FILE STRIPING

The text below describes how files from Ceph file system clients are stored across objects stored in RADOS.

CEPH_FILE_LAYOUT

Ceph distributes (stripes) the data for a given file across a number of underlying objects. The way file data is mapped to those objects is defined by the ceph_file_layout structure. The data distribution is a modified RAID 0, where data is striped across a set of objects up to a (per-file) fixed size, at which point another set of objects holds the file's data. The second set also holds no more than the fixed amount of data, and then another set is used, and so on.

Defining some terminology will go a long way toward explaining the way file data is laid out across Ceph objects.

- file

  A collection of contiguous data, named from the perspective of the Ceph client (i.e., a file on a Linux system using Ceph storage). The data for a file is divided into fixed-size "stripe units," which are stored in ceph "objects."

- stripe unit

  The size (in bytes) of a block of data used in the RAID 0 distribution of a file. All stripe units for a file have equal size. The last stripe unit is typically incomplete–i.e. it represents the data at the end of the file as well as unused "space" beyond it up to the end of the fixed stripe unit size.

- stripe count

  The number of consecutive stripe units that constitute a RAID 0 "stripe" of file data.

- stripe

  A contiguous range of file data, RAID 0 striped across "stripe count" objects in fixed-size "stripe unit" blocks.

- object

  A collection of data maintained by Ceph storage. Objects are used to hold portions of Ceph client files.

- object set

  A set of objects that together represent a contiguous portion of a file.

Three fields in the ceph_file_layout structure define this mapping:

```
u32 fl_stripe_unit;
u32 fl_stripe_count;
u32 fl_object_size;
```

(They are actually maintained in their on-disk format, __le32.)

The role of the first two fields should be clear from the definitions above.

The third field is the maximum size (in bytes) of an object used to back file data. The object size is a multiple of the stripe unit.

A file's data is blocked into stripe units, and consecutive stripe units are stored on objects in an object set. The number of objects in a set is the same as the stripe count. No object storing file data will exceed the file's designated object size, so after some fixed number of complete stripes, a new object set is used to store subsequent file data.

Note that by default, Ceph uses a simple striping strategy in which object_size equals stripe_unit and stripe_count is 1. This simply puts one stripe_unit in each object.

Here's a more complex example:

```
file size = 1 trillion = 1000000000000 bytes

fl_stripe_unit = 64KB = 65536 bytes
fl_stripe_count = 5 stripe units per stripe
fl_object_size = 64GB = 68719476736 bytes
```

This means:

```
file stripe size = 64KB * 5 = 320KB = 327680 bytes
each object holds 64GB / 64KB = 1048576 stripe units
file object set size = 64GB * 5 = 320GB = 343597383680 bytes
    (also 1048576 stripe units * 327680 bytes per stripe unit)
```

So the file’s 1 trillion bytes can be divided into complete object sets, then complete stripes, then complete stripe units, and finally a single incomplete stripe unit:

```
- 1 trillion bytes / 320GB per object set = 2 complete object sets
    (with 312805232640 bytes remaining)
- 312805232640 bytes / 320KB per stripe = 954605 complete stripes
    (with 266240 bytes remaining)
- 266240 bytes / 64KB per stripe unit = 4 complete stripe units
    (with 4096 bytes remaining)
- and the final incomplete stripe unit holds those 4096 bytes.
```

The ASCII art below attempts to capture this:

```
   _________   _________   _________   _________   _________
  /object  0\ /object  1\ /object  2\ /object  3\ /object  4\
  +=========+ +=========+ +=========+ +=========+ +=========+
  |  stripe | |  stripe | |  stripe | |  stripe | |  stripe |
o |   unit  | |   unit  | |   unit  | |   unit  | |   unit  | stripe 0
b |     0   | |     1   | |     2   | |     3   | |     4   |
j |---------| |---------| |---------| |---------| |---------|
e |  stripe | |  stripe | |  stripe | |  stripe | |  stripe |
c |   unit  | |   unit  | |   unit  | |   unit  | |   unit  | stripe 1
t |     5   | |     6   | |     7   | |     8   | |     9   |
  |---------| |---------| |---------| |---------| |---------|
s |     .   | |     .   | |     .   | |     .   | |     .   |
e       .           .           .           .           .
t |     .   | |     .   | |     .   | |     .   | |     .   |
  |---------| |---------| |---------| |---------| |---------|
0 |  stripe | |  stripe | |  stripe | |  stripe | |  stripe | stripe
  |   unit  | |   unit  | |   unit  | |   unit  | |   unit  | 1048575
  | 5242875 | | 5242876 | | 5242877 | | 5242878 | | 5242879 |
  \=========/ \=========/ \=========/ \=========/ \=========/

   _________   _________   _________   _________   _________
  /object  5\ /object  6\ /object  7\ /object  8\ /object  9\
  +=========+ +=========+ +=========+ +=========+ +=========+
  |  stripe | |  stripe | |  stripe | |  stripe | |  stripe | stripe
o |   unit  | |   unit  | |   unit  | |   unit  | |   unit  | 1048576
b | 5242880 | | 5242881 | | 5242882 | | 5242883 | | 5242884 |
j |---------| |---------| |---------| |---------| |---------|
e |  stripe | |  stripe | |  stripe | |  stripe | |  stripe | stripe
c |   unit  | |   unit  | |   unit  | |   unit  | |   unit  | 1048577
t | 5242885 | | 5242886 | | 5242887 | | 5242888 | | 5242889 |
  |---------| |---------| |---------| |---------| |---------|
s |     .   | |     .   | |     .   | |     .   | |     .   |
e       .           .           .           .           .
t |     .   | |     .   | |     .   | |     .   | |     .   |
  |---------| |---------| |---------| |---------| |---------|
1 |  stripe | |  stripe | |  stripe | |  stripe | |  stripe | stripe
  |   unit  | |   unit  | |   unit  | |   unit  | |   unit  | 2097151
  | 10485755| | 10485756| | 10485757| | 10485758| | 10485759|
  \=========/ \=========/ \=========/ \=========/ \=========/

   _________   _________   _________   _________   _________
  /object 10\ /object 11\ /object 12\ /object 13\ /object 14\
  +=========+ +=========+ +=========+ +=========+ +=========+
  |  stripe | |  stripe | |  stripe | |  stripe | |  stripe | stripe
o |   unit  | |   unit  | |   unit  | |   unit  | |   unit  | 2097152
b | 10485760| | 10485761| | 10485762| | 10485763| | 10485764|
j |---------| |---------| |---------| |---------| |---------|
e |  stripe | |  stripe | |  stripe | |  stripe | |  stripe | stripe
c |   unit  | |   unit  | |   unit  | |   unit  | |   unit  | 2097153
t | 10485765| | 10485766| | 10485767| | 10485768| | 10485769|
  |---------| |---------| |---------| |---------| |---------|
s |     .   | |     .   | |     .   | |     .   | |     .   |
e       .           .           .           .           .
```

```
t |     .   | |     .   | |     .   | |     .   | |     .   |
  |---------| |---------| |---------| |---------| |---------|
2 |  stripe | |  stripe | |  stripe | |  stripe | |  stripe | stripe
  |   unit  | |   unit  | |   unit  | |   unit  | |   unit  | 3051756
  | 15258780| | 15258781| | 15258782| | 15258783| | 15258784|
  |---------| |---------| |---------| |---------| |---------|
  |  stripe | |  stripe | |  stripe | |  stripe | | (partial| (partial
  |   unit  | |   unit  | |   unit  | |   unit  | |  stripe | stripe
  | 15258785| | 15258786| | 15258787| | 15258788| |  unit)  | 3051757)
  \=========/ \=========/ \=========/ \=========/ \=========/
```