs reads and writes to the various data repositories by invoking their respective cloud-side services. The *Cloud File System Service* handles requests for file blocks, MAC files, and the Merkle tree (stored in our implementation in an NTFS file system). Operations on file blocks are executed asynchronously at the portal. Sequential access operations to the same file can potentially arrive out of order at the cloud. (Re-ordering can occur in transit on the network, as our portal and cloud machines are each equipped with three network cards.) To reduce disk spinning, the Cloud File System Service orders requests to the same file in increasing order by block offset.

6.2 Portal

The portal interacts with multiple clients. Clients issue file system calls to the *Portal Service*, shown at the bottom of the portal component in Figure 1. The portal executes client operations in parallel: Each operation is executed in a thread pool as a user-scheduled task with asynchronous steps. When an operation is waiting for a long running step such as disk and network I/O, the task is paused and the current thread switches to another task. This allows thousands of simultaneously active operations to be handled by the thread pool with a small number of threads. In our setup, the thread pool had 16 threads—one for each virtual CPU core, for maximum parallelism.

Operations don't interact directly with the cloud, but instead with the Merkle Tree and Block Caches. All data and meta-data requested by the caches is downloaded from the cloud via the *Storage Interface* in

the portal, shown in the middle of the portal in Figure 1. While in use by an active operation, blocks and nodes are retained in the cache. Prior to being cached, however, blocks and nodes downloaded from the cloud are checked for integrity by the *Integrity Checker* components.

Our implementation benefits from multi-core functionality available in most modern computers. Operations performed on active blocks in the cache are split into atomic operations (e.g., hash update for a tree node, check MAC for a data block or compact nodes in file version trees). These are inserted into various priority queues maintained at the portal. Multiple threads seize blocks from these queues, lock them and execute the atomic operations. Operations are always started in order, but may complete out of order. However, our implementation ensures that the effect of the operations on the system is the same as if they were executed by a single thread in order.

6.2.1 Merkle tree cache

The Merkle Tree Cache in the portal is Iris's most complex component. Much of the design effort and complexity of Iris lies in the caching strategy for recently accessed portions of the tree. We designed a generic, efficient Merkle Tree Cache that ensures consistency across thousands of simultaneous asynchronous client operations.

When an operation accesses the cache, it first locks it using a mutex and unlocks it when its done. All of the operations are designed such that they access the cache for a very short period of time for tasks such as changing the value of a few fields of a Merkle tree node. To ensure a high degree of parallelism, the Merkle tree mutex is never locked while an operation waits for a long running step such as network and disk I/O.

When executing operations in parallel, a real challenge is to handle dependencies among tree nodes and maintain data structure consistency and integrity. We do this by imposing several orderings of operations. Nodes are brought into the cache in a top-down order and are evicted in a bottom-up order. The top-down ordering is necessary because when a node is read from the untrusted storage, it can only be verified once all of its ancestors have also been cached in and verified. Likewise, a node can only be written out to the untrusted storage after the hash of its subtree has been computed. If multiple nodes in a sub-tree are modified, the Merkle Tree Cache will only hash the shared path to the root once, thereby significantly reducing the number of hashes that need to be performed.

Phases. To enforce the ordering, each node is always in one of the following phases: Reading, Verifying, Neutral, Compacting, UpdatingHash, or Writing. A node always traverses these phases in order and only after its parent or children have reached a certain phase. For example, a node only enters the verifying phase after its parent has completed the verifying phase. The Reading and Verifying phases are applied top-down and the Compacting, UpdatingHash, and Writing phases are applied bottom-up. When a node is in the Neutral phase, it is in the cache and available to be used by operations.

Pinning. Operations oftentimes need to access multiple nodes. For example, a WriteFile operation needs to access the path in the version tree that descends all the way to the version node corresponding to a specific block. The operations first *pin* all of the nodes they need and then proceed to execute. If a node is pinned that is not currently in the cache, the operation is paused and resumed when all of its pinned nodes have been loaded into the cache. Once a node is pinned, it is not cached out until it is unpinned (e.g., when the operation completes). A node may be pinned multiple times, in which case it must be unpinned the same number of times until it is considered in the unpinned state and may be cached out.

If a node is pinned, its ancestors, sibling, and siblings of the ancestors are automatically *indirectly* pinned. This is necessary because if the node is modified, the indirectly pinned nodes will be needed when updating the hashes of the path to that node.

Eviction. When the cache reaches its maximum allowed size, it repeatedly evicts least-recently-used (LRU) leaf nodes, causing a bottom-up wave of evictions. Evicting a node consists of transitioning its phase from the Neutral to Compacting. The node then goes through the UpdatingHash and Writing phases until it is finally removed from the cache. If a node and its subtree were not modified, then the UpdatingHash and Writing phases are skipped.

6.2.2 Other components

The Block Cache functions much like the Merkle Tree Cache except that blocks don't have parents/children so there are no dependencies between blocks.

The Merkle Tree and Block Caches keep track of two items per node/block: The old and new data. The old data is the value of the node/block when it was fetched from the cloud. The new data is its value after it was (possibly) modified by an operation. When a node/block is evicted, the portal computes the difference of the byte representations of the old and new data and updates the parities.

Another component of the portal is the auditing module. This service, periodically invoked by the portal, transmits a PoR challenge to the cloud and receives and verifies the response, consisting simply of a set of randomly selected data blocks in the file system and their associated Merkle tree paths. The portal also maintains a repository of *Parities* to recover from file-system corruptions detected in a PoR, seen in the portal cache module in Figure 1. Parities undergo frequent modification: Multiple parities are updated with every file-block write. Thus, the Parities repository sits in the main memory of the portal.

The portal can include a checkpointing service that backs up data stored in the main memory at the portal to local permanent storage. To enable recovery in the advent of a portal crash, checkpointed data can be periodically transmitted to the cloud (with a MAC for integrity). While we have not implemented this component, it can rely on well known checkpointing techniques.

7 Experimental evaluation

We ran several experiments to test different aspects of Iris. We first describe our setup and then present our results. Two machines ran the full end-to-end system implementation described in Section 6: The Portal and the Cloud.

Portal Computer. The Portal computer has an Intel Core i7 processor and 12 GB of RAM. The experiments were run on Windows 7 64-bit installed on a rotational disk, but no data was written to the Portal's hard drive for the purpose of our experiments.

Cloud Computer. The Cloud computer has seven rotational hard drives with 1TB of storage each. The file system and MAC files reside on these disks. The disks are used as separate devices and are not configured as a RAID array. This configuration mimics a cloud where each disk could potentially be on a separate physical machine. The operating system (Windows 7 64-bit) runs on an separate additional hard drive to avoid interfering with our experiment.

Networking. Because our file system can handle very large throughput, we used three 1Gbps cables to connect the two computers. Each computer had one network port on the motherboard and two additional network cards. After accounting for networking overhead, the 3 Gbps combined connections between the two computers can handle about 280 MB/s of data transfer as our experiments show.

Configuration. In our configuration, write operations originate from clients (simulated as threads on the Portal). Then they are processed by the Portal and multiplexed over the three network connections. Finally,

data reaches the Cloud computer and is written to the appropriate disk. Reads are similarly processed, but the data flow is in the opposite direction (from the Cloud machine to the Portal).

Simulated Latency. To obtain more realistic results, we deliberately simulated 20ms round-trip time (RTT) latency between the clients and Portal, and 100ms RTT latency between the Portal and Cloud. This setting aims to resemble the scenario where the clients and Portal are both part of the same corporate network and the Cloud is a data center located elsewhere on the same continent.

7.1 Workloads

To evaluate Iris, we used the following workloads. Each workload was recorded as a trace and played back exactly under different parameterizations of our system.

- **Tar/Untar** (directory structure): A workload to measure the performance of operations that access and modify a realistic directory structure. The tarball consists of the source code for the entire Linux kernel (about 420 MB, 37,000 files, and 2,300 directories).
- **IOZone** (various file access patterns): This workload tested the performance of combining various access patterns such as read, write, reread/rewrite, random read/write, backwards read, and strided read. We used the popular I/O benchmarking tool IOZone [1].
- **Sequential Read/Write** (throughput): Measures the performance of sequentially reading/writing ten files simultaneously, each of size 10 GB.
- **Random Read/Write** (seeks): Measures the performance of randomly reading/writing ten files simultaneously, each of size 1 GB. Reads and writes are uniformly random, and trigger seeks with almost every operation. For the random read workload, the file is first randomly written and then only the random read portion of the trace is benchmarked.

7.2 Results

Our experimental results show how Iris performs under the above workloads on the full end-to-end system described in Section 6. We note that even with seven hard drives for storage and three 1 Gbps network links between the Portal and Cloud, under no workload was the Portal the bottleneck. Depending on the workload, the limiting factor was either the network or hard drives.

Varying the Merkle Tree Cache Size: The parallel Merkle Tree Cache is crucial for the performance of our system. The cache allows the Portal to perform file operations without having to read and write entire Merkle tree paths from the server for each operation. The asynchronous cache also allows for pausing operations that are waiting to retrieve Merkle tree nodes while other operations actively use the cache.

Multiple paths can be loaded into the Merkle Tree Cache at once while maintaining consistency. In order to demonstrate the usefulness of the cache, in this experiment we varied its size (i.e., how many nodes it can hold at once) and we timed each of the workloads under different cache sizes. The results are in Figure 6.

Interpretation: As demonstrated in the figure, the Tar, Untar, and IOZone workloads greatly benefit from having a Merkle tree cache of 5 to 10 MB (about 10,000 to 20,000 nodes), whereas the sequential and random read/write workloads are mostly unaffected by the cache size.

The reason is quite simple: The Tar, Untar, and IOZone benchmarks frequently revisit the same part of the Merkle tree. For example, the Tar/Untar workloads often read/write multiple files within the same

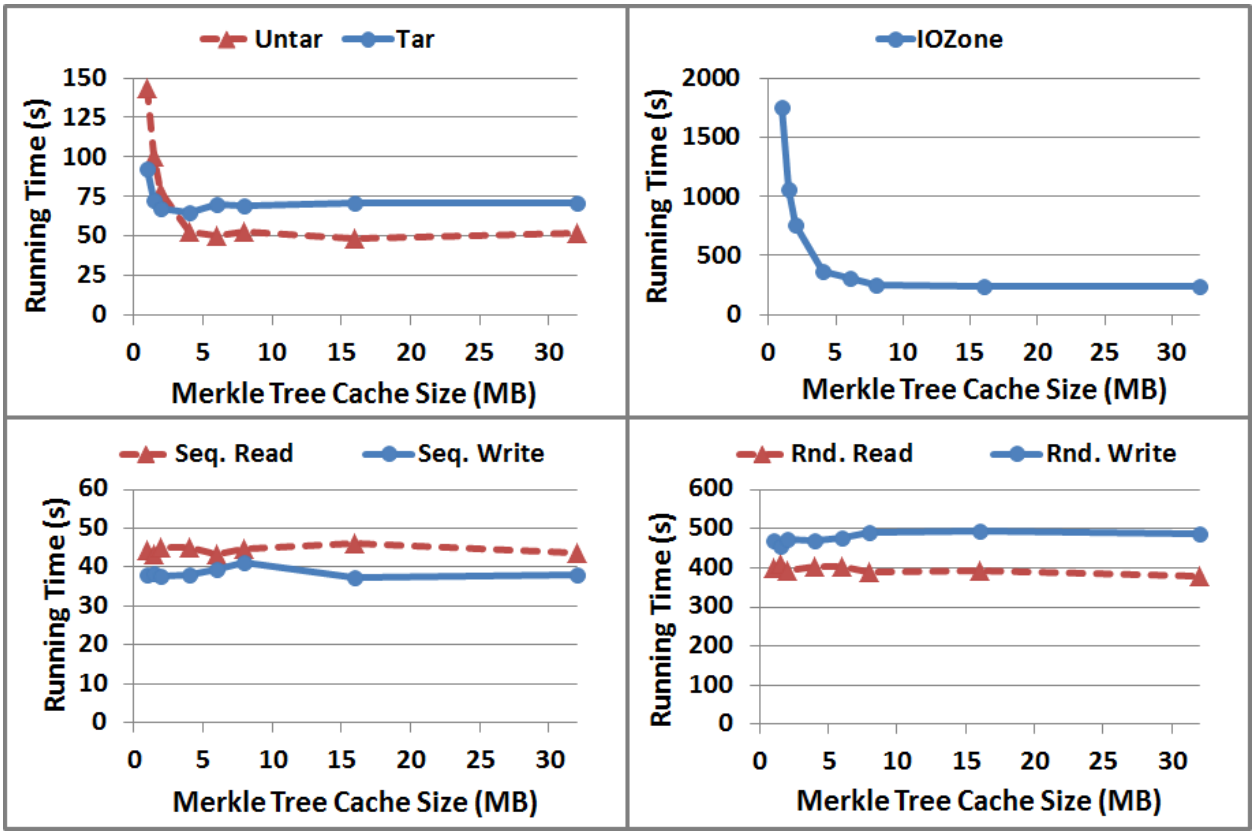

Figure 6: Workloads under different Merkle Tree Cache sizes. In each plot, the horizontal axis is the Merkle Tree Cache size (in MB) and the vertical axis is the time (in seconds) for the workload to complete.

directory (and hence their Merkle tree paths share many nodes). Likewise, the random write portion of the IOZone benchmark creates a file with a large uncompactable Merkle tree which is then read sequentially and the sequential read portion of the workload yields an in-order traversal of the Merkle tree that is significantly sped up by the cache.

On the other hand, the sequential read and write workloads generate version tree nodes that are quickly compacted. Hence the Merkle Tree Cache only needs to hold a few dozen nodes at a time. The random read/write workloads are extremely intensive on the Cloud's disks. Almost every operation causes a seek, so the Cloud's disks are the bottleneck. Because the random read/write operations are executed very slowly by the Cloud's disks and the Portal parallelizes requests for the Merkle tree nodes, there is plenty of time for the Merkle tree nodes to be fetched without delaying the workload.

Scalability: In Iris, all operations are handled by the same thread pool and each file has its own queue of pending/active operations. From the Portal's perspective, there is little difference between each operation being issued by a different client and all operations being issued by the same client. Most of the overhead of having multiple clients comes from having to manage multiple TCP sockets and their associated buffers.

We wanted then to show that Iris can easily scale to 100 clients accessing it simultaneously. To maximize both strain on the Portal's CPU and the number of cryptographic operations performed, each client generated a sequential access pattern. (With more seek-intensive access, the bottleneck would be disk seeks on the Cloud.)

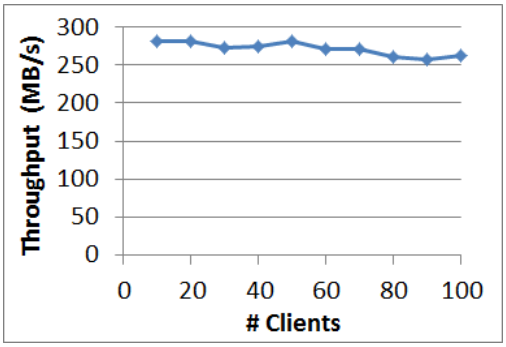

Figure 7: Avg sequential read & write speed.

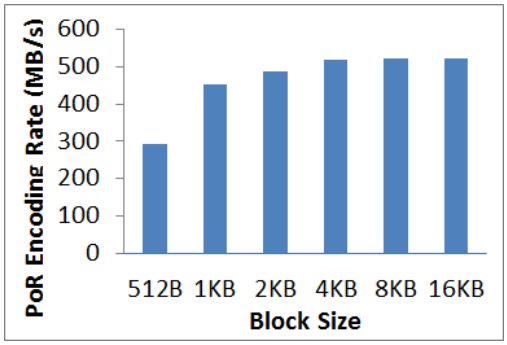

Figure 8: PoR Encoding Rate.

	Total Latency (ms)		Network I/O		Cloud Disk I/O		Portal Processing	
Portal Cache:	Hot	Cold	Hot	Cold	Hot	Cold	Hot	Cold
Create file in directory of depth 0	20.0	20.0	20.0	20.0	0.0	0.0	0.0	0.0
Create file in directory of depth 1	20.0	144.0	20.0	120.0	0.0	9.6	0.0	14.4
Create file in directory of depth 2	20.0	254.0	20.0	220.0	0.0	16.5	0.0	17.5
Create file in directory of depth 3	20.0	363.0	20.0	320.0	0.0	23.4	0.0	19.6
List directory with 10 files at depth 1	27.7	678.9	20.0	620.3	0.0	32.6	7.7	26.0
Write 1 MB file at depth 1, wait completed	24.8	138.8	20.0	120.0	0.0	0.0	4.8	18.8
Read 1 MB file at depth 1	20.0	284.2	20.0	220.0	0.0	43.7	0.0	20.5

Figure 9: Latency for different operations in Iris.

We averaged the sequential read and sequential write speeds for 10 to 100 clients. Figure 7 shows the results. As can be seen, Iris consistently reads/writes at 250 MB/s to 280 MB/s. The slight performance degradation for 100 clients is due to the fact that many files are accessed at once and that causes a larger portion of disk seeks.

Latency: Figure 9 shows the latency for several basic operations in Iris. The latency is measured under two scenarios: when the portal cache is hot and cold. A hot cache means that the cache already contains all of the data (Merkle tree nodes and blocks) necessary to perform the operation on the portal alone. A cold cache means that all of the data has been evicted from the portal's cache.

The bulk of the latency (over 84%) comes from the portal-cloud and client-portal network latencies. Our results show that the latency introduced by the portal for integrity checking and cache management (denoted as portal processing time) is much smaller in comparison: less than 14% for a cold cache and less than 29% for a hot cache.

The 1 MB read operation takes about half of the time of the 1 MB write operation because the portal notifies the client that the write operation has completed while uploading the file to the cloud in the background. For the read operation, the portal must first read the file from the cloud.

The high cold cache latency for high depth operations (e.g., create depth 3 and list directory) is due to the fact that each file is represented as a separate node in the Merkle tree and tree paths are fetched one node at a time. It should be noted that this latency can be significantly reduced by having the portal fetch all nodes in a path in parallel or grouping multiple files into a single file node.

PoR Encoding Rate: Finally, we measure the rate at which the Portal can perform erasure-encoding for file-system recovery if auditing detects corruption. Figure 8 shows encoding speeds for data blocks of different sizes.

8 Conclusions

We have presented Iris, an authenticated file-system designed to outsource enterprise-class file systems to the cloud. Iris goes beyond basic data-integrity verification to achieve two stronger properties: File freshness and retrievability. Using a lightweight, tenant-side portal as a point of aggregation, Iris efficiently processes asynchronous requests from multiple clients transparently, i.e., with no underlying file-system interface changes.

Iris achieves a degree of end-to-end optimization possible only through a carefully crafted, holistic architecture, one of the systems's major contributions. Iris's architecture also relies on several technical novelties: The authenticating data-structure design and management, caching techniques, sequential-file-access optimizations, and a new erasure code enabling the first efficient dynamic PoR.

In practice, a common impediment to security-system deployment is performance overhead. It is our hope to see Iris become the first authenticated file-system to break through this barrier thanks to a combination of strong integrity assurances with high performance.

References

[1] IOzone filesystem benchmark. www.iozone.org. 2011.

[2] `www.memcached.org`.

[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. Usenix, 2002.

[4] G. Ateniese, R. Burns, R. Curtmola, J. Herring, L. Kissner, Z. Peterson, and D. Song. Provable data possession at untrusted stores. In *14th ACM CCS*, pages 598–609, 2007.

[5] M. Blaze. A cryptographic file system for Unix. In *Proc. First ACM Conference on Computer and Communication Security (CCS 1993)*, pages 9–16, 1993.

[6] K. Bowers, A. Juels, and A. Oprea. Proofs of retrievability: Theory and implementation. In *Proc. ACM Cloud Computing Security Workshop (CCSW 2009)*, 2009.

[7] G. Cattaneo, L. Catuogno, A. Del Sorbo, and P. Persiano. The design and implementation of a transparent cryptographic file system for Unix. pages 199–212, 2001.

[8] Y. Chen and R. Sion. To cloud or not to cloud? musings on costs and viability. In *ACM Symposium on Cloud Computing (SOCC)*, 2011.

[9] Y. Dodis, S. Vadhan, and D. Wichs. Proofs of retrievability via hardness amplification. In *Proc. 6th IACR TCC*, volume 5444 of *LNCS*, pages 109–127, 2009.

[10] C. Erway, A. Kupcu, C. Papamanthou, and R. Tamassia. Dynamic provable data possession. In *Proc. ACM Conference on Computer and Communications Security (CCS 2009)*, 2009.

[11] A. J. Feldman, W. P. Zeller, M. J. Freedman, and E. W. Felten. Sporc: Group collaboration using untrusted cloud resources. In *Proc. OSDI*, 2010.

[12] K. Fu. Group sharing and random access in cryptographic storage file systems. Master's thesis, Massachusetts Institute of Technology, 1999.

[13] K. Fu, F. Kaashoek, and D. Mazieres. Fast and secure distributed read-only file system. *ACM Transactions on Computer Systems*, 20:1–24, 2002.

[14] R. Geambasu, J. P. John, S. D. Gribble, T. Kohno, and H. M. Levy. Keypad: An auditing file system for theft-prone devices. In *Proc. European Conference on Computer Systems (EuroSys)*, 2011.

[15] E. Goh, H. Shacham, N. Modadugu, and D. Boneh. SiRiUS: Securing remote untrusted storage. In *Proc. Network and Distributed Systems Security Symposium (NDSS 2003)*, pages 131–145, 2003.

[16] M. T. Goodrich, C. Papamanthou, R. Tamassia, and N. Triandopoulos. Athos: Efficient authentication of outsourced file systems. In *Proc. Information Security Conference 2008*, 2008.

[17] A. Juels and B. Kaliski. PORs: Proofs of retrievability for large files. In *Proc. ACM Conference on Computer and Communications Security (CCS 2007)*, pages 584–597, 2007.

[18] M. Kallahalla, E. Riedel, R. Swaminathan, Q. Wang, and K. Fu. Plutus: Scalable secure file sharing on untrusted storage. In *Proc. 2nd USENIX Conference on File and Storage Technologies (FAST)*, 2003.

[19] S. Kamara, C. Papamanthou, and T. Roeder. Cs2: A searchable cryptographic cloud storage system. Technical Report MSR-TR-2011-58, Microsoft, 2011.

[20] J. Li, M. Krohn, D. Mazieres, and D. Shasha. Secure untrusted data repository. In *Proc. 6th Symposium on Operating System Design and Implementation (OSDI)*, pages 121–136. Usenix, 2004.

[21] P. Mahajan, S. Setty, S. Lee, A. Clement, L. Alvisi, M. Dahlin, and M. Walfish. Depot: Cloud storage with minimal trust. In *Proc. OSDI*, 2010.

[22] E. Miller, D. Long, W. Freeman, and B. Reed. Strong security for distributed file systems. In *Proc. 1st USENIX Conference on File and Storage Technologies (FAST)*, 2002.

[23] A. Oprea and M. K. Reiter. Integrity checking in cryptographic file systems with constant trusted storage. In *Proc. Usenix Security Symposium 2007*, 2007.

[24] R. Pletka and C. Cachin. Cryptographic security for a high-performance distributed file system. In *Proc. 24th IEEE Conf. on Mass Storage Systems and Technologies (MSST 2007)*, 2007.

[25] R. A. Popa, J. Lorch, D. Molnar, H. J. Wang, and L. Zhuang. Enabling security in cloud storage SLAs with CloudProof. In *Proc. 2011 USENIX Annual Technical Conference (USENIX)*, 2011.

[26] H. Shacham and B. Waters. Compact proofs of retrievability. In *Proc. ASIACRYPT*, volume 5350 of *LNCS*, pages 90–107, 2008.

[27] A. Shokrollahi. Ldpc codes: An introduction. *Digital Fountain, Inc., Tech. Rep*, page 2, 2003.

[28] A. Shraer, C. Cachin, A. Cidon, I. Keidar, Y. Michalevsky, and D. Shaket. Venus: Verification for untrusted cloud storage. In *Proc. Workshop on Cloud Computing Security.*, 2010.

[29] P. Stanica. Good lower and upper bounds on binomial coefficients. *Journal of Inequalities in Pure and Applied Mathematics*, 2:Article 30, 2001.

[30] C. A. Stein, J. H. Howard, and M. Selzer. Unifying file system protection. In *Proc. USENIX Annual Technical Conference*, 2001.

[31] C. Studholme and I. F. Blake. Random matrices and codes for the erasure channel. *Algorithmica*, 56:605–620, 2010.

[32] Q. Wang, C. Wang, J. Li, K. Ren, and W. Lou. Enabling public verifiability and data dynamics for storage security in cloud computing. In *Proc. 14th European Symposium on Research in Computer Security (ESORICS 2009)*, 2009.

[33] Q. Zheng and S. Xu. Fair and dynamic proofs of retrievability. In *Proc. 1st ACMM Conference on Data and Application Security and Privacy (CODASPY)*, 2011.

A Detailed analysis of erasure code construction

We will first explain and analyze the new erasure code together with its encoding and decoding for a single stripe, i.e., $m = 1$. Next we will generalize and analyze multiple stripes, we will prove a tight bound on the recovery failure probability ρ and compare our bounds with our analysis of a single stripe. Finally, we apply the bound to derive the theorem used in our analysis for practical example parameters (Theorem 1).

A.1 Single Stripe ($m = 1$)

Encoding: For the purpose of erasure coding, we view data blocks or tree nodes as identifier-value pairs $\delta = (\delta_{id}; \delta_{val})$, where $\delta_{val} = (\delta_1, \ldots, \delta_b)$ is a sequence of b bits δ_1, ..., δ_b and δ_{id} is a unique identifier. To randomize the mapping from data blocks to parity blocks, we use a keyed hash function $H_k(.)$ that maps an identifier δ_{id} to a binary pseudo random vector $\theta = (\theta_1, \ldots, \theta_p)$ of p bits.

The parity structure is represented by a $b \times p$ binary matrix P. Initially, no blocks are encoded into P and P equals the all-zero matrix. We describe two operations; adding a block δ into P and subtracting a block δ from P. To add δ, the client computes $H_k(\delta_{id}) = \theta = (\theta_1, \ldots, \theta_p)$, constructs the $b \times p$ matrix $A = \delta_{val} \otimes \theta = \{\delta_i \theta_j\}_{i,j}$ and updates the parity structure to $P \leftarrow P + A$. If the portal subtracts δ, then A is subtracted from P, that is, $P \leftarrow P - A$, or equivalently $P \leftarrow P + A$, since addition is modulo 2.

Notice that P corresponds to blocks in B if

$$P = \sum_{\delta \in B} \delta_{val} \otimes H_k(\delta_{id}). \tag{1}$$

Let B_{val} be the matrix for which the columns correspond δ_{val}, and let B_{id} be the matrix for which the rows correspond to $H_k(\delta_{id})$. Then (1) can be rewritten as

$$P = B_{val} B_{id}. \tag{2}$$

Notice that B_{val} has b rows and B_{id} has p columns. Since $H_k(.)$ is a keyed hash function, B_{id} is a *pseudo random* binary matrix. Finally, notice that the parity structure P is a binary $b \times p$ matrix which requires $s = b \cdot p$ bits storage.

Decoding: We assume that the portal retrieves complete blocks from the server for which the correctness can be verified (blocks are part of an authenticated data structure which can be used to verify the integrity and freshness of blocks). Those blocks that verify correctly are subtracted from P. Let I be the set of all block identities of blocks that are corrupted (did not verify correctly) or missing (that were not send by the server). The blocks corresponding to I are called *erasures*. During recovery the portal knows all the identities of blocks that were added to the parity structure P, that is, the portal is able to reconstruct set I. Let B be the set of blocks that correspond to the unique identities in I. After subtraction of the correct blocks from P, matrix P corresponds to (1) where the δ_{id}'s are known to the client and the δ_{val}'s need to be reconstructed by the portal during decoding. This can be done by a simple Gaussian elimination. Since the portal stores the parity structure, the portal knows P. The portal knows I and therefore B_{id}. See (2), the portal needs to solve the linear system of equations $P = B_{val} B_{id}$ for B_{val}. This is only possible if B_{id} has a right inverse (such that Gaussian elimination works fine).

If the number of blocks in B (that is, the number of rows in B_{id}) is more than a (the number of columns of B_{id}), then B_{id} does not have a right inverse. If the number of erasures, that is, the number of blocks in B, is less than or equal to p, then B_{id} may have a right inverse: over *random* binary matrices B_{id} with equal

probability of a 1 or 0 in each of its entries, the probability that all rows in B_{id} are linearly independent is equal to

$$\prod_{i=0}^{|B|-1}(1-2^{-(p-i)}) \geq 1-\sum_{i=0}^{|B|-1}2^{-(p-i)} \geq 1-2^{-(p-|B|)}. \quad (3)$$

Summarizing, the probability that erasure decoding fails is at most $2^{-(p-|B|)}$.

Analysis: In our dynamic POR solution, the client queries/challenges random blocks, which the server needs to correctly respond to (the client verifies the responses by using the authenticated data structure). If all n file blocks are in a single stripe and if the client checks c random challenge response pairs, then, given $j = |B|$ erasures, the probability (taken over a uniform distribution of challenge response pairs) that none of the challenge response pairs detects an erasure is equal to

$$\begin{aligned}\binom{n-j}{c}/\binom{n}{c} &= \binom{n-c}{j}/\binom{n}{j} \\ &\leq \frac{n-c}{n}\cdots\frac{n-c-j+1}{n-j+1} \leq (1-c/n)^j.\end{aligned}$$

For $c/n < 1/2$, given j erasures, the probability of decoding failure while all c challenge response pairs verify correctly is at most (see (3))

$$\min\{2^{-(p-j)},1\}(1-c/n)^j \leq (1-c/n)^p \leq e^{-pc/n}. \quad (4)$$

Hence, the probability ρ that not all blocks can be fully recovered, that is, the probability that erasure decoding fails while all c challenge response pairs verify correctly is at most

$$\rho \leq e^{-pc/n}. \quad (5)$$

E.g., for $c = l\sqrt{n}$ and $p = h\sqrt{n}$, $\rho \leq e^{-lh}$, storage is $s = b \cdot p = O(\sqrt{n})$ bits and verification bandwidth equals $c = O(\sqrt{n})$ number of challenge response pairs.

If a file block is written, then its old version is subtracted from parity structure P and the new version is added to P. These operations are efficient if matrix $A = \delta_{val} \otimes H_k(\delta_{id})$ in (1) can be represented and computed in an efficient way. The length of vector δ_{val} is the size b of a block value; b is a fixed system parameter (e.g., b equals 4KB or 32768 bits). Vector $H_k(\delta_{id})$ has size[1] $p = h\sqrt{n}$. Vector $H_k(\delta_{id})$ has an efficient representation if it has $O(\log p) = O(\log n)$ ones in which case updating P only costs $O(b \log p)$ XOR operations together with one hash evaluation. That is, the fraction of ones in $H_k(\delta_{id})$ is $O((\log p)/p)$. As a result matrix B_{id} is a random *sparse* binary matrix in which each entry is equal to 1 with probability $O((\log p)/p)$.

For a random sparse binary matrix B_{id}, inequality (3) may not hold. The rank of sparse matrices over finite fields has been well studied, see [31] for a survey. If each entry is chosen independently and at random to be 1 with probability $q \leq (\ln p)/p$, then the rank properties of B_{id} are not indistinguishable from the purely random case $q = 1/2$. Based on extensive simulation and literature, [31] *conjectures that for $q > 2(\ln p)/p$ and p large enough the rank properties of B_{id} are indistinguishable from the purely random case*, in particular a bound similar to (3) should hold.

During decoding the client needs to solve the linear system of equations $P = B_{val}B_{id}$ for B_{val}. If the server imposes $j = |B| = p$ erasures, then probability ρ of recovery failure is maximized (see (4)). For p

[1] If we design $p = O(\log n)$, then $H_k(\delta_{id})$ can be efficiently computed. However, for $p = O(\log n)$, ρ is small only if c scales linearly in n (see (4)), which is not practical.

erasures straightforward Gaussian elimination needs $p^2 = h^2 n$ storage (matrix B_{id} has $j = |B| = p$ rows and p columns). Notice that we only allow $O(\sqrt{n})$ storage at the client's site. In order to improve on the amount of required storage for Gaussian elimination, the client needs to use B_{id}'s sparse structure.

Belief propagation [27] works well (in $O(p)$ time and within the allowed storage) for any sparse matrix B_{id} that is sufficiently rectangular. In our application we need an exponentially small upper bound on ρ, so, we need to know an accurate estimate on the probability that decoding fails which current literature does not provide:

Belief propagation: If the number of rows in B_{id} is $j \leq 1/q = O(p/\log p)$ with $q \approx 2(\ln p)/p$, then B_{id} is expected to have a positive fraction of columns with a single 1. For such a sparse rectangular $j \times p$ matrix B_{id}, Gaussian elimination is efficient: As explained in [27], we may use belief propagation over the Binary Erasure Channel (BEC) for LDPC codes. The generator matrix of the LDPC code is the $n \times p$ matrix with rows $H_k(\delta_{id})$ for each of the n blocks δ (notice that it has B_{id} as a submatrix). The construction of matrix B_{id} is equivalent to the first iteration in the belief propagation algorithm. The decoding graph after the first iteration is a bipartite graph with nodes representing the rows of B_{id} and nodes representing the columns of B_{id}; a row node and a column node are connected by an edge only if the corresponding entry in B_{id} is 1. Since qjp equals the number of 1s in B_{id}, that is, the number of edges in the decoding graph after the first iteration, the remaining iterations cost $qjp = O(p)$ blocks storage and run in $qjp = O(p)$ time (as is explained in [27], each edge is considered once).

We notice that for $j \leq 1/q$, it is very likely that before every iteration the decoding graph is represented by a matrix having a positive fraction of columns with a single 1. This is a necessary condition for belief propagation to successfully complete erasure decoding. This corresponds to the *independence assumption* in [27], which is the assumption that each iteration is statistically independent of previous iterations in that the decoding graph at each iteration has the same edge degree distributions for row nodes and column nodes. Based on the independence assumption, a condition for successful erasure decoding using edge degree distributions can be derived (inequality (7) in [27]). In particular, if $j \leq 1/p$, then the independence assumption implies successful erasure decoding. The independence assumption is correct for the r first iterations only if the neighborhood of a row node in the bipartite graph represented by B_{id} up to depth r is a tree. After r iterations at most a small fraction of erasures (less than a constant) needs to be decoded using Gaussian elimination. It is likely that this can be done within the available amount of storage. However, if the set of linear equations after r iterations does not have full rank, then Gaussian elimination will fail.

Belief propagation works well for any sparse matrix B_{id} that is sufficiently rectangular. Unfortunately, current literature does not provide an exact analysis. Gaussian elimination fails with an exponentially small probability. For this reason, an analysis based on the independence assumption closely fits simulation results for LDPC codes in communication theory. However, in our application we need an exponentially small upper bound on ρ, so, we need to know an accurate estimate on the probability that Gaussian elimination fails. For example, suppose that f is such that for $j \leq p/(f \log p)$ (the larger f the more rectangular B_{id}) the probability of failing Gaussian elimination is $\leq 2^{-p/(f \log p)-j}$. Then, by using the arguments from which (4) and (5) are obtained,

$$\rho \leq (1 - c/n)^{p/(f \log p)} \leq e^{-pc/(nf \log p)}.$$

A.2 Multiple Stripes

Rather than using belief propagation, we design a specific sparse structure in combination with 1) an efficient decoding algorithm that meets the storage requirements, together with 2) an efficient updating algorithm having $u = O(\log n)$, and for which 3) we can provide a tight upper bound on ρ without assuming any

conjecture.

We propose to split the single stripe into m independent stripes each being a code word having $p = O(\log n)$ parities. Each block is assigned to exactly one stripe, hence, if the client updates a block, then on average only $u = p/2 = O(\log n)$ parities of the corresponding stripe need to be updated (the multi-striped structure is indeed sparse).

We need a keyed hash function $H_k(.)$ that maps an identity δ_{id} to a pseudo random bit string representing a pair $(\theta_{ind}; \theta)$, where $\theta = (\theta_1, \ldots, \theta_p)$ and θ_{ind} is the index of the stripe to which block $\delta = (\delta_{id}; \delta_{val})$ is added.

Each stripe i, $1 \leq i \leq m$, has its own parity structure $P[i]$. To add or subtract a block δ, the client computes $H_k(\delta_{id}) = (\theta_{ind}; \theta)$, constructs $A = \delta_{val} \otimes \theta$ as before, and updates $P[\theta_{ind}] \leftarrow P[\theta_{ind}] + A$.

Erasure decoding of the multi-striped structure consists of decoding each stripe separately. Successful decoding involves m times a Gaussian elimination, each time computing the right inverse of a $(\leq p) \times p$ matrix costing at most $p^2 = (a/m)^2 = O((\log n)^2)$ XOR operations. Decoding can be done within the allowed storage.

Recovery failure probability: The recovery failure probability ρ is equal to the probability that erasure decoding fails while all challenge response pairs verify correctly. As a technical preliminary, for integers p and u and $0 \leq \beta \leq 1$, let $R(p, u, \beta)$ be the probability that a binary $\beta p \times p$ rectangular sparse matrix in which each entry is chosen independently and at random to be 1 with probability u/p does not have full rank. The following theorem expresses the upper bound on ρ in terms of $R(p, u, \beta)$.

In the purely random case $u = p/2$ and $R(p, u = p/2, \beta) \leq 2^{-(1-\beta)p}$, see (3). Based on extensive simulation and literature, [31] states the conjecture that for $u > 2 \ln p$ and β sufficiently close to 1, $R(p, u, \beta) \approx R(p, p/2, \beta) \leq 2^{-(1-\beta)p}$.

We define $h(x) = -x \log_2 x - (1 - x) \log_2(1 - x)$ as the binary entropy function.

Theorem 2 *For any value of technical parameters $0 \leq \beta, \epsilon \leq 1$ and $0 \leq \kappa \leq e^{1/4}/(2\pi)$,*

$$\rho \leq \frac{s}{pb} \cdot \left[\begin{array}{l} \frac{e^{1/8}}{\sqrt{\pi}} \sqrt{\frac{(2-\epsilon)^3(1-\kappa)^2}{2\kappa(\kappa+\epsilon-\epsilon\kappa)}} n 2^{-(1-h((1-\epsilon)/(2-\epsilon)))\cdot\beta p} + \\ + R(p, u, \beta) + e^{-(1-\epsilon)\beta cs/(nb)} + e^{-(1-\kappa)cpb/s} \end{array} \right].$$

For $\kappa \leq c/n$, the inequality holds with the term $e^{-(1-\kappa)cpb/s}$ removed.

By substituting $R(p, u, \beta) \leq 2^{-(1-\beta)p}$ and $\beta = 1/(2 - h((1-\epsilon)/(2-\epsilon)))$, we obtain the next theorem, where we use the monotonically increasing function $f \in [0, 1) \rightarrow [0, \infty)$ and monotonically decreasing function $g \in [0, 1] \rightarrow [0, 1]$ defined as

$$\begin{aligned} f(\epsilon) &= (\ln 2)(1 - h((1-\epsilon)/(2-\epsilon)))/(1-\epsilon), \\ g(\epsilon) &= (1-\epsilon)/(2 - h((1-\epsilon)/(2-\epsilon))). \end{aligned}$$

Theorem 3 *Let $u = p/2$ or let $u > 2 \ln p$ and $\beta = 1/(2 - h((1-\epsilon)/(2-\epsilon)))$ sufficiently close to 1 such that $R(p, u, \beta) \approx R(p, p/2, \beta) \leq 2^{-(1-\beta)p}$. Then, for any value of technical parameters $0 \leq \epsilon \leq 1$ and $0 \leq \kappa \leq e^{1/4}/(2\pi)$,*

$$\rho \leq \frac{s}{pb} \cdot \left[\begin{array}{l} \left(1 + \frac{e^{1/8}}{\sqrt{\pi}} \sqrt{\frac{(2-\epsilon)^3(1-\kappa)^2}{2\kappa(\kappa+\epsilon-\epsilon\kappa)}} n\right) e^{-f(\epsilon)g(\epsilon)\cdot p} + \\ + e^{-g(\epsilon)\cdot cs/(nb)} + e^{-(1-\kappa)cpb/s} \end{array} \right].$$

For $\kappa \leq c/n$, the inequality holds with the term $e^{-(1-\kappa)cpb/s}$ removed.

We will first prove Theorem 3 from which the proof of Theorem 2 will follow as a direct consequence.

Proof of Theorem 3: Suppose that the client checks $c \leq n$ random challenge response pairs such that c/m blocks within each of the m stripes are verified. Suppose that there are j erasures. We want to compute the probability ρ that not all blocks can be fully recovered given c challenge response pairs and j erasures. The recovery failure probability ρ is equal to the probability that, given c challenge response pairs and j erasures, erasure decoding fails while all challenge response pairs verify correctly. We will prove a tight upper bound on ρ.

Let w_i be the number of blocks added to a stripe i. Notice that w_i is binomially distributed with length n and probability $1/m$; the expected number of blocks added to a single stripe is $w = n/m$ with standard deviation $\sqrt{(1-1/m)n/m} \leq \sqrt{w}$. By using Chernoff's bounds for the lower and upper tail of the binomial distribution, we obtain, for $t \geq 0$,

$$Pr(w_i < w - t\sqrt{w}) < e^{-t^2/2}$$

and

$$\begin{aligned} Pr(w_i > w + t\sqrt{w}) &< (e^{t/\sqrt{w}}/(1+t/\sqrt{w})^{(1+t/\sqrt{w})})^w \\ &\leq e^{-t^2/4} \text{ for } t/\sqrt{w} < 2e-1. \end{aligned}$$

If $e^{-t^2/4} = \rho/r$, then $|w - w_i| \leq 2\sqrt{(\log r/\rho)/w}$ with probability $\geq 1 - \rho/r$. If $w \gg \log(r/\rho)$, then the approximation $w_i = w = n/m$ for each stripe holds with probability $\geq 1 - \rho/r$. In our analysis we use this approximation. We choose not to include the bounds on w_i given by $|w - w_i| \leq 2\sqrt{(\log r/\rho)/w}$ in order to keep our analysis presentable.

Let $Pr(j_1, \ldots, j_m)$ with $j = j_1 + \ldots + j_m$ be the probability that the distribution of the j erasures over stripes is such that, for $1 \leq i \leq m$, stripe i has j_i erasures. Since the keyed hash function outputs pseudo random sequences, the server cannot distinguish the actual assignment of blocks to stripes from a random assignment. Therefore,

$$Pr(j_1, \ldots, j_m) = \left[\prod_{i=1}^{m} \binom{w}{j_i}\right] / \binom{wm}{j_1 + \ldots + j_m}.$$

We define J as the set of sequences of non-negative integers that sum up to j (notice that J has sequences of variable length).

Our analysis for a single stripe generalizes: given a distribution of $(j_1, \ldots, j_m) \in J$ erasures, the probability of decoding failure is equal to

$$Pr(\text{failure}|j_1, \ldots, j_m) \leq \sum_{i=1}^{m} \min\{1, 2^{-(p-j_i)}\}. \tag{6}$$

If $j_i \leq w - c/m$ for all $1 \leq i \leq m$, then the probability of not detecting any erasure during the verification of all challenge response pairs is equal to

$$Pr(\text{no-detection}|j_1, \ldots, j_m) = \prod_{i=1}^{m} \left[\binom{w-j_i}{c/m} / \binom{w}{c/m}\right]. \tag{7}$$

If there exists an index i such that $j_i > w - c/m$, then $Pr(\text{no-detection}|j_1, \ldots, j_m) = 0$.

Since "failure" and "no-detection" are independent statistical events,

$$\rho = \sum_{(j_1,\ldots,j_m)\in J} \begin{array}{l} Pr(j_1,\ldots,j_m) \\ \cdot Pr(\text{failure}|j_1,\ldots,j_m) \\ \cdot Pr(\text{no-detection}|j_1,\ldots,j_m). \end{array} \tag{8}$$

The following lemmas derive tight upper bounds on each of the three probabilities in ρ.

Lemma 1 *Let* $0 \leq x \leq 1$ *and let*

$$\mathcal{A} \subseteq \left\{ (j_1, j_2) \textit{ s.t. } \binom{w}{j_1}\binom{w}{j_2} / \binom{2w}{j_1+j_2} \leq x \right\}.$$

Define

$$z = \max_{(j_1,j_2)\in\mathcal{A}} j_1 + j_2.$$

Then,

$$\sum_{(j_1,\ldots,j_m)\in J \textit{ s.t. } (j_1,j_2)\in\mathcal{A}} Pr(j_1,\ldots,j_m) \leq xz. \tag{9}$$

Proof. We first substitute

$$Pr(j_1,\ldots,j_m) = \frac{\binom{w}{j_1}\binom{w}{j_2}}{\binom{2w}{j_1+j_2}} \frac{\binom{2w}{j_1+j_2}\binom{w}{j_3}\cdots\binom{w}{j_m}}{\binom{2w+w+\ldots+w}{(j_1+j_2)+j_3+\ldots+j_m}}$$

into the left side of inequality (9). By using the stated assumption on set $\mathcal{A}$, this yields the upper bound

$$\sum_{(j_1,\ldots,j_m)\in J \text{ s.t. } (j_1,j_2)\in\mathcal{A}} x \frac{\binom{2w}{j_1+j_2}\binom{w}{j_3}\cdots\binom{w}{j_m}}{\binom{2w+w+\ldots+w}{(j_1+j_2)+j_3+\ldots+j_m}}$$

Since $|\{(j_1, j_2) \in \mathcal{A} \text{ s.t. } j_1 + j_2 = j'\}| \leq z$ for any j', we obtain the upper bound

$$\sum_{(j',j_3,\ldots,j_m)\in J} xz \frac{\binom{2w}{j'}\binom{w}{j_3}\cdots\binom{w}{j_m}}{\binom{2w+w+\ldots+w}{j'+j_3+\ldots+j_m}} = xz.$$

QED

Lemma 2 *For* $j_1 \leq (1-\kappa)w$ *with* $0 \leq \kappa \leq e^{1/4}/(2\pi)$ *and* $j_2 \leq \gamma j_1$ *with* $0 \leq \gamma \leq 1$,

$$\frac{\binom{w}{j_1}\binom{w}{j_2}}{\binom{2w}{j_1+j_2}} < \frac{e^{1/8}}{\sqrt{\pi}} \sqrt{\frac{1+\gamma}{2\kappa(1-\gamma(1-\kappa))}} 2^{-(1-h(\gamma/(1+\gamma)))\cdot j_1}.$$

Proof. The limiting case $j_2 = 0$ with $\gamma = 0$ follows from

$$\binom{w}{j_1} / \binom{2w}{j_1} \quad \leq \quad (1 - w/(2w))^{j_1} = 2^{-j_1}.$$

For the general case we use the following upper and lower bound from [29, Theorem 2.6 with $n = 1$]: for $v_0 > v_1 \geq 1$,

$$\frac{e^{-1/8}}{\sqrt{2\pi}} C(v_0, v_1) < \binom{v_0}{v_1} < \frac{1}{\sqrt{2\pi}} C(v_0, v_1),$$

where

$$C(v_0, v_1) = \frac{v_0^{v_0+1/2}}{v_1^{v_1+1/2}(v_0 - v_1)^{v_0-v_1+1/2}}.$$

For $j_2 \neq 0$ (implying $j_1 \neq 0$), these bounds yield (after a reordering of terms) the upper bound

$$\frac{\binom{w}{j_1}\binom{w}{j_2}}{\binom{2w}{j_1+j_2}} < \begin{array}{l} \frac{e^{1/8}}{\sqrt{\pi}} \sqrt{w} \sqrt{\frac{j_1+j_2}{2j_1j_2}} \sqrt{\frac{(w-j_1)+(w-j_2)}{2(w-j_1)(w-j_2)}} \\ \qquad \cdot L\left[\frac{j_1-j_2}{j_1+j_2}\right]^{-(j_1+j_2)/2} \\ \qquad \cdot L\left[\frac{(w-j_2)-(w-j_1)}{(w-j_2)+(w-j_1)}\right]^{-((w-j_2)+(w-j_1))/2} \end{array}$$

where

$$L[y] = (1+y)^{1+y}(1-y)^{1-y} = 2^{2(1-h((1-y)/2))}$$

for the binary entropy function $h(.)$.

The lemma follows after applying each of the following bounds: $L[y] \geq 1$, from $(j_1 - j_2)/(j_1 + j_2) \geq (1-\gamma)/(1+\gamma)$ we obtain $L[(j_1-j_2)/(j_1+j_2)] \geq L[(1-\gamma)/(1+\gamma)]$, $(j_1+j_2)/2 \geq j_1/2$, from $1 \leq j_2 \leq \gamma j_1$ we obtain $\sqrt{(j_1+j_2)/(2j_1j_2)} \leq \sqrt{(1+\gamma)/2}$, and from $0 \leq j_1 \leq (1-\kappa)w$ and $0 \leq j_2 \leq \gamma j_1 \leq \gamma(1-\kappa)w$ we obtain $\sqrt{(2w - j_1 - j_2)/(2(w-j_1)(w-j_2))} \leq 1/\sqrt{\kappa(1-\gamma(1-\kappa))w}$. QED

From the two previous lemmas we obtain:

Lemma 3 *Let $0 \leq \beta, \gamma \leq 1$ and $0 \leq \kappa \leq e^{1/4}/(2\pi)$, and define*

$$\mathcal{B} = \left\{ (j_1, \ldots, j_m) \in J \text{ s.t. } \begin{array}{l} \beta p \leq j_1 \leq (1-\kappa)w \\ \text{and } \exists_i j_i \leq \gamma j_1 \end{array} \right\}.$$

Then, $\sum_{(j_1,\ldots,j_m)\in\mathcal{B}} Pr(j_1, \ldots, j_m)$ is at most equal to

$$\frac{e^{1/8}}{\sqrt{\pi}} \sqrt{\frac{(1+\gamma)^3(1-\kappa)^2}{2\kappa(1-\gamma(1-\kappa))}} n 2^{-(1-h(\gamma/(1+\gamma)))\cdot\beta p}.$$

Proof. Let x be the upper bound of Lemma 2 with j_1 lower bounded by βp. By symmetry arguments, the bound in Lemma 2 holds for $j_1 \leq (1-\kappa)w$ and any index i such that $j_i \leq \gamma j_1$. So, set $\mathcal{B}$ is a subset of

$$\bigcup_{i=2}^{m} \left\{ (j_1, \ldots, j_m) \in J \text{ s.t. } \begin{array}{l} \beta p \leq j_1 \leq (1-\kappa)w, \\ j_i \leq \gamma j_1 \text{ and} \\ \binom{w}{j_1}\binom{w}{j_i}/\binom{2w}{j_1+j_i} \leq x \end{array} \right\}.$$

For each index pair $(1, i)$ in this union, we define a set $\mathcal{A}$ as in Lemma 1, where $z \leq (1+\gamma)(1-\kappa)w$. Application of Lemma 1 for each pair $(1, i)$ proves that

$$\sum_{(j_1,\ldots,j_m)\in\mathcal{B}} Pr(j_1, \ldots, j_m) \leq (m-1)xz \leq (1+\gamma)(1-\kappa)nx.$$

QED

Lemma 4 *Let* $0 \leq \beta, \gamma \leq 1$*, and define*

$$\mathcal{B} = \left\{ (j_1, \ldots, j_m) \in J \text{ s.t. } \begin{array}{l} \beta p \leq j_1 \\ \text{and } \forall_i j_i \geq \gamma j_1 \end{array} \right\}.$$

Then, for $(j_1, \ldots, j_m) \in \mathcal{B}$*,*

$$Pr(\textit{no-detection}|j_1, \ldots, j_m) \leq e^{-\gamma\beta cpm/n}.$$

Proof. If $(j_1, \ldots, j_m) \in \mathcal{B}$, then all $j_i \geq \gamma\beta p$. Hence, probability $Pr(\text{no-detection}|j_1, \ldots, j_m)$ is equal to (notice that $c \leq mw = n$)

$$\begin{aligned} \prod_{i=1}^{m} \left[\binom{w - j_i}{c/m} / \binom{w}{c/m} \right] &\leq \left[(1 - (c/m)/w)^{\gamma\beta p} \right]^m \\ &\leq e^{-\gamma\beta cpm/n}. \end{aligned}$$

QED

Lemma 5 *Let* $j_1 > (1 - \kappa)w$ *with* $0 \leq \kappa \leq e^{1/4}/(2\pi)$*. For* $(j_1, \ldots, j_m) \in J$ *and* $\kappa > c/n$*,*

$$Pr(\textit{no-detection}|j_1, \ldots, j_m) \leq e^{-(1-\kappa)c/m}.$$

If $\kappa \leq c/n$*, then* $Pr(\textit{no-detection}|j_1, \ldots, j_m) = 0$*.*

Proof. If $\kappa \leq c/n$, then $j_1 \geq (1-\kappa)w \geq (1-c/n)w = w - c/m$, hence, $Pr(\text{no-detection}|j_1, \ldots, j_m) = 0$. For $\kappa > c/n$, $Pr(\text{no-detection}|j_1, \ldots, j_m)$ is at most equal to, see (7),

$$\begin{aligned} \binom{w - j_1}{c/m} / \binom{w}{c/m} &\leq (1 - (c/m)/w)^{j_1} \\ &\leq e^{-j_1 c/n} \leq e^{-(1-\kappa)c/m}. \end{aligned}$$

QED

Lemma 6 *For any* $0 \leq \beta, \gamma \leq 1$ *and* $0 \leq \kappa \leq e^{1/4}/(2\pi)$*,*

$$\rho \leq m \cdot \left[\begin{array}{l} \frac{e^{1/8}}{\sqrt{\pi}} \sqrt{\frac{(1+\gamma)^3(1-\kappa)^2}{2\kappa(1-\gamma(1-\kappa))}} n 2^{-(1-h(\gamma/(1+\gamma)))\cdot\beta p} + \\ +2^{-(1-\beta)p} + e^{-\gamma\beta cpm/n} + e^{-(1-\kappa)c/m} \end{array} \right].$$

For $\kappa \leq c/n$*, we may remove the term* $e^{-(1-\kappa)c/m}$ *from the bound.*

Proof. We first notice that if all $j_i \leq \beta p$, $1 \leq i \leq m$, then $Pr(\text{failure}|j_1, \ldots, j_m)$ is at most equal to, see (6),

$$\sum_{i=1}^{m} \min\{1, 2^{-(p-j_i)}\} \leq m 2^{-(1-\beta)p}. \tag{10}$$

Secondly, by symmetry Lemmas 3, 4 and 5 hold for j_1 replaced by any j_i. By combining all observations, ρ, see (8), is at most the right side of (10) plus m times the sum of the upper bounds stated in Lemmas 3, 4 and 5. This proves the lemma. QED

Theorem 3 follows immediately from Lemma 6 by choosing β such that $(1-h(\gamma/(1+\gamma)))\beta = 1-\beta$, that is,

$$\beta = 1/(2-h(\gamma/(1+\gamma))),$$

and by substituting $m = s/(pb)$ and $\gamma = 1-\epsilon$, for $0 \leq \epsilon \leq 1$.

Proof of Theorem 2: Rank properties only play a role in (6) and in (10). By replacing the bound in (10) by $m \cdot R(p,u,\beta)$ we obtain Theorem 2.

Asymptotic: We will now analyze the strength of the derived upper bound on ρ and argue that for large n, the bound is tight.

Let $\kappa = c/n$. Notice that $(2-\epsilon)^3(1-\kappa)^2 \leq 8$ and $2\kappa(\kappa+\epsilon-\epsilon\kappa) \geq 2\kappa^2$. For $\epsilon \geq f^{-1}(cs/(npb))$, $f(\epsilon)g(\epsilon) > g(\epsilon)cs/(npb)$ and the terms in the upper bound of ρ collapse leading to

$$\rho \leq \frac{s}{pb}(2 + \frac{2 \cdot e^{1/8}}{\sqrt{\pi}}\frac{n^2}{c})e^{-g(\epsilon)\cdot cs/(nb)}.$$

Notice that if we have a single stripe, then $s = pb$ and $f^{-1}(cs/(npb)) = f^{-1}(c/n)$ which is close to 0 for $c = O(\sqrt{n})$. Notice that $g(0) = 1$, so the bound in the theorem corresponds to (5).

A more precise analysis uses the Taylor expansion around $1/2$ of the binary entropy function,

$$h(x) = 1 - \frac{1}{2\ln 2}\sum_{i=1}^{\infty}\frac{(1-2x)^{2i}}{i(2i-1)},$$

hence, for $0 \leq \epsilon \leq 1$,

$$h\left(\frac{1-\epsilon}{2-\epsilon}\right) = 1 - \frac{1}{2\ln 2}\sum_{i=1}^{\infty}\frac{(\epsilon/(2-\epsilon))^{2i}}{i(2i-1)} \leq 1 - \frac{\epsilon^2}{8\ln 2}$$

and

$$\begin{aligned} h\left(\frac{1-\epsilon}{2-\epsilon}\right) &= 1 - \frac{1}{2\ln 2}\sum_{i=1}^{\infty}\frac{(\epsilon/(2-\epsilon))^{2i}}{i(2i-1)} \\ &\geq 1 - \frac{1}{2\ln 2}\sum_{i=1}^{\infty}(\epsilon/(2-\epsilon))^{2i} \\ &= 1 - \frac{1}{2\ln 2}\left(\frac{1}{1-\frac{\epsilon}{2-\epsilon}} - 1\right) \\ &= 1 - \frac{\epsilon}{4\ln 2(1-\epsilon)}. \end{aligned}$$

The upper bound proves

$$f(\epsilon) \geq \epsilon^2/(8(1-\epsilon)) \geq \epsilon^2/8.$$

So, $\epsilon^2/8 \geq cs/(npb)$ implies $\epsilon \geq f^{-1}(cs/(npb))$. The lower bound proves

$$g(\epsilon) \geq (1-\epsilon)/\left(1 + \frac{\epsilon}{4\ln 2(1-\epsilon)}\right) \geq 1 - \frac{1+4\ln 2}{4\ln 2}\epsilon.$$

So,

$$\rho \quad \leq \quad \frac{s}{pb}(2+\frac{2\cdot e^{1/8}}{\sqrt{\pi}}\frac{n^2}{c})e^{-(1-\frac{1+4\ln 2}{4\ln 2}\sqrt{8cs/(npb)})\cdot cs/(nb)}.$$

For a single stripe, $s = pb$ and we obtain

$$\rho \leq (2+1.278\cdot\frac{n^2}{c})e^{-(1-3.849\cdot\sqrt{c/n})\cdot cp/n}$$

showing into what extend our bounding techniques weakened the bound in (5).

Remark: In a further refinement of the multi-striping structure, we may decide to assign each block to two arbitrary stripes each having half the number of parities (such that the cost of updating parities remains the same). Such a "two-dimensional" striping structure has the characteristics of a product code; successful erasure decoding is more likely. Notice that a "multi-dimensional" striping structure tends to become the random sparse structure as discussed for the single stripe with the bound (5).

A.3 Example Parameters

If we set $\kappa = e^{1/4}/(2\pi) = 0.204$ and $\epsilon = 3/4$ (resulting in $f(\epsilon) = 1.11$, $g(\epsilon) = 0.196$, and $\beta = 0.687 \ll 1$), the upper bound of Theorem 3 yields

$$\rho \leq \frac{s}{pb}\cdot\left[(1+1.24\cdot n)e^{-p/4.6}+e^{-cs/(5.1\cdot nb)}+e^{-0.8\cdot cpb/s}\right].$$

Notice that under the condition $s/b \leq 2.0\cdot\sqrt{np}$, we have $cs/(5.1\cdot nb) \leq 0.8\cdot cpb/s$. Thus, letting $c = 5.1\cdot(nb/s)\cdot(l+\ln(s/(bp)))$ for technical parameter l, we obtain the following corollary:

Corollary 1 *Let $u = p/2$. Then, for any $l > 1$, if $s/b \leq 2.0\cdot\sqrt{np}$, $c = 5.1\cdot(nb/s)\cdot(l+\ln(s/(bp)))$ and $p \geq 4.6\cdot(l+\ln(1.24\cdot n)+\ln(s/(pb)))$, then*

$$\rho \leq 3\cdot e^{-l}.$$

If we set $\kappa = e^{1/4}/(2\pi) = 0.204$ and $\epsilon = 1/3$ (resulting in $f(\epsilon) = 0.03$, $g(\epsilon) = 0.648$, and $\beta = 0.972$), we obtain the following corollary:

Corollary 2 *Let $u > 2\ln p$ such that $R(p,u,0.972) \approx R(p,p/2,0.972) \leq 2^{-(1-0.972)p}$. Then, for any $l > 1$, if $s/b \leq 2.0\cdot\sqrt{np}$, $c = 1.54\cdot(nb/s)\cdot(l+\ln(s/(bp)))$ and $p \geq 51.45\cdot(l+\ln(1.71\cdot n)+\ln(s/(pb)))$, then*

$$\rho \leq 3\cdot e^{-l}.$$

If condition $s/b \leq 2.0\cdot\sqrt{np}$ in Corollary 1 is not satisfied, then we need to use $\kappa = c/n$ in Theorem 3. Together with $\epsilon = 3/4$, this gives the slightly weaker bound $\rho \leq \frac{s}{bp}\cdot\left[(1+1.27\cdot\frac{n(n-c)}{\sqrt{c(3n+c)}})e^{-p/4.6}+e^{-cs/(5.1\cdot nb)}\right]$:

Corollary 3 *Let $u = p/2$. Then, for any $l > 1$, if $c = 5.1\cdot(nb/s)\cdot(l+\ln(s/(bp)))$ and $p \geq 4.6\cdot(l+\ln(1.27\cdot\frac{n(n-c)}{\sqrt{c(3n+c)}})+\ln(s/(2pb)))$, then*

$$\rho \leq 3\cdot e^{-l}.$$

The three corollaries combined prove Theorem 1.