Computing facilities for small physics analysis group

Damian Reynolds[1] [dlreynol@rcf.rhic.bnl.gov]

Andrey Y Shevel[2] [Andrey.Shevel@pnpi.spb.ru]

PHENIX technical note

Abstract

The small physics group (3-15 persons) might use a number of computing facilities for the production analysis/simulation, developing/testing, teaching. The most recent instance of the cluster for Nuclear Chemistry Group at **SUNYSB** is briefly described. It is discussed the features of different types of computing facilities: collaboration computing facilities, group local computing cluster, cloud computing. The author(s) emphasize the growing variety of different computing possibilities including recently emerged and growing role of the group owned computing cluster of micro size.

Introduction

Usually members of a physics group have computer accounts on large computing facilities which are supported by the physics collaboration. Such the facilities have certain rules: who can the access to the computing installation in which scale, and for which purpose.

From time to time small physics group needs something different - more agile and flexible computing infrastructure. The small physics group might demand for more or less centralized computing facilities under group control for several purposes:

- to keep common data (papers, drafts, programs, fraction of experimental data, etc);
- to test new/modified simulation or/and analysis software/algorithms;
- to give the account for short time visitors/students who needs to do something in analysis;
- any other possible requirements, in particular as good gateway for remote large computing cluster(s).

Obviously such small computing installation is used to be the good complement to large computing facility.

The computing needs can be considered in various ways [from point of view of the small group]:

- to use big[3] centralized cluster (here we mean collaboration cluster)
  - advantages:
    - everything is done, no problems for the group in hardware and raw maintenance (security, user registration, etc)
  - disadvantage:
    - relatively long registration process (depends on the cluster and organization) and obtaining permission to use computing resources (disk space, etc);

1 State University of New York (campus Stony Brook)
2 Petersburg Nuclear Physics Institute (Russia)
3 The cluster sizes: big, large = more than 1000 machines; middle size = until 1000; small = until 100; micro = O(10).

        - you are under rules implemented on the cluster (those may not be suitable for each person, for example for short term visitor/student);
        - due to a range of limitations of the data access the job turn around time might not be optimized;
        - collaboration cluster performance might be changed significantly due to variation of user job traffic, for example prior a conference;
- cloud computing
    - advantages:
        - it is possible to use as much (almost) computing nodes as you like in concrete day; no worry about hardware and maintenance;
    - disadvantage:
        - you have to pay for that with the policy ***pay as you go***[4].
        - the bandwidth between computing nodes and between computing nodes and storage might not be high as well as other parameters might not be attractive.

- own group local cluster
    - advantages:
        - you can do what you want and how you want; students have a lot of possibilities to try a range of methods how to organize the cluster and how to organize the computing; also it is good point where several students could get experience in collaborative work; it is good for teaching.
    - disadvantages:
        - you need to have all required hardware and install base OS and application software;
        - also you need to organize everything working (including stable power, environment temperature, etc, etc);
        - if interests in the group are split the optimization for one subgroup may not be conducive for the other subgroup;

Here it is assumed that that physics group is using more than one cluster to get the computing task done. Further in this paper we will analyze the own local computing cluster and cloud computing facilities: now and in nearest 2-5 years.

Usually small physics group has limited financial resources. This fact does impose many restrictions on the cluster architecture.

The cluster has to be:

- cheap (consideration on the true cluster ownership cost is in [15]);
- reliable hardware;
- not demanding intensive watching/maintenance.

4 ***Pay as you go*** – the policy when customer has to pay for the really used resources (CPU time, disk memory, etc)

Other requirements – the implication of the desire to decrease the maintenance efforts:

- compatibility (architecture and base OS) with collaboration cluster environment (**RACF** for example);
- same set of application software as on the collaboration cluster.

From above we see that group owned computing cluster is not possible to be large or even midrange, it is quite small = micro cluster. The good configuration of the group owned cluster might consist of 5-15 modern machines (multicore CPUs, 3 GB of main memory per core, 10-20 TB per machine of disk space). Such the group cluster can help to get more flexibility when using several remote computing facilities: collaboration cluster(s), public cloud computing, etc.

The situation in different physics groups might differ from each other. Here we shall discuss the concrete group cluster solution for Nuclear Chemistry Group (**NCG**) at **SUNYSB**/Chemistry.

Local computing cluster

The computing cluster in **NCG** is appeared in 2000 or bit earlier. At that time all the machines (30+) have 512MB of main memory and Dual 500 MHz CPUs. This cluster was used for program development, test analysis, student work, etc. More than 70 registered users and around 5-7 are quite active. Starting from 2005 the cluster was running almost without support (no stable person who was even just watching the cluster). In case of problems somebody came to the room and reboot the cluster. Later the cluster has been partly upgraded/modified. Nevertheless cluster was degrading with the time (machine by machine and disk by disk).

To 2010 the cluster came with around 17 machines of different sizes of main memory (0.5-2.0 GB) frequencies (0.5-3.1 GHz), computing power and production year, fortunately all CPUs are from Intel. The CPU compatibility in our case is good feature indeed, because it permits just to copy many libraries from **PHENIX** as it is to save time and labor. After the cluster access machine and network switches became dead it was decided to reconfigure the cluster in quite short time with available second hand components: machines, network switches, cables, etc. By fortune we were able to recover most of user home directories from the Tape Library Subsystem - Qualstar **TLS**-4660 Tape Library, 4 AIT-2 drives, 63 slots, around 3 TB (production year around 1999 or earlier). In the past such the **TLS** was used as backup device. However after years it requires to be replaced or refurbished which in turn requires human efforts and money. As the result we can not rely on this device in backup procedure. That means no cluster backup is planned anymore. In desire to reduce the volume of maintenance we left in operation as minimum as possible in cluster hardware. Each user on this cluster has at least several opportunities to keep backup copy of his home directory (or critically important data) in the collaboration computing facilities (**RACF**), who has an account, or keep the copy in cloud memory [2-4].

To reduce downtime for the cluster it is good to buy and install special equipment **KVM switch over IP** [23] to do many control actions (switching **on** and **off** of any machine in the cluster, get access to the console of any machine, etc) remotely over Internet. In another words the group might use remote help from external experts. However in our case the idea is not implemented yet.

Each machine in cluster has at least two Ethernet ports, so we are able to use two network segments: one for external network and one for private network. External network segment is used for **SSH**, **NIS**, batch control protocols. Private segment is used for **NFS**. The **NFS** mounted areas are used for data and home directories which are located on different disk drives connected to different machines. Such the solution decreased the probability of congestion (bottleneck effect).

The cluster has exactly same directory tree as **PHENIX** has. From time to time the directories from **PHENIX** are mirrored to the cluster directory tree. **AFS** is not in use in the cluster.

Mirror procedure is very simple

rsync -Rrltvz --rsh=ssh shevel@rftpexp01.rhic.bnl.gov:/afs/rhic.bnl.gov/phenix/PHENIX_LIB/sys/i386_sl5 /data04/Ram/

rsync -Rrltvz --rsh=ssh shevel@rftpexp01.rhic.bnl.gov:/afs/rhic.bnl.gov/i386_sl5/opt/phenix /data04/Ram/

rsync -Rrltvz --rsh=ssh shevel@rftpexp01.rhic.bnl.gov:/afs/rhic.bnl.gov/x8664_sl5/opt/phenix /data04/Ram/

rsync -Rrltvz --rsh=ssh shevel@rftpexp01.rhic.bnl.gov:/afs/rhic.bnl.gov/rcassoft/x8664_sl5/cernlib /data04/Ram/

On each computing node we have

/afs -> /data04/Ram/afs

It gives us the same view for directory tree (and content) as in **PHENIX**.

As the batch system we use pair of torque/maui from http://www.supercluster.org.

Due to security reasons (remember, the cluster has no regular maintenance) the cluster is available from only specifically defined network domains.

To keep some knowledge about the cluster we have prepared several things:

- short description for system administration (in google doc);
- short description of the used hardware (in google doc);
- the table of cable interconnections (in google doc);
- mailing list in google for all active users (now 17 users);
- several web pages on google [https://sites.google.com/site/ramdata2009/].

Above information is very useful when you need to move the cluster in another room/location [we experienced with such the moving needs several times].

Because the cluster is located in relatively large room with good ventilation there is no needs for air conditioner. After years of experience we found that the University electric power grid is quite stable [1-3 unplanned interruption per year]. On other hand earlier used **UPS**s require the change of batteries quite often (at least every two years). In small group without regular watching such the replace is not feasible. Obviously from time to time the cluster suffered from loss of electric power. It was observed that no data file was lost due to such outages.

The basic **OS** (Scientific Linux with same RPM set as on **RACF**) installation procedure and basic configuration is semiautomatic: there are couple of scripts with use of kickstart as initial step and another step consisting of bash script for post kickstart configuration. We do not use any virtualization feature.

With the time the cluster became interesting for more than one small group at **SUNYSB**/Chemistry.

Finally two students were agree to help each other to watch the cluster (when they have the time) and help to other users.

Everything what was done can be considered in some assumptions and in according with modern terminology as volunteer cloud of type **PAAS** [**P**latform **A**s **A** **S**ervice].

To show cluster status for volunteers we used simple monitoring system – **Ganglia** [13]. The ganglia server has been deployed on separate computer outside the cluster. Several other services like web server, twiki (not only for cluster) were created outside cluster. In general our intention was to distribute among independent computers almost all services to decrease the number of single point of failures (**SPOF**). Even if cluster is completely down we are able to find out information about the cluster behaviour with information available on **Ganglia**.

Why we did not use available packages like **Rocks** [14] in this cluster? Several reasons were taken

into account:

- we planned to keep our cluster as close as possible to **PHENIX** directory structure and to set of **RPM** packages which are installed on **RACF**;
- **Rocks** uses in particular local **DHCP** server (in the cluster). This fact is not hailed by university network administrators especially for small computing systems which are not maintained on regular base; in cluster computers we carefully removed all components like **DHCP**, **DNS** service, etc. Minor misconfiguration in this area might lead to serious and unexpected troubles in completely different part of the university network.
- also history reasons were taken place (the cluster had set of bash script to perform required configuration before other options like **Rocks** became available).

Remarkable step for the cluster running was creating discussion group (mailing list) for cluster users on Google.com. Such the location for discussion mailing archive is quite reliable and may be most natural choice for small physics group. Such the mailing feature is used to keep people informed for everything happened around cluster: new installations, problems, advices, etc. Obviously users do help each other sending the information in this mailing list. In our circumstances the users mailing list does form king of thinking engine for various methods how to use the cluster for concrete tasks. Fortunately such the mailing list and mailing archive does require no attention to keep it up (at least up to now). Using the Google.com for mailing lists we do not care where exactly our mailing archive is located therefore we often say that it is located somewhere in ***cloud***.

The cloud computing is hot topic in **IT** last 2-3 years. Many successful experiments with clouds have been performed [1, 16, 18, 20]. Obvious question is how group owned cluster does relate to cloud computing? May be in nearest 2-3 years the cluster of micro size will not help to small physics group? First of all we try to take more close look for cloud computing.

Cloud computing

The **cloud computing** is quite not trivial paradigm which has a lot of instances in government and private sectors. The quote below is part of **cloud computing** definition I copied from [11].

Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

That was just beginning of the definition[5] but it gives main idea. The reconfiguration of the computer system resources on the fly with web **API** became possible after virtual machines architecture (**VM**) **CPU**s together with virtualization software tools like **Xen [**21**]**, **K**ernel-based **V**irtual **M**achine **(KVM) [**22**]**, and other come into reality.

Let us be more specific when talking on the computing cloud service from user's point of view. There are many models to use cloud computing, in particular: private/corporate cloud computing, public computing. At this point we will consider so called ***public cloud computing*** – cloud computing which is available to almost anybody in Internet. Also there are different kinds of features available in public computing clouds, for example **Amazon** ec2/s3 [6,7].

May be simplest instance for several public clouds is sync service, where your local file or directory is in automatic sync with some cloud location. Usually physicist has more than one computer

5 Whole definition is explained in two pages or so.

(desktop, laptop, communicator, etc). It is very suitable when we have same directory for all computer devices. Technically that was and is possible with old **AFS,** for example. However to keep the same content for the directory with heavy and complicated **AFS** service (security, **AFS** servers infrastructure, etc) is not easy. For specific data (e.g. browser bookmarks) more interesting to keep all your computers in sync with the tool like **Xmark** [8] - specific type of cloud service. Also you might need to keep in sync relatively small fraction of the data with the tool like **Dropbox** [2]. Not bad to tell that minimum use of the service is free of charge. Mentioned tools (i.e. public space clouds) are very useful even you have just one laptop. The laptop might become broken or lost, but your data will be kept on mentioned clouds and you can use them from another computer.

I described several public clouds which are very popular among many users. However many tens of similar services are available with little difference in character and style of service with different policy to pay for the service, e.g. [3,4].

Some physicists are afraid to use public cloud computing service because the public cloud is out of their control (for instance the service could be down forever due to business or/and political issues). That is true. On other hand we can consider the control capability as the reliability of the access to the cloud. Can we think that public cloud service is 100% reliable all the time (including risks rooted in business)? The answer is **no.** Unfortunately **w**e have to say the same about any other instance of computing service of any kind. At the same time the small groups do have often not so reliable local computing which depends on unstable enthusiast activity. In many cases for even middle term time frame (2-5 years) local computing service is most probably less reliable than public cloud computing service. If you are worrying for the reliability of you data being safe - the obvious conclusion is to use both.

Among other public cloud services we can pay attention on ability to keep in cloud not only mailing lists, but any type documents, web pages, wiki pages, etc [10]. Not surprising that even small **Twiki** servers might migrate to **Google** [19]. The clouds are also used for normal computing as well.

Several successful testbeds with using the cloud computing for production simulation in **HEP** have been carried out, e.g. **ATLAS** [1] and **STAR** [16] (latter work has many deep and smart observations of the experience with computing grid and cloud computing architectures). The success does depend on a lot of details, in particular on the computing infrastructure components and theirs parameters which are “under hood” of computing cloud. In work [17] authors were urged to do additional conversions of **VM** images, may be due to the lack of the open standards in the field. In other cases [16,18] authors do find that tested public cloud has not so good computer hardware parameters as they expected. Recently mentioned works are dealing with consideration of high scale computing which is out of scope of the paper, but this can be used as more advanced example. Obviously if it is quite feasible for large scale computing it is really possible for relatively small computing needs. Also it has to be taken into account computing cloud initiatives and plans in government [5].

When you use public cloud computing facility with policy ***pay as you go*** it is possible to save money if you use cloud computing only when it is really required. For example, before important conference you might need to do a lot of simulation with 1000 or more jobs in parallel. That means you might rent in cloud 1000 or more CPU cores and pay for that. However after the conference is ended you can download all cloud configuration to group owned computing cluster and keep it until next conference. By this you are freeing public cloud i.e. you do not pay anymore for the resources. Before next conference you can upload your cloud configuration from the group owned cluster to public cloud and use it again for same or corrected job runs.

In other words the group owned cluster is used in described scenario as important gateway to public cloud computing. The number of public cloud computing instances is growing significantly each year. That means the importance of suitable gateway to different clouds for small physics group is

growing as well.

Conclusion

The small computing/information installations are already on the way to use the clouds. Significantly more is coming. The moving to the cloud does eliminate for small physics group cluster hardware maintenance activity, but not application software and data structure maintenance. Also to achieve maximum effect of using the cloud you can not ignore good understanding of cloud hardware and OS architecture.

With using more than one remote computing resource the importance of the group owned cluster of micro size to keep the fraction of the data, program libraries, configuration files, etc. is growing.

References

1. Jan-Philip Gehrcke et al, ATALS@AWS, http://iopscience.iop.org/1742-6596/219/5/052020
2. Dropbox http://www.dropbox.com/
3. A Drive http://www.adrive.com/
4. WindowsLive http://live.com/
5. Vivek Kundra, U.S. Chief Information Officer, “Federal Cloud Computing Strategy”, http://www.cio.gov/documents/Federal-Cloud-Computing-Strategy.pdf
6. Amazon Elastic Compute Cloud (Amazon EC2) - http://aws.amazon.com/ec2/
7. Amazon Simple Storage Service (Amazon S3) - http://aws.amazon.com/s3/
8. Free Bookmark Sync - http://www.xmarks.com/
9. Cloud computing at TeraGrid - http://www.rcac.purdue.edu/teragrid/resources/
10. Google facilities – http://www.google.com
11. NIST definition of Cloud Computing - http://www.nist.gov/itl/cloud/upload/cloud-def-v15.pdf
12. Red Hat Cloud Foundations - http://www.redhat.com/solutions/cloud/foundations/
13. Ganglia - http://ganglia.sourceforge.net/
14. Rocks - http://www.rocksclusters.org
15. The True Cost of HPC Cluster Ownership - http://www.clustermonkey.net//content/view/262/1/
16. Jerome Lauret et al. *From grid to cloud: STAR experience* http://computing.ornl.gov/workshops/scidac2010/papers/data_j_lauret.pdf
17. Jerome Lauret et al. *Contextualization in practice: the Clemson experience* http://pos.sissa.it/archive/conferences/093/027/ACAT2010_027.pdf
18. Keith R. Jackson et al. *Performance analysis of high performance computing applications on the Amazon Web Services Cloud* http://www.lbl.gov/cs/CSnews/cloudcomBP.pdf
19. Migrate your Twiki to Google Sites (using Google Sites API and Perl) http://blog.famzah.net/2010/05/30/migrate-your-twiki-to-google-sites-using-google-sites-api-and-perl/
20. **HEPiX** Fall 2010, Cornell University, Ithaca, New York, 1-5 November, 2010

http://cdsweb.cern.ch/record/1307061/files/HEPiX%20Trip%20Report%20Fall%202010.pdf

21. **Xen** http://www.xen.org/
22. **KVM** http://www.linux-kvm.org/page/Main_Page
23. **KVM switch** http://en.wikipedia.org/wiki/KVM_switch