Lawrence Livermore National Laboratory

LMT
Lustre Monitoring Tools

April 13, 2011

Christopher Morrone

Lawrence Livermore National Laboratory, P. O. Box 808, Livermore, CA 94551

This work performed under the auspices of the U.S. Department of Energy by
Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344

LLNL-PRES-479171

LMT Mission

- Collect and display real-time and historical information about Lustre filesystem activity.

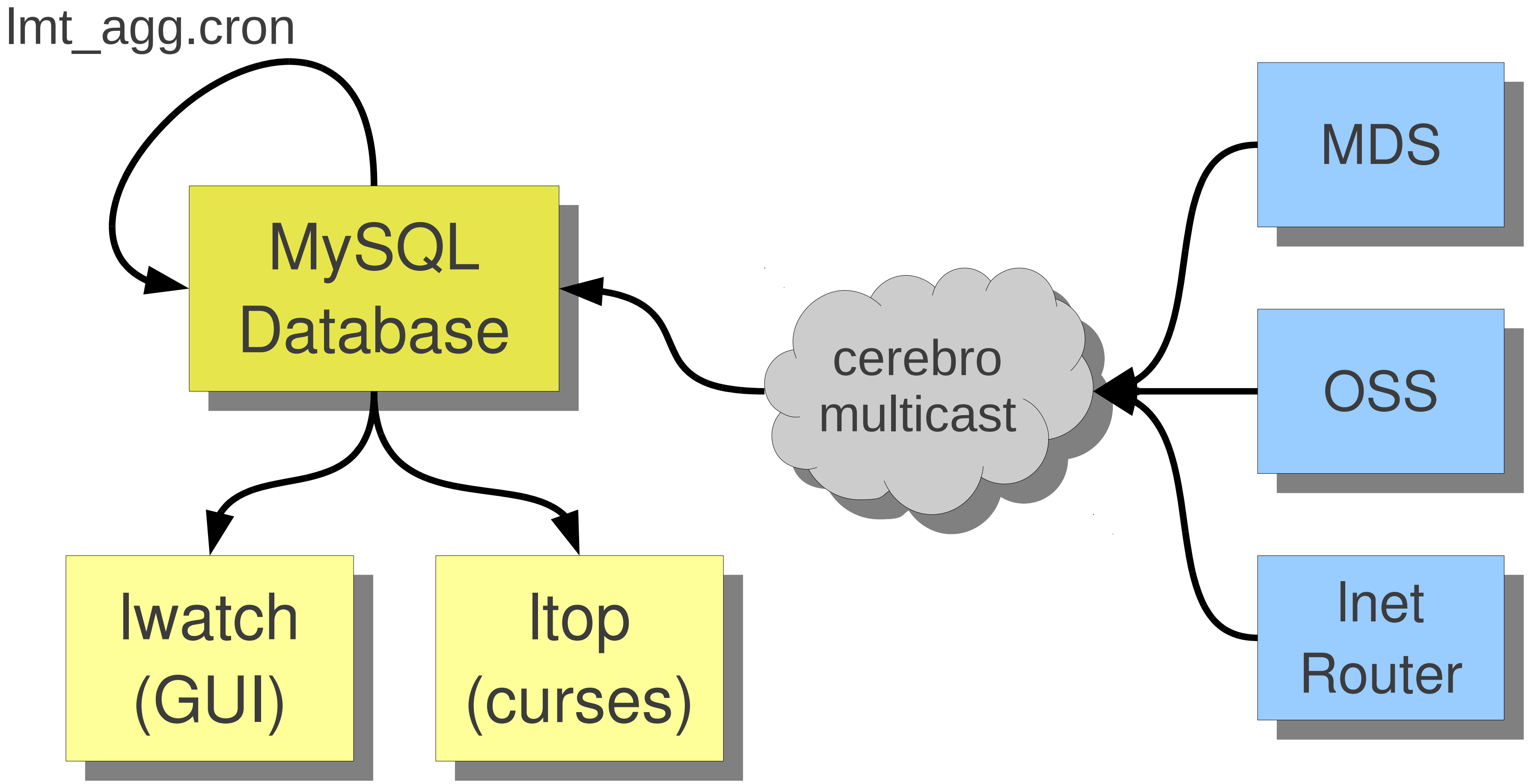
LMT History

- LMT Version 1
  - In-house python prototype
- LMT Version 2
  - Rewrite of LMT in C, Java, Perl, Sh
  - Cerebro for data collection
  - Incorporated MySQL for logging data, proved very useful [Uselton 2009 CUG]
  - ltop text utility
  - lwatch GUI in Java
- LMT Version 3
  - Major improvments by Jim Garlick

Cerebro overview

- Cluster monitoring daemon, tools and libraries
- Inspired by ganglia
- Uses multicast
- Dynamic module interface (plugins) for adding “metrics”
  - /usr/lib/cerebro
- Current metrics:
  - cerebro_metric_lmt_mdt
  - cerebro_metric_lmt_ost
  - cerebro_metric_lmt_router
  - cerebro_metric_lmt_osc

LMT 2 architecture



Problems in LMT version 2

- Lustre config must be expressed in an odd language, then pre-loaded into MySQL.
- Nothing functions until both MySQL and cerebro are up.
- Poor error handling and logging make debug difficult.
- There are two overlapping config files in odd locations.
- The cerebro module code is prototype quality and brittle.

Improved in Version 3

- Lustre config is automatically determined on the fly.
- ltop now functions as soon as cerebro is up.
- MySQL is actually optional now.
- Error handling and logging are rewritten/improved.
- Cerebro module code has been refactored/rewritten.
- There is a single new config file: /etc/lmt/lmt.conf
- More data is collected/shown in ltop

LUA-based lmt.conf

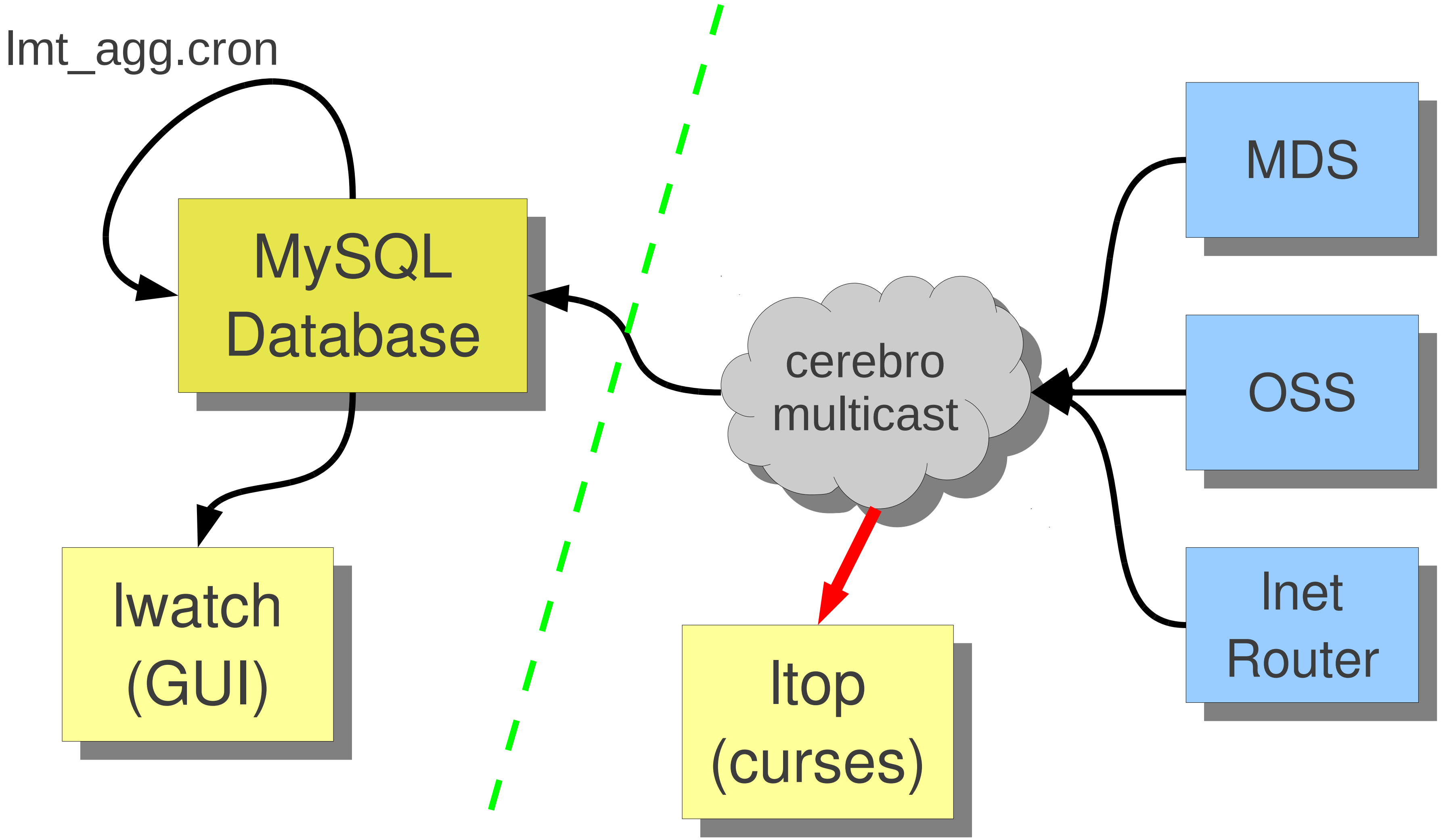
```
lmt_cbr_debug = 0
lmt_proto_debug = 0
lmt_db_debug = 0
lmt_db_host = nil
lmt_db_port = 0
lmt_db_rouser = "lwatchclient"
lmt_db_ropasswd = nil
lmt_db_rwuser = "lwatchadmin"

f = io.open("/etc/lmt/rwpasswd")
if (f) then
   lmt_db_rwpasswd = f:read("*all")
   f:close()
else
   lmt_db_rwpasswd = nil
end
```

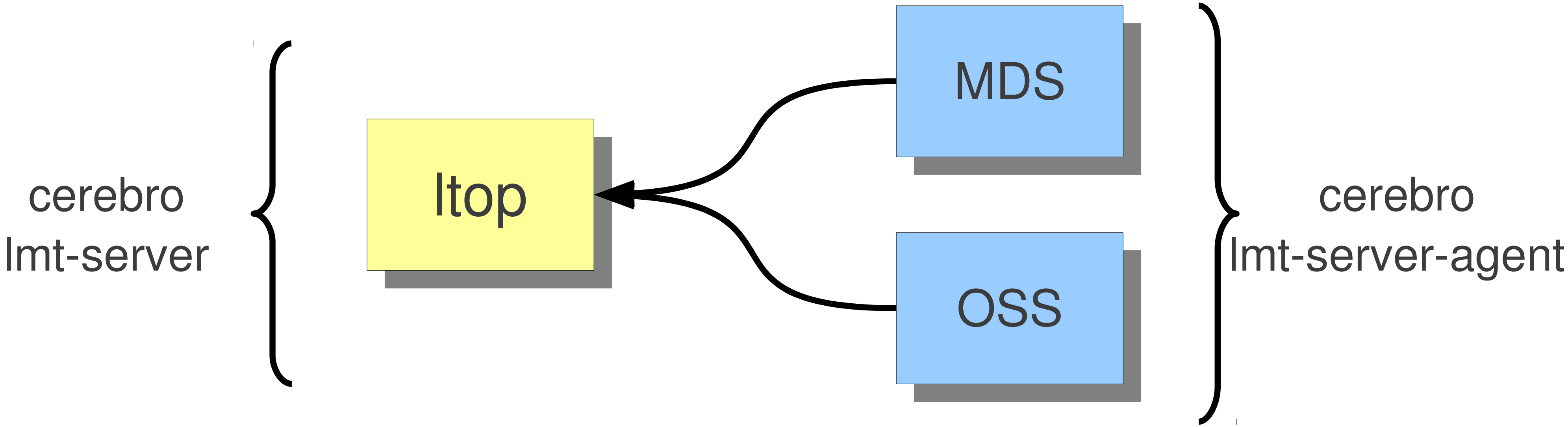
Unchanged in Version 3

- The architecture is the same (except ltop).
- The database schema is unchanged.
- The lwatch/lstat java clients are unchanged (moved to separate lmt-gui package).
- Cron aggregation scripts that convert high → low-res
- MySQL sample data still exist (kludge!).

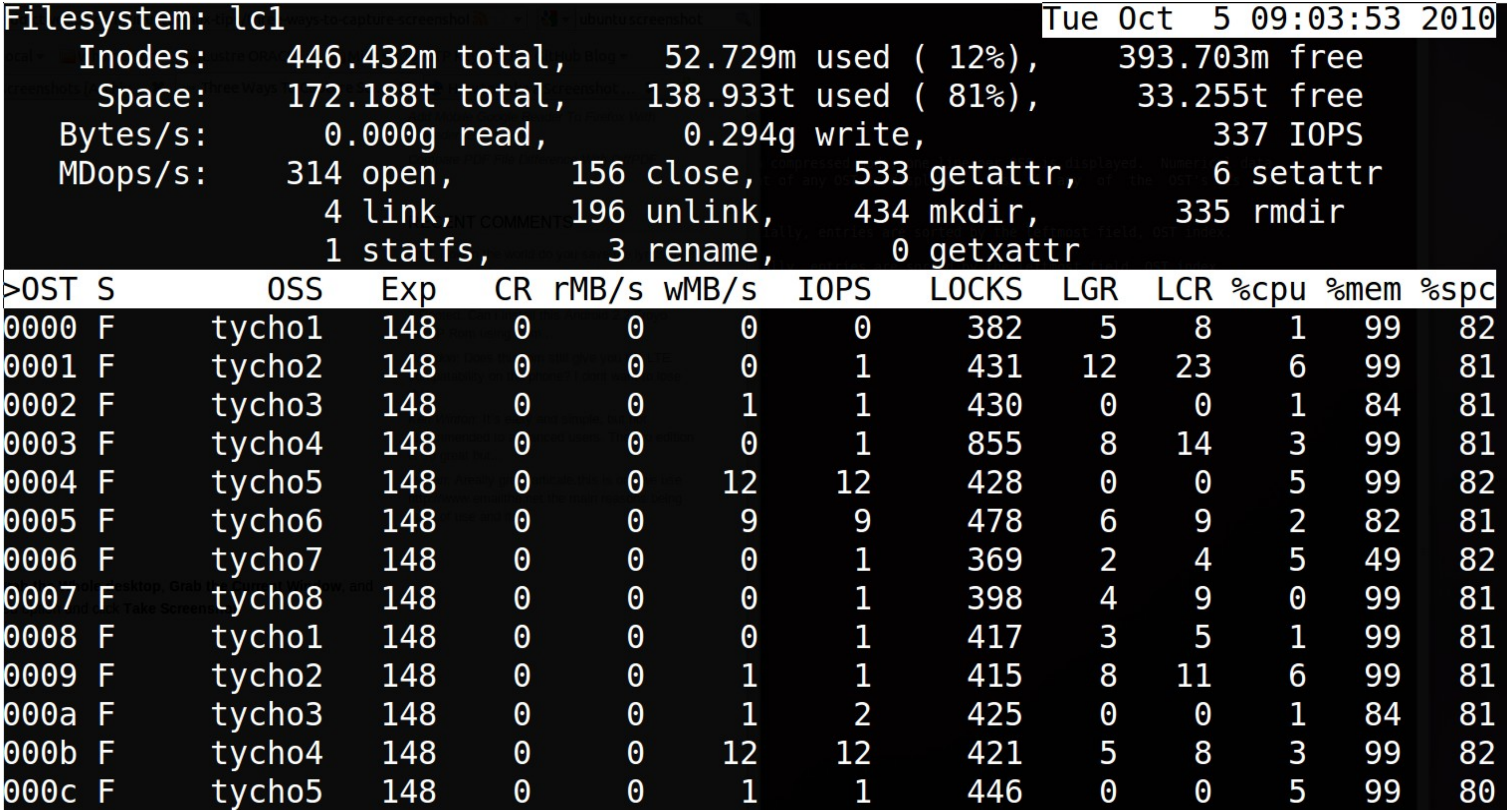
LMT 3 architecture



Simple setup for ltop

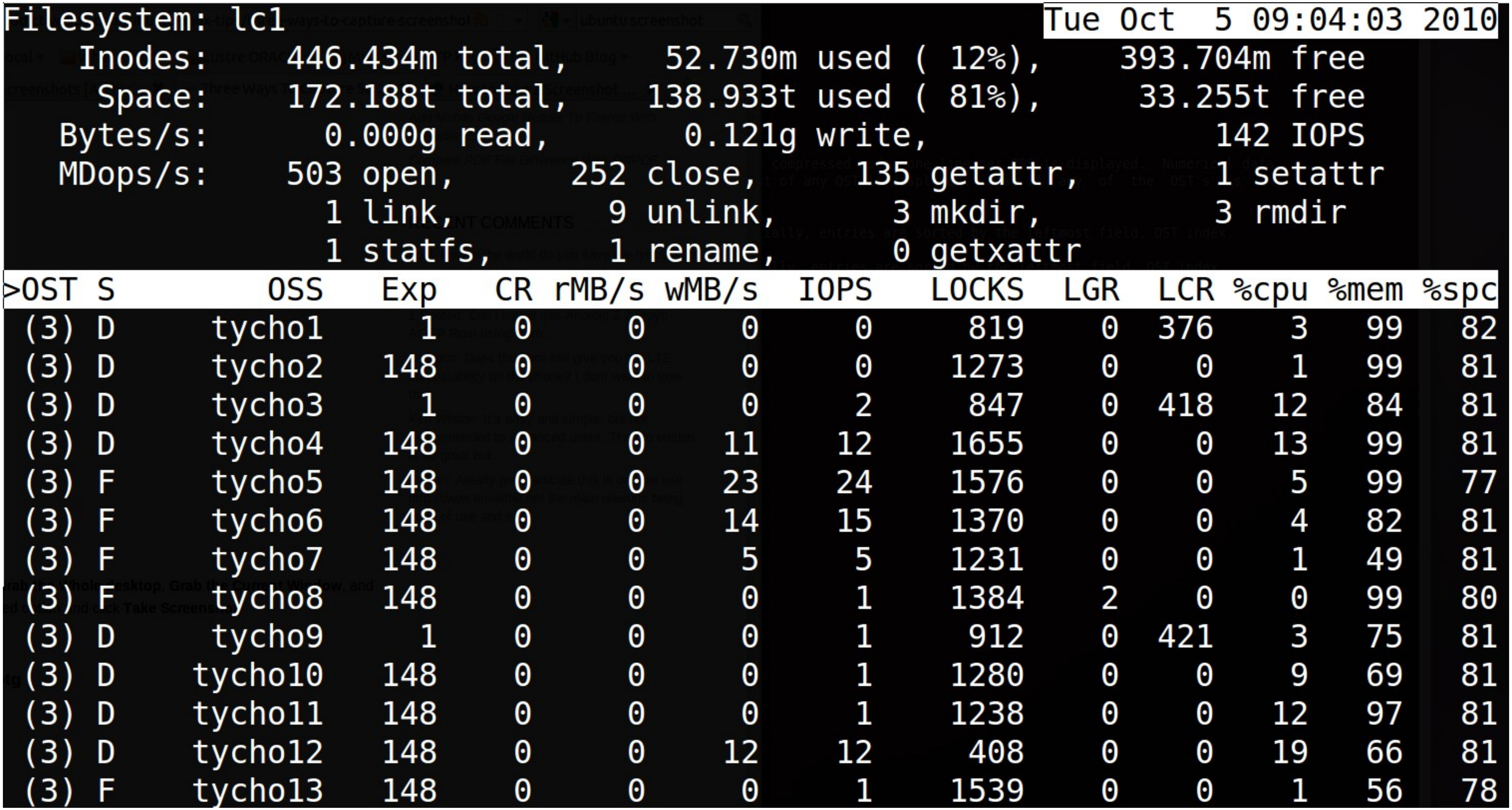


ltop screenshot



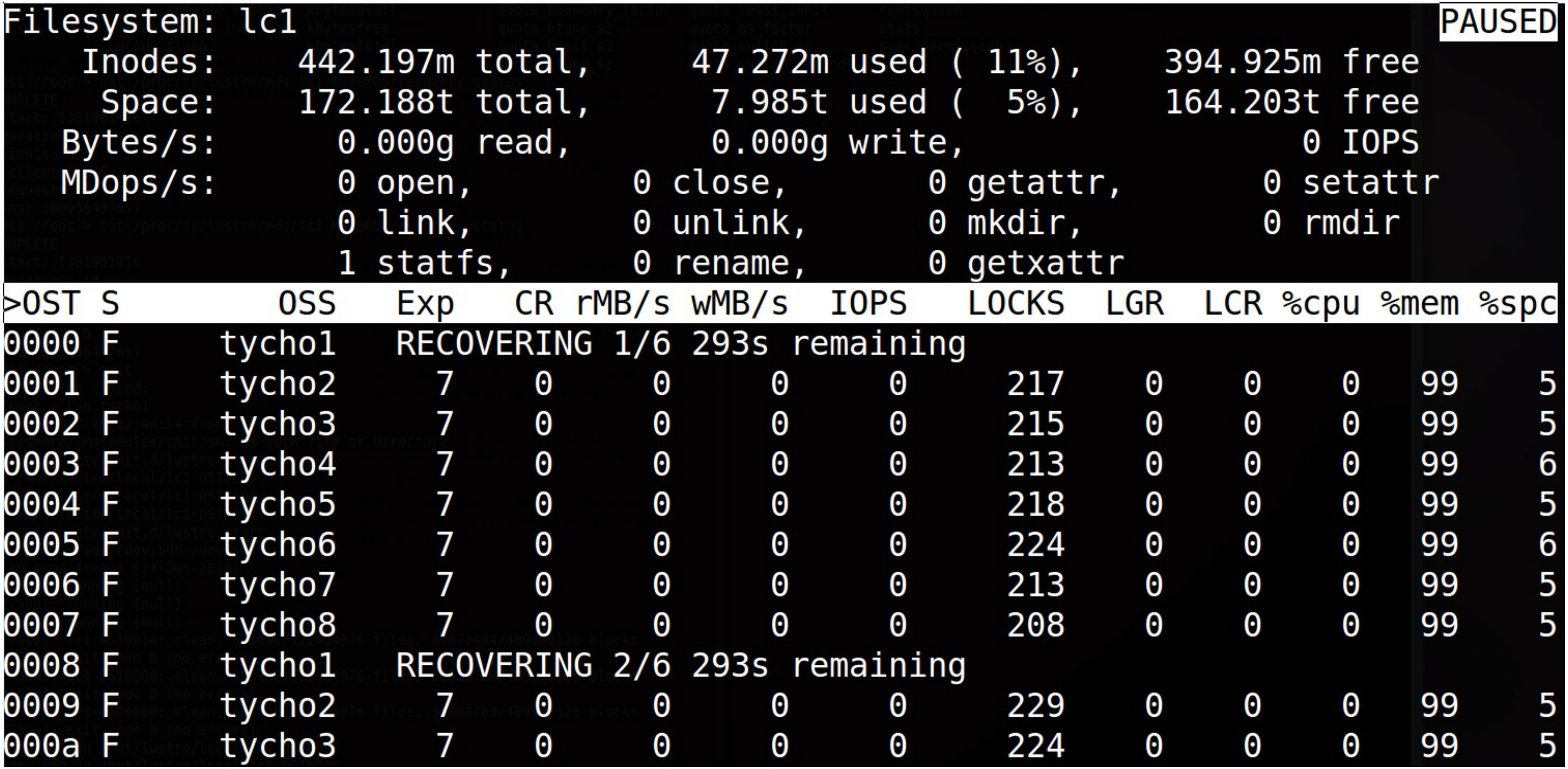
```
Filesystem: lc1                                        Tue Oct  5 09:03:53 2010
    Inodes: 446.432m total,     52.729m used ( 12%),    393.703m free
     Space: 172.188t total,    138.933t used ( 81%),     33.255t free
   Bytes/s:   0.000g read,       0.294g write,               337 IOPS
   MDops/s:   314 open,        156 close,    533 getattr,       6 setattr
                4 link,        196 unlink,   434 mkdir,       335 rmdir
                1 statfs,        3 rename,     0 getxattr
>OST S        OSS   Exp   CR rMB/s wMB/s  IOPS   LOCKS  LGR  LCR %cpu %mem %spc
0000 F     tycho1   148    0     0     0     0     382    5    8    1   99   82
0001 F     tycho2   148    0     0     0     1     431   12   23    6   99   81
0002 F     tycho3   148    0     0     1     1     430    0    0    1   84   81
0003 F     tycho4   148    0     0     0     1     855    8   14    3   99   81
0004 F     tycho5   148    0     0    12    12     428    0    0    5   99   82
0005 F     tycho6   148    0     0     9     9     478    6    9    2   82   81
0006 F     tycho7   148    0     0     0     1     369    2    4    5   49   82
0007 F     tycho8   148    0     0     0     1     398    4    9    0   99   81
0008 F     tycho1   148    0     0     0     1     417    3    5    1   99   81
0009 F     tycho2   148    0     0     1     1     415    8   11    6   99   81
000a F     tycho3   148    0     0     1     2     425    0    0    1   84   81
000b F     tycho4   148    0     0    12    12     421    5    8    3   99   82
000c F     tycho5   148    0     0     1     1     446    0    0    5   99   80
```

ltop ost “compressed” view



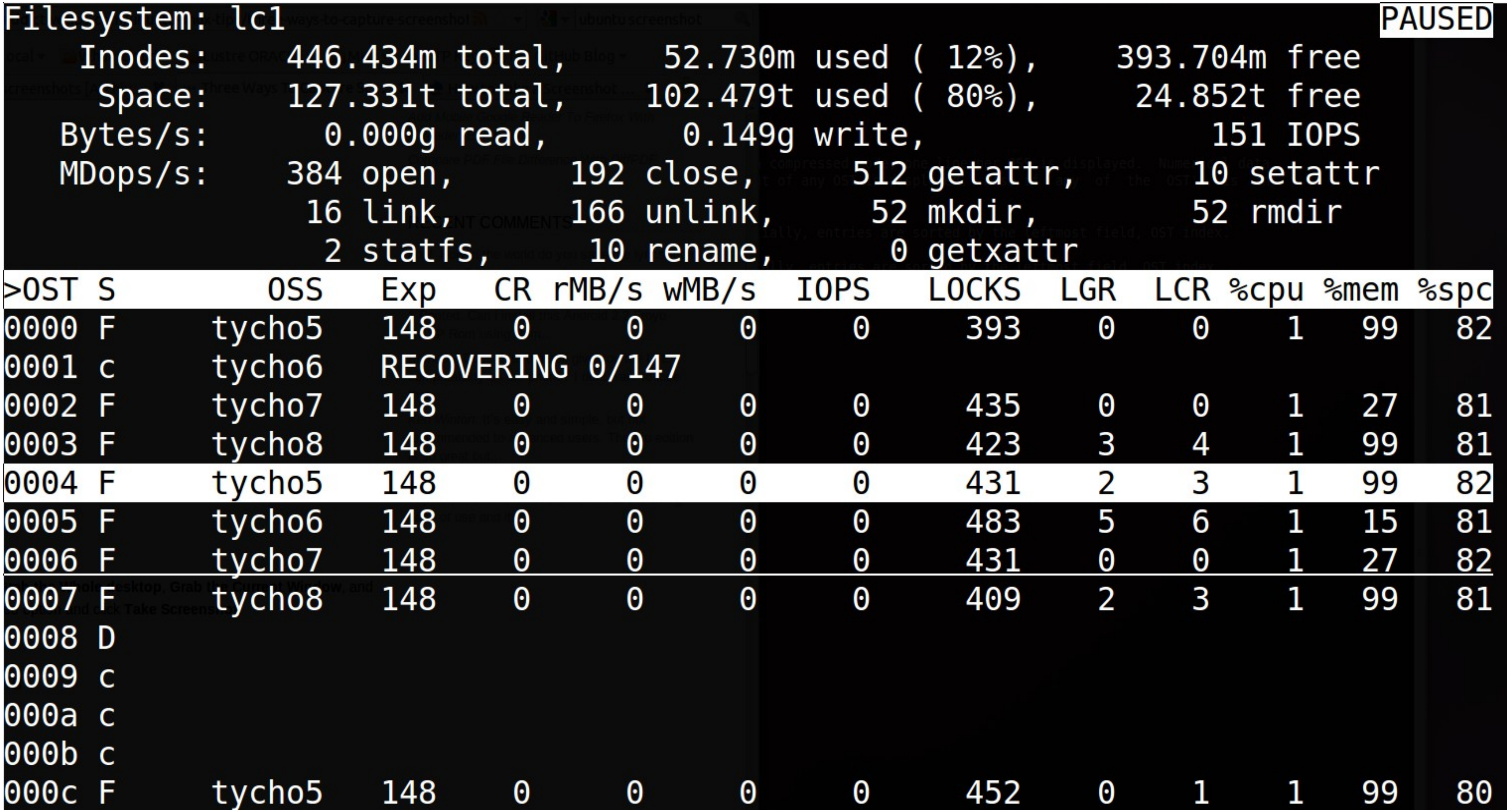
```
Filesystem: lc1                                            Tue Oct  5 09:04:03 2010
    Inodes:  446.434m total,    52.730m used ( 12%),    393.704m free
     Space:  172.188t total,   138.933t used ( 81%),     33.255t free
   Bytes/s:    0.000g read,      0.121g write,              142 IOPS
   MDops/s:  503 open,      252 close,  135 getattr,        1 setattr
               1 link,        9 unlink,     3 mkdir,        3 rmdir
               1 statfs,      1 rename,     0 getxattr
>OST S         OSS   Exp   CR rMB/s wMB/s  IOPS   LOCKS  LGR  LCR %cpu %mem %spc
 (3) D     tycho1     1    0     0     0     0     819    0  376    3   99   82
 (3) D     tycho2   148    0     0     0     0    1273    0    0    1   99   81
 (3) D     tycho3     1    0     0     0     2     847    0  418   12   84   81
 (3) D     tycho4   148    0     0    11    12    1655    0    0   13   99   81
 (3) F     tycho5   148    0     0    23    24    1576    0    0    5   99   77
 (3) F     tycho6   148    0     0    14    15    1370    0    0    4   82   81
 (3) F     tycho7   148    0     0     5     5    1231    0    0    1   49   81
 (3) F     tycho8   148    0     0     0     1    1384    2    0    0   99   80
 (3) D     tycho9     1    0     0     0     1     912    0  421    3   75   81
 (3) D    tycho10   148    0     0     0     1    1280    0    0    9   69   81
 (3) D    tycho11   148    0     0     0     1    1238    0    0   12   97   81
 (3) D    tycho12   148    0     0    12    12     408    0    0   19   66   81
 (3) F    tycho13   148    0     0     0     1    1539    0    0    1   56   78
```

ltop recovery example



```
Filesystem: lc1                                                                PAUSED
    Inodes:    442.197m total,     47.272m used ( 11%),    394.925m free
     Space:    172.188t total,      7.985t used (  5%),    164.203t free
   Bytes/s:      0.000g read,       0.000g write,                 0 IOPS
   MDops/s:      0 open,        0 close,       0 getattr,       0 setattr
                 0 link,        0 unlink,      0 mkdir,         0 rmdir
                 1 statfs,      0 rename,      0 getxattr
>OST S        OSS    Exp    CR rMB/s wMB/s   IOPS    LOCKS   LGR   LCR %cpu %mem %spc
0000 F     tycho1    RECOVERING 1/6 293s remaining
0001 F     tycho2      7     0     0     0      0      217     0     0    0   99    5
0002 F     tycho3      7     0     0     0      0      215     0     0    0   99    5
0003 F     tycho4      7     0     0     0      0      213     0     0    0   99    6
0004 F     tycho5      7     0     0     0      0      218     0     0    0   99    5
0005 F     tycho6      7     0     0     0      0      224     0     0    0   99    6
0006 F     tycho7      7     0     0     0      0      213     0     0    0   99    5
0007 F     tycho8      7     0     0     0      0      208     0     0    0   99    5
0008 F     tycho1    RECOVERING 2/6 293s remaining
0009 F     tycho2      7     0     0     0      0      229     0     0    0   99    5
000a F     tycho3      7     0     0     0      0      224     0     0    0   99    5
```

ltop screenshot



```
Filesystem: lc1                                                        PAUSED
    Inodes: 446.434m total,    52.730m used ( 12%),    393.704m free
     Space: 127.331t total,   102.479t used ( 80%),     24.852t free
   Bytes/s:   0.000g read,      0.149g write,               151 IOPS
   MDops/s:     384 open,        192 close,   512 getattr,   10 setattr
                 16 link,        166 unlink,    52 mkdir,     52 rmdir
                  2 statfs,       10 rename,     0 getxattr
>OST S        OSS    Exp    CR rMB/s wMB/s  IOPS   LOCKS  LGR  LCR %cpu %mem %spc
0000 F     tycho5    148     0     0     0     0     393    0    0    1   99   82
0001 c     tycho6    RECOVERING 0/147
0002 F     tycho7    148     0     0     0     0     435    0    0    1   27   81
0003 F     tycho8    148     0     0     0     0     423    3    4    1   99   81
0004 F     tycho5    148     0     0     0     0     431    2    3    1   99   82
0005 F     tycho6    148     0     0     0     0     483    5    6    1   15   81
0006 F     tycho7    148     0     0     0     0     431    0    0    1   27   82
0007 F     tycho8    148     0     0     0     0     409    2    3    1   99   81
0008 D
0009 c
000a c
000b c
000c F     tycho5    148     0     0     0     0     452    0    1    1   99   80
```

Get LMT

- LMT Code
  - https://github.com/chaos/lmt
  - https://github.com/chaos/lmt-gui
- LMT Wiki
  - https://github.com/chaos/lmt/wiki
- Cerebro Code
  - https://github.com/chaos/cerebro