Internet Engineering Task Force (IETF)
Request for Comments: 7862
Category: Standards Track
ISSN: 2070-1721

T. Haynes
Primary Data

November 2016

Network File System (NFS) Version 4 Minor Version 2 Protocol

Abstract

This document describes NFS version 4 minor version 2; it describes the protocol extensions made from NFS version 4 minor version 1. Major extensions introduced in NFS version 4 minor version 2 include the following: Server-Side Copy, Application Input/Output (I/O) Advise, Space Reservations, Sparse Files, Application Data Blocks, and Labeled NFS.

Status of This Memo

This is an Internet Standards Track document.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Further information on Internet Standards is available in Section 2 of RFC 7841.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc7862.

Table of Contents

1. Introduction

The NFS version 4 minor version 2 (NFSv4.2) protocol is the third minor version of the NFS version 4 (NFSv4) protocol. The first minor version, NFSv4.0, is described in [RFC7530], and the second minor version, NFSv4.1, is described in [RFC5661].

As a minor version, NFSv4.2 is consistent with the overall goals for NFSv4, but NFSv4.2 extends the protocol so as to better meet those goals, based on experiences with NFSv4.1. In addition, NFSv4.2 has adopted some additional goals, which motivate some of the major extensions in NFSv4.2.

1.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

1.2. Scope of This Document

This document describes the NFSv4.2 protocol as a set of extensions to the specification for NFSv4.1. That specification remains current and forms the basis for the additions defined herein. The specification for NFSv4.0 remains current as well.

It is necessary to implement all the REQUIRED features of NFSv4.1 before adding NFSv4.2 features to the implementation. With respect to NFSv4.0 and NFSv4.1, this document does not:

- describe the NFSv4.0 or NFSv4.1 protocols, except where needed to contrast with NFSv4.2
- modify the specification of the NFSv4.0 or NFSv4.1 protocols
- clarify the NFSv4.0 or NFSv4.1 protocols -- that is, any clarifications made here apply only to NFSv4.2 and not to NFSv4.0 or NFSv4.1

NFSv4.2 is a superset of NFSv4.1, with all of the new features being optional. As such, NFSv4.2 maintains the same compatibility that NFSv4.1 had with NFSv4.0. Any interactions of a new feature with NFSv4.1 semantics is described in the relevant text.

The full External Data Representation (XDR) [RFC4506] for NFSv4.2 is presented in [RFC7863].

1.3. NFSv4.2 Goals

A major goal of the enhancements provided in NFSv4.2 is to take common local file system features that have not been available through earlier versions of NFS and to offer them remotely. These features might

- already be available on the servers, e.g., sparse files
- be under development as a new standard, e.g., SEEK pulls in both SEEK_HOLE and SEEK_DATA
- be used by clients with the servers via some proprietary means, e.g., Labeled NFS

NFSv4.2 provides means for clients to leverage these features on the server in cases in which such leveraging had previously not been possible within the confines of the NFS protocol.

1.4. Overview of NFSv4.2 Features

1.4.1. Server-Side Clone and Copy

A traditional file copy of a remotely accessed file, whether from one server to another or between locations in the same server, results in the data being put on the network twice -- source to client and then client to destination. New operations are introduced to allow unnecessary traffic to be eliminated:

- The intra-server CLONE feature allows the client to request a synchronous cloning, perhaps by copy-on-write semantics.
- The intra-server COPY feature allows the client to request the server to perform the copy internally, avoiding unnecessary network traffic.
- The inter-server COPY feature allows the client to authorize the source and destination servers to interact directly.

As such copies can be lengthy, asynchronous support is also provided.

1.4.2. Application Input/Output (I/O) Advise

Applications and clients want to advise the server as to expected I/O behavior. Using IO_ADVISE (see Section 15.5) to communicate future I/O behavior such as whether a file will be accessed sequentially or randomly, and whether a file will or will not be accessed in the near future, allows servers to optimize future I/O requests for a file by, for example, prefetching or evicting data. This operation can be used to support the posix_fadvise() [posix_fadvise] function. In addition, it may be helpful to applications such as databases and video editors.

1.4.3. Sparse Files

Sparse files are files that have unallocated or uninitialized data blocks as holes in the file. Such holes are typically transferred as zeros when read from the file. READ_PLUS (see Section 15.10) allows a server to send back to the client metadata describing the hole, and DEALLOCATE (see Section 15.4) allows the client to punch holes into a file. In addition, SEEK (see Section 15.11) is provided to scan for the next hole or data from a given location.

1.4.4. Space Reservation

When a file is sparse, one concern that applications have is ensuring that there will always be enough data blocks available for the file during future writes. ALLOCATE (see Section 15.1) allows a client to request a guarantee that space will be available. Also, DEALLOCATE (see Section 15.4) allows the client to punch a hole into a file, thus releasing a space reservation.

1.4.5. Application Data Block (ADB) Support

Some applications treat a file as if it were a disk and as such want to initialize (or format) the file image. The WRITE_SAME operation (see Section 15.12) is introduced to send this metadata to the server to allow it to write the block contents.

1.4.6. Labeled NFS

While both clients and servers can employ Mandatory Access Control (MAC) security models to enforce data access, there has been no protocol support for interoperability. A new file object attribute, sec_label (see Section 12.2.4), allows the server to store MAC labels on files, which the client retrieves and uses to enforce data access (see Section 9.5.3). The format of the sec_label accommodates any MAC security system.

1.4.7. Layout Enhancements

In the parallel NFS implementations of NFSv4.1 (see Section 12 of [RFC5661]), the client cannot communicate back to the metadata server any errors or performance characteristics with the storage devices. NFSv4.2 provides two new operations to do so: LAYOUTERROR (see Section 15.6) and LAYOUTSTATS (see Section 15.7), respectively.

1.5. Enhancements to Minor Versioning Model

In NFSv4.1, the only way to introduce new variants of an operation was to introduce a new operation. For instance, READ would have to be replaced or supplemented by, say, either READ2 or READ_PLUS. With the use of discriminated unions as parameters for such functions in NFSv4.2, it is possible to add a new "arm" (i.e., a new entry in the union and a corresponding new field in the structure) in a subsequent minor version. It is also possible to move such an operation from OPTIONAL/RECOMMENDED to REQUIRED. Forcing an implementation to adopt each arm of a discriminated union at such a time does not meet the spirit of the minor versioning rules. As such, new arms of a discriminated union MUST follow the same guidelines for minor

versioning as operations in NFSv4.1 -- i.e., they may not be made REQUIRED. To support this, a new error code, NFS4ERR_UNION_NOTSUPP, allows the server to communicate to the client that the operation is supported but the specific arm of the discriminated union is not.

2. Minor Versioning

NFSv4.2 is a minor version of NFSv4 and is built upon NFSv4.1 as documented in [RFC5661] and [RFC5662].

NFSv4.2 does not modify the rules applicable to the NFSv4 versioning process and follows the rules set out in [RFC5661] or in Standards Track documents updating that document (e.g., in an RFC based on [NFSv4-Versioning]).

NFSv4.2 only defines extensions to NFSv4.1, each of which may be supported (or not) independently. It does not

- introduce infrastructural features
- make existing features MANDATORY to NOT implement
- change the status of existing features (i.e., by changing their status among OPTIONAL, RECOMMENDED, REQUIRED)

The following versioning-related considerations should be noted.

- When a new case is added to an existing switch, servers need to report non-support of that new case by returning NFS4ERR_UNION_NOTSUPP.
- As regards the potential cross-minor-version transfer of stateids, Parallel NFS (pNFS) (see Section 12 of [RFC5661]) implementations of the file-mapping type may support the use of an NFSv4.2 metadata server (see Sections 1.7.2.2 and 12.2.2 of [RFC5661]) with NFSv4.1 data servers. In this context, a stateid returned by an NFSv4.2 COMPOUND will be used in an NFSv4.1 COMPOUND directed to the data server (see Sections 3.2 and 3.3).

3. pNFS Considerations for New Operations

The interactions of the new operations with non-pNFS functionality are straightforward and are covered in the relevant sections. However, the interactions of the new operations with pNFS are more complicated. This section provides an overview.

3.1. Atomicity for ALLOCATE and DEALLOCATE

Both ALLOCATE (see Section 15.1) and DEALLOCATE (see Section 15.4) are sent to the metadata server, which is responsible for coordinating the changes onto the storage devices. In particular, both operations must either fully succeed or fail; it cannot be the case that one storage device succeeds whilst another fails.

3.2. Sharing of Stateids with NFSv4.1

An NFSv4.2 metadata server can hand out a layout to an NFSv4.1 storage device. Section 13.9.1 of [RFC5661] discusses how the client gets a stateid from the metadata server to present to a storage device.

3.3. NFSv4.2 as a Storage Protocol in pNFS: The File Layout Type

A file layout provided by an NFSv4.2 server may refer to either (1) a storage device that only implements NFSv4.1 as specified in [RFC5661] or (2) a storage device that implements additions from NFSv4.2, in which case the rules in Section 3.3.1 apply. As the file layout type does not provide a means for informing the client as to which minor version a particular storage device is providing, the client will have to negotiate this with the storage device via the normal Remote Procedure Call (RPC) semantics of major and minor version discovery. For example, as per Section 16.2.3 of [RFC5661], the client could try a COMPOUND with a minorversion field value of 2; if it gets NFS4ERR_MINOR_VERS_MISMATCH, it would drop back to 1.

3.3.1. Operations Sent to NFSv4.2 Data Servers

In addition to the commands listed in [RFC5661], NFSv4.2 data servers MAY accept a COMPOUND containing the following additional operations: IO_ADVISE (see Section 15.5), READ_PLUS (see Section 15.10), WRITE_SAME (see Section 15.12), and SEEK (see Section 15.11), which will be treated like the subset specified as "Operations Sent to NFSv4.1 Data Servers" in Section 13.6 of [RFC5661].

Additional details on the implementation of these operations in a pNFS context are documented in the operation-specific sections.

4. Server-Side Copy

The server-side copy features provide mechanisms that allow an NFS client to copy file data on a server or between two servers without the data being transmitted back and forth over the network through the NFS client. Without these features, an NFS client would copy data from one location to another by reading the data from the source server over the network and then writing the data back over the network to the destination server.

If the source object and destination object are on different file servers, the file servers will communicate with one another to perform the COPY operation. The server-to-server protocol by which this is accomplished is not defined in this document.

The copy feature allows the server to perform the copying either synchronously or asynchronously. The client can request synchronous copying, but the server may not be able to honor this request. If the server intends to perform asynchronous copying, it supplies the client with a request identifier that the client can use to monitor the progress of the copying and, if appropriate, cancel a request in progress. The request identifier is a stateid representing the internal state held by the server while the copying is performed. Multiple asynchronous copies of all or part of a file may be in progress in parallel on a server; the stateid request identifier allows monitoring and canceling to be applied to the correct request.

4.1. Protocol Overview

The server-side copy offload operations support both intra-server and inter-server file copies. An intra-server copy is a copy in which the source file and destination file reside on the same server. In an inter-server copy, the source file and destination file are on different servers. In both cases, the copy may be performed synchronously or asynchronously.

In addition, the CLONE operation provides COPY-like functionality in the intra-server case, which is both synchronous and atomic in that other operations may not see the target file in any state between the state before the CLONE operation and the state after it.

Throughout the rest of this document, the NFS server containing the source file is referred to as the "source server" and the NFS server to which the file is transferred as the "destination server". In the case of an intra-server copy, the source server and destination server are the same server. Therefore, in the context of an intra-server copy, the terms "source server" and "destination server" refer to the single server performing the copy.

The new operations are designed to copy files or regions within them. Other file system objects can be copied by building on these operations or using other techniques. For example, if a user wishes to copy a directory, the client can synthesize a directory COPY operation by first creating the destination directory and the individual (empty) files within it and then copying the contents of the source directory's files to files in the new destination directory.

For the inter-server copy, the operations are defined to be compatible with the traditional copy authorization approach. The client and user are authorized at the source for reading. Then, they are authorized at the destination for writing.

4.1.1. COPY Operations

CLONE: Used by the client to request a synchronous atomic COPY-like operation. (Section 15.13)

COPY_NOTIFY: Used by the client to request the source server to authorize a future file copy that will be made by a given destination server on behalf of the given user. (Section 15.3)

COPY: Used by the client to request a file copy. (Section 15.2)

OFFLOAD_CANCEL: Used by the client to terminate an asynchronous file copy. (Section 15.8)

OFFLOAD_STATUS: Used by the client to poll the status of an asynchronous file copy. (Section 15.9)

CB_OFFLOAD: Used by the destination server to report the results of an asynchronous file copy to the client. (Section 16.1)

4.1.2. Requirements for Operations

Inter-server copy, intra-server copy, and intra-server clone are each OPTIONAL features in the context of server-side copy. A server may choose independently to implement any of them. A server implementing any of these features may be REQUIRED to implement certain operations. Other operations are OPTIONAL in the context of a particular feature (see Table 5 in Section 13) but may become REQUIRED, depending on server behavior. Clients need to use these operations to successfully copy a file.

For a client to do an intra-server file copy, it needs to use either the COPY or the CLONE operation. If COPY is used, the client MUST support the CB_OFFLOAD operation. If COPY is used and it returns a stateid, then the client MAY use the OFFLOAD_CANCEL and OFFLOAD_STATUS operations.

For a client to do an inter-server file copy, it needs to use the COPY and COPY_NOTIFY operations and MUST support the CB_OFFLOAD operation. If COPY returns a stateid, then the client MAY use the OFFLOAD_CANCEL and OFFLOAD_STATUS operations.

If a server supports the intra-server COPY feature, then the server MUST support the COPY operation. If a server's COPY operation returns a stateid, then the server MUST also support these operations: CB_OFFLOAD, OFFLOAD_CANCEL, and OFFLOAD_STATUS.

If a server supports the CLONE feature, then it MUST support the CLONE operation and the clone_blksize attribute on any file system on which CLONE is supported (as either source or destination file).

If a source server supports the inter-server COPY feature, then it MUST support the COPY_NOTIFY and OFFLOAD_CANCEL operations. If a destination server supports the inter-server COPY feature, then it MUST support the COPY operation. If a destination server's COPY operation returns a stateid, then the destination server MUST also support these operations: CB_OFFLOAD, OFFLOAD_CANCEL, COPY_NOTIFY, and OFFLOAD_STATUS.

Each operation is performed in the context of the user identified by the Open Network Computing (ONC) RPC credential in the RPC request containing the COMPOUND or CB_COMPOUND request. For example, an OFFLOAD_CANCEL operation issued by a given user indicates that a specified COPY operation initiated by the same user is to be canceled. Therefore, an OFFLOAD_CANCEL MUST NOT interfere with a copy of the same file initiated by another user.

An NFS server MAY allow an administrative user to monitor or cancel COPY operations using an implementation-specific interface.

4.2. Requirements for Inter-Server Copy

The specification of the inter-server copy is driven by several requirements:

- The specification MUST NOT mandate the server-to-server protocol.
- The specification MUST provide guidance for using NFSv4.x as a copy protocol. For those source and destination servers willing to use NFSv4.x, there are specific security considerations that the specification MUST address.
- The specification MUST NOT mandate preconfiguration between the source and destination servers. Requiring that the source and destination servers first have a "copying relationship" increases the administrative burden. However, the specification MUST NOT preclude implementations that require preconfiguration.
- The specification MUST NOT mandate a trust relationship between the source and destination servers. The NFSv4 security model requires mutual authentication between a principal on an NFS client and a principal on an NFS server. This model MUST continue with the introduction of COPY.

4.3. Implementation Considerations

4.3.1. Locking the Files

Both the source file and the destination file may need to be locked to protect the content during the COPY operations. A client can achieve this by a combination of OPEN and LOCK operations. That is, either share locks or byte-range locks might be desired.

Note that when the client establishes a lock stateid on the source, the context of that stateid is for the client and not the destination. As such, there might already be an outstanding stateid, issued to the destination as the client of the source, with the same value as that provided for the lock stateid. The source MUST interpret the lock stateid as that of the client, i.e., when the destination presents it in the context of an inter-server copy, it is on behalf of the client.

4.3.2. Client Caches

In a traditional copy, if the client is in the process of writing to the file before the copy (and perhaps with a write delegation), it will be straightforward to update the destination server. With an inter-server copy, the source has no insight into the changes cached on the client. The client SHOULD write the data back to the source. If it does not do so, it is possible that the destination will receive a corrupt copy of the file.

4.4. Intra-Server Copy

To copy a file on a single server, the client uses a COPY operation. The server may respond to the COPY operation with the final results of the copy, or it may perform the copy asynchronously and deliver the results using a CB_OFFLOAD callback operation. If the copy is performed asynchronously, the client may poll the status of the copy using OFFLOAD_STATUS or cancel the copy using OFFLOAD_CANCEL.

A synchronous intra-server copy is shown in Figure 1. In this example, the NFS server chooses to perform the copy synchronously. The COPY operation is completed, either successfully or unsuccessfully, before the server replies to the client's request. The server's reply contains the final result of the operation.

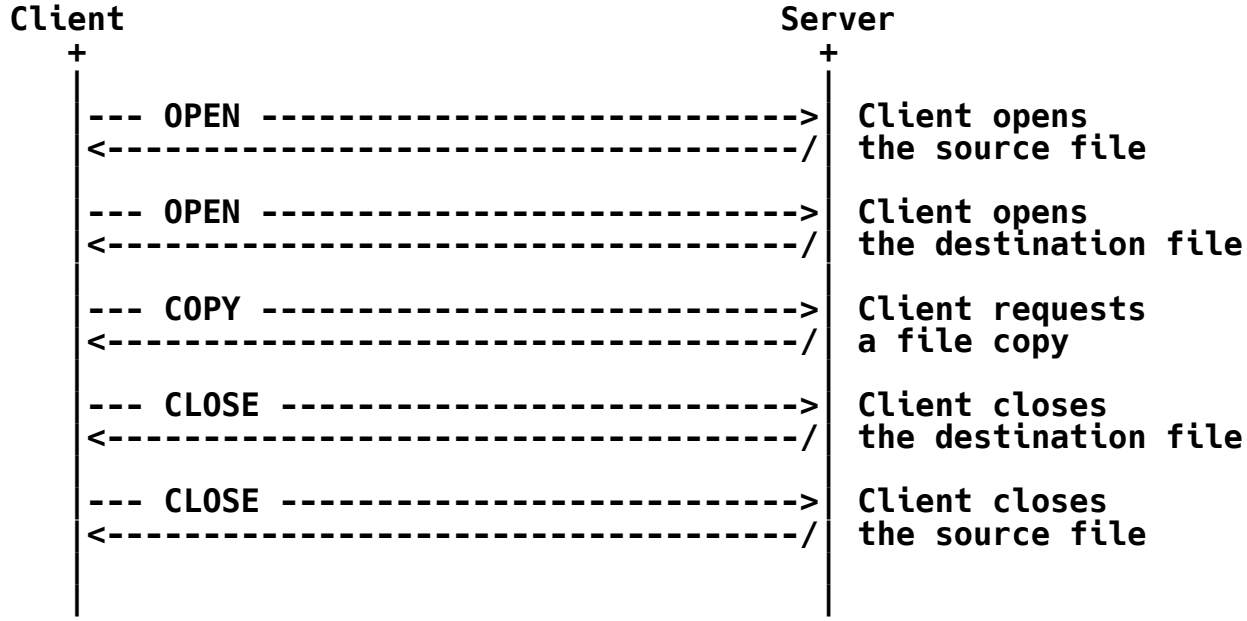


Figure 1: A Synchronous Intra-Server Copy

An asynchronous intra-server copy is shown in Figure 2. In this example, the NFS server performs the copy asynchronously. The server's reply to the copy request indicates that the COPY operation was initiated and the final result will be delivered at a later time. The server's reply also contains a copy stateid. The client may use this copy stateid to poll for status information (as shown) or to cancel the copy using an OFFLOAD_CANCEL. When the server completes the copy, the server performs a callback to the client and reports the results.

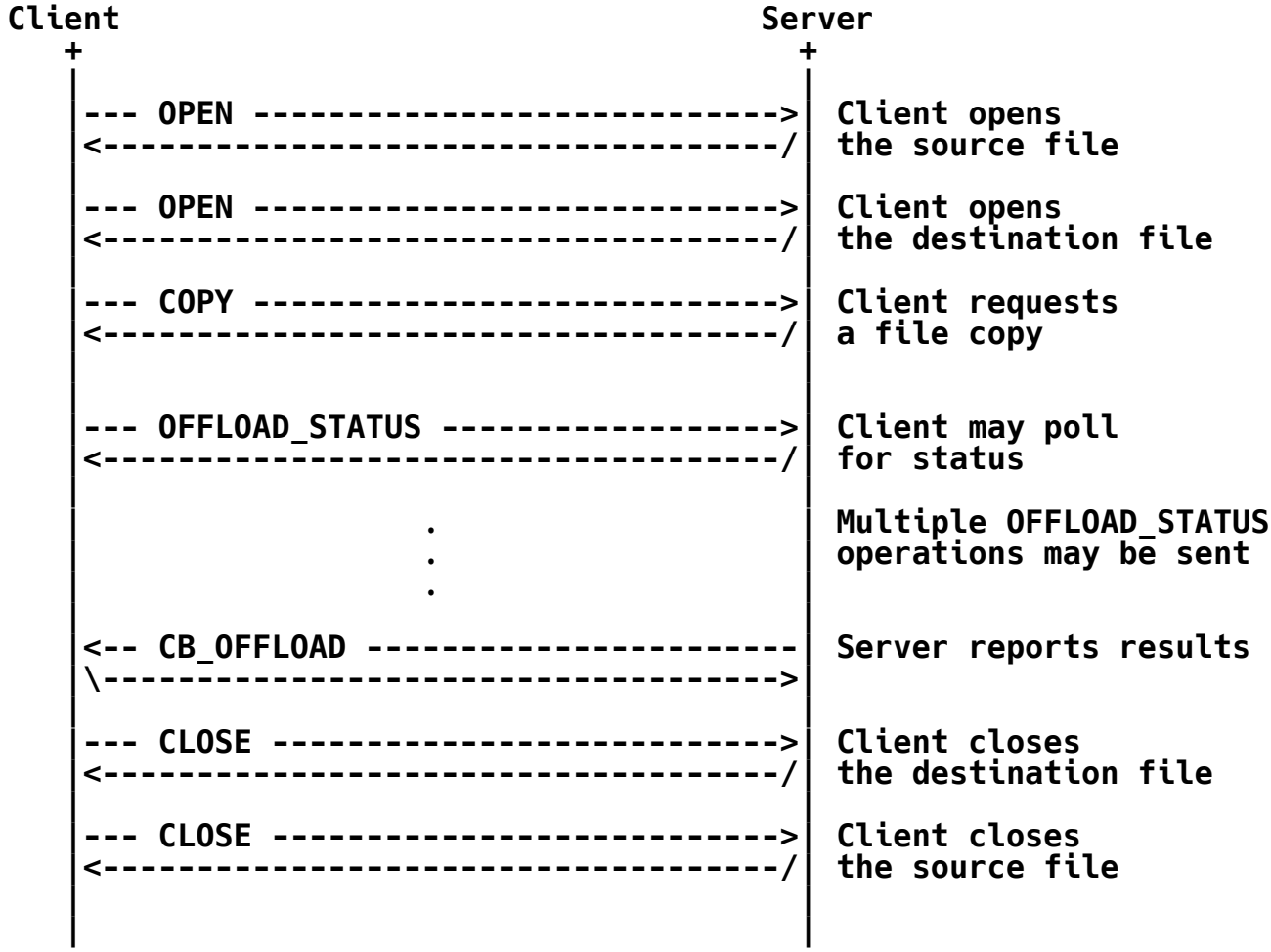


Figure 2: An Asynchronous Intra-Server Copy

4.5. Inter-Server Copy

A copy may also be performed between two servers. The copy protocol is designed to accommodate a variety of network topologies. As shown in Figure 3, the client and servers may be connected by multiple networks. In particular, the servers may be connected by a specialized, high-speed network (network 192.0.2.0/24 in the diagram) that does not include the client. The protocol allows the client to set up the copy between the servers (over network 203.0.113.0/24 in the diagram) and for the servers to communicate on the high-speed network if they choose to do so.

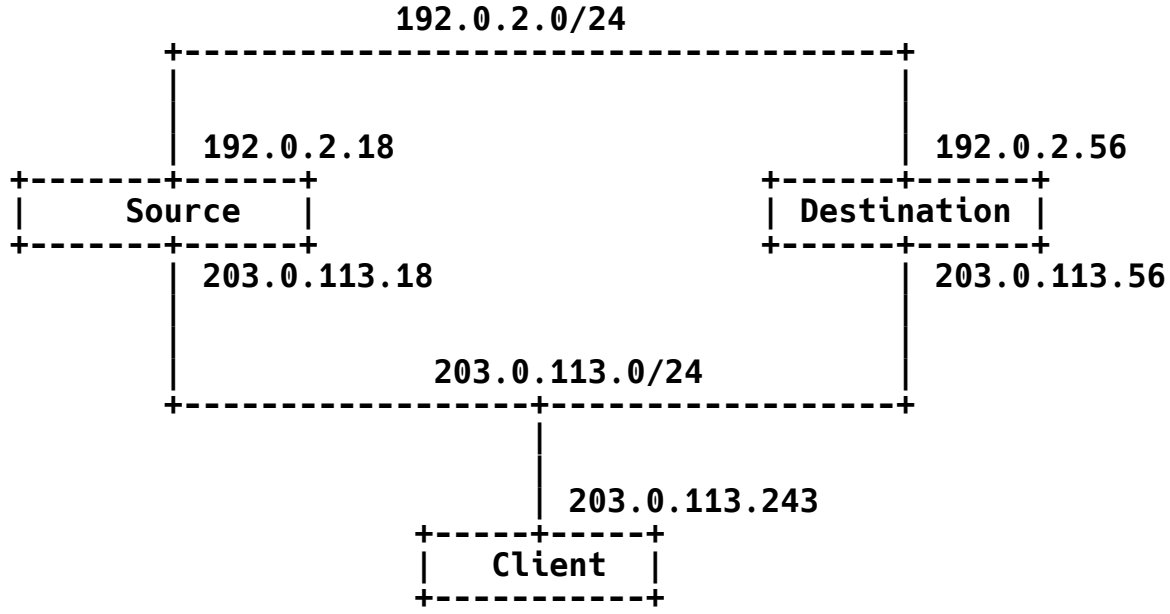


Figure 3: An Example Inter-Server Network Topology

For an inter-server copy, the client notifies the source server that a file will be copied by the destination server using a COPY_NOTIFY operation. The client then initiates the copy by sending the COPY operation to the destination server. The destination server may perform the copy synchronously or asynchronously.

A synchronous inter-server copy is shown in Figure 4. In this case, the destination server chooses to perform the copy before responding to the client's COPY request.

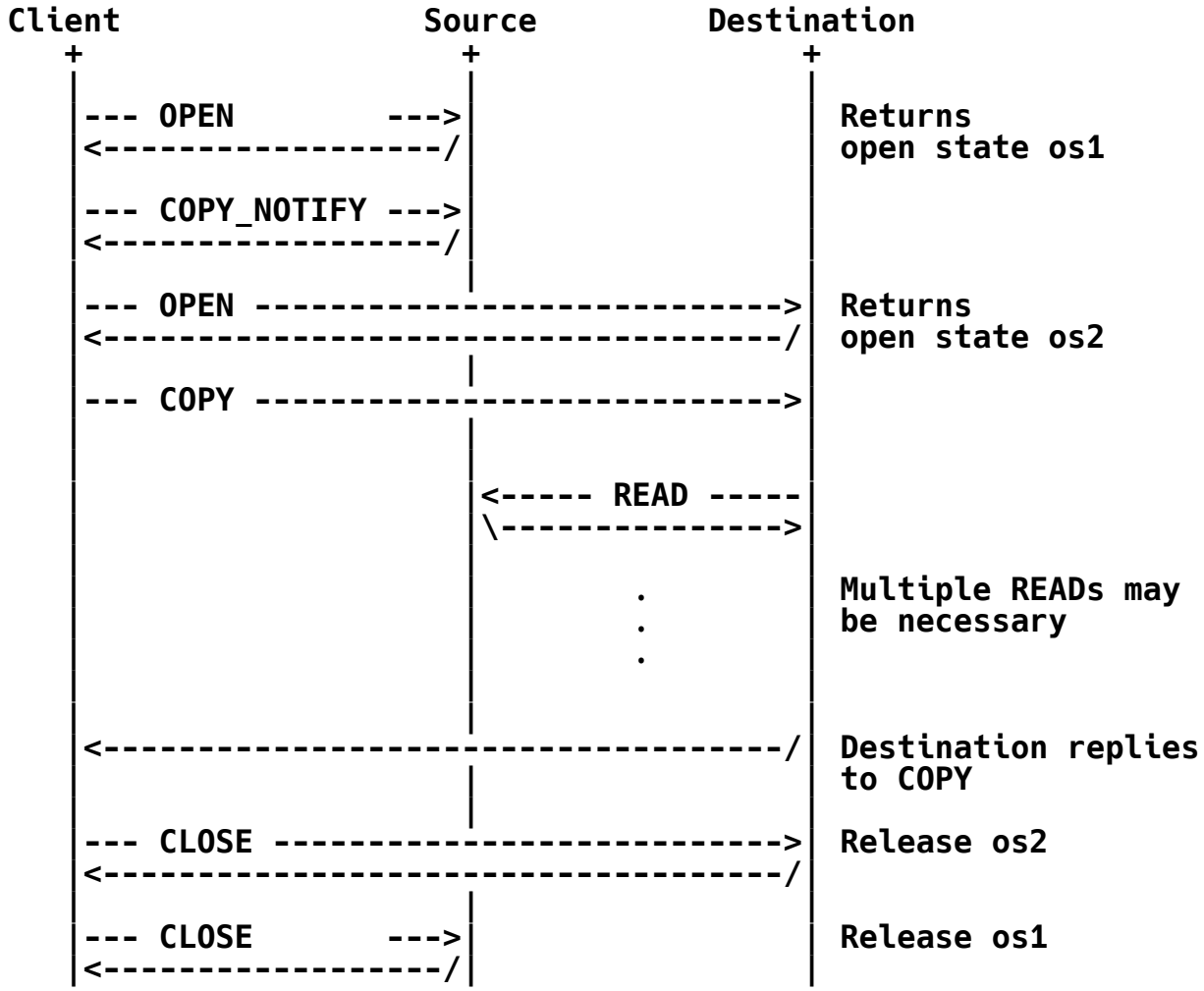
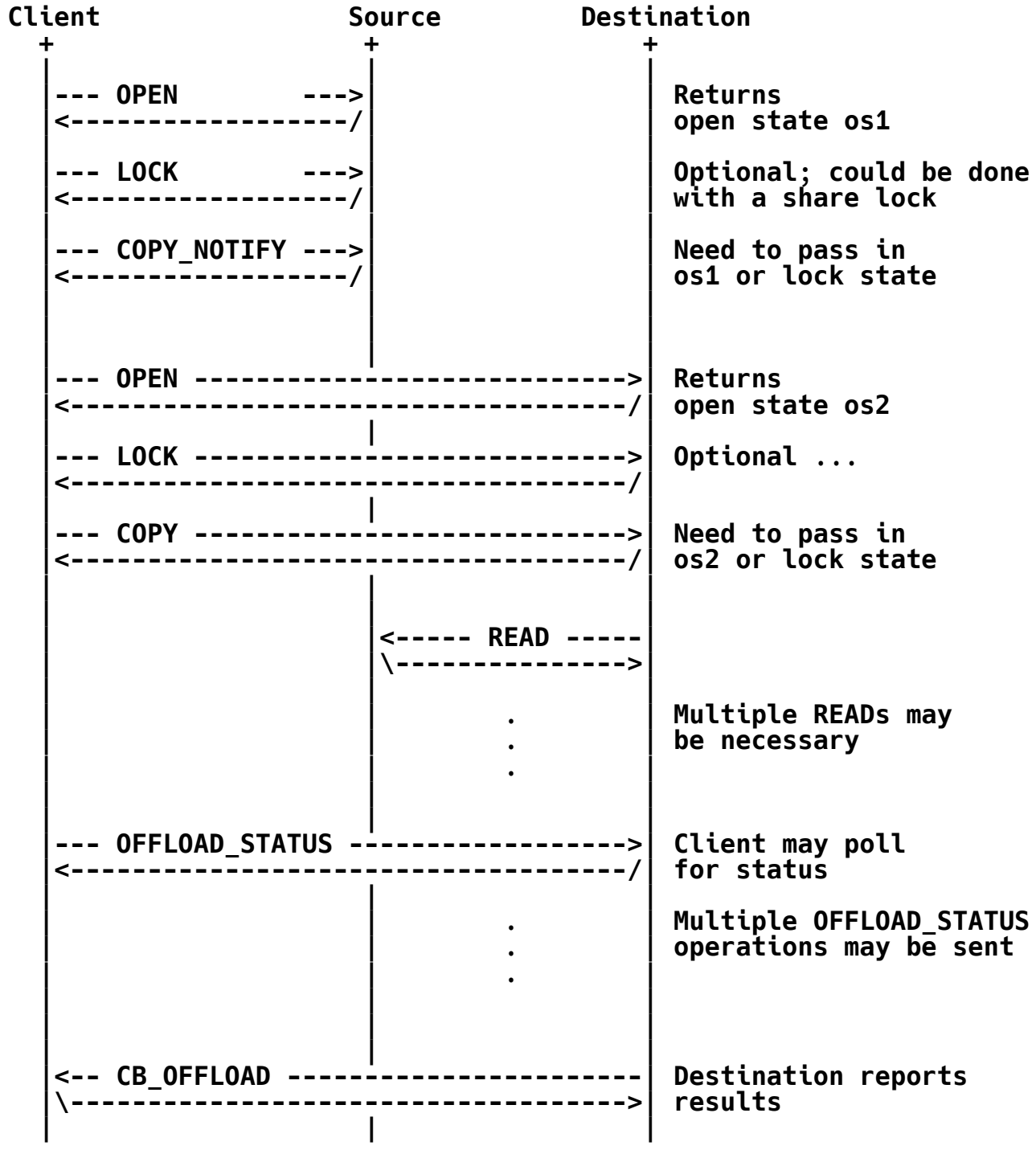


Figure 4: A Synchronous Inter-Server Copy

An asynchronous inter-server copy is shown in Figure 5. In this case, the destination server chooses to respond to the client's COPY request immediately and then perform the copy asynchronously.



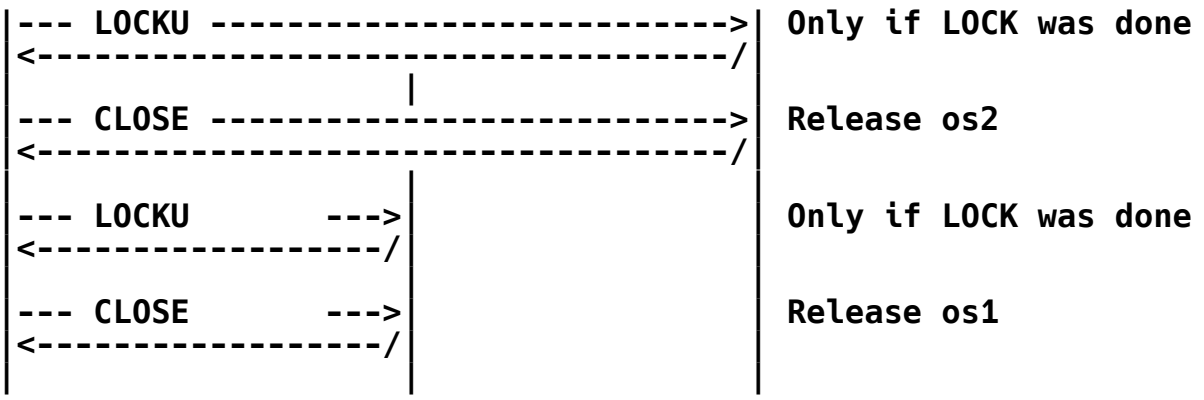


Figure 5: An Asynchronous Inter-Server Copy

4.6. Server-to-Server Copy Protocol

The choice of what protocol to use in an inter-server copy is ultimately the destination server's decision. However, the destination server has to be cognizant that it is working on behalf of the client.

4.6.1. Considerations on Selecting a Copy Protocol

The client can have requirements over both the size of transactions and error recovery semantics. It may want to split the copy up such that each chunk is synchronously transferred. It may want the copy protocol to copy the bytes in consecutive order such that upon an error the client can restart the copy at the last known good offset. If the destination server cannot meet these requirements, the client may prefer the traditional copy mechanism such that it can meet those requirements.

4.6.2. Using NFSv4.x as the Copy Protocol

The destination server MAY use standard NFSv4.x (where x >= 1) operations to read the data from the source server. If NFSv4.x is used for the server-to-server copy protocol, the destination server can use the source filehandle and ca_src_stateid provided in the COPY request with standard NFSv4.x operations to read data from the source server. Note that the ca_src_stateid MUST be the cnr_stateid returned from the source via the COPY_NOTIFY (Section 15.3).

4.6.3. Using an Alternative Copy Protocol

In a homogeneous environment, the source and destination servers might be able to perform the file copy extremely efficiently using specialized protocols. For example, the source and destination servers might be two nodes sharing a common file system format for the source and destination file systems. Thus, the source and destination are in an ideal position to efficiently render the image of the source file to the destination file by replicating the file system formats at the block level. Another possibility is that the source and destination might be two nodes sharing a common storage area network, and thus there is no need to copy any data at all; instead, ownership of the file and its contents might simply be reassigned to the destination. To allow for these possibilities, the destination server is allowed to use a server-to-server copy protocol of its choice.

In a heterogeneous environment, using a protocol other than NFSv4.x (e.g., HTTP [RFC7230] or FTP [RFC959]) presents some challenges. In particular, the destination server is presented with the challenge of accessing the source file given only an NFSv4.x filehandle.

One option for protocols that identify source files with pathnames is to use an ASCII hexadecimal representation of the source filehandle as the filename.

Another option for the source server is to use URLs to direct the destination server to a specialized service. For example, the response to COPY_NOTIFY could include the URL <ftp://s1.example.com:9999/_FH/0x12345>, where 0x12345 is the ASCII hexadecimal representation of the source filehandle. When the destination server receives the source server's URL, it would use "_FH/0x12345" as the filename to pass to the FTP server listening on port 9999 of s1.example.com. On port 9999 there would be a special instance of the FTP service that understands how to convert NFS filehandles to an open file descriptor (in many operating systems, this would require a new system call, one that is the inverse of the makefh() function that the pre-NFSv4 MOUNT service needs).

Authenticating and identifying the destination server to the source server is also a challenge. One solution would be to construct unique URLs for each destination server.

4.7. netloc4 - Network Locations

The server-side COPY operations specify network locations using the netloc4 data type shown below (see [RFC7863]):

<CODE BEGINS>

```
enum netloc_type4 {
        NL4_NAME        = 1,
        NL4_URL         = 2,
        NL4_NETADDR     = 3
};

union netloc4 switch (netloc_type4 nl_type) {
        case NL4_NAME:          utf8str_cis nl_name;
        case NL4_URL:           utf8str_cis nl_url;
        case NL4_NETADDR:       netaddr4    nl_addr;
};
```

<CODE ENDS>

If the netloc4 is of type NL4_NAME, the nl_name field MUST be specified as a UTF-8 string. The nl_name is expected to be resolved to a network address via DNS, the Lightweight Directory Access Protocol (LDAP), the Network Information Service (NIS), /etc/hosts, or some other means. If the netloc4 is of type NL4_URL, a server URL [RFC3986] appropriate for the server-to-server COPY operation is specified as a UTF-8 string. If the netloc4 is of type NL4_NETADDR, the nl_addr field MUST contain a valid netaddr4 as defined in Section 3.3.9 of [RFC5661].

When netloc4 values are used for an inter-server copy as shown in Figure 3, their values may be evaluated on the source server, destination server, and client. The network environment in which these systems operate should be configured so that the netloc4 values are interpreted as intended on each system.

4.8. Copy Offload Stateids

A server may perform a copy offload operation asynchronously. An asynchronous copy is tracked using a copy offload stateid. Copy offload stateids are included in the COPY, OFFLOAD_CANCEL, OFFLOAD_STATUS, and CB_OFFLOAD operations.

A copy offload stateid will be valid until either (A) the client or server restarts or (B) the client returns the resource by issuing an OFFLOAD_CANCEL operation or the client replies to a CB_OFFLOAD operation.

A copy offload stateid's seqid MUST NOT be zero. In the context of a copy offload operation, it is inappropriate to indicate "the most recent copy offload operation" using a stateid with a seqid of zero (see Section 8.2.2 of [RFC5661]). It is inappropriate because the stateid refers to internal state in the server and there may be several asynchronous COPY operations being performed in parallel on the same file by the server. Therefore, a copy offload stateid with a seqid of zero MUST be considered invalid.

4.9. Security Considerations for Server-Side Copy

All security considerations pertaining to NFSv4.1 [RFC5661] apply to this section; as such, the standard security mechanisms used by the protocol can be used to secure the server-to-server operations.

NFSv4 clients and servers supporting the inter-server COPY operations described in this section are REQUIRED to implement the mechanism described in Section 4.9.1.1 and to support rejecting COPY_NOTIFY requests that do not use the RPC security protocol (RPCSEC_GSS) [RFC7861] with privacy. If the server-to-server copy protocol is based on ONC RPC, the servers are also REQUIRED to implement [RFC7861], including the RPCSEC_GSSv3 "copy_to_auth", "copy_from_auth", and "copy_confirm_auth" structured privileges. This requirement to implement is not a requirement to use; for example, a server may, depending on configuration, also allow COPY_NOTIFY requests that use only AUTH_SYS.

If a server requires the use of an RPCSEC_GSSv3 copy_to_auth, copy_from_auth, or copy_confirm_auth privilege and it is not used, the server will reject the request with NFS4ERR_PARTNER_NO_AUTH.

4.9.1. Inter-Server Copy Security

4.9.1.1. Inter-Server Copy via ONC RPC with RPCSEC_GSSv3

When the client sends a COPY_NOTIFY to the source server to expect the destination to attempt to copy data from the source server, it is expected that this copy is being done on behalf of the principal (called the "user principal") that sent the RPC request that encloses the COMPOUND procedure that contains the COPY_NOTIFY operation. The user principal is identified by the RPC credentials. A mechanism that allows the user principal to authorize the destination server to perform the copy, lets the source server properly authenticate the destination's copy, and does not allow the destination server to exceed this authorization is necessary.

An approach that sends delegated credentials of the client's user principal to the destination server is not used for the following reason. If the client's user delegated its credentials, the destination would authenticate as the user principal. If the destination were using the NFSv4 protocol to perform the copy, then the source server would authenticate the destination server as the user principal, and the file copy would securely proceed. However, this approach would allow the destination server to copy other files. The user principal would have to trust the destination server to not do so. This is counter to the requirements and therefore is not considered.

Instead, a feature of the RPCSEC_GSSv3 protocol [RFC7861] can be used: RPC-application-defined structured privilege assertion. This feature allows the destination server to authenticate to the source server as acting on behalf of the user principal and to authorize the destination server to perform READs of the file to be copied from the source on behalf of the user principal. Once the copy is complete, the client can destroy the RPCSEC_GSSv3 handles to end the authorization of both the source and destination servers to copy.

For each structured privilege assertion defined by an RPC application, RPCSEC_GSSv3 requires the application to define a name string and a data structure that will be encoded and passed between client and server as opaque data. For NFSv4, the data structures specified below MUST be serialized using XDR.

Three RPCSEC_GSSv3 structured privilege assertions that work together to authorize the copy are defined here. For each of the assertions, the description starts with the name string passed in the rp_name field of the rgss3_privs structure defined in Section 2.7.1.4 of [RFC7861] and specifies the XDR encoding of the associated structured data passed via the rp_privilege field of the structure.

copy_from_auth: A user principal is authorizing a source principal ("nfs@<source>") to allow a destination principal ("nfs@<destination>") to set up the copy_confirm_auth privilege required to copy a file from the source to the destination on behalf of the user principal. This privilege is established on the source server before the user principal sends a COPY_NOTIFY operation to the source server, and the resultant RPCSEC_GSSv3 context is used to secure the COPY_NOTIFY operation.

<CODE BEGINS>

```
struct copy_from_auth_priv {
        secret4             cfap_shared_secret;
        netloc4             cfap_destination;
        /* the NFSv4 user name that the user principal maps to */
        utf8str_mixed       cfap_username;
};
```

<CODE ENDS>

cfap_shared_secret is an automatically generated random number secret value.

copy_to_auth: A user principal is authorizing a destination principal ("nfs@<destination>") to set up a copy_confirm_auth privilege with a source principal ("nfs@<source>") to allow it to copy a file from the source to the destination on behalf of the user principal. This privilege is established on the destination server before the user principal sends a COPY operation to the destination server, and the resultant RPCSEC_GSSv3 context is used to secure the COPY operation.

<CODE BEGINS>

```
struct copy_to_auth_priv {
        /* equal to cfap_shared_secret */
        secret4              ctap_shared_secret;
        netloc4              ctap_source<>;
        /* the NFSv4 user name that the user principal maps to */
        utf8str_mixed        ctap_username;
};
```

<CODE ENDS>

ctap_shared_secret is the automatically generated secret value used to establish the copy_from_auth privilege with the source principal. See Section 4.9.1.1.1.

copy_confirm_auth: A destination principal ("nfs@<destination>") is confirming with the source principal ("nfs@<source>") that it is authorized to copy data from the source. This privilege is established on the destination server before the file is copied from the source to the destination. The resultant RPCSEC_GSSv3 context is used to secure the READ operations from the source to the destination server.

<CODE BEGINS>

```
struct copy_confirm_auth_priv {
        /* equal to GSS_GetMIC() of cfap_shared_secret */
        opaque              ccap_shared_secret_mic<>;
        /* the NFSv4 user name that the user principal maps to */
        utf8str_mixed       ccap_username;
};
```

<CODE ENDS>

4.9.1.1.1. Establishing a Security Context

When the user principal wants to copy a file between two servers, if it has not established copy_from_auth and copy_to_auth privileges on the servers, it establishes them as follows:

- As noted in [RFC7861], the client uses an existing RPCSEC_GSSv3 context termed the "parent" handle to establish and protect RPCSEC_GSSv3 structured privilege assertion exchanges. The copy_from_auth privilege will use the context established between the user principal and the source server used to OPEN the source file as the RPCSEC_GSSv3 parent handle. The copy_to_auth privilege will use the context established between the user principal and the destination server used to OPEN the destination file as the RPCSEC_GSSv3 parent handle.

- A random number is generated to use as a secret to be shared between the two servers. Note that the random number SHOULD NOT be reused between establishing different security contexts. The resulting shared secret will be placed in the copy_from_auth_priv cfap_shared_secret field and the copy_to_auth_priv ctap_shared_secret field. Because of this shared_secret, the RPCSEC_GSS3_CREATE control messages for copy_from_auth and copy_to_auth MUST use a Quality of Protection (QoP) of rpc_gss_svc_privacy.

- An instance of copy_from_auth_priv is filled in with the shared secret, the destination server, and the NFSv4 user id of the user principal and is placed in rpc_gss3_create_args assertions[0].privs.privilege. The string "copy_from_auth" is placed in assertions[0].privs.name. The source server unwraps the rpc_gss_svc_privacy RPCSEC_GSS3_CREATE payload and verifies that the NFSv4 user id being asserted matches the source server's mapping of the user principal. If it does, the privilege is established on the source server as <copy_from_auth, user id, destination>. The field "handle" in a successful reply is the RPCSEC_GSSv3 copy_from_auth "child" handle that the client will use in COPY_NOTIFY requests to the source server.

- An instance of copy_to_auth_priv is filled in with the shared secret, the cnr_source_server list returned by COPY_NOTIFY, and the NFSv4 user id of the user principal. The copy_to_auth_priv instance is placed in rpc_gss3_create_args assertions[0].privs.privilege. The string "copy_to_auth" is placed in assertions[0].privs.name. The destination server unwraps the rpc_gss_svc_privacy RPCSEC_GSS3_CREATE payload and verifies that the NFSv4 user id being asserted matches the destination server's mapping of the user principal. If it does, the privilege is established on the destination server as <copy_to_auth, user id, source list>. The field "handle" in a successful reply is the RPCSEC_GSSv3 copy_to_auth child handle that the client will use in COPY requests to the destination server involving the source server.

As noted in Section 2.7.1 of [RFC7861] ("New Control Procedure - RPCSEC_GSS_CREATE"), both the client and the source server should associate the RPCSEC_GSSv3 child handle with the parent RPCSEC_GSSv3 handle used to create the RPCSEC_GSSv3 child handle.

4.9.1.1.2. Starting a Secure Inter-Server Copy

When the client sends a COPY_NOTIFY request to the source server, it uses the privileged copy_from_auth RPCSEC_GSSv3 handle. cna_destination_server in the COPY_NOTIFY MUST be the same as cfap_destination specified in copy_from_auth_priv. Otherwise, the COPY_NOTIFY will fail with NFS4ERR_ACCESS. The source server verifies that the privilege <copy_from_auth, user id, destination> exists and annotates it with the source filehandle, if the user principal has read access to the source file and if administrative policies give the user principal and the NFS client read access to the source file (i.e., if the ACCESS operation would grant read access). Otherwise, the COPY_NOTIFY will fail with NFS4ERR_ACCESS.

When the client sends a COPY request to the destination server, it uses the privileged copy_to_auth RPCSEC_GSSv3 handle. ca_source_server list in the COPY MUST be the same as ctap_source list specified in copy_to_auth_priv. Otherwise, the COPY will fail with NFS4ERR_ACCESS. The destination server verifies that the privilege <copy_to_auth, user id, source list> exists and annotates it with the source and destination filehandles. If the COPY returns a wr_callback_id, then this is an asynchronous copy and the wr_callback_id must also must be annotated to the copy_to_auth privilege. If the client has failed to establish the copy_to_auth privilege, it will reject the request with NFS4ERR_PARTNER_NO_AUTH.

If either the COPY_NOTIFY operation or the COPY operations fail, the associated copy_from_auth and copy_to_auth RPCSEC_GSSv3 handles MUST be destroyed.

4.9.1.1.3. Securing ONC RPC Server-to-Server Copy Protocols

After a destination server has a copy_to_auth privilege established on it and it receives a COPY request, if it knows it will use an ONC RPC protocol to copy data, it will establish a copy_confirm_auth privilege on the source server prior to responding to the COPY operation, as follows:

- Before establishing an RPCSEC_GSSv3 context, a parent context needs to exist between nfs@<destination> as the initiator principal and nfs@<source> as the target principal. If NFS is to be used as the copy protocol, this means that the destination server must mount the source server using RPCSEC_GSSv3.

- An instance of copy_confirm_auth_priv is filled in with information from the established copy_to_auth privilege. The value of the ccap_shared_secret_mic field is a GSS_GetMIC() of the ctap_shared_secret in the copy_to_auth privilege using the parent handle context. The ccap_username field is the mapping of the user principal to an NFSv4 user name ("user"@"domain" form) and MUST be the same as the ctap_username in the copy_to_auth privilege. The copy_confirm_auth_priv instance is placed in rpc_gss3_create_args assertions[0].privs.privilege. The string "copy_confirm_auth" is placed in assertions[0].privs.name.

- The RPCSEC_GSS3_CREATE copy_from_auth message is sent to the source server with a QoP of rpc_gss_svc_privacy. The source server unwraps the rpc_gss_svc_privacy RPCSEC_GSS3_CREATE payload and verifies the cap_shared_secret_mic by calling GSS_VerifyMIC() using the parent context on the cfap_shared_secret from the established copy_from_auth privilege, and verifies that the ccap_username equals the cfap_username.

- If all verifications succeed, the copy_confirm_auth privilege is established on the source server as <copy_confirm_auth, shared_secret_mic, user id>. Because the shared secret has been verified, the resultant copy_confirm_auth RPCSEC_GSSv3 child handle is noted to be acting on behalf of the user principal.

- If the source server fails to verify the copy_from_auth privilege, the COPY_NOTIFY operation will be rejected with NFS4ERR_PARTNER_NO_AUTH.

- If the destination server fails to verify the copy_to_auth or copy_confirm_auth privilege, the COPY will be rejected with NFS4ERR_PARTNER_NO_AUTH, causing the client to destroy the associated copy_from_auth and copy_to_auth RPCSEC_GSSv3 structured privilege assertion handles.

- All subsequent ONC RPC READ requests sent from the destination to copy data from the source to the destination will use the RPCSEC_GSSv3 copy_confirm_auth child handle.

Note that the use of the copy_confirm_auth privilege accomplishes the following:

- If a protocol like NFS is being used with export policies, the export policies can be overridden if the destination server is not authorized to act as an NFS client.

- Manual configuration to allow a copy relationship between the source and destination is not needed.

4.9.1.1.4. Maintaining a Secure Inter-Server Copy

If the client determines that either the copy_from_auth or the copy_to_auth handle becomes invalid during a copy, then the copy MUST be aborted by the client sending an OFFLOAD_CANCEL to both the source and destination servers and destroying the respective copy-related context handles as described in Section 4.9.1.1.5.

4.9.1.1.5. Finishing or Stopping a Secure Inter-Server Copy

Under normal operation, the client MUST destroy the copy_from_auth and the copy_to_auth RPCSEC_GSSv3 handle once the COPY operation returns for a synchronous inter-server copy or a CB_OFFLOAD reports the result of an asynchronous copy.

The copy_confirm_auth privilege is constructed from information held by the copy_to_auth privilege and MUST be destroyed by the destination server (via an RPCSEC_GSS3_DESTROY call) when the copy_to_auth RPCSEC_GSSv3 handle is destroyed.

The copy_confirm_auth RPCSEC_GSS3 handle is associated with a copy_from_auth RPCSEC_GSS3 handle on the source server via the shared secret and MUST be locally destroyed (there is no RPCSEC_GSS3_DESTROY, as the source server is not the initiator) when the copy_from_auth RPCSEC_GSSv3 handle is destroyed.

If the client sends an OFFLOAD_CANCEL to the source server to rescind the destination server's synchronous copy privilege, it uses the privileged copy_from_auth RPCSEC_GSSv3 handle, and the cra_destination_server in the OFFLOAD_CANCEL MUST be the same as the name of the destination server specified in copy_from_auth_priv. The source server will then delete the <copy_from_auth, user id, destination> privilege and fail any subsequent copy requests sent under the auspices of this privilege from the destination server. The client MUST destroy both the copy_from_auth and the copy_to_auth RPCSEC_GSSv3 handles.

If the client sends an OFFLOAD_STATUS to the destination server to check on the status of an asynchronous copy, it uses the privileged copy_to_auth RPCSEC_GSSv3 handle, and the osa_stateid in the OFFLOAD_STATUS MUST be the same as the wr_callback_id specified in the copy_to_auth privilege stored on the destination server.

If the client sends an OFFLOAD_CANCEL to the destination server to cancel an asynchronous copy, it uses the privileged copy_to_auth RPCSEC_GSSv3 handle, and the oaa_stateid in the OFFLOAD_CANCEL MUST be the same as the wr_callback_id specified in the copy_to_auth privilege stored on the destination server. The destination server will then delete the <copy_to_auth, user id, source list> privilege and the associated copy_confirm_auth RPCSEC_GSSv3 handle. The client MUST destroy both the copy_to_auth and copy_from_auth RPCSEC_GSSv3 handles.

4.9.1.2. Inter-Server Copy via ONC RPC without RPCSEC_GSS

ONC RPC security flavors other than RPCSEC_GSS MAY be used with the server-side copy offload operations described in this section. In particular, host-based ONC RPC security flavors such as AUTH_NONE and AUTH_SYS MAY be used. If a host-based security flavor is used, a minimal level of protection for the server-to-server copy protocol is possible.

The biggest issue is that there is a lack of a strong security method to allow the source server and destination server to identify themselves to each other. A further complication is that in a multihomed environment the destination server might not contact the source server from the same network address specified by the client in the COPY_NOTIFY. The cnr_stateid returned from the COPY_NOTIFY can be used to uniquely identify the destination server to the source server. The use of the cnr_stateid provides initial authentication of the destination server but cannot defend against man-in-the-middle attacks after authentication or against an eavesdropper that observes the opaque stateid on the wire. Other secure communication techniques (e.g., IPsec) are necessary to block these attacks.

Servers SHOULD reject COPY_NOTIFY requests that do not use RPCSEC_GSS with privacy, thus ensuring that the cnr_stateid in the COPY_NOTIFY reply is encrypted. For the same reason, clients SHOULD send COPY requests to the destination using RPCSEC_GSS with privacy.

5. Support for Application I/O Hints

Applications can issue client I/O hints via posix_fadvise() [posix_fadvise] to the NFS client. While this can help the NFS client optimize I/O and caching for a file, it does not allow the NFS server and its exported file system to do likewise. The IO_ADVISE procedure (Section 15.5) is used to communicate the client file access patterns to the NFS server. The NFS server, upon receiving an IO_ADVISE operation, MAY choose to alter its I/O and caching behavior but is under no obligation to do so.

Application-specific NFS clients such as those used by hypervisors and databases can also leverage application hints to communicate their specialized requirements.

6. Sparse Files

A sparse file is a common way of representing a large file without having to utilize all of the disk space for it. Consequently, a sparse file uses less physical space than its size indicates. This means the file contains "holes", byte ranges within the file that contain no data. Most modern file systems support sparse files, including most UNIX file systems and Microsoft's New Technology File System (NTFS); however, it should be noted that Apple's Hierarchical File System Plus (HFS+) does not. Common examples of sparse files include Virtual Machine (VM) OS/disk images, database files, log files, and even checkpoint recovery files most commonly used by the High-Performance Computing (HPC) community.

In addition, many modern file systems support the concept of "unwritten" or "uninitialized" blocks, which have uninitialized space allocated to them on disk but will return zeros until data is written to them. Such functionality is already present in the data model of the pNFS block/volume layout (see [RFC5663]). Uninitialized blocks can be thought of as holes inside a space reservation window.

If an application reads a hole in a sparse file, the file system must return all zeros to the application. For local data access there is little penalty, but with NFS these zeros must be transferred back to the client. If an application uses the NFS client to read data into memory, this wastes time and bandwidth as the application waits for the zeros to be transferred.

A sparse file is typically created by initializing the file to be all zeros. Nothing is written to the data in the file; instead, the hole is recorded in the metadata for the file. So, an 8G disk image might be represented initially by a few hundred bits in the metadata (on UNIX file systems, the inode) and nothing on the disk. If the VM then writes 100M to a file in the middle of the image, there would now be two holes represented in the metadata and 100M in the data.

No new operation is needed to allow the creation of a sparsely populated file; when a file is created and a write occurs past the current size of the file, the non-allocated region will either be a hole or be filled with zeros. The choice of behavior is dictated by the underlying file system and is transparent to the application. However, the abilities to read sparse files and to punch holes to reinitialize the contents of a file are needed.

Two new operations -- DEALLOCATE (Section 15.4) and READ_PLUS (Section 15.10) -- are introduced. DEALLOCATE allows for the hole punching, where an application might want to reset the allocation and reservation status of a range of the file. READ_PLUS supports all the features of READ but includes an extension to support sparse files. READ_PLUS is guaranteed to perform no worse than READ and can dramatically improve performance with sparse files. READ_PLUS does not depend on pNFS protocol features but can be used by pNFS to support sparse files.

6.1. Terminology

Regular file: An object of file type NF4REG or NF4NAMEDATTR.

Sparse file: A regular file that contains one or more holes.

Hole: A byte range within a sparse file that contains all zeros. A hole might or might not have space allocated or reserved to it.

6.2. New Operations

6.2.1. READ_PLUS

READ_PLUS is a new variant of the NFSv4.1 READ operation [RFC5661]. Besides being able to support all of the data semantics of the READ operation, it can also be used by the client and server to efficiently transfer holes. Because the client does not know in advance whether a hole is present or not, if the client supports READ_PLUS and so does the server, then it should always use the READ_PLUS operation in preference to the READ operation.

READ_PLUS extends the response with a new arm representing holes to avoid returning data for portions of the file that are initialized to zero and may or may not contain a backing store. Returning actual data blocks corresponding to holes wastes computational and network resources, thus reducing performance.

When a client sends a READ operation, it is not prepared to accept a READ_PLUS-style response providing a compact encoding of the scope of holes. If a READ occurs on a sparse file, then the server must expand such data to be raw bytes. If a READ occurs in the middle of a hole, the server can only send back bytes starting from that offset. By contrast, if a READ_PLUS occurs in the middle of a hole, the server can send back a range that starts before the offset and extends past the requested length.

6.2.2. DEALLOCATE

The client can use the DEALLOCATE operation on a range of a file as a hole punch, which allows the client to avoid the transfer of a repetitive pattern of zeros across the network. This hole punch is a result of the unreserved space returning all zeros until overwritten.

7. Space Reservation

Applications want to be able to reserve space for a file, report the amount of actual disk space a file occupies, and free up the backing space of a file when it is not required.

One example is the posix_fallocate() operation [posix_fallocate], which allows applications to ask for space reservations from the operating system, usually to provide a better file layout and reduce overhead for random or slow-growing file-appending workloads.

Another example is space reservation for virtual disks in a hypervisor. In virtualized environments, virtual disk files are often stored on NFS-mounted volumes. When a hypervisor creates a virtual disk file, it often tries to preallocate the space for the file so that there are no future allocation-related errors during the operation of the VM. Such errors prevent a VM from continuing execution and result in downtime.

Currently, in order to achieve such a guarantee, applications zero the entire file. The initial zeroing allocates the backing blocks, and all subsequent writes are overwrites of already-allocated blocks. This approach is not only inefficient in terms of the amount of I/O done; it is also not guaranteed to work on file systems that are log-structured or deduplicated. An efficient way of guaranteeing space reservation would be beneficial to such applications.

The new ALLOCATE operation (see Section 15.1) allows a client to request a guarantee that space will be available. The ALLOCATE operation guarantees that any future writes to the region it was successfully called for will not fail with NFS4ERR_NOSPC.

Another useful feature is the ability to report the number of blocks that would be freed when a file is deleted. Currently, NFS reports two size attributes:

size The logical file size of the file.

space_used The size in bytes that the file occupies on disk.

While these attributes are sufficient for space accounting in traditional file systems, they prove to be inadequate in modern file systems that support block-sharing. In such file systems, multiple inodes (the metadata portion of the file system object) can point to a single block with a block reference count to guard against premature freeing. Having a way to tell the number of blocks that would be freed if the file was deleted would be useful to applications that wish to migrate files when a volume is low on space.

Since virtual disks represent a hard drive in a VM, a virtual disk can be viewed as a file system within a file. Since not all blocks within a file system are in use, there is an opportunity to reclaim blocks that are no longer in use. A call to deallocate blocks could result in better space efficiency; less space might be consumed for backups after block deallocation.

The following attribute and operation can be used to resolve these issues:

space_freed This attribute reports the space that would be freed when a file is deleted, taking block-sharing into consideration.

DEALLOCATE This operation deallocates the blocks backing a region of the file.

If space_used of a file is interpreted to mean the size in bytes of all disk blocks pointed to by the inode of the file, then shared blocks get double-counted, over-reporting the space utilization. This also has the adverse effect that the deletion of a file with shared blocks frees up less than space_used bytes.

On the other hand, if space_used is interpreted to mean the size in bytes of those disk blocks unique to the inode of the file, then shared blocks are not counted in any file, resulting in under-reporting of the space utilization.

For example, two files, A and B, have 10 blocks each. Let six of these blocks be shared between them. Thus, the combined space utilized by the two files is 14 * BLOCK_SIZE bytes. In the former case, the combined space utilization of the two files would be reported as 20 * BLOCK_SIZE. However, deleting either would only result in 4 * BLOCK_SIZE being freed. Conversely, the latter interpretation would report that the space utilization is only 8 * BLOCK_SIZE.

Using the space_freed attribute (see Section 12.2.2) is helpful in solving this problem. space_freed is the number of blocks that are allocated to the given file that would be freed on its deletion. In the example, both A and B would report space_freed as 4 * BLOCK_SIZE and space_used as 10 * BLOCK_SIZE. If A is deleted, B will report space_freed as 10 * BLOCK_SIZE, as the deletion of B would result in the deallocation of all 10 blocks.

Using the space_freed attribute does not solve the problem of space being over-reported. However, over-reporting is better than under-reporting.

8. Application Data Block Support

At the OS level, files are contained on disk blocks. Applications are also free to impose structure on the data contained in a file and thus can define an Application Data Block (ADB) to be such a structure. From the application's viewpoint, it only wants to handle ADBs and not raw bytes (see [Strohm11]). An ADB is typically

comprised of two sections: header and data. The header describes the characteristics of the block and can provide a means to detect corruption in the data payload. The data section is typically initialized to all zeros.

The format of the header is application specific, but there are two main components typically encountered:

1. An Application Data Block Number (ADBN), which allows the application to determine which data block is being referenced. This is useful when the client is not storing the blocks in contiguous memory, i.e., a logical block number.

2. Fields to describe the state of the ADB and a means to detect block corruption. For both pieces of data, a useful property would be that the allowed values are specially selected so that, if passed across the network, corruption due to translation between big-endian and little-endian architectures is detectable. For example, 0xf0dedef0 has the same (32 wide) bit pattern in both architectures, making it inappropriate.

Applications already impose structures on files [Strohm11] and detect corruption in data blocks [Ashdown08]. What they are not able to do is efficiently transfer and store ADBs. To initialize a file with ADBs, the client must send each full ADB to the server, and that must be stored on the server.

This section defines a framework for transferring the ADB from client to server and presents one approach to detecting corruption in a given ADB implementation.

8.1. Generic Framework

The representation of the ADB needs to be flexible enough to support many different applications. The most basic approach is no imposition of a block at all, which entails working with the raw bytes. Such an approach would be useful for storing holes, punching holes, etc. In more complex deployments, a server might be supporting multiple applications, each with their own definition of the ADB. One might store the ADBN at the start of the block and then have a guard pattern to detect corruption [McDougall07]. The next might store the ADBN at an offset of 100 bytes within the block and have no guard pattern at all, i.e., existing applications might already have well-defined formats for their data blocks.

The guard pattern can be used to represent the state of the block, to protect against corruption, or both. Again, it needs to be able to be placed anywhere within the ADB.

Both the starting offset of the block and the size of the block need to be represented. Note that nothing prevents the application from defining different-sized blocks in a file.

8.1.1. Data Block Representation

<CODE BEGINS>

```
struct app_data_block4 {
        offset4         adb_offset;
        length4         adb_block_size;
        length4         adb_block_count;
        length4         adb_reloff_blocknum;
        count4          adb_block_num;
        length4         adb_reloff_pattern;
        opaque          adb_pattern<>;
};
```

<CODE ENDS>

The app_data_block4 structure captures the abstraction presented for the ADB. The additional fields present are to allow the transmission of adb_block_count ADBs at one time. The adb_block_num is used to convey the ADBN of the first block in the sequence. Each ADB will contain the same adb_pattern string.

As both adb_block_num and adb_pattern are optional, if either adb_reloff_pattern or adb_reloff_blocknum is set to NFS4_UINT64_MAX, then the corresponding field is not set in any of the ADBs.

8.2. An Example of Detecting Corruption

In this section, an example ADB format is defined in which corruption can be detected. Note that this is just one possible format and means to detect corruption.

Consider a very basic implementation of an operating system's disk blocks. A block is either data or an indirect block that allows for files that are larger than one block. It is desired to be able to initialize a block. Lastly, to quickly unlink a file, a block can be marked invalid. The contents remain intact; this would enable the OS application in question to undelete a file.

The application defines 4K-sized data blocks, with an 8-byte block counter occurring at offset 0 in the block, and with the guard pattern occurring at offset 8 inside the block. Furthermore, the guard pattern can take one of four states:

0xfeedface - This is the FREE state and indicates that the ADB format has been applied.

0xcafedead - This is the DATA state and indicates that real data has been written to this block.

0xe4e5c001 - This is the INDIRECT state and indicates that the block contains block counter numbers that are chained off of this block.

0xba1ed4a3 - This is the INVALID state and indicates that the block contains data whose contents are garbage.

Finally, it also defines an 8-byte checksum starting at byte 16 that applies to the remaining contents of the block (see [Baira08] for an example of using checksums to detect data corruption). If the state is FREE, then that checksum is trivially zero. As such, the application has no need to transfer the checksum implicitly inside the ADB -- it need not make the transfer layer aware of the fact that there is a checksum (see [Ashdown08] for an example of checksums used to detect corruption in application data blocks).

Corruption in each ADB can thus be detected:

- If the guard pattern is anything other than one of the allowed values, including all zeros.
- If the guard pattern is FREE and any other byte in the remainder of the ADB is anything other than zero.
- If the guard pattern is anything other than FREE, then if the stored checksum does not match the computed checksum.
- If the guard pattern is INDIRECT and one of the stored indirect block numbers has a value greater than the number of ADBs in the file.
- If the guard pattern is INDIRECT and one of the stored indirect block numbers is a duplicate of another stored indirect block number.

As can be seen, the application can detect errors based on the combination of the guard pattern state and the checksum but also can detect corruption based on the state and the contents of the ADB.

This last point is important in validating the minimum amount of data incorporated into the generic framework. That is, the guard pattern is sufficient in allowing applications to design their own corruption detection.

Finally, it is important to note that none of these corruption checks occur in the transport layer. The server and client components are totally unaware of the file format and might report everything as being transferred correctly, even in cases where the application detects corruption.

8.3. An Example of READ_PLUS

The hypothetical application presented in Section 8.2 can be used to illustrate how READ_PLUS would return an array of results. A file is created and initialized with 100 4K ADBs in the FREE state with the WRITE_SAME operation (see Section 15.12):

```
   WRITE_SAME {0, 4K, 100, 0, 0, 8, 0xfeedface}
```

Further, assume that the application writes a single ADB at 16K, changing the guard pattern to 0xcafedead; then there would be in memory:

```
    0K ->   (4K - 1) : 00 00 00 00 ... fe ed fa ce 00 00 ... 00
    4K ->   (8K - 1) : 00 00 00 01 ... fe ed fa ce 00 00 ... 00
    8K ->  (12K - 1) : 00 00 00 02 ... fe ed fa ce 00 00 ... 00
   12K ->  (16K - 1) : 00 00 00 03 ... fe ed fa ce 00 00 ... 00
   16K ->  (20K - 1) : 00 00 00 04 ... ca fe de ad 00 00 ... 00
   20K ->  (24K - 1) : 00 00 00 05 ... fe ed fa ce 00 00 ... 00
   24K ->  (28K - 1) : 00 00 00 06 ... fe ed fa ce 00 00 ... 00
      ...
  396K -> (400K - 1) : 00 00 00 63 ... fe ed fa ce 00 00 ... 00
```

And when the client did a READ_PLUS of 64K at the start of the file, it could get back a result of data:

```
    0K ->   (4K - 1) : 00 00 00 00 ... fe ed fa ce 00 00 ... 00
    4K ->   (8K - 1) : 00 00 00 01 ... fe ed fa ce 00 00 ... 00
    8K ->  (12K - 1) : 00 00 00 02 ... fe ed fa ce 00 00 ... 00
   12K ->  (16K - 1) : 00 00 00 03 ... fe ed fa ce 00 00 ... 00
   16K ->  (20K - 1) : 00 00 00 04 ... ca fe de ad 00 00 ... 00
   20K ->  (24K - 1) : 00 00 00 05 ... fe ed fa ce 00 00 ... 00
   24K ->  (28K - 1) : 00 00 00 06 ... fe ed fa ce 00 00 ... 00
      ...
   62K ->  (64K - 1) : 00 00 00 15 ... fe ed fa ce 00 00 ... 00
```

8.4. An Example of Zeroing Space

A simpler use case for WRITE_SAME is applications that want to efficiently zero out a file, but do not want to modify space reservations. This can easily be achieved by a call to WRITE_SAME without an ADB block numbers and pattern, e.g.:

```
WRITE_SAME {0, 1K, 10000, 0, 0, 0, 0}
```

9. Labeled NFS

Access control models such as UNIX permissions or Access Control Lists (ACLs) are commonly referred to as Discretionary Access Control (DAC) models. These systems base their access decisions on user identity and resource ownership. In contrast, Mandatory Access Control (MAC) models base their access control decisions on the label on the subject (usually a process) and the object it wishes to access [RFC4949]. These labels may contain user identity information but usually contain additional information. In DAC systems, users are free to specify the access rules for resources that they own. MAC models base their security decisions on a system-wide policy -- established by an administrator or organization -- that the users do not have the ability to override. In this section, a MAC model is added to NFSv4.2.

First, a method is provided for transporting and storing security label data on NFSv4 file objects. Security labels have several semantics that are met by NFSv4 recommended attributes such as the ability to set the label value upon object creation. Access control on these attributes is done through a combination of two mechanisms. As with other recommended attributes on file objects, the usual DAC checks, based on the ACLs and permission bits, will be performed to ensure that proper file ownership is enforced. In addition, a MAC system MAY be employed on the client, server, or both to enforce additional policy on what subjects may modify security label information.

Second, a method is described for the client to determine if an NFSv4 file object security label has changed. A client that needs to know if a label on a file or set of files is going to change SHOULD request a delegation on each labeled file. In order to change such a security label, the server will have to recall delegations on any file affected by the label change, so informing clients of the label change.

An additional useful feature would be modification to the RPC layer used by NFSv4 to allow RPCs to assert client process subject security labels and enable the enforcement of Full Mode as described in Section 9.5.1. Such modifications are outside the scope of this document (see [RFC7861]).

9.1. Definitions

Label Format Specifier (LFS): an identifier used by the client to establish the syntactic format of the security label and the semantic meaning of its components. LFSs exist in a registry associated with documents describing the format and semantics of the label.

Security Label Format Selection Registry: the IANA registry (see [RFC7569]) containing all registered LFSs, along with references to the documents that describe the syntactic format and semantics of the security label.

Policy Identifier (PI): an optional part of the definition of an LFS. The PI allows clients and servers to identify specific security policies.

Object: a passive resource within the system that is to be protected. Objects can be entities such as files, directories, pipes, sockets, and many other system resources relevant to the protection of the system state.

Subject: an active entity, usually a process that is requesting access to an object.

MAC-Aware: a server that can transmit and store object labels.

MAC-Functional: a client or server that is Labeled NFS enabled. Such a system can interpret labels and apply policies based on the security system.

Multi-Level Security (MLS): a traditional model where objects are given a sensitivity level (Unclassified, Secret, Top Secret, etc.) and a category set (see [LB96], [RFC1108], [RFC2401], and [RFC4949]).

(Note: RFC 2401 has been obsoleted by RFC 4301, but we list RFC 2401 here because RFC 4301 does not discuss MLS.)

9.2. MAC Security Attribute

MAC models base access decisions on security attributes bound to subjects (usually processes) and objects (for NFS, file objects). This information can range from a user identity for an identity-based MAC model, sensitivity levels for MLS, or a type for type enforcement. These models base their decisions on different criteria, but the semantics of the security attribute remain the same. The semantics required by the security attribute are listed below:

- MUST provide flexibility with respect to the MAC model.
- MUST provide the ability to atomically set security information upon object creation.
- MUST provide the ability to enforce access control decisions on both the client and the server.
- MUST NOT expose an object to either the client or server namespace before its security information has been bound to it.

NFSv4 implements the MAC security attribute as a recommended attribute. This attribute has a fixed format and semantics, which conflicts with the flexible nature of security attributes in general. To resolve this, the MAC security attribute consists of two components. The first component is an LFS, as defined in [RFC7569], to allow for interoperability between MAC mechanisms. The second component is an opaque field, which is the actual security attribute data. To allow for various MAC models, NFSv4 should be used solely as a transport mechanism for the security attribute. It is the responsibility of the endpoints to consume the security attribute and make access decisions based on their respective models. In addition, creation of objects through OPEN and CREATE allows the security attribute to be specified upon creation. By providing an atomic create and set operation for the security attribute, it is possible to enforce the second and fourth requirements listed above. The recommended attribute FATTR4_SEC_LABEL (see Section 12.2.4) will be used to satisfy this requirement.

9.2.1. Delegations

In the event that a security attribute is changed on the server while a client holds a delegation on the file, both the server and the client MUST follow the NFSv4.1 protocol (see Section 10 of [RFC5661]) with respect to attribute changes. It SHOULD flush all changes back to the server and relinquish the delegation.

9.2.2. Permission Checking

It is not feasible to enumerate all possible MAC models and even levels of protection within a subset of these models. This means that the NFSv4 client and servers cannot be expected to directly make access control decisions based on the security attribute. Instead, NFSv4 should defer permission checking on this attribute to the host system. These checks are performed in addition to existing DAC and ACL checks outlined in the NFSv4 protocol. Section 9.5 gives a specific example of how the security attribute is handled under a particular MAC model.

9.2.3. Object Creation

When creating files in NFSv4, the OPEN and CREATE operations are used. One of the parameters for these operations is an fattr4 structure containing the attributes the file is to be created with. This allows NFSv4 to atomically set the security attribute of files upon creation. When a client is MAC-Functional, it must always provide the initial security attribute upon file creation. In the event that the server is MAC-Functional as well, it should determine by policy whether it will accept the attribute from the client or instead make the determination itself. If the client is not MAC-Functional, then the MAC-Functional server must decide on a default label. A more in-depth explanation can be found in Section 9.5.

9.2.4. Existing Objects

Note that under the MAC model, all objects must have labels. Therefore, if an existing server is upgraded to include Labeled NFS support, then it is the responsibility of the security system to define the behavior for existing objects.

9.2.5. Label Changes

Consider a Guest Mode system (Section 9.5.3) in which the clients enforce MAC checks and the server has only a DAC security system that stores the labels along with the file data. In this type of system, a user with the appropriate DAC credentials on a client with poorly configured or disabled MAC labeling enforcement is allowed access to the file label (and data) on the server and can change the label.

Clients that need to know if a label on a file or set of files has changed SHOULD request a delegation on each labeled file so that a label change by another client will be known via the process described in Section 9.2.1, which must be followed: the delegation will be recalled, which effectively notifies the client of the change.

Note that the MAC security policies on a client can be such that the client does not have access to the file unless it has a delegation.

9.3. pNFS Considerations

The new FATTR4_SEC_LABEL attribute is metadata information, and as such the storage device is not aware of the value contained on the metadata server. Fortunately, the NFSv4.1 protocol [RFC5661] already has provisions for doing access-level checks from the storage device to the metadata server. In order for the storage device to validate the subject label presented by the client, it SHOULD utilize this mechanism.

9.4. Discovery of Server Labeled NFS Support

The server can easily determine that a client supports Labeled NFS when it queries for the FATTR4_SEC_LABEL label for an object. Further, it can then determine which LFS the client understands. The client might want to discover whether the server supports Labeled NFS and which LFS the server supports.

The following COMPOUND MUST NOT be denied by any MAC label check:

```
PUTROOTFH, GETATTR {FATTR4_SEC_LABEL}
```

Note that the server might have imposed a security flavor on the root that precludes such access. That is, if the server requires Kerberized access and the client presents a COMPOUND with AUTH_SYS, then the server is allowed to return NFS4ERR_WRONGSEC in this case. But if the client presents a correct security flavor, then the server MUST return the FATTR4_SEC_LABEL attribute with the supported LFS filled in.

9.5. MAC Security NFS Modes of Operation

A system using Labeled NFS may operate in three modes (see Section 4 of [RFC7204]). The first mode provides the most protection and is called "Full Mode". In this mode, both the client and server implement a MAC model allowing each end to make an access control decision. The second mode is a subset of the Full Mode and is called "Limited Server Mode". In this mode, the server cannot enforce the

labels, but it can store and transmit them. The remaining mode is called the "Guest Mode"; in this mode, one end of the connection is not implementing a MAC model and thus offers less protection than Full Mode.

9.5.1. Full Mode

Full Mode environments consist of MAC-Functional NFSv4 servers and clients and may be composed of mixed MAC models and policies. The system requires that both the client and server have an opportunity to perform an access control check based on all relevant information within the network. The file object security attribute is provided using the mechanism described in Section 9.2.

Fully MAC-Functional NFSv4 servers are not possible in the absence of RPCSEC_GSSv3 [RFC7861] support for client process subject label assertion. However, servers may make decisions based on the RPC credential information available.

9.5.1.1. Initial Labeling and Translation

The ability to create a file is an action that a MAC model may wish to mediate. The client is given the responsibility to determine the initial security attribute to be placed on a file. This allows the client to make a decision as to the acceptable security attribute to create a file with before sending the request to the server. Once the server receives the creation request from the client, it may choose to evaluate if the security attribute is acceptable.

Security attributes on the client and server may vary based on MAC model and policy. To handle this, the security attribute field has an LFS component. This component is a mechanism for the host to identify the format and meaning of the opaque portion of the security attribute. A Full Mode environment may contain hosts operating in several different LFSs. In this case, a mechanism for translating the opaque portion of the security attribute is needed. The actual translation function will vary based on MAC model and policy and is outside the scope of this document. If a translation is unavailable for a given LFS, then the request MUST be denied. Another recourse is to allow the host to provide a fallback mapping for unknown security attributes.

9.5.1.2. Policy Enforcement

In Full Mode, access control decisions are made by both the clients and servers. When a client makes a request, it takes the security attribute from the requesting process and makes an access control decision based on that attribute and the security attribute of the

object it is trying to access. If the client denies that access, an RPC to the server is never made. If, however, the access is allowed, the client will make a call to the NFS server.

When the server receives the request from the client, it uses any credential information conveyed in the RPC request and the attributes of the object the client is trying to access to make an access control decision. If the server's policy allows this access, it will fulfill the client's request; otherwise, it will return NFS4ERR_ACCESS.

Future protocol extensions may also allow the server to factor into the decision a security label extracted from the RPC request.

Implementations MAY validate security attributes supplied over the network to ensure that they are within a set of attributes permitted from a specific peer and, if not, reject them. Note that a system may permit a different set of attributes to be accepted from each peer.

9.5.2. Limited Server Mode

A Limited Server mode (see Section 4.2 of [RFC7204]) consists of a server that is label aware but does not enforce policies. Such a server will store and retrieve all object labels presented by clients and will utilize the methods described in Section 9.2.5 to allow the clients to detect changing labels, but may not factor the label into access decisions. Instead, it will expect the clients to enforce all such access locally.

9.5.3. Guest Mode

Guest Mode implies that either the client or the server does not handle labels. If the client is not Labeled NFS aware, then it will not offer subject labels to the server. The server is the only entity enforcing policy and may selectively provide standard NFS services to clients based on their authentication credentials and/or associated network attributes (e.g., IP address, network interface). The level of trust and access extended to a client in this mode is configuration specific. If the server is not Labeled NFS aware, then it will not return object labels to the client. Clients in this environment may consist of groups implementing different MAC model policies. The system requires that all clients in the environment be responsible for access control checks.

9.6. Security Considerations for Labeled NFS

Depending on the level of protection the MAC system offers, there may be a requirement to tightly bind the security attribute to the data.

When only one of the client or server enforces labels, it is important to realize that the other side is not enforcing MAC protections. Alternate methods might be in use to handle the lack of MAC support, and care should be taken to identify and mitigate threats from possible tampering outside of these methods.

An example of this is that a server that modifies READDIR or LOOKUP results based on the client's subject label might want to always construct the same subject label for a client that does not present one. This will prevent a non-Labeled NFS client from mixing entries in the directory cache.

10. Sharing Change Attribute Implementation Characteristics with NFSv4 Clients

Although both the NFSv4 [RFC7530] and NFSv4.1 [RFC5661] protocols define the change attribute as being mandatory to implement, there is little in the way of guidance as to its construction. The only mandated constraint is that the value must change whenever the file data or metadata changes.

While this allows for a wide range of implementations, it also leaves the client with no way to determine which is the most recent value for the change attribute in a case where several RPCs have been issued in parallel. In other words, if two COMPOUNDs, both containing WRITE and GETATTR requests for the same file, have been issued in parallel, how does the client determine which of the two change attribute values returned in the replies to the GETATTR requests corresponds to the most recent state of the file? In some cases, the only recourse may be to send another COMPOUND containing a third GETATTR that is fully serialized with the first two.

NFSv4.2 avoids this kind of inefficiency by allowing the server to share details about how the change attribute is expected to evolve, so that the client may immediately determine which, out of the several change attribute values returned by the server, is the most recent. change_attr_type is defined as a new recommended attribute (see Section 12.2.3) and is a per-file system attribute.

11. Error Values

NFS error numbers are assigned to failed operations within a COMPOUND (COMPOUND or CB_COMPOUND) request. A COMPOUND request contains a number of NFS operations that have their results encoded in sequence in a COMPOUND reply. The results of successful operations will consist of an NFS4_OK status followed by the encoded results of the operation. If an NFS operation fails, an error status will be entered in the reply and the COMPOUND request will be terminated.

11.1. Error Definitions

Error	Number	Description
NFS4ERR_BADLABEL	10093	Section 11.1.3.1
NFS4ERR_OFFLOAD_DENIED	10091	Section 11.1.2.1
NFS4ERR_OFFLOAD_NO_REQS	10094	Section 11.1.2.2
NFS4ERR_PARTNER_NO_AUTH	10089	Section 11.1.2.3
NFS4ERR_PARTNER_NOTSUPP	10088	Section 11.1.2.4
NFS4ERR_UNION_NOTSUPP	10090	Section 11.1.1.1
NFS4ERR_WRONG_LFS	10092	Section 11.1.3.2

Table 1: Protocol Error Definitions

11.1.1. General Errors

This section deals with errors that are applicable to a broad set of different purposes.

11.1.1.1. NFS4ERR_UNION_NOTSUPP (Error Code 10090)

One of the arguments to the operation is a discriminated union, and while the server supports the given operation, it does not support the selected arm of the discriminated union.

11.1.2. Server-to-Server Copy Errors

These errors deal with the interaction between server-to-server copies.

11.1.2.1. NFS4ERR_OFFLOAD_DENIED (Error Code 10091)

The COPY offload operation is supported by both the source and the destination, but the destination is not allowing it for this file. If the client sees this error, it should fall back to the normal copy semantics.

11.1.2.2. NFS4ERR_OFFLOAD_NO_REQS (Error Code 10094)

The COPY offload operation is supported by both the source and the destination, but the destination cannot meet the client requirements for either consecutive byte copy or synchronous copy. If the client sees this error, it should either relax the requirements (if any) or fall back to the normal copy semantics.

11.1.2.3. NFS4ERR_PARTNER_NO_AUTH (Error Code 10089)

The source server does not authorize a server-to-server COPY offload operation. This may be due to the client's failure to send the COPY_NOTIFY operation to the source server, the source server receiving a server-to-server copy offload request after the copy lease time expired, or some other permission problem.

The destination server does not authorize a server-to-server COPY offload operation. This may be due to an inter-server COPY request where the destination server requires RPCSEC_GSSv3 and it is not used, or some other permissions problem.

11.1.2.4. NFS4ERR_PARTNER_NOTSUPP (Error Code 10088)

The remote server does not support the server-to-server COPY offload protocol.

11.1.3. Labeled NFS Errors

These errors are used in Labeled NFS.

11.1.3.1. NFS4ERR_BADLABEL (Error Code 10093)

The label specified is invalid in some manner.

11.1.3.2. NFS4ERR_WRONG_LFS (Error Code 10092)

The LFS specified in the subject label is not compatible with the LFS in the object label.

11.2. New Operations and Their Valid Errors

This section contains a table that gives the valid error returns for each new NFSv4.2 protocol operation. The error code NFS4_OK (indicating no error) is not listed but should be understood to be returnable by all new operations. The error values for all other operations are defined in Section 15.2 of [RFC5661].

Operation	Errors
ALLOCATE	NFS4ERR_ACCESS, NFS4ERR_ADMIN_REVOKED, NFS4ERR_BADXDR, NFS4ERR_BAD_STATEID, NFS4ERR_DEADSESSION, NFS4ERR_DELAY, NFS4ERR_DELEG_REVOKED, NFS4ERR_DQUOT, NFS4ERR_EXPIRED, NFS4ERR_FBIG, NFS4ERR_FHEXPIRED, NFS4ERR_GRACE, NFS4ERR_INVAL, NFS4ERR_IO, NFS4ERR_ISDIR, NFS4ERR_MOVED, NFS4ERR_NOFILEHANDLE, NFS4ERR_NOSPC, NFS4ERR_NOTSUPP, NFS4ERR_OLD_STATEID, NFS4ERR_OPENMODE, NFS4ERR_OP_NOT_IN_SESSION, NFS4ERR_REP_TOO_BIG, NFS4ERR_REP_TOO_BIG_TO_CACHE, NFS4ERR_REQ_TOO_BIG, NFS4ERR_RETRY_UNCACHED_REP, NFS4ERR_ROFS, NFS4ERR_SERVERFAULT, NFS4ERR_STALE, NFS4ERR_SYMLINK, NFS4ERR_TOO_MANY_OPS, NFS4ERR_WRONG_TYPE
CLONE	NFS4ERR_ACCESS, NFS4ERR_ADMIN_REVOKED, NFS4ERR_BADXDR, NFS4ERR_BAD_STATEID, NFS4ERR_DEADSESSION, NFS4ERR_DELAY, NFS4ERR_DELEG_REVOKED, NFS4ERR_DQUOT, NFS4ERR_EXPIRED, NFS4ERR_FBIG, NFS4ERR_FHEXPIRED, NFS4ERR_GRACE, NFS4ERR_INVAL, NFS4ERR_IO, NFS4ERR_ISDIR, NFS4ERR_MOVED, NFS4ERR_NOFILEHANDLE, NFS4ERR_NOSPC, NFS4ERR_NOTSUPP, NFS4ERR_OLD_STATEID, NFS4ERR_OPENMODE, NFS4ERR_OP_NOT_IN_SESSION, NFS4ERR_REP_TOO_BIG, NFS4ERR_REP_TOO_BIG_TO_CACHE, NFS4ERR_REQ_TOO_BIG, NFS4ERR_RETRY_UNCACHED_REP, NFS4ERR_ROFS, NFS4ERR_SERVERFAULT, NFS4ERR_STALE, NFS4ERR_SYMLINK, NFS4ERR_TOO_MANY_OPS, NFS4ERR_WRONG_TYPE, NFS4ERR_XDEV

COPY	NFS4ERR_ACCESS, NFS4ERR_ADMIN_REVOKED, NFS4ERR_BADXDR, NFS4ERR_BAD_STATEID, NFS4ERR_DEADSESSION, NFS4ERR_DELAY, NFS4ERR_DELEG_REVOKED, NFS4ERR_DQUOT, NFS4ERR_EXPIRED, NFS4ERR_FBIG, NFS4ERR_FHEXPIRED, NFS4ERR_GRACE, NFS4ERR_INVAL, NFS4ERR_IO, NFS4ERR_ISDIR, NFS4ERR_LOCKED, NFS4ERR_MOVED, NFS4ERR_NOFILEHANDLE, NFS4ERR_NOSPC, NFS4ERR_OFFLOAD_DENIED, NFS4ERR_OLD_STATEID, NFS4ERR_OPENMODE, NFS4ERR_OP_NOT_IN_SESSION, NFS4ERR_PARTNER_NO_AUTH, NFS4ERR_PARTNER_NOTSUPP, NFS4ERR_PNFS_IO_HOLE, NFS4ERR_PNFS_NO_LAYOUT, NFS4ERR_REP_TOO_BIG, NFS4ERR_REP_TOO_BIG_TO_CACHE, NFS4ERR_REQ_TOO_BIG, NFS4ERR_RETRY_UNCACHED_REP, NFS4ERR_ROFS, NFS4ERR_SERVERFAULT, NFS4ERR_STALE, NFS4ERR_SYMLINK, NFS4ERR_TOO_MANY_OPS, NFS4ERR_WRONG_TYPE
COPY_NOTIFY	NFS4ERR_ACCESS, NFS4ERR_ADMIN_REVOKED, NFS4ERR_BADXDR, NFS4ERR_BAD_STATEID, NFS4ERR_DEADSESSION, NFS4ERR_DELAY, NFS4ERR_DELEG_REVOKED, NFS4ERR_EXPIRED, NFS4ERR_FHEXPIRED, NFS4ERR_GRACE, NFS4ERR_INVAL, NFS4ERR_IO, NFS4ERR_ISDIR, NFS4ERR_LOCKED, NFS4ERR_MOVED, NFS4ERR_NOFILEHANDLE, NFS4ERR_OLD_STATEID, NFS4ERR_OPENMODE, NFS4ERR_OP_NOT_IN_SESSION, NFS4ERR_PNFS_IO_HOLE, NFS4ERR_PNFS_NO_LAYOUT, NFS4ERR_REP_TOO_BIG, NFS4ERR_REP_TOO_BIG_TO_CACHE, NFS4ERR_REQ_TOO_BIG, NFS4ERR_RETRY_UNCACHED_REP, NFS4ERR_SERVERFAULT, NFS4ERR_STALE, NFS4ERR_SYMLINK, NFS4ERR_TOO_MANY_OPS, NFS4ERR_WRONG_TYPE
DEALLOCATE	NFS4ERR_ACCESS, NFS4ERR_ADMIN_REVOKED, NFS4ERR_BADXDR, NFS4ERR_BAD_STATEID, NFS4ERR_DEADSESSION, NFS4ERR_DELAY, NFS4ERR_DELEG_REVOKED, NFS4ERR_EXPIRED, NFS4ERR_FBIG, NFS4ERR_FHEXPIRED, NFS4ERR_GRACE, NFS4ERR_INVAL, NFS4ERR_IO, NFS4ERR_ISDIR, NFS4ERR_MOVED, NFS4ERR_NOFILEHANDLE, NFS4ERR_NOTSUPP, NFS4ERR_OLD_STATEID, NFS4ERR_OPENMODE, NFS4ERR_OP_NOT_IN_SESSION, NFS4ERR_REP_TOO_BIG, NFS4ERR_REP_TOO_BIG_TO_CACHE,

	NFS4ERR_REQ_TOO_BIG, NFS4ERR_RETRY_UNCACHED_REP, NFS4ERR_ROFS, NFS4ERR_SERVERFAULT, NFS4ERR_STALE, NFS4ERR_SYMLINK, NFS4ERR_TOO_MANY_OPS, NFS4ERR_WRONG_TYPE
GETDEVICELIST	NFS4ERR_NOTSUPP
IO_ADVISE	NFS4ERR_ACCESS, NFS4ERR_ADMIN_REVOKED, NFS4ERR_BADXDR, NFS4ERR_BAD_STATEID, NFS4ERR_DEADSESSION, NFS4ERR_DELAY, NFS4ERR_DELEG_REVOKED, NFS4ERR_EXPIRED, NFS4ERR_FBIG, NFS4ERR_FHEXPIRED, NFS4ERR_GRACE, NFS4ERR_INVAL, NFS4ERR_IO, NFS4ERR_ISDIR, NFS4ERR_MOVED, NFS4ERR_NOFILEHANDLE, NFS4ERR_NOTSUPP, NFS4ERR_OLD_STATEID, NFS4ERR_OP_NOT_IN_SESSION, NFS4ERR_RETRY_UNCACHED_REP, NFS4ERR_SERVERFAULT, NFS4ERR_STALE, NFS4ERR_SYMLINK, NFS4ERR_TOO_MANY_OPS, NFS4ERR_WRONG_TYPE
LAYOUTERROR	NFS4ERR_ADMIN_REVOKED, NFS4ERR_BADXDR, NFS4ERR_BAD_STATEID, NFS4ERR_DEADSESSION, NFS4ERR_DELAY, NFS4ERR_DELEG_REVOKED, NFS4ERR_EXPIRED, NFS4ERR_FHEXPIRED, NFS4ERR_GRACE, NFS4ERR_INVAL, NFS4ERR_ISDIR, NFS4ERR_MOVED, NFS4ERR_NOFILEHANDLE, NFS4ERR_NOTSUPP, NFS4ERR_NO_GRACE, NFS4ERR_OLD_STATEID, NFS4ERR_OP_NOT_IN_SESSION, NFS4ERR_REP_TOO_BIG, NFS4ERR_REP_TOO_BIG_TO_CACHE, NFS4ERR_REQ_TOO_BIG, NFS4ERR_RETRY_UNCACHED_REP, NFS4ERR_SERVERFAULT, NFS4ERR_STALE, NFS4ERR_TOO_MANY_OPS, NFS4ERR_UNKNOWN_LAYOUTTYPE, NFS4ERR_WRONG_CRED, NFS4ERR_WRONG_TYPE
LAYOUTSTATS	NFS4ERR_ADMIN_REVOKED, NFS4ERR_BADXDR, NFS4ERR_BAD_STATEID, NFS4ERR_DEADSESSION, NFS4ERR_DELAY, NFS4ERR_DELEG_REVOKED, NFS4ERR_EXPIRED, NFS4ERR_FHEXPIRED, NFS4ERR_GRACE, NFS4ERR_INVAL, NFS4ERR_ISDIR, NFS4ERR_MOVED, NFS4ERR_NOFILEHANDLE, NFS4ERR_NOTSUPP, NFS4ERR_NO_GRACE, NFS4ERR_OLD_STATEID, NFS4ERR_OP_NOT_IN_SESSION, NFS4ERR_REP_TOO_BIG, NFS4ERR_REP_TOO_BIG_TO_CACHE, NFS4ERR_REQ_TOO_BIG, NFS4ERR_RETRY_UNCACHED_REP, NFS4ERR_SERVERFAULT, NFS4ERR_STALE,

	NFS4ERR_TOO_MANY_OPS, NFS4ERR_UNKNOWN_LAYOUTTYPE, NFS4ERR_WRONG_CRED, NFS4ERR_WRONG_TYPE
OFFLOAD_CANCEL	NFS4ERR_ADMIN_REVOKED, NFS4ERR_BADXDR, NFS4ERR_BAD_STATEID, NFS4ERR_COMPLETE_ALREADY, NFS4ERR_DEADSESSION, NFS4ERR_DELAY, NFS4ERR_EXPIRED, NFS4ERR_GRACE, NFS4ERR_NOTSUPP, NFS4ERR_OLD_STATEID, NFS4ERR_OP_NOT_IN_SESSION, NFS4ERR_SERVERFAULT, NFS4ERR_TOO_MANY_OPS
OFFLOAD_STATUS	NFS4ERR_ADMIN_REVOKED, NFS4ERR_BADXDR, NFS4ERR_BAD_STATEID, NFS4ERR_COMPLETE_ALREADY, NFS4ERR_DEADSESSION, NFS4ERR_DELAY, NFS4ERR_EXPIRED, NFS4ERR_GRACE, NFS4ERR_NOTSUPP, NFS4ERR_OLD_STATEID, NFS4ERR_OP_NOT_IN_SESSION, NFS4ERR_SERVERFAULT, NFS4ERR_TOO_MANY_OPS
READ_PLUS	NFS4ERR_ACCESS, NFS4ERR_ADMIN_REVOKED, NFS4ERR_BADXDR, NFS4ERR_BAD_STATEID, NFS4ERR_DEADSESSION, NFS4ERR_DELAY, NFS4ERR_DELEG_REVOKED, NFS4ERR_EXPIRED, NFS4ERR_FHEXPIRED, NFS4ERR_GRACE, NFS4ERR_INVAL, NFS4ERR_IO, NFS4ERR_ISDIR, NFS4ERR_LOCKED, NFS4ERR_MOVED, NFS4ERR_NOFILEHANDLE, NFS4ERR_NOTSUPP, NFS4ERR_OLD_STATEID, NFS4ERR_OPENMODE, NFS4ERR_OP_NOT_IN_SESSION, NFS4ERR_PARTNER_NO_AUTH, NFS4ERR_PNFS_IO_HOLE, NFS4ERR_PNFS_NO_LAYOUT, NFS4ERR_REP_TOO_BIG, NFS4ERR_REP_TOO_BIG_TO_CACHE, NFS4ERR_REQ_TOO_BIG, NFS4ERR_RETRY_UNCACHED_REP, NFS4ERR_SERVERFAULT, NFS4ERR_STALE, NFS4ERR_SYMLINK, NFS4ERR_TOO_MANY_OPS, NFS4ERR_WRONG_TYPE
SEEK	NFS4ERR_ACCESS, NFS4ERR_ADMIN_REVOKED, NFS4ERR_BADXDR, NFS4ERR_BAD_STATEID, NFS4ERR_DEADSESSION, NFS4ERR_DELAY, NFS4ERR_DELEG_REVOKED, NFS4ERR_EXPIRED, NFS4ERR_FHEXPIRED, NFS4ERR_GRACE, NFS4ERR_INVAL, NFS4ERR_IO, NFS4ERR_ISDIR, NFS4ERR_LOCKED, NFS4ERR_MOVED, NFS4ERR_NOFILEHANDLE, NFS4ERR_NOTSUPP, NFS4ERR_OLD_STATEID, NFS4ERR_OPENMODE, NFS4ERR_OP_NOT_IN_SESSION, NFS4ERR_PNFS_IO_HOLE, NFS4ERR_PNFS_NO_LAYOUT, NFS4ERR_REP_TOO_BIG, NFS4ERR_REP_TOO_BIG_TO_CACHE, NFS4ERR_REQ_TOO_BIG, NFS4ERR_RETRY_UNCACHED_REP,

	NFS4ERR_SERVERFAULT, NFS4ERR_STALE, NFS4ERR_SYMLINK, NFS4ERR_TOO_MANY_OPS, NFS4ERR_UNION_NOTSUPP, NFS4ERR_WRONG_TYPE
WRITE_SAME	NFS4ERR_ACCESS, NFS4ERR_ADMIN_REVOKED, NFS4ERR_BADXDR, NFS4ERR_BAD_STATEID, NFS4ERR_DEADSESSION, NFS4ERR_DELAY, NFS4ERR_DELEG_REVOKED, NFS4ERR_DQUOT, NFS4ERR_EXPIRED, NFS4ERR_FBIG, NFS4ERR_FHEXPIRED, NFS4ERR_GRACE, NFS4ERR_INVAL, NFS4ERR_IO, NFS4ERR_ISDIR, NFS4ERR_LOCKED, NFS4ERR_MOVED, NFS4ERR_NOFILEHANDLE, NFS4ERR_NOSPC, NFS4ERR_NOTSUPP, NFS4ERR_OLD_STATEID, NFS4ERR_OPENMODE, NFS4ERR_OP_NOT_IN_SESSION, NFS4ERR_PNFS_IO_HOLE, NFS4ERR_PNFS_NO_LAYOUT, NFS4ERR_REP_TOO_BIG, NFS4ERR_REP_TOO_BIG_TO_CACHE, NFS4ERR_REQ_TOO_BIG, NFS4ERR_RETRY_UNCACHED_REP, NFS4ERR_ROFS, NFS4ERR_SERVERFAULT, NFS4ERR_STALE, NFS4ERR_SYMLINK, NFS4ERR_TOO_MANY_OPS, NFS4ERR_WRONG_TYPE

Table 2: Valid Error Returns for Each New Protocol Operation

11.3. New Callback Operations and Their Valid Errors

This section contains a table that gives the valid error returns for each new NFSv4.2 callback operation. The error code NFS4_OK (indicating no error) is not listed but should be understood to be returnable by all new callback operations. The error values for all other callback operations are defined in Section 15.3 of [RFC5661].

Callback Operation	Errors
CB_OFFLOAD	NFS4ERR_BADHANDLE, NFS4ERR_BADXDR, NFS4ERR_BAD_STATEID, NFS4ERR_DELAY, NFS4ERR_OP_NOT_IN_SESSION, NFS4ERR_REP_TOO_BIG, NFS4ERR_REP_TOO_BIG_TO_CACHE, NFS4ERR_REQ_TOO_BIG, NFS4ERR_RETRY_UNCACHED_REP, NFS4ERR_SERVERFAULT, NFS4ERR_TOO_MANY_OPS

Table 3: Valid Error Returns for Each New Protocol Callback Operation

12. New File Attributes

12.1. New RECOMMENDED Attributes - List and Definition References

The list of new RECOMMENDED attributes appears in Table 4. The meanings of the columns of the table are:

Name: The name of the attribute.

Id: The number assigned to the attribute. In the event of conflicts between the assigned number and [RFC7863], the latter is authoritative, but in such an event, it should be resolved with errata to this document and/or [RFC7863]. See [IESG08] for the errata process.

Data Type: The XDR data type of the attribute.

Acc: Access allowed to the attribute.

R means read-only (GETATTR may retrieve, SETATTR may not set).

W means write-only (SETATTR may set, GETATTR may not retrieve).

R W means read/write (GETATTR may retrieve, SETATTR may set).

Defined in: The section of this specification that describes the attribute.

Name	Id	Data Type	Acc	Defined in
clone_blksize	77	uint32_t	R	Section 12.2.1
space_freed	78	length4	R	Section 12.2.2
change_attr_type	79	change_attr_type4	R	Section 12.2.3
sec_label	80	sec_label4	R W	Section 12.2.4

Table 4: New RECOMMENDED Attributes

12.2. Attribute Definitions

12.2.1. Attribute 77: clone_blksize

The clone_blksize attribute indicates the granularity of a CLONE operation.

12.2.2. Attribute 78: space_freed

space_freed gives the number of bytes freed if the file is deleted. This attribute is read-only and is of type length4. It is a per-file attribute.

12.2.3. Attribute 79: change_attr_type

<CODE BEGINS>

```
enum change_attr_type4 {
            NFS4_CHANGE_TYPE_IS_MONOTONIC_INCR         = 0,
            NFS4_CHANGE_TYPE_IS_VERSION_COUNTER        = 1,
            NFS4_CHANGE_TYPE_IS_VERSION_COUNTER_NOPNFS = 2,
            NFS4_CHANGE_TYPE_IS_TIME_METADATA          = 3,
            NFS4_CHANGE_TYPE_IS_UNDEFINED              = 4
};
```

<CODE ENDS>

change_attr_type is a per-file system attribute that enables the NFSv4.2 server to provide additional information about how it expects the change attribute value to evolve after the file data or metadata has changed. While Section 5.4 of [RFC5661] discusses per-file system attributes, it is expected that the value of change_attr_type will not depend on the value of "homogeneous" and will only change in the event of a migration.

NFS4_CHANGE_TYPE_IS_MONOTONIC_INCR: The change attribute value MUST monotonically increase for every atomic change to the file attributes, data, or directory contents.

NFS4_CHANGE_TYPE_IS_VERSION_COUNTER: The change attribute value MUST be incremented by one unit for every atomic change to the file attributes, data, or directory contents. This property is preserved when writing to pNFS data servers.

NFS4_CHANGE_TYPE_IS_VERSION_COUNTER_NOPNFS: The change attribute value MUST be incremented by one unit for every atomic change to the file attributes, data, or directory contents. In the case where the client is writing to pNFS data servers, the number of increments is not guaranteed to exactly match the number of WRITEs.

NFS4_CHANGE_TYPE_IS_TIME_METADATA: The change attribute is implemented as suggested in [RFC7530] in terms of the time_metadata attribute.

NFS4_CHANGE_TYPE_IS_UNDEFINED: The change attribute does not take values that fit into any of these categories.

If either NFS4_CHANGE_TYPE_IS_MONOTONIC_INCR, NFS4_CHANGE_TYPE_IS_VERSION_COUNTER, or NFS4_CHANGE_TYPE_IS_TIME_METADATA is set, then the client knows at the very least that the change attribute is monotonically increasing, which is sufficient to resolve the question of which value is the most recent.

If the client sees the value NFS4_CHANGE_TYPE_IS_TIME_METADATA, then by inspecting the value of the "time_delta" attribute it additionally has the option of detecting rogue server implementations that use time_metadata in violation of the specification.

If the client sees NFS4_CHANGE_TYPE_IS_VERSION_COUNTER, it has the ability to predict what the resulting change attribute value should be after a COMPOUND containing a SETATTR, WRITE, or CREATE. This again allows it to detect changes made in parallel by another client. The value NFS4_CHANGE_TYPE_IS_VERSION_COUNTER_NOPNFS permits the same, but only if the client is not doing pNFS WRITEs.

Finally, if the server does not support change_attr_type or if NFS4_CHANGE_TYPE_IS_UNDEFINED is set, then the server SHOULD make an effort to implement the change attribute in terms of the time_metadata attribute.

12.2.4. Attribute 80: sec_label

```
<CODE BEGINS>

typedef uint32_t  policy4;

struct labelformat_spec4 {
        policy4 lfs_lfs;
        policy4 lfs_pi;
};

struct sec_label4 {
        labelformat_spec4       slai_lfs;
        opaque                  slai_data<>;
};

<CODE ENDS>
```

The FATTR4_SEC_LABEL contains an array of two components, with the first component being an LFS. It serves to provide the receiving end with the information necessary to translate the security attribute into a form that is usable by the endpoint. Label Formats assigned an LFS may optionally choose to include a Policy Identifier field to allow for complex policy deployments. The LFS and the Security Label Format Selection Registry are described in detail in [RFC7569]. The translation used to interpret the security attribute is not specified as part of the protocol, as it may depend on various factors. The second component is an opaque section that contains the data of the attribute. This component is dependent on the MAC model to interpret and enforce.

In particular, it is the responsibility of the LFS specification to define a maximum size for the opaque section, slai_data<>. When creating or modifying a label for an object, the client needs to be guaranteed that the server will accept a label that is sized correctly. By both client and server being part of a specific MAC model, the client will be aware of the size.

13. Operations: REQUIRED, RECOMMENDED, or OPTIONAL

Tables 5 and 6 summarize the operations of the NFSv4.2 protocol and the corresponding designations of REQUIRED, RECOMMENDED, and OPTIONAL to implement or MUST NOT implement. The "MUST NOT implement" designation is reserved for those operations that were defined in either NFSv4.0 or NFSv4.1 and MUST NOT be implemented in NFSv4.2.

For the most part, the REQUIRED, RECOMMENDED, or OPTIONAL designation for operations sent by the client is for the server implementation. The client is generally required to implement the operations needed for the operating environment that it serves. For example, a read-only NFSv4.2 client would have no need to implement the WRITE operation and is not required to do so.

The REQUIRED or OPTIONAL designation for callback operations sent by the server is for both the client and server. Generally, the client has the option of creating the backchannel and sending the operations on the forechannel that will be a catalyst for the server sending callback operations. A partial exception is CB_RECALL_SLOT; the only way the client can avoid supporting this operation is by not creating a backchannel.

Since this is a summary of the operations and their designation, there are subtleties that are not presented here. Therefore, if there is a question regarding implementation requirements, the operation descriptions themselves must be consulted, along with other relevant explanatory text within either this specification or the NFSv4.1 specification [RFC5661].

The abbreviations used in the second and third columns of Tables 5 and 6 are defined as follows:

REQ: REQUIRED to implement

REC: RECOMMENDED to implement

OPT: OPTIONAL to implement

MNI: MUST NOT implement

For the NFSv4.2 features that are OPTIONAL, the operations that support those features are OPTIONAL, and the server MUST return NFS4ERR_NOTSUPP in response to the client's use of those operations when those operations are not implemented by the server. If an OPTIONAL feature is supported, it is possible that a set of operations related to the feature become REQUIRED to implement. The third column of the tables designates the feature(s) and if the operation is REQUIRED or OPTIONAL in the presence of support for the feature.

The OPTIONAL features identified and their abbreviations are as follows:

pNFS: Parallel NFS

FDELG: File Delegations

DDELG: Directory Delegations

COPYra: Intra-server Server-Side Copy

COPYer: Inter-server Server-Side Copy

ADB: Application Data Blocks

Operation	REQ, REC, OPT, or MNI	Feature (REQ, REC, or OPT)
ACCESS	REQ	
ALLOCATE	OPT	
BACKCHANNEL_CTL	REQ	
BIND_CONN_TO_SESSION	REQ	
CLONE	OPT	
CLOSE	REQ	
COMMIT	REQ	
COPY	OPT	COPYer (REQ), COPYra (REQ)
COPY_NOTIFY	OPT	COPYer (REQ)
CREATE	REQ	
CREATE_SESSION	REQ	
DEALLOCATE	OPT	
DELEGPURGE	OPT	FDELG (REQ)
DELEGRETURN	OPT	FDELG, DDELG, pNFS (REQ)
DESTROY_CLIENTID	REQ	
DESTROY_SESSION	REQ	
EXCHANGE_ID	REQ	
FREE_STATEID	REQ	
GETATTR	REQ	
GETDEVICEINFO	OPT	pNFS (REQ)
GETDEVICELIST	MNI	pNFS (MNI)
GETFH	REQ	
GET_DIR_DELEGATION	OPT	DDELG (REQ)
ILLEGAL	REQ	
IO_ADVISE	OPT	
LAYOUTCOMMIT	OPT	pNFS (REQ)
LAYOUTERROR	OPT	pNFS (OPT)
LAYOUTGET	OPT	pNFS (REQ)
LAYOUTRETURN	OPT	pNFS (REQ)
LAYOUTSTATS	OPT	pNFS (OPT)
LINK	OPT	
LOCK	REQ	
LOCKT	REQ	
LOCKU	REQ	
LOOKUP	REQ	
LOOKUPP	REQ	
NVERIFY	REQ	
OFFLOAD_CANCEL	OPT	COPYer (OPT), COPYra (OPT)
OFFLOAD_STATUS	OPT	COPYer (OPT), COPYra (OPT)

OPEN	REQ	
OPENATTR	OPT	
OPEN_CONFIRM	MNI	
OPEN_DOWNGRADE	REQ	
PUTFH	REQ	
PUTPUBFH	REQ	
PUTROOTFH	REQ	
READ	REQ	
READDIR	REQ	
READLINK	OPT	
READ_PLUS	OPT	
RECLAIM_COMPLETE	REQ	
RELEASE_LOCKOWNER	MNI	
REMOVE	REQ	
RENAME	REQ	
RENEW	MNI	
RESTOREFH	REQ	
SAVEFH	REQ	
SECINFO	REQ	
SECINFO_NO_NAME	REC	pNFS file layout (REQ)
SEEK	OPT	
SEQUENCE	REQ	
SETATTR	REQ	
SETCLIENTID	MNI	
SETCLIENTID_CONFIRM	MNI	
SET_SSV	REQ	
TEST_STATEID	REQ	
VERIFY	REQ	
WANT_DELEGATION	OPT	FDELG (OPT)
WRITE	REQ	
WRITE_SAME	OPT	ADB (REQ)

Table 5: Operations

Operation	REQ, REC, OPT, or MNI	Feature (REQ, REC, or OPT)
CB_GETATTR	OPT	FDELG (REQ)
CB_ILLEGAL	REQ	
CB_LAYOUTRECALL	OPT	pNFS (REQ)
CB_NOTIFY	OPT	DDELG (REQ)
CB_NOTIFY_DEVICEID	OPT	pNFS (OPT)
CB_NOTIFY_LOCK	OPT	
CB_OFFLOAD	OPT	COPYer (REQ), COPYra (REQ)
CB_PUSH_DELEG	OPT	FDELG (OPT)
CB_RECALL	OPT	FDELG, DDELG, pNFS (REQ)
CB_RECALL_ANY	OPT	FDELG, DDELG, pNFS (REQ)
CB_RECALL_SLOT	REQ	
CB_RECALLABLE_OBJ_AVAIL	OPT	DDELG, pNFS (REQ)
CB_SEQUENCE	OPT	FDELG, DDELG, pNFS (REQ)
CB_WANTS_CANCELLED	OPT	FDELG, DDELG, pNFS (REQ)

Table 6: Callback Operations

14. Modifications to NFSv4.1 Operations

14.1. Operation 42: EXCHANGE_ID - Instantiate the client ID

14.1.1. ARGUMENT

<CODE BEGINS>

```
/* new */
const EXCHGID4_FLAG_SUPP_FENCE_OPS      = 0x00000004;
```

<CODE ENDS>

14.1.2. RESULT

Unchanged

14.1.3. MOTIVATION

Enterprise applications require guarantees that an operation has either aborted or completed. NFSv4.1 provides this guarantee as long as the session is alive: simply send a SEQUENCE operation on the same slot with a new sequence number, and the successful return of SEQUENCE indicates that the previous operation has completed. However, if the session is lost, there is no way to know when any operations in progress have aborted or completed. In hindsight, the NFSv4.1 specification should have mandated that DESTROY_SESSION either abort or complete all outstanding operations.

14.1.4. DESCRIPTION

A client SHOULD request the EXCHGID4_FLAG_SUPP_FENCE_OPS capability when it sends an EXCHANGE_ID operation. The server SHOULD set this capability in the EXCHANGE_ID reply whether the client requests it or not. It is the server's return that determines whether this capability is in effect. When it is in effect, the following will occur:

- The server will not reply to any DESTROY_SESSION invoked with the client ID until all operations in progress are completed or aborted.

- The server will not reply to subsequent EXCHANGE_ID operations invoked on the same client owner with a new verifier until all operations in progress on the client ID's session are completed or aborted.

- In implementations where the NFS server is deployed as a cluster, it does support client ID trunking, and the EXCHGID4_FLAG_SUPP_FENCE_OPS capability is enabled, then a session ID created on one node of the storage cluster MUST be destroyable via DESTROY_SESSION. In addition, DESTROY_CLIENTID and an EXCHANGE_ID with a new verifier affect all sessions, regardless of what node the sessions were created on.

14.2. Operation 48: GETDEVICELIST - Get all device mappings for a file system

14.2.1. ARGUMENT

<CODE BEGINS>

```
struct GETDEVICELIST4args {
        /* CURRENT_FH: object belonging to the file system */
        layouttype4     gdla_layout_type;

        /* number of device IDs to return */
        count4          gdla_maxdevices;

        nfs_cookie4     gdla_cookie;
        verifier4       gdla_cookieverf;
};
```

<CODE ENDS>

14.2.2. RESULT

<CODE BEGINS>

```
struct GETDEVICELIST4resok {
        nfs_cookie4             gdlr_cookie;
        verifier4               gdlr_cookieverf;
        deviceid4               gdlr_deviceid_list<>;
        bool                    gdlr_eof;
};

union GETDEVICELIST4res switch (nfsstat4 gdlr_status) {
case NFS4_OK:
        GETDEVICELIST4resok     gdlr_resok4;
default:
        void;
};
```

<CODE ENDS>

14.2.3. MOTIVATION

The GETDEVICELIST operation was introduced in [RFC5661] specifically to request a list of devices at file system mount time from block layout type servers. However, the use of the GETDEVICELIST operation introduces a race condition versus notification about changes to pNFS device IDs as provided by CB_NOTIFY_DEVICEID. Implementation experience with block layout servers has shown that there is no need

for GETDEVICELIST. Clients have to be able to request new devices using GETDEVICEINFO at any time in response to either a new deviceid in LAYOUTGET results or the CB_NOTIFY_DEVICEID callback operation.

14.2.4. DESCRIPTION

Clients and servers MUST NOT implement the GETDEVICELIST operation.

15. NFSv4.2 Operations

15.1. Operation 59: ALLOCATE - Reserve space in a region of a file

15.1.1. ARGUMENT

<CODE BEGINS>

```
struct ALLOCATE4args {
        /* CURRENT_FH: file */
        stateid4        aa_stateid;
        offset4         aa_offset;
        length4         aa_length;
};
```

<CODE ENDS>

15.1.2. RESULT

<CODE BEGINS>

```
struct ALLOCATE4res {
        nfsstat4        ar_status;
};
```

<CODE ENDS>

15.1.3. DESCRIPTION

Whenever a client wishes to reserve space for a region in a file, it calls the ALLOCATE operation with the current filehandle set to the filehandle of the file in question, and with the start offset and length in bytes of the region set in aa_offset and aa_length, respectively.

CURRENT_FH must be a regular file. If CURRENT_FH is not a regular file, the operation MUST fail and return NFS4ERR_WRONG_TYPE.

The aa_stateid MUST refer to a stateid that is valid for a WRITE operation and follows the rules for stateids in Sections 8.2.5 and 18.32.3 of [RFC5661].

The server will ensure that backing blocks are reserved to the region specified by aa_offset and aa_length, and that no future writes into this region will return NFS4ERR_NOSPC. If the region lies partially or fully outside the current file size, the file size will be set to aa_offset + aa_length implicitly. If the server cannot guarantee this, it must return NFS4ERR_NOSPC.

The ALLOCATE operation can also be used to extend the size of a file if the region specified by aa_offset and aa_length extends beyond the current file size. In that case, any data outside of the previous file size will return zeros when read before data is written to it.

It is not required that the server allocate the space to the file before returning success. The allocation can be deferred; however, it must be guaranteed that it will not fail for lack of space. The deferral does not result in an asynchronous reply.

The ALLOCATE operation will result in the space_used and space_freed attributes being increased by the number of bytes reserved, unless they were previously reserved or written and not shared.

15.2. Operation 60: COPY - Initiate a server-side copy

15.2.1. ARGUMENT

<CODE BEGINS>

```
struct COPY4args {
        /* SAVED_FH: source file */
        /* CURRENT_FH: destination file */
        stateid4        ca_src_stateid;
        stateid4        ca_dst_stateid;
        offset4         ca_src_offset;
        offset4         ca_dst_offset;
        length4         ca_count;
        bool            ca_consecutive;
        bool            ca_synchronous;
        netloc4         ca_source_server<>;
};
```

<CODE ENDS>

15.2.2. RESULT

```
<CODE BEGINS>

struct write_response4 {
        stateid4        wr_callback_id<1>;
        length4         wr_count;
        stable_how4     wr_committed;
        verifier4       wr_writeverf;
};

struct copy_requirements4 {
        bool            cr_consecutive;
        bool            cr_synchronous;
};

struct COPY4resok {
        write_response4         cr_response;
        copy_requirements4      cr_requirements;
};

union COPY4res switch (nfsstat4 cr_status) {
case NFS4_OK:
        COPY4resok              cr_resok4;
case NFS4ERR_OFFLOAD_NO_REQS:
        copy_requirements4      cr_requirements;
default:
        void;
};

<CODE ENDS>
```

15.2.3. DESCRIPTION

The COPY operation is used for both intra-server and inter-server copies. In both cases, the COPY is always sent from the client to the destination server of the file copy. The COPY operation requests that a range in the file specified by SAVED_FH be copied to a range in the file specified by CURRENT_FH.

Both SAVED_FH and CURRENT_FH must be regular files. If either SAVED_FH or CURRENT_FH is not a regular file, the operation MUST fail and return NFS4ERR_WRONG_TYPE.

SAVED_FH and CURRENT_FH must be different files. If SAVED_FH and CURRENT_FH refer to the same file, the operation MUST fail with NFS4ERR_INVAL.

If the request is for an inter-server copy, the source-fh is a filehandle from the source server and the COMPOUND procedure is being executed on the destination server. In this case, the source-fh is a foreign filehandle on the server receiving the COPY request. If either PUTFH or SAVEFH checked the validity of the filehandle, the operation would likely fail and return NFS4ERR_STALE.

If a server supports the inter-server copy feature, a PUTFH followed by a SAVEFH MUST NOT return NFS4ERR_STALE for either operation. These restrictions do not pose substantial difficulties for servers. CURRENT_FH and SAVED_FH may be validated in the context of the operation referencing them and an NFS4ERR_STALE error returned for an invalid filehandle at that point.

The ca_dst_stateid MUST refer to a stateid that is valid for a WRITE operation and follows the rules for stateids in Sections 8.2.5 and 18.32.3 of [RFC5661]. For an inter-server copy, the ca_src_stateid MUST be the cnr_stateid returned from the earlier COPY_NOTIFY operation, while for an intra-server copy ca_src_stateid MUST refer to a stateid that is valid for a READ operation and follows the rules for stateids in Sections 8.2.5 and 18.22.3 of [RFC5661]. If either stateid is invalid, then the operation MUST fail.

The ca_src_offset is the offset within the source file from which the data will be read, the ca_dst_offset is the offset within the destination file to which the data will be written, and the ca_count is the number of bytes that will be copied. An offset of 0 (zero) specifies the start of the file. A count of 0 (zero) requests that all bytes from ca_src_offset through EOF be copied to the destination. If concurrent modifications to the source file overlap with the source file region being copied, the data copied may include all, some, or none of the modifications. The client can use standard NFS operations (e.g., OPEN with OPEN4_SHARE_DENY_WRITE or mandatory byte-range locks) to protect against concurrent modifications if the client is concerned about this. If the source file's EOF is being modified in parallel with a COPY that specifies a count of 0 (zero) bytes, the amount of data copied is implementation dependent (clients may guard against this case by specifying a non-zero count value or preventing modification of the source file as mentioned above).

If the source offset or the source offset plus count is greater than the size of the source file, the operation MUST fail with NFS4ERR_INVAL. The destination offset or destination offset plus count may be greater than the size of the destination file. This allows the client to issue parallel copies to implement operations such as

<CODE BEGINS>

```
% cat file1 file2 file3 file4 > dest
```

<CODE ENDS>

If the ca_source_server list is specified, then this is an inter-server COPY operation and the source file is on a remote server. The client is expected to have previously issued a successful COPY_NOTIFY request to the remote source server. The ca_source_server list MUST be the same as the COPY_NOTIFY response's cnr_source_server list. If the client includes the entries from the COPY_NOTIFY response's cnr_source_server list in the ca_source_server list, the source server can indicate a specific copy protocol for the destination server to use by returning a URL that specifies both a protocol service and server name. Server-to-server copy protocol considerations are described in Sections 4.6 and 4.9.1.

If ca_consecutive is set, then the client has specified that the copy protocol selected MUST copy bytes in consecutive order from ca_src_offset to ca_count. If the destination server cannot meet this requirement, then it MUST return an error of NFS4ERR_OFFLOAD_NO_REQS and set cr_consecutive to be FALSE. Likewise, if ca_synchronous is set, then the client has required that the copy protocol selected MUST perform a synchronous copy. If the destination server cannot meet this requirement, then it MUST return an error of NFS4ERR_OFFLOAD_NO_REQS and set cr_synchronous to be FALSE.

If both are set by the client, then the destination SHOULD try to determine if it can respond to both requirements at the same time. If it cannot make that determination, it must set to TRUE the one it can and set to FALSE the other. The client, upon getting an NFS4ERR_OFFLOAD_NO_REQS error, has to examine both cr_consecutive and cr_synchronous against the respective values of ca_consecutive and ca_synchronous to determine the possible requirement not met. It MUST be prepared for the destination server not being able to determine both requirements at the same time.

Upon receiving the NFS4ERR_OFFLOAD_NO_REQS error, the client has to determine whether it wants to re-request the copy with a relaxed set of requirements or revert to manually copying the data. If it decides to manually copy the data and this is a remote copy, then the client is responsible for informing the source that the earlier COPY_NOTIFY is no longer valid by sending it an OFFLOAD_CANCEL.

If the operation does not result in an immediate failure, the server will return NFS4_OK.

If the wr_callback_id is returned, this indicates that an asynchronous COPY operation was initiated and a CB_OFFLOAD callback will deliver the final results of the operation. The wr_callback_id stateid is termed a "copy stateid" in this context. The server is given the option of returning the results in a callback because the data may require a relatively long period of time to copy.

If no wr_callback_id is returned, the operation completed synchronously and no callback will be issued by the server. The completion status of the operation is indicated by cr_status.

If the copy completes successfully, either synchronously or asynchronously, the data copied from the source file to the destination file MUST appear identical to the NFS client. However, the NFS server's on-disk representation of the data in the source file and destination file MAY differ. For example, the NFS server might encrypt, compress, deduplicate, or otherwise represent the on-disk data in the source and destination files differently.

If a failure does occur for a synchronous copy, wr_count will be set to the number of bytes copied to the destination file before the error occurred. If cr_consecutive is TRUE, then the bytes were copied in order. If the failure occurred for an asynchronous copy, then the client will have gotten the notification of the consecutive copy order when it got the copy stateid. It will be able to determine the bytes copied from the coa_bytes_copied in the CB_OFFLOAD argument.

In either case, if cr_consecutive was not TRUE, there is no assurance as to exactly which bytes in the range were copied. The client MUST assume that there exists a mixture of the original contents of the range and the new bytes. If the COPY wrote past the end of the file on the destination, then the last byte written to will determine the new file size. The contents of any block not written to and past the original size of the file will be as if a normal WRITE extended the file.

15.3. Operation 61: COPY_NOTIFY - Notify a source server of a future copy

15.3.1. ARGUMENT

<CODE BEGINS>

```
struct COPY_NOTIFY4args {
        /* CURRENT_FH: source file */
        stateid4        cna_src_stateid;
        netloc4         cna_destination_server;
};
```

<CODE ENDS>

15.3.2. RESULT

<CODE BEGINS>

```
struct COPY_NOTIFY4resok {
        nfstime4        cnr_lease_time;
        stateid4        cnr_stateid;
        netloc4         cnr_source_server<>;
};

union COPY_NOTIFY4res switch (nfsstat4 cnr_status) {
case NFS4_OK:
        COPY_NOTIFY4resok       resok4;
default:
        void;
};
```

<CODE ENDS>

15.3.3. DESCRIPTION

This operation is used for an inter-server copy. A client sends this operation in a COMPOUND request to the source server to authorize a destination server identified by cna_destination_server to read the file specified by CURRENT_FH on behalf of the given user.

The cna_src_stateid MUST refer to either open or locking states provided earlier by the server. If it is invalid, then the operation MUST fail.

The cna_destination_server MUST be specified using the netloc4 network location format. The server is not required to resolve the cna_destination_server address before completing this operation.

If this operation succeeds, the source server will allow the cna_destination_server to copy the specified file on behalf of the given user as long as both of the following conditions are met:

- The destination server begins reading the source file before the cnr_lease_time expires. If the cnr_lease_time expires while the destination server is still reading the source file, the destination server is allowed to finish reading the file. If the cnr_lease_time expires before the destination server uses READ or READ_PLUS to begin the transfer, the source server can use NFS4ERR_PARTNER_NO_AUTH to inform the destination server that the cnr_lease_time has expired.

- The client has not issued an OFFLOAD_CANCEL for the same combination of user, filehandle, and destination server.

The cnr_lease_time is chosen by the source server. A cnr_lease_time of 0 (zero) indicates an infinite lease. To avoid the need for synchronized clocks, copy lease times are granted by the server as a time delta. To renew the copy lease time, the client should resend the same copy notification request to the source server.

The cnr_stateid is a copy stateid that uniquely describes the state needed on the source server to track the proposed COPY. As defined in Section 8.2 of [RFC5661], a stateid is tied to the current filehandle, and if the same stateid is presented by two different clients, it may refer to different states. As the source does not know which netloc4 network location the destination might use to establish the COPY operation, it can use the cnr_stateid to identify that the destination is operating on behalf of the client. Thus, the source server MUST construct copy stateids such that they are distinct from all other stateids handed out to clients. These copy stateids MUST denote the same set of locks as each of the earlier delegation, locking, and open states for the client on the given file (see Section 4.3.1).

A successful response will also contain a list of netloc4 network location formats called cnr_source_server, on which the source is willing to accept connections from the destination. These might not be reachable from the client and might be located on networks to which the client has no connection.

This operation is unnecessary for an intra-server copy.

15.4. Operation 62: DEALLOCATE - Unreserve space in a region of a file

15.4.1. ARGUMENT

<CODE BEGINS>

```
struct DEALLOCATE4args {
        /* CURRENT_FH: file */
        stateid4        da_stateid;
        offset4         da_offset;
        length4         da_length;
};
```

<CODE ENDS>

15.4.2. RESULT

<CODE BEGINS>

```
struct DEALLOCATE4res {
        nfsstat4        dr_status;
};
```

<CODE ENDS>

15.4.3. DESCRIPTION

Whenever a client wishes to unreserve space for a region in a file, it calls the DEALLOCATE operation with the current filehandle set to the filehandle of the file in question, and with the start offset and length in bytes of the region set in da_offset and da_length, respectively. If no space was allocated or reserved for all or parts of the region, the DEALLOCATE operation will have no effect for the region that already is in unreserved state. All further READs from the region passed to DEALLOCATE MUST return zeros until overwritten.

CURRENT_FH must be a regular file. If CURRENT_FH is not a regular file, the operation MUST fail and return NFS4ERR_WRONG_TYPE.

The da_stateid MUST refer to a stateid that is valid for a WRITE operation and follows the rules for stateids in Sections 8.2.5 and 18.32.3 of [RFC5661].

Situations may arise where da_offset and/or da_offset + da_length will not be aligned to a boundary for which the server does allocations or deallocations. For most file systems, this is the block size of the file system. In such a case, the server can deallocate as many bytes as it can in the region. The blocks that cannot be deallocated MUST be zeroed.

DEALLOCATE will result in the space_used attribute being decreased by the number of bytes that were deallocated. The space_freed attribute may or may not decrease, depending on the support and whether the blocks backing the specified range were shared or not. The size attribute will remain unchanged.

15.5. Operation 63: IO_ADVISE - Send client I/O access pattern hints to the server

15.5.1. ARGUMENT

<CODE BEGINS>

```
enum IO_ADVISE_type4 {
        IO_ADVISE4_NORMAL                       = 0,
        IO_ADVISE4_SEQUENTIAL                   = 1,
        IO_ADVISE4_SEQUENTIAL_BACKWARDS         = 2,
        IO_ADVISE4_RANDOM                       = 3,
        IO_ADVISE4_WILLNEED                     = 4,
        IO_ADVISE4_WILLNEED_OPPORTUNISTIC       = 5,
        IO_ADVISE4_DONTNEED                     = 6,
        IO_ADVISE4_NOREUSE                      = 7,
        IO_ADVISE4_READ                         = 8,
        IO_ADVISE4_WRITE                        = 9,
        IO_ADVISE4_INIT_PROXIMITY               = 10
};

struct IO_ADVISE4args {
        /* CURRENT_FH: file */
        stateid4        iaa_stateid;
        offset4         iaa_offset;
        length4         iaa_count;
        bitmap4         iaa_hints;
};
```

<CODE ENDS>

15.5.2. RESULT

```
<CODE BEGINS>

struct IO_ADVISE4resok {
        bitmap4 ior_hints;
};

union IO_ADVISE4res switch (nfsstat4 ior_status) {
case NFS4_OK:
        IO_ADVISE4resok resok4;
default:
        void;
};

<CODE ENDS>
```

15.5.3. DESCRIPTION

The IO_ADVISE operation sends an I/O access pattern hint to the server for the owner of the stateid for a given byte range specified by iar_offset and iar_count. The byte range specified by iaa_offset and iaa_count need not currently exist in the file, but the iaa_hints will apply to the byte range when it does exist. If iaa_count is 0, all data following iaa_offset is specified. The server MAY ignore the advice.

The following are the allowed hints for a stateid holder:

IO_ADVISE4_NORMAL There is no advice to give. This is the default behavior.

IO_ADVISE4_SEQUENTIAL Expects to access the specified data sequentially from lower offsets to higher offsets.

IO_ADVISE4_SEQUENTIAL_BACKWARDS Expects to access the specified data sequentially from higher offsets to lower offsets.

IO_ADVISE4_RANDOM Expects to access the specified data in a random order.

IO_ADVISE4_WILLNEED Expects to access the specified data in the near future.

IO_ADVISE4_WILLNEED_OPPORTUNISTIC Expects to possibly access the data in the near future. This is a speculative hint, and therefore the server should prefetch data or indirect blocks only if it can be done at a marginal cost.

IO_ADVISE_DONTNEED Expects that it will not access the specified data in the near future.

IO_ADVISE_NOREUSE Expects to access the specified data once and then not reuse it thereafter.

IO_ADVISE4_READ Expects to read the specified data in the near future.

IO_ADVISE4_WRITE Expects to write the specified data in the near future.

IO_ADVISE4_INIT_PROXIMITY Informs the server that the data in the byte range remains important to the client.

Since IO_ADVISE is a hint, a server SHOULD NOT return an error and invalidate an entire COMPOUND request if one of the sent hints in iar_hints is not supported by the server. Also, the server MUST NOT return an error if the client sends contradictory hints to the server, e.g., IO_ADVISE4_SEQUENTIAL and IO_ADVISE4_RANDOM in a single IO_ADVISE operation. In these cases, the server MUST return success and an ior_hints value that indicates the hint it intends to implement. This may mean simply returning IO_ADVISE4_NORMAL.

The ior_hints returned by the server is primarily for debugging purposes, since the server is under no obligation to carry out the hints that it describes in the ior_hints result. In addition, while the server may have intended to implement the hints returned in ior_hints, the server may need to change its handling of a given file -- for example, because of memory pressure, additional IO_ADVISE hints sent by other clients, or heuristically detected file access patterns.

The server MAY return different advice than what the client requested. Some examples include another client advising of a different I/O access pattern, another client employing a different I/O access pattern, or inability of the server to support the requested I/O access pattern.

Each issuance of the IO_ADVISE operation overrides all previous issuances of IO_ADVISE for a given byte range. This effectively follows a strategy of "last hint wins" for a given stateid and byte range.

Clients should assume that hints included in an IO_ADVISE operation will be forgotten once the file is closed.

15.5.4. IMPLEMENTATION

The NFS client may choose to issue an IO_ADVISE operation to the server in several different instances.

The most obvious is in direct response to an application's execution of posix_fadvise(). In this case, IO_ADVISE4_WRITE and IO_ADVISE4_READ may be set, based upon the type of file access specified when the file was opened.

15.5.5. IO_ADVISE4_INIT_PROXIMITY

The IO_ADVISE4_INIT_PROXIMITY hint is non-POSIX in origin and can be used to convey that the client has recently accessed the byte range in its own cache. That is, it has not accessed it on the server but has accessed it locally. When the server reaches resource exhaustion, knowing which data is more important allows the server to make better choices about which data to, for example, purge from a cache or move to secondary storage. It also informs the server as to which delegations are more important, because if delegations are working correctly, once delegated to a client and the client has read the content for that byte range, a server might never receive another READ request for that byte range.

The IO_ADVISE4_INIT_PROXIMITY hint can also be used in a pNFS setting to let the client inform the metadata server as to the I/O statistics between the client and the storage devices. The metadata server is then free to use this information about client I/O to optimize the data storage location.

This hint is also useful in the case of NFS clients that are network-booting from a server. If the first client to be booted sends this hint, then it keeps the cache warm for the remaining clients.

15.5.6. pNFS File Layout Data Type Considerations

The IO_ADVISE considerations for pNFS are very similar to the COMMIT considerations for pNFS (see Section 13.7 of [RFC5661]). That is, as with COMMIT, some NFS server implementations prefer that IO_ADVISE be done on the storage device, and some prefer that it be done on the metadata server.

For the file's layout type, NFSv4.2 includes an additional hint, NFL42_CARE_IO_ADVISE_THRU_MDS, which is valid only on metadata servers running NFSv4.2 or higher. ("NFL" stands for "NFS File Layout".) Any file's layout obtained from an NFSv4.1 metadata server MUST NOT have NFL42_UFLG_IO_ADVISE_THRU_MDS set. Any file's layout

obtained with an NFSv4.2 metadata server MAY have NFL42_UFLG_IO_ADVISE_THRU_MDS set. However, if the layout utilizes NFSv4.1 storage devices, the IO_ADVISE operation cannot be sent to them.

If NFL42_UFLG_IO_ADVISE_THRU_MDS is set, the client MUST send the IO_ADVISE operation to the metadata server in order for it to be honored by the storage device. Once the metadata server receives the IO_ADVISE operation, it will communicate the advice to each storage device.

If NFL42_UFLG_IO_ADVISE_THRU_MDS is not set, then the client SHOULD send an IO_ADVISE operation to the appropriate storage device for the specified byte range. While the client MAY always send IO_ADVISE to the metadata server, if the server has not set NFL42_UFLG_IO_ADVISE_THRU_MDS, the client should expect that such an IO_ADVISE is futile. Note that a client SHOULD use the same set of arguments on each IO_ADVISE sent to a storage device for the same open file reference.

The server is not required to support different advice for different storage devices with the same open file reference.

15.5.6.1. Dense and Sparse Packing Considerations

The IO_ADVISE operation MUST use the iar_offset and byte range as dictated by the presence or absence of NFL4_UFLG_DENSE (see Section 13.4.4 of [RFC5661]).

For example, if NFL4_UFLG_DENSE is present, then (1) a READ or WRITE to the storage device for iaa_offset 0 really means iaa_offset 10000 in the logical file and (2) an IO_ADVISE for iaa_offset 0 means iaa_offset 10000 in the logical file.

For example, if NFL4_UFLG_DENSE is absent, then (1) a READ or WRITE to the storage device for iaa_offset 0 really means iaa_offset 0 in the logical file and (2) an IO_ADVISE for iaa_offset 0 means iaa_offset 0 in the logical file.

For example, if NFL4_UFLG_DENSE is present, the stripe unit is 1000 bytes and the stripe count is 10, and the dense storage device file is serving iar_offset 0. A READ or WRITE to the storage device for iaa_offsets 0, 1000, 2000, and 3000 really means iaa_offsets 10000, 20000, 30000, and 40000 (implying a stripe count of 10 and a stripe unit of 1000), and then an IO_ADVISE sent to the same storage device with an iaa_offset of 500 and an iaa_count of 3000 means that the IO_ADVISE applies to these byte ranges of the dense storage device file:

- 500 to 999
- 1000 to 1999
- 2000 to 2999
- 3000 to 3499

That is, the contiguous range 500 to 3499, as specified in IO_ADVISE.

It also applies to these byte ranges of the logical file:

- 10500 to 10999 (500 bytes)
- 20000 to 20999 (1000 bytes)
- 30000 to 30999 (1000 bytes)
- 40000 to 40499 (500 bytes)

(total 3000 bytes)

For example, if NFL4_UFLG_DENSE is absent, the stripe unit is 250 bytes, the stripe count is 4, and the sparse storage device file is serving iaa_offset 0. Then, a READ or WRITE to the storage device for iaa_offsets 0, 1000, 2000, and 3000 really means iaa_offsets 0, 1000, 2000, and 3000 in the logical file, keeping in mind that in the storage device file byte ranges 250 to 999, 1250 to 1999, 2250 to 2999, and 3250 to 3999 are not accessible. Then, an IO_ADVISE sent to the same storage device with an iaa_offset of 500 and an iaa_count of 3000 means that the IO_ADVISE applies to these byte ranges of the logical file and the sparse storage device file:

- 500 to 999 (500 bytes) - no effect
- 1000 to 1249 (250 bytes) - effective
- 1250 to 1999 (750 bytes) - no effect
- 2000 to 2249 (250 bytes) - effective
- 2250 to 2999 (750 bytes) - no effect
- 3000 to 3249 (250 bytes) - effective
- 3250 to 3499 (250 bytes) - no effect

(subtotal 2250 bytes) - no effect
(subtotal 750 bytes) - effective
(grand total 3000 bytes) - no effect + effective

If neither the NFL42_UFLG_IO_ADVISE_THRU_MDS flag nor the NFL4_UFLG_DENSE flag is set in the layout, then any IO_ADVISE request sent to the data server with a byte range that overlaps stripe units that the data server does not serve MUST NOT result in the status NFS4ERR_PNFS_IO_HOLE. Instead, the response SHOULD be successful, and if the server applies IO_ADVISE hints on any stripe units that overlap with the specified range, those hints SHOULD be indicated in the response.

15.6. Operation 64: LAYOUTERROR - Provide errors for the layout

15.6.1. ARGUMENT

<CODE BEGINS>

```
struct device_error4 {
        deviceid4       de_deviceid;
        nfsstat4        de_status;
        nfs_opnum4      de_opnum;
};

struct LAYOUTERROR4args {
        /* CURRENT_FH: file */
        offset4                 lea_offset;
        length4                 lea_length;
        stateid4                lea_stateid;
        device_error4           lea_errors<>;
};
```

<CODE ENDS>

15.6.2. RESULT

<CODE BEGINS>

```
struct LAYOUTERROR4res {
        nfsstat4        ler_status;
};
```

<CODE ENDS>

15.6.3. DESCRIPTION

The client can use LAYOUTERROR to inform the metadata server about errors in its interaction with the layout (see Section 12 of [RFC5661]) represented by the current filehandle, client ID (derived from the session ID in the preceding SEQUENCE operation), byte range (lea_offset + lea_length), and lea_stateid.

Each individual device_error4 describes a single error associated with a storage device, which is identified via de_deviceid. If the layout type (see Section 12.2.7 of [RFC5661]) supports NFSv4 operations, then the operation that returned the error is identified via de_opnum. If the layout type does not support NFSv4 operations, then either (1) it MAY choose to map the operation onto one of the allowed operations that can be sent to a storage device with the file layout type (see Section 3.3) or (2) it can signal no support for operations by marking de_opnum with the ILLEGAL operation. Finally, the NFS error value (nfsstat4) encountered is provided via de_status and may consist of the following error codes:

NFS4ERR_NXIO: The client was unable to establish any communication with the storage device.

NFS4ERR_*: The client was able to establish communication with the storage device and is returning one of the allowed error codes for the operation denoted by de_opnum.

Note that while the metadata server may return an error associated with the layout stateid or the open file, it MUST NOT return an error in the processing of the errors. If LAYOUTERROR is in a COMPOUND before LAYOUTRETURN, it MUST NOT introduce an error other than what LAYOUTRETURN would already encounter.

15.6.4. IMPLEMENTATION

There are two broad classes of errors: transient and persistent. The client SHOULD strive to only use this new mechanism to report persistent errors. It MUST be able to deal with transient issues by itself. Also, while the client might consider an issue to be persistent, it MUST be prepared for the metadata server to consider such issues to be transient. A prime example of this is if the metadata server fences off a client from either a stateid or a filehandle. The client will get an error from the storage device and might relay either NFS4ERR_ACCESS or NFS4ERR_BAD_STATEID back to the metadata server, with the belief that this is a hard error. If the metadata server is informed by the client that there is an error, it can safely ignore that. For the metadata server, the mission is accomplished in that the client has returned a layout that the metadata server had most likely recalled.

The client might also need to inform the metadata server that it cannot reach one or more of the storage devices. While the metadata server can detect the connectivity of both of these paths:

- metadata server to storage device
- metadata server to client

it cannot determine if the client and storage device path is working. As with the case of the storage device passing errors to the client, it must be prepared for the metadata server to consider such outages as being transitory.

Clients are expected to tolerate transient storage device errors, and hence clients SHOULD NOT use the LAYOUTERROR error handling for device access problems that may be transient. The methods by which a client decides whether a device access problem is transient or persistent are implementation specific but may include retrying I/Os to a data server under appropriate conditions.

When an I/O to a storage device fails, the client SHOULD retry the failed I/O via the metadata server. In this situation, before retrying the I/O, the client SHOULD return the layout, or the affected portion thereof, and SHOULD indicate which storage device or devices was problematic. The client needs to do this when the storage device is being unresponsive in order to fence off any failed write attempts and ensure that they do not end up overwriting any later data being written through the metadata server. If the client does not do this, the metadata server MAY issue a layout recall callback in order to perform the retried I/O.

The client needs to be cognizant that since this error handling is optional in the metadata server, the metadata server may silently ignore this functionality. Also, as the metadata server may consider some issues the client reports to be expected, the client might find it difficult to detect a metadata server that has not implemented error handling via LAYOUTERROR.

If a metadata server is aware that a storage device is proving problematic to a client, the metadata server SHOULD NOT include that storage device in any pNFS layouts sent to that client. If the metadata server is aware that a storage device is affecting many clients, then the metadata server SHOULD NOT include that storage device in any pNFS layouts sent out. If a client asks for a new layout for the file from the metadata server, it MUST be prepared for the metadata server to return that storage device in the layout. The metadata server might not have any choice in using the storage device, i.e., there might only be one possible layout for the system.

Also, in the case of existing files, the metadata server might have no choice regarding which storage devices to hand out to clients.

The metadata server is not required to indefinitely retain per-client storage device error information. The metadata server is also not required to automatically reinstate the use of a previously problematic storage device; administrative intervention may be required instead.

15.7. Operation 65: LAYOUTSTATS - Provide statistics for the layout

15.7.1. ARGUMENT

<CODE BEGINS>

```
struct layoutupdate4 {
        layouttype4             lou_type;
        opaque                  lou_body<>;
};

struct io_info4 {
        uint64_t        ii_count;
        uint64_t        ii_bytes;
};

struct LAYOUTSTATS4args {
        /* CURRENT_FH: file */
        offset4                 lsa_offset;
        length4                 lsa_length;
        stateid4                lsa_stateid;
        io_info4                lsa_read;
        io_info4                lsa_write;
        deviceid4               lsa_deviceid;
        layoutupdate4           lsa_layoutupdate;
};
```

<CODE ENDS>

15.7.2. RESULT

<CODE BEGINS>

```
struct LAYOUTSTATS4res {
        nfsstat4        lsr_status;
};
```

<CODE ENDS>

15.7.3. DESCRIPTION

The client can use LAYOUTSTATS to inform the metadata server about its interaction with the layout (see Section 12 of [RFC5661]) represented by the current filehandle, client ID (derived from the session ID in the preceding SEQUENCE operation), byte range (lsa_offset and lsa_length), and lsa_stateid. lsa_read and lsa_write allow non-layout-type-specific statistics to be reported. lsa_deviceid allows the client to specify to which storage device the statistics apply. The remaining information the client is presenting is specific to the layout type and presented in the lsa_layoutupdate field. Each layout type MUST define the contents of lsa_layoutupdate in their respective specifications.

LAYOUTSTATS can be combined with IO_ADVISE (see Section 15.5) to augment the decision-making process of how the metadata server handles a file. That is, IO_ADVISE lets the server know that a byte range has a certain characteristic, but not necessarily the intensity of that characteristic.

The statistics are cumulative, i.e., multiple LAYOUTSTATS updates can be in flight at the same time. The metadata server can examine the packet's timestamp to order the different calls. The first LAYOUTSTATS sent by the client SHOULD be from the opening of the file. The choice of how often to update the metadata server is made by the client.

Note that while the metadata server may return an error associated with the layout stateid or the open file, it MUST NOT return an error in the processing of the statistics.

15.8. Operation 66: OFFLOAD_CANCEL - Stop an offloaded operation

15.8.1. ARGUMENT

<CODE BEGINS>

```
struct OFFLOAD_CANCEL4args {
        /* CURRENT_FH: file to cancel */
        stateid4        oca_stateid;
};
```

<CODE ENDS>

15.8.2. RESULT

<CODE BEGINS>

```
struct OFFLOAD_CANCEL4res {
        nfsstat4        ocr_status;
};
```

<CODE ENDS>

15.8.3. DESCRIPTION

OFFLOAD_CANCEL is used by the client to terminate an asynchronous operation, which is identified by both CURRENT_FH and the oca_stateid. That is, there can be multiple OFFLOAD_CANCEL operations acting on the file, and the stateid will identify to the server exactly which one is to be stopped. Currently, there are only two operations that can decide to be asynchronous: COPY and WRITE_SAME.

In the context of server-to-server copy, the client can send OFFLOAD_CANCEL to either the source or destination server, albeit with a different stateid. The client uses OFFLOAD_CANCEL to inform the destination to stop the active transfer and uses the stateid it got back from the COPY operation. The client uses OFFLOAD_CANCEL and the stateid it used in the COPY_NOTIFY to inform the source to not allow any more copying from the destination.

OFFLOAD_CANCEL is also useful in situations in which the source server granted a very long or infinite lease on the destination server's ability to read the source file and all COPY operations on the source file have been completed.

15.9. Operation 67: OFFLOAD_STATUS - Poll for the status of an asynchronous operation

15.9.1. ARGUMENT

<CODE BEGINS>

```
struct OFFLOAD_STATUS4args {
        /* CURRENT_FH: destination file */
        stateid4        osa_stateid;
};
```

<CODE ENDS>

15.9.2. RESULT

<CODE BEGINS>

```
struct OFFLOAD_STATUS4resok {
        length4         osr_count;
        nfsstat4        osr_complete<1>;
};

union OFFLOAD_STATUS4res switch (nfsstat4 osr_status) {
case NFS4_OK:
        OFFLOAD_STATUS4resok            osr_resok4;
default:
        void;
};
```

<CODE ENDS>

15.9.3. DESCRIPTION

OFFLOAD_STATUS can be used by the client to query the progress of an asynchronous operation, which is identified by both CURRENT_FH and the osa_stateid. If this operation is successful, the number of bytes processed is returned to the client in the osr_count field.

If the optional osr_complete field is present, the asynchronous operation has completed. In this case, the status value indicates the result of the asynchronous operation. In all cases, the server will also deliver the final results of the asynchronous operation in a CB_OFFLOAD operation.

The failure of this operation does not indicate the result of the asynchronous operation in any way.

15.10. Operation 68: READ_PLUS - READ data or holes from a file

15.10.1. ARGUMENT

```
<CODE BEGINS>

struct READ_PLUS4args {
        /* CURRENT_FH: file */
        stateid4        rpa_stateid;
        offset4         rpa_offset;
        count4          rpa_count;
};

<CODE ENDS>
```

15.10.2. RESULT

```
<CODE BEGINS>

enum data_content4 {
        NFS4_CONTENT_DATA = 0,
        NFS4_CONTENT_HOLE = 1
};

struct data_info4 {
        offset4         di_offset;
        length4         di_length;
};

struct data4 {
        offset4         d_offset;
        opaque          d_data<>;
};

union read_plus_content switch (data_content4 rpc_content) {
case NFS4_CONTENT_DATA:
        data4           rpc_data;
case NFS4_CONTENT_HOLE:
        data_info4      rpc_hole;
default:
        void;
};
```

```
   /*
    * Allow a return of an array of contents.
    */
   struct read_plus_res4 {
           bool                    rpr_eof;
           read_plus_content       rpr_contents<>;
   };

   union READ_PLUS4res switch (nfsstat4 rp_status) {
   case NFS4_OK:
           read_plus_res4  rp_resok4;
   default:
           void;
   };

   <CODE ENDS>
```

15.10.3. DESCRIPTION

The READ_PLUS operation is based upon the NFSv4.1 READ operation (see Section 18.22 of [RFC5661]) and similarly reads data from the regular file identified by the current filehandle.

The client provides an rpa_offset of where the READ_PLUS is to start and an rpa_count of how many bytes are to be read. An rpa_offset of zero means that data will be read starting at the beginning of the file. If rpa_offset is greater than or equal to the size of the file, the status NFS4_OK is returned with di_length (the data length) set to zero and eof set to TRUE.

The READ_PLUS result is comprised of an array of rpr_contents, each of which describes a data_content4 type of data. For NFSv4.2, the allowed values are data and hole. A server MUST support both the data type and the hole if it uses READ_PLUS. If it does not want to support a hole, it MUST use READ. The array contents MUST be contiguous in the file.

Holes SHOULD be returned in their entirety -- clients must be prepared to get more information than they requested. Both the start and the end of the hole may exceed what was requested. If data to be returned is comprised entirely of zeros, then the server SHOULD return that data as a hole instead.

The server may elect to return adjacent elements of the same type. For example, if the server has a range of data comprised entirely of zeros and then a hole, it might want to return two adjacent holes to the client.

If the client specifies an rpa_count value of zero, the READ_PLUS succeeds and returns zero bytes of data. In all situations, the server may choose to return fewer bytes than specified by the client. The client needs to check for this condition and handle the condition appropriately.

If the client specifies data that is entirely contained within a hole of the file (i.e., both rpa_offset and rpa_offset + rpa_count are within the hole), then the di_offset and di_length returned MAY be for the entire hole. If the owner has a locked byte range covering rpa_offset and rpa_count entirely, the di_offset and di_length MUST NOT be extended outside the locked byte range. This result is considered valid until the file is changed (detected via the change attribute). The server MUST provide the same semantics for the hole as if the client read the region and received zeros; the implied hole's contents lifetime MUST be exactly the same as any other read data.

If the client specifies data by an rpa_offset that begins in a non-hole of the file but extends into a hole (the rpa_offset + rpa_count is in the hole), the server should return an array comprised of both data and a hole. The client MUST be prepared for the server to return a short read describing just the data. The client will then issue another READ_PLUS for the remaining bytes, to which the server will respond with information about the hole in the file.

Except when special stateids are used, the stateid value for a READ_PLUS request represents a value returned from a previous byte-range lock or share reservation request or the stateid associated with a delegation. The stateid identifies the associated owners, if any, and is used by the server to verify that the associated locks are still valid (e.g., have not been revoked).

If the read ended at the end of the file (formally, in a correctly formed READ_PLUS operation, if rpa_offset + rpa_count is equal to the size of the file) or the READ_PLUS operation extends beyond the size of the file (if rpa_offset + rpa_count is greater than the size of the file), eof is returned as TRUE; otherwise, it is FALSE. A successful READ_PLUS of an empty file will always return eof as TRUE.

If the current filehandle is not an ordinary file, an error will be returned to the client. In the case that the current filehandle represents an object of type NF4DIR, NFS4ERR_ISDIR is returned. If the current filehandle designates a symbolic link, NFS4ERR_SYMLINK is returned. In all other cases, NFS4ERR_WRONG_TYPE is returned.

For a READ_PLUS with a stateid value of all bits equal to zero, the server MAY allow the READ_PLUS to be serviced subject to mandatory byte-range locks or the current share deny modes for the file. For a READ_PLUS with a stateid value of all bits equal to one, the server MAY allow READ_PLUS operations to bypass locking checks at the server.

On success, the current filehandle retains its value.

15.10.3.1. Note on Client Support of Arms of the Union

It was decided not to add a means for the client to inform the server as to which arms of READ_PLUS it would support. In a later minor version, it may become necessary for the introduction of a new operation that would allow the client to inform the server as to whether it supported the new arms of the union of data types available in READ_PLUS.

15.10.4. IMPLEMENTATION

In general, the IMPLEMENTATION notes for READ in Section 18.22.4 of [RFC5661] also apply to READ_PLUS.

15.10.4.1. Additional pNFS Implementation Information

With pNFS, the semantics of using READ_PLUS remains the same. Any data server MAY return a hole result for a READ_PLUS request that it receives. When a data server chooses to return such a result, it has the option of returning information for the data stored on that data server (as defined by the data layout), but it MUST NOT return results for a byte range that includes data managed by another data server.

If mandatory locking is enforced, then the data server must also ensure that only information that is within the owner's locked byte range is returned.

15.10.5. READ_PLUS with Sparse Files: Example

The following table describes a sparse file. For each byte range, the file contains either non-zero data or a hole. In addition, the server in this example will only create a hole if it is greater than 32K.

Byte Range	Contents
0-15999	Hole
16K-31999	Non-Zero
32K-255999	Hole
256K-287999	Non-Zero
288K-353999	Hole
354K-417999	Non-Zero

Table 7: Sparse File

Under the given circumstances, if a client was to read from the file with a maximum read size of 64K, the following will be the results for the given READ_PLUS calls. This assumes that the client has already opened the file, acquired a valid stateid ("s" in the example), and just needs to issue READ_PLUS requests.

1. READ_PLUS(s, 0, 64K) --> NFS_OK, eof = FALSE, <data[0,32K], hole[32K,224K]>. Since the first hole is less than the server's minimum hole size, the first 32K of the file is returned as data and the remaining 32K is returned as a hole that actually extends to 256K.

2. READ_PLUS(s, 32K, 64K) --> NFS_OK, eof = FALSE, <hole[32K,224K]>. The requested range was all zeros, and the current hole begins at offset 32K and is 224K in length. Note that the client should not have followed up the previous READ_PLUS request with this one, as the hole information from the previous call extended past what the client was requesting.

3. READ_PLUS(s, 256K, 64K) --> NFS_OK, eof = FALSE, <data[256K, 288K], hole[288K, 354K]>. Returns an array of the 32K data and the hole, which extends to 354K.

4. READ_PLUS(s, 354K, 64K) --> NFS_OK, eof = TRUE, <data[354K, 418K]>. Returns the final 64K of data and informs the client that there is no more data in the file.

15.11. Operation 69: SEEK - Find the next data or hole

15.11.1. ARGUMENT

```
<CODE BEGINS>

enum data_content4 {
        NFS4_CONTENT_DATA = 0,
        NFS4_CONTENT_HOLE = 1
};

struct SEEK4args {
        /* CURRENT_FH: file */
        stateid4        sa_stateid;
        offset4         sa_offset;
        data_content4   sa_what;
};

<CODE ENDS>
```

15.11.2. RESULT

```
<CODE BEGINS>

struct seek_res4 {
        bool            sr_eof;
        offset4         sr_offset;
};

union SEEK4res switch (nfsstat4 sa_status) {
case NFS4_OK:
        seek_res4       resok4;
default:
        void;
};

<CODE ENDS>
```

15.11.3. DESCRIPTION

SEEK is an operation that allows a client to determine the location of the next data_content4 in a file. It allows an implementation of the emerging extension to the lseek(2) function to allow clients to determine the next hole whilst in data or the next data whilst in a hole.

From the given sa_offset, find the next data_content4 of type sa_what in the file. If the server cannot find a corresponding sa_what, then the status will still be NFS4_OK, but sr_eof would be TRUE. If the server can find the sa_what, then the sr_offset is the start of that content. If the sa_offset is beyond the end of the file, then SEEK MUST return NFS4ERR_NXIO.

All files MUST have a virtual hole at the end of the file. That is, if a file system does not support sparse files, then a COMPOUND with {SEEK 0 NFS4_CONTENT_HOLE;} would return a result of {SEEK 1 X;}, where "X" was the size of the file.

SEEK must follow the same rules for stateids as READ_PLUS (Section 15.10.3).

15.12. Operation 70: WRITE_SAME - WRITE an ADB multiple times to a file

15.12.1. ARGUMENT

<CODE BEGINS>

```
enum stable_how4 {
        UNSTABLE4       = 0,
        DATA_SYNC4      = 1,
        FILE_SYNC4      = 2
};

struct app_data_block4 {
        offset4         adb_offset;
        length4         adb_block_size;
        length4         adb_block_count;
        length4         adb_reloff_blocknum;
        count4          adb_block_num;
        length4         adb_reloff_pattern;
        opaque          adb_pattern<>;
};

struct WRITE_SAME4args {
        /* CURRENT_FH: file */
        stateid4        wsa_stateid;
        stable_how4     wsa_stable;
        app_data_block4 wsa_adb;
};
```

<CODE ENDS>

15.12.2. RESULT

```
<CODE BEGINS>

struct write_response4 {
        stateid4        wr_callback_id<1>;
        length4         wr_count;
        stable_how4     wr_committed;
        verifier4       wr_writeverf;
};

union WRITE_SAME4res switch (nfsstat4 wsr_status) {
case NFS4_OK:
        write_response4         resok4;
default:
        void;
};

<CODE ENDS>
```

15.12.3. DESCRIPTION

The WRITE_SAME operation writes an application data block to the regular file identified by the current filehandle (see WRITE SAME (10) in [T10-SBC2]). The target file is specified by the current filehandle. The data to be written is specified by an app_data_block4 structure (Section 8.1.1). The client specifies with the wsa_stable parameter the method of how the data is to be processed by the server. It is treated like the stable parameter in the NFSv4.1 WRITE operation (see Section 18.32.3 of [RFC5661]).

A successful WRITE_SAME will construct a reply for wr_count, wr_committed, and wr_writeverf as per the NFSv4.1 WRITE operation results. If wr_callback_id is set, it indicates an asynchronous reply (see Section 15.12.3.1).

As it is an OPTIONAL operation, WRITE_SAME has to support NFS4ERR_NOTSUPP. As it is an extension of WRITE, it has to support all of the errors returned by WRITE. If the client supports WRITE_SAME, it MUST support CB_OFFLOAD.

If the server supports ADBs, then it MUST support the WRITE_SAME operation. The server has no concept of the structure imposed by the application. It is only when the application writes to a section of the file does order get imposed. In order to detect corruption even before the application utilizes the file, the application will want to initialize a range of ADBs using WRITE_SAME.

When the client invokes the WRITE_SAME operation, it wants to record the block structure described by the app_data_block4 into the file.

When the server receives the WRITE_SAME operation, it MUST populate adb_block_count ADBs in the file, starting at adb_offset. The block size will be given by adb_block_size. The ADBN (if provided) will start at adb_reloff_blocknum, and each block will be monotonically numbered, starting from adb_block_num in the first block. The pattern (if provided) will be at adb_reloff_pattern of each block and will be provided in adb_pattern.

The server SHOULD return an asynchronous result if it can determine that the operation will be long-running (see Section 15.12.3.1). Once either the WRITE_SAME finishes synchronously or the server uses CB_OFFLOAD to inform the client of the asynchronous completion of the WRITE_SAME, the server MUST return the ADBs to clients as data.

15.12.3.1. Asynchronous Transactions

ADB initialization may cause a server to decide to service the operation asynchronously. If it decides to do so, it sets the stateid in wr_callback_id to be that of the wsa_stateid. If it does not set the wr_callback_id, then the result is synchronous.

When the client determines that the reply will be given asynchronously, it should not assume anything about the contents of what it wrote until it is informed by the server that the operation is complete. It can use OFFLOAD_STATUS (Section 15.9) to monitor the operation and OFFLOAD_CANCEL (Section 15.8) to cancel the operation. An example of an asynchronous WRITE_SAME is shown in Figure 6. Note that, as with the COPY operation, WRITE_SAME must provide a stateid for tracking the asynchronous operation.

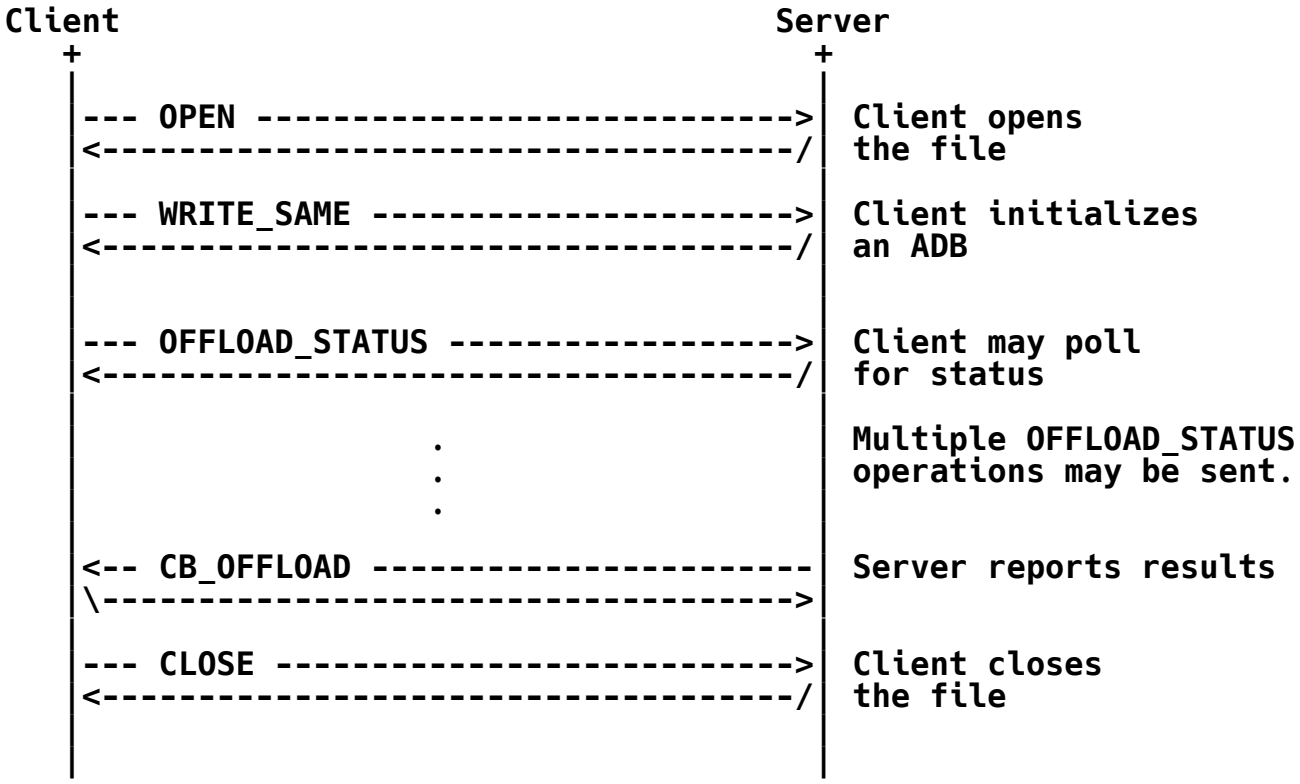


Figure 6: An Asynchronous WRITE_SAME

When CB_OFFLOAD informs the client of the successful WRITE_SAME, the write_response4 embedded in the operation will provide the necessary information that a synchronous WRITE_SAME would have provided.

Regardless of whether the operation is asynchronous or synchronous, it MUST still support the COMMIT operation semantics as outlined in Section 18.3 of [RFC5661]. That is, COMMIT works on one or more WRITE operations, and the WRITE_SAME operation can appear as several WRITE operations to the server. The client can use locking operations to control the behavior on the server with respect to long-running asynchronous WRITE_SAME operations.

15.12.3.2. Error Handling of a Partially Complete WRITE_SAME

WRITE_SAME will clone adb_block_count copies of the given ADB in consecutive order in the file, starting at adb_offset. An error can occur after writing the Nth ADB to the file. WRITE_SAME MUST appear to populate the range of the file as if the client used WRITE to transfer the instantiated ADBs. That is, the contents of the range will be easy for the client to determine in the case of a partially complete WRITE_SAME.

15.13. Operation 71: CLONE - Clone a range of a file into another file

15.13.1. ARGUMENT

<CODE BEGINS>

```
struct CLONE4args {
        /* SAVED_FH: source file */
        /* CURRENT_FH: destination file */
        stateid4        cl_src_stateid;
        stateid4        cl_dst_stateid;
        offset4         cl_src_offset;
        offset4         cl_dst_offset;
        length4         cl_count;
};
```

<CODE ENDS>

15.13.2. RESULT

<CODE BEGINS>

```
struct CLONE4res {
        nfsstat4        cl_status;
};
```

<CODE ENDS>

15.13.3. DESCRIPTION

The CLONE operation is used to clone file content from a source file specified by the SAVED_FH value into a destination file specified by CURRENT_FH without actually copying the data, e.g., by using a copy-on-write mechanism.

Both SAVED_FH and CURRENT_FH must be regular files. If either SAVED_FH or CURRENT_FH is not a regular file, the operation MUST fail and return NFS4ERR_WRONG_TYPE.

The ca_dst_stateid MUST refer to a stateid that is valid for a WRITE operation and follows the rules for stateids in Sections 8.2.5 and 18.32.3 of [RFC5661]. The ca_src_stateid MUST refer to a stateid that is valid for a READ operation and follows the rules for stateids in Sections 8.2.5 and 18.22.3 of [RFC5661]. If either stateid is invalid, then the operation MUST fail.

The cl_src_offset is the starting offset within the source file from which the data to be cloned will be obtained, and the cl_dst_offset is the starting offset of the target region into which the cloned data will be placed. An offset of 0 (zero) indicates the start of the respective file. The number of bytes to be cloned is obtained from cl_count, except that a cl_count of 0 (zero) indicates that the number of bytes to be cloned is the count of bytes between cl_src_offset and the EOF of the source file. Both cl_src_offset and cl_dst_offset must be aligned to the clone block size (Section 12.2.1). The number of bytes to be cloned must be a multiple of the clone block size, except in the case in which cl_src_offset plus the number of bytes to be cloned is equal to the source file size.

If the source offset or the source offset plus count is greater than the size of the source file, the operation MUST fail with NFS4ERR_INVAL. The destination offset or destination offset plus count may be greater than the size of the destination file.

If SAVED_FH and CURRENT_FH refer to the same file and the source and target ranges overlap, the operation MUST fail with NFS4ERR_INVAL.

If the target area of the CLONE operation ends beyond the end of the destination file, the offset at the end of the target area will determine the new size of the destination file. The contents of any block not part of the target area will be the same as if the file size were extended by a WRITE.

If the area to be cloned is not a multiple of the clone block size and the size of the destination file is past the end of the target area, the area between the end of the target area and the next multiple of the clone block size will be zeroed.

The CLONE operation is atomic in that other operations may not see any intermediate states between the state of the two files before the operation and after the operation. READs of the destination file will never see some blocks of the target area cloned without all of them being cloned. WRITEs of the source area will either have no effect on the data of the target file or be fully reflected in the target area of the destination file.

The completion status of the operation is indicated by cr_status.

16. NFSv4.2 Callback Operations

16.1. Operation 15: CB_OFFLOAD - Report the results of an asynchronous operation

16.1.1. ARGUMENT

<CODE BEGINS>

```
struct write_response4 {
        stateid4        wr_callback_id<1>;
        length4         wr_count;
        stable_how4     wr_committed;
        verifier4       wr_writeverf;
};

union offload_info4 switch (nfsstat4 coa_status) {
case NFS4_OK:
        write_response4 coa_resok4;
default:
        length4         coa_bytes_copied;
};

struct CB_OFFLOAD4args {
        nfs_fh4         coa_fh;
        stateid4        coa_stateid;
        offload_info4   coa_offload_info;
};
```

<CODE ENDS>

16.1.2. RESULT

<CODE BEGINS>

```
struct CB_OFFLOAD4res {
        nfsstat4        cor_status;
};
```

<CODE ENDS>

16.1.3. DESCRIPTION

CB_OFFLOAD is used to report to the client the results of an asynchronous operation, e.g., server-side COPY or WRITE_SAME. The coa_fh and coa_stateid identify the transaction, and the coa_status indicates success or failure. The coa_resok4.wr_callback_id MUST NOT be set. If the transaction failed, then the coa_bytes_copied contains the number of bytes copied before the failure occurred. The coa_bytes_copied value indicates the number of bytes copied but not which specific bytes have been copied.

If the client supports any of the following operations:

COPY: for both intra-server and inter-server asynchronous copies

WRITE_SAME: for ADB initialization

then the client is REQUIRED to support the CB_OFFLOAD operation.

There is a potential race between the reply to the original transaction on the forechannel and the CB_OFFLOAD callback on the backchannel. Section 2.10.6.3 of [RFC5661] describes how to handle this type of issue.

Upon success, the coa_resok4.wr_count presents for each operation:

COPY: the total number of bytes copied

WRITE_SAME: the same information that a synchronous WRITE_SAME would provide

17. Security Considerations

NFSv4.2 has all of the security concerns present in NFSv4.1 (see Section 21 of [RFC5661]), as well as those present in the server-side copy (see Section 4.9) and in Labeled NFS (see Section 9.6).

18. IANA Considerations

The IANA considerations for Labeled NFS are addressed in [RFC7569].

19. References

19.1. Normative References

[posix_fadvise]
The Open Group, "Section 'posix_fadvise()' of System Interfaces of The Open Group Base Specifications Issue 7", IEEE Std 1003.1, 2016 Edition (HTML Version), ISBN 1937218812, September 2016, <http://www.opengroup.org/>.

[posix_fallocate]
The Open Group, "Section 'posix_fallocate()' of System Interfaces of The Open Group Base Specifications Issue 7", IEEE Std 1003.1, 2016 Edition (HTML Version), ISBN 1937218812, September 2016, <http://www.opengroup.org/>.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <http://www.rfc-editor.org/info/rfc2119>.

[RFC3986] Berners-Lee, T., Fielding, R., and L. Masinter, "Uniform Resource Identifier (URI): Generic Syntax", STD 66, RFC 3986, DOI 10.17487/RFC3986, January 2005, <http://www.rfc-editor.org/info/rfc3986>.

[RFC5661] Shepler, S., Ed., Eisler, M., Ed., and D. Noveck, Ed., "Network File System (NFS) Version 4 Minor Version 1 Protocol", RFC 5661, DOI 10.17487/RFC5661, January 2010, <http://www.rfc-editor.org/info/rfc5661>.

[RFC5662] Shepler, S., Ed., Eisler, M., Ed., and D. Noveck, Ed., "Network File System (NFS) Version 4 Minor Version 1 External Data Representation Standard (XDR) Description", RFC 5662, DOI 10.17487/RFC5662, January 2010, <http://www.rfc-editor.org/info/rfc5662>.

[RFC7569] Quigley, D., Lu, J., and T. Haynes, "Registry Specification for Mandatory Access Control (MAC) Security Label Formats", RFC 7569, DOI 10.17487/RFC7569, July 2015, <http://www.rfc-editor.org/info/rfc7569>.

[RFC7861] Adamson, A. and N. Williams, "Remote Procedure Call (RPC) Security Version 3", RFC 7861, DOI 10.17487/RFC7861, November 2016, <http://www.rfc-editor.org/info/rfc7861>.

[RFC7863] Haynes, T., "Network File System (NFS) Version 4 Minor Version 2 External Data Representation Standard (XDR) Description", RFC 7863, DOI 10.17487/RFC7863, November 2016, <http://www.rfc-editor.org/info/rfc7863>.

19.2. Informative References

[Ashdown08] Ashdown, L., "Chapter 15: Validating Database Files and Backups", Oracle Database Backup and Recovery User's Guide 11g Release 1 (11.1), August 2008, <http://download.oracle.com/docs/cd/B28359_01/backup.111/b28270/rcmvalid.htm>.

[Baira08] Bairavasundaram, L., Goodson, G., Schroeder, B., Arpaci-Dusseau, A., and R. Arpaci-Dusseau, "An Analysis of Data Corruption in the Storage Stack", Proceedings of the 6th USENIX Symposium on File and Storage Technologies (FAST '08), 2008, <http://www.usenix.org/events/fast08/tech/full_papers/bairavasundaram/bairavasundaram.pdf>.

[IESG08] IESG, "IESG Processing of RFC Errata for the IETF Stream", July 2008, <https://www.ietf.org/iesg/statement/errata-processing.html>.

[LB96] LaPadula, L. and D. Bell, "MITRE Technical Report 2547, Volume II", Journal of Computer Security, Volume 4, Issue 2-3, 239-263, IOS Press, Amsterdam, The Netherlands, January 1996.

[McDougall07] McDougall, R. and J. Mauro, "Section 11.4.3: Detecting Memory Corruption", Solaris Internals: Solaris 10 and OpenSolaris Kernel Architecture, 2nd Edition, 2007.

[NFSv4-Versioning] Noveck, D., "Rules for NFSv4 Extensions and Minor Versions", Work in Progress, draft-ietf-nfsv4-versioning-07, October 2016.

[RFC959] Postel, J. and J. Reynolds, "File Transfer Protocol", STD 9, RFC 959, DOI 10.17487/RFC0959, October 1985, <http://www.rfc-editor.org/info/rfc959>.

[RFC1108] Kent, S., "U.S. Department of Defense Security Options for the Internet Protocol", RFC 1108, DOI 10.17487/RFC1108, November 1991, <http://www.rfc-editor.org/info/rfc1108>.

[RFC2401] Kent, S. and R. Atkinson, "Security Architecture for the Internet Protocol", RFC 2401, DOI 10.17487/RFC2401, November 1998, <http://www.rfc-editor.org/info/rfc2401>.

[RFC4506] Eisler, M., Ed., "XDR: External Data Representation Standard", STD 67, RFC 4506, DOI 10.17487/RFC4506, May 2006, <http://www.rfc-editor.org/info/rfc4506>.

[RFC4949] Shirey, R., "Internet Security Glossary, Version 2", FYI 36, RFC 4949, DOI 10.17487/RFC4949, August 2007, <http://www.rfc-editor.org/info/rfc4949>.

[RFC5663] Black, D., Fridella, S., and J. Glasgow, "Parallel NFS (pNFS) Block/Volume Layout", RFC 5663, DOI 10.17487/RFC5663, January 2010, <http://www.rfc-editor.org/info/rfc5663>.

[RFC7204] Haynes, T., "Requirements for Labeled NFS", RFC 7204, DOI 10.17487/RFC7204, April 2014, <http://www.rfc-editor.org/info/rfc7204>.

[RFC7230] Fielding, R., Ed., and J. Reschke, Ed., "Hypertext Transfer Protocol (HTTP/1.1): Message Syntax and Routing", RFC 7230, DOI 10.17487/RFC7230, June 2014, <http://www.rfc-editor.org/info/rfc7230>.

[RFC7530] Haynes, T., Ed., and D. Noveck, Ed., "Network File System (NFS) Version 4 Protocol", RFC 7530, DOI 10.17487/RFC7530, March 2015, <http://www.rfc-editor.org/info/rfc7530>.

[Strohm11] Strohm, R., "Chapter 2: Data Blocks, Extents, and Segments", Oracle Database Concepts 11g Release 1 (11.1), January 2011, <http://download.oracle.com/docs/cd/B28359_01/server.111/b28318/logical.htm>.

[T10-SBC2] Elliott, R., Ed., "ANSI INCITS 405-2005, Information Technology - SCSI Block Commands - 2 (SBC-2)", November 2004, <ftp://www.t10.org/t10/document.05/05-344r0.pdf>.

Acknowledgments

Tom Haynes would like to thank NetApp, Inc. for its funding of his time on this project.

For the topic "sharing change attribute implementation characteristics with NFSv4 clients", the original document was by Trond Myklebust.

For the NFS server-side copy, the original document was by James Lentini, Mike Eisler, Deepak Kenchammana, Anshul Madan, and Rahul Iyer. Tom Talpey co-authored an unpublished version of that document. It was also reviewed by a number of individuals: Pranoop Erasani, Tom Haynes, Arthur Lent, Trond Myklebust, Dave Noveck, Theresa Lingutla-Raj, Manjunath Shankararao, Satyam Vaghani, and Nico Williams. Anna Schumaker's early prototyping experience helped us avoid some traps. Also, both Olga Kornievskaia and Andy Adamson brought implementation experience to the use of copy stateids in the inter-server copy. Jorge Mora was able to optimize the handling of errors for the result of COPY.

For the NFS space reservation operations, the original document was by Mike Eisler, James Lentini, Manjunath Shankararao, and Rahul Iyer.

For the sparse file support, the original document was by Dean Hildebrand and Marc Eshel. Valuable input and advice was received from Sorin Faibish, Bruce Fields, Benny Halevy, Trond Myklebust, and Richard Scheffenegger.

For the application I/O hints, the original document was by Dean Hildebrand, Mike Eisler, Trond Myklebust, and Sam Falkner. Some early reviewers included Benny Halevy and Pranoop Erasani.

For Labeled NFS, the original document was by David Quigley, James Morris, Jarrett Lu, and Tom Haynes. Peter Staubach, Trond Myklebust, Stephen Smalley, Sorin Faibish, Nico Williams, and David Black also contributed in the final push to get this accepted.

Christoph Hellwig was very helpful in getting the WRITE_SAME semantics to model more of what T10 was doing for WRITE SAME (10) [T10-SBC2]. And he led the push to get space reservations to more closely model the posix_fallocate() operation.

Andy Adamson picked up the RPCSEC_GSSv3 work, which enabled both Labeled NFS and server-side copy to provide more secure options.

Christoph Hellwig provided the update to GETDEVICELIST.

Jorge Mora provided a very detailed review and caught some important issues with the tables.

During the review process, Talia Reyes-Ortiz helped the sessions run smoothly. While many people contributed here and there, the core reviewers were Andy Adamson, Pranoop Erasani, Bruce Fields, Chuck Lever, Trond Myklebust, David Noveck, Peter Staubach, and Mike Kupfer.

Elwyn Davies was the General Area Reviewer for this document, and his insights as to the relationship of this document and both [RFC5661] and [RFC7530] were very much appreciated!

Author's Address

Thomas Haynes
Primary Data, Inc.
4300 El Camino Real Ste 100
Los Altos, CA 94022
United States of America

Phone: +1 408 215 1519
Email: thomas.haynes@primarydata.com