Usage of LSF for batch farms at CERN

Dr. Veronique Lefebure, Dr. Ulrich Schwickerath

CERN-IT FIO/FS, CERN, Route de Meyrin, 1211 Geneva 23

E-mail: `Veronique.Lefebure@cern.ch, Ulrich.Schwickerath@cern.ch`

Abstract. CERN uses Platforms Load Sharing Facility (LSF)[1] since 1998 to manage the large batch system installations. Since that time, the farm has increased significantly, and commodity based hardware running GNU/Linux has replaced other Unix flavors on specialized hardware. In this paper we will present how the system is set up nowadays. We will briefly report on issues seen in the past, and actions which have been taken to resolve them. In this context the status of the evaluation of the most recent version of this product, LSF 7.0, is presented, and the planned migration scenario is described.

1. Introduction

The largest cluster of batch nodes at CERN, the lxbatch farm, is a heterogenous cluster of GNU/Linux machines of different vendors and with different hardware configuration. At the time of writing this document it contains around 3000 machines, the majority still being single core double CPU nodes. The majority of these machines are shared between different groups and communities at CERN, as well as the LHC computing grid. Within the last two years the capacity has doubled twice. CERN is expecting to grow by another factor of two in terms of CPU cores in the next years. The new machines will be equipped with multi core CPUs. The rapid growth has unveiled several issues which had to be resolved. The most important once are described, and work-arounds and solutions are shown. The present hardware is already two years old, and it's warranty is going to expire. Moreover, from the experiences of the last grow it seems unlikely that these machine will be able to stand another doublication of resources. In addition, with the startup of LHC, the number of users is expected to grow. A significant upgrade of the hardware is therefor required. At the same time the opportunity is taken to also upgrade the LSF software. The status of the evaluation of the most recent version of LSF (7.0) is described, and a migration scenario is developed.

2. The structure of the batch system at CERN

The CERN shared batch resources consist essentially of one big LSF instance. It is currently running LSF 6.1, and contains about 3000 machines at the time of writing this document, including submission only hosts. The instance contains nodes with different operating systems, namely

- SLC3 (32bit)
- SLC4 (32bit)
- SLC4 (64bit)

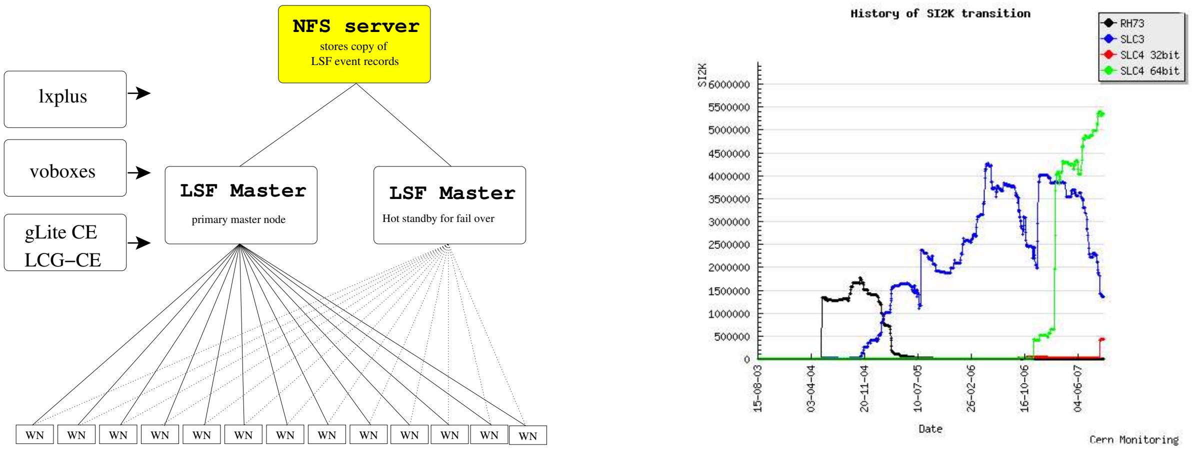

Figure 1. The batch system architecture (left) and evolution of the computing power and operating systems (right)

As the support for Scientific Linux 3 will expire in autumn 2007, this part of the cluster will be rapidly retired and migrated to SLC4. Nodes which are not capable to host a 64bit operating system will be moved into the SLC4/32bit part. Submission hosts include

- CERNs interactive login cluster, lxplus
- VO - boxes
- Computing Elements (CE)

Access to the lxplus cluster is granted via secure shell. The interactive nodes can be used for software development and, to a certain amount, debugging of experiment software. Apart from that it serves as a gateway directly to lxbatch but also to the LHC computing GRID.

VO boxes are dedicated specialized machines with high availability, which are used by specific VOs only. Like lxplus, these nodes offer interactive access via secure shell, but they are available only to a limited number of users. Computing elements serve as an interface between the LHC computing grid and the local batch farm.

There are two machines dedicated to LSF master services. The LSF job scheduler runs on one of them while the second serves as a hot standby machine. The information of the running jobs is shared between the two nodes using a NFS volume. However, the system is set up in a way that the master always keeps a local copy. In case the NFS server fails, the fail over mechanism is lost but the system continues to run. In case the primary master fails, the hot standby overtakes the services from the primary LSF master node, using the shared event files from NFS. A load sharing between the two is currently not possible, though. Fig. 1 shows a schematic al view of the described architecture.

All nodes which have access to the batch system are managed by the Quattor [3] tool kit. Dynamic adding of hosts is not used, in order to ensure a proper system configuration. System monitoring is done using LEMON [3]. A new Lemon sensor allows the monitoring of the CPU/Wall clock time ratio, based on all finished jobs within the last 24 hours. A bunch additional scripts scripts detect problems like orphaned jobs, stalled file transfers and jobs which do not use any CPU time[4].

2.1. Fair share

Fair share scheduling is used to ensure that each community which uses the batch system gets a previously agreed fraction of the available resources. Groups of users inside these communities

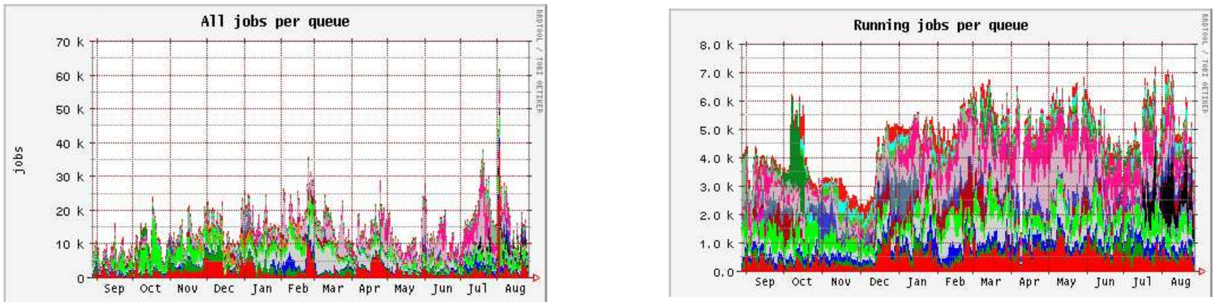

Figure 2. Number of queued (upper graph) and running (lower graph) jobs

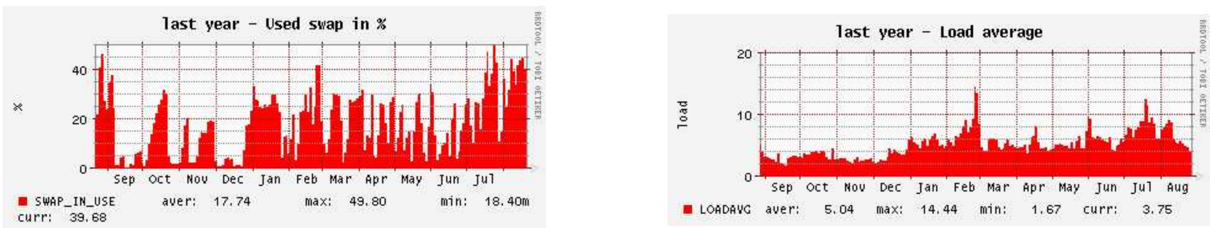

Figure 3. Swap space usage (upper graph) and load average (lower graph) of the LSF master node in the last year.

can get a fraction of the total allocation assigned to the community. Examples of such groups could be ordinary users, production users, grid users and software managers. The definition of sub-groups happens inside the LSF configuration[1]. The fraction that each group gets can be changed within 24h on demand. At the long term, it is foreseen to give the experiments the ability to change these fractions them-self. Share allocations and sub group fractions are stored in a database. The information is used to automatically recreate and deploy the corresponding LSF configuration files once per day. A LSF reconfiguration is initiated automatically in case of configuration changes. At the same time hosts can be added or removed from the LSF configuration, based on information stored in the Configuration Data Base (CDB)[3].

2.2. Usage patterns

Most of the jobs at CERN are single threaded sequential jobs. In principle it is possible to run parallel jobs although the demand for it is very small. Based on this use pattern, the number of job slots per machine has been chosen as

$$n_{slots} = n_{cores} + 1$$

where n_{slots} is the number of job slots and n_{cores} is the total number of CPU cores of the machine. Recent machines come with 2 GB of memory per CPU core which is sufficient for the typical jobs we see on lxbatch at the time of writing this document.

3. Limitations

As the CERN computer center keeps growing rapidly, being the largest center in the LCG environment, regular reviews of any scalability issues are required. At the time of writing this

[1] There is no support for secondary UNIX groups at CERN, and the primary group ID is assigned to the community/experiment

document, the typical number of jobs in the system is between 20k and 30k, and the typical job throughput is around 20k per day or above.

3.1. LSF master hardware

LSF master services currently run on 2.8 GHz dual Xeon machines, with 4 GB RAM, and mirrored system disks with redundant power supplies. All three master nodes in1 are secured by a UPS and a Diesel Generator, to ensure reliable running in case of power glitches. The current master nodes still use a SLC3 operating system. Problems have been observed in particular if more than 50 000 jobs are in the system. Under these conditions the primary master can run out of physical memory, and performance is heavily degraded.

A further increase in worker nodes and/or users requires the LSF master hardware to be replaced by more powerful machines. Most important seem a significant upgrade in main memory. A 64bit machine with 8 cores and 16 GB of memory seems adequate for the expected work load of the next 3 years.

3.2. Software

The new hardware requires a 64bit operating system. The support for SLC3 will end in autumn 2007, therefore the new operating system will be based on SLC4/64bit. In the past, problems have been seen if users or software components hammered the batch system with batch system status queries. A step towards a more reliable system must therefore also try to reduce known sources of such queries. Last but not least, LSF6.1 is a rather old version of Platform LSF with scalability limits around 5000 hosts[2]. Therefore, an update to the newest version is advisable.

3.3. Grid software and information system

Computing Elements (CE) serve as a gateway between the local batch system and the GRID. Each CE runs an information provider process, which frequently queries the batch system. While in the original design it was foreseen to have only one such machine per cluster, the number of CE nodes had to be increased drastically to allow CERN to cope with the work load. The so called GRIDCE cluster currently contains about 24 machines, including test nodes. As a result, the number of batch system queries per second increased drastically. Under high load conditions, this caused performance problems with the batch system, in particular in the response time, with the effect that query requests started to pile up.

To cope with this situation, the interface of the information system to the local LSF batch system was rewritten from scratch. Batch system query results are now stored in cache files on a network file system which is shared between all the CEs nodes. Apart from the reduction of the load on the batch system, this also has the advantage that the status reported by different CEs (which all look at the same batch system nodes) become synchronized. A disadvantage is that an additional machine was needed to provide the network disk.

The rewritten LSF information provider plugin is now part of the official distribution of the EGEE middleware distribution.

4. Tests and upgrade plans

From the observations described above it seems clear that CERN is currently running close to the limits of the system, and that there is a need to upgrade both hardware and software. The latest version of LSF, LSF7.0, has been packaged and installed on a test instance. On this instance, the following tests were performed:

- scalability in term of number of jobs
- scalability in terms of batch system queries, with a reasonable number of jobs in the system

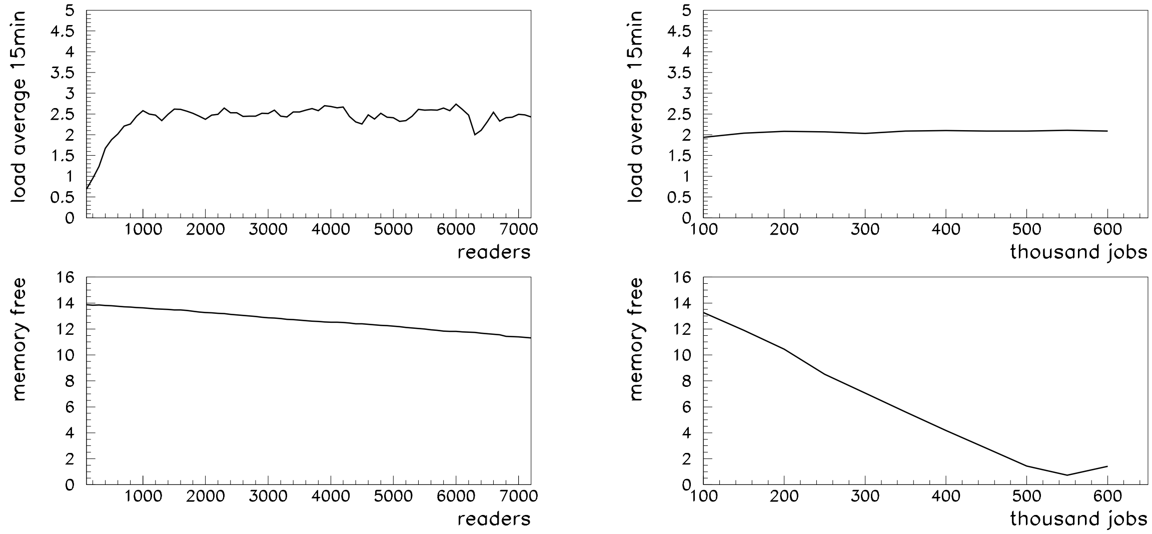

Figure 4. LSF 7 tests: Load average and free memory as a function of the number of processes querying the batch system (left) and the number of jobs (right) in the system.

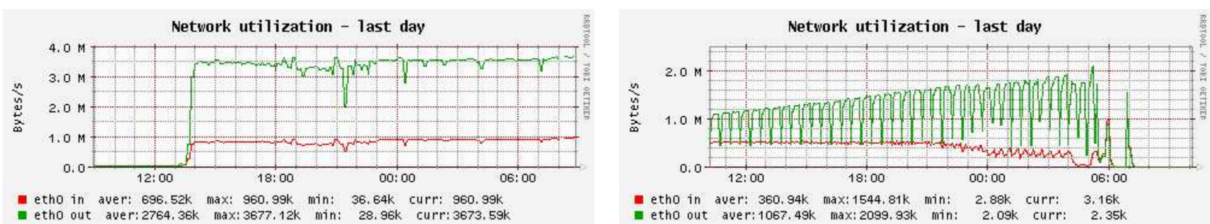

Figure 5. Network utilization during the batch system query test (left) and the number of jobs test (right). The time scale on the X-axis is arbitrary.

4.1. Hardware selection

Fig. 3 shows the swap memory usage of the current primary LSF master node during the past year. Periods where the machine runs out of physical memory are clearly increasing, and the amount of the used swap space suggests that the main memory should be at least doubled. For address space reasons a 64bit machine should be used. While 16GB of memory seems advisable, the situation with the number of core is not obvious. The high load situations which have been seen on the production systems could not be reproduced with a new dual Quadcore machine and 16GB of memory. However, besides of LSF master functionality the production master also executes a bunch of cron jobs and other tasks, which could be dedicated to additional cores. The new hardware itself has to be as reliable as the previous was, with mirrored disk and redundant power supplies. Primary master and hot standby must not be on the same switch, to make the system more robust against switch failures.

4.2. LSF 7 tests

A test environment has been set up with a master node of the target configuration. Two main tests have been done. For the first test, batch worker nodes were added as submission only hosts to the test instance. On each batch node 4 processes were started which queried the batch system in a endless loop, using bjobs and bacct requests. The left hand side of Fig. 4 and 5 show the load average, free memory and the network utilization of the LSF master machine during this test. It can be seen that the network link saturates rapidly, and the load remains stable. The network interface speed turns out to be the bottle neck in this configuration.

In the second test small jobs were queued in the system, to test the memory requirements as a function of the number of jobs in the system. The jobs themself were very simple sequential jobs. It was not attempted to do any optimization by, for example, making use of job arrays during this test. Submission was done in blocks of 50,000 jobs. After each submission, the number of jobs in the batch system was queried. The result of this test is shown in Fig. 4 and 5. The structure in the network utilization in 5 is caused by the batch system query which was done to verify the number of jobs in the system after each submission. Above about 600,000 jobs the system became unresponsive. In the test 500,000 jobs corresponded to about 12 GByte of used memory, leaving 3 GByte for other tasks. This result is consistent with the statements in Platforms release notes [1]. As a result of this test, the number of jobs should be limited to at most 500,000 jobs.

5. Migration scenarios

The basic architecture which is in place now has proven to be reliable. The fact that LSF6 server nodes can talk to a LSF7 master node allows for a rolling update, making use of the existing fail-over mechanism. In a first step, the hot standby node will be retired and replaced by a new machine running the new software. This step is fully transparent for the users. Then, the master node will be brought down. The batch services will stop for some time until the hot standby, now running the new software, takes over. Services continue, with a LSF 7 master on new hardware, and LSF 6.1 server nodes. The next step is the replacement of the original primary master by the new hardware. When it resumes it will take over again from the hot standby machine, and the migration of the master nodes is done. Next, the batch worker nodes have to be migrated. The new software packages will be deployed on the nodes, and LSF restarted. Also this operation is expected to be transparent for running user jobs. During the migration of the LSF master machine, some log files are to be transfered from the old to the new master. Therefore, and in order to reduce the traffic on the NFS server during the migration, the batch system should be paused during the master migration, although, strictly speaking, even that may not be necessary.

It is planned to practice the migration procedure at least once in a test environment, once the detailed procedure is fixed.

6. Summary

The batch farms at CERN have doubled their capacity twice within the last two years, without changing the LSF master hardware. Scalability issues have been addressed by drastically reducing the number of batch system queries. A required update of the Grid middleware is now part of the EGEE software releases. However, another doublication of worker nodes requires a significant update of both hardware and LSF software while keeping the base architecture of the system. It is expected that these updates will be sufficient to guarantee reliable batch services for the first years of LHC running.

7. References

[1] Release Notes for Platform LSF Version 7.0, Nov. 2006, PLATFORM corporation.
[2] Release Notes for Platform LSF Version 6.1, Feb. 2005 , PLATFORM corporation. http://www.platform.com/services/ support/docs/lsfdoc61/release_notes.html
[3] Extremely Large Fabric management system, http://elfms.web.cern.ch/elfms
[4] Tim Bell, CERN, priv. comm.