Journal of Physics: Conference Series **331** (2011) 072064 doi:10.1088/1742-6596/331/7/072064

A Job Monitoring and Accounting Tool for the LSF Batch System

Subir Sarkar† and Sonia Taneja
Scuola Normale Superiore, Pisa & INFN, Sezione di Pisa, Italy

Abstract. This paper presents a web based job monitoring and group-and-user accounting tool for the LSF Batch System. The user oriented job monitoring displays a simple and compact quasi real-time overview of the batch farm for both local and Grid jobs. For Grid jobs the Distinguished Name (DN) of the Grid users is shown. The overview monitor provides the most up-to-date status of a batch farm at any time. The accounting tool works with the LSF accounting log files. The accounting information is shown for a few pre-defined time periods by default. However, one can also compute the same information for any arbitrary time window. The tool already proved to be an extremely useful means to validate more extensive accounting tools available in the Grid world. Several sites have already been using the present tool and more sites running the LSF batch system have shown interest. We shall discuss the various aspects that make the tool essential for site administrators and end-users alike and outline the current status of development as well as future plans.

1. Introduction

A web based job and group-and-user accounting monitor for the LSF Batch system has been developed for both local and Grid jobs. Both site administrators and end-users will find the monitor useful in spotting problems with jobs at the earliest so that CPU time and hence resource utilisation can be optimised. Moreover, it offers an independent way to validate more extensive monitoring and accounting tools available in the Grid world.

The monitor provides the most up-to-date overview of the state of a batch farm at any time. The user oriented view shows a compact status summary of the local batch farm. For Grid jobs, the Distinguished Name (DN) of the Grid certificate is used to identify users uniquely. The accounting tool solely depends on the LSF accounting log files and does not need any external database support. The accounting monitor primarily shows information for a few pre-defined time intervals. However one can also get the same information for any arbitrary time window using a command line tool in order to prepare monthly reports etc. It might be noted that accounting information for recently completed jobs, e.g. jobs completed in the last 12 hours or so, can be reliably obtained only with such a local tool.

A few sites have been using the present tool and more sites running the LSF batch system have shown interest. We shall discuss the various features that make the monitor indispensable for site administrators and end-users alike and outline the current status of development and plans for the near future.

† Corresponding author, email : subir.sarkar@cern.ch

2. Objective and Monitorable

The monitoring system has been built with the aim to publish at regular intervals the most reliable data on availability of resources, current usage, job flow, usage history, job efficiency, and share utilisation at a batch system. We describe below the various monitorable quantities already available in the present monitor:

- Overview - broadly classified as (a) CPU resource availability, usage, and user share, and (b) job status at the batch farm. To suit both site administrators and end-users, in addition to overall job status we also collect information for each (1) group/Virtual Organisation (VO), (2) Computing Element, and (3) user/Distinguished Name (DN). Each individual sub-set shows total, running, pending, and held jobs in the queue, *average CPU efficiency*, defined as a ratio of total CPU time over total wall-time, number of jobs with very low CPU efficiency and *jobflow*, an indicator of farm activity defined as the number of submitted, dispatched, and completed jobs in the last hour. For groups we also show the *wall-time share*, which is a ratio of the total wall-time used by a group over the overall wall-time used at the farm at the time when information is collected.
- History - time series plots for (a) resource availability, and occupancy at the farm, and (b) number of running and pending jobs, and CPU efficiency for the whole farm as well as supported groups/VOs.
- Accounting - group-and-user accounting shows important quantities like the (a) job share and success rate, (b) CPU usage and efficiency, (c) wall-time share, and (d) average time in seconds to wait in queue etc. for different time intervals for groups/VOs and individual users.

3. Description

The LSF job and accounting monitor implements a simple design shown in Figure 1.

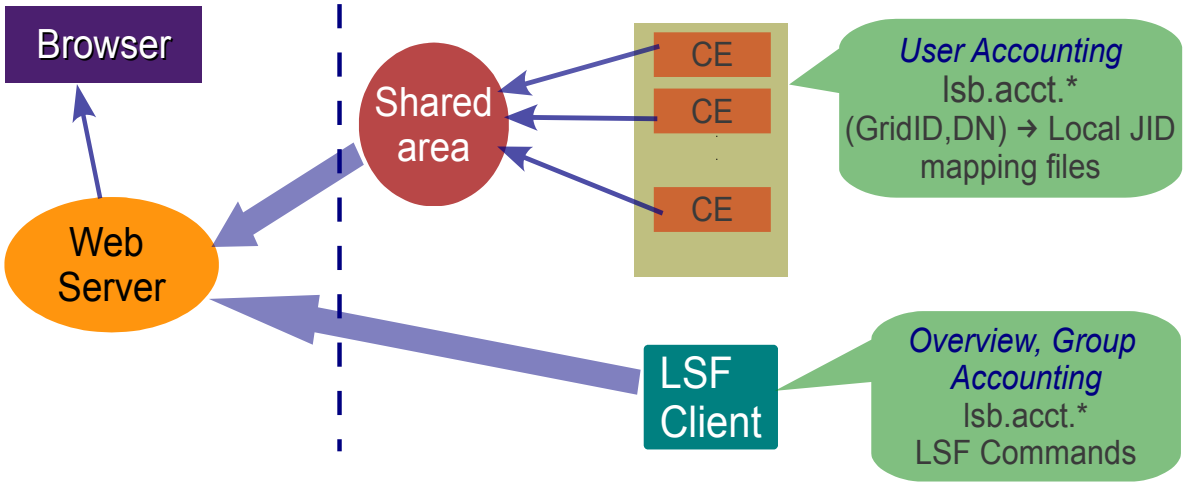


Figure 1. Schematic view of the monitoring framework.

The information collector consists of three distinct components: (1) a sensor that uses the LSF commands and Grid job definition files to get the relevant farm and job overview data, (2) a simple parser that reads the LSF accounting files to collect the group accounting information for pre-defined time periods, and (3) a parser that correlates LSF accounting data with the user DN to get the user accounting statistics. The sensors must run on services that can access

the relevant files. The monitoring information is eventually published as web pages and also as XML documents for further processing by applications like central monitors in an experiment [1], widget-based monitoring tools etc.

The collector side sensors are written in Perl. Several Perl modules are required, namely (a) RRDtool to save and plot historical information about job overview, (b) A templating engine to generate HTML, (c) GD and ImageMagick to generate accounting charts, and (d) XML tools etc., that must be installed. The sensors run as cron jobs. The web interface uses two different JavaScript libraries, namely *Ext JS* [2] for the job overview and group accounting monitor and *jQuery* [3] for the user accounting monitor, in order to build a rich user interface with tabbed view, table with advanced features like sorting, searching etc. and Ajax calls. The monitor supports only standard features and all the major browsers are supported.

4. Web Interface

In the following subsections we shall briefly discuss the various parts of the monitoring interface. Note that presently there are two different web pages (a) job overview and group accounting monitor, and (b) user accounting monitor.

4.1. Farm Overview and Usage History

Figure 2 shows (a) resource availability and overall and group/VO job status at a farm in several tables and (b) evolution of the same information with time, in different time bins. The overview

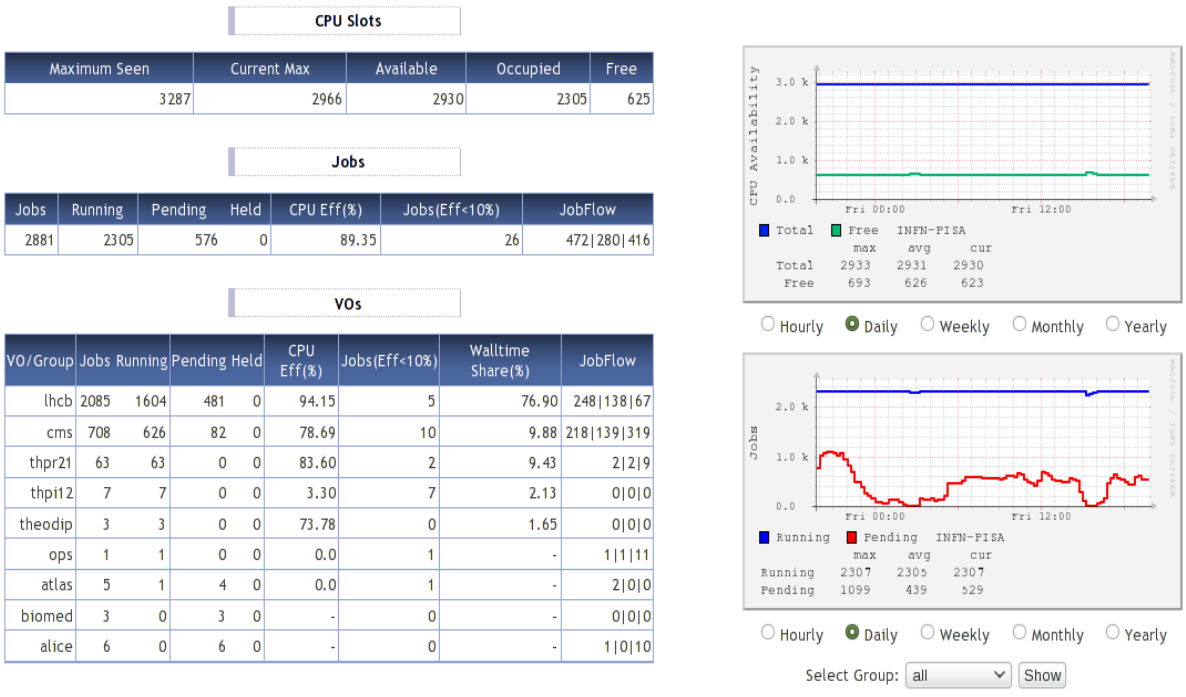


Figure 2. Farm overview : The tables show resource availability and job status. The plots show historical information about resource availability and usage. One can select different pre-defined time bins for both the plots and a group from a drop-down list for the job status plot.

tables usually help site administrators to find problematic jobs belonging to a group/VO, site Computing Element, or an individual user. The time series plots give users a rough idea about

what to expect from the farm in the next hours and provides rough accounting information for the groups/VOs.

4.2. User Oriented View

Figure 3 shows the status of jobs belonging to individual users. Users are uniquely identified by the Distinguished Name (DN) of the Grid Certificate that was used to submit Grid jobs. This view makes the monitor useful for end-users. Grid job submission files are parsed in order to map a local job to a remote user.

User DN

Group	Jobs	Running	Pending	Held	CPU Eff(%)	Jobs(Eff<10%)	JobFlow	DN
cms	1377	625	752	0	14.45	127	0|0|29	/DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=ceballos/CN=488892/CN=Guillelmo Gomez Ceballos Retuerto
theodip	442	442	0	0	53.29	202	0|1|0	/C=IT/O=INFN/OU=Personal Certificate/L=Pisa/CN=Claudio Bonati
compchem	890	432	458	0	61.64	184	121|0|204	/DC=es/DC=irisgrid/O=ehu/CN=ernesto.garcia
theoinfn	421	421	0	0	65.22	117	1|96|7	/C=IT/O=INFN/OU=Personal Certificate/L=Pisa/CN=vincenzo alba
theoinfn	295	295	0	0	29.55	194	0|0|5	/C=IT/O=INFN/OU=Personal Certificate/L=Pisa/CN=Giacomo Ceccarelli
theodip	322	187	135	0	38.38	9	80|185|75	/C=IT/O=INFN/OU=Personal Certificate/L=Pisa/CN=Mario Alberto Annunziata
cms	121	121	0	0	1.50	121	0|0|5	/O=Grid/O=NorduGrid/OU=helsinki.fi/CN=Matti Kortelainen
cms	56	44	12	0	42.91	1	0|0|3	/DC=org/DC=doegrids/OU=People/CN=Yi Chen 27675
cms	102	25	77	0	1.02	14	0|14|9	/DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=asciaba/CN=430796/CN=Andrea Sciaba
cms	21	21	0	0	3.15	21	0|0|0	/DC=org/DC=doegrids/OU=People/CN=Christoph Paus 966610
cms	45	16	29	0	11.83	1	0|0|0	/DC=org/DC=doegrids/OU=Services/CN=uscmspilot05/glidein-1.t2.ucsd.edu
cms	45	14	31	0	21.32	2	0|0|0	/DC=org/DC=doegrids/OU=Services/CN=uscmspilot48/glidein-1.t2.ucsd.edu
cms	41	13	28	0	17.05	5	0|0|0	/DC=org/DC=doegrids/OU=Services/CN=uscmspilot08/glidein-1.t2.ucsd.edu
cms	39	12	27	0	22.06	2	0|0|0	/DC=org/DC=doegrids/OU=Services/CN=uscmspilot06/glidein-1.t2.ucsd.edu
cms	22	12	10	0	83.02	2	0|0|0	/C=RU/O=RDIG/OU=users/OU=sinp.msu.ru/CN=Andrey Belyaev
cms	37	12	25	0	21.82	2	0|0|1	/DC=org/DC=doegrids/OU=Services/CN=uscmspilot47/glidein-1.t2.ucsd.edu
cms	258	10	248	0	53.82	5	2|0|4	/C=IT/O=INFN/OU=Personal Certificate/L=Bari/CN=Vincenzo Spinoso

Figure 3. User oriented view : The Distinguished Name (DN) of the Grid Certificate identifies the Grid users uniquely.

4.3. Group Accounting

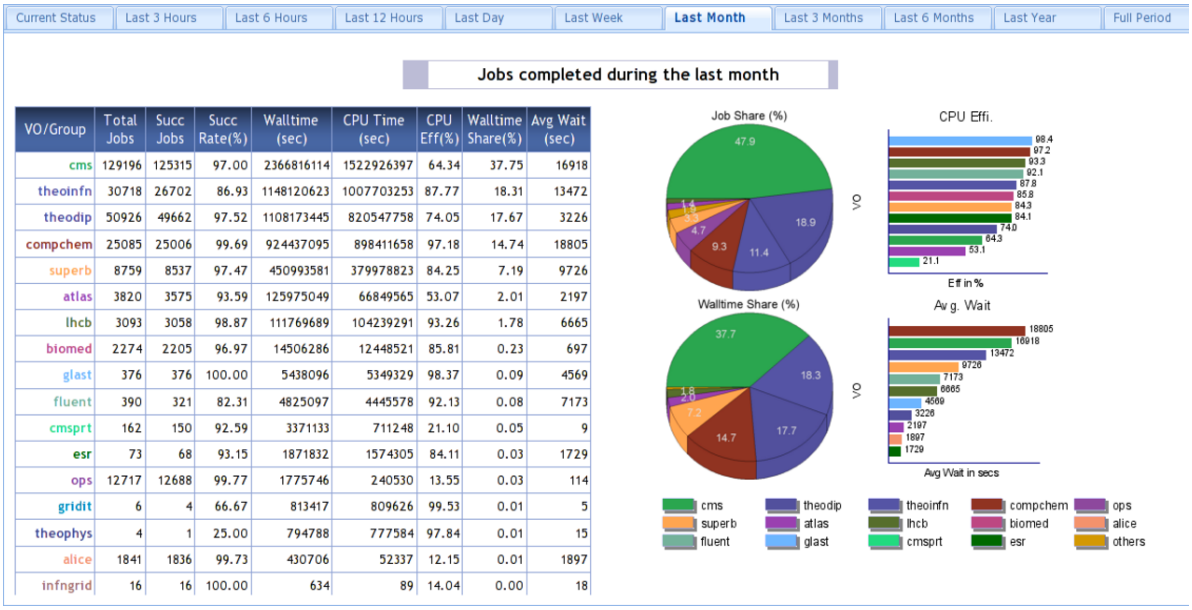
Figure 4 shows the accounting information for each group that finished a minimum number of jobs at the farm, using a number of tabs for different time periods. In addition to the table that presents the various numbers, important quantities are also presented as pie- and bar-charts for the majors groups for each period.

4.4. User Accounting

Figure 5 shows the accounting information in tabular form for each individual user who finished a minimum number of jobs at the farm. A number of tabbed panels are used to present information for different time intervals. The table supports search, sort, and pagination capabilities. A Grid user belonging to several groups/VOs is accounted for correctly. Note that only the Grid jobs are considered in user accounting.

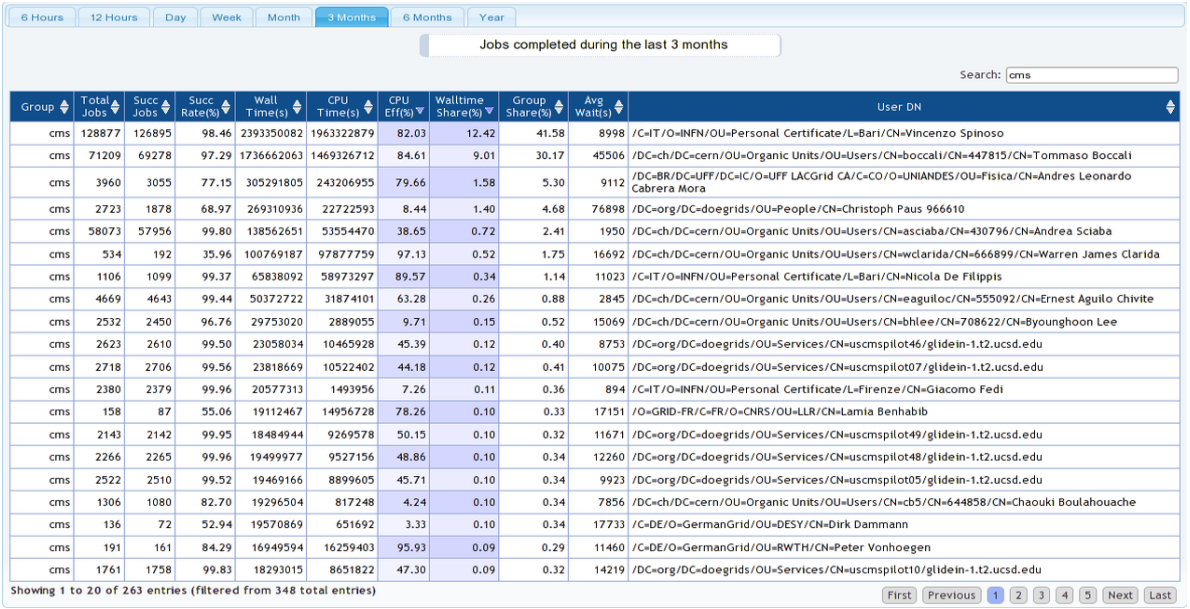
5. Site Monitors

A number of sites running different version of the LSF batch systems have deployed the monitoring system, as shown in table 1 :



VO/Group	Total Jobs	Succ Jobs	Succ Rate(%)	Walltime (sec)	CPU Time (sec)	CPU Eff(%)	Walltime Share(%)	Avg Wait (sec)
cms	129196	125315	97.00	2366816114	1522926397	64.34	37.75	16918
theoinfn	30718	26702	86.93	1148120623	1007703253	87.77	18.31	13472
theodip	50926	49662	97.52	1108173445	820547758	74.05	17.67	3226
compchem	25085	25006	99.69	924437095	898411658	97.18	14.74	18805
superb	8759	8537	97.47	450993581	379978823	84.25	7.19	9726
atlas	3820	3575	93.59	125975049	66849565	53.07	2.01	2197
lhcb	3093	3058	98.87	111769689	104239291	93.26	1.78	6665
biomed	2274	2205	96.97	14506286	12448521	85.81	0.23	697
glast	376	376	100.00	5438096	5349329	98.37	0.09	4569
fluent	390	321	82.31	4825097	4445578	92.13	0.08	7173
cmsprt	162	150	92.59	3371133	711248	21.10	0.05	9
esr	73	68	93.15	1871832	1574305	84.11	0.03	1729
ops	12717	12688	99.77	1775746	240530	13.55	0.03	114
gridit	6	4	66.67	813417	809626	99.53	0.01	5
theophys	4	1	25.00	794788	777584	97.84	0.01	15
alice	1841	1836	99.73	430706	52337	12.15	0.01	1897
infngrid	16	16	100.00	634	89	14.04	0.00	18

Figure 4. Group accounting monitor: The table shows, for each time period, accounting information for each group while the charts compare the important parameters for the major groups in that period.



Group	Total Jobs	Succ Jobs	Succ Rate(%)	Wall Time(s)	CPU Time(s)	CPU Eff(%)	Walltime Share(%)	Group Share(%)	Avg Wait(s)	User DN
cms	128877	126895	98.46	2393350082	1963322879	82.03	12.42	41.58	8998	/C=IT/O=INFN/OU=Personal Certificate/L=Bari/CN=Vincenzo Spinoso
cms	71209	69278	97.29	1736662063	1469326712	84.61	9.01	30.17	45506	/DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=boccali/CN=447815/CN=Tommaso Boccali
cms	3960	3055	77.15	305291805	243206955	79.66	1.58	5.30	9112	/DC=BR/DC=UFF/DC=IC/O=UFF LACGrid CA/C=CO/O=UNIANDES/OU=Fisica/CN=Andres Leonardo Cabrera Mora
cms	2723	1878	68.97	269310936	22722593	8.44	1.40	4.68	76898	/DC=org/DC=doegrids/OU=People/CN=Christoph Paus 966610
cms	58073	57956	99.80	138562651	53554470	38.65	0.72	2.41	1950	/DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=asciaba/CN=430796/CN=Andrea Sciaba
cms	534	192	35.96	100769187	97877759	97.13	0.52	1.75	16692	/DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=wclarida/CN=666899/CN=Warren James Clarida
cms	1106	1099	99.37	65838092	58973297	89.57	0.34	1.14	11023	/C=IT/O=INFN/OU=Personal Certificate/L=Bari/CN=Nicola De Filippis
cms	4669	4643	99.44	50372722	31874101	63.28	0.26	0.88	2845	/DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=eaguiloc/CN=555092/CN=Ernest Aguilo Chivite
cms	2532	2450	96.76	29753020	2889055	9.71	0.15	0.52	15069	/DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=bhlee/CN=708622/CN=Byounghoon Lee
cms	2623	2610	99.50	23058034	10465928	45.39	0.12	0.40	8753	/DC=org/DC=doegrids/OU=Services/CN=uscmspilot46/glidein-1.t2.ucsd.edu
cms	2718	2706	99.56	23818669	10522402	44.18	0.12	0.41	10075	/DC=org/DC=doegrids/OU=Services/CN=uscmspilot07/glidein-1.t2.ucsd.edu
cms	2380	2379	99.96	20577313	1493956	7.26	0.11	0.36	894	/C=IT/O=INFN/OU=Personal Certificate/L=Firenze/CN=Giacomo Fedi
cms	158	87	55.06	19112467	14956728	78.26	0.10	0.33	17151	/O=GRID-FR/C=FR/O=CNRS/OU=LLR/CN=Lamia Benhabib
cms	2143	2142	99.95	18484944	9269578	50.15	0.10	0.32	11671	/DC=org/DC=doegrids/OU=Services/CN=uscmspilot49/glidein-1.t2.ucsd.edu
cms	2266	2265	99.96	19499977	9527156	48.86	0.10	0.34	12260	/DC=org/DC=doegrids/OU=Services/CN=uscmspilot48/glidein-1.t2.ucsd.edu
cms	2522	2510	99.52	19469166	8899605	45.71	0.10	0.34	9923	/DC=org/DC=doegrids/OU=Services/CN=uscmspilot05/glidein-1.t2.ucsd.edu
cms	1306	1080	82.70	19296504	817248	4.24	0.10	0.34	7856	/DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=cb5/CN=644858/CN=Chaouki Boulahouache
cms	136	72	52.94	19570869	651692	3.33	0.10	0.34	17733	/C=DE/O=GermanGrid/OU=DESY/CN=Dirk Dammann
cms	191	161	84.29	16949594	16259403	95.93	0.09	0.29	11460	/C=DE/O=GermanGrid/OU=RWTH/CN=Peter Vonhoegen
cms	1761	1758	99.83	18293015	8651822	47.30	0.09	0.32	14219	/DC=org/DC=doegrids/OU=Services/CN=uscmspilot10/glidein-1.t2.ucsd.edu

Showing 1 to 20 of 263 entries (filtered from 348 total entries)
First Previous 1 2 3 4 5 Next Last

Figure 5. User accounting monitor: The Distinguished Name (DN) of the Grid Certificate identifies the Grid users uniquely.

6. Deployment

The monitoring system is in active development. New features are added regularly while the existing ones get refined. The job overview and group accounting software is distributed as a tarball that can be installed easily and configured extensively. We briefly describe below the

T2_IT_Pisa	http://farmsmon.pi.infn.it/lsfmon
	http://farmsmon.pi.infn.it/users/users.html
T2_IT_Legnaro	http://t2.lnl.infn.it/lsfmon
U. Harvard Cluster	https://software.rc.fas.harvard.edu/lsfmon [protected]

Table 1. The monitor has been deployed at a few sites. The monitoring instances are not yet synchronised to the latest development of the web interface.

deployment procedure:

```
> wget http://sarkar.web.cern.ch/sarkar/dist/lsfmon_v1.8.1.tgz
> tar xzvf lsfmon_v1.8.1.tgz -C /opt
> ln -s lsfmon_v1.8.1 lsfmon
> cd /opt/lsfmon/install
```

A configuration file `app.cfg` included in the distribution has to be suitably modified according to the site setup. The following script

```
> ./setup.sh
```

will generate the necessary files (scripts, cron jobs, application level configuration file etc.). The application level configuration file has a lot of options that might be changed to affect both the content and look-and-feel of the monitor. Further details can be found in reference [5]. The user accounting software will soon be available as a separate package with documentation.

7. Conclusion

The present tool compliments existing global monitors effectively and helps both site administrators and users to spot and eventually fix problems early and reliably. Accounting information for recently completed jobs can only be obtained with such a local tool. The overview monitor inspired development of similar tools for Condor and PBS batch systems [4] and more importantly an XML based uniform local job monitoring framework in the CMS experiment [1].

The monitor now has basic support for parallel jobs. It has been noted that more work is needed for better support of local job accounting and to adapt the monitor to changes in complex farm organisation. As we mentioned earlier, the user accounting monitor is an independent application at present. We plan to merge the user accounting monitor with the rest in near future.

Acknowledgment

The work has been partially funded under contract 2008MHENNA_002 of Programmi di Ricerca Scientifica di Rilevante Interesse Nazionale (Italy). We acknowledge contribution from Massimo Biasotto on LSF 7 accounting file parser and Claudio Strizzolo on HTML validation and font optimisation respectively. We are thankful to Tommaso Boccali, Alberto Ciampa, Enrico Mazzoni, and Satish Patel for helpful suggestions and support.

References

[1] C. Grandi et. al., CMS Distributed Computing Integration in the LHC sustained operations era, PO-WED-061 (Contribution to CHEP 2010, Taipei)
[2] http://www.sencha.com
[3] http://jquery.com
[4] https://twiki.cern.ch/twiki/bin/viewauth/CMS/AnalysisOpsT2Monitoring
[5] http://cern.ch/sarkar/doc/lsfmon.html