Mpi tests fail on both ceph-fuse and kclient

Running teuthology using the following yaml demonstrates this problem:

```
interactive-on-error: true
roles:
  - [mon.a, mon.b, mon.c, mds.a, osd.0, osd.1, osd.2]
  - [client.2]
  - [client.1]
  - [client.0]
overrides:
  ceph:
    conf:
      mds:
        debug ms: 1
        debug mds: 20
    client:
      debug ms: 1
      debug client: 20
overrides:
  ceph:
    fs: xfs
    conf:
    osd:
      osd sloppy crc: true
      osd op thread timeout: 60
# make sure we get the same MPI version on all hosts
os_type: ubuntu
os_version: "14.04"
tasks:
  - chef: null
  - install:
  - ceph:
  - kclient:
    pexec:
    clients:
      - cd $TESTDIR
        - wget http://ceph.com/qa/fsx-mpi.c
        - mpicc fsx-mpi.c -o fsx-mpi
        - rm fsx-mpi.c
        - ln -s $TESTDIR/mnt.* $TESTDIR/gmnt
        - ssh_keys:
        - mpi:
          exec: sudo $TESTDIR/fsx-mpi -o 1MB -N 50000 -p 10000 -l 1048576 $TESTDIR/gmnt
          workdir: $TESTDIR/gmnt
          - pexec:
            all:
```
At the point where the teuthology-task goes interactive, ssh to one of the clients. df gmnt shows:


running the following command (from the cephtest directory) shows the corresponding following failure:

```
COMMAND:


FAILURE:

Warning: Permanently added '10.214.138.156' (ECDSA) to the list of known hosts.
Warning: Permanently added '10.214.138.142' (ECDSA) to the list of known hosts.

skipping zero size read
truncating to largest ever: 0x7cccb
READ BAD DATA: offset = 0x2568c, size = 0xe8a1
OFFSET GOOD BAD RANGE
0x2568c 0x068f 0x0000 0x e838
operation# (mod 256) for the bad data unknown, check HOLE and EXTEND ops
LOG DUMP (8 total operations):
1 (1 mod 256): SKIPPED (no operation)
2 (2 mod 256): WRITE 0x1c748 thru 0xa6f7b (0x8a834 bytes) HOLE ***WWWW
3 (3 mod 256): WRITE 0x7ea33 thru 0x77b31 (0x8203 bytes)
4 (4 mod 256): READ 0x21437 thru 0x42e18 (0x219e2 bytes) ***RRRR***
5 (5 mod 256): MAPREAD 0x6f92f thru 0x77b31 (0x8203 bytes)
6 (6 mod 256): WRITE 0x1860 thru 0x8dc6f (0x8c410 bytes) ***WWWW
7 (7 mod 256): TRUNCATE DOWN from 0xa6f7c to 0x7cccb
8 (8 mod 256): READ 0x2568c thru 0x33f2c (0x8e410 bytes) ***RRRR***
Correct content saved for comparison
(maybe hexdump "mnt.1" vs "mnt.1.fsxgood"

===================================================================================
= BAD TERMINATION OF ONE OF YOUR APPLICATION PROCESSES
= EXIT CODE: 110
= CLEANING UP REMAINING PROCESSES
= YOU CAN IGNORE THE BELOW CLEANUP MESSAGES
===================================================================================

[proxy:0:2@vpm067] HYD_pmcd_pmip_control_cmd_cb (/pm/pmiserv/pmip_cb.c:886): assert (!closed) failed
[proxy:0:2@vpm067] HYDT_dmxu_poll_wait_for_event (/tools/demux/demux_poll.c:77): callback returned error status
[proxy:0:2@vpm067] main (/pm/pmiserv/pmip.c:206): demux engine error waiting for event

This test completes when run on a non-ceph device:

```
mpiexec -f /home/ubuntu/cephtest/mpi-hosts -wdir /home/ubuntu/cephtest/gmnt sudo /home/ubuntu/cephtest/fsx-mpi -o 1MB -N 50000 -p 10000 -l 1048576 /tmp/foobar

Similar problems can be seen if one uses ceph-fuse instead of kclient.
Running teuthology using the following yaml demonstrates this problem:

```
interactive-on-error: true
roles:
- [mon.a, mon.b, mon.c, mds.a, osd.0, osd.1, osd.2]
- [client.2]
- [client.1]
- [client.0]
overrides:
  ceph:
  conf:
    mds:
      debug ms: 1
      debug mds: 20
    client:
      debug ms: 1
      debug client: 20
overrides:
  ceph:
  fs: xfs
  conf:
  osd:
    osd sloppy crc: true
    osd op thread timeout: 60
# make sure we get the same MPI version on all hosts
os_type: ubuntu
os_version: "14.04"
tasks:
- chef: null
- install:
- ceph:
- kclient:
- pexec:
  clients:
  - cd $TESTDIR
  - wget http://ceph.com/qa/fsx-mpi.c
  - mpicc fsx-mpi.c -o fsx-mpi
  - rm fsx-mpi.c
  - ln -s $TESTDIR/mnt.* $TESTDIR/gmnt
- ssh_keys:
- mpi:
  exec: sudo $TESTDIR/fsx-mpi -o 1MB -N 50000 -p 10000 -l 1048576 $TESTDIR/gmnt
  workdir: $TESTDIR/gmnt
- pexec:
  all:
  - rm -rf $TESTDIR/gmnt
  - rm -rf $TESTDIR/fsx-mpi
```

At the point where the teuthology-task goes interactive, ssh to one of the clients. df gmnt shows:

```
```

running the following command (from the cephtest directory) shows the corresponding following failure:

```
COMMAND:

```
FAILURE:

Warning: Permanently added '10.214.138.156' (ECDSA) to the list of known hosts.
Warning: Permanently added '10.214.138.142' (ECDSA) to the list of known hosts.

skipping zero size read
truncating to largest ever: 0x7cccb

READ BAD DATA: offset = 0x2568c, size = 0xe8a1
OFFSET      GOOD     BAD     RANGE
0x2568c    0x068f    0x0000    0xe838

operation# (mod 256): SKIPPED (no operation)
2(2 mod 256): WRITE 0x1c748 thru 0xa6f7b (0x8a834 bytes) HOLE ***WWW
3(3 mod 256): WRITE 0x7ea33 thru 0x7fc14 (0x11e2 bytes)
4(4 mod 256): READ 0x21437 thru 0x42e18 (0x219e2 bytes) ***RRRR***
5(5 mod 256): MAPREAD 0x6f92f thru 0x77b31 (0x8203 bytes)
6(6 mod 256): WRITE 0x1860 thru 0x8dc6f (0x8c410 bytes) ***WWW
7(7 mod 256): TRUNCATE DOWN from 0xa6f7c to 0x7cccb
8(8 mod 256): READ 0x2568c thru 0x33f2c (0xe8a1 bytes) ***RRRR***

Correct content saved for comparison
(maybe hexdump "mnt.1" vs "mnt.1.fsxgood")

===============================================================================
   BAD TERMINATION OF ONE OF YOUR APPLICATION PROCESSES
   CLEAN UP REMAINING PROCESSES
   YOU CAN IGNORE THE BELOW CLEANUP MESSAGES
===============================================================================

[proxy:0:1@vpm152] HYD_pmcd_pmip_control_cmd_cb (/pm/pmiserv/pmip_cb.c:886): assert (!closed) failed
[proxy:0:1@vpm152] HYDT_dmxu_poll_wait_for_event (/tools/demux/demux_poll.c:77): callback returned error status
[proxy:0:1@vpm152] main (/pm/pmiserv/pmip.c:206): demux engine error waiting for event
[proxy:0:2@vpm067] HYD_pmcd_pmip_control_cmd_cb (/pm/pmiserv/pmip_cb.c:886): assert (!closed) failed
[proxy:0:2@vpm067] HYDT_dmxu_poll_wait_for_event (/tools/demux/demux_poll.c:77): callback returned error status
[proxy:0:2@vpm067] main (/pm/pmiserv/pmip.c:206): demux engine error waiting for event

This test completes when run on a non-ceph device:
<pre>
mpiexec -i /home/ubuntu/cephtest/mpi-hosts -wdir /home/ubuntu/cephtest/gmnt sudo /home/ubuntu/cephtest/fsx-mpi -o 1MB -N 50000 -p 10000 -l 1048576 /tmp/foobar
</pre>

Similar problems can be seen if one uses ceph-fuse instead of kclient.

11/25/2015
Is this a new failure on the existing cluster, or something that's come up on the new cluster?

If I'm understanding my quick skim correctly, this is not "MPI tests are failing" but "this mpi-fsx" test is failing, right? The suites/fs/multiclient/tasks/mdtest.yaml fragment is still being executed regularly and seems to be doing fine, whereas the suites/fs/multiclient/tasks/fsx-mpi.yaml.disabled fragment that I think you're trying to get going here was disabled a long time ago (and maybe it really is turning up a bug in CephFS).

please remove the sudo before /home/ubuntu/cephtest/fsx-mpi. Otherwise rank of all processes will be zero

the test passes after applying above patch and adding "-W -R" options to fsx-mpi, memory mapped IOs are disabled because cephfs don't synchronize memory mapped data among different hosts.