ARC cache for ATLAS analysis jobs

Mattias Wadenstein

Hepix 2010 Fall Meeting

2010-11-03, Cornell

ARC basics

- Jobs come in with a list of input and output
  - Pairs of local name and URL
  - (“input.data” “srm://srm.ndgf.org/data/1”)
  - The session dir (job's cwd) will have “input.data” available for reading
- Input files are cached based on the URL
  - Unless requested otherwise
  - Some verification made before use

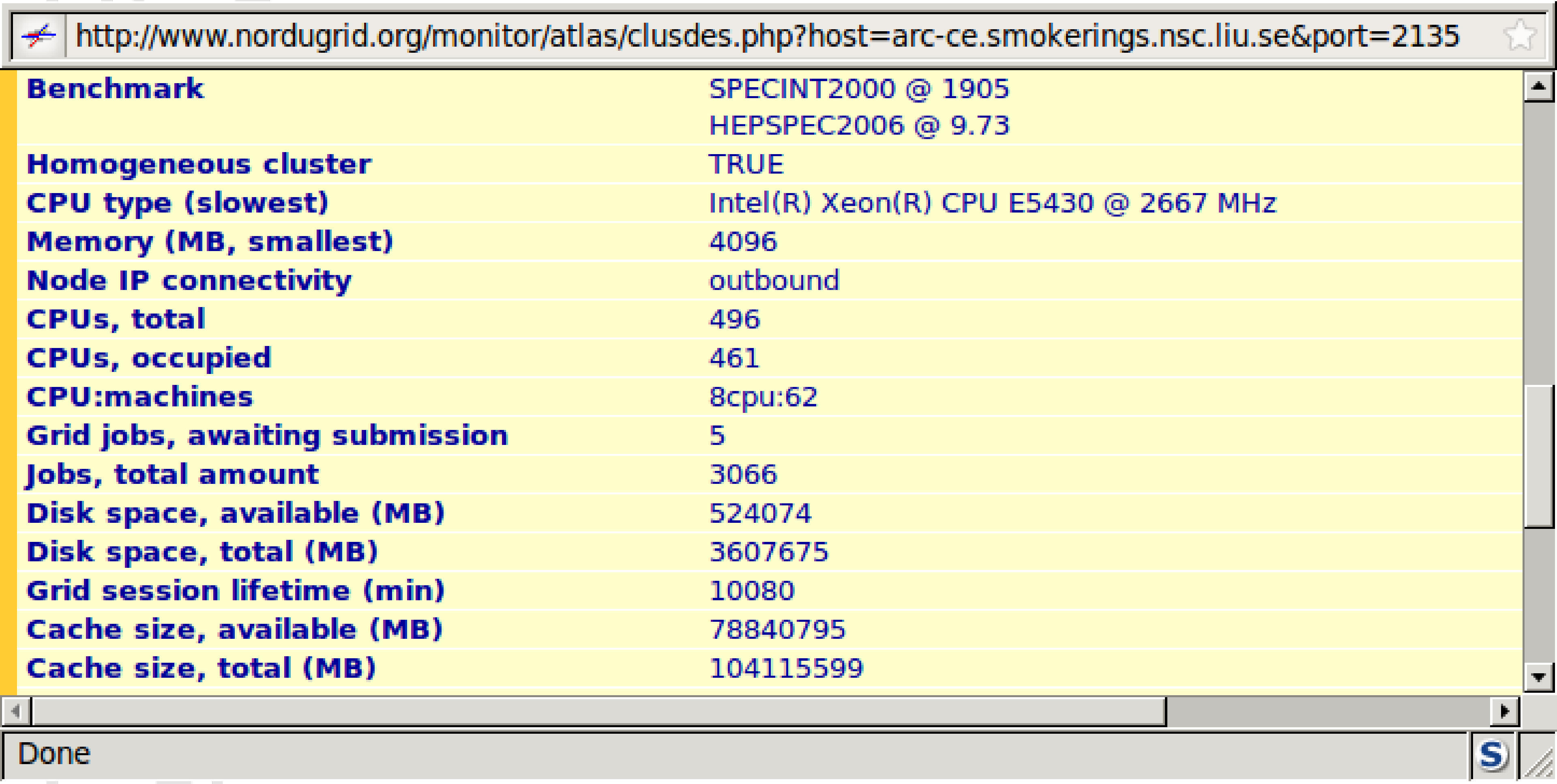
Cache design

- A set of shared filesystems
  - Lustre, GPFS and NFS in common use
  - For NFS deployment, one filesystem per raidset and server makes sense
  - Cleaned LRU by atime
    - Doesn't need to be exact, lazy atime in GPFS etc works fine
    - Cleaned by stand-alone script, can be run from cron on storage servers
- ACIX publishes cache content (hash)

In practice

- Most Atlas tier2 storage in the NDGF cloud is cache (or up-shifted to t1 storage)
- About 1PB cache total in the cloud
  - Roughly 100TB at each cluster
  - Those clusters support ~500 concurrent jobs
- 90% hitrate in ACIX (with no pre-staging)
- Average 3 replicas per file
  - skewed by files like DBRelease, pilot tarball in all jobs and on all sites
- A 120TB cache took 2 months to fill

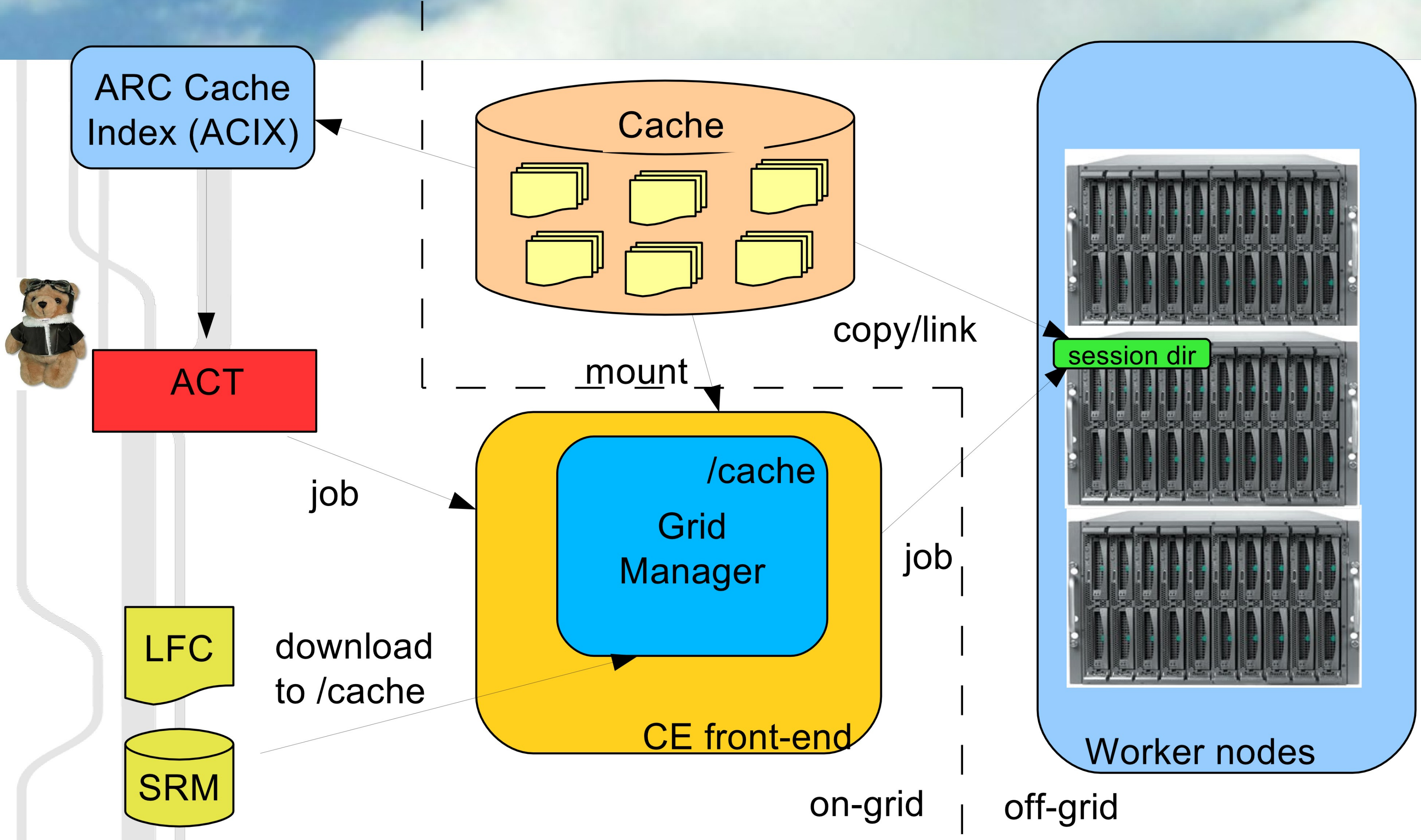
Cache in infosys



http://www.nordugrid.org/monitor/atlas/clusdes.php?host=arc-ce.smokerings.nsc.liu.se&port=2135

Benchmark	SPECINT2000 @ 1905 HEPSPEC2006 @ 9.73
Homogeneous cluster	TRUE
CPU type (slowest)	Intel(R) Xeon(R) CPU E5430 @ 2667 MHz
Memory (MB, smallest)	4096
Node IP connectivity	outbound
CPUs, total	496
CPUs, occupied	461
CPU:machines	8cpu:62
Grid jobs, awaiting submission	5
Jobs, total amount	3066
Disk space, available (MB)	524074
Disk space, total (MB)	3607675
Grid session lifetime (min)	10080
Cache size, available (MB)	78840795
Cache size, total (MB)	104115599

ARC and ATLAS



ACT Cache-based brokering

- For each job create weighted list of clusters
  - Then try clusters in order until job is accepted
- Weight is #files already cached times a random number in [0,1) range
  - Random makes not all tasks go to one cluster with popular files already cached
  - And jobs to new clusters with empty caches
  - Obvious tunable for more throughput etc

ATLAS comments

- Atlas production manager comments:

When large analysis tasks are submitted (1k jobs with 20TB input each), it usually turns out the same dataset is used many times by the same user and many users do the analysis on the same dataset.

So, for the first large task, the inputs are downloaded directly, but then they are reused many times. The reason for this is simple: when a large MC task is finished, many users try to use it before it is replicated to other clouds.

Many user also run reconstruction jobs on ESDs which are not replicated at all.

– Andrej Filipcic

Cache statistics – 30TB cluster

```
Usage statistics: /export/jcache02-fs02/data
Total deletable files found: 3287 (11 files locked)
Total size of deletable files found: 1 TB (3 GB locked)
Used space on file system: 1 TB / 1 TB (70.05%)
At size (% of total)  Newest file               Oldest file
129 GB (10%)          Tue Nov  2 15:58:46 2010  Mon Nov  1 16:47:11 2010
258 GB (20%)          Mon Nov  1 16:40:31 2010  Mon Nov  1 07:18:29 2010
381 GB (30%)          Mon Nov  1 07:18:18 2010  Mon Nov  1 04:48:30 2010
510 GB (40%)          Mon Nov  1 04:38:26 2010  Mon Nov  1 03:21:42 2010
635 GB (50%)          Mon Nov  1 03:21:13 2010  Mon Nov  1 00:52:31 2010
764 GB (60%)          Mon Nov  1 00:50:12 2010  Sun Oct 31 22:48:30 2010
886 GB (70%)          Sun Oct 31 22:47:19 2010  Sat Oct 30 22:09:01 2010
1014 GB (80%)         Sat Oct 30 22:09:00 2010  Fri Oct 29 17:58:14 2010
1 TB (90%)            Fri Oct 29 17:57:54 2010  Thu Oct 28 20:16:47 2010
1 TB (100%)           -                         Thu Oct 28 15:06:04 2010
```

Cache statistics – 100TB cluster

```
Usage statistics: /arc/cache/c1
Total deletable files found: 17023 (2906 files locked or in use)
Total size of deletable files found: 2 TB (134 GB locked or in use)
Used space on file system: 3 TB / 17 TB (17.32%)
At size (% of total)  Newest file               Oldest file
300 GB (10%)          Mon Nov  1 14:08:15 2010  Sat Oct  2 23:14:17 2010
600 GB (20%)          Sat Oct  2 23:12:48 2010  Tue Sep 28 00:49:05 2010
900 GB (30%)          Tue Sep 28 00:42:32 2010  Wed Sep 22 19:37:15 2010
1 TB (40%)            Wed Sep 22 19:36:31 2010  Wed Sep 22 11:15:39 2010
1 TB (50%)            Wed Sep 22 11:13:53 2010  Wed Sep 22 06:05:50 2010
1 TB (60%)            Wed Sep 22 06:02:49 2010  Tue Sep 21 12:25:59 2010
2 TB (70%)            Tue Sep 21 11:22:50 2010  Mon Sep 20 14:49:15 2010
2 TB (80%)            Mon Sep 20 14:48:45 2010  Sun Sep 19 23:10:53 2010
2 TB (90%)            Sun Sep 19 22:44:10 2010  Sun Sep 19 13:17:58 2010
2 TB (100%)           Sun Sep 19 13:16:51 2010  Thu Sep 16 19:39:35 2010
```

Questions

Questions?