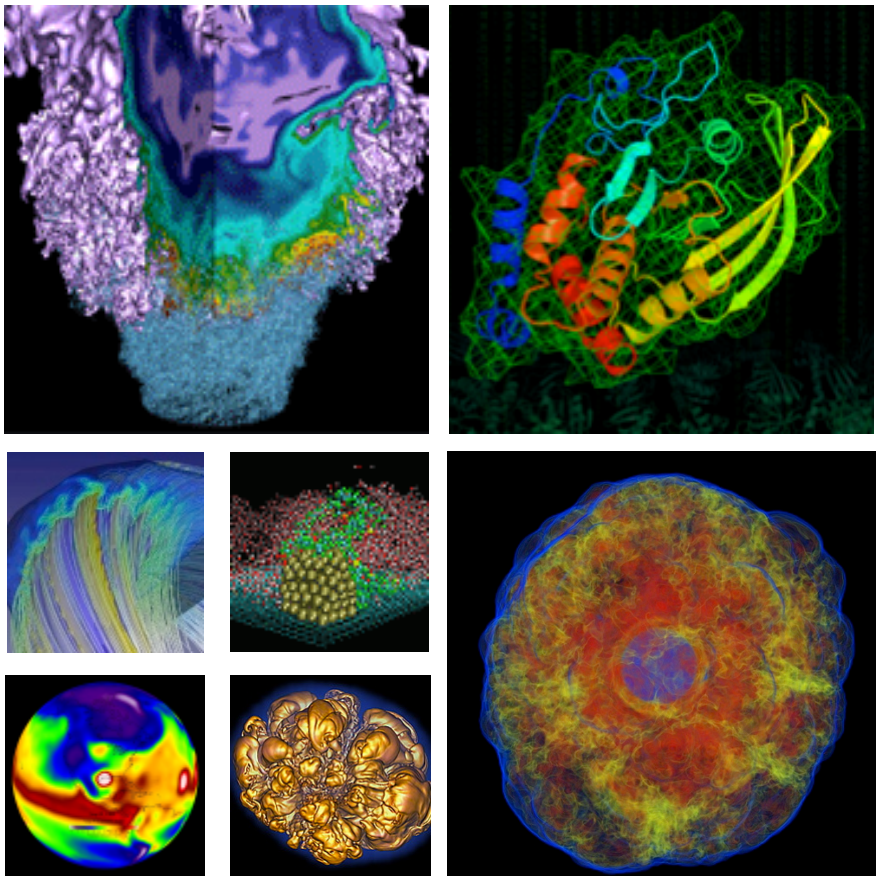
Using Resource Utilization Reporting to Collect DVS Usage Statistics



Tina Butler

NERSC Computational Systems Group

CUG 2014
May 7, 2014

NERSC 40 YEARS at the FOREFRONT 1974-2014

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

Outline

- Historical Cray accounting and utilization
- Resource Utilization Reporting
- The need for a custom plugin
- Design and implementation
- Further work

Historical Cray Accounting

- **Cray System Accounting (CSA)**
  - Available with UNICOS on vector machines
  - Provided job-level and project-level accounting and metrics
    - System and user CPU times
    - Memory highwater and averages
    - Block and character I/O counts
  - Became open-source Comprehensive System Accounting under SGI
  - Still supported with Cray Linux Environment (CLE), but does not scale and not all functions are implemented

Historical Cray Accounting, ctd

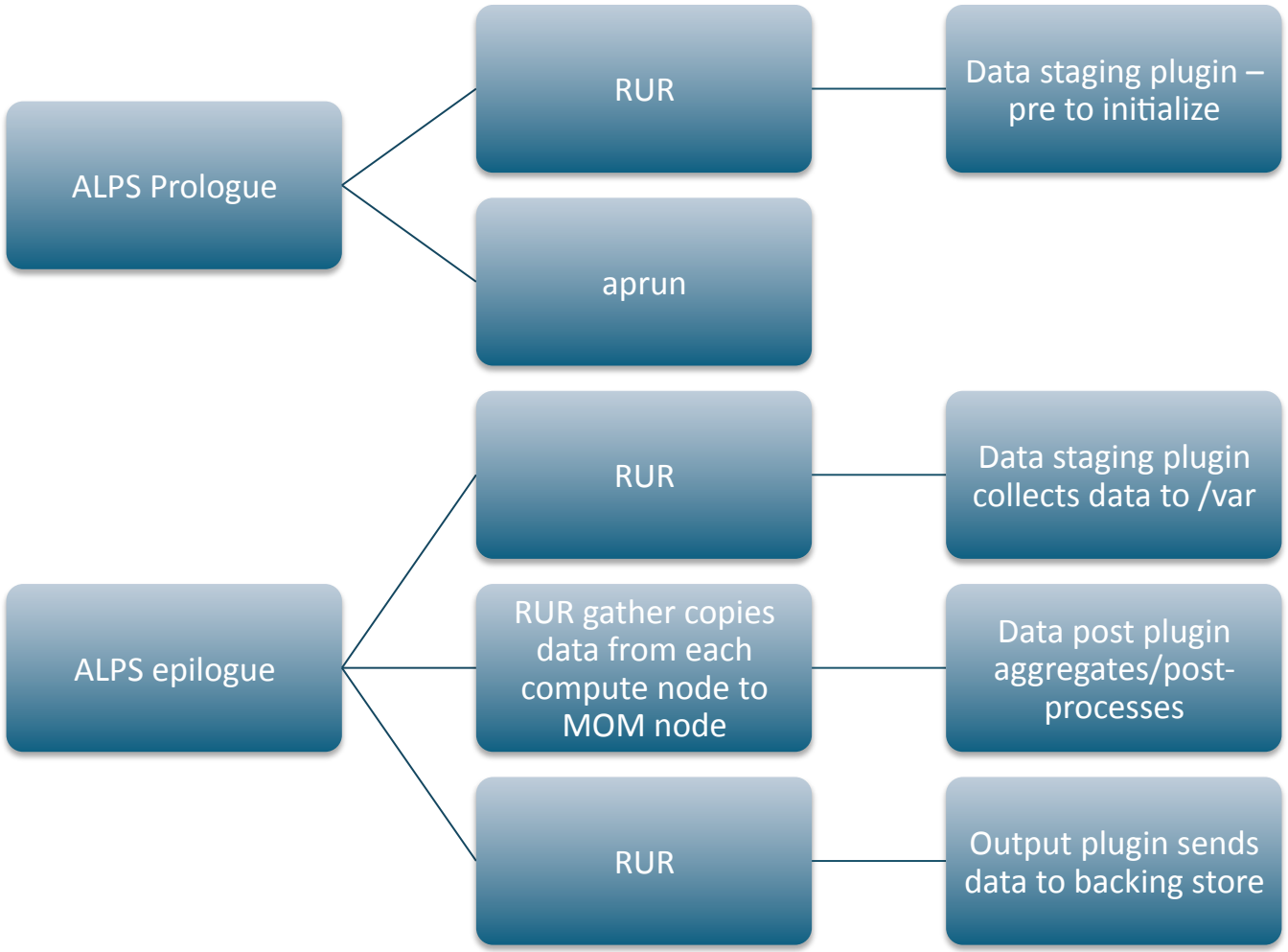
- **Mazama**
  - did not scale well on SMW
- **Application Resource Utilization (ARU)**
  - Released with CLE 4
  - Provides basic process accounting per aprun
  - Integrated with ALPS
  - Not extensible
  - Output to syslog or flat file
  - When aprun terminates with an error, no metrics – hitting wallclock is an error

<150>1 2014-04-17T00:00:05.982308-07:00 c5-0c2s4n3 apsys 19438 p0-20140403t113614 [alps_msgs@34] apid=28108121, Finishing, user=56395, batch_id=7447167.hopque01, exit_code=0, exitcode_array=0, exitsignal_array=0, utime=521, stime=41, maxrss=1425528, inblocks=443257, outblocks=801443, cpus=24, start=Wed Apr 16 23:50:43 2014, stop=Thu Apr 17 00:00:05 2014, cmd=smoothing

Resource Utilization Reporting (RUR)

- **Replacement for ARU**
  - A scalable, flexible and extensible framework for collecting data from compute nodes
  - Features a site-customizable plugin architecture
  - Launched by ALPS prologue and epilogue, but not tightly integrated with ALPS like ARU
  - Written in python, but custom plugins don't have to be in python

RUR Architecture/Workflow



Included Plugins

- **Cray currently provides 3 sets of data plugins**
  - taskstats – basic process accounting, essentially a replacement for ARU; kernel rusage data
  - gpustat – utilization statistics for NVIDIA gpus on XK and XC systems
  - energy - power utilization statistics, XC only
- **Two types of output plugins are provided in CLE 4.2**
  - llm - syslogs RUR output using the Lightweight Log Manager
  - file – writes RUR output to a designated flat file
- **New user output plugin with CLE 5.1**
  - Outputs directly to user directory when environment variable set and plugin enabled.

Installing and configuring RUR

- **RUR is installed by default in CLE, but not enabled**
  - RUR is enabled by adding it to the apsys stanza of /etc/alps.conf (/etc/opt/cray/alps/alps.conf in CLE 5.x)
- **CLE must be configured to use /dsl as default on the compute node**
- **RUR plugins are defined/configured in /etc/opt/cray/rur/rur.conf**
  - Data and output plugins are turned on and off
  - Custom plugins made known to the RUR framework
- **'Managing System Software for the Cray Linux Environment' S-2393**

The need for a custom DVS plugin

- **NERSC gathers utilization metrics from a broad set of sources to characterize system resource usage by user applications.**
- **NERSC provides users cross-platform storage via the GPFS-based NERSC Global Filesystem (NGF).**
  - NGF actually consists of multiple filesystem instances resident on different storage hardware with different block sizes, access, and performance characteristics

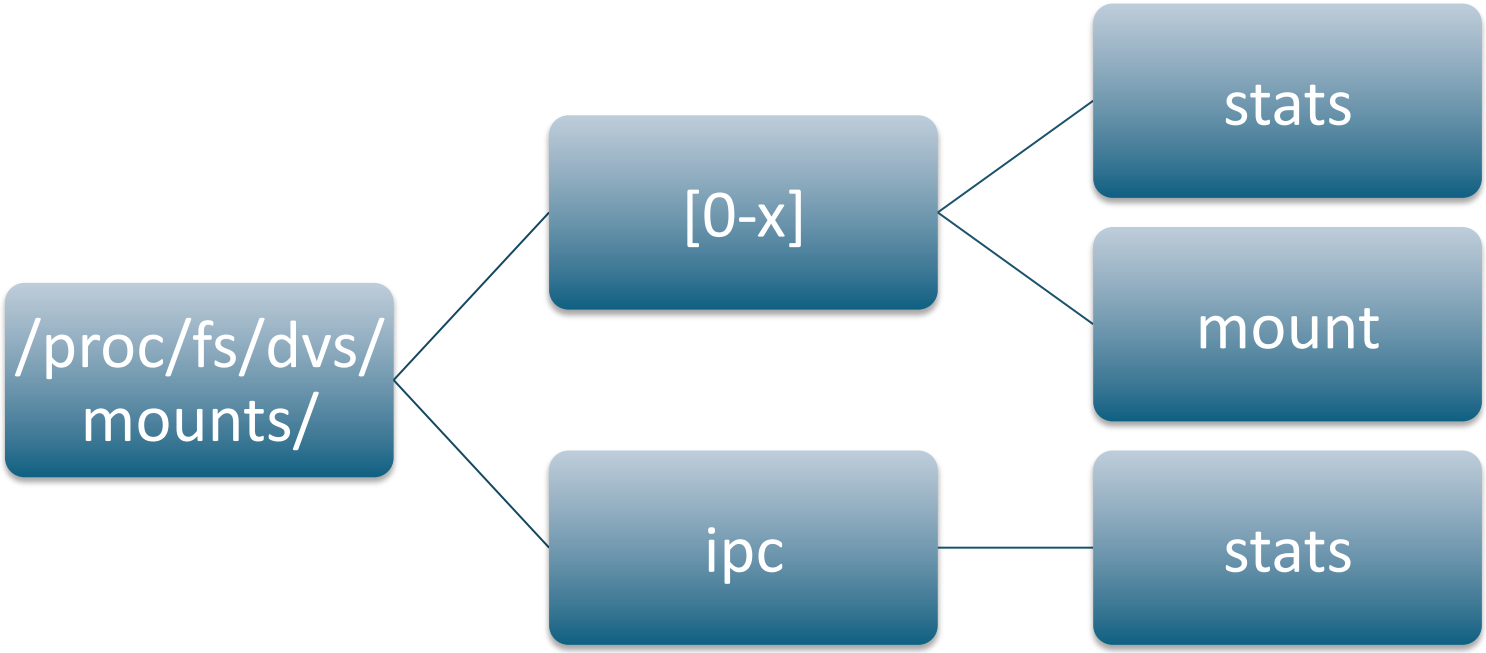
- **/global/syscom, bs=65536**
- **/global/common, bs=65536**
- **/global/u1, bs=131072**
- **/global/u2, bs=131072**
- **/global/dna, bs=1048576**
- **/global/project, bs=4194304, RDMA**
- **/global/projectb, bs=1048576, RDMA**
- **/global/scratch2, bs=8388608, RDMA**

The need for a custom DVS plugin, ctd

- Cray systems access NGF from compute and MOM nodes using the Data Virtualization Service (DVS)
- DVS collects client-side per-mount point request statistics and client-side IPC statistics on compute nodes
- In order to assess NGF usage and performance on a user application level it is desirable to collect DVS client statistics for each NGF mount point
- RUR provides a mechanism for collecting this data

DVS Statistics

- **Statistics are in the /proc filesystem**
- **On clients, stats are collected per mount point**
- **Stats files are initialized by writing '2' to the file**



DVS Client Statistics

Example of /proc/fs/dvs/mounts/[0-n]/stats:

RQ_LOOKUP: 8994092 0	RQ_OPEN: 68151 0
RQ_CLOSE: 68151 0	RQ_READDIR: 23753 0
RQ_CREATE: 698 0	RQ_UNLINK: 337 0
RQ_LSEEK: 0 0	RQ_IOCTL: 0 0
RQ_FLUSH: 0 0	RQ_RELEASE: 0 0
RQ_FSYNC: 0 0	RQ_FASYNC: 0 0
RQ_LOCK: 0 0	RQ_LINK: 0 0
RQ_SYMLINK: 2 0	RQ_MKDIR: 12 0
RQ_RMDIR: 0 0	RQ_MKNOD: 0 0
RQ_RENAME: 37 0	RQ_READLINK: 27312 0
RQ_TRUNCATE: 6 0	RQ_SETATTR: 2074 0
RQ_GETATTR: 313266 0	RQ_PARALLEL_READ: 19034471 0
RQ_PARALLEL_WRITE: 1408148 77	RQ_STATFS: 11 0
RQ_READPAGE_ASYNC: 4555 0	RQ_READPAGE_DATA: 4555 0
RQ_GETEOI: 0 0	RQ_INITFS: 0 0
RQ_SETXATTR: 236 0	RQ_GETXATTR: 49 0
RQ_LISTXATTR: 0 0	RQ_REMOVEXATTR: 0 0
RQ_VERIFYFS: 0 0	RQ_GET_LANE_INFO: 0 0
RQ_RO_CACHE_DISABLE: 0 0	RQ_PERMISSION: 5329 0
read_min_max: 0 4616704	write_min_max: 18388608
IPC requests: 0 0	IPC async requests: 0 0
IPC replies: 0 0	Open files: 0

Example of /proc/fs/dvs/mounts/[0-n]/mount:

local-mount /global/project
remote-path /global/project
options(rw,blksize=4194304,nodename=c3-0c0s4n0:c7-2c2s6n3,nocache,nodatasync,noclosesync,retry,failover,userenv,clusterfs,killprocess,nobulk_rw,noatomic,nodeferopens,no_distribute_create_ops,no_ro_cache,maxnodes=1,nnodes=2,magic=0x47504653)
active_nodes c3-0c0s4n0 c7-2c2s6n3
inactive_nodes
remote-magic 0x47504653

The dvs plugin

- Written in python
- dvs staging plugin –pre zeroes dvs client counters for each mount point
- After the application runs the dvs staging plugin walks the directories /proc/fs/dvs/mounts/[0-x] to collect the contents of the stats and mount files
- Statistics are written to /var/spool/RUR/dvs.*apid* on each compute node
- dvs post, running on the MOM node, copies the compute node data to an aggregate output file and passes it to the RUR framework to pass to output plugins

RUR output from dvs

uid: 18639, apid: 450546, jobid: 14941.grace01.nersc.gov, cmdname: /bin/hostname dvs dvs['/global/
scratch2', ' 4172 0', ' 149 0', ' 149 0', ' 0 0', ' 149 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', '
0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 746 0', ' 0 0', ' 892 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', '
0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 8 3360', ' 0 0', ' 0 0', ' 0 0', ' 0']['/
project', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', '
0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 1 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0',
' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0']['/global/u2', ' 1376 0', ' 0 0', '
0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0',
' 0 0', ' 0 0', ' 0 0', ' 1 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0
0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0']['/global/u1', ' 14343 0', ' 552 0', ' 552 0', ' 0 0', ' 0
0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', '
0 0', ' 553 0', ' 1104 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0
0', ' 0 0', ' 0 177', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0']['/global/common', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0
0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', '
1 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0',
' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0']['/dsl', ' 34035 0', ' 47428 0', ' 47329 0', ' 628 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0',
' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 772 0', ' 0 0', ' 0 0', ' 483 0', ' 0 0',
' 0 0', ' 0 0', ' 16440 0', ' 16440 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 8
32768', ' 0 0', ' 0 0', ' 0 0', ' 0 0', ' 99']

Some comments about RUR

- The RUR gather stage expects a single line file per node
- Output plugins expect to output single line per application
- Currently, output plugins are active for all enabled data plugins, i.e., you can't tie a data plugin to a specific output plugin.
- For debugging, errors are logged to /var/log/apsys on the MOM node for the aprun.

Further work

- **Currently only collecting mount point statistics**
  - Dvs post needs more work
  - IPC stats also desirable
  - Have to assess which IPC data is most useful
- **Incorporate DVS client statistics into the NERSC job completion database**

Acknowledgements

- This work was supported by the Director, Office of Science, Office of Advance Scientific Computing Research of the U.S. Department of Energy under contract No. DEAC02-05CH11231.

References

- Introduction to Cray Data Virtualization Service, S-0005-51-1
- Managing System Software for the Cray Linux Environment, S-2393-4202

Thank you.

RUR output from dvs

uid: 18639, apid: 450525, jobid: 14937.grace01.nersc.gov, cmdname: /bin/hostname dvs dvs['/global/scratch2', ['RQ_LOOKUP: 4172 0', 'RQ_OPEN: 149 0', 'RQ_CLOSE: 149 0', 'RQ_READDIR: 0 0', 'RQ_CREATE: 149 0', 'RQ_UNLINK: 0 0', 'RQ_LSEEK: 0 0', 'RQ_IOCTL: 0 0', 'RQ_FLUSH: 0 0', 'RQ_RELEASE: 0 0', 'RQ_FSYNC: 0 0', 'RQ_FASYNC: 0 0', 'RQ_LOCK: 0 0', 'RQ_LINK: 0 0', 'RQ_SYMLINK: 0 0', 'RQ_MKDIR: 0 0', 'RQ_RMDIR: 0 0', 'RQ_MKNOD: 0 0', 'RQ_RENAME: 0 0', 'RQ_READLINK: 0 0', 'RQ_TRUNCATE: 0 0', 'RQ_SETATTR: 0 0', 'RQ_GETATTR: 746 0', 'RQ_PARALLEL_READ: 0 0', 'RQ_PARALLEL_WRITE: 892 0', 'RQ_STATFS: 0 0', 'RQ_READPAGE_ASYNC: 0 0', 'RQ_READPAGE_DATA: 0 0', 'RQ_GETEOI: 0 0', 'RQ_INITFS: 0 0', 'RQ_SETXATTR: 0 0', 'RQ_GETXATTR: 0 0', 'RQ_LISTXATTR: 0 0', 'RQ_REMOVEXATTR: 0 0', 'RQ_VERIFYFS: 0 0', 'RQ_GET_LANE_INFO: 0 0', 'RQ_RO_CACHE_DISABLE: 0 0', 'RQ_PERMISSION: 0 0', 'read_min_max: 0 0', 'write_min_max: 8 3360', 'IPC requests: 0 0', 'IPC async requests: 0 0', 'IPC replies: 0 0', 'Open files: 0']] ['/project', ['RQ_LOOKUP: 0 0', 'RQ_OPEN: 0 0', 'RQ_CLOSE: 0 0', 'RQ_READDIR: 0 0', 'RQ_CREATE: 0 0', 'RQ_UNLINK: 0 0', 'RQ_LSEEK: 0 0', 'RQ_IOCTL: 0 0', 'RQ_FLUSH: 0 0', 'RQ_RELEASE: 0 0', 'RQ_FSYNC: 0 0', 'RQ_FASYNC: 0 0', 'RQ_LOCK: 0 0', 'RQ_LINK: 0 0', 'RQ_SYMLINK: 0 0', 'RQ_MKDIR: 0 0', 'RQ_RMDIR: 0 0', 'RQ_MKNOD: 0 0', 'RQ_RENAME: 0 0', 'RQ_READLINK: 0 0', 'RQ_TRUNCATE: 0 0', 'RQ_SETATTR: 0 0', 'RQ_GETATTR: 1 0', 'RQ_PARALLEL_READ: 0 0', 'RQ_PARALLEL_WRITE: 0 0', 'RQ_STATFS: 0 0', 'RQ_READPAGE_ASYNC: 0 0', 'RQ_READPAGE_DATA: 0 0', 'RQ_GETEOI: 0 0', 'RQ_INITFS: 0 0', 'RQ_SETXATTR: 0 0', 'RQ_GETXATTR: 0 0', 'RQ_LISTXATTR: 0 0', 'RQ_REMOVEXATTR: 0 0', 'RQ_VERIFYFS: 0 0', 'RQ_GET_LANE_INFO: 0 0', 'RQ_RO_CACHE_DISABLE: 0 0', 'RQ_PERMISSION: 0 0', 'read_min_max: 0 0', 'write_min_max: 0 0', 'IPC requests: 0 0', 'IPC async requests: 0 0', 'IPC replies: 0 0', 'Open files: 0']]['/global/u2', ['RQ_LOOKUP: 0 0', 'RQ_OPEN: 0 0', 'RQ_CLOSE: 0 0', 'RQ_READDIR: 0 0', 'RQ_CREATE: 0 0', 'RQ_UNLINK: 0 0', 'RQ_LSEEK: 0 0', 'RQ_IOCTL: 0 0', 'RQ_FLUSH: 0 0', 'RQ_RELEASE: 0 0', 'RQ_FSYNC: 0 0', 'RQ_FASYNC: 0 0', 'RQ_LOCK: 0 0', 'RQ_LINK: 0 0', 'RQ_SYMLINK: 0 0', 'RQ_MKDIR: 0 0', 'RQ_RMDIR: 0 0', 'RQ_MKNOD: 0 0', 'RQ_RENAME: 0 0', 'RQ_READLINK: 0 0', 'RQ_TRUNCATE: 0 0', 'RQ_SETATTR: 0 0', 'RQ_GETATTR: 1 0', 'RQ_PARALLEL_READ: 0 0', 'RQ_PARALLEL_WRITE: 0 0', 'RQ_STATFS: 0 0', 'RQ_READPAGE_ASYNC: 0 0', 'RQ_READPAGE_DATA: 0 0', 'RQ_GETEOI: 0 0', 'RQ_INITFS: 0 0', 'RQ_SETXATTR: 0 0', 'RQ_GETXATTR: 0 0', 'RQ_LISTXATTR: 0 0', 'RQ_REMOVEXATTR: 0 0', 'RQ_VERIFYFS: 0 0', 'RQ_GET_LANE_INFO: 0 0', 'RQ_RO_CACHE_DISABLE: 0 0', 'RQ_PERMISSION: 0 0', 'read_min_max: 0 0', 'write_min_max: 0 0', 'IPC requests: 0 0', 'IPC async requests: 0 0', 'IPC replies: 0 0', 'Open files: 0']]['/global/u1', ['RQ_LOOKUP: 14295 0', 'RQ_OPEN: 552 0', 'RQ_CLOSE: 552 0', 'RQ_READDIR: 0 0', 'RQ_CREATE: 0 0', 'RQ_UNLINK: 0 0', 'RQ_LSEEK: 0 0', 'RQ_IOCTL: 0 0', 'RQ_FLUSH: 0 0', 'RQ_RELEASE: 0 0', 'RQ_FSYNC: 0 0', 'RQ_FASYNC: 0 0', 'RQ_LOCK: 0 0', 'RQ_LINK: 0 0', 'RQ_SYMLINK: 0 0', 'RQ_MKDIR: 0 0', 'RQ_RMDIR: 0 0', 'RQ_MKNOD: 0 0', 'RQ_RENAME: 0 0', 'RQ_READLINK: 0 0', 'RQ_TRUNCATE: 0 0', 'RQ_SETATTR: 0 0', 'RQ_GETATTR: 553 0', 'RQ_PARALLEL_READ: 1104 0', 'RQ_PARALLEL_WRITE: 0 0', 'RQ_STATFS: 0 0', 'RQ_READPAGE_ASYNC: 0 0', 'RQ_READPAGE_DATA: 0 0', 'RQ_GETEOI: 0 0', 'RQ_INITFS: 0 0', 'RQ_SETXATTR: 0 0', 'RQ_GETXATTR: 0 0', 'RQ_LISTXATTR: 0 0', 'RQ_REMOVEXATTR: 0 0', 'RQ_VERIFYFS: 0 0', 'RQ_GET_LANE_INFO: 0 0', 'RQ_RO_CACHE_DISABLE: 0 0', 'RQ_PERMISSION: 0 0', 'read_min_max: 0 177', 'write_min_max: 0 0', 'IPC requests: 0 0', 'IPC async requests: 0 0', 'IPC replies: 0 0', 'Open files: 0']] ['/global/common', ['RQ_LOOKUP: 0 0', 'RQ_OPEN: 0 0', 'RQ_CLOSE: 0 0', 'RQ_READDIR: 0 0', 'RQ_CREATE: 0 0', 'RQ_UNLINK: 0 0', 'RQ_LSEEK: 0 0', 'RQ_IOCTL: 0 0', 'RQ_FLUSH: 0 0', 'RQ_RELEASE: 0 0', 'RQ_FSYNC: 0 0', 'RQ_FASYNC: 0 0', 'RQ_LOCK: 0 0', 'RQ_LINK: 0 0', 'RQ_SYMLINK: 0 0', 'RQ_MKDIR: 0 0', 'RQ_RMDIR: 0 0', 'RQ_MKNOD: 0 0', 'RQ_RENAME: 0 0', 'RQ_READLINK: 0 0', 'RQ_TRUNCATE: 0 0', 'RQ_SETATTR: 0 0', 'RQ_GETATTR: 1 0', 'RQ_PARALLEL_READ: 0 0', 'RQ_PARALLEL_WRITE: 0 0', 'RQ_STATFS: 0 0', 'RQ_READPAGE_ASYNC: 0 0', 'RQ_READPAGE_DATA: 0 0', 'RQ_GETEOI: 0 0', 'RQ_INITFS: 0 0', 'RQ_SETXATTR: 0 0', 'RQ_GETXATTR: 0 0', 'RQ_LISTXATTR: 0 0', 'RQ_REMOVEXATTR: 0 0', 'RQ_VERIFYFS: 0 0', 'RQ_GET_LANE_INFO: 0 0', 'RQ_RO_CACHE_DISABLE: 0 0', 'RQ_PERMISSION: 0 0', 'read_min_max: 0 0', 'write_min_max: 0 0', 'IPC requests: 0 0', 'IPC async requests: 0 0', 'IPC replies: 0 0', 'Open files: 0']]['/dsl', ['RQ_LOOKUP: 20229 0', 'RQ_OPEN: 31121 0', 'RQ_CLOSE: 31024 0', 'RQ_READDIR: 356 0', 'RQ_CREATE: 0 0', 'RQ_UNLINK: 0 0', 'RQ_LSEEK: 0 0', 'RQ_IOCTL: 0 0', 'RQ_FLUSH: 0 0', 'RQ_RELEASE: 0 0', 'RQ_FSYNC: 0 0', 'RQ_FASYNC: 0 0', 'RQ_LOCK: 0 0', 'RQ_LINK: 0 0', 'RQ_SYMLINK: 0 0', 'RQ_MKDIR: 0 0', 'RQ_RMDIR: 0 0', 'RQ_MKNOD: 0 0', 'RQ_RENAME: 0 0', 'RQ_READLINK: 449 0', 'RQ_TRUNCATE: 0 0', 'RQ_SETATTR: 0 0', 'RQ_GETATTR: 347 0', 'RQ_PARALLEL_READ: 0 0', 'RQ_PARALLEL_WRITE: 0 0', 'RQ_STATFS: 0 0', 'RQ_READPAGE_ASYNC: 9956 0', 'RQ_READPAGE_DATA: 9956 0', 'RQ_GETEOI: 0 0', 'RQ_INITFS: 0 0', 'RQ_SETXATTR: 0 0', 'RQ_GETXATTR: 0 0', 'RQ_LISTXATTR: 0 0', 'RQ_REMOVEXATTR: 0 0', 'RQ_VERIFYFS: 0 0', 'RQ_GET_LANE_INFO: 0 0', 'RQ_RO_CACHE_DISABLE: 0 0', 'RQ_PERMISSION: 0 0', 'read_min_max: 8 32768', 'write_min_max: 0 0', 'IPC requests: 0 0', 'IPC async requests: 0 0', 'IPC replies: 0 0', 'Open files: 98']]