Development, installation and maintenance of Elix-II, a 180 nodes diskless cluster running thin-OSCAR

Michel Barrette[a] , Xavier Barnabé-Thériault[a], Mehdi Bozzo-Rey[a],
Carol Gauthier[a], Francis Giraldeau[a], Benoît des Ligneris[a*],
Jean-Philippe Turcotte[a], Patrick Vachon[a], Alain Veilleux[a]

[a]Centre de Calcul Scientifique,
Université de Sherbroke, Québec, Canada

A short presentation of the first diskless cluster developed at Sherbrooke University physics department (Elix) will be made. Then we will present the development of Elix-II, a 180 nodes diskless cluster installed and managed with the Open Source Cluster Application Resource (OSCAR) framework. The need to support diskless cluster for OSCAR lead us to create the thin-OSCAR workgroup, we will present the specific requirements for this particular architecture type (diskless and systemless) for Beowulf clusters. Some examples of scientific problems that can be solved on a diskless cluster will also be given.

Une brève présentation du premier cluster sans disque réalisé au département de physique de l'Université de Sherbrooke, Elix, sera faite. Ensuite, nous présenterons Elix-II, un cluster sans disque de 180 processeurs installé et maintenu à l'aide du projet Open Source Cluster Application Ressource (OSCAR). La nécessité d'adapter le projet OSCAR aux systèmes dans disques nous a poussé à créer le groupe de travail thin-OSCAR. Nous présenterons les prérequis pour ce genre d'architecture (sans disque-dur ou sans système sur le disque-dur) ainsi que des exemples de problèmes scientifiques qui ont été résolus sur les clusters sans disque Elix et Elix-II.

Introduction

Diskless clusters are a recent evolution of the general Beowulf idea which consist of removing all computer parts that are not directly useful for scientific computation. Harddisks are useful for data storage but not directly for computation. Specifically, disk access time is very slow compared to RAM access time and even slower compared to cache access time. Some very successful data-mining company [1] decided, for efficiency reason, to have all their database in RAM and therefore prove the viability of the RAM-based approach on a very large scale (more than 10.000 nodes).

We use the same approach for scientific computing and we wanted to prove, on a smaller scale, that diskless cluster is a viable solution for numerous type of numerical problems that can be used in a large variety of scientific fields.

The Scientific Computing Center [2] of Sherbrooke University developed a unique knowledge and experience relative to diskless clusters. The first diskless cluster we built, Elix (1999) will be presented from a hardware and software perspective. This cluster is used for solid state physics computations that will be briefly presented.

For our most recent cluster, Elix-2, we decided to use the OSCAR [3] framework that brings us « the best known methods for building, programming, and using clusters ». The realization of Elix-II will be detailed from a hardware perspective (realization of our own blades, out of band node power management system, . . .) and software perspective. Our implication in OSCAR grew to the point that we are now core members of the OSCAR work-group. We created and lead the thin-OSCAR [4] workgroup in order to address the specific problems of supporting diskless and systemless nodes inside the OSCAR framework.

*Corresponding author : benoit@des.ligneris.net

1 Elix

This is the first Linux cluster in Sherbrooke University. The project was started at the end of 1998.

1.1 Hardware organization

A bunch of standard desktop boxes were purchased and interconnected with standard 100 Mb Ethernet network. The first phase was 16 boxes of dual Pentium-II 400 MHz whereas the second phase was 30 boxes of mono Pentium-III 667 MHz and a dual Pentium-II 400 MHz as a server for a total of 64 CPUs usable for calculation into 46 nodes.

Figure 1. Picture of the Elix Rack with nodes

1.2 Software implementation

We used the standard rootnfs mechanism [5] to boot our diskless client from a dedicated server that is used as NFS server, PBS server (no calculation on the server), DHCP and TFTP [6] server. The boot process for a node is presented figure 2.

For each node, we duplicated completely all files on the server so that they have all their complete file-system available via NFS except for the `/usr` directory that was common to all nodes and mounted in read-only mode. The `/home` directory was also common to each nodes but, as it's the case for user files, this directory was readable and writable by all clients. The total size of a node image (without the `/usr` and `/home` directory) was around 40 Mb. The `/usr` directory for the nodes size was less than 200 Mb.

The only delicate part of this cluster was the boot and reboot process (it happens often because power failures are frequent at Sherbrooke University and the UPS for the whole cluster provides only a few minutes of additional uptime). Indeed, the **tftp** server we used was not optimized for high-performance and it was impossible to boot the 46 nodes at the same time.

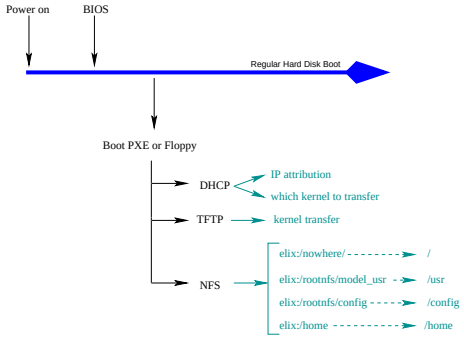

Figure 2. Boot process of an Elix node

1.3 Scientific problems solved

The Elix cluster has been used extensively by professor Tremblay's team to solve numerous problems relative to quantum state of matter [7–9] and to quantum computing [10, 11]. Quantum Monte Carlo simulations were the main numerical tool but simulated annealing and other computationally intensive algorithms were used. The Professor Senechal's team [12, 13] uses Elix for Cluster Perturbation Theory (CPT) and extensive sparse matrix inversion (Lanczös method, see, e.g. [14]) as well as for post-processing.

Table 1 summarize the number of people that used the Elix cluster at the physics department from it's early beginning up to now. The table 2 gives the aggregated statistics

Study level	Number	Year(s)
Undergraduate students	5	2000 - 2002
M. Sc. Students	5	2001-2002
Ph. D. Students	4	2000-2002
Post Doctoral Fellows	3	2002
Professors	2	2000-2002

TAB. 1
Usage of Elix cluster (people)

Group	Usage (CPU-hours)
A.M. Tremblay's group	135110
D. Senechal's group	140971
Other	12794

TAB. 2
Usage of Elix cluster (hours) from march 2001 up to December 2002

(per group) from march 2001 to the end of December 2002.

Numerous results of these scientific calculations were presented at several international conferences. Also, the Elix cluster was a key tools for the Master's thesis of three students. A patented technology has been developed using numerical computation made on the Elix cluster. Moreover, it gives the occasion to others students (experimentalists and theoreticians alike) to try and develop several numerical intensive programs that would not have been useful otherwise because of huge calculation time and scarcity of computing resources for experimentalists.

While we could believe that the Elix usefulness would come to an end abruptly with the coming of its bigger and more efficient successor, Elix2, the cluster is still used by postdoctoral fellows, graduate and Ph. D. students and therefore has shown an exceptional longevity for a computing resource. The cluster will be upgraded soon to more powerful processors.

Another side effect of the presence of the Elix cluster is that the numerical computation course (for undergrad students) is now given on Linux workstation so that students are now more efficient than ever on any Linux cluster as they are more familiar with the operating system and all the excellent open source tools available for programming. Most of theorists students that are the core users of Elix work under Linux so that they can test their programs on their workstation and then use the cluster to do the massive computation.

All in all, we can conclude that the Elix cluster gave access to very efficient and cost effective computing resource. The total cost of the cluster, including cooling system, research, development and installation is evaluated to 90.000 $ (Canadian). The cost per CPU is around 1400 $ (CA). The diskless cluster did not limit that much the range of problems that were addressed and a wide range of numerical simulation has been made : optimization (simulated annealing, quantum Monte-Carlo, matrix inversion, nume-

rical treatment (FFT, fit, integration, ...), finite differences and finite elements. The only tasks that can't, of course, run on a diskless clusters are I/O intensive tasks requiring a disk per node.

2 Elix-2

With the great success of the Elix cluster, from a scientific, sysadmin and « builder of cluster » point of vue, the creation of its successor, called Elix2 was decided. This new cluster was funded by professors Tremblay (Sherbrooke University) and Nelson (Bishop's University). The total number of node will be 180. For this bigger cluster, we decided to look at the available tools for cluster installation and management.

OSCAR was tested and seemed to be able to fulfill our needs. It's image management system (the System Installation Tools) was something we did not have and each image creation was a (painful) manual step (basically, tries and errors until it works !). We had some kind of remote node execution but the simplicity and power of the C3 tools seduced us. Finally, OSCAR uses PERL for it's internal work (as well as shell scripts) so we know, because OSCAR was an open source project, that we will be able to code any missing features.

The next step was to add support for diskless nodes for OSCAR and this lead us to create the thin-OSCAR workgroup. Elix2 is the model for a bigger and already funded project (the Mammoth project) that will have more than 1000 nodes. Elix2 is a proof of concept for the Mammoth project and, while we use thin-OSCAR internally, we plan to submit all the written code to peer review by integrating it's functionality into OSCAR.

2.1 Hardware organization

The first Elix was built with standard desktop boxes with their floppy disk (we did not know PXE well by the time of Elix creation). This time, we decided to get rid of floppy drives and, because no more space was available in the Elix room, we decided to use a more compact vertical blade solution. This solution allows us to put more computers per square foot and, as a consequence, reduce the total cost of ownership of the cluster.

All this work was done in cooperation with the RITTAL [15] company that help us from the early stage of design, prototype and final system.

2.1.1 Blade design

While a lot of clustering companies sell blades, they are not particularly cost effective. When we made our calculations for the best performance/cost ratio, the home-made blade solution was definitely the winner. In order to do that, we designed an aluminum plate able to hold the motherboard (standard ATX or micro ATX) and a regular size power-supply. Those components are very cost effective and that is the main reason for our choice. We had some trouble designing a clean way to fix the power supply and we decided to produce custom components only for this specific part.

The front panel of the blade is also in aluminum and has all the slots (network, power, ...) accessible so that maintenance of nodes can be made very easily : the only operation required to access all those slots is to open the rack door. A picture of an assembled blade is presented figure 3.

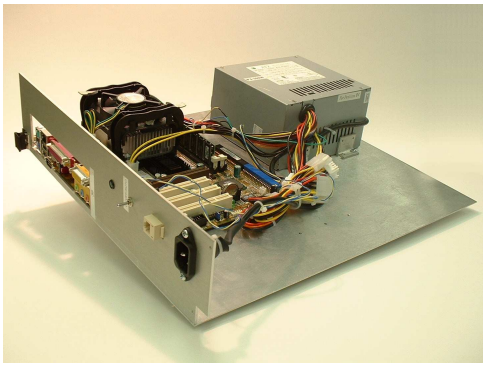

Figure 3. Picture of a Elix2 blade

2.1.2 Rack Design

As mentioned previously, RITTAL helped us to design the rack. Each rack holds 3 stages of blades that are then placed back to back. The total capacity of a rack is then 36 (3 stages, 6 by stage, back to back) with room available at the bottom of the rack for network hardware (switch HP4000 , 40 100 Mb ports with GB up-link). We use hardened power-bar (8 outlets, 20A) to power all the nodes and the switches. A picture of second stage of the first rack is presented figure 4.

Figure 4. Picture of the Elix-2 Rack

2.1.3 *Out of band node power management*

This particular feature, developed with the CEGEP of Sherbrooke [16], allows us to remotely switch the power of any given node. In order to have this control on remote nodes, two contacts have to be made on the power supply. Then, each node can have three power states : automatic (i.e. controlled by the out of band node power management system), always on or always off (i.e. no out of band node power management). As a consequence, it is possible to switch on or off a node manually (for security reason) or remotely with the out of band control system. At the time of this writing, the out of band management system is a prototype and has to be installed on the production cluster. However, all power supply has been linked to the control system so that the only requirement on installation will be to plug the out of band control system into the nodes slots.

2.2 Software organization

All the installation process is based on OSCAR [3] and the server is strictly identical to a regular OSCAR server. We were very pleased to use the System Installation Suite [17], which is a core component of OSCAR, in order to build nodes images. Of course, we did not use the regular installation process, we wrote a script that allows us to transform a regular image into a minimal ramdisk, that is then transferred to the node via NFS. The detailed procedure of how to build a root ramdisk and even a root-raid-ramdisk is explained in another article presented at this conference [18].

2.2.1 *OSCAR diskless : proof of concept*

One of the big advantage of diskless clusters is, of course, the removal of the only mechanical component of a regular system : the hard disk. The removal of a moving part, which is one of the major cause of node failure in regular clusters, provides the global cluster with a better availability. Another major advantage of diskless and systemless cluster is the easier manageability of the whole cluster. All the files are located on a so-called master node and are transferred via the network. Update of nodes can be made « live » from the master node and all gets updated at the same time (reboot of node is only necessary when updating kernel or because of power failure). It is harder for entropy to grow in diskless or systemless clusters because all the nodes share exactly the same image. In clusters with disks, it is necessary to rebuild nodes frequently to make sure that all nodes are up, running and responding when doing any update (we had some bad experience of this entropic behavior with our IBM-SP3 of 64 nodes).

2.2.2 *thin-OSCAR working group : design an implementation*

On the old Elix cluster, there was no bundled software solution that makes an easy way to install, use and maintain diskless clusters. Now, the team uses thin-OSCAR as the development framework for Elix II. This solution is now available for the advantage of all the open source and scientific community. See the thin-OSCAR web site for additionnal information.

Acknowledgments

We would like to thank Intel for their generous donation and technical help. The RITTAL company was very helpful for the design and implementation of our home-made blade and specific rack needs.

References

1. Google Search Engine
`http://www.google.ca/press/highlights.html`
2. Sherbrooke University Centre de Calcul Scientifique
`http://ccs.usherbrooke.ca/`
3. Open Source Cluster Application Resource (OSCAR)
`http://oscar.sf.net/`
4. thin-OSCAR work-group. Support for diskless and systemless cluster for OSCAR
`http://thin-oscar.ccs.usherbrooke.ca/`
5. NFS-Root mini-HOWTO
`http://www.tldp.org/HOWTO/mini/NFS-Root.html`
6. TFTP standard (RFC 1350)
`ftp://ftp.rfc-editor.org/in-notes/rfc1350.txt`
7. B. Kyung, S. Allen, A.-M. S. Tremblay, Pairing fluctuations and pseudogaps in the attractive Hubbard model, Phys. Rev. B 64, 075116/1-15, 2001
8. B. Kyung, J.S. Landry, D. Poulin, A.-M.S. Tremblay, Comment on "Absence of a Slater Transition in the Two-Dimensional Hubbard Model, submitted for publication `cond-mat/0112273`
9. B. Kyung, J.S. Landry and A.-M.S. Tremblay, "How antiferromagnetic fluctuations both help and hinder d-wave superconductivity", submitted for publication
10. A. Blais, "Quantum network optimization" Phys. Rev. A 64, 022312 (2001).
11. A. Blais and A.-M.S. Tremblay, Effect of noise on geometric logic gates for quantum computation, Accepted in Phys. Rev. A
12. D. Sénéchal, D. Perez and M. Pioro-Ladrière, The spectral weight of the Hubbard model through cluster perturbation theory, Phys. Rev. Lett. 84 (2000) 522-525.
13. D. Sénéchal, D. Perez and D. Plouffe, Cluster Perturbation Theory for Hubbard models, Phys. Rev. B 66, 075129 (2002)
14. Exact diagonalization methods for quantum systems H.Q. Lin and J.E. Gubernatis Computers in Physics, vol. 7, no 4, Jul/Aug. 1993, p. 400.
15. RITTAL Canada : web site
`http://www.rittal.ca/`
16. Remote Power Management of cluster nodes,
to be published.
17. System Installation Suite website
`http://www.sisuite.org/`
18. Root Raid in Ram How-To, Mehdi Bozzo-Rey, Michel Barrette, Benoît des Ligneris submitted to HPCS2003 (2003).