

HPSS tools developed at ECMWF

F. Dequen

May 2004

francis.dequen@ecmwf.int

ECMWF Monitoring Suite

- **ECMWF has developed many years ago a monitoring suite allowing operators and analysts to keep watch on its data handling systems behaviour.**
 - **Initially used to aggregate information from errpt, file system usage, TSM activity.**
 - **Now include information from HPSS.**
- Uses whpss from Tim Starrin (Big thank you to him).**
- Use several in house tools.**

Query_hpss_status

- Use hpssadm to obtain status of the SCs, and see how much data is being transmitted on devices.
- Compare the HPSS view of tape drives with the robot view (e.g. acsls).
- Extrapolate from these the bandwidth seen on the tape and disk devices.
- Allow to detect slow rates, idle drives, discrepancies between robot and hpss view.
- It has been updated to integrate some new 5.1 hpssadm feature (e.g. a much more powerful “device list” command)

→ Thank you, Vicky and Deryl.

HPSS Tools, May 2004 3



Query_hpss_output

HPSS Tape Drives report			Wed Jun 04 12:09:39 GMT+00:00 2003								
Drive Aix	Id	device name	Volume	Read MB/s	Write MB/s	Mover (Device)	Admin St	Oper. St	Drive State	Robot Status	Volume
	2100	/dev/rmt2100	F0972900	0.000	0.000	Unlocked	Enabled	Enabled	Enabled	online in use	F09729
	2101	/dev/rmt2101	F1109100	0.000	0.000	Unlocked	Enabled	Enabled	Enabled	online in use	F11091
	2102	/dev/rmt2102		0.000	0.000	Unlocked	Enabled	Disabled	Disabled		
	2103	/dev/rmt2103		0.000	0.000	Unlocked	Enabled	Disabled	Disabled		
	2200	/dev/rmt2200	F1115300	0.000	0.000	Unlocked	Enabled	Enabled	Enabled	online in use	F11153
	2201	/dev/rmt2201	F1503100	0.000	10.917	Unlocked	Enabled	Enabled	Enabled	online in use	F15031
	2202	/dev/rmt2202	F1020900	0.000	0.000	Unlocked	Enabled	Enabled	Enabled	online in use	F10209
	2203	/dev/rmt2203	F1086300	17.917	0.000	Unlocked	Enabled	Enabled	Enabled	online in use	F10863
	3001	/dev/rmt3001	M0005500	0.000	17.401	Unlocked	Enabled	Enabled	Enabled	UP	M00055
	3002	/dev/rmt3002	M0076300	0.000	2.339	Unlocked	Enabled	Enabled	Enabled	UP	M00763

- The version 5 output is a bit different, but includes info about disks.
- A new sort field allows to display drives by user defined categories (e.g. drives in one silo, disks for a given SC...)

HPSS Tools, May 2004 4



Interpret_gk_log

- Use the gatekeeper to track opening and closing of bitfiles.
- Convert the bitfile_id of these bitfiles into a file name.
- Display a list of the opened files.

```
Bitfiles currently in use.
=====
Monitoring started on 20030604:115240                               Number of files opened:  8

File Name  UID                               Host      Time Opened
=====
./marse4mnth/1/fc/19730100/sfc/37295.20030524.123706 marser  hdrv06    20030604:11531
./marse4mnth/1/an/19570900/sfc/37052.20030522.005317 marser  hdrv06    20030604:11524
/marsrdenfo/ec74/pf/20030213/sfc/645367.20030604.115444.tmp marsrd  hdrg04    20030604:11544
marsrdseas/ed9t/fc/20011001/pt/1/644355.20030604.115455.tmp marsrd  hdrg04    20030604:11545
./marse4wave/1/an/19640201/sfc/37415.20030522.212918 marser  hdrv06    20030604:11524
./marsrdenfo/edgh/cf/19530127/pt/644341.20030604.115447.tmp marsrd  hdrg04    20030604:11545
./marse4oper/1/an/19641001/pl/33455.20030510.082353 marser  hdrv06    20030604:11535
./marsodenfo/12/cf/20020414/sfc/124099.20030604.115316.tmp marsod  athos5-ge 20030604:11532
```

pvljob

- Ever been in a situation where you want to know what is using this tape, or that drive, or why some user requests are waiting for access to a tape which seems to be mounted?
- Have you ever seen PVL jobs for which tapes are mounted, but no activity seems to take place for long period of time?
- Solution:
 - Use the gui or hpssadm to find out the links between drives, tape volumes and pvl jobs.
 - Fastidious, long, no obvious differentiation between disk/tape pvl jobs.
 - Use ECMWF's pvljob tool.

pvljob

- pvljob gives a list of pvl jobs, the volumes these use or request, and which drives are allocated to the jobs.
- By default, only tape related jobs are displayed.
- List can be ordered by tape, pvid, drive, or start time.

pvljob [-p | -v | -d | -t | -c] [-a]

```
pvljob -v
```

PVL Job	Volume	Drive	Creation time	Commit Time
831226	B1218500	3002	2004-04-14 15:50:12	2004-04-14 15:50:12
831227	B1218600	3003	2004-04-14 15:50:18	2004-04-14 15:50:18
812760	C2101300	2375	2004-04-12 00:02:40	1970-01-01 00:00:00
812760	C2101700	2378	2004-04-12 00:02:40	1970-01-01 00:00:00
831329	C2116000	2379	2004-04-14 16:03:56	2004-04-14 16:03:56

HPSS Tools, May 2004 7



Disk_allocation_stats

- A couple of incidents where a SC is less than 90% full, but files can not be stored in it anymore.
- High purge Treshold was not reached, no purges taking place.
 - Very stable situation, with access denial for hours.
- Issue was disk fragmentation.
- We recovered by lowering the high purge tresholds and allowing files to span on more segments.
- During investigations, we created a tool to show the list of free extends on a disk or in an SC.

HPSS Tools, May 2004 8



Disk_allocation_stat

```
disk_allocation_stat [-s small] [-l large ] [-x] [-d]
                    {-v volume | -c scid [-D] } -i subsys
```

```
disk_allocation_stat -i 3 -s 8 -l 40 -c 1331
```

Summary info for Class 1331

```
Clustersize (MB)      : 0.5
Volume size (MB)      : 262144
Total in use (cluster) : 1281503
Total in use (MB)     : 640751.5
Percentage in use     : 81.48%
Total free (cluster)  : 291361
Total free (MB)       : 145680.5
Number of free clusters: 12996
Percentage free       : 18.52%
Avg free size (MB)   : 11.2096414281317
Largest cluster (MB) : 6057
```

Free extents <= 8 MB

```
Number of extents    : 10007
Size (MB)            : 29170.5
Percentage of volume(s): 3.71%
```

Free extents >= 40 MB

```
Number of extents    : 370
Size (MB)            : 65584.5
Percentage of volume(s): 8.34%
```

HPSS Tools, May 2004 9



Disk_allocation_stat

```
disk_allocation_stat -i 3 -s 12 -l 512 -v E64A8200 -d
```

```
Volume E64A8200      Storage Class 1341
1st extent length(xt)  nxt extent length (MB)  type volume      SCid Fl
=====
58709      6827      65536      27308.000 Free E64A8200      1341 L
55257      3068      58325      12272.000 Free E64A8200      1341 L
53954      563       54517      2252.000 Free E64A8200      1341 L
54598      531       55129      2124.000 Free E64A8200      1341 L
52582      110       52692      440.000 Free E64A8200      1341
53701      98        53799      392.000 Free E64A8200      1341
52088      58        52146      232.000 Free E64A8200      1341
53015      53        53068      212.000 Free E64A8200      1341
```

HPSS Tools, May 2004 10



Release_knots / count_knots

- Are we keeping data on disk for as long as we would like?
- What is the impact of adding some disk space to a SC?
- Release_knots / count_knots give a feel about the “disk survival age” of files in cache.

Count_knots

```
count_knots /etc/knots/input
Report time: 2004/04/14 16:02:50
```

<u>COS</u>	<u>First disk copy</u>	<u>Age-days</u>	<u>Last tape-only</u>
301	20040122-174229	82.93	
311	20040122-174229	82.93	
321	20040213-004900	61.63	20040212-004901
331	20040309-124901	36.13	20040309-004901
341	20040408-124901	6.13	20040408-004902
401	20040122-174229	82.93	
411	20040122-174229	82.93	
421	20040213-004900	61.63	20040212-004901
431	20040309-124901	36.13	20040309-004901
441	20040408-124901	6.13	20040408-004902

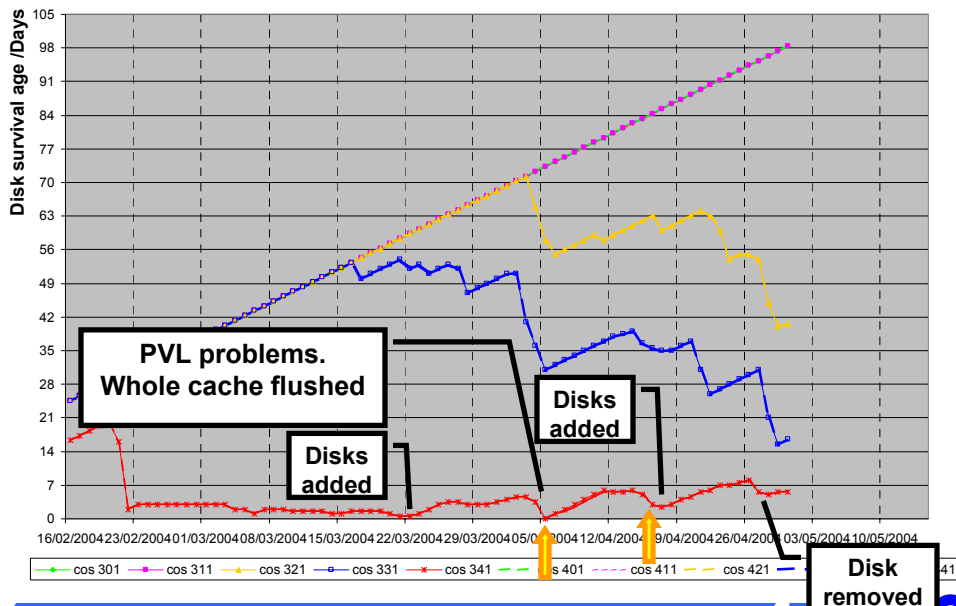
Release_knots/Count_knots

- Implementation:

- We create every 12 hours small files (or knots) in each of the COSes that we want to monitor. (release_knots).
- Using scrub, we check the status of these files, whether they are still on disk or not, and generate a small report (count_knots).

- Assumes that purge is done by last access time.

Release/Count knots



Migration performances reporting

- Our experience with disk-to-tape hierarchies is still limited.
- We are still tuning our environment.
- One possible issue may be the performances of migration/purge.
- Mps_reporter provides good information about this
 - A lot of information, difficult to get an “at a glance” feel.
 - Only provides throughput info for completed migrations.
- Mps_report_formatter addresses these issues to some extend.

mps_report_formatter

```
cat /var/hpss/mps/mps3/mps3.20040427 |mps_reporter|mps_report_formatter -s -c 1331

2004/04/27 01:47:28 1083030448 Disk Purge Summary <Succeeded> 1331 Total
39432 MB Speed 938.869MB/s Bitfiles 1202/ 1216

2004/04/27 02:25:34 1083032734 Disk Migration Summary <Succeeded> 1331 Total
69068 MB Speed 11.158MB/s Bitfiles 2298/ 2367

...

2004/04/27 14:30:36 1083076236 Disk Purge Summary <Succeeded> 1331 Total
39890 MB Speed 664.833MB/s Bitfiles 1623/ 1632

Total data Migrated for SC 1331 169787.397 MB
Total data Purged for SC 1331 197937.500 MB
```

mps_report_formatter

- Without `-s` parameter, additional information provided for individual file copy.
- Estimation of the throughput rates for the complete SC, complete thread, last file processed.

```
cat /var/hpss/mps/mps3/mps3.20040427 |mps_reporter|mps_report_formatter
```

Time		Time (sec)	Operation	Result
SC	Thread	MB	Total MB (SC)	SC MB/s th MB/s Sht MBs
2004/04/27 00:09:20		1083024560	Disk Migration for a Bitfile	<Succeeded>
1341	00002d4e	84.415	10195.057	18.205 5.843 3.837
2004/04/27 00:09:22		1083024562	Disk Migration for a Bitfile	<Succeeded>
1341	00002b4c	67.365	10262.422	18.261 10.077 9.624

Migration performances reporting

- Sometimes, the information provided by `mps_reporter` is not sufficient, and core server tracing need to be turned on.
- A lot of information is generated in these conditions.
- `interpret_delog_migration` and `interpret_delog_migration_post` extract from the `delog` output some useful information related to the migration, and reports
 - Interesting events. (Tape mount requests and completion, start and end of processing of a given bitfile migration)
 - Level of migration parallelism. (number of concurrent migration processes, by SC)

interpret_delog_migration

elaps	Date Time	Request id	1301	1311	1321	1331	1341	Migration	Event		
Details											
time	th tp th tp th tp th tp th tp										
=====											
6385	04/27 16:14:05	1486278673	0	0	0	2	1	2	0	1321 --> 398	process file
	/ecfs/op ./oparch/qscat/science/L2B/2000/216/QS_S2B05853.20002171233										
6385	04/27 16:14:05	1486268049	0	0	0	2	1	2	1	4 0 1331 --> 398	Tape mount request
6385	04/27 16:14:05	1486278617	0	0	0	2	0	2	1	4 0 1321 --> 398	Tape mount done
	82 secs										
6386	04/27 16:14:06	1486268053	0	0	0	2	0	1	1	4 0 1331 --> 398	end_file 1331 --> 398
	Rate: 6.909 Size: 17280000 Mount:None Posit: 0 Total: 2 sec										
6387	04/27 16:14:07	1486278673	0	0	0	1	0	1	1	4 0 1321 --> 398	end_file 1321 --> 398
	Rate: 3.422 Size: 8215365 Mount:None Posit: 0 Total: 2 sec										