If you get on github and look for storage-benchmarks you'll see my tools to deploy fio etc.

Although I think doing a performance test on a 3 node cluster doesn't have a lot of value.
Well I guess you can compare against non EC pools.

1) Compare against non EC pools
2) Test the stability and recovery of an EC pool.
3) Also how much space we get back using an EC pool?
4) How bad is it when an OSD goes out on an EC pool?
5) Also how does it deal with corruption of one of the OSD's etc
6) What does the memory utilization look like when doing EC pools?
7) Are writes faster in EC pools vs non EC pools?
8) How slow are reads in EC pools vs non EC pools?

================================================================

http://docs.ceph.com/docs/hammer/dev/erasure-coded-pool/
http://docs.ceph.com/docs/master/rados/operations/erasure-code/
http://www.networkcomputing.com/storage/raid-vs-erasure-coding/1792588127

http://ceph.com/pgcalc/

**ERASURE CODED POOL INFO:**

- **The simplest erasure coded pool is equivalent to RAID5 and requires at least three hosts.**
- **Erasure coded pools require more resources than replicated pools and lack some functionalities such as partial writes. To overcome these limitations, it is recommended to set a cache tier before the erasure coded pool.**
- **You can't use erasure coded pools directly with RBD. They're only suitable for use with RGW or as the base pool for a replicated cache pool**
- **http://docs.ceph.com/docs/master/rados/operations/erasure-code/#erasure-coded-pool-and-cache-tiering**

- **chunk**
  when the encoding function is called, it returns chunks of the same size. Data chunks which can be concatenated to reconstruct the original object and coding chunks which can be used to rebuild a lost chunk.
  **K**
  the number of data chunks, i.e. the number of chunks in which the original object is divided. For instance if K = 2 a 10KB object will be divided into K objects of 5KB each.
  **M**
  the number of coding chunks, i.e. the number of additional chunks computed by the encoding functions. If there are 2 coding chunks, it means 2 OSDs can be out without losing data.

======================================================================

**CLUSTER INFO:**

| | |
|---|---|
| Pistore-cc38-e04.ece.comcast.net | 10.251.1.151 |
| Pistore-cc38-e05.ece.comcast.net | 10.251.1.152 |
| Pistore-cc38-e06.ece.comcast.net | 10.251.1.153 |

| | |
|---|---|
| Pistoremon-cc38-d01.ece.comcast.net | 10.251.1.71 |
| Pistoremon-cc38-d02.ece.comcast.net | 10.251.1.72 |
| Pistoremon-cc38-d03.ece.comcast.net | 10.251.1.68 |

```
[root@pistore-cc38-e05 ~]# ceph version
ceph version 0.94.5 (9764da52395923e0b32908d83a9f7304401fee43)

[root@pistore-cc38-e04 ~]# ceph -s
    cluster f709ca2f-2369-4e94-80c0-b9e8e5e20a49
     health HEALTH_OK
     monmap e3: 3 mons at
{pistoremon2-cc38-d01=10.251.1.71:6790/0,pistoremon2-cc38-d02=10.251.1.72:6790/0,pistoremon2-cc38-d03=10.
251.1.68:6790/0}
            election epoch 12, quorum 0,1,2 pistoremon2-cc38-d03,pistoremon2-cc38-d01,pistoremon2-cc38-d02
     osdmap e2811: 215 osds: 215 up, 215 in
      pgmap v62219: 13904 pgs, 18 pools, 102 GB data, 159 kobjects
            233 GB used, 1172 TB / 1173 TB avail
               13904 active+clean
[root@pistore-cc38-e04 ~]# ceph pg stat
v62262: 13904 pgs: 13904 active+clean; 102 GB data, 233 GB used, 1172 TB / 1173 TB avail
```

```
[root@pistore-cc38-e05 ~]# ceph osd pool create ecpool 12 12 erasure
pool 'ecpool' created

[root@pistore-cc38-e05 ~]# rados df
pool name                 KB      objects       clones     degraded      unfound           rd        rd KB           wr        wr KB
.intent-log                0            0            0            0            0            0            0            0            0
.log                       0            0            0            0            0            0            0            0            0
.rgw                       1            4            0            0            0           44           32           20            6
.rgw.buckets       103871081       162300            0            0            0            0            0       162312    103871081
.rgw.buckets.index         0            2            0            0            0           28           24           14            0
.rgw.control               0            8            0            0            0            0            0            0            0
.rgw.gc                    0           32            0            0            0         9888         9856         6592            0
.rgw.root                  1            3            0            0            0           81           54            3            3
```

```
.usage             0        0        0        0        0        0        0        0        0
.users             0        0        0        0        0        0        0        0        0
.users.email        0        0        0        0        0        0        0        0        0
.users.swift        0        0        0        0        0        0        0        0        0
.users.uid          1        2        0        0        0       21       18       14        1
cinder_backups          0        0        0        0        0        0        0        0        0
cinder_volumes     3156295      793        0        0        0      209      259    26215   3136179
ecpool             0        0        0        0        0        0        0        0        0
glance             0        0        0        0        0        0        0        0        0
rbd                0        0        0        0        0        0        0        0        0
  total used      254002392      163144
  total avail   1259309442628
  total space   1259563445020
```

```
[root@pistore-cc38-e05 ~]# ceph osd crush rule create-erasure ecruleset
created ruleset ecruleset at 2
```

```
[root@pistore-cc38-e05 ~]# ceph osd pool delete ecpool ecpool --yes-i-really-really-mean-it
pool 'ecpool' removed
```

**CREATE the POOL:**

```
[root@pistore-cc38-e05 ~]# ceph osd pool create ecpool 4096 4096 erasure default ecruleset
pool 'ecpool' created
```

```
[root@pistore-cc38-e05 ~]# ceph osd stat
     osdmap e2811: 215 osds: 215 up, 215 in

[root@pistore-cc38-e05 ~]# ceph -s
    cluster f709ca2f-2369-4e94-80c0-b9e8e5e20a49
     health HEALTH_OK
     monmap e3: 3 mons at
{pistoremon2-cc38-d01=10.251.1.71:6790/0,pistoremon2-cc38-d02=10.251.1.72:6790/0,pistoremon2-cc38-d03=10.251.1.68:6790/0}
            election epoch 12, quorum 0,1,2 pistoremon2-cc38-d03,pistoremon2-cc38-d01,pistoremon2-cc38-d02
     osdmap e2811: 215 osds: 215 up, 215 in
      pgmap v59997: 13904 pgs, 18 pools, 102 GB data, 159 kobjects
            233 GB used, 1172 TB / 1173 TB avail
               13904 active+clean
```

```
[root@pistore-cc38-e05 ~]# ceph osd pool ls
rbd
cinder_volumes
cinder_backups
glance
```

```
.rgw
.rgw.control
.rgw.gc
.log
.intent-log
.usage
.users
.users.email
.users.swift
.users.uid
.rgw.buckets
.rgw.root
.rgw.buckets.index
ecpool
```

**VERIFY CRUSHMAP**:

```
[root@pistore-cc38-e05 tmp]# ceph osd getcrushmap -o crushmap.raw
got crush map from osdmap epoch 2811
[root@pistore-cc38-e05 tmp]# crushtool -d crushmap.raw -o crushmap.decompiled
[root@pistore-cc38-e05 tmp]# vi crushmap.decompiled

# begin crush map
tunable choose_local_tries 0
tunable choose_local_fallback_tries 0
tunable choose_total_tries 50
tunable chooseleaf_descend_once 1
tunable straw_calc_version 1

# devices
device 0 osd.0
device 1 osd.1
<SNIP>
root default {
        id -1           # do not change unnecessarily
        # weight 1173.898
        alg straw
        hash 0  # rjenkins1
        item pistore-cc38-e04 weight 387.659
        item pistore-cc38-e05 weight 393.119
        item pistore-cc38-e06 weight 393.119
}

# rules
rule replicated_ruleset {
        ruleset 0
        type replicated
        min_size 1
        max_size 10
        step take default
        step chooseleaf firstn 0 type host
        step emit
}
rule erasure-code {
        ruleset 1
        type erasure
        min_size 3
        max_size 3
        step set_chooseleaf_tries 5
        step set_choose_tries 100
        step take default
        step chooseleaf indep 0 type host
        step emit
}
rule ecruleset {
        ruleset 2
        type erasure
        min_size 3
        max_size 3
        step set_chooseleaf_tries 5
        step set_choose_tries 100
        step take default
        step chooseleaf indep 0 type host
        step emit
}

# end crush map
```

**min_size**
Description: If a pool makes fewer replicas than this number, CRUSH will NOT select this rule.
Type: Integer
Purpose: A component of the rule mask.
Required: Yes
Default: 1

**max_size**
Description: If a pool makes more replicas than this number, CRUSH will NOT select this rule.
Type: Integer
Purpose: A component of the rule mask.
Required: Yes
Default: 10

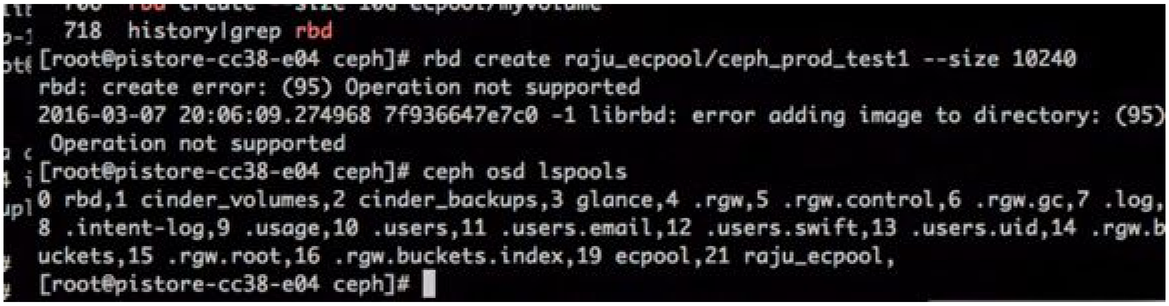
```
[root@pistore-cc38-e04 tmp]# ceph osd erasure-code-profile get default
```

```
directory=/usr/lib64/ceph/erasure-code
k=2
m=1
plugin=jerasure
technique=reed_sol_van
```

**CREATE BLOCK DEVICE**

- **rbd create ecpool/some-name --size 10240**
- **[root@pistore-cc38-e04 tmp]# rbd create ecpool/paul-test --size 102400**
    - **rbd: create error: (95) Operation not supported**
      **2016-03-07 20:14:47.279073 7ffbed4317c0 -1 librbd: error adding image to directory: (95) Operation not supported**

- **http://lists.ceph.com/pipermail/ceph-users-ceph.com/2014-July/041442.html**
  **You can't use erasure coded pools directly with RBD. They're only suitable for use with RGW or as the base pool for a replicated cache pool, and you need to be very careful/specific with the configuration. I believe this is well-documented, so check it out! :)**
  **-Greg**



```
  718  history|grep rbd
[root@pistore-cc38-e04 ceph]# rbd create raju_ecpool/ceph_prod_test1 --size 10240
rbd: create error: (95) Operation not supported
2016-03-07 20:06:09.274968 7f936647e7c0 -1 librbd: error adding image to directory: (95)
 Operation not supported
[root@pistore-cc38-e04 ceph]# ceph osd lspools
0 rbd,1 cinder_volumes,2 cinder_backups,3 glance,4 .rgw,5 .rgw.control,6 .rgw.gc,7 .log,
8 .intent-log,9 .usage,10 .users,11 .users.email,12 .users.swift,13 .users.uid,14 .rgw.b
uckets,15 .rgw.root,16 .rgw.buckets.index,19 ecpool,21 raju_ecpool,
[root@pistore-cc38-e04 ceph]# 
```

**ADD TIERING:**

```
ceph osd tier add ecpool hot-storage
ceph osd tier cache-mode hot-storage writeback
ceph osd tier set-overlay ecpool hot-storage
ceph osd tier set-overlay ecpool hot-storage2
```

- ceph osd pool create hot-storage2 128 128
- ceph osd tier add ecpool hot-storage2
- ceph osd tier cache-mode hot-storage2 writeback
- ceph osd tier set-overlay ecpool hot-storage2
- 
- rbd --pool ecpool create --size 10 myvolume
- ceph df
- rbd --pool ecpool create --size 10 volume1

- rbd --pool ecpool create --size 10 volume2

```
rados bench -p <pool_name> <seconds> <write|seq|rand>
rados bench -p ecpool 10 write --no-cleanup
rados bench -p ecpool 10 seq
rados bench -p ecpool 10 rand
```

**SPACE Savings- 35%**

```
        ecpool          19      27       0       781T       2
        raju_ecpool       21      20       0       781T       2
        hot-storage       25      529      0       586T       9
        hot-storage2       26    15104M      0       586T     3786
```

**IO TESTS**

```
[root@pistore-cc38-e04 ~]# rados bench -p ecpool 10 write --no-cleanup
 Maintaining 16 concurrent writes of 4194304 bytes for up to 10 seconds or 0 objects
 Object prefix: benchmark_data_pistore-cc38-e04.ece.comcast._752424
   sec Cur ops   started  finished  avg MB/s  cur MB/s  last lat   avg lat
     0       0         0         0         0         0         -         0
     1      16       366       350    1399.5      1400 0.0504525 0.0444853
     2      16       743       727   1453.65      1508 0.0455481 0.0434804
     3      16      1104      1088   1450.37      1444  0.042901 0.0437978
     4      15      1480      1465   1464.74      1508 0.0437389 0.0434479
     5      16      1854      1838   1470.16      1492 0.0387672 0.0433496
     6      15      2253      2238   1491.77      1600 0.0402353 0.0427585
     7      16      2638      2622   1498.06      1536 0.0450289 0.0425719
     8      15      3020      3005   1502.29      1532 0.0435767 0.0424847
     9      16      3396      3380   1502.02      1500  0.039936 0.0425092
    10      16      3776      3760    1503.8      1520 0.0423127 0.0424683
 Total time run:         10.022655
Total writes made:      3776
Write size:             4194304
Bandwidth (MB/sec):     1506.986

Stddev Bandwidth:       456.319
Max bandwidth (MB/sec): 1600
Min bandwidth (MB/sec): 0
Average Latency:        0.0424547
Stddev Latency:         0.00717848
Max latency:            0.276606
Min latency:            0.0189369

[root@pistore-cc38-e04 ~]# rados bench -p ecpool 10 seq
   sec Cur ops   started  finished  avg MB/s  cur MB/s  last lat   avg lat
     0       0         0         0         0         0         -         0
    1      16       892       876   3501.92      3504 0.0188736 0.0180087
    2      16      1759      1743   3484.76      3468 0.0155447 0.0182008
    3      16      2651      2635   3511.68      3568 0.0186046 0.0181565
    4      16      3582      3566   3564.41      3724 0.0147051 0.0179154
 Total time run:        4.215088
Total reads made:     3776
Read size:            4194304
Bandwidth (MB/sec):    3583.318

Average Latency:       0.0178299
Max latency:           0.261218
Min latency:           0.00752362

[root@pistore-cc38-e04 ~]# rados bench -p ecpool 10 rand
   sec Cur ops   started  finished  avg MB/s  cur MB/s  last lat   avg lat
     0       0         0         0         0         0         -         0
     1      16       887       871   3482.97      3484 0.0188287 0.0181858
     2      16      1747      1731   3461.31      3440 0.0226137  0.018387
     3      16      2639      2623   3496.67      3568  0.019687 0.0182443
     4      16      3544      3528   3527.39      3620 0.0150313 0.0180995
     5      16      4454      4438   3549.78      3640 0.0158618 0.0179898
     6      16      5378      5362   3574.08      3696  0.020528 0.0178719
     7      16      6308      6292   3594.87      3720 0.0196542 0.0177721
     8      16      7210      7194   3596.47      3608 0.0162085 0.0177744
     9      16      8138      8122   3609.26      3712 0.0169863 0.0177134
    10      16      9068      9052   3620.25      3720 0.0162538 0.0176595
 Total time run:        10.016762
Total reads made:     9068
Read size:            4194304
Bandwidth (MB/sec):    3621.130

Average Latency:       0.0176602
Max latency:           0.258535
Min latency:           0.00681743
```

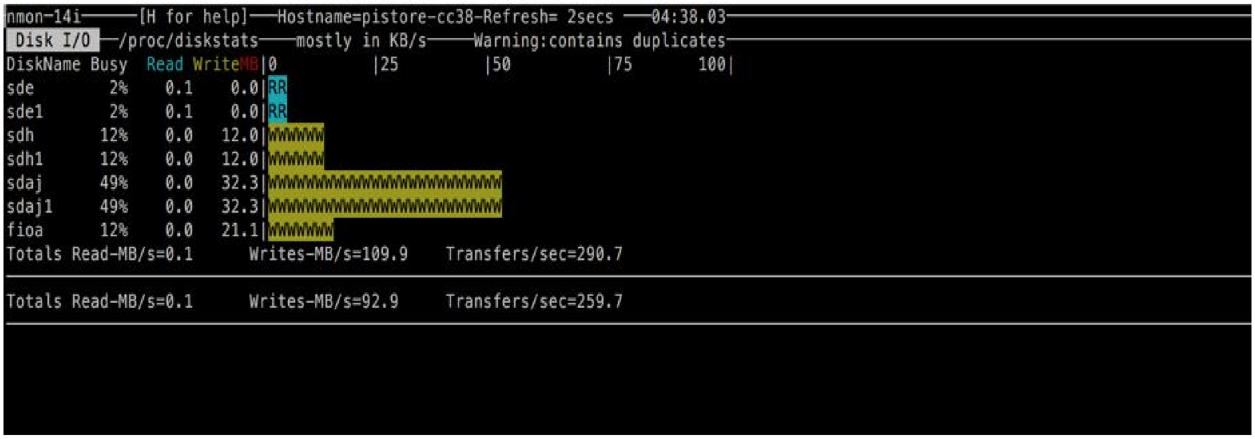
**FLUSH CACHE TIER**

```
[root@pistore-cc38-e04 ~]# rados -p hot-storage2  cache-flush-evict-all
```

**4GB/minute = 70MB/sec**

```
┌nmon─14i─────────[H for help]───Hostname=pistore-cc38-Refresh= 2secs
───04:33.52──────────────────────────────────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┐
│ Top Processes Procs=1072 mode=3 (1=Basic, 3=Perf 4=Size
5=I/O)────────────────────────────────────────────────────────────────────────────────
────────────────────────────────────────────────────────────────────────────────│
│  PID      %CPU    Size     Res    Res     Res     Res    Shared    Faults  Command
│
│           Used     KB     Set   Text   Data    Lib    KB     Min   Maj
```

```
|
|  126083    31.8 2246244  443312   11120 2134244      0 14964  30152      0 ceph-osd
|
|  209805    18.3 2301392  306316   11120 2188700      0 14492  14290      0 ceph-osd
|
|   13372     8.4 12862696  281544    2156 12183204      0  5548      4      0 beam.smp
|
|  882362     3.3   16668    5600     108    6108      0  1028    501      0 nmon_x86_64_rhe
|
|  106949     2.8 2155736  323848   11120 2044820      0 14616    447      0 ceph-osd
|
|  181555     2.8 2507324  647360   11120 2395284      0 14304   1056      0 ceph-osd
```



```
nmon-14i───────[H for help]───Hostname=pistore-cc38-Refresh= 2secs ───04:38.03──────────
 Disk I/O ──/proc/diskstats────mostly in KB/s─────Warning:contains duplicates─────────────
DiskName Busy  Read WriteMB|0          |25         |50          |75       100|
sde        2%    0.1    0.0|RR
sde1       2%    0.1    0.0|RR
sdh       12%    0.0   12.0|WWWWWW
sdh1      12%    0.0   12.0|WWWWWW
sdaj      49%    0.0   32.3|WWWWWWWWWWWWWWWWWWWWWWWWW
sdaj1     49%    0.0   32.3|WWWWWWWWWWWWWWWWWWWWWWWWW
fioa      12%    0.0   21.1|WWWWWWW
Totals Read-MB/s=0.1      Writes-MB/s=109.9     Transfers/sec=290.7
──────────────────────────────────────────────────────────────────────────────────────────
Totals Read-MB/s=0.1      Writes-MB/s=92.9      Transfers/sec=259.7
──────────────────────────────────────────────────────────────────────────────────────────
```

-

**LIST RBD devices in POOL:**

```
[root@pistore-cc38-e04 NMON]# rbd ls hot-storage2
myvolume
volume1
volume2
volume3
```

**DELETE an RBD VOLUME:**

```
[root@pistore-cc38-e04 NMON]# rbd rm volume3 -p hot-storage2
Removing image: 100% complete...done.

[root@pistore-cc38-e04 NMON]# rbd ls hot-storage2
myvolume
volume1
volume2
```

**WRITES**
**HOT-STORAGE2**

```
[root@pistore-cc38-e04 NMON]# rados bench -p hot-storage2 60  write --no-cleanup
```

```
Total time run:         60.027280
Total writes made:      10697
Write size:             4194304
Bandwidth (MB/sec):     712.809

Stddev Bandwidth:       250.413
Max bandwidth (MB/sec): 1348
Min bandwidth (MB/sec): 0
Average Latency:        0.0897722
Stddev Latency:         0.09571
Max latency:            0.516921
Min latency:            0.0219688
Bandwidth (MB/sec):     712.809
Bandwidth (MB/sec):     657.356
Bandwidth (MB/sec):     685.950
```

## ECPOOL

```
Total time run:         60.046008
Total writes made:      9336
Write size:             4194304
Bandwidth (MB/sec):     621.923

Stddev Bandwidth:       216.429
Max bandwidth (MB/sec): 1144
Min bandwidth (MB/sec): 0
Average Latency:        0.102897
Stddev Latency:         0.116177
Max latency:            0.635549
Min latency:            0.0344002

Bandwidth (MB/sec):     621.923
Bandwidth (MB/sec):     598.219
```

## RUN FIO TESTS

```
[root@pistore-cc38-e04 ~]# rbd --pool ecpool ls
myvolume
volume1
volume2
```

**FIO SERVER: pistore-cc38-e01** 10.251.1.148

## Test Recovery from Failed OSD

## CLEANUP

```
[root@pistore-cc38-e04 ~]# ceph osd pool delete hot-storage hot-storage --yes-i-really-really-mean-it
Error EBUSY: pool 'hot-storage' is a tier of 'raju_ecpool'
[root@pistore-cc38-e04 ~]# ceph osd pool delete raju_ecpool raju_ecpool --yes-i-really-really-mean-it
Error EBUSY: pool 'raju_ecpool' has tiers hot-storage


[root@pistore-cc38-e04 ~]# ceph osd tier remove hot-storage raju_ecpool
pool 'raju_ecpool' is now (or already was) not a tier of 'hot-storage'

[root@pistore-cc38-e04 ~]# ceph osd tier cache-mode hot-storage forward
set cache-mode for pool 'hot-storage' to forward

[root@pistore-cc38-e04 ~]# rados -p hot-storage ls
rb.0.20f99.238e1f29.000000000000
rb.0.8593c.238e1f29.000000000001
rb.0.8593c.238e1f29.000000000000
rb.0.20f99.238e1f29.000000000001
rb.0.c0a41.2ae8944a.000000000000
rbd_id.myvolume
rbd_children
rajuvolume2.rbd
rb.0.8593c.238e1f29.000000000002
rbd_id.rajuvolume3
rbd_directory
rb.0.20f99.238e1f29.000000000002
rbd_id.rajuvolume2
rb.0.c0a3e.238e1f29.000000000002
rb.0.c0a3e.238e1f29.000000000000
rb.0.c0a41.2ae8944a.000000000001
myvolume.rbd
rajuvolume3.rbd
rb.0.c0a41.2ae8944a.000000000002
rajuvolume1.rbd
rb.0.c0a3e.238e1f29.000000000001
rbd_id.rajuvolume1

[root@pistore-cc38-e04 ~]# rados -p hot-storage cache-flush-evict-all
        rb.0.20f99.238e1f29.000000000000
        rb.0.8593c.238e1f29.000000000001
        rb.0.8593c.238e1f29.000000000000
        rb.0.20f99.238e1f29.000000000001
        rb.0.c0a41.2ae8944a.000000000000
        rbd_id.myvolume
        rbd_children
        rajuvolume2.rbd
        rb.0.8593c.238e1f29.000000000002
        rbd_id.rajuvolume3
        rbd_directory
        rb.0.20f99.238e1f29.000000000002
        rbd_id.rajuvolume2
        rb.0.c0a3e.238e1f29.000000000002
        rb.0.c0a3e.238e1f29.000000000000
        rb.0.c0a41.2ae8944a.000000000001
        myvolume.rbd
        rajuvolume3.rbd
        rb.0.c0a41.2ae8944a.000000000002
        rajuvolume1.rbd
        rb.0.c0a3e.238e1f29.000000000001
        rbd_id.rajuvolume1

[root@pistore-cc38-e04 ~]# ceph osd tier remove-overlay raju_ecpool
there is now (or already was) no overlay for 'raju_ecpool'

[root@pistore-cc38-e04 ~]# ceph osd tier remove raju_ecpool hot-storage
pool 'hot-storage' is now (or already was) not a tier of 'raju_ecpool'

[root@pistore-cc38-e04 ~]# ceph osd pool delete raju_ecpool raju_ecpool --yes-i-really-really-mean-it
pool 'raju_ecpool' removed
[root@pistore-cc38-e04 ~]# ceph osd pool delete hot-storage hot-storage --yes-i-really-really-mean-it
pool 'hot-storage' removed
```