Object-based pNFS Operations
draft-ietf-nfsv4-pnfs-obj-04

Status of this Memo

By submitting this Internet-Draft, each author represents that any applicable patent or other IPR claims of which he or she is aware have been or will be disclosed, and any of which he or she becomes aware will be disclosed, in accordance with Section 6 of BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt.

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

This Internet-Draft will expire on March 8, 2008.

Copyright Notice

Copyright (C) The IETF Trust (2007).

Abstract

This Internet-Draft provides a description of the object-based pNFS extension for NFSv4. This is a companion to the main pnfs specification in the NFSv4 Minor Version 1 Internet Draft, which is currently draft-ietf-nfsv4-minorversion1-13.txt.
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 [1].

Table of Contents

1.  Introduction .............................................. 4
2.  Object Storage Device Addressing and Discovery ............ 4
   2.1.  pnfs_osd_addr_type4 ................................. 5
   2.2.  pnfs_osd_deviceaddr4 ............................... 6
3.  Object-Based Layout ........................................ 6
   3.1.  pnfs_osd_layout4 .................................... 7
   3.1.1.  pnfs_osd_objid4 .................................. 7
   3.1.2.  pnfs_osd_version4 .................................. 8
   3.1.3.  pnfs_osd_object_cred4 ............................. 8
   3.1.4.  pnfs_osd RAID_algorithm4 ......................... 10
   3.1.5.  pnfs_osd_data map4 ............................... 10
   3.2.  Data Mapping Schemes ................................ 11
   3.2.1.  Simple Striping .................................... 11
   3.2.2.  Nested Striping .................................... 12
   3.2.3.  Mirroring ......................................... 13
3.3.  RAID Algorithms ......................................... 14
   3.3.1.  PNFS_OSD_RAID_0 ................................... 14
   3.3.2.  PNFS_OSD_RAID_4 ................................... 14
   3.3.3.  PNFS_OSD_RAID_5 ................................... 15
   3.3.4.  PNFS_OSD_RAID_PQ .................................. 15
   3.3.5.  RAID Usage and implementation notes ............... 16
4.  Object-Based Layout Update ................................ 16
   4.1.  pnfs_osd_layoutupdate4 ............................... 16
   4.1.1.  pnfs_osd_deltaspaceused4 ......................... 17
   4.1.2.  pnfs_osd_errno4 ................................... 17
   4.1.3.  pnfs_osd_ioerr4 ................................... 18
5.  Object-Based Creation Layout Hint .......................... 19
   5.1.  pnfs_osd_layouthint4 ............................... 19
6.  Layout Segments ............................................ 20
   6.1.  CB_LAYOUTRECALL and LAYOUTRETURN .................... 20
   6.2.  LAYOUTCOMMIT ....................................... 21
7.  Recalling Layouts ......................................... 21
   7.1.  CB_RECALL_ANY ....................................... 22
8.  Client Fencing ............................................ 22
9.  Security Considerations ................................... 23
   9.1.  OSD Security Data Types ............................... 24
   9.2.  The OSD Security Protocol ............................ 24
   9.3.  Protocol Privacy Requirements ........................ 25
   9.4.  Revoking Capabilities ................................. 26
1. Introduction

In pNFS, the file server returns typed layout structures that describe where file data is located. There are different layouts for different storage systems and methods of arranging data on storage devices. This document describes the layouts used with object-based storage devices (OSD) that are accessed according to the iSCSI/OSD storage protocol standard (SNIA T10/1355-D [2]).

An "object" is a container for data and attributes, and files are stored in one or more objects. The OSD protocol specifies several operations on objects, including READ, WRITE, FLUSH, GET ATTRIBUTES, SET ATTRIBUTES, CREATE and DELETE. However, using the object-based layout the client only uses the READ, WRITE, GET ATTRIBUTES and FLUSH commands. The other commands are only used by the pNFS server.

An object-based layout for pNFS includes object identifiers, capabilities that allow clients to READ or WRITE those objects, and various parameters that control how file data is striped across their component objects. The OSD protocol has a capability-based security scheme that allows the pNFS server to control what operations and what objects can be used by clients. This scheme is described in more detail in the Security Considerations section (Section 9).

2. Object Storage Device Addressing and Discovery

Data operations to an OSD require the client to know the "address" of each OSD’s root object. The root object is synonymous with SCSI logical unit. The client specifies SCSI logical units to its SCSI stack using a representation local to the client. Because these representations are local, GETDEVICEINFO must return information that can be used by the client to select the correct local representation.

In the block world, a set offset (logical block number or track/sector) contains a disk label. This label identifies the disk uniquely. In contrast, an OSD has a standard set of attributes on its root object. For device identification purposes the OSD System ID (root information attribute number 3) and/or OSD Name (root information attribute number 9) are used as the label. These appear in the pnfs_osd_deviceaddr4 type below under the "systemid" and "osdname" fields.

In some situations, SCSI target discovery may need to be driven based on information contained in the GETDEVICEINFO response. One example of this is iSCSI targets that are not known to the client until a layout has been requested. Eventually iSCSI will adopt ANSI T10 SAM-3, at which time the World Wide Name (WWN aka, EUI-64/EUI-128)
naming conventions can be specified. In addition, Fibre Channel (FC) SCSI targets have a unique WWN. Although these FC targets have already been discovered, some implementations may want to specify the WWN in addition to the label. This information appears as the "target" and "lun" fields in the pnfs_osd_deviceaddr4 type described below.

The systemid is used by the client, along with the object credential to sign each request with the request integrity check value. This method protects the client from unintentionally accessing a device if the device address mapping was changed (or revoked). The server computes the capability_key using its own view of the systemid associated with the respective deviceid present in the credential. If the client's view of the deviceid mapping is stale, the client will use the wrong systemid (which must be system-wide unique) and the I/O request to the OSD will fail to pass the integrity check verification.

To recover from this condition the client should report the error via LAYOUTCOMMIT, return the layout using LAYOUTRETURN, and invalidate all the device address mappings associated with this layout. The client can then ask for a new layout if it wishes using LAYOUTGET and resolve the referenced deviceids using GETDEVICEINFO or GETDEVICELIST.

The server MUST provide either the systemid, the OSD name, or both. When the OSD name is present the client SHOULD get the root information attributes whenever it establishes communication with the OSD and verify that the OSD name it got from the OSD matches the one sent by the metadata server. If the systemid was not given by the server it MUST be taken from the OSD-provided attribute; note that in this case the OSD GET ATTRIBUTES operation must be performed with the NOSEC security method.

2.1. pnfs_osd_addr_type4

The following enum specifies the manner in which a scsi target can be specified. The target can be specified as a network address, as an Internet Qualified Name (IQN), or by the World-Wide Name (WWN) of the target.

```c
enum pnfs_obj_addr_type4 {
    OBJ_TARGET_NETADDR = 1,
    OBJ_TARGET_IQN     = 2,
    OBJ_TARGET_WWN     = 3
};
```
2.2. pnfs_osd_deviceaddr4

The specification for an object device address is as follows:

```c
struct pnfs_osd_deviceaddr4 {
    union target switch (pnfs_osd_addr_type4 type) {
        case OBJ_TARGET_NETADDR:
            pnfs_netaddr4  netaddr;

        case OBJ_TARGET_IQN:
            string          iqn<>;

        case OBJ_TARGET_WWN:
            string          wwn<>;

        default:
            void;
    }
    uint64_t            lun;
    opaque              systemid<>;
    opaque              osdname<>;
};
```

3. Object-Based Layout

The `layout4` type is defined in the NFSv4.1 draft [6] as follows:

```c
enum layouttype4 {
    LAYOUT4_NFSV4_1_FILES   = 1,
    LAYOUT4_OSD2_OBJECTS    = 2,
    LAYOUT4_BLOCK_VOLUME    = 3
};

struct layout_content4 {
    layouttype4             loc_type;
    opaque                  loc_body<>;
};

struct layout4 {
    offset4                 lo_offset;
    length4                 lo_length;
    layoutiomode4           lo_iomode;
    layout_content4         lo_content;
};
```

This document defines structure associated with the `layouttype4`
value, LAYOUT4_OSD2_OBJECTS. The NFSv4.1 draft [6] specifies the loc_body structure as an XDR type "opaque". The opaque layout is uninterpreted by the generic pNFS client layers, but obviously must be interpreted by the object-storage layout driver. This document defines the structure of this opaque value, pnfs_osd_layout4.

3.1. pnfs_osd_layout4

```c
struct pnfs_osd_layout4 {
    pnfs_osd_data_map4      map;
    pnfs_osd_object_cred4   components<>;
};
```

The pnfs_osd_layout4 structure specifies a layout over a set of component objects. The components field is an array of object identifiers and security credentials that grant access to each object. The organization of the data is defined by the pnfs_osd_data_map4 type that specifies how the file’s data is mapped onto the component objects (i.e., the striping pattern). The data placement algorithm that maps file data onto component objects assume that each component object occurs exactly once in the array of components. Therefore, component objects MUST appear in the component array only once.

Note that the layout depends on the file size, which the client learns from the generic return parameters of LAYOUTGET, by doing GETATTR commands to the metadata server. The client uses the file size to decide if it should fill holes with zeros, or return a short read. Striping patterns can cause cases where component objects are shorter than other components because a hole happens to correspond to the last part of the component object.

3.1.1. pnfs_osd_objid4

An object is identified by a number, somewhat like an inode number. The object storage model has a two level scheme, where the objects within an object storage device are grouped into partitions.

```c
struct pnfs_osd_objid4 {
    deviceid4       device_id;
    uint64_t        partition_id;
    uint64_t        object_id;
};
```

The pnfs_osd_objid4 type is used to identify an object within a partition on a specified object storage device. "device_id" selects the object storage device from the set of available storage devices. The device is identified with the deviceid4 type, which is an index.
into addressing information about that device returned by the GETDEVICELIST and GETDEVICEINFO pnfs operations. Within an OSD, a partition is identified with a 64-bit number, "partition_id". Within a partition, an object is identified with a 64-bit number, "object_id". Creation and management of partitions is outside the scope of this standard, and is a facility storage provided by the object file system.

3.1.2. pnfs_osd_version4

enum pnfs_osd_version4 {
    PNFS_OSD_MISSING    = 0,
    PNFS_OSD_VERSION_1  = 1,
    PNFS_OSD_VERSION_2  = 2
};

The osd_version is used to indicate the OSD protocol version or whether an object is missing (i.e., unavailable). Some layout schemes encode redundant information and can compensate for missing components, but the data placement algorithm needs to know what parts are missing.

At this time the OSD standard is at version 1.0, and we anticipate a version 2.0 of the standard ([SNIA T10/1729-D [7]]). The second generation OSD protocol has additional proposed features to support more robust error recovery, snapshots, and byte-range capabilities. Therefore, the OSD version is explicitly called out in the information returned in the layout. (This information can also be deduced by looking inside the capability type at the format field, which is the first byte. The format value is 0x1 for an OSD v1 capability. However, it seems most robust to call out the version explicitly.)

3.1.3. pnfs_osd_object_cred4

enum pnfs_osd_cap_key_sec4 {
    PNFS_OSD_CAP_KEY_SEC_NONE = 0,
    PNFS_OSD_CAP_KEY_SEC_SSV  = 1,
};

struct pnfs_osd_object_cred4 {
    pnfs_osd_objid4         object_id;
    pnfs_osd_version4       osd_version;
    pnfs_osd_cap_key_sec4   cap_key_sec;
    opaque                  capability_key<>
    opaque                  capability<>
};
The pnfs_osd_object_cred4 structure is used to identify each component comprising the file. The object_id identifies the component object, the osd_version represents the osd protocol version, or whether that component is unavailable, and the capability and capability key, along with the systemid from the pnfs_osd_deviceaddr, provide the OSD security credentials needed to access that object. The cap_key_sec value denotes the method used to secure the capability_key (see Section 9.1 for more details).

To comply with the OSD security requirements the capability key SHOULD be transferred securely to prevent eavesdropping (see Section 9). Therefore, a client SHOULD either issue the LAYOUTGET operation via RPCSEC_GSS with the privacy service or to previously establish an SSV for the sessions via the NFSv4.1 SET_SSV operation. The pnfs_osd_cap_key_sec4 type is used to identify the method used by the server to secure the capability key.

- **PNFS_OSD_CAP_KEY_SEC_NONE** denotes that the capability_key is not encrypted in which case the client SHOULD issue the LAYOUTGET operation with RPCSEC_GSS with the privacy service or the NFSv4.1 transport should be secured by using methods that are external to NFSv4.1 like the use of IPSEC [8] for transporting the NFSV4.1 protocol.

- **PNFS_OSD_CAP_KEY_SEC_SSV** denotes that the capability_key contents are encrypted using the SSV GSS context and the capability key as inputs to the GSS_Wrap() function (see [3]) with the conf_req_flag set to TRUE. The client MUST use the secret SSV key as part of the client’s GSS context to decrypt the capability key using the value of the capability_key field as the input_message to the GSS_unwrap() function. Note that to prevent eavesdropping of the SSV key the client SHOULD issue SET_SSV via RPCSEC_GSS with the privacy service.

The actual method chosen depends on whether the client established a SSV key with the server and whether it issued the LAYOUTGET operation with the RPCSEC_GSS privacy method. Naturally, if the client did not establish a SSV key via SET_SSV the server MUST use the PNFS_OSD_CAP_KEY_SEC_NONE method. Otherwise, if the LAYOUTGET operation was not issued with the RPCSEC_GSS privacy method the server SHOULD secure the capability_key with the PNFS_OSD_CAP_KEY_SEC_SSV method. The server MAY use the PNFS_OSD_CAP_KEY_SEC_SSV method also when the LAYOUTGET operation was issued with the RPCSEC_GSS privacy method.
3.1.4. **pnfs_osd_raid_algorithm4**

```c
enum pnfs_osd_raid_algorithm4 {
    PNFS_OSD_RAID_0     = 1,
    PNFS_OSD_RAID_4     = 2,
    PNFS_OSD_RAID_5     = 3,
    PNFS_OSD_RAID_PQ    = 4     /* Reed-Solomon P+Q */
};
```

*pnfs_osd_raid_algorithm4* represents the data redundancy algorithm used to protect the file’s contents. See Section 3.3 for more details.

3.1.5. **pnfs_osd_data_map4**

```c
struct pnfs_osd_data_map4 {
    length4                     stripe_unit;
    uint32_t                    group_width;
    uint32_t                    group_depth;
    uint32_t                    mirror_cnt;
    pnfs_osd_raid_algorithm4    raid_algorithm;
};
```

The *pnfs_osd_data_map4* structure parameterizes the algorithm that maps a file’s contents over the component objects. Instead of limiting the system to simple striping scheme where loss of a single component object results in data loss, the map parameters support mirroring and more complicated schemes that protect against loss of a component object.

The *stripe_unit* is the number of bytes placed on one component before advancing to the next one in the list of components. The number of bytes in a full stripe is stripe_unit times the number of components. In some raid schemes, a stripe includes redundant information (i.e., parity) that lets the system recover from loss or damage to a component object.

The *group_width* and *group_depth* parameters allow a nested striping pattern. If there is no nesting, then *group_width* and *group_depth* MUST be zero. Otherwise, the *group_width* defines the width of a data stripe, and the *group_depth* defines how many stripes are accessed before advancing to the next group of components in the list of component objects for the file. The size of the components array MUST be a multiple of *group_width*.

The *mirror_cnt* is used to replicate a file by replicating its component objects. If there is no mirroring, then *mirror_cnt* MUST be 0. If *mirror_cnt* is greater than zero, then the size of the
component array MUST be a multiple of (mirror_cnt+1).

See Section 3.2 for more details.

3.2. Data Mapping Schemes

This section describes the different data mapping schemes in detail.

3.2.1. Simple Striping

The object layout always uses a "dense" layout as described in the pNFS document. This means that the second stripe unit of the file starts at offset 0 of the second component, rather than at offset stripe_unit bytes. After a full stripe has been written, the next stripe unit is appended to the first component object in the list without any holes in the component objects. The mapping from the logical offset within a file (L) to do the component object C and object-specific offset O is defined by the following equations:

\[
L = \text{logical offset into the file} \\
W = \text{total number of components} \\
S = W \times \text{stripe\_unit} \\
N = \frac{L}{S} \\
C = \frac{(L-(N*S))}{\text{stripe\_unit}} \\
O = (N\times\text{stripe\_unit})+(L\%\text{stripe\_unit})
\]

In these equations, S is the number of bytes in a full stripe, and N is the stripe number. C is an index into the array of components, so it selects a particular object storage device. Both N and C count from zero. O is the offset within the object that corresponds to the file offset. Note that this computation does not accommodate the same object appearing in the component array multiple times.

For example, consider an object striped over four devices, <D0 D1 D2 D3>. The stripe_unit is 4096 bytes. The stripe width S is thus 4 * 4096 = 16384.
Offset 0:
N = 0 / 16384 = 0
C = 0-0/4096 = 0 (D0)
O = 0*4096 + (0%4096) = 0

Offset 4096:
N = 4096 / 16384 = 0
C = (4096-(0*16384)) / 4096 = 1 (D1)
O = (0*4096)+(4096%4096) = 0

Offset 9000:
N = 9000 / 16384 = 0
C = (9000-(0*16384)) / 4096 = 2 (D2)
O = (0*4096)+(9000%4096) = 808

Offset 132000:
N = 132000 / 16384 = 8
C = (132000-(8*16384)) / 4096 = 0
O = (8*4096) + (132000%4096) = 33696

3.2.2. Nested Striping

The group_width and group_depth parameters allow a nested striping pattern. If there is no nesting, then group_width and group_depth MUST be zero. Otherwise, the group_width defines the width of a data stripe, and the group_depth defines how many stripes are written before advancing to the next group of components in the list of component objects for the file. The size of the components array MUST be a multiple of group_width. The math used to map from a file offset to a component object and offset within that object is shown below. The computations map from the logical offset L to the component index C and offset relative O within that component object.

L = logical offset into the file
W = total number of components
S = stripe_unit * group_depth * W
T = stripe_unit * group_depth * group_width
U = stripe_unit * group_width
M = L / S
G = (L - (M * S)) / T
H = (L - (M * S)) % T
N = H / U
C = (H - (N * U)) / stripe_unit + G * group_width
O = L % stripe_unit + N * stripe_unit + M * group_depth * stripe_unit

In these equations, S is the number of bytes striped across all component objects before the pattern repeats. T is the number of bytes striped within a group of component objects before advancing to
the next group. \( U \) is the number of bytes in a stripe within a group. \( M \) is the "major" (i.e., across all components) stripe number, and \( N \) is the "minor" (i.e., across the group) stripe number. \( G \) counts the groups from the beginning of the major stripe, and \( H \) is the byte offset within the group.

For example, consider an object striped over 100 devices with a group_width of 10, a group_depth of 50, and a stripe_unit of 1 MB. In this scheme, 500 MB are written to the first 10 components, and 5000 MB is written before the pattern wraps back around to the first component in the array.

Offset 0:
\[
\begin{align*}
W &= 100 \\
S &= 1 \text{ MB} \times 50 \times 100 = 5000 \text{ MB} \\
T &= 1 \text{ MB} \times 50 \times 10 = 500 \text{ MB} \\
U &= 1 \text{ MB} \times 10 = 10 \text{ MB} \\
M &= 0 / 5000 \text{ MB} = 0 \\
G &= (0 - (0 \times 5000 \text{ MB})) / 500 \text{ MB} = 0 \\
H &= (0 - (0 \times 5000 \text{ MB})) \% 500 \text{ MB} = 0 \\
N &= 0 / 10 \text{ MB} = 0 \\
C &= (0 - (0 \times 10 \text{ MB})) / 1 \text{ MB} + 0 \times 10 = 0 \\
O &= 0 \% 1 \text{ MB} + 0 \times 1 \text{ MB} + 0 \times 50 \times 1 \text{ MB} = 0
\end{align*}
\]

Offset 27 MB:
\[
\begin{align*}
M &= 27 \text{ MB} / 5000 \text{ MB} = 0 \\
G &= (27 \text{ MB} - (0 \times 5000 \text{ MB})) / 500 \text{ MB} = 0 \\
H &= (27 \text{ MB} - (0 \times 5000 \text{ MB})) \% 500 \text{ MB} = 27 \text{ MB} \\
N &= 27 \text{ MB} / 10 \text{ MB} = 2 \\
C &= (27 \text{ MB} - (2 \times 10 \text{ MB})) / 1 \text{ MB} + 0 \times 10 = 7 \\
O &= 27 \text{ MB} \% 1 \text{ MB} + 2 \times 1 \text{ MB} + 0 \times 50 \times 1 \text{ MB} = 2 \text{ MB}
\end{align*}
\]

Offset 7232 MB:
\[
\begin{align*}
M &= 7232 \text{ MB} / 5000 \text{ MB} = 1 \\
G &= (7232 \text{ MB} - (1 \times 5000 \text{ MB})) / 500 \text{ MB} = 4 \\
H &= (7232 \text{ MB} - (1 \times 5000 \text{ MB})) \% 500 \text{ MB} = 232 \text{ MB} \\
N &= 232 \text{ MB} / 10 \text{ MB} = 23 \\
C &= (232 \text{ MB} - (23 \times 10 \text{ MB})) / 1 \text{ MB} + 4 \times 10 = 42 \\
O &= 7232 \text{ MB} \% 1 \text{ MB} + 23 \times 1 \text{ MB} + 1 \times 50 \times 1 \text{ MB} = 73 \text{ MB}
\end{align*}
\]

3.2.3. Mirroring

The mirror_cnt is used to replicate a file by replicating its component objects. If there is no mirroring, then mirror_cnt MUST be 0. If mirror_cnt is greater than zero, then the size of the component array MUST be a multiple of (mirror_cnt+1). Thus, for a classic mirror on two objects, mirror_cnt is one. If group_width is also non-zero, then the size MUST be a multiple of group_width *
Replicas are adjacent in the components array, and the value C produced by the above equations is not a direct index into the components array. Instead, the following equations determine the replica component index RC\textsubscript{i}, where i ranges from 0 to mirror\_cnt.

\[
C = \text{component index for striping or two-level striping}
\]

\[
i \text{ranges from 0 to mirror\_cnt, inclusive}
\]

\[
RC_i = C \times (\text{mirror\_cnt}+1) + i
\]

### 3.3. RAID Algorithms

`pnfs\_osd\_raid\_algorithm4` determines the algorithm and placement of redundant data. This section defines the different RAID algorithms.

#### 3.3.1. PNFS\_OSD\_RAID\_0

PNFS\_OSD\_RAID\_0 means there is no parity data, so all bytes in the component objects are data bytes located by the above equations for C and O. If a component object is unavailable, the pNFS client can choose to return NULLs for the missing data, or it can retry the READ against the pNFS server, or it can return an EIO error.

#### 3.3.2. PNFS\_OSD\_RAID\_4

PNFS\_OSD\_RAID\_4 means that the last component object, or the last in each group if group\_width is > zero, contains parity information computed over the rest of the stripe with an XOR operation. If a component object is unavailable, the client can read the rest of the stripe units in the damaged stripe and recompute the missing stripe unit by XORing the other stripe units in the stripe. Or the client can replay the READ against the pNFS server which will presumably perform the reconstructed read on the client’s behalf.

When parity is present in the file, then there is an additional computation to map from the file offset L to the offset that accounts for embedded parity, L’. First compute L’, and then use L’ in the above equations for C and O.

\[
L = \text{file offset, not accounting for parity}
\]

\[
P = \text{number of parity devices in each stripe}
\]

\[
W = \text{group\_width, if not zero, else size of component array}
\]

\[
N = \frac{L}{(W-P \times \text{stripe\_unit})}
\]

\[
L' = N \times (W \times \text{stripe\_unit}) + (L \% (W-P \times \text{stripe\_unit}))
\]
3.3.3. PNFS_OSD_RAID_5

PNFS_OSD_RAID_5 means that the position of the parity data is rotated on each stripe. In the first stripe, the last component holds the parity. In the second stripe, the next-to-last component holds the parity, and so on. In this scheme, all stripe units are rotated so that I/O is evenly spread across objects as the file is read sequentially. The rotated parity layout is illustrated here, with numbers indicating the stripe unit.

```
0 1 2 P
4 5 P 3
8 P 6 7
P 9 a b
```

To compute the component object $C$, first compute the offset that accounts for parity $L'$ and use that to compute $C$. Then rotate $C$ to get $C'$. Finally, increase $C'$ by one if the parity information comes at or before $C'$ within that stripe. The following equations illustrate this by computing $I$, which is the index of the component that contains parity for a given stripe.

$L = \text{file offset, not accounting for parity}$
$W = \text{group\_width, if not zero, else size of component array}$
$N = \frac{L}{(W-1 \times \text{stripe\_unit})}$
(Compute $L'$ as describe above)
(Compute $C$ based on $L'$ as described above)
$C' = (C - (N \times W)) \mod W$
$I = W - (N \times W) - 1$
if $(C' <= I)$ {
  $C'++$
}

3.3.4. PNFS_OSD_RAID_PQ

PNFS_OSD_RAID_PQ is a double-parity scheme that uses the Reed-Solomon P+Q encoding scheme. In this layout, the last two component objects hold the $P$ and $Q$ data, respectively. $P$ is parity computed with XOR, and $Q$ is a more complex equation that is not described here. The equations given above for embedded parity can be used to map a file offset to the correct component object by setting the number of parity components to 2 instead of 1 for RAID4 or RAID5. Clients may simply choose to read data through the metadata server if two components are missing or damaged.

Issue: This scheme also has a RAID_4 like layout where the ECC blocks are stored on the same components on every stripe and a rotated, RAID-5 like layout where the stripe units are rotated. Should we
make the following properties orthogonal: RAID_4 or RAID_5 (i.e.,
non-rotated or rotated), and then have the number of parity
components and the associated algorithm be the orthogonal parameter?

3.3.5. RAID Usage and implementation notes

RAID layouts with redundant data in their stripes require additional
serialization of updates to ensure correct operation. Otherwise, if
two clients simultaneously write to the same logical range of an
object, the result could include different data in the same ranges of
mirrored tuples, or corrupt parity information. It is the
responsibility of the metadata server to enforce serialization
requirements such as this. For example, the metadata server may do
so by not granting overlapping write layouts within mirrored objects.

4. Object-Based Layout Update

layoutupdate4 is used in the LAYOUTCOMMIT operation to convey updates
to the layout and additional information to the metadata server. It
is defined in the NFSv4.1 draft [6] as follows:

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

The layoutupdate4 type is an opaque value at the generic pNFS client
level. If the lou_type layout type is LAYOUT4_OSD2_OBJECTS, then the
lou_body opaque value is defined by the pnfs_osd_layoutupdate4 type.

4.1. pnfs_osd_layoutupdate4

    struct pnfs_osd_layoutupdate4 {
        pnfs_osd_deltaspaceused4    delta_space_used;
        pnfs_osd_ioerr4             ioerr<>;
    };

Object-Based pNFS clients are not allowed to modify the layout.
"delta_space_used" is used to convey capacity usage information back
to the metadata server and, in case OSD I/O operations failed,
"ioerr" is used to report these errors to the metadata server.
4.1.1. pnfs_osd_deltaspaceused4

union pnfs_osd_deltaspaceused4 switch (bool valid) {
    case TRUE:
        int64_t delta; /* Bytes consumed by write activity */
    case FALSE:
        void;
};

pnfs_osd_deltaspaceused4 is used to convey space utilization information at the time of LAYOUTCOMMIT. For the file system to properly maintain capacity used information, it needs to track how much capacity was consumed by WRITE operations performed by the client. In this protocol, the OSD returns the capacity consumed by a write, which can be different than the number of bytes written because of internal overhead like block-based allocation and indirect blocks, and the client reflects this back to the pNFS server so it can accurately track quota. The pNFS server can choose to trust this information coming from the clients and therefore avoid querying the OSDs at the time of LAYOUTCOMMIT. If the client is unable to obtain this information from the OSD, it simply returns invalid delta_space_used.

4.1.2. pnfs_osd_errno4

enum pnfs_osd_errno4 {
    PNFS_OSD_ERR_EIO            = 1,
    PNFS_OSD_ERR_NOT_FOUND      = 2,
    PNFS_OSD_ERR_NO_SPACE       = 3,
    PNFS_OSD_ERR_BAD_CRED       = 4,
    PNFS_OSD_ERR_NO_ACCESS      = 5,
    PNFS_OSD_ERR_UNREACHABLE    = 6,
    PNFS_OSD_ERR_RESOURCE       = 7
};

pnfs_osd_errno4 is used to represent error types when read/write errors are reported to the metadata server. The error codes serve as hints to the metadata server that may help it in diagnosing the exact reason for the error and in repairing it.

- PNFS_OSD_ERR_EIO indicates the operation failed because the Object Storage Device experienced a failure trying to access the object. The most common source of these errors is media errors, but other internal errors might cause this. In this case, the metadata server should go examine the broken object more closely, hence it should be used as the default error code.
o PNFS_OSD_ERR_NOT_FOUND indicates the object ID specifies an object that does not exist on the Object Storage Device.

o PNFS_OSD_ERR_NO_SPACE indicates the operation failed because the Object Storage Device ran out of free capacity during the operation.

o PNFS_OSD_ERR_BAD_CRED indicates the security parameters are not valid. The primary cause of this is that the capability has expired, or the access policy tag (a.k.a, capability version number) has been changed to revoke capabilities. The client will need to return the layout and get a new one with fresh capabilities.

o PNFS_OSD_ERR_NO_ACCESS indicates the capability does not allow the requested operation. This should not occur in normal operation because the metadata server should give out correct capabilities, or none at all.

o PNFS_OSD_ERR_UNREACHABLE indicates the client did not complete the I/O operation at the Object Storage Device due to a communication failure. Whether the I/O operation was executed by the OSD or not is undetermined.

o PNFS_OSD_ERR_RESOURCE indicates the client did not issue the I/O operation due to a local problem on the initiator (i.e. client) side, e.g., when running out of memory. The client MUST guarantee that the OSD command was never dispatched to the OSD.

4.1.3. pnfs_osd_ioerr4

```c
struct pnfs_osd_ioerr4 {
    pnfs_osd_objid4     component;
    length4             comp_offset;
    length4             comp_length;
    bool                iswrite;
    pnfs_osd_errno4     errno;
};
```

The pnfs_osd_ioerr4 structure is used to return error indications for objects that generated errors during data transfers. These are hints to the metadata server that there are problems with that object. For each error, "component", "comp_offset", and "comp_length" represent the object and byte range within the component object in which the error occurred. "iswrite" is set to "true" if the failed OSD operation was data modifying, and "errno" represents the type of error.
5. Object-Based Creation Layout Hint

The layouthint4 type is defined in the NFSv4.1 draft [6] as follows:

```c
struct layouthint4 {
    layouttype4 loh_type;
    opaque loh_body<>;
};
```

The layouthint4 structure is used by the client to pass in a hint about the type of layout it would like created for a particular file. If the loh_type layout type is LAYOUT4_OSD2_OBJECTS, then the loh_body opaque value is defined by the pnfs_osd_layouthint4 type.

5.1. pnfs_osd_layouthint4

```c
union num_comps_hint4 switch (bool valid) {
    case TRUE: uint32_t num_comps;
    case FALSE: void;
};

union stripe_unit_hint4 switch (bool valid) {
    case TRUE: length4 stripe_unit;
    case FALSE: void;
};

union group_width_hint4 switch (bool valid) {
    case TRUE: uint32_t group_width;
    case FALSE: void;
};

union group_depth_hint4 switch (bool valid) {
    case TRUE: uint32_t group_depth;
    case FALSE: void;
};

union mirror_cnt_hint4 switch (bool valid) {
    case TRUE: uint32_t mirror_cnt;
    case FALSE: void;
};
```
union raid_algorithm_hint4 switch (bool valid) {
    case TRUE:
        pnfs_osd_raid_algorithm4 raid_algorithm;
    case FALSE:
        void;
};

struct pnfs_osd_layouthint4 {
    num_comps_hint4 num_comps_hint;
    stripe_unit_hint4 stripe_unit_hint;
    group_width_hint4 group_width_hint;
    group_depth_hint4 group_depth_hint;
    mirror_cnt_hint4 mirror_cnt_hint;
    raid_algorithm_hint4 raid_algorithm_hint;
};

This type conveys hints for the desired data map. All parameters are optional so the client can give values for only the parameters it cares about, e.g. it can provide a hint for the desired number of mirrored components, regardless of the the raid algorithm selected for the file. The server should make an attempt to honor the hints but it can ignore any or all of them at its own discretion and without failing the respective create operation.

The num_comps hint can be used to limit the total number of component objects comprising the file. All other hints correspond directly to the different fields of pnfs_osd_data_map4.

6. Layout Segments

The pnfs layout operations operate on logical byte ranges. There is no requirement in the protocol for any relationship between byte ranges used in LAYOUTGET to acquire layouts and byte ranges used in CB_LAYOUTRECALL, LAYOUTCOMMIT, or LAYOUTRETURN. However, using OSD capabilities poses limitations on these operations since the capabilities associated with layout segments cannot be merged or split. The following guidelines should be followed for proper operation of object-based layouts.

6.1. CB_LAYOUTRECALL and LAYOUTRETURN

In general, the object-based layout driver should keep track of each layout segment it got, keeping record of the segment’s iomode, offset, and length. The server should allow the client to get
multiple overlapping layout segments but is free to recall the layout to prevent overlap.

In response to CB_LAYOUTRECALL, the client should return all layout segments matching the given iomode and overlapping with the recalled range. When returning the layouts for this byte range with LAYOUTRETURN the client MUST NOT return a sub-range of a layout segment it has; each LAYOUTRETURN sent MUST completely cover at least one outstanding layout segment.

The server, in turn, should release any segment that exactly matches the clientid, iomode, and byte range given in LAYOUTRETURN. If no exact match is found then the server should release all layout segments matching the clientid and iomode and that are fully contained in the returned byte range. If none are found and the byte range is a subset of an outstanding layout segment with for the same clientid and iomode, then the client can be considered malfunctioning and the server SHOULD recall all layouts from this client to reset its state. If this behavior repeats the server SHOULD deny all LAYOUTGETs from this client.

6.2. LAYOUTCOMMIT

LAYOUTCOMMIT is only used by object-based pNFS to convey modified attributes hints and/or to report I/O errors to the MDS. Therefore, the offset and length in LAYOUTCOMMIT4args are reserved for future use and should be set to 0. However, component byte ranges in the optional pnfs_osd_ioerr4 structure are used for recovering the object and MUST be set by the client to cover all failed I/O operations to the component.

7. Recalling Layouts

The object-based metadata server should recall outstanding layouts in the following cases:

- When the file’s security policy changes, i.e. ACLs or permission mode bits are set.
- When the file’s aggregation map changes, rendering outstanding layouts invalid.
- When there are sharing conflicts. For example, the server will issue stripe aligned layout segments for RAID-5 objects. To prevent corruption of the file’s parity, Multiple clients must not hold valid write layouts for the same stripes. An outstanding RW layout should be recalled when a conflicting LAYOUTGET is received.
from a different client for LAYOUTIOMODE4_RW and for a byte-range overlapping with the outstanding layout segment.

7.1. CB_RECALL_ANY

The metadata server can use the CB_RECALL_ANY callback operation to notify the client to return some or all of its layouts. The NFSv4.1 draft \[6\] defines the following types:

```c
const RCA4_TYPE_MASK_OBJ_LAYOUT_MIN = 8;
const RCA4_TYPE_MASK_OBJ_LAYOUT_MAX = 11;

struct CB_RECALL_ANY4args {
    uint32_t craa_objects_to_keep;
    bitmap4 craa_type_mask;
};
```

Typically, CB_RECALL_ANY will be used to recall client state when the server needs to reclaim resources. The `craa_type_mask` bitmap specifies the type of resources that are recalled and the `craa_objects_to_keep` value specifies how many of the recalled objects the client is allowed to keep. The object-based layout type mask flags are defined as follows. They represent the iomode of the recalled layouts. In response, the client SHOULD return layouts of the recalled iomode that it needs the least, keeping at most `craa_objects_to_keep` object-based layouts.

```c
const PNFS_OSD_RCA4_TYPE_MASK_READ  = RCA4_TYPE_MASK_OBJ_LAYOUT_MIN;
const PNFS_OSD_RCA4_TYPE_MASK_RW    = RCA4_TYPE_MASK_OBJ_LAYOUT_MIN+1;
const PNFS_OSD_RCA4_TYPE_MASK_ANY   = RCA4_TYPE_MASK_OBJ_LAYOUT_MIN+2;
```

The `PNFS_OSD_RCA4_TYPE_MASK_READ` flag notifies the client to return layouts of iomode LAYOUTIOMODE4_READ. Similarly, the `PNFS_OSD_RCA4_TYPE_MASK_RW` flag notifies the client to return layouts of iomode LAYOUTIOMODE4_RW. The `PNFS_OSD_RCA4_TYPE_MASK_ANY` flag notifies the client to return layouts of either iomode.

8. Client Fencing

In cases where clients are uncommunicative and their lease has expired or when clients fail to return recalled layouts in a timely manner the server MAY revoke client layouts and/or device address mappings and reassign these resources to other clients. To avoid data corruption, the metadata server MUST fence off the revoked clients from the respective objects as described in Section 9.4.
9. Security Considerations

The pNFS extension partitions the NFSv4 file system protocol into two parts, the control path and the data path (storage protocol). The control path contains all the new operations described by this extension; all existing NFSv4 security mechanisms and features apply to the control path. The combination of components in a pNFS system is required to preserve the security properties of NFSv4 with respect to an entity accessing data via a client, including security countermeasures to defend against threats that NFSv4 provides defenses for in environments where these threats are considered significant.

The metadata server enforces the file access-control policy at LAYOUTGET time. The client should use suitable authorization credentials for getting the layout for the requested iomode (READ or RW) and the server verifies the permissions and ACL for these credentials, possibly returning NFS4ERR_ACCESS if the client is not allowed the requested iomode. If the LAYOUTGET operation succeeds the client receives, as part of the layout, a set of object capabilities allowing it I/O access to the specified objects corresponding to the requested iomode. When the client acts on I/O operations on behalf of its local users it MUST authenticate and authorize the user by issuing respective OPEN and ACCESS calls to the metadata server, similarly to having NFSv4 data delegations. If access is allowed the client uses the corresponding (READ or RW) capabilities to perform the I/O operations at the object-storage devices. When the metadata server receives a request to change file’s permissions or ACL it SHOULD recall all layouts for that file and it MUST change the capability version attribute on all objects comprising the file to implicitly invalidate any outstanding capabilities before committing to the new permissions and ACL. Doing this will ensure that clients re-authorize their layouts according to the modified permissions and ACL by requesting new layouts. Recalling the layouts in this case is courtesy of the server intended to prevent clients from getting an error on I/Os done after the capability version changed.

The object storage protocol MUST implement the security aspects described in version 1 of the T10 OSD protocol definition [2]. The standard defines four security methods: NOSEC, CAPKEY, CMDRSP, and ALLDATA. To provide minimum level of security allowing verification and enforcement of the server access control policy using the layout security credentials, the NOSEC security method MUST NOT be used for I/O operation. It MAY only be used to get the System ID attribute when the metadata server provided only the OSD name with the device address. The remainder of this section gives an overview of the security mechanism described in that standard. The goal is to give
the reader a basic understanding of the object security model. Any discrepancies between this text and the actual standard are obviously to be resolved in favor of the OSD standard.

9.1. OSD Security Data Types

There are three main data types associated with object security: a capability, a credential, and security parameters. The capability is a set of fields that specifies an object and what operations can be performed on it. A credential is a signed capability. Only a security manager that knows the secret device keys can correctly sign a capability to form a valid credential. In pNFS, the file server acts as the security manager and returns signed capabilities (i.e., credentials) to the pNFS client. The security parameters are values computed by the issuer of OSD commands (i.e., the client) that prove they hold valid credentials. The client uses the credential as a signing key to sign the requests it makes to OSD, and puts the resulting signatures into the security_parameters field of the OSD command. The object storage device uses the secret keys it shares with the security manager to validate the signature values in the security parameters.

The security types are opaque to the generic layers of the pNFS client. The credential contents are defined as opaque within the pnfs_osd_object_cred4 type. Instead of repeating the definitions here, the reader is referred to section 4.9.2.2 of the OSD standard.

9.2. The OSD Security Protocol

The object storage protocol relies on a cryptographically secure capability to control accesses at the object storage devices. Capabilities are generated by the metadata server, returned to the client, and used by the client as described below to authenticate their requests to the Object Storage Device (OSD). Capabilities therefore achieve the required access and open mode checking. They allow the file server to define and check a policy (e.g., open mode) and the OSD to enforce that policy without knowing the details (e.g., user IDs and ACLs).

Since capabilities are tied to layouts, and since they are used to enforce access control, when the file ACL or mode changes the outstanding capabilities MUST be revoked to enforce the new access permissions. The server SHOULD recall layouts to allow clients to gracefully return their capabilities before the access permissions change.

Each capability is specific to a particular object, an operation on that object, a byte range w/in the object (in OSDv2), and has an
explicit expiration time. The capabilities are signed with a secret key that is shared by the object storage devices (OSD) and the metadata managers. Clients do not have device keys so they are unable to forge the signatures in the security parameters. The combination of a capability, the OSD system id, and a signature is called a "credential" in the OSD specification.

The details of the security and privacy model for Object Storage are defined in the T10 OSD standard. The following sketch of the algorithm should help the reader understand the basic model.

LAYOUTGET returns a CapKey and a Cap which, together with the OSD SystemID, are also called a credential. It is a capability and a signature over that capability and the SystemID. The OSD Standard refers to the CapKey as the "Credential integrity check value" and to the ReqMAC as the "Request integrity check value".

\[ \text{CapKey} = \text{MAC}\langle\text{SecretKey}\rangle(\text{Cap}, \text{SystemID}) \]
\[ \text{Credential} = \{\text{Cap}, \text{SystemID}, \text{CapKey}\} \]

The client uses CapKey to sign all the requests it issues for that object using the respective Cap. In other words, the Cap appears in the request to the storage device, and that request is signed with the CapKey as follows:

\[ \text{ReqMAC} = \text{MAC}\langle\text{CapKey}\rangle(\text{Req}, \text{ReqNonce}) \]
\[ \text{Request} = \{\text{Cap}, \text{Req}, \text{ReqNonce}, \text{ReqMAC}\} \]

The following is sent to the OSD: \{Cap, Req, ReqNonce, ReqMAC\}. The OSD uses the SecretKey it shares with the metadata server to compare the ReqMAC the client sent with a locally computed value:

\[ \text{LocalCapKey} = \text{MAC}\langle\text{SecretKey}\rangle(\text{Cap}, \text{SystemID}) \]
\[ \text{LocalReqMAC} = \text{MAC}\langle\text{LocalCapKey}\rangle(\text{Req}, \text{ReqNonce}) \]

and if they match the OSD assumes that the capabilities came from an authentic metadata server and allows access to the object, as allowed by the Cap.

9.3. Protocol Privacy Requirements

Note that if the server LAYOUTGET reply, holding CapKey and Cap, is snooped by another client, it can be used to generate valid OSD requests (within the Cap access restrictions).

To provide the required privacy requirements for the capability key returned by LAYOUTGET, the GSS-API can be used, e.g. by using the RPCSEC_GSS privacy method to send the LAYOUTGET operation or by using
the SSV key to encrypt the capability_key using the GSS_Wrap() function. Two general ways to provide privacy in the absence of GSS-API that are independent of NFSv4 are either an isolated network such as a VLAN or a secure channel provided by IPsec [8].

9.4. Revoking Capabilities

At any time, the metadata server may invalidate all outstanding capabilities on an object by changing its POLICY ACCESS TAG attribute. The value of the POLICY ACCESS TAG is part of a capability, and it must match the state of the object attribute. If they do not match, the OSD rejects accesses to the object with the sense key set to ILLEGAL REQUEST and an additional sense code set to INVALID FIELD IN CDB. When a client attempts to use a capability and is rejected this way, it should issue a LAYOUTCOMMIT for the object and specify PNFS_OSD_BAD_CRED in the ioerr parameter. The client may elect to issue a compound LAYOUTRETURN/LAYOUTGET (or LAYOUTCOMMIT/ LAYOUTRETURN/LAYOUTGET) to attempt to fetch a refreshed set of capabilities.

The metadata server may elect to change the access policy tag on an object at any time, for any reason (with the understanding that there is likely an associated performance penalty, especially if there are outstanding layouts for this object). The metadata server MUST revoke outstanding capabilities when any one of the following occurs:

- the permissions on the object change,
- a conflicting mandatory byte-range lock is granted, or
- a layout is revoked and reassigned to another client

A pNFS client will typically hold one layout for each byte range for either READ or READ/WRITE. The client’s credentials are checked by the metadata server at LAYOUTGET time and it is the client’s responsibility to enforce access control among multiple users accessing the same file. It is neither required nor expected that the pNFS client will obtain a separate layout for each user accessing a shared object. The client SHOULD use OPEN and ACCESS calls to check user permissions when performing I/O so that the server’s access control policies are correctly enforced. The result of the ACCESS operation may be cached while the client holds a valid layout as the server is expected to recall layouts when the file’s access permissions or ACL change.
10. IANA Considerations

As described in the NFSv4.1 draft [6], new layout type numbers will be requested from IANA. This document defines the protocol associated with the existing layout type number, LAYOUT4_OSD2_OBJECTS, and it requires no further actions for IANA.

11. References

11.1. Normative References


11.2. Informative References


Appendix A. Acknowledgments

Todd Pisek was a co-editor of the initial drafts for this document.
Authors’ Addresses

Benny Halevy  
Panasas, Inc.  
1501 Reedsdale St. Suite 400  
Pittsburgh, PA  15233  
USA  
Phone: +1-412-323-3500  
Email: bhalevy@panasas.com  
URI:  http://www.panasas.com/

Brent Welch  
Panasas, Inc.  
6520 Kaiser Drive  
Fremont, CA  95444  
USA  
Phone: +1-650-608-7770  
Email: welch@panasas.com  
URI:  http://www.panasas.com/

Jim Zelenka  
Panasas, Inc.  
1501 Reedsdale St. Suite 400  
Pittsburgh, PA  15233  
USA  
Phone: +1-412-323-3500  
Email: jimz@panasas.com  
URI:  http://www.panasas.com/
Full Copyright Statement

Copyright (C) The IETF Trust (2007).

This document is subject to the rights, licenses and restrictions contained in BCP 78, and except as set forth therein, the authors retain all their rights.

This document and the information contained herein are provided on an "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE REPRESENTS OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY, THE IETF TRUST AND THE INTERNET ENGINEERING TASK FORCE DISCLAIM ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Intellectual Property

The IETF takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights. Information on the procedures with respect to rights in RFC documents can be found in BCP 78 and BCP 79.

Copies of IPR disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at http://www.ietf.org/ipr.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement this standard. Please address the information to the IETF at ietf-ipr@ietf.org.

Acknowledgment

Funding for the RFC Editor function is provided by the IETF Administrative Support Activity (IASA).