LS-DYNA on Linux-Clusters at EDAG
Use Case

John U.S. Hanlon, IT-Administrator, EDAG Engineering & Design AG
Bernward Platz, Manager Software-Development, Teraport GmbH

1 Abstract

At EDAG, the increasing demand for crash simulations with LS-DYNA required the expansion of the existing compute resources in the beginning of 2002. As at this point there was a stable and performant MPP-version of LS-DYNA available, EDAG decided - despite of the new technology - to purchase a Linux-cluster consisting of 8 nodes and 16 processors. Meanwhile EDAG has deployed over 14 clusters at three different locations. Whereas at the beginning, purchase costs and performance played a major role, due to the increasing complexity further demands became important which were realised by the EDAG partner, Teraport:

- fast, convenient and reproducible installation and administration of the cluster systems
- parallel usage of several DYNA versions with various MPI-libraries
- easy extensibility
- convenient usage of cluster resources via web technologies
- high-available access to the cluster resources
- handling of high data transfer
- integration of further FEM-applications

The paper describes a use case which reports the concepts of the realisation of these requirements as well as the problems such as hardware quality during the installation and setup of the cluster environment. At the end, a perspective concerning further possibilities of development such as web-based workflow optimization and result management will be shown.

2 Introduction

Until the beginning of August 2002 EDAG solely used workstations and servers running HP-UX for CAE computations, that where distributed to the different hardware resources through the queuing and load sharing tool LSF (Load Sharing Facility), a product of Platform Computing. In 2001 EDAG heard of other companies already using Linux systems for their simulations successfully. Finally when the existing capacity of CPUs where not sufficient any more and new hardware had to be ordered, EDAG decided to use servers equipped with AMD Athlon processors and running Linux operating system.

3 Chronological Summary

The first Linux cluster, that was brought online in February 2002, consisted of 8 Athlon dual processor machines manufactured by Fujitsu-Siemens Computers. One special management node was used for the administration of the cluster. For the purpose of providing fileserver functions and a LSF installation a single CPU machine was sufficient. The cluster was seamlessly integrated into the existing environment, using NIS and LSF services of the CAE department's network and was installed by Teraport. It was designed for the exclusive use with LS-Dyna. Due to the size of the models every computation was intended to always use all 16 processors of the cluster making use of the MPP version of LS-Dyna. From the external point of view the cluster was supposed to look like a single 16 CPU machine. This besides the wish to reduce license fees was the reason to install LSF on the management node and the first compute node only. The first compute node, that is also responsible for the decomposition of the models, distributes the calculation to the other nodes using MPI. Figure 1 makes this design clearer.

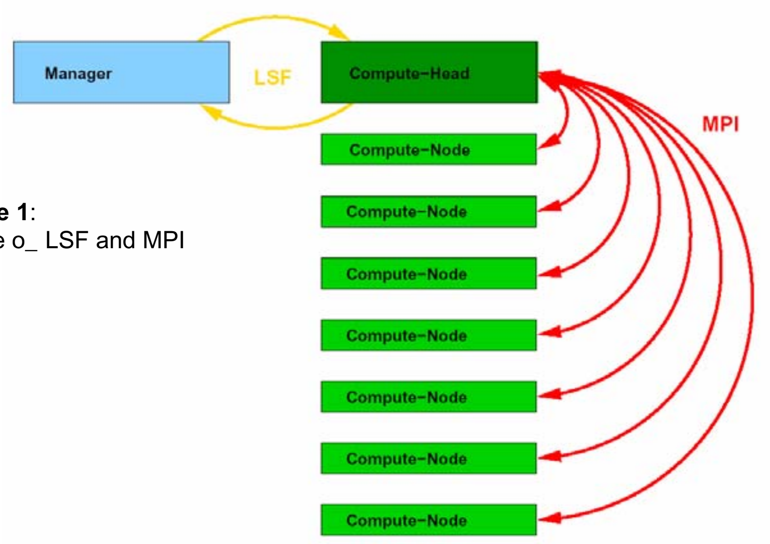


Figure 1:
Usage o_ LSF and MPI

Calculations were complete up to 3 times faster than on the Unix servers. Convinced of the productive efficiency of the cluster 2 more Fujitsu-Siemens clusters were purchased in May 2002. Because EDAG toyed with the idea to distribute clusters to different locations at a later date, the new clusters were connected to the EDAG network through their own private networks. The only way to access a cluster was through the management node, from where the computation jobs were started. By this means the cluster could be moved simply from one location to another and at the same time access to the cluster could easily be restricted. The next cluster followed soon (see figure 2). Because of the project situation, hardware partly had to be purchased at very short notice. This led to ordering hardware from other manufacturers, because the time to delivery of the most of the hardware manufactures averages out at 6 to 8 weeks. As it turned out later on the different delivery times are caused by different kinds of quality management of the companies. Finally in September 2002 there were 10 Linux clusters installed at EDAG. In order to be able to move clusters rapidly from one location to another, all clusters were.

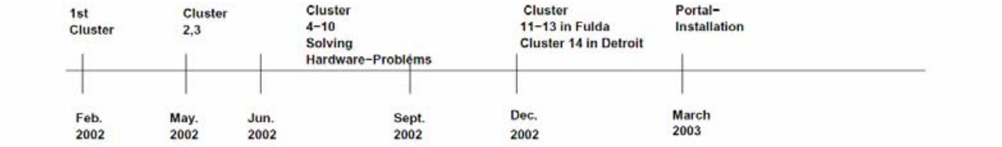


Figure 2: Timeline

set up as independent systems each with its one management node, user administration and LSF system. For the system administrators this meant an increasing effort to maintain the systems, for the users this resulted in the problem to pick out the best and free cluster for every single computation.

With the last 3 clusters being set up and brought online in December 2002, a consolidation of the cluster became necessary resulting in 3 clusters at to locations. Each of the three clusters was organized with one management node. See figure 3 for a view on this concept.

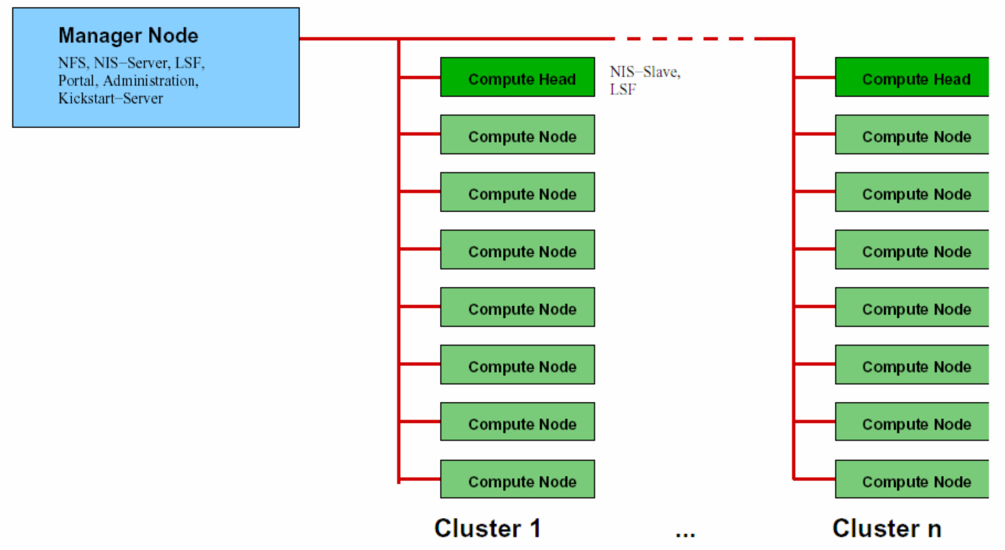


Figure 3: Teraport Linuxcluster - Setup

4 Concepts and Problems

4.1 Administration and Installation

From the beginning on administration of the clusters was based on a concept, that made it possible to install and configure services and applications fast and reproducible on all cluster machines. Teraport's administration tool TPAdmin allows parallel installation and configuration of applications like for example LS-Dyna on all compute nodes from one central management node without having to login to the nodes. The cluster appears like one single machine. Teraport has developed a special mechanism based on scripts with which applications can be installed remotely. The applications are available in form of RPM packets, that are built by Teraport. The versions of the application to be installed as well as the destination hosts are configured through configuration files that are stored on the management node. RPM packages provide for a consistent installation for all applications. It is no problem to install different versions of an application in parallel. It is also account for the dependencies of different versions of an application on different library versions (for example LAM-MPI). With the combination of scripts, configuration files and RPM packages application can be distributed rapidly and without any interaction.

TPAdmin is not used for the installation of the operating system. On the contrary it requires an installed operating system. The operating system (RedHat) is installed with RedHat's kickstart method. This method basically consists of a kickstart file, that contains the configuration (network, filesystem, software packages) for the hosts. This configuration is gathered from a central kickstart server and interpreted by the host that is to be installed after booting from CD, floppy, harddisk or network. Then the Installation takes place automatically. Teraport has extended this method for clusters, so far that the complete cluster is defined and the kickstart files are generated for all nodes of the cluster with help of a special cluster configuration file.

The combination kickstart/TPAdmin allows setting up a whole cluster completely and fully functional within one hour. In case that one node fails, it is re-integrated into the cluster in less than 30 minutes, even if the hardware needs to be replaced.

4.2 Hardware

The first three clusters were based on Athlon MP 1900+ and 2000+ mainboards (Tyan Thunder). The hardware setup of the machines was designed by Fujitsu-Siemens. As already mentioned above, the next five clusters were ordered from a different hardware vendor due to time constraints. The new hardware turned out to be very unstable. Teraport traced back the reason for this instability to the fact that among other things the power supplies were not sufficiently dimensioned, only providing 300W. The harwdare manufacturer was not easy to convince of this fact and attempted with other more or less questionable changes to stabilize the hardware. Also the delivered quality left some things to be desired, as for example forgotten or incorrect graphics adapters. Only when the hardware was change (460 Watt power supplies instead of the 300 Watt ones, Tyan Thunder mainboards instead of Tyan Tiger) these problems could be solved.

Recapitulatory it can be said, that longer delivery periods of some hardware vendors are certainly caused by an increased standard of quality management. Thus putting up with longer delivery times is economically advantageous in the end.

4.3 Scalability and High Availability

When clusters are integrated under ONE management node it has to be taken care that this node is no single point of failure, because it provides file service, user administration (NIS) and LSF service. Hence it has to be ensured, that a failure of the management node does not cause computations to be interrupted. This can be achieved by configuring the clusters to run computations on their own local filesystem and copy the results back to the fileserver after the calculation is finished. NIS can be set up redundant by using NIS slave servers. Finally also LSF provides the option tp be implemented redundantly by configuring the head nodes (= first node in a cluster) as possible LSF master server (see figure 3). In order to make the access to the cluster redundant, i.e. it is possible to submit jobs and access data although the management node fails, the management node has to be designed as two separate machines controlling each other. If one machine fails the other takes over all the services. This concept (see figure 4) will be implemented at EDAG in the next months.

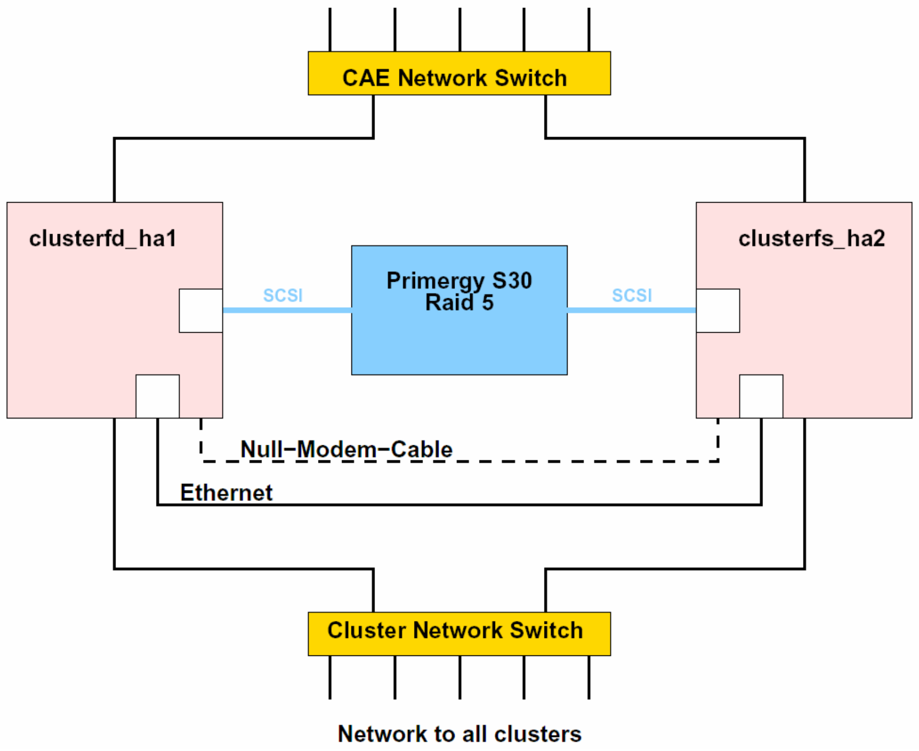
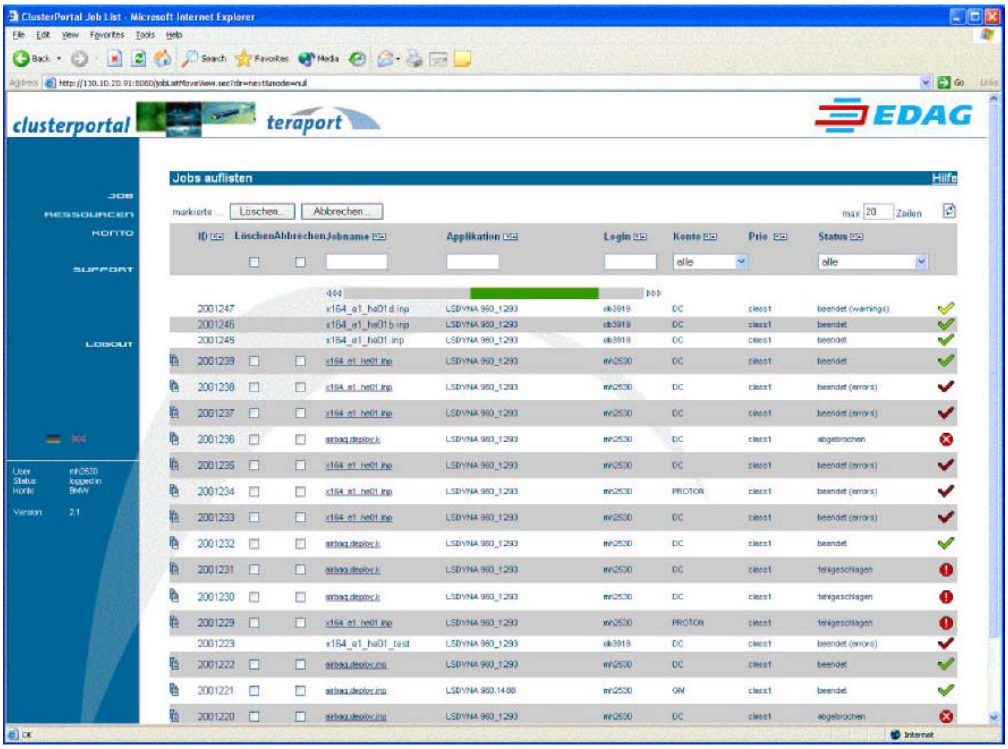


Figure 4: HA-Server setup

4.4 User Convenience

Besides the administrative side also the easy usability of the cluster has to be remembered. Normally the user would have to cope not only with the job management system but also with building up the MPI environment and copying input and output files. Until end of March 2003 EDAG used a script based solution, that had to be updated with every new application or version. Beyond this there is the disadvantage that users have to log into the management node in order to submit jobs. With Teraport's Clusterportal, a web based access system, most CAE applications like LSDyna can be used through a graphical user interface. To make the change to the gui easier, the portal was integrated into the existing scripts via its own command line interface, so that the user can submit his jobs with the same script as before. Now he can do this from his workstation and he doesn't have to login into the management node. The Job is submitted tp the portal using http. The Computation can further on be controlled through the Web-GUI. This provides the user with his usual environment. Additionally he can make use of the portal's advantages.


ClusterPortal Job List - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://130.10.20.91:8080/jobListMoveView.sec?dir=next&mode=null
Go Links
clusterportal
teraport
EDAG
JOB
RESSOURCEN
KONTO
SUPPORT
LOGOUT
Jobs auflisten
Hilfe
markierte ...
Löschen...
Abbrechen...
max 20 Zeilen
Done
Internet

User mh2530
Status logged in
Konto BMW

Version 2.1

ID	Löschen	Abbrechen	Jobname	Applikation	Login	Konto	Prio	Status	
	□	□				alle		alle	
2001247			x164_e1_he01d.inp	LSDYNA 960_1293	db3919	DC	class1	beendet (warnings)	✓
2001246			x164_e1_he01b.inp	LSDYNA 960_1293	db3919	DC	class1	beendet	✓
2001245			x164_e1_he01.inp	LSDYNA 960_1293	db3919	DC	class1	beendet	✓
2001239	□	□	x164_e1_he01.inp	LSDYNA 960_1293	mh2530	DC	class1	beendet	✓
2001238	□	□	x164_e1_he01.inp	LSDYNA 960_1293	mh2530	DC	class1	beendet (errors)	✓
2001237	□	□	x164_e1_he01.inp	LSDYNA 960_1293	mh2530	DC	class1	beendet (errors)	✓
2001236	□	□	airbag.deploy.k	LSDYNA 960_1293	mh2530	DC	class1	abgebrochen	✗
2001235	□	□	x164_e1_he01.inp	LSDYNA 960_1293	mh2530	DC	class1	beendet (errors)	✓
2001234	□	□	x164_e1_he01.inp	LSDYNA 960_1293	mh2530	PROTON	class1	beendet (errors)	✓
2001233	□	□	x164_e1_he01.inp	LSDYNA 960_1293	mh2530	DC	class1	beendet (errors)	✓
2001232	□	□	airbag.deploy.k	LSDYNA 960_1293	mh2530	DC	class1	beendet	✓
2001231	□	□	airbag.deploy.k	LSDYNA 960_1293	mh2530	DC	class1	fehlgeschlagen	!
2001230	□	□	airbag.deploy.k	LSDYNA 960_1293	mh2530	DC	class1	fehlgeschlagen	!
2001229	□	□	x164_e1_he01.inp	LSDYNA 960_1293	mh2530	PROTON	class1	fehlgeschlagen	!
2001223			x164_e1_he01_test	LSDYNA 960_1293	db3919	DC	class1	beendet (errors)	✓
2001222	□	□	airbag.deploy.inp	LSDYNA 960_1293	mh2530	DC	class1	beendet	✓
2001221	□	□	airbag.deploy.inp	LSDYNA 960.1488	mh2530	GM	class1	beendet	✓
2001220	□	□	airbag.deploy.inp	LSDYNA 960_1293	mh2530	DC	class1	abgebrochen	✗

Figure 5: Teiraport's Clusterportal

5 Summary and Outlook

After having the hardware problems described above, EDAG can now make use of a stable, simple to administrate and comfortable to handle cluster system, which can be easily expanded at any time. Failing nodes can be replaced without problems and then be newly installed.
Still there are things to be improved:

- LSF: The whole functionality of LSF is rather unused. Hence the question is, if LSF is really needed. By using a different job management system like PBS expense could be reduced significantly without having to turn down functionality. The changeover from LSF to PBS is planned for 2003. Because the portal hides the underlying job management system, a migration is smoothly to accomplish.
- High Availability: In Fulda all clusters are supposed to be managed by one management node, which makes it even more important to set up this node highly available.
- Integration of HP servers and other locations: It should be possible to control all clusters through the portal. Furthermore it should play a role, if a cluster runs with LSF or PBS. The Clusterportal provides appropriate possibilities.
- Usage of other applications than LS-Dyna: Through the portal applications like Abaqus, Pamcrash and Nastran can also used. This means one uniform way to access all applications.
- Result management: The portal is not only supposed to work as access system for clusters, but also as platform for analysis ans validation of computations as well as for communication of the result.