Cryptographic Security for a High-Performance Distributed File System

Roman Pletka *
AdNovum Informatik AG
CH-8005 Zürich, Switerland
roman@pletka.ch

Christian Cachin
IBM Zurich Research Laboratory
CH-8803 Rüschlikon, Switzerland
cca@zurich.ibm.com

Abstract

Storage systems are increasingly subject to attacks. Cryptographic file systems mitigate the danger of exposing data by using encryption and integrity protection methods and guarantee end-to-end security for their clients. This paper describes a generic design for cryptographic file systems and its realization in a distributed storage-area network (SAN) file system. Key management is integrated with the meta-data service of the SAN file system. The implementation supports file encryption and integrity protection through hash trees. Both techniques have been implemented in the client file system driver. Benchmarks demonstrate that the overhead is noticeable for some artificially constructed use cases, but that it is very small for typical file system applications.

1. Introduction

Security is quickly becoming a mandatory feature of data storage systems. Today, storage space is typically provided by complex networked systems. These networks have traditionally been confined to data centers in physically secured locations. But with the availability of high-speed LANs and storage networking protocols such as FCIP [30] and iSCSI [32], these networks are becoming virtualized and open to access from user machines. Hence, clients may access the storage devices directly, and the existing static security methods no longer make sense. New, dynamic security mechanisms are required for protecting stored data in virtualized and networked storage systems.

A secure storage system should protect the confidentiality and the integrity of the stored data. In distributed storage systems, data exists in two different forms, leading also to different exposures to unauthorized access:

Data in flight: Data that is in transit on a network, between clients, servers, and storage devices. Unauthorized access may occur from other nodes on the network. These attacks and their countermeasures are similar to the situation for other communication channels, for which cryptographic protection is widely available.

Data at rest: Data that resides on a storage device. An attacker may physically access the storage device or send appropriate commands over the network. If the network is not secure, these commands may also be initiated by clients that are authorized to access other parts of the networked storage system. Data at rest differs from data in flight because it is sometimes harder to transparently apply cryptographic protection that expands the data length, like appending a few bytes of integrity checks to a stored data block. Furthermore, data at rest must be accessible in arbitrary order, unlike data on a communication channel that is only read in the order it was written. In such cases, new cryptographic methods are needed for protecting data at rest.

Data at rest is generally considered to be at higher risk than data in flight, because an attacker has more time and flexibility to access it. Moreover, new regulations such as Sarbanes-Oxley, HIPAA, and Basel II also dictate the use of encryption for data at rest.

Storage systems use a layered architecture, and cryptographic protection can be applied on any layer. For example, one popular approach used today is to encrypt data at the level of the block-storage device, either in the storage device itself, by an appliance on the storage network [14], or by a virtual device driver in the operating system (e.g., encryption using the loopback device in Linux). The advantage is that file systems can use the encrypted devices without modifications, but the drawback is that such file systems cannot extend the cryptographic security to its users. The reason is that any file-system client can access the storage space in its unprotected form, and that access control and key administration take place below the file system.

In this paper, we address encryption at the file-system

*Work done at IBM Zurich Research Laboratory.

level. We describe the design and implementation of cryptographic protection methods in a high-performance distributed file system. After introducing a generic model for secure file systems, we show how it can be implemented using SAN.FS, a SAN file system from IBM [25]. Our design addresses confidentiality protection by data encryption and integrity protection by means of hash trees. A key part of this paper is the discussion of the implementation and an evaluation of its performance. The model itself as well as our design choices are generic and can be applied to other distributed file systems.

Encryption in the file system maintains the end-to-end principle in the sense that stored data is protected at the level of the file-system users, and not at the infrastructure level, as is the case with block-level encryption for data at rest and storage-network encryption for data in flight. Moreover, an optimally secure distributed storage architecture should minimize the use of cryptographic operations and avoid unnecessary decryption and re-encryption of data as long as the data does not leave the file system. This can be achieved by performing encryption and integrity protection of data directly on the clients in the file system, thereby eliminating the need to separately encrypt the data in flight between clients and storage devices. Given the processing capacity of typical workstations today, encryption and integrity verification add only a small overhead to the cost of file-system operations, as our benchmarks demonstrate.

Distributed file systems like SAN.FS and cluster file systems are usually optimized for performance, capacity, and reliability. For example, in SAN.FS and in the recent pNFS effort [15], meta-data operations are separated from the data path for increasing scalability. From a security perspective, such an approach might sometimes be suboptimal or even make it impossible to provide end-to-end security. This work shows that cryptographic security can be added to high-performance distributed file systems at minimal additional performance cost. Although our work was done in the context of SAN.FS, our findings apply also to other distributed file systems.

The remainder of this paper is organized as follows. Section 2 introduces a general model for secure file systems and discusses related work. Then, Section 3 describes the design of SAN.FS and how cryptographic extensions can be added to it. Section 4 provides more details about our implementation of cryptographic extensions to SAN.FS. Section 5 shows our performance results and Section 6 concludes the paper.

2. Model and Related Work

This section first presents an abstract model of a distributed file system, introduces cryptographic distributed file systems, and reviews previous work in the area.

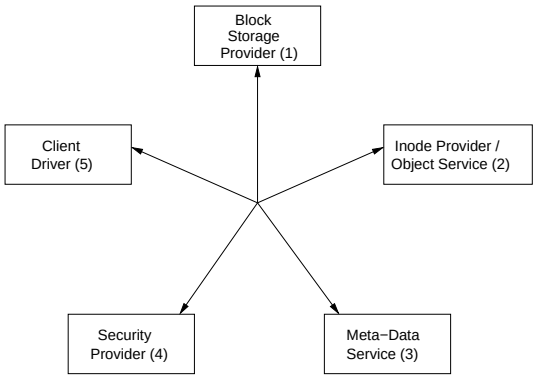

Figure 1. Components of a distributed file system.

2.1. File System Components

File systems are complex programs designed for storing data on persistent storage devices such as disks. A file system manages the space available on the storage devices, provides the abstraction of files, which are data containers that can grow or shrink and have a name and other meta-data associated to them, and manages the files by organizing them into a hierarchical directory structure.

Internally, most file systems distinguish at least the following five components as shown in Figure 1: (1) a *block-storage provider* that serves as a bulk data store and operates only on fixed-size blocks; (2) an *inode provider* (or *object-storage service*), which provides a flat space of storage containers of variable size; (3) a *meta-data service*, handling abstractions such as directories and file attributes and coordinating concurrent data access; (4) a *security provider* responsible for security and access-control features; and (5) a *client driver* that uses all other components to realize the file system abstraction to the operating system on the client machine.

The first three components correspond to the layered design of typical file systems, i.e., data written to disk in a file system traverses the file-system layer, the object layer, and the block layer in that order. The security provider is usually needed by all three layers. In most modern operating systems, the block-storage provider is implemented as a block device in the operating system, and therefore not part of the file system.

In traditional file systems, all components reside on the same host in one module. With the advent of high-speed networks, it has become feasible to integrate file system components across several machines into distributed file systems, which allow concurrent access to the data. A network can be inserted between any or all of the com-

ponents, in principle, and the networks themselves can be shared. For example, in storage-area networks only the storage provider is accessed over a network; in distributed file systems such as NFS and AFS, the client uses a network to access a file server, which contains storage, inode, and meta-data providers. The security provider can be an independent entity in AFS and in NFSv4.

The NASD architecture [10] and its successor Object Store [3] propose network access to the object-storage service. Compared with accessing a block-storage provider over the network, this design simplifies the security architecture. The security model for object storage [2] assumes that the device is trusted to enforce access control on a per-object basis. The security provider is realized as an independent entity, accessed over a network. Object storage is an emerging technology, and, to our knowledge, distributed file systems in which clients directly access object-storage devices are not yet widely available.

In SAN.FS, on which we focus in the remainder of this paper, clients access the storage devices directly over a SAN (i.e., using Fibre Channel or iSCSI). All meta-data operations are delegated to a dedicated server, which is accessed using TCP/IP over a local-area network (LAN).

2.2. Cryptographic File Systems

Cryptographic file systems encrypt and/or protect the integrity of the stored data using encryption and data authentication. Cryptography is used because the underlying storage provider is not trusted to prevent unauthorized access to the data. For example, the storage provider may use removable media or must be accessed over a network, and therefore proper access control cannot be enforced; another common example of side-channels to the data are broken disks that are being replaced.

In a system using encryption, access to the keys gives access to the data. Therefore, it is important that the *security provider* manages the encryption keys for the file system. Introducing a separate key management service, which has to be synchronized with the security provider providing access control information, only complicates matters. Analogously, the security provider should be responsible for managing integrity reference values, such as hashes of all files.

Cryptographic file systems exist in two forms: either as an enhancement within an existing physical file system that uses an underlying block-storage provider, or as a virtual file system that must be mounted over another (virtual or physical) file system. The first approach results in *monolithic* cryptographic file systems that can be optimized for performance. The second approach results in *stackable* or *layered* file systems [35], whose advantage lies in the isolation of the encryption functionality from the details of a physical file system. In this way, the encryption layer can be reused for many physical file systems. But because the operating system must maintain a copy of the data on each layer, stackable file systems are generally slower than monolithic ones.

2.3. Previous Work on Cryptographic File Systems

A considerable number of prototype and production cryptographic file systems have been developed in the past 15 years. We refer to the excellent surveys by Wright *et al.* [34] and by Kher and Kim [18] for more details, and mention only the most important systems here.

Most early cryptographic file systems are layered and use the NFS protocol for accessing a lower-layer file system: CFS [4] uses an NFS loopback server in user space and provides per-directory keys that are derived from passphrases; TCFS [6] uses a modified NFS client in the kernel and utilizes a hierarchical key management scheme, in which per-user master keys to protect per-file keys are maintained. SFS [22, 23, 24] is a distributed cryptographic file system also using the NFS interfaces, which is available for several Unix variants. These systems do not contain an explicit security provider responsible for key management, and delegate much of that work to the user.

SUNDR [21] is a distributed file system that works with a completely untrusted storage server. It guarantees that clients can detect any violation of integrity and consistency, as long as they see file updates of each other, but it cannot prevent modifications to the stored data by the server. SUNDR provides file integrity protection using hash trees, and makes frequent use of digital signatures.

Another system that protects file integrity is I3FS [28], a layered file system designed to detect malicious file modifications in connection with intrusions. It needs a second authorization mechanism apart from the normal file system authorization and acts as a Tripwire-like [19] intrusion-detection system built into the kernel.

SFS-RO [7] and Chefs [8] are two systems protecting file integrity using hash trees designed for read-only data distribution, where update are only possible by using off-line operations. Like I3FS, they do not implement the standard file system interface and require special commands for write operations.

A cryptographic file system has also been implemented using secure network-attached disks (SNAD) [27]. SNAD storage devices are a hybrid design, providing traditional block storage as well as features commonly found in object-storage devices and file servers. In contrast to traditional storage providers, SNAD devices require strong client authentication for any operation and also perform data verification on the content. Data is encrypted by the

clients before sending it to the SNAD and authenticated using per-block digital signatures or per-block secret-key authentication (MAC).

Microsoft Windows 2000 and later editions contain an extension of NTFS called EFS [31], which provides file encryption with shared and group access. It relies on the security provider in the Windows operating system for user authentication and key management. As it is built into NTFS, it represents a monolithic solution.

Some more recent cryptographic file systems follow the layered approach: NCryptfs [33] and eCryptFS [13] are two native Linux file systems, which are implemented in the kernel and use stacking at the VFS layer based on the FiST framework [36]. EncFS [12] for Linux is implemented in user-space relying on Linux's file system in user space module (FUSE). FUSE intercepts system calls at the VFS layer and redirects them to the daemon in user space. NCryptfs, eCryptFS, and EncFS currently provide only manual key management on a per-file system basis, but the eCryptFS design includes support for a sophisticated key management scheme with per-file encryption keys and shared access using public-key cryptography.

Farsite [1] and Oceanstore [20] are two storage systems designed for wide-area and scalable file sharing; they protect confidentiality and integrity through various techniques, including encryption and hash trees, and use decentralized administration. They differ in their trust model and in their performance characteristics from the kernel-level file systems considered here.

Except for Windows EFS and apart from using a stackable file system on top of a networked file system such as NFS or AFS, there are currently no distributed cryptographic file systems that offer high performance and allow file sharing and concurrent access to encrypted files.

All file systems mentioned support confidentiality through encryption, SUNDR and SFS-RO provide only data integrity through hash functions and digital signatures, and Farsite, Oceanstore, and Chefs support both.

3. Design

This section presents the SAN File System (SAN.FS) and our design for turning SAN.FS into a cryptographic file system supporting confidentiality and integrity.

3.1. SAN.FS

SAN File System (SAN.FS) from IBM, also known as *Storage Tank*, implements a distributed file system on a SAN, providing shared access to virtualized storage devices for a large heterogeneous set of clients, combined with policy-based file allocation [25]. It is scalable because the clients access the storage devices directly over the SAN. This is achieved by separating meta-data operations from the data path and by breaking up the traditional client-server architecture into three components, as shown in Figure 2.

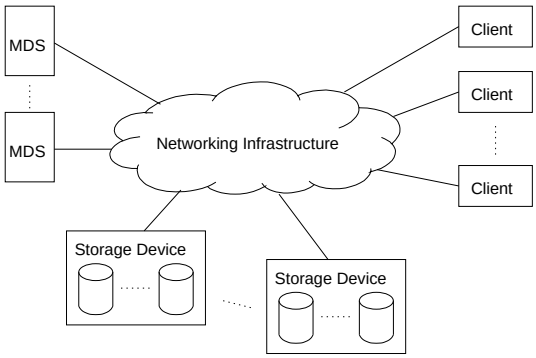

Figure 2. The architecture of SAN.FS.

The three components of SAN.FS are the following:

1. A client driver, which comes in several variations, as a VFS provider for Unix-style operating systems such as Linux and AIX, or as an installable file system for Microsoft Windows. The client driver also implements an object service (according to Section 2.1) as an intermediate layer.

2. The meta-data server (MDS), which runs on a dedicated cluster of nodes, implements all meta-data service abstractions such as directories and file meta-data, and performs lock administration for file sharing.

3. The storage devices, which are standard SAN-attached storage servers that implement a block-storage service. Note that SAN.FS does not contain a security provider, but delegates this function to the clients.

In SAN.FS, all bulk data traffic flows directly between a client and the storage devices over the SAN. The client communicates with the MDS over a LAN using TCP/IP for allocating storage space, locating data on the SAN, performing meta-data operations, and coordinating concurrent file access. The protocol between the client and the MDS is known as the *SAN.FS protocol* [16]. The MDS is responsible for data layout on the storage devices. It also implements a distributed locking protocol in which leases are given to clients for performing operations on the data [5, 16]. As the clients heavily rely on local data caching to boost performance, the MDS essentially implements a cache controller for the distributed data caches at all clients in SAN.FS.

SAN.FS maintains access control information such as file access permissions for Unix and the security descriptor for Windows in the meta-data, but leaves its interpretation up to the client operating system [16]. In order to implement proper access control for all users of a SAN.FS installation, one must therefore ensure that only trusted client machines connect to the MDS and to the SAN. It is possible to share files between Windows and Unix.

3.2. Cryptographic SAN.FS

The goal of our cryptographic SAN.FS design is to provide end-to-end confidentiality and integrity protection for the data stored by the users on the SAN.FS clients such that all cryptographic operations occur only once in the data path. We assume that the MDS is trusted to maintain cryptographic keys for encryption and reference values for integrity protection, and does not expose them to unauthorized clients. We also assume that the clients properly enforce file access control. Storage devices and other entities with access to the SAN are untrusted entities that potentially attempt to violate the security policy. Hence, using the terminology of Section 2.1, the meta-data provider also implements the security provider.

Corresponding with the design goals of SAN.FS, the client also performs the cryptographic operations and sends the protected data over the SAN to the storage devices. Encryption keys and integrity reference values are stored by the MDS as extensions of the file meta-data. The links between clients and the MDS are protected using IPsec or Kerberos. The encryption and integrity protection methods are described later in this section.

A guideline for our design was to leave the storage devices unmodified. This considerably simplifies deployment with the existing, standardized storage devices without incurring additional performance degradation. But a malicious device with access to the SAN can destroy stored data by overwriting it, because the storage devices are not capable of checking access permissions. Cryptographic integrity protection in the file system can detect such modifications, but not prevent them.

We remark that an alternative type of storage device, providing strong access control to the data, is available with object storage [2, 3]. It prevents any unauthorized modification to the data by other nodes on the SAN. Our design is orthogonal to the security design of object storage, and could easily be integrated in a SAN file system using object-storage devices.

Confidentiality Protection. The confidentiality protection mechanism encrypts the data to be stored on the clients with a symmetric cryptosystem, using a per-file encryption key. Each disk-level data block is encrypted with the AES block cipher in CBC mode, with an initialization vector derived from the file object identifier and from the offset of the block in the file and the per-file key. These choices ensure that all initialization vectors are distinct.

Instead of CBC mode, it would also be possible to use a tweakable encryption mode, such as those being considered for standardization in the IEEE P1619 effort. These modes offer better protection against active attacks on the stored data, because even a small change to an encrypted block will cause the recovered plaintext to look random and completely independent of the original plaintext. With CBC mode, an attacker can have some influence on the recovered plaintext, when no additional integrity protection method is used. Despite this deficiency, we chose CBC mode because it offers better performance (essentially twice the speed of tweakable encryption when implemented in software) and because our integrity protection scheme provides complete defense against modifications to the stored data.

The file encryption key is unique to every file and stored as part of a file's meta-data. As such a key is short (typically 16–32 bytes), the changes to the MDS for adding it are small. The key can be chosen by either the MDS or the client.

Integrity Protection. The integrity protection mechanism detects unauthorized modification of data at rest or data in flight by keeping a cryptographic hash or "digest" of every file. The hash value is short, typically 20–64 bytes with the SHA family of hash functions, and is stored together with the file meta-data by the MDS. All clients writing to the file also update the hash value at the MDS, and clients reading file data verify that any data read from storage matches the hash value obtained from the MDS. An error is reported if the data does not match the hash value.

The hash function is not applied to the complete file at once, because the hash value would have to be recomputed from scratch whenever only a part of the file changes, and data could only be verified after reading the entire file. This would incur a prohibitive overhead for large files. It is important to use a data structure that allows verification and manipulation of hash values with an effort that is roughly proportional to the amount of data affected.

The well-known solution to this problem is to create a *hash tree*, also known as *Merkle tree* [26], and to store it together with the file. A hash tree is computed from the file data by applying the hash function to every data block in the file independently and storing the resulting hash values in the leaves of the tree. The value of every interior node in the hash tree is computed by applying the hash function to the values of its children. The value at the root of the tree, which is called the *root hash value*, then represents a unique cryptographic digest of the data in the file.

A single file-data block can be verified by computing the

hash value of the block in the leaf node and by recomputing all tree nodes on the path from the leaf to the root. To recompute an interior node, all sibling nodes must be read from storage. The analogous procedure works for updates. Using hash trees, the cost of a read or a write operation of integrity-protected files is logarithmic in the length of the file (in the worst case), instead of proportional to the entire file length.

The question where to store the hash-tree data must be addressed. Conceptually, the hash tree is part of the meta-data, as it contains information *about* the file data. But apart from the root hash, no part of the hash tree must be protected [26]. Moreover, the hash tree must be updated along with every data operation and its size is proportional to the size of the file, so it resembles file data. This suggests that it should be stored together with the file data. SAN.FS, for example, uses a file-block size of 4 kB. With SHA-256 as hash function, a hash tree of degree 16 takes about 1% of the size of the corresponding file for large files.

Moreover, the SAN.FS protocol between the clients and the MDS is optimized for small messages that typically are of constant size. The protocol would require major modifications to handle the data traffic and the storage space needed by hash trees.

Therefore, we store hash-tree data on the untrusted storage space and only save the root hash value on the MDS together with the meta-data. We allocate a separate file object per file for storing hash-tree data. The existing functions for acquiring and accessing storage space can therefore be exploited for storing the hash tree. The file is visible at the object layer, but filtered out from the normal file system view of the clients. The SAN.FS distributed locking protocol is modified such that the hash tree object is tied to the corresponding data file and always covered by the locks for the data file. Without adding such a link, deadlocks might occur.

4. Implementation

We have implemented a prototype of the cryptographic SAN file system design in Linux. This section describes the extensions of the SAN.FS protocol, the modifications to the MDS, and the implementation of encryption and integrity protection in the client file system driver. The storage devices have not been modified.

4.1. SAN.FS Protocol

The clients communicate with the MDS using the SAN.FS protocol version 2.1 [16]. The SAN.FS protocol implements reliable message delivery and defines requests and transactions. Either participant can send request messages to the other participant with commands that can be executed quickly. Transactions consisting of four messages are only initiated by the client and only for executing operations that result in state changes on the server.

The SAN.FS protocol defines multiple types of locks that can be acquired by clients on file and directory objects. The *data locks* on files are relevant for the cryptographic operations. A data lock protects meta-data and file data cached locally by a client from concurrent access by other clients.

A data lock on a file object is typically held in either *shared read* or *exclusive* mode. It applies to the entire file and allows the client to read or to modify file data, respectively. When the server grants a data lock on a file object to a client, it sends along the object attributes, such as file size and access permissions.

A set of *cryptographic attributes* has been added to the object attributes. The cryptographic attributes contain the type of cryptographic protection applied (encryption, integrity, or both), the encryption method, the encryption key, the hash method, the root hash value, and the identifier of the hash-tree file object. As the object attributes are always passed to the client with a granted data lock, the client driver knows all necessary information to perform the cryptographic operations.

The most important extensions in the protocol occur for creating a file and for accessing the object attributes.

Creating a file object: When the client sends a request to create a file object, it can also specify the desired cryptographic attributes. These flags take effect for the newly created file unless the server is configured to override them. The root hash value is left empty at this time.

Accessing file object attributes: When a client requests the acquisition of a data lock to access a file, it also receives the cryptographic attributes as part of the response from the MDS. A client holding an exclusive data lock on a file object is also allowed to modify the cryptographic attributes, for example to turn on encryption. Usually the client modifies only the root hash value in accordance with the data that it writes. When the client returns an exclusive data lock to the MDS, the root hash value has to be consistent with the hash tree and the data in the file.

Apart from extending the SAN.FS protocol to handle the cryptographic data, the protocol traffic between the MDS and the client must be cryptographically protected on the network. This can either be achieved by establishing a secure IPsec tunnel between the client and the MDS or by using Kerberos to encrypt and authenticate the messages between the client and the MDS. Both forms have been implemented. IPsec can be used transparently for SAN.FS

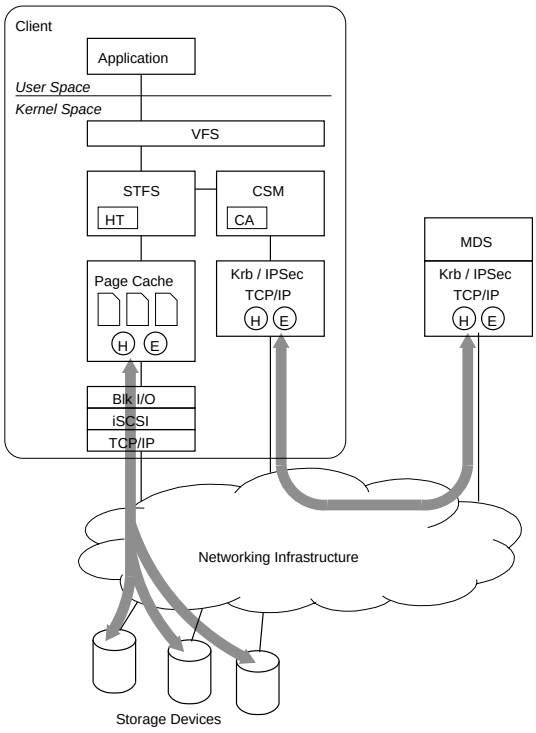

Figure 3. Design of cryptographic SAN.FS. The encircled "E" and "H" denote encryption and hash-tree operations, respectively.

because IPsec can be configured at the operating system level; it requires the server to maintain client authentication keys, however. Using Kerberos takes a small number of changes to the client driver and the MDS, and relies on the Kerberos KDC for key management.

4.2. Meta-Data Server

Only minimal changes to the MDS were required. The cryptographic attributes are stored together with existing attributes of every file object. In contrast to most operations, where the MDS merely responds to client requests, the MDS implementation takes an active role in setting the cryptographic attributes: It can be configured to enforce that the encryption and integrity protection flags be turned on or off, and to mandate the choice of particular encryption and hash methods. This allows the administrator to specify a uniform policy for the cryptographic protection applied to the file system.

The MDS can also generate an encryption key upon creation of a new file. It contains a cryptographically strong pseudorandom generator for this purpose.

4.3. Client Driver

Most of the cryptographic extensions are located in the client driver, because it performs the cryptographic operations on the bulk data. The SAN.FS client driver we used is implemented as a Linux kernel module for the 2.6.6 kernel[1]. The structure of the driver is shown in Figure 3. It consists of two main parts:

StorageTank file system driver (STFS): The STFS module contains the platform-dependent layer of the driver and implements the interface to the VFS layer of the Linux kernel. It handles reading and writing of file data from and to the page cache and the block devices.

Client state manager (CSM): The CSM is the part of the driver that interacts with the MDS using the SAN.FS protocol. It maintains the object attributes, including the cryptographic attributes. The CSM code is platform-independent and portable across all SAN.FS client driver implementations. It uses a generic interface for platform-dependent services of the operating system (not shown in the figure). Note that the CSM is not involved in reading or writing file object data.

As for any other block-device-based file system, the cached file data is maintained by the Linux page cache. All cryptographic operations operate on blocks of 4 kB at a time, which is the smallest unit of data allocation in SAN.FS. Conveniently, the page size in Linux is 4 kB or a multiple of it, so that the cryptographic operations do not have to span multiple paging operations.

Our cryptographic operations take place at the bottom of the client driver on the data path, immediately above the block-device layer. Read and write requests from the file system result in paging requests that are processed asynchronously by a pager module implemented in the client driver. The pager module consists of multiple threads for sending requests to read data from storage and for writing dirty pages back to disk. (The SAN.FS implementation in Linux does not use the Linux kernel's `pdflush` daemon.)

More concretely, when a pageout thread writes out a page of an encrypted and integrity-protected file, the thread first encrypts the page and hashes the resulting ciphertext to obtain the leaf value for the hash tree. It stores the ciphertext in a buffer page that must be allocated for the request. Then it dispatches a write request from the buffer page to the block device, according to the data layout.

For pagein requests, integrity verification and decryption take place analogously in a pagein kernel thread (in

[1] A current release of the SAN.FS client reference implementation for Linux on i386 is available from `http://www-03.ibm.com/servers/storage/software/virtualization/sfs/implementation.html` as of September 2006.

kernel-thread context), after the block device has completed the I/O request and the page has been brought in completely (in interrupt context).

The other modifications concern the CSM and its data structures, through which the link to the MDS storing the hash information is established. The extension mainly deals with processing the cryptographic attributes (CA).

The driver uses the cryptographic functions in the Linux kernel crypto API for encrypting and for hashing data. This approach enables the use of a wide range of cryptographic algorithms and dedicated hardware accelerators supporting this interface.

Implementing encryption and decryption is straightforward, but the hash-tree operations require some sophisticated algorithms. The hash-tree (HT) data is buffered in the page cache, and for every node in the tree, two flags are maintained that denote whether the node has passed verification and whether a node is dirty because a write operation to a page invalidated it. Using these flags, a pagein operation only needs to verify some nodes along the path to the root until it encounters a node that has already been verified. A pageout operation on a dirty page writes a new hash value into a leaf of the tree. The internal nodes of the hash tree are only recomputed after all dirty pages that it spans have been written out. When a file is processed sequentially (for reading or writing), buffering the hash tree in this way results in a constant processing overhead per page operation [9].

One complication that arises is that to verify the integrity of a page during a pagein operation, all corresponding hash-tree data must be ready before the page arrives and its hash value can be compared with the value in the leaf node. Because verification occurs in a kernel thread when the I/O operation is completed, it is not possible to start additional I/O operations for reading hash-tree data or allocating more memory in this context. Therefore, our implementation serializes the operations and ensures that all necessary hash-tree data is available before the pagein request is dispatched to the block device.

The design is also illustrated in Figure 3, where an encircled "E" stands for encryption and an encircled "H" stands for integrity protection operations. The arrows depict the flow of the protected data.

4.4. Hash Tree Layout

This section completes the description of the cryptographic SAN.FS client driver by illustrating the layout of the hash-tree data.

To compute the hash tree, a file is divided into 4 kB blocks, corresponding to the Linux page size. We recall the construction of a k-ary Merkle tree using a hash function $H()$: Every leaf node stores the output of H applied to a data page of length b bytes, and every internal node stores the hash value computed on the concatenation of the hash values in its children.

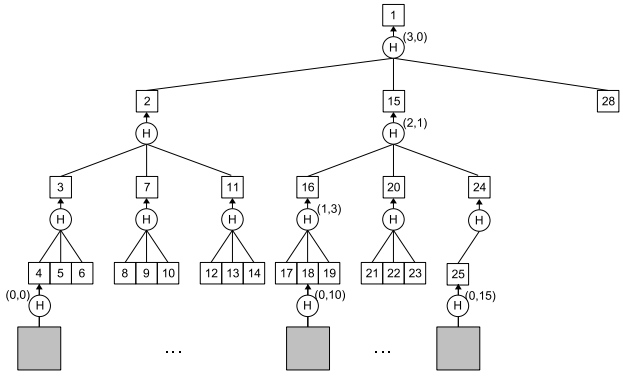

Figure 4. A ternary hash tree with four levels, numbered from 0 to 3 according to their height. The small squares represent the nodes of the tree and contain the node index in pre-order enumeration. The nodes at level 0 are the leaf nodes and are computed by hashing a single data block (grey squares). Levels 1–3 contain internal nodes, and level 3 contains the root hash.

Suppose the tree has depth t. A *level* of the tree consists of the set of nodes with the same distance from the root. Levels are numbered according to their *height* in a drawing of the inverted tree as shown in Figure 4. The height of the root node is t. Every other node has height $h-1$ if its parent has height h. Hence, leaves have height 0. The j-th node (from the left) with height h in the tree can be identified by the tuple (h, j).

As the maximum file size in SAN.FS is fixed (2^{64} bytes), the maximum depth of the hash tree can be computed in advance, given the degree k. A high degree k results in a flat tree structure and has therefore similar unfavorable properties as using a single hash value for the whole file. If k is small, the tree is deeper and therefore requires more space as well as more integrity operations during verification, especially with random-access workloads. After some experimentation with those parameters, we chose $k = 16$ and obtain a tree of depth 13 in our implementation. The complete tree with maximum depth is constructed implicitly, but every level contains only as many allocated nodes as are needed to represent the allocated blocks of the file. This choice simplifies the design of the hash-tree algorithms, in particular with respect to padding and file holes.

In particular, no leaf nodes are allocated for data blocks

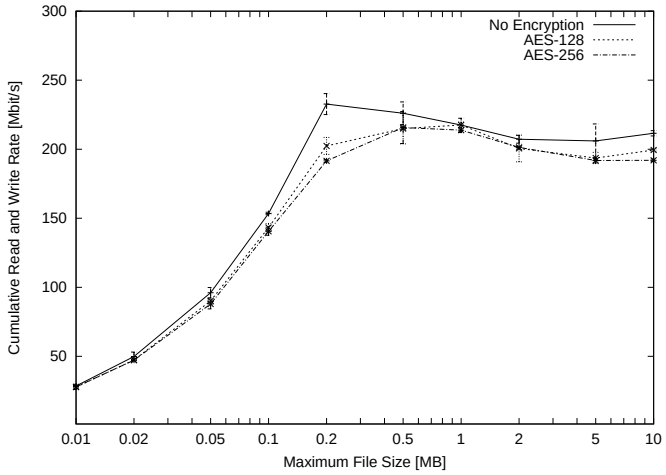

Figure 5. Encryption performance using Postmark, with varying maximum file sizes.

beyond the length of the file or for data blocks in holes. As reading such blocks would return the all-zero string according to the file-system semantics, we treat them as all-zero blocks for computing the hash tree. To prevent the allocation of hash-tree nodes covering empty file areas, the same heuristic encoding scheme is used for hash-tree nodes by the implementation: When a hash node is read from the hash-tree file object and returns the all-zero string, it is interpreted as the hash value resulting at that particular height of the tree when file data of only zeroes is hashed. This ensures that all leave nodes in the subtree rooted at this node contain only the hash of a block of zeroes and need not be allocated either. The node values for all-zero file data can be precomputed for all levels in the driver.

To serialize the hash tree, several choices are available: for example, level-by-level enumeration with two-dimensional identifiers of the form (h, j) or enumeration according to a recursive tree-traversal algorithm. Because of our choice to always implicitly maintain the hash tree for the maximum file size, enumerating the nodes according to a pre-order tree traversal is advantageous.

Figure 4 shows the typical case of contiguous file data starting at offset 0 using a ternary hash three with four levels. As can be verified easily, all hash-tree nodes that have to be allocated are also in a contiguous region in pre-order enumeration, starting with the root node at index 1. Using the heuristic encoding above, no unnecessary tree nodes have to be allocated for such files; all nodes that are to the left of the path from the highest leaf node to the root node correspond to the hash value of the all-zero file data, which are not allocated.

The nodes of the hash tree are serialized by traversing the tree in pre-order and writing every node to the file in that sequence. Some simple algorithms can be used to calculate the index of a node in pre-order enumeration from their two-dimensional identifier.

5. Performance Analysis

In this section, we report on performance measurements performed with the prototype using encryption and integrity protection. Here we give only results for Postmark [17], a benchmark creating realistic workloads, and for a synthetic benchmark, which reads and writes large amounts of data sequentially. A more detailed account of the evaluation can be found in the full paper [29].

Our testbed consists of two storage servers (one for the meta data and one for the data to be stored), an MDS, and a client. All machines are IBM x335/6 and x345/6 systems with 2 hyper-threaded Intel Xeon CPUs each and clock speeds from 2.8–3.2 GHz. The client has 3 GB RAM. The meta-data storage server contains a single drive. The data storage server contains 14 drives, organized in two RAID 5EE arrays with seven drives each, in an IBM storage expansion EXP-400 using the IBM ServeRAID 6m RAID controller. All disks are IBM Ultra320 SCSI disks with 73.4 GB capacity and running at 10k RPM. The storage devices are connected with iSCSI to the MDS and the test client over a single switched Gigabit-Ethernet.

Confidentiality Protection. Postmark is a benchmark for file-system applications and generates a file-system load similar to an Internet mail, web, or news server. It creates a large number of small sequential transactions. The read and write operations generated by the transactions are parallelized by the kernel. Figure 5 shows the cumulative read and write rate reported by Postmark v1.51, as a func-

	Unprotected [Mbit/s]	Encrypted (AES-128) [Mbit/s]	Encrypted (AES-256) [Mbit/s]	Integrity-protected [Mbit/s]
Read	458	310	279	303
Write	388	283	247	384

Table 1. Performance comparison for reading and writing large amounts of data sequentially.

tion of the maximal file size parameter. The minimum file size is being fixed to 1 kB and the maximum file size varies from 10 kB to 10 MB. In this test, Postmark is configured to create 2000 files with sizes equally distributed between the minimum and maximum configured file size and executes 5000 transactions on them. All other parameters are set to their default values in Postmark. Each curve represents the average of 11 differently seeded test runs. The 95% confidence interval is also shown, and is mostly centered closely around the mean.

It is clear that the smaller the files are, the larger is the fraction of meta-data operations. Up to a maximum file size of 200 kB, the performance is limited by the large number of meta-data operations. Above this size, we reach the limitations of the storage devices. In general we can see that the overhead for confidentiality protection is small in this benchmark and lies in the range of 5%–20%.

A second test consists of reading and writing large amount of sequential data using the Unix `dd` command. Eight files of size 1 GB each are written and read concurrently in blocks of 4 kB. The eight files are organized into two groups of four, and each group is stored on one of the RAID arrays, to avoid the disks being the performance bottleneck. The goal is to keep the file system overhead minimal in order to measure the actual end-to-end read/write performance. There are four kernel threads for pageout and pagein operations, which allows us to exploit all four available CPUs visible in Linux for encryption.

The read and write rates for AES-128 and AES-256 encryption are displayed in the second and third columns of Table 1. They are calculated from the average execution time of the eight `dd` commands, which was measured using the Unix `time` command. It is evident that for such large amounts of data, the available CPU power and CPU-to-memory bandwidth become a bottleneck for performing cryptographic operations. During reads the storage bandwidth is reduced by 32% for AES-128 and 39% for AES-256, compared to not using encryption; during writes, the reduction is about 27% for AES-128 and 36% for AES-256, respectively. The measurement, however, represents an artificial worst case for a file system. Additional tests revealed that the performance using iSCSI nullio-mode, where no data is stored on disk, achieves about 800 Mbit/s for reading and about 720 Mbit/s for writing of unencrypted data, thus saturating the Gigabit Ethernet (including the TCP/IP and iSCSI overhead).

Integrity Protection. We describe measurements with the same two benchmarks as for encryption. We ran Postmark and applied integrity protection using SHA-256. The third column of Table 2 shows the reported throughput in terms of a cumulative read and write rate for a maximum file size of 20 MB and a total number of 1000 data files. The "unprotected" case corresponds to the results reported in Figure 5. The table also shows the performance of encryption and integrity protection combined.

For the other test involving large sequential reads and writes, the third column of Table 1 contains the summarized timings with SHA-256 for integrity protection. The test uses the same setup as above. Writing shows no significant overhead because the hash tree is calculated and written to disk only after all file data has been written and therefore not included in the reported time. The hash tree size is about 1% of the size of the file. In contrast, the read operations are slower, because the hash tree data is pre-fetched and this incurs a larger latency for many low-level page-read operations. Reading may also generate a pseudo-random read access pattern to the hash-tree file.

The results show that encryption has a smaller impact on performance than integrity protection. This is actually not surprising because integrity protection involves much more complexity. Recall that our implementation first reads all hash-tree nodes necessary to verify a data page before it issues the read operation for the data page. This ensures that the completion of the page-read operation does not block because of missing data. Executing these two steps sequentially simplifies implementation but doubles the network latency of reads. Furthermore, managing the cached hash tree in memory takes some time as well.

6. Conclusion

We have presented a security architecture for cryptographic distributed file systems and its implementation in IBM SAN.FS. By protecting data on the clients before storing it on a SAN, no additional cryptography operations are necessary to secure the data in-flight on the SAN. Moreover, no additional computations by storage devices and no changes to the storage devices are required. The architec-

	Unprotected		Integrity-protected		Difference
	[MBit/s]	[%]	[MBit/s]	[%]	[%]
Unencrypted	219		156		-28.7
Encrypted with AES-128	202	-8.0	147	-5.8	-27.1
Encrypted with AES-256	198	-9.6	141	-9.7	-28.8

Table 2. Performance of integrity protection and combined encryption and integrity protection using Postmark (cumulative read and write rate). The "Unprotected" columns show the throughput without integrity protection, without encryption, with AES-128 encryption, and with AES-256 encryption. The second column denotes the relative performance loss due to using encryption. Analogously, the columns under the heading "Integrity-protected" show the rates with integrity protection applied. The fifth column "Difference" shows the relative loss due to applying integrity protection for each of the encryption choices.

ture can also be integrated with future storage devices that support access control, like object storage [3].

The implementation in SAN.FS as a monolithic cryptographic file system shows that sustained high performance can be achieved. By carefully integrating the cryptographic operations in the appropriate places of the file system driver, the overhead is actually almost not noticeable in a typical file-server environment. This is consistent with earlier benchmarks of cryptographic file systems in different environments [27, 34].

Our approach has three distinct advantages over previous systems. First, by centralizing the key management on an on-line trusted server (the MDS in our case), we gain efficiency because key management can be done with symmetric cryptography. In contrast, key management schemes performed entirely by the users, as in SFS [23] or in Windows EFS [31], requires the use of public-key cryptography.

Secondly, we believe that cryptographic integrity protection is an important requirement, even though many users of secure file systems first concentrate on encryption. Since integrity protection is also considerably more complex than encryption alone, most cryptographic file systems available today do not support it. Some systems, like SiRiUS [11], always hash entire files, and will not perform well with large files.

And, last but not least, many past designs of cryptographic file systems have chosen to simplify the implementation by using the layered approach. This limits their performance because they must maintain several data buffers. Some process data in user space, which involves copying the data in and out of the kernel multiple times. Although building the cryptographic operations into the kernel requires more work, our results show that it performs well.

Still, there is room for improvement in our design and implementation. Our hash tree implementation should include more sophisticated locking mechanisms in order to be able to read hash-tree data and file data in parallel instead of sequentially. Since an update to a file triggers at least two write operations at the block layer (one in the file data and one on hash tree data), a client crash during such an operation may violate the atomicity of the update in a more severe way than in ordinary file systems. For instance, it may be that the file data is written correctly but the integrity information is not, and an integrity violation will result upon reading. Providing a graceful recovery procedure for this situation poses an interesting and challenging open problem. Furthermore, our choice of a 16-ary hash tree was somewhat arbitrary. In ongoing work, we are exploring different hash tree topologies and alternative ways to store the hash tree. Preliminary results show that these two factors impact the file system performance.

References

[1] 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," in *Proc. 5th Symp. Operating Systems Design and Implementation (OSDI)*, 2002.

[2] A. Azagury, R. Canetti, M. Factor, S. Halevi, E. Henis, D. Naor, N. Rinetzky, O. Rodeh, and J. Satran, "A two layered approach for securing an object store network," in *Proc. 1st International IEEE Security in Storage Workshop (SISW 2002)*, 2002.

[3] A. Azagury, V. Dreizin, M. Factor, E. Henis, D. Naor, N. Rinetzky, O. Rodeh, J. Satran, A. Tavory, and L. Yerushalmi, "Towards an object store," in *Proc. IEEE/NASA Conference on Mass Storage Systems and Technologies (MSST 2003)*, pp. 165–177, 2003.

[4] M. Blaze, "A cryptographic file system for Unix," in *Proc. 1st ACM Conference on Computer and Communications Security*, Nov. 1993.

[5] R. C. Burns, R. M. Rees, L. J. Stockmeyer, and D. D. E. Long, "Scalable session locking for a distributed file system," *Cluster Computing*, vol. 4, pp. 295–306, Oct. 2001.

[6] G. Cattaneo, L. Catuogno, A. D. Sorbo, and P. Persiono, "The design and implementation of a transparent cryptographic filesystem for UNIX," in *Proc. USENIX Annual Technical Conference: FREENIX Track*, pp. 199–212, June 2001.

[7] K. Fu, F. Kaashoek, and D. Mazières, "Fast and secure distributed read-only file system," *ACM Transactions on Computer Systems*, vol. 20, pp. 1–24, Feb. 2002.

[8] K. E. Fu, *Integrity and Access Control in Untrusted Content Distribution Networks*. PhD thesis, Department of Electrical Engineering and Computer Science, MIT, Sept. 2005.

[9] B. Gassend, G. E. Suh, D. Clarke, M. van Dijk, and S. Devadas, "Caches and hash trees for efficient memory integrity verification," in *Proc. 9th Intl. Symposium on High-Performance Computer Architecture (HPCA '03)*, 2003.

[10] G. A. Gibson, D. F. Nagle, K. Amiri, F. W. Chang, E. Feinberg, H. Gobioff, C. Lee, B. Ozceri, E. Riedel, and D. Rochberg, "A case for network-attached secure disks," Tech. Rep. CMU-CS-96-142, School of Computer Science, Carnegie Mellon University, 1996.

[11] E.-J. Goh, H. Shacham, N. Modadugu, and D. Boneh, "SiRiUS: Securing remote untrusted storage," in *Proc. 10th Network and Distributed System Security Symposium (NDSS)*, pp. 131–145, Feb. 2003.

[12] V. Gough, "EncFS: Encrypted file system." `http://arg0.net/wiki/encfs`, July 2003.

[13] M. A. Halcrow *et al.*, "eCryptfs: An enterprise-class cryptographic filesystem for Linux." `http://ecryptfs.sourceforge.net/`, 2005.

[14] L. G. Harbaugh, "Encryption appliances reviewed," *Storage Magazine*, Jan. 2006.

[15] D. Hildebrand and P. Honeyman, "Exporting storage systems in a scalable manner with pNFS," in *Proc. 22nd IEEE/13th NASA Goddard Conference on Mass Storage Systems and Technologies (MSST)*, Apr. 2005.

[16] "IBM TotalStorage SAN File System Draft Protocol Specification 2.1." Available from `http://www-07.ibm.com/storage/in/software/virtualisation/sfs/protocol.html`, Sept. 2004.

[17] J. Katcher, "Postmark: A new file system benchmark," Technical Report TR3022, Network Appliance, 1997.

[18] V. Kher and Y. Kim, "Securing distributed storage: Challenges, techniques, and systems," in *Proc. Workshop on Storage Security and Survivability (StorageSS)*, 2005.

[19] G. H. Kim and E. H. Spafford, "The design and implementation of Tripwire: A file system integrity checker," in *Proc. 2nd ACM Conference on Computer and Communications Security*, pp. 18–29, 1994.

[20] J. Kubiatowicz, D. Bindel, Y. Chen, S. Czerwinski, P. Eaton, D. Geels, *et al.*, "OceanStore: An architecture for global-scale persistent storage," in *Proc. Ninth International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2000)*, Nov. 2000.

[21] J. Li, M. Krohn, D. Mazires, and D. Shasha, "Secure untrusted data repository (SUNDR)," in *Proc. 6th Symp. Operating Systems Design and Implementation (OSDI)*, pp. 121–136, 2004.

[22] D. Mazières, "A toolkit for user-level file systems," in *Proc. USENIX Annual Technical Conference*, June 2001.

[23] D. Mazières *et al.*, "Self-certifying file system." `http://www.fs.net/`, 2003.

[24] D. Mazières, M. Kaminsky, F. Kaashoek, and E. Witchel, "Separating key management from file system security," in *Proc. 17th ACM Symposium on Operating System Principles (SOSP '99)*, 1999.

[25] J. Menon, D. A. Pease, R. Rees, L. Duyanovich, and B. Hillsberg, "IBM Storage Tank — a heterogeneous scalable SAN file system," *IBM Systems Journal*, vol. 42, no. 2, pp. 250–267, 2003.

[26] R. C. Merkle, "A digital signature based on a conventional encryption function," in *Advances in Cryptology: CRYPTO '87* (C. Pomerance, ed.), vol. 293 of *Lecture Notes in Computer Science*, Springer, 1988.

[27] E. L. Miller, W. E. Freeman, D. D. E. Long, and B. C. Reed, "Strong security for network-attached storage," in *Proc. USENIX Conference on File and Storage Technologies (FAST 2002)*, 2002.

[28] S. Patil, A. Kashyap, G. Sivathanu, and E. Zadok, "I3FS: An In-Kernel Integrity Checker and Intrusion Detection File System," in *Proc. 18th USENIX Large Installation System Administration Conference (LISA 2004)*, pp. 69–79, Nov. 2004.

[29] R. Pletka and C. Cachin, "Cryptographic security for a high-performance distributed file system," Research Report RZ 3661, IBM Research, Sept. 2006.

[30] M. Rajagopal, E. G. Rodriguez, and R. Weber, "Fibre channel over TCP/IP (FCIP)." RFC 3821, July 2004.

[31] M. Russinovich, "Inside encrypting file system," *Windows & .NET magazine*, June–July 1999.

[32] J. Satran, K. Meth, C. Sapuntzakis, M. Chadalapaka, and E. Zeidner, "Internet small computer systems interface (iSCSI)." RFC 3720, Apr. 2004.

[33] C. P. Wright, M. Martino, and E. Zadok, "NCryptfs: A secure and convenient cryptographic file system," in *Proc. Annual USENIX Technical Conference*, pp. 197–210, June 2003.

[34] C. P. Wright, J. Dave, and E. Zadok, "Cryptographic file systems performance: What you don't know can hurt you," in *Proc. 2nd IEEE Security in Storage Workshop*, pp. 47–61, Oct. 2003.

[35] E. Zadok, R. Iyer, N. Joukov, G. Sivathanu, and C. P. Wright, "On incremental file system development," *ACM Transactions on Storage*, vol. 2, pp. 161–196, May 2006.

[36] E. Zadok and J. Nieh, "FiST: A language for stackable file systems," in *Proc. USENIX Annual Technical Conference*, June 2000.