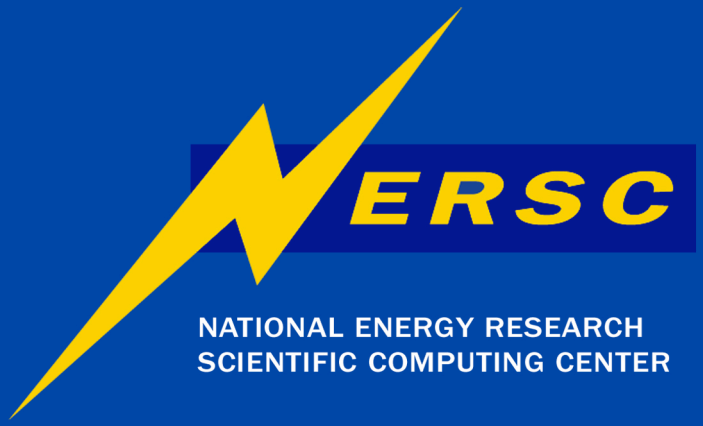


National Energy Research Scientific Computing Center (NERSC)

Quotas and Backups

Lustre User Group, 2008

Nicholas P. Cardo
NERSC Center Division, LBNL
April 29, 2008

Words of Wisdom

"Strive for perfection in everything. Take the best that exists and make it better. If it does not exist, create it. Accept nothing nearly right or good enough."

Sir Henry Royce

co-founder of Rolls-Royce

Disk Quota Facts

- A necessary evil. *If there's space it will be filled.*
- File systems are getting bigger. *~18TB and ~350TB. 1 TB disks are a reality!*
- A full file system = disaster for all.
- Computers are getting bigger and faster. *Can generate significant amounts of data quickly.*

Have You Seen…

```
$ lfs quota -u cardo /scratch
Disk quotas for user cardo (uid 18599):
     Filesystem  blocks   quota   limit   grace   files   quota   limit
grace
       /scratch 1687752564       0       0           26386       0       0
nid04116_mds_UUID
                    5844          102400           26386                0
      ost0_UUID 26740924          26828800
      ost1_UUID 26742568          26828800
      ost2_UUID 29325236          29388800
      ost3_UUID 29587756          29696000
      ost4_UUID 20114840          20172800
      ost5_UUID 20109372          20172800
      ost6_UUID 17650140          17715200
      ost7_UUID 17440000          17510400
      ost8_UUID 17466164          17612800
      ost9_UUID 17465452          17612800
     ost10_UUID 17721944          17817600
     ost11_UUID 17503604          17612800
```

But Wait…

```
ost12_UUID 18630136          18739200
ost13_UUID 18838112          18944000
ost14_UUID 12885324          13004800
ost15_UUID 13327864          13414400
ost16_UUID 12512808          12595200
ost17_UUID 12378468          12492800
ost18_UUID 12682500          12800000
ost19_UUID 12542668          12595200
ost20_UUID 18609748          18739200
ost21_UUID 18847072          18944000
ost22_UUID 18780612          18841600
ost23_UUID 18468464          18534400
ost24_UUID 18403884          18534400
ost25_UUID 18397220          18534400
ost26_UUID 18164172          18227200
ost27_UUID 20948288          21094400
ost28_UUID 25838008          25907200
ost29_UUID 21449944          21504000
ost30_UUID 21601908          21708800
ost31_UUID 21438272          21504000
ost32_UUID 9859796          9932800
ost33_UUID 9557404          9625600
ost34_UUID 9386612          9523200
```

Hold that thought…

```
ost35_UUID 7070468          7168000
ost36_UUID 15934728          16076800
ost37_UUID 15983992          16076800
ost38_UUID 16117040          16179200
ost39_UUID 15984508          16076800
ost40_UUID 15984072          16076800
ost41_UUID 15953536          16076800
ost42_UUID 15876940          15974400
ost43_UUID 17717568          17817600
ost44_UUID 18019332          18124800
ost45_UUID 21947952          22016000
ost46_UUID 23795304          23859200
ost47_UUID 21806784          21913600
ost48_UUID 21503088          21606400
ost49_UUID 22613668          22732800
ost50_UUID 22611272          22732800
ost51_UUID 22880244          22937600
ost52_UUID 22888436          23040000
ost53_UUID 21471056          21606400
ost54_UUID 22192236          22323200
ost55_UUID 22205908          22323200
ost56_UUID 22211820          22323200
ost57_UUID 22212324          22323200
ost58_UUID 24001288          24064000
```

Are we there yet?

```
ost59_UUID 23684924          23756800
ost60_UUID 23689340          23756800
ost61_UUID 23684780          23756800
ost62_UUID 21192552          21299200
ost63_UUID 21262224          21401600
ost64_UUID 32693736          32768000
ost65_UUID 32693688          32768000
ost66_UUID 32990524          33075200
ost67_UUID 32921404          32972800
ost68_UUID 26616088          26726400
ost69_UUID 26652080          26726400
ost70_UUID 26910636          27033600
ost71_UUID 26910940          27033600
ost72_UUID 29576136          29696000
ost73_UUID 29560724          29696000
ost74_UUID 29063060          29184000
ost75_UUID 29359560          29491200
ost76_UUID 27041908          27136000
ost77_UUID 27015836          27136000
ost78_UUID 26981072          27033600
ost79_UUID 26844660          26931200
```

Now what was I looking for?

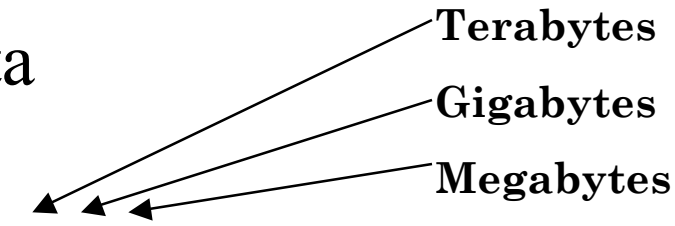
Lustre API Can Help

```
/*
*   Get the current quota limits
*/
qctl.qc_cmd  = LUSTRE_Q_GETQUOTA;
qctl.qc_type = USRQUOTA;
qctl.qc_id   = uid;

rc = llapi_quotactl(fs,&qctl);

if (rc != 0) {
   fprintf(stderr,"%s: fatal quotactl error \"%s (%d)\"\n",ProgName,strerror(errno),errno);

} else {
  /*
   *  Transfer the data
   */
   ui->ui_uid        = qctl.qc_id;
   ui->ui_shardlimit = qctl.qc_dqblk.dqb_bhardlimit     / 1024 / 1024; /* in GB */
   ui->ui_curspace   = toqb(qctl.qc_dqblk.dqb_curspace) / 1024 / 1024;
   ui->ui_ihardlimit = qctl.qc_dqblk.dqb_ihardlimit;
   ui->ui_curinodes  = qctl.qc_dqblk.dqb_curinodes;
}
```

Objectives

- Present quota usage to users in an easy to read manner. Include all quotas in one display, regardless of filesystem type. *(myquota)*
- Produce a report for all users usage that is sorted by either inodes or storage. *(quotarpt)*
- Easy to use command line quota management with exception tracking. *(chquota)*

myquota



$ myquota [[-u username] | [-g groupname]] [-TGM]

```
Displaying quota usage for user cardo:
                ---------- Space (GB) ----------  ------------ Inode -------------
FileSystem       Usage     Quota    InDoubt  Grace   Usage     Quota    InDoubt  Grace
----------      -------   -------   -------  -----  -------   -------  -------  -----
scratch          1610      2048        -       -     26386    100000        -       -
u0                  3       100        -       -      1523     25000        -       -
project             0         -        0       -      1033         -      208       -
```

Not Lustre

Supports:

Lustre	**EXT3**
GPFS	**NFS mounted GPFS**
XFS	

quotarpt

$ quotarpt [-f filesystem] [-n count] [-s | -i]

```
Filesystem: /u0
Report Type: Space
Report Date: Thu Apr 24 08:42:52 2008

          ---- Space (GBs) --- Inode ---
Username  Usage  Quota   Usage    Quota
-------- ------ ------ -------- --------
User1       137    140      997    15000
User2        54     60     4933    15000
User3        48     50    11427    15000
User4        47     50     4818    15000
User5        34     50     2937    15000
User6        33     50    77484    85000
User7        33     50     2906    15000
User8        32     50     2883    15000
User9        25     50    14904    15000
user10       24     50     5547    15000
```

chquota

```
$ chquota <actions> <parameters>
Valid Actions are:
   -a ........ autoclear expired requests
   -c ........ override default quotas
   -n ........ new quota entries, sets defaults
   -r ........ reset a quota back to defaults
   -R ........ Report all overrides

Valid parameters are:
   -u <user> ...... username
   -g <group> ..... groupname
   -i <limit> ..... inode limit
   -f <filesys> ... file system
   -s <limit> ..... space limit with unit (T|G|M)
   -t <ticket> .... trouble ticket
   -e <expire> .... expiration date
```

		Space Quota			Inode Quota			
Username	Q	GigaBs	Expiration	Ticket	Inodes	Expiration	Ticket	Filesystem
user1	U	1000	05/01/2008	071022-000055	-	--/--/----	-	/scratch
user2	U	-	--/--/----	-	1000000	12/14/2008	071219-000044	/scratch
user3	U	3072	07/10/2008	080103-000010	300000	01/12/2009	080214-000035	/scratch
user4	U	1024	07/06/2008	080201-000043	-	--/--/----	-	/scratch
user5	U	3072	05/22/2008	080212-000114	-	--/--/----	-	/scratch
user6	U	3072	11/16/2008	071116-000028	5000000	11/16/2008	071116-000028	/scratch
uesr7	U	910	06/19/2008	071206-000020	1300000	06/19/2008	071206-000020	/scratch
uesr8	U	750	09/30/2008	071203-000035	120000	01/31/2009	071220-000166	/scratch
user9	U	1024	01/12/2009	080312-000030	-	--/--/----	-	/scratch
user10	U	1024	07/04/2009	080331-000143	100000	06/15/2008	071214-000004	/scratch
user11	U	2048	06/15/2008	071214-000020	-	--/--/----	-	/scratch
user12	U	1024	11/27/2008	080314-000071	-	--/--/----	-	/scratch
user13	U	1024	04/04/2009	080404-000005	-	--/--/----	-	/scratch
user14	U	20480	07/07/2009	080326-000051	5000000	07/07/2009	080326-000051	/scratch
user15	U	2048	05/08/2009	080408-000071	-	--/--/----	-	/scratch

Room for Improvement

- Consider Q_GETNEXT functionality for sequential listing of all quota entries.
- Consider Q_GETALL functionality to retrieve all quota information for all entries in one operation.
- What's with KiloBytes??? Support units through TeraBytes.
- Better documentation.

File System Backups

- There will be file system failures. *It is inevitable.*
- Need to backup/restore "large" amounts of data. *The definition of "large" seems to increase annually.*
- Backups and Restores need to be "fast". *The definition of "fast" increases as the amount of data increases.*

Did You Know About…

- e2scan: a fast inode scanner that runs on the MDS.
- llbackup: a backup/restore utility that can perform a filesystem backup in multiple parallel streams.

e2scan

```
Usage: /usr/sbin/e2scan {-l | -f} [ options ] device-filename
Modes:  -l: list recently changed files
        -f: create file database
Options:
        -a groups: readahead 'groups' inode tables (default 1)
        -b blocks: buffer 'blocks' inode table blocks
        -C chdir: list files relative to 'chdir' in filesystem
        -d database: output database filename (default e2scan.db)
        -D: list not only files, but directories as well
        -n filename: list files newer than 'filename'
        -N date: list files newer than 'date'
           (default 1 day, 0 for all files)
        -o outfile: output file list to 'outfile'
```

llbackup

```
llbackup: backup a list of files, running on multiple nodes
usage: llbackup [-chjvxz] [-C directory] [-e rsh] [-i outputlist]
                [-l logdir] [-n nodes] [-s splitmb] [-T tar] -f outputfile
        -c create archive
        -C directory: relative directory for filenames (default PWD)
        -e rsh: specify the passwordless remote shell (default ssh -q -x)
        -f outputfile: specify base output filename for backup
        -h: print this help message and exit
        -i inputfile: list of files to backup (default stdin)
        -j: use bzip2 compression on output file(s)
        -l logdir: directory for output logs
        -n nodes: comma-separated list of nodes to run backups
        -s splitmb: target size for backup chunks  (default 8192MiB)
        -S splitcount: number of files sent to each client  (default 200)
        -t: list table of contents of tarfile
        -T tar: specify the backup command (default tar)
        -v: be verbose - list all files being processed
        -V: print version number and exit
        -x: extract files instead of backing them up
        -z: use gzip compression on output file(s)
```

Parameters

```
today=`date +%Y%m%d`

CHUNK=51200
FS=u0
MDS=nid04140
MDSDEV=/dev/sda

SSHOPTS="-o StrictHostKeyChecking=no -o ConnectTimeout=15 -q -x "

NODELIST="nid04100 nid04103 nid04104 nid04107 nid04108 nid04111 nid04112
nid04115 nid04119 nid04580 nid04583 nid04584 nid04587 nid04588 nid04591
nid04595"
```

Build the Node List

```
for node in $NODELIST
do
        nd=`ssh $SSHOPTS $node hostname`
        if [ "$nd" != "" ]
        then
                NodeList1="$NodeList1 $nd"
        fi
done

NodeList1=`echo $NodeList1 | sed 's/ /,/g'`
NodeList="$NodeList1,$NodeList1"
```

Perform the Inode Scan

```
echo "Generating backup list"
ssh -q -x $MDS "time /usr/sbin/e2scan -D -l -a 1 -N 0 -o $BACKUPLIST $MDSDEV"
```

Now Backup Files

```
echo "Performing backup..."

$LLBACKUP -C /$FS -f $BACKUPDIR/$FS -i $BACKUPLIST -n "$NodeList" \
          -l $BACKUPDIR -s $CHUNK
```

To Complete the Utility

- Add tuning for your system. *(chunk, files, stripes…)*
- Move backup files to archival storage. *(hpss)*
- Setup backup schedule. *(after a while an incremental will equal a full)*

Room for Improvement

- All backup nodes must be up. Down nodes won't be skipped. Relatively simple test…
- e2scan can be so much more… Add an option to print a one-liner containing the inode contents.
- Documentation

Questions?